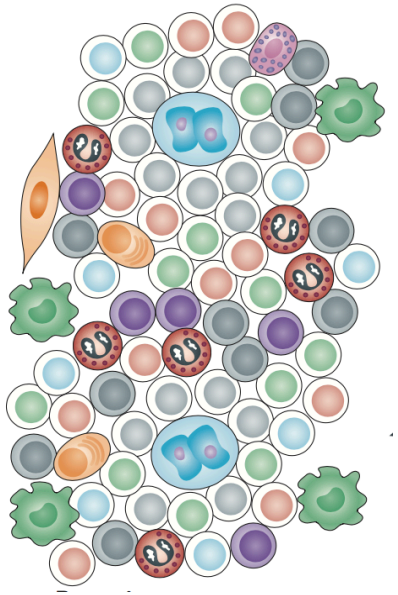
Deconvolution methods to dissect immunologic profile of the tumor

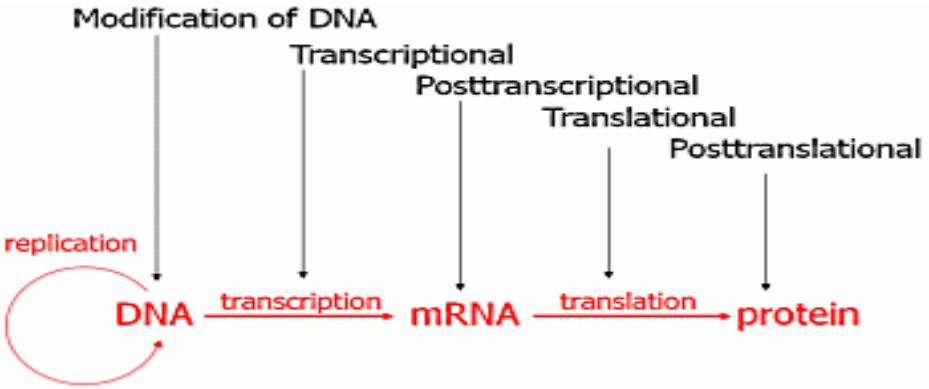
Petri Pölönen, M.Sc
University of Eastern Finland
Institute of Biomedicine, school of medicine
Bioinformatics-Systems genomics group (Merja Heinäniemi PhD)

UNIVERSITY OF EASTERN FINLAND

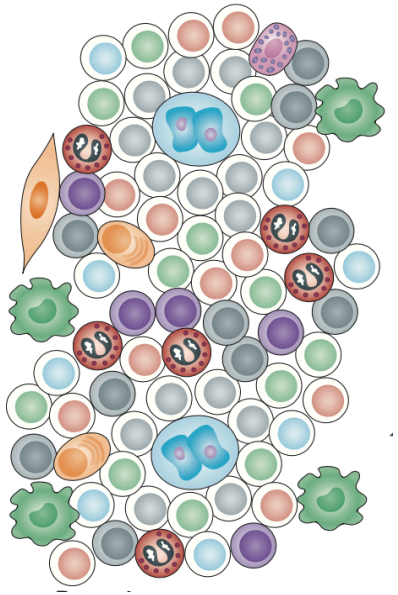
“SOURCE separation, or deconvolution, is the problem of estimating individual signal components from their mixtures. This problem arises when source signals are transmitted through a mixing channel and the mixed sensor readings are observed.”



“SOURCE separation, or deconvolution, is the problem of estimating individual signal components from their mixtures. This problem arises when source signals are transmitted through a mixing channel and the mixed sensor readings are observed.”



RNA-seq
microarray



All cells have same DNA, but they have different mRNA and protein levels

mRNA expression

Samples

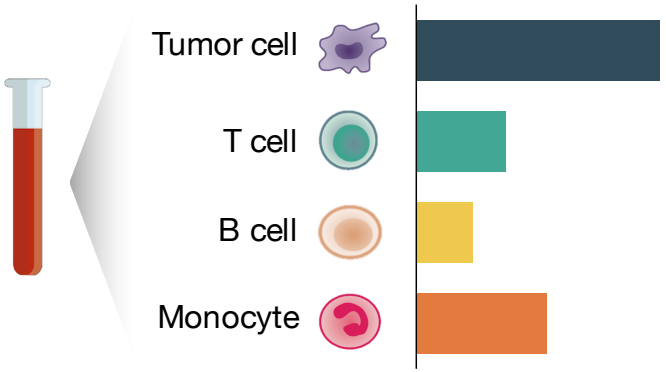
Genes

Deconvolution methods

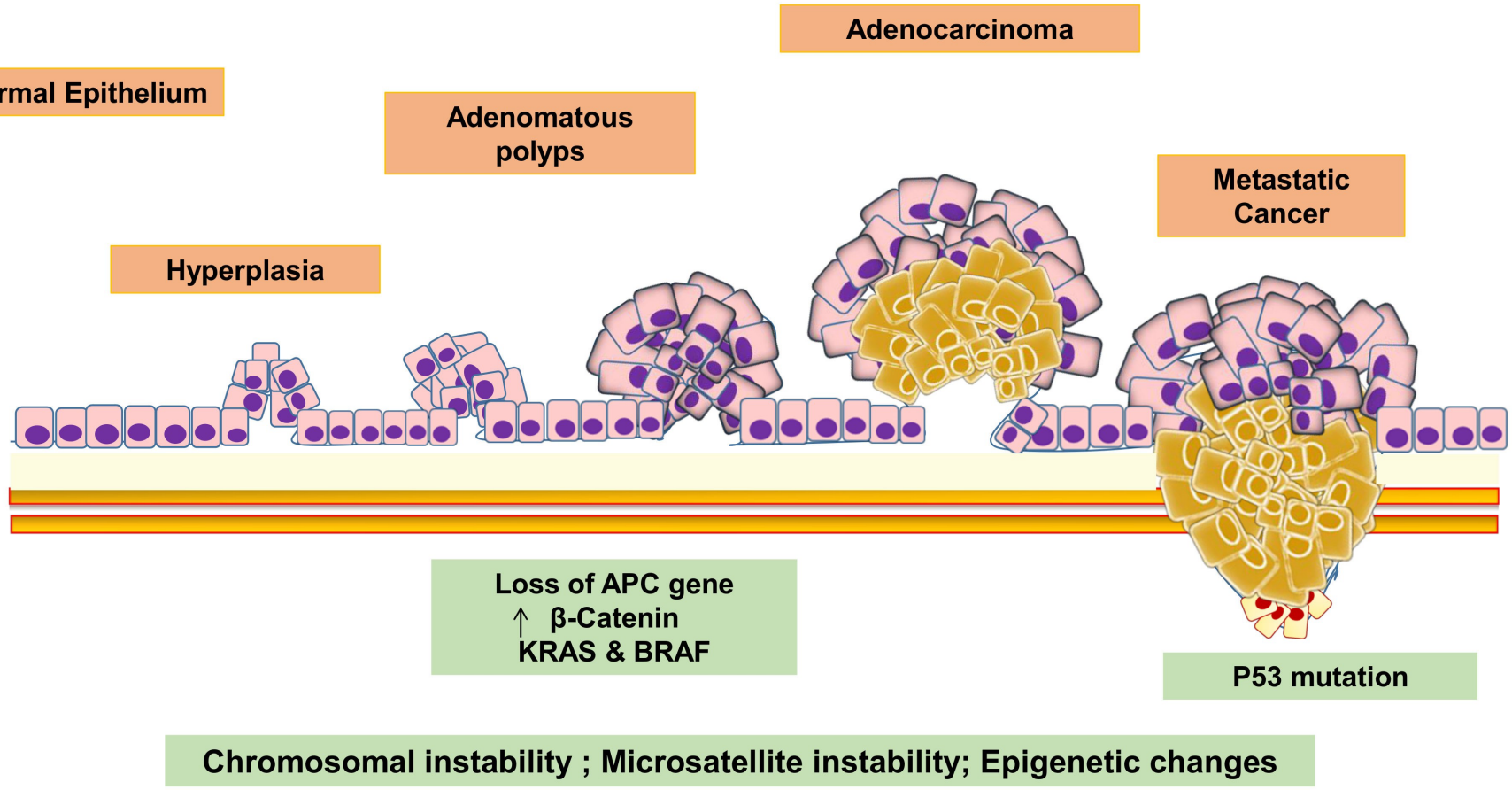
Intuition

- Models assume *linearity*: the expression signature of the mixture is a weighted sum of the expression profile for its constitutive cell-types →approximation that is based on the linearity assumption
- Seek to distinguish features (genes) that closely conform to the linearity assumption, from the rest of the genes.

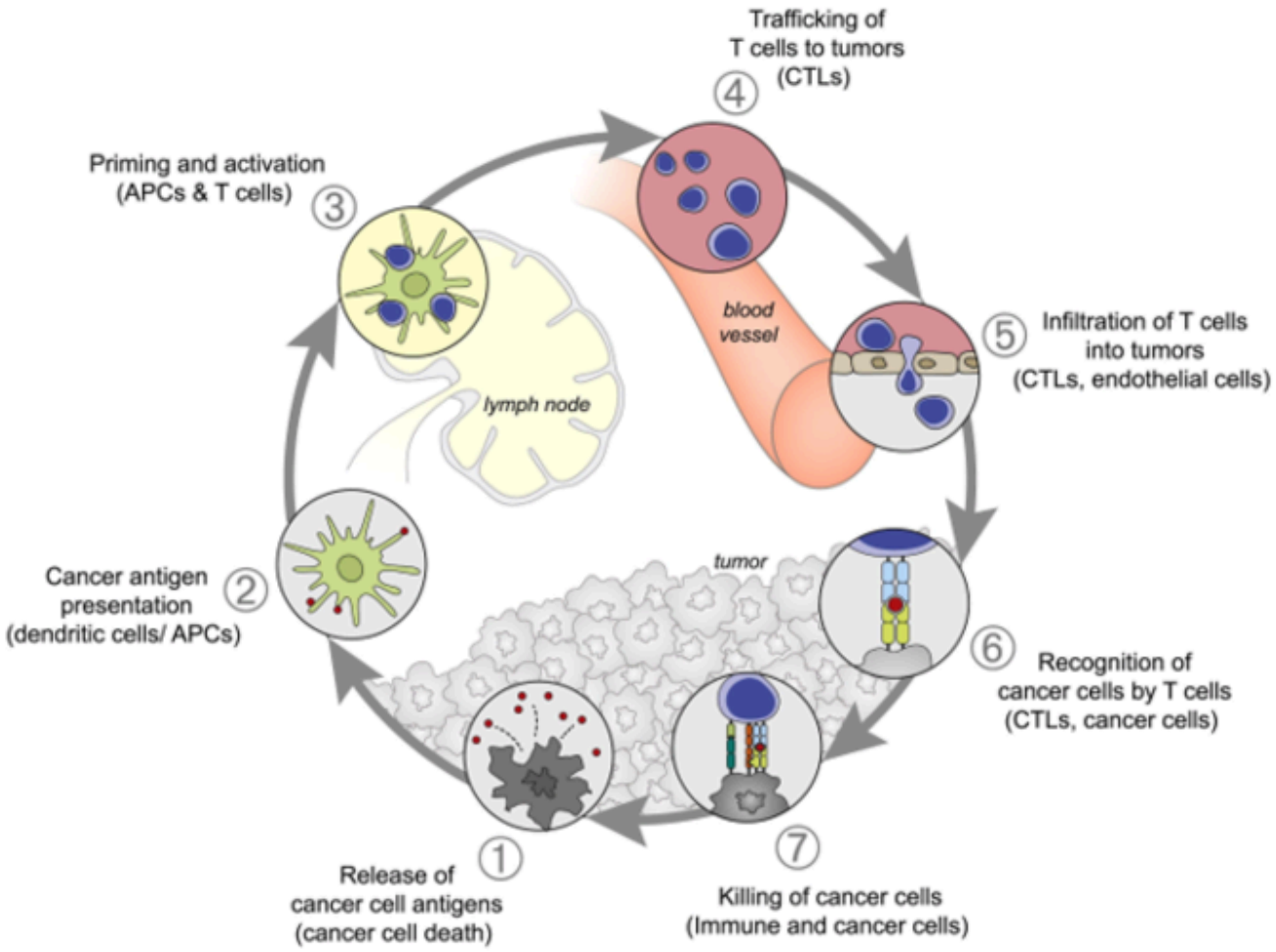
Mix:
90 % Cells tumor
3 % Cells T-Cells
1 % Cells B-Cells
6 % Cells monocyte
Total 100%



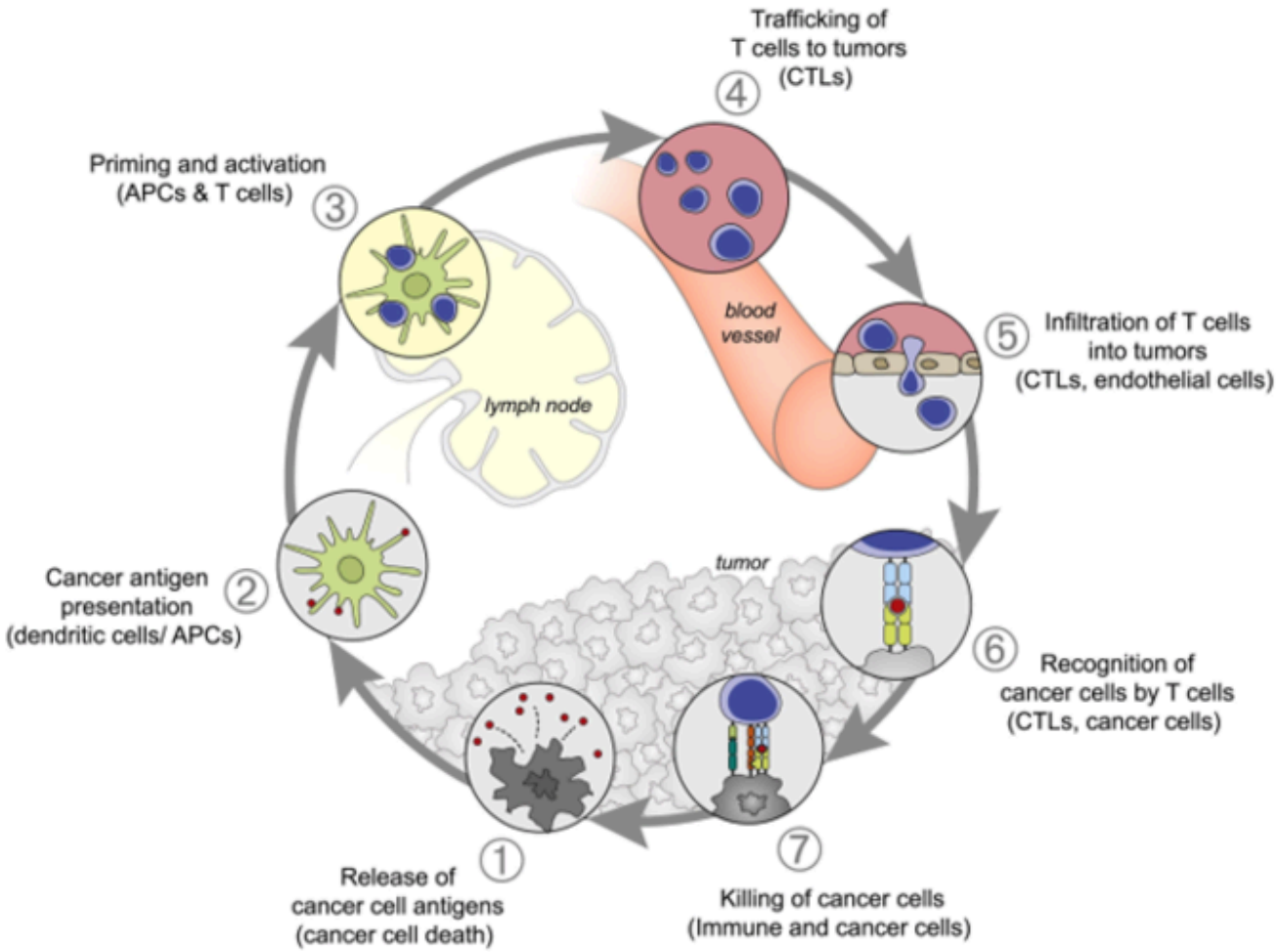
Cancer development

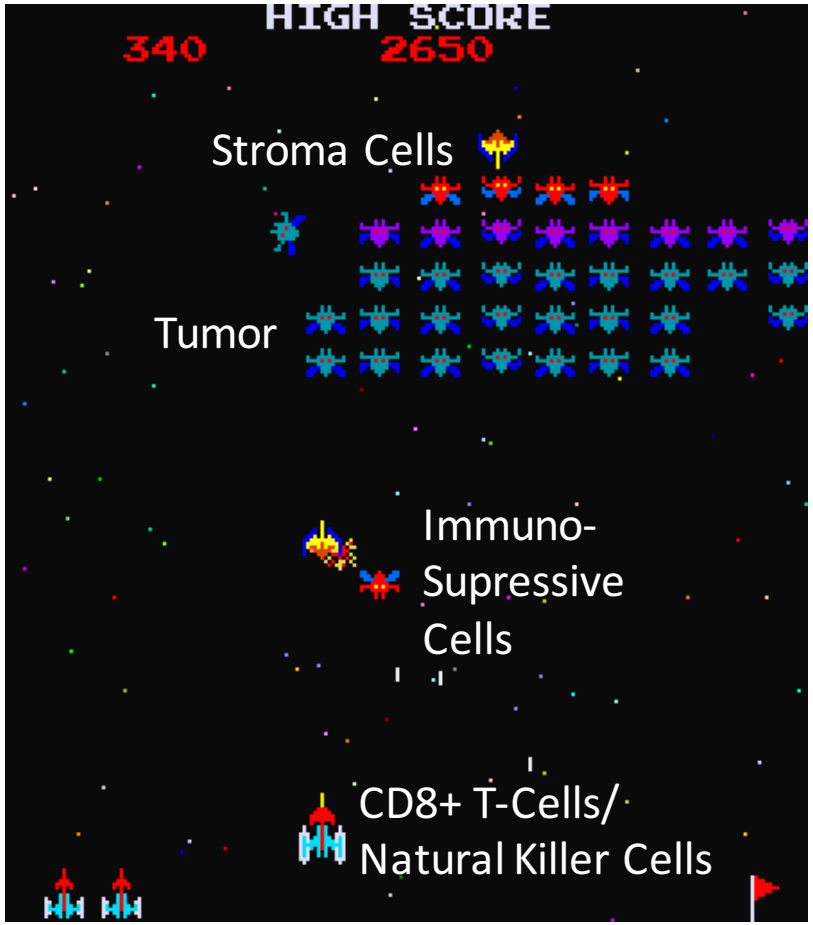


Tumor microenvironment matters!

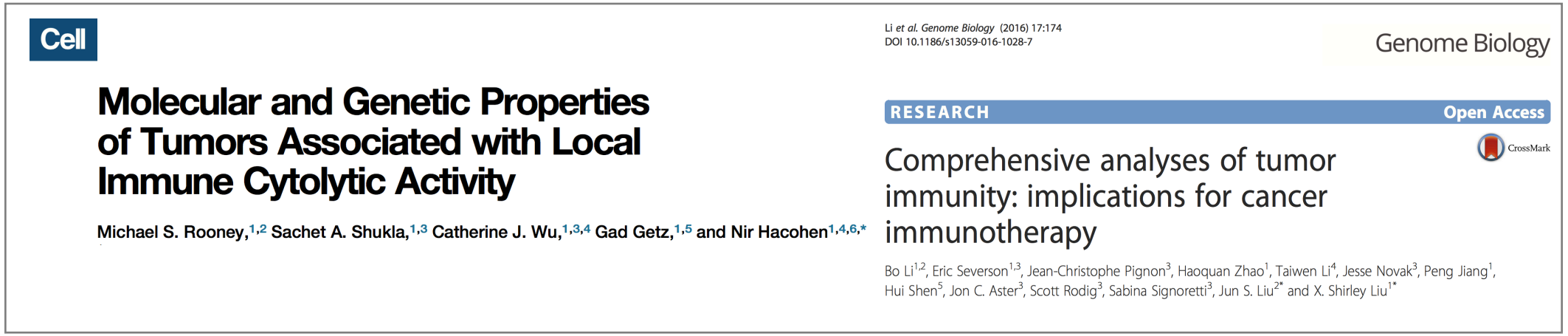


Tumor microenvironment matters!





Why is immunology so hot right now?



Cell

Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity

Michael S. Rooney,[1,2] Sachet A. Shukla,[1,3] Catherine J. Wu,[1,3,4] Gad Getz,[1,5] and Nir Hacohen[1,4,6,*]

Li *et al. Genome Biology* (2016) 17:174
DOI 10.1186/s13059-016-1028-7

Genome Biology

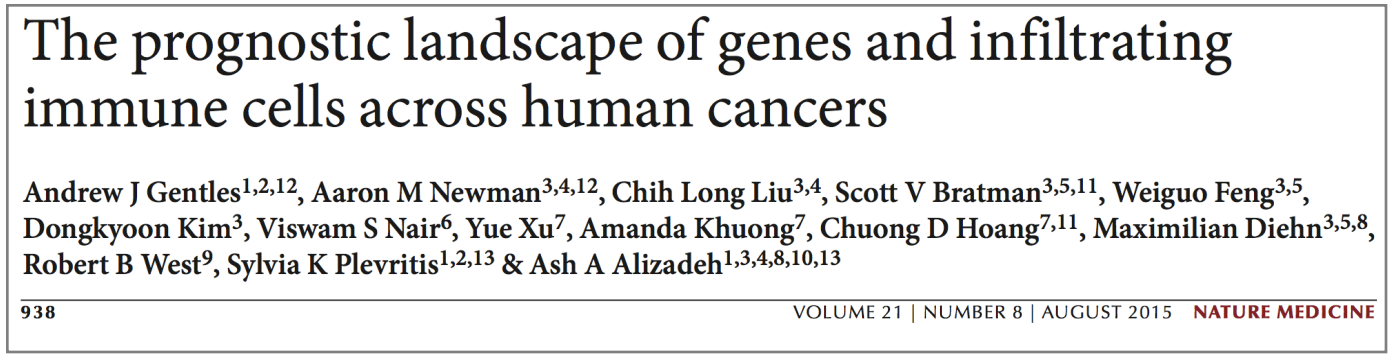
RESEARCH Open Access

CrossMark

Comprehensive analyses of tumor immunity: implications for cancer immunotherapy

Bo Li[1,2], Eric Severson[1,3], Jean-Christophe Pignon[3], Haoquan Zhao[1], Taiwen Li[4], Jesse Novak[3], Peng Jiang[1], Hui Shen[5], Jon C. Aster[3], Scott Rodig[3], Sabina Signoretti[3], Jun S. Liu[2*] and X. Shirley Liu[1*]

Associations to genetic properties (mutations, mutation load, copy number)



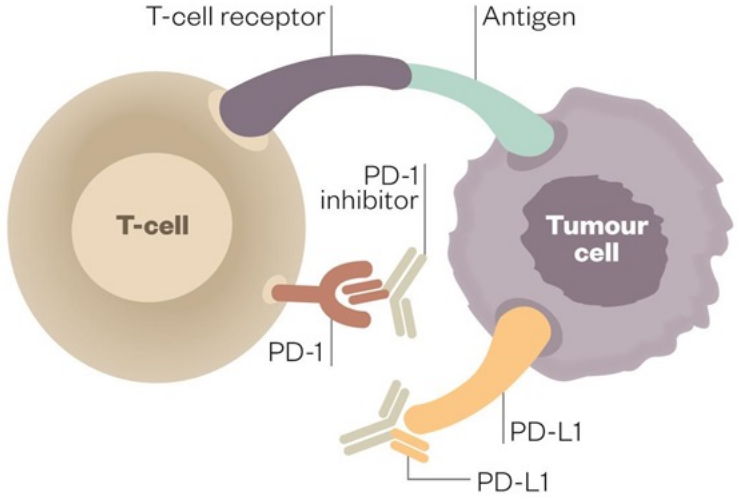
The prognostic landscape of genes and infiltrating immune cells across human cancers

Andrew J Gentles[1,2,12], Aaron M Newman[3,4,12], Chih Long Liu[3,4], Scott V Bratman[3,5,11], Weiguo Feng[3,5], Dongkyoon Kim[3], Viswam S Nair[6], Yue Xu[7], Amanda Khuong[7], Chuong D Hoang[7,11], Maximilian Diehn[3,5,8], Robert B West[9], Sylvia K Plevritis[1,2,13] & Ash A Alizadeh[1,3,4,8,10,13]

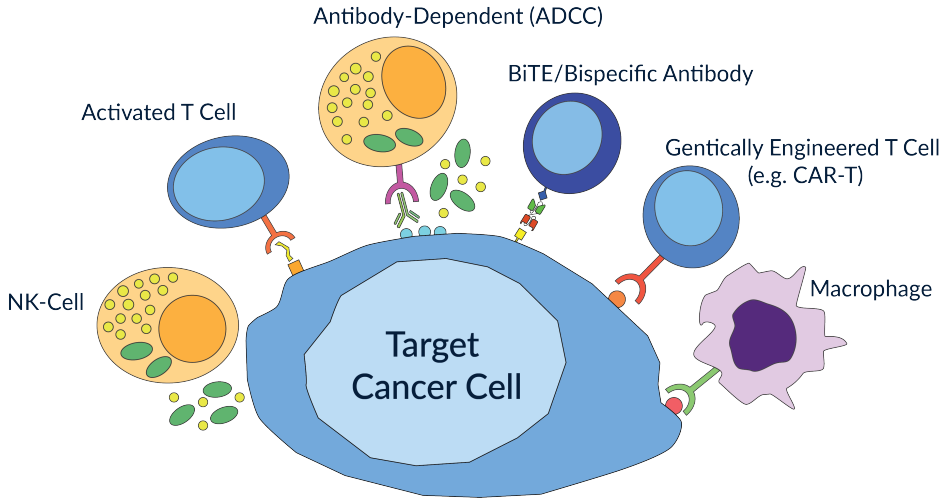
938 VOLUME 21 | NUMBER 8 | AUGUST 2015 NATURE MEDICINE

An increasing proportion of malignant cells, as well as a growing fraction of tumor infiltrating lymphocytes compared to surrounding cells, directly influence tumor growth, metastasis, and clinical outcomes for patients

Why is immunology so hot right now?



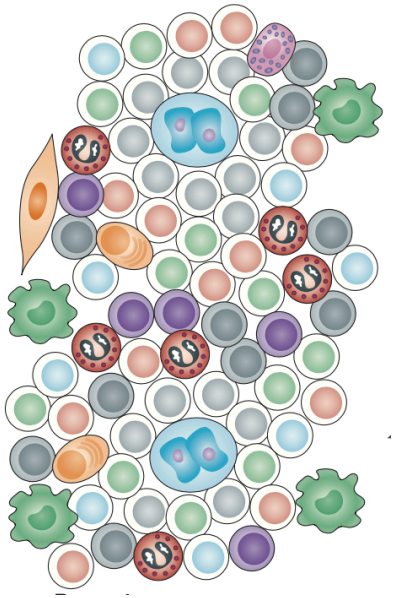
Make T-cells more efficient in killing tumor cells

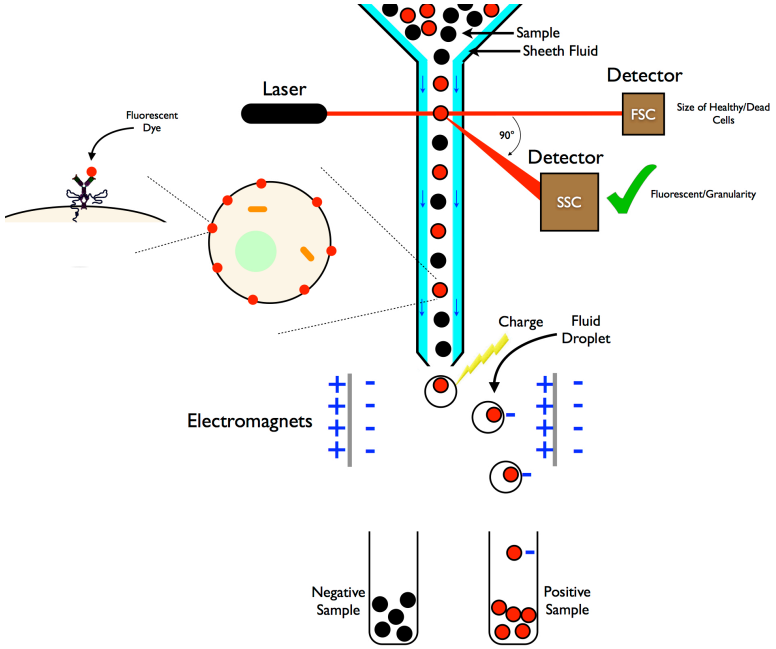


Make T-cells that are engineered to kill cancer cells

Profiling immune cells using **proteins**: Cell sorting

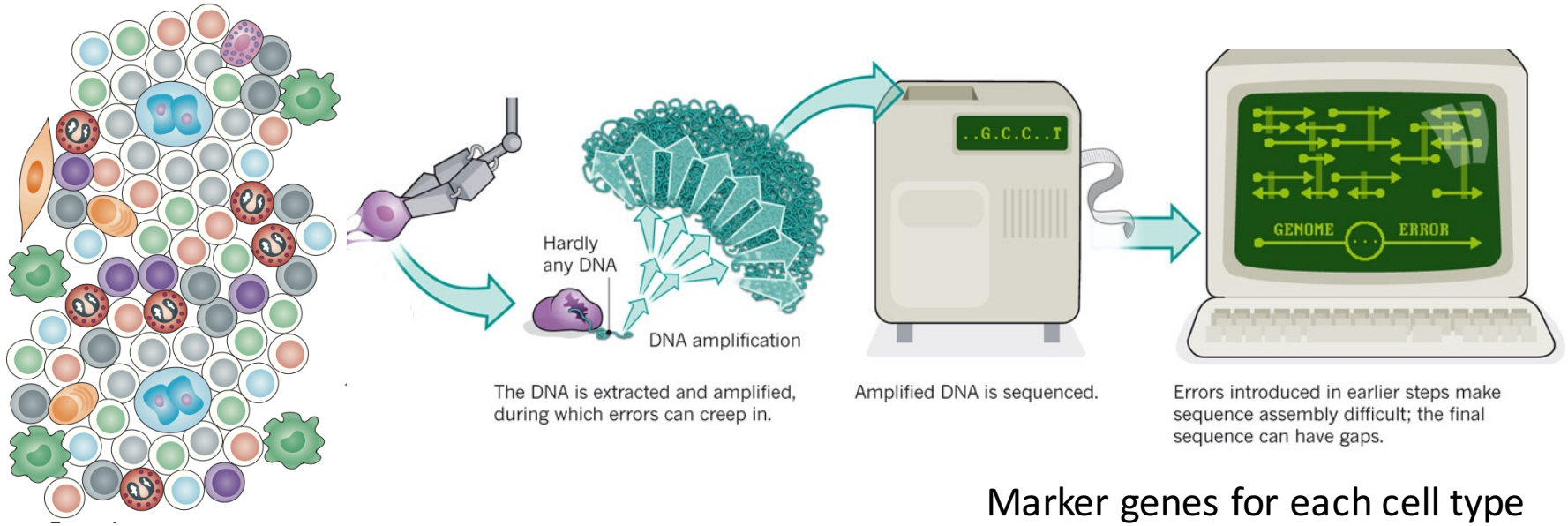
- Cell isolation entails a loss of system perspective and is not currently feasible for all cell subsets





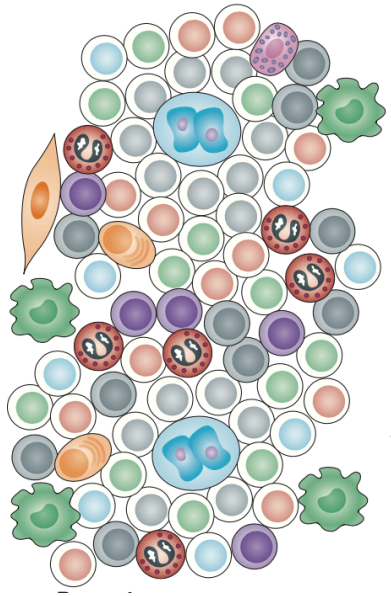
FACS: Depends on surface markers, max 8 markers at a time

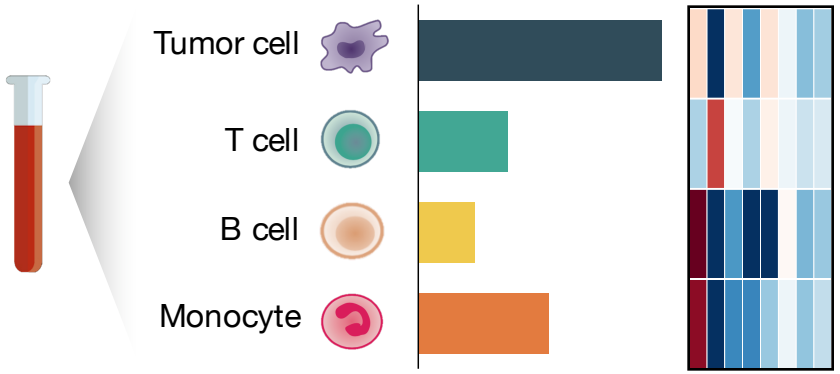
Profiling immune cells using single cell **mRNA** sequencing



Marker genes for each cell type
Technical artefacts
Markers not always detected accurately

Profiling immune cells using **bulk mRNA** sequencing, **deconvolution** to infer fractions



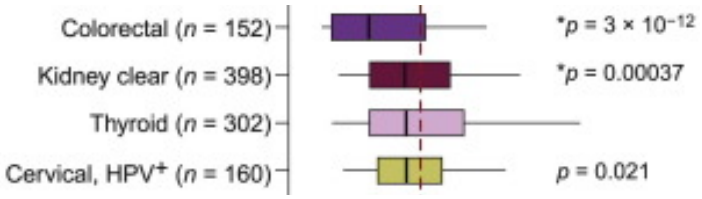


Data exists, cheap, data analysis straight forward

•Computational deconvolution methods can provide a cell-centered system perspective

Methods

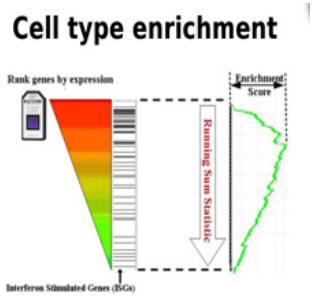
Statistics



Geometric Mean

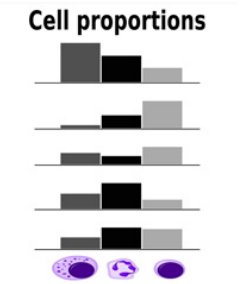
Marker based

Unsupervised Learning



Gene set based

Supervised Learning



Deconvolution

Ease of result interpretation

Complexity, Computation time, Assumptions made about the data, Amount of training data needed

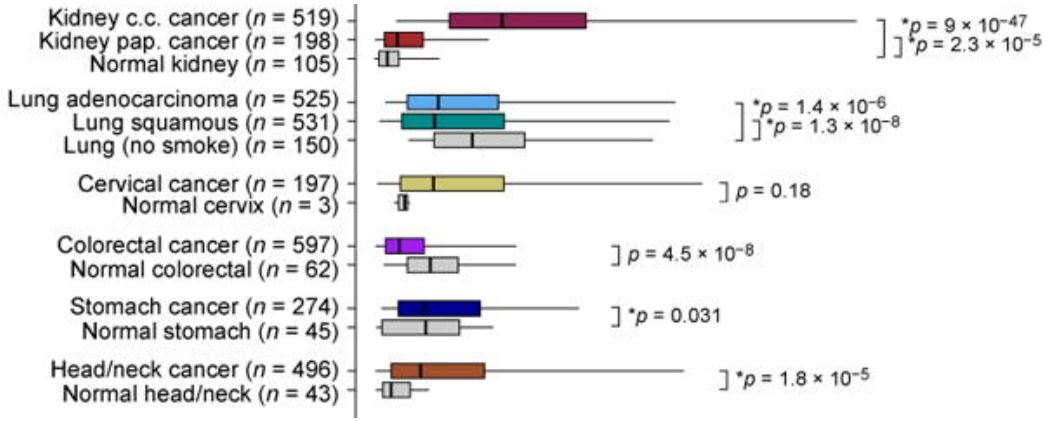
Amount of data transformed and distorted

Marker gene based methods:

- Geometric mean:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

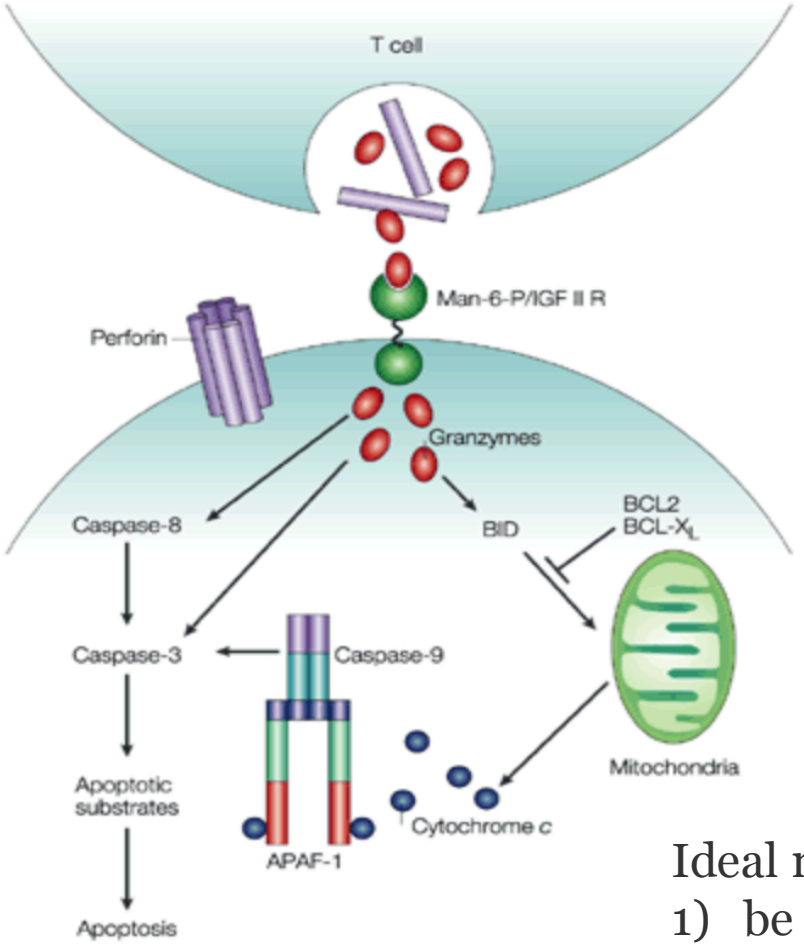
- Subjective to noise
- Gene selection can be biased, depends on previous knowledge

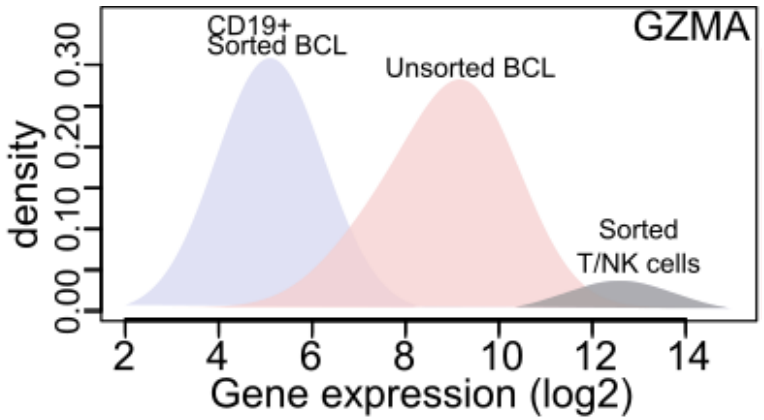


Cytolytic activity (geometric mean of GZMA-PRF1) in solid tumors

Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell. 2015;160(1–2):48–61.

Making holes in the cancer cells, or starting programmed cell death of cancer cells





Pölönen & Dufva et. al (unpublished)

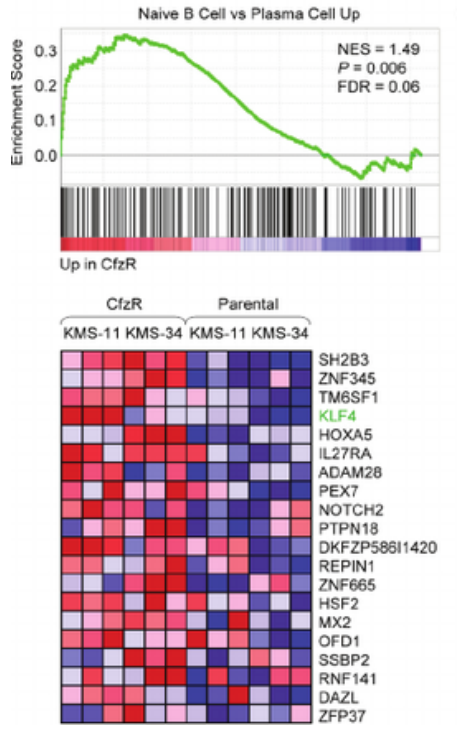
Ideal marker:

1) be highly informative of the cell type in which it is expressed
2) shows low variance due to spatiotemporal changes in the environment (changes in time, microenvironment or experiment)

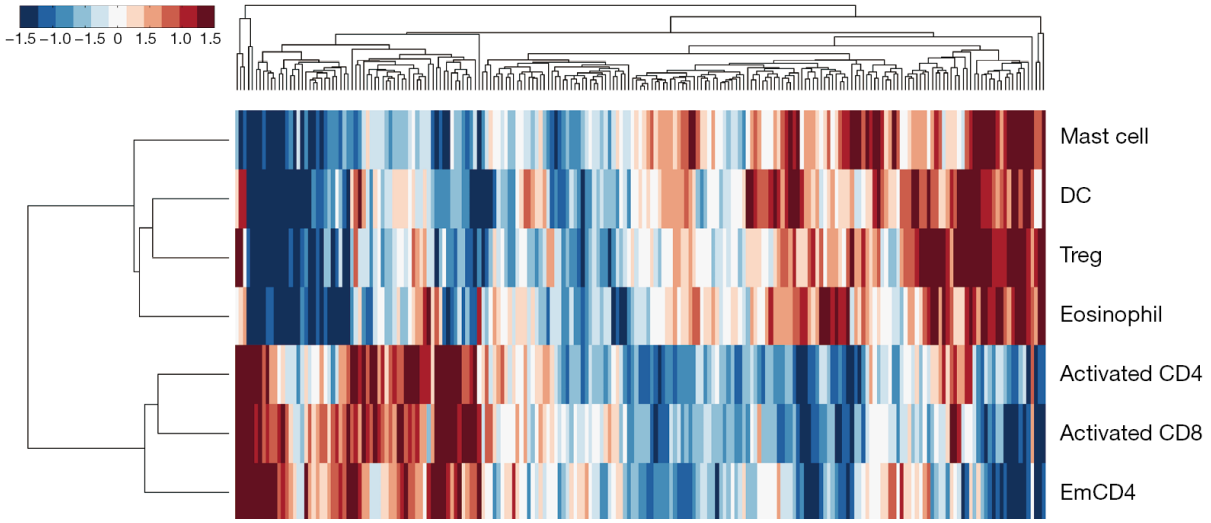
Gene set based method using unsupervised learning:

For example Bindea et. al. gene sets, derived from immune cell differential expression analysis

- GSEA (two group comparison)



- ssGSEA/GSVA (single sample enrichment of a gene set)



Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39:782–795.]

Deconvolution methods, supervised learning

Biological data properties:

- *Biological variations*
  - Different developmental stages
  - *Response to environment change*
- *Technical variations*
- Number of parameters is greater than the number of samples ($p \ll n$)
- Features are highly correlated
- Some genes are ubiquitously expressed in all cell-types to perform housekeeping functions, whereas other genes exhibit *specificity* for one, or a group of cell-types

Machine learning challenges:

→*Noisy data (Training and test data)*

→*Feature selection*

- Seek to distinguish features (genes) that closely conform to the linearity assumption, from the rest of the (variable) genes.

→*Over-fitting*

- Model is excessively complex, such as having too many parameters relative to the number of observations

→*Collinearity*

- Is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy

Different models used previously

Regression methods:

- Ordinary Least Squares (OLS) regression (Abbas *et al.)*
- non-negative least squares (NNLS) regression (Qiao et al.)
- quadratic programming (QP) (Gong et. al)
- Non-negative matrix factorization (NMF), coupled with regression (Venet et al., Repsilber et al., Zuckerman et al., Gajoux et al.)
- robust linear regression (RLR) and v-SVR regression (Newman et al.)

Bayesian methods:

- Bayesian prior and MCMC sampling (Erkkilä et al.)
- Latent Dirichlet Allocation (LDA)

Reference	Method	Loss	Non-negativity	Sum-to-one	Regularizer
Abbas *et al.* (2009)	Ordinary Least Squares (OLS)	$\mathcal{L}_2$	Imp	Imp	-
Gong *et al.* (2011)	Quadratic Programming	$\mathcal{L}_2$	Exp	Exp	-
Qiao *et al.* (2012)	Non-negative Least Squares (NNLS)	$\mathcal{L}_2$	Exp	Imp	-
DCQ– Altboum *et al.* (2014)	Elastic Net	$\mathcal{L}_2$	Imp	Imp	$\mathcal{L}_1 / \mathcal{L}_2$
RLR– Newman *et al.* (2015)	Robust Linear Regression (RLR)	Huber	Imp	Imp	-
CIBERSORT– Newman *et al.* (2015)	ν-SVR	ϵ-insensitive	Imp	Imp	$\mathcal{L}_2$

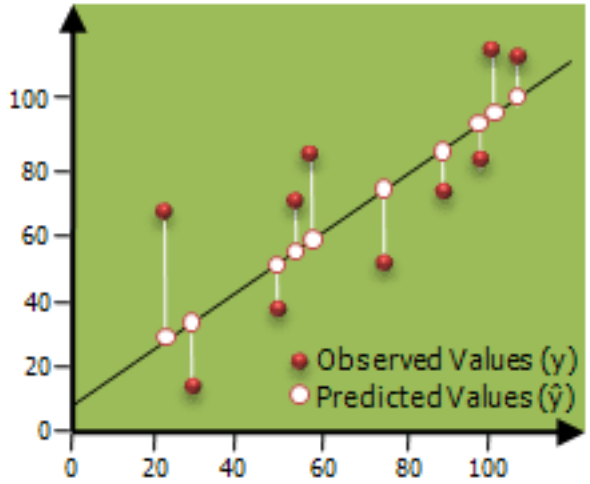
Formal definition

$\mathcal{L}$ cost of estimation error

m Number of samples

ri Fitting error

Simple example, Ordinary Least Squares



$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{m} \mathcal{L}(r_i)$$ Optimization problem, minimizes the sum of estimation error over all samples

Formal definition

$\mathcal{L}$ cost of estimation error
yi expression level of a gene in the mixture
$\boldsymbol{xi}$ expression level of the same gene in the reference cell type
w^T fraction of each cell type in the mixture, weight vector

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{m} \mathcal{L}(r_i)$$

Fitting error

$$r_i = y_i - f(\boldsymbol{x}_i)$$

Observed value

f(X) is a function that minimizes the aggregation error over samples, f(Xi) is a value of estimated value.

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

Linear regression finds a linear function that can predict observed values

$$r_i = y_i - \boldsymbol{w}^T \boldsymbol{x}_i$$

Formal definition

$\mathcal{L}$ cost of estimation error
yi expression level of a gene in the mixture
xi expression level of the same gene in the reference cell type
w^T fraction of each cell type in the mixture, weight vector

$r_i = y_i - \boldsymbol{w}^T \boldsymbol{x}_i$ Fitting error

Objective function $\underset{w \in \mathbb{R}^k}{\textbf{argmin}} \left\{ \underbrace{\sum_{i=1}^{m} \mathcal{L}(y_i - \boldsymbol{w}^T \boldsymbol{x}_i)}_{\text{Overall loss}} \right\}$

Finding weights (w^T) for X to minimize the prediction error

Formal definition

$\mathcal{L}$ cost of estimation error
yi expression level of a gene in the mixture
xi expression level of the same gene in the reference cell type
w^T fraction of each cell type in the mixture
λ parameter controls the relative importance of estimation error versus regularization (regularization parameter)

$r_i = y_i - \boldsymbol{w}^T \boldsymbol{x}_i$ Fitting error

Objective function $\underset{w \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{m} \mathcal{L}(y_i - \boldsymbol{w}^T \boldsymbol{x}_i)}_{\text{Overall loss}} \right\}$

loss function $\mathcal{L}$

Function that maps values of one or more variables onto a real number intuitively representing some "cost" associated with the event. An optimization problem seeks to minimize a loss function.

https://www.youtube.com/watch?v=euhATa4wgzo

https://www.youtube.com/watch?v=iSfcRku6euQ

Choice of Loss (or Cost) Functions

- *Ordinary Least Squares (OLS)* $\mathcal{L}_2(r_i) = r_i^2 = (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$
- Absolute deviation loss $\mathcal{L}_1(r_i) = |r_i| = |y_i - \boldsymbol{w}^T \boldsymbol{x}_i|$
- Huber's loss function $$\mathcal{L}^{(M)}_{\text{Huber}}(r_i) = \begin{cases} r_i^2, & \text{if } |r_i| \le M \\ M\,(2|r_i| - M), & \text{otherwise} \end{cases}$$
- *Support Vector Regression* $$\mathcal{L}^{(\epsilon)}_{\epsilon}(r_i) = \mathbf{max}\,(0, |r_i| - \epsilon) = \begin{cases} 0, & \text{if } |r_i| \le \epsilon \\ |r_i| - \epsilon, & \text{otherwise.} \end{cases}$$

Formal definition

$\mathcal{L}$ cost of estimation error
yi expression level of a gene in the mixture
xi expression level of the same gene in the reference cell type
w^T fraction of each cell type in the mixture
λ parameter controls the relative importance of estimation error versus regularization (regularization parameter)

$r_i = y_i - \boldsymbol{w}^T \boldsymbol{x}_i$ Fitting error

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{m} \mathcal{L}(r_i)$$

$$f_w(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

$$r_i = y_i - f(\boldsymbol{x}_i)$$

$$\underset{w \in \mathbb{R}^k}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{m} \mathcal{L}(y_i - \boldsymbol{w}^T \boldsymbol{x}_i)}_{\text{Overall loss}} + \underbrace{\lambda \mathcal{R}(\boldsymbol{w})}_{\text{Regularizer}} \right\}$$

regularizer function $\mathcal{R}$

Function for selecting the preferred level of model complexity so your models are better at predicting (generalizing). Too complex model can overfit or be too simple and underfit, either way giving poor predictions

https://www.youtube.com/watch?v=C79kIYkKZ1g

Deconvolution methods

Choice of Regularizers

- L1/norm-1 regularizer/Lasso
- L2/norm-2 regularizer/ridge regression
- *L1+ L2/elastic net*
- *Group Lasso*

$\mathcal{R}$ regularizer function

w regression coefficient vector

$$\mathcal{R}_1(\boldsymbol{w}) = \|\boldsymbol{w}\|_1 = \sum_{i=1}^{k} |w_i|$$

$$\mathcal{R}_2(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2 = \sum_{i=1}^{k} w_i^2$$

$$\mathcal{R}_{\text{elastic}}(\boldsymbol{w}) = \alpha\mathcal{R}_1(\boldsymbol{w}) + (1-\alpha)\mathcal{R}_2(\boldsymbol{w})$$

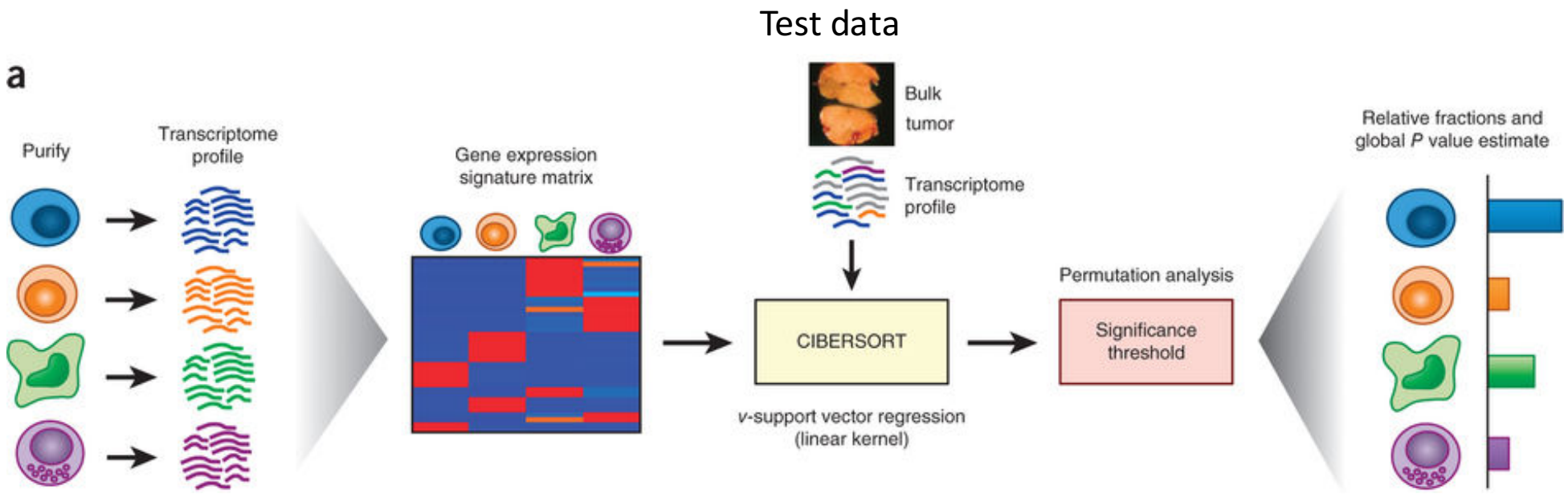
$$\mathcal{R}_{\text{group}} = \sum_{G_i} \mathcal{L}_2(\boldsymbol{w}(G_i))$$

Objective Functions Used in Practice

- Ordinary Least Squares (OLS)
- Ridge Regression
- Least Absolute Selection and Shrinkage Operator (LASSO) Regression
- Robust Regression
- Support Vector Regression

Deconvolution methods

Example: CIBERSORT



Training data

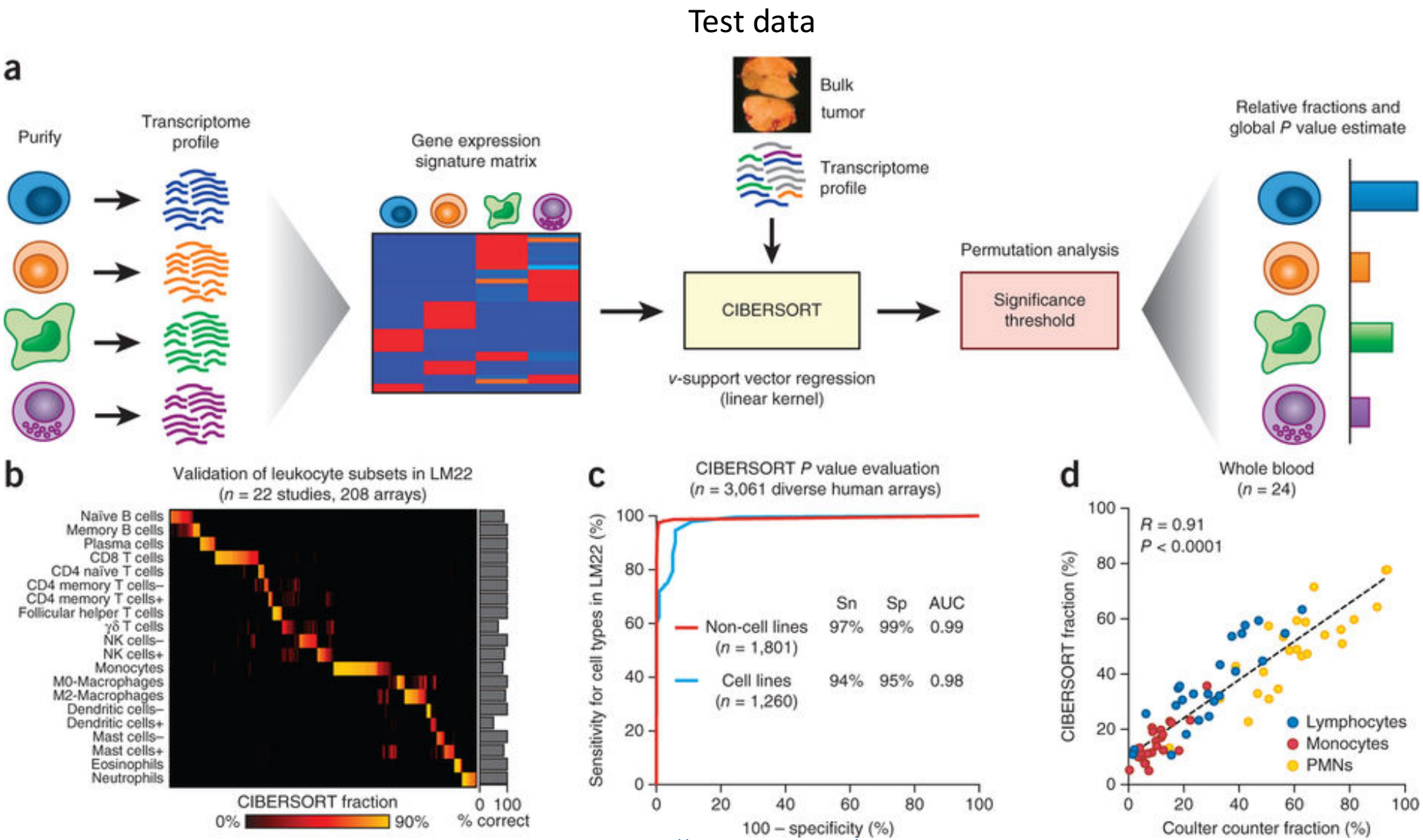
Algorithm

Significance Estimation

Output

Deconvolution methods

Example: CIBERSORT



Benchmark data set

https://www.youtube.com/watch?v=OAI6eAyP-yo
(AUC-ROC curve explained)

Validations, independent method

Cibersort algorithm

- nu–support vector regression (ν-SVR)
- Robustness to noise and overfitting owing to both a linear loss function and feature selection of genes from the signature matrix
- Tolerance to collinearity via utilization of the L2-norm penalty function.
- Produces an empirical P-value for the deconvolution using Monte Carlo sampling

DECONVOLUTION WAR!

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7. 2015

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

DECONVOLUTION WAR!

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7. 2016

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

DECONVOLUTION WAR!

CIBERSORT,
22 cell type
deconvolution

TIMER,
6 cell type
deconvolution

TIMER correspondence:
critique to CIBERSORT

"CIBERSORT produces non-biological negative correlations mainly due to collinearity"

"CIBERSORT is only applicable to microarray data"

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7. 2016

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

DECONVOLUTION WAR!

CIBERSORT,
22 cell type
deconvolution

TIMER,
6 cell type
deconvolution

TIMER correspondence:
critique to CIBERSORT

"CIBERSORT produces non-biological negative correlations mainly due to collinearity"

"CIBERSORT is only applicable to microarray data"

CIBERSORT correspondence:
critique to TIMER

TIMER has "problems with marker gene selection and/or data normalization"

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7. 2016

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

DECONVOLUTION WAR!

2016 — CIBERSORT, 22 cell type deconvolution

Genes exhibit *specificity* for one, or a group of cell-types

2016 — TIMER, 6 cell type deconvolution

2017 — TIMER correspondence: critique to CIBERSORT

"CIBERSORT produces non-biological negative correlations mainly due to collinearity" *Collinearity*

"CIBERSORT is only applicable to microarray data" *Technical Variation*

2017 — CIBERSORT correspondence: critique to TIMER

TIMER has "problems with **marker gene selection** and/or data **normalization**" *Feature selection, Technical Variation*

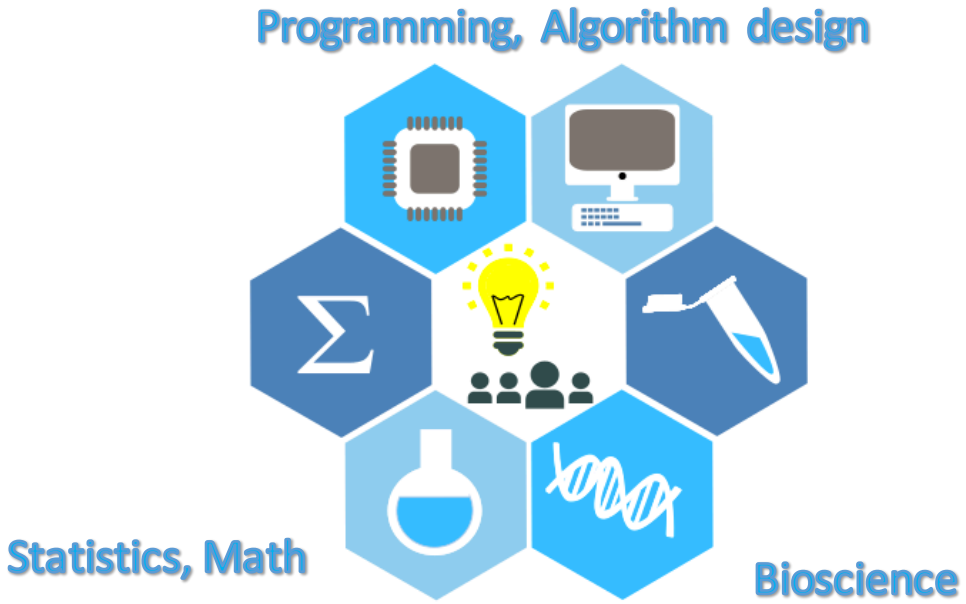
2017 — TIMER correspondence: critique to CIBERSORT

"we will show in this response that the inclusion of **highly similar features** contributes as significantly as normalization, if not more, to the observed artificial negative correlations among the estimates obtained by CIBERSORT. "

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7. 2016

Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol. 2016;17:174.

Bioinformatics is a challenging multidisciplinary field



For reading and references

- Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25:571–8. doi: 10.1016/j.coi.2013.09.015. **REVIEW, biological overview of the problem**
- Mohammadi S, Zuckerman N, Goldsmith A, Grama A. A critical survey of deconvolution methods for separating cell-types in complex tissues. arXiv. 2015: 1510.04583 **Technical REVIEW of methods**

Questions:

Consider you are working with deconvolution to dissect immunologic profile of the tumor:

- Why deconvolution can be useful and what is the main output of the algorithm?
- Which steps are needed to develop a deconvolution algorithm and reason why?
  Hint: Point is not to list all different algorithms that could be used, but to point out the key steps and reasons why they are needed, as referred below.

Training data

Test data

Algorithm
- Cost function
- Regularization
- Feature Selection

Benchmarking/Validation