HBC
Harvard Chan Bioinformatics Core

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Deconvoluting bulk RNA-seq data to measure cell populations

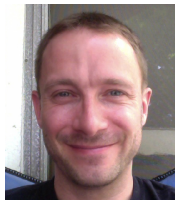
John Hutchinson
The Harvard Chan Bioinformatics Core

The Harvard Chan Bioinformatics Core

Shannan Ho Sui
Director

John Hutchinson
Associate Director

Victor Barrera



Rory Kirchner

Zhu Zhuo

Preeti Bhetariya

Meeta Mistry

Mary Piper

Jihe Liu

Radhika Khetani
Training Director

Ilya Sytchev

James Billingsley

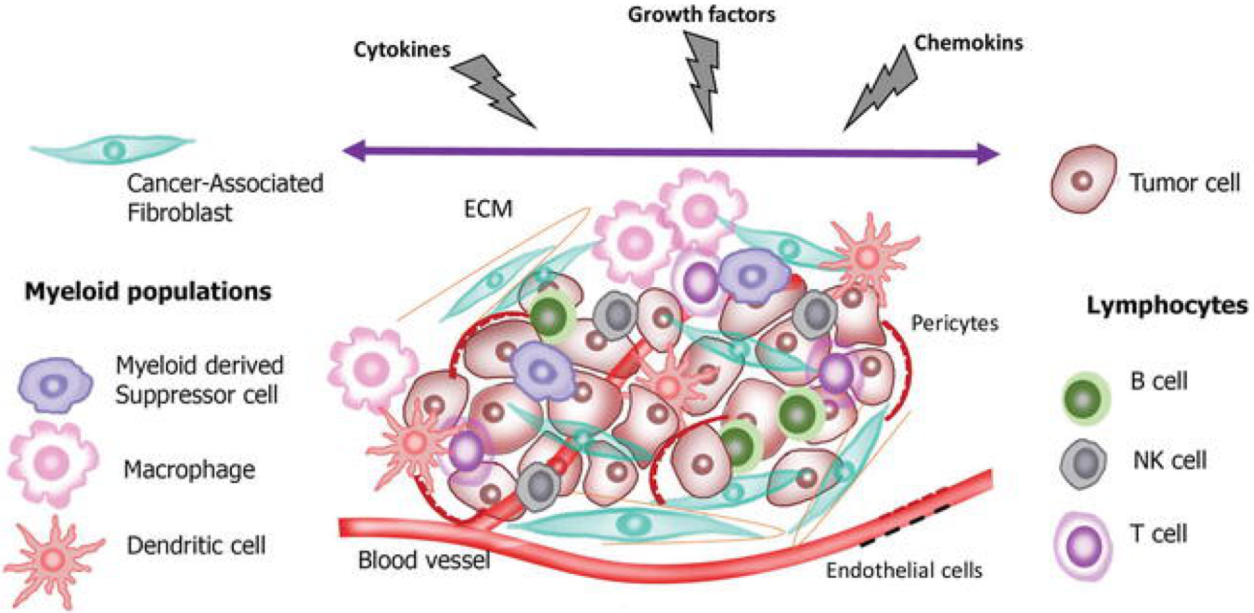
Sergey Naumenko

Joon Yoon

Peter Kraft
Faculty Advisor

Why would immuno-oncologists care about cell composition?

- Thereapeutic decisions require knowledge of complex tumor microenvironment
  - Cell types and proportions?

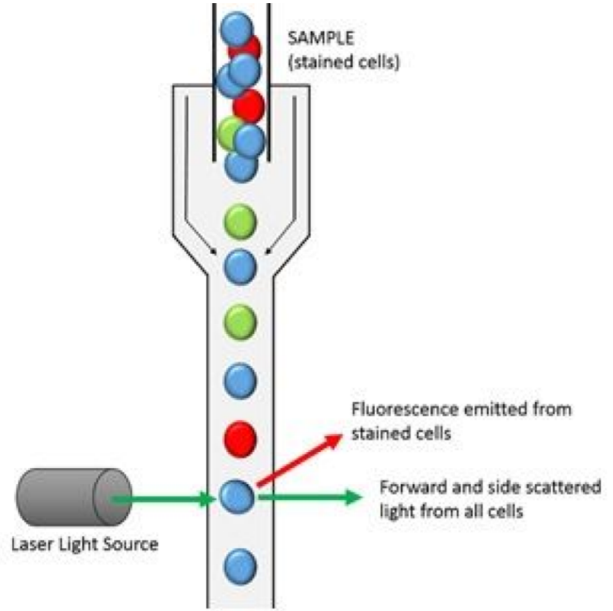


Approaches

- Cell sorting
  - FACS
  - CyTOF
- IHC/IF
  - Cell staining
- Bulk Transcriptomics
  - Microarrays
  - RNA-seq
- Single cell RNA-seq
  - Transcriptomics of single cells
- Combinations?
  - Spatial transcriptomics?

Approaches - Cell sorting

- Sort and label cells using cell type specific antigens
- Detect labels on cells
  - CyTOF - time-of-flight mass spectrometry
  - FACS - fluorescent activation cell sorting



Pros

- Known technology, established infrastructure
- Comparatively cheap (non CyTOF)

Cons

- Limited markers (max 50 for CyTOF)
- CyTOF antibodies are expensive
- Potential disaggregation issues

Approaches - IHC/IF

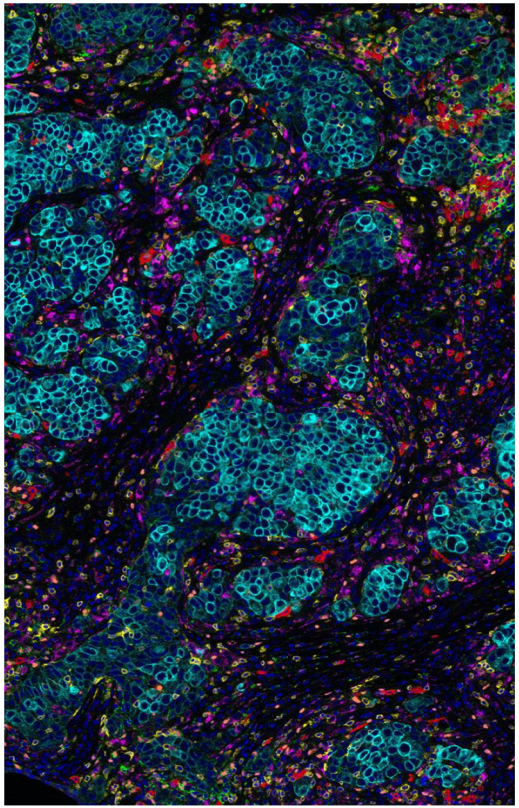
Sectioning and staining for cell type specific markers

Pros

- Known technology, established infrastructure
- Comparatively cheap
- Lots of FFPE and frozen tissue samples available

Cons

- Sections only, hard/expensive to assay entire tumor
- Limited to a few markers per section



Approaches - Single Cell RNA-seq

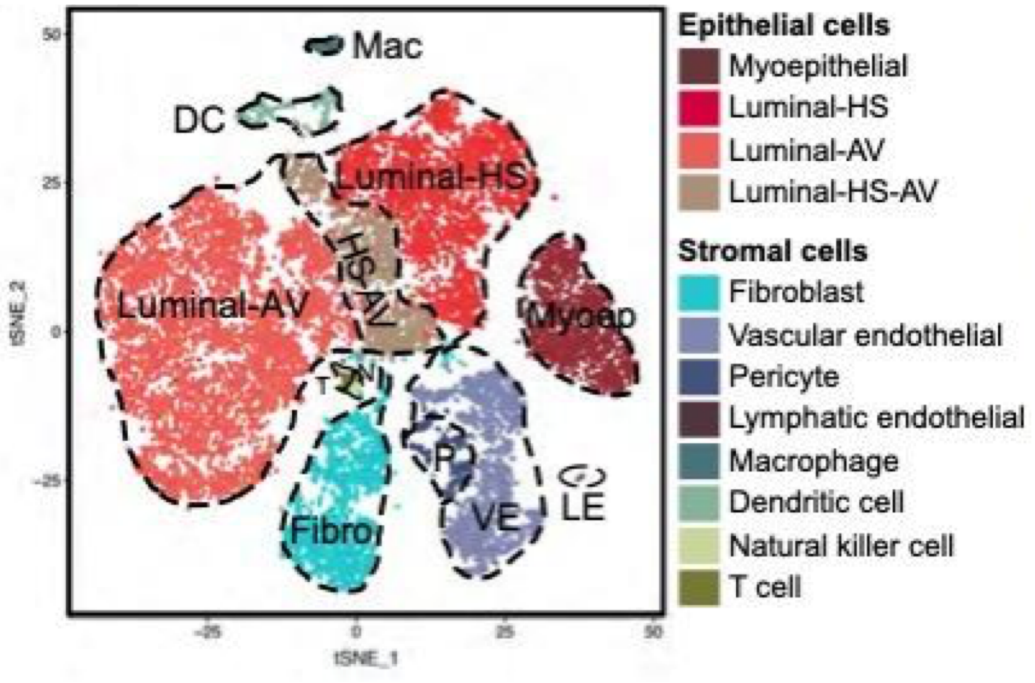
10X, InDrops, DropSeq, SmartSeq

Pros

- Powerful
- Effective

Cons

- Expensive
- Disaggregation bias
- Can’t always identify the cells
  - Marker issues



Li, C. M.-C. *et al.* Aging-associated alterations in the mammary gland revealed by single-cell RNA sequencing. *bioRxiv* 773408 (2019). doi:10.1101/773408

Approaches - Bulk RNA-seq

Bulk RNA-seq = all cells within mixture contribute to final expression levels

Pros

- Can assay entire sample at once
- Can help identify transcription changes in individual cell types
- Huge amount of data out there already
- Cheap(er)

Cons

- Hard to do well

Bulk sample analysis is just like putting a fruit salad into a blender - the taste is an average of all ingredients.

Analyzing single cells is like tasting each individual piece of fruit to gain a much more nuanced understanding of the composition of the fruit salad

(Graphic blatantly stolen from the Qiagen website)

$200/sample (Novogene)

$4000-10000/sample

Can we computationally figure out what went into the mixture?

Approaches – Two main types

1. Deconvolution
    1. Partial or full
2. Marker based measurements

Methods – Some popular approaches

One review listed 64 approaches!

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)

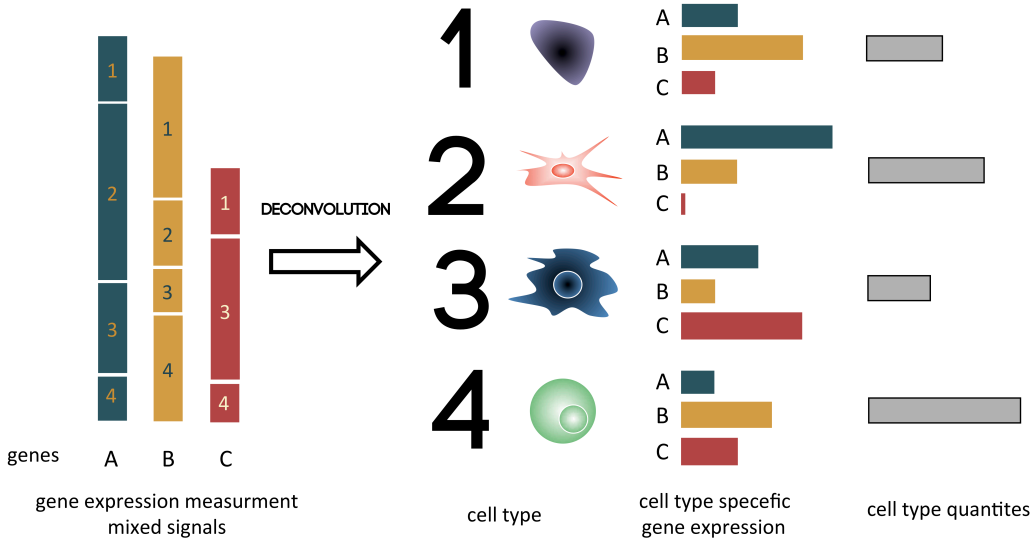
Deconvolution methods – unmixing the smoothie

How many strawberries, kiwis, pineapples and oranges went into the salad?

Deconvolution methods

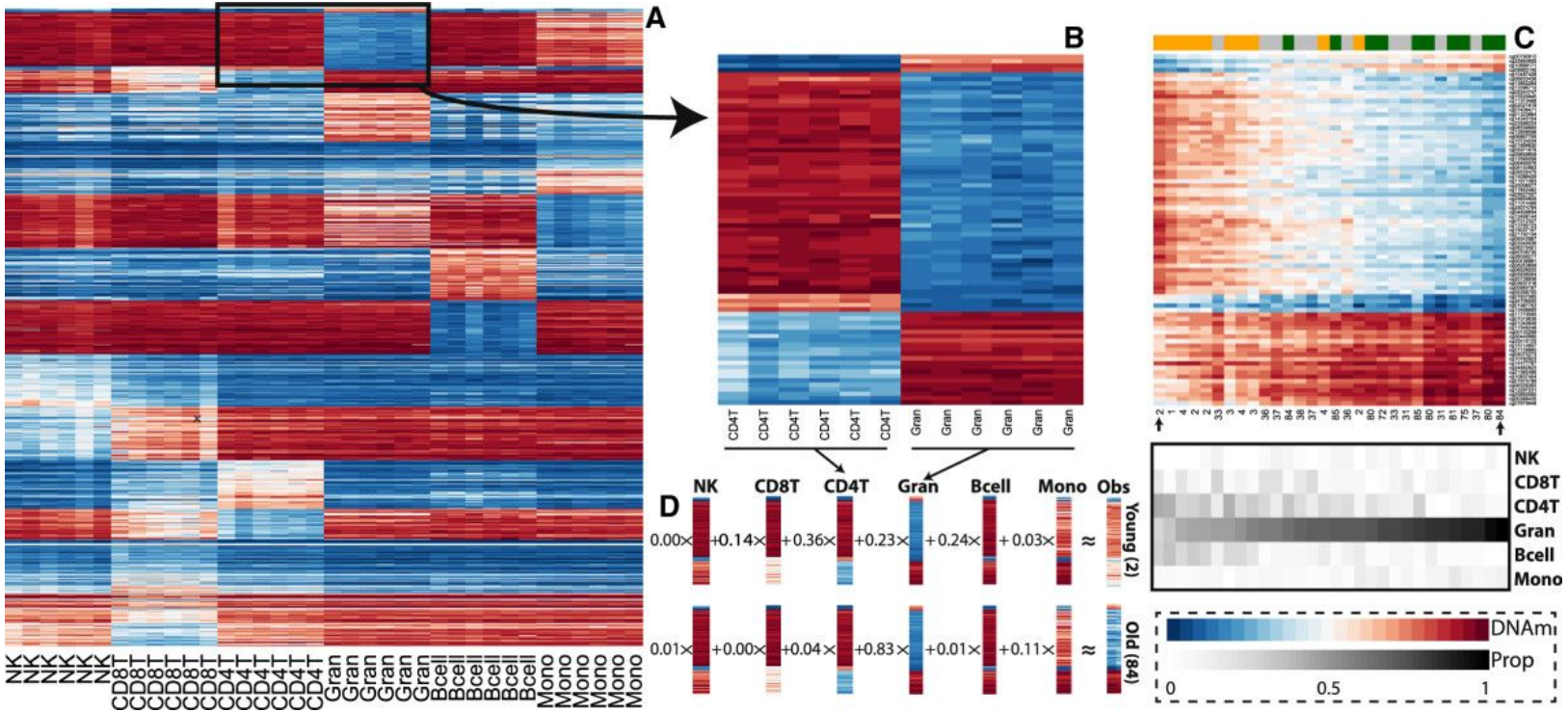
- Complicated math
- “a system of equations that describe the expression of each gene in a heterogeneous sample as a linear combination of the expression levels of that gene across the different cell subsets present in the sample, weighted by their relative cell fractions”

(Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).



Successful deconvolution in a related technology

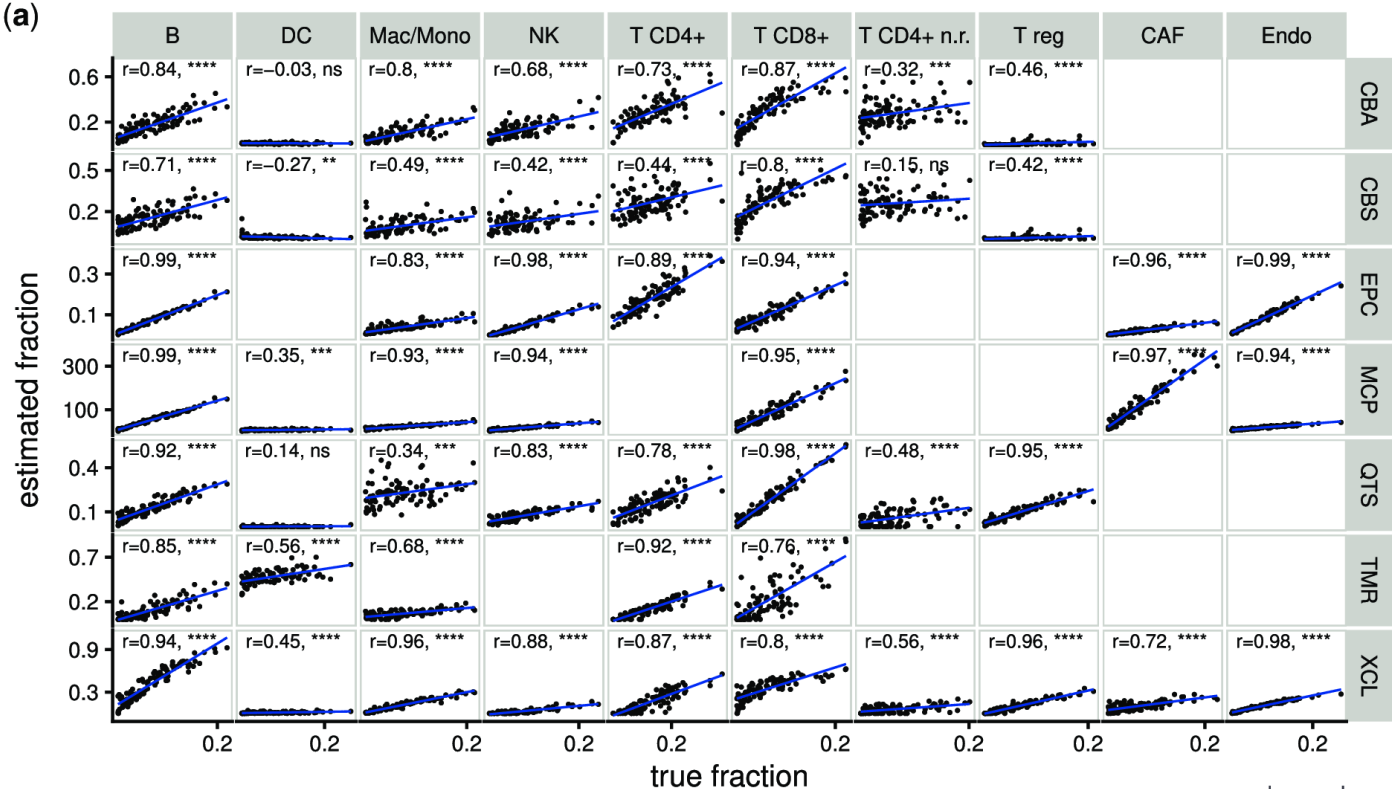
Changes in DNA methylation in PBMCs during aging driven entirely by changes in cell composition



Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).

Deconvolutions don't always work well

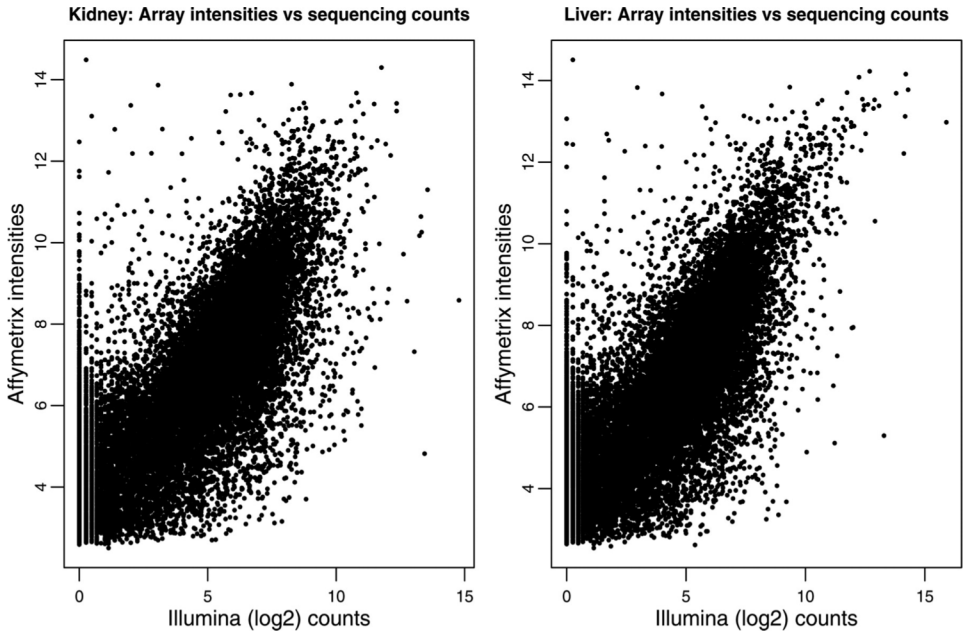
- simulated data sets drawn from scRNA-seq data



Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Issues - Technological biases

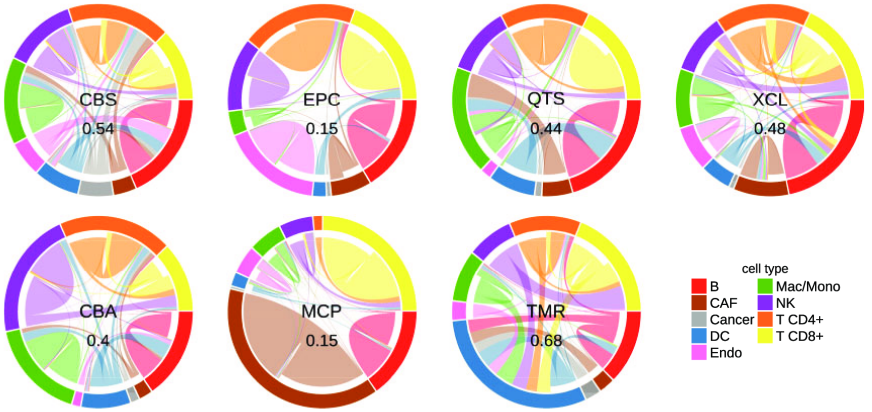
- Some of the methods rely on microarray based cell type references
- Microarrays = probe intensities
  - continuous measure, best modeled by normal distribution after log transformation
- RNA-seq – read counts
  - count based measure, best modeled by negative binomial distribution of raw counts
- Can transform RNA-seq data to better fit microarray (normal) distributions but count based methods would be better



John C. Marioni et al. Genome Res. 2008;18:1509-1517

Issues – “spillover”

- Closely related cell types have similar cell signatures
- scores that predict enrichment of one cell type may also predict enrichment of another cell type
    - other cell type might not even be present



Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Issues – Effects of unknown cell types

There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

- Donald Rumsfeld

- you can't measure something you don't know is there

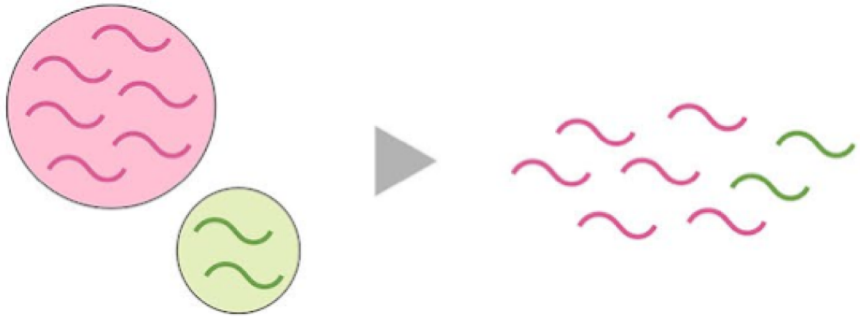
- "spillover" from unidentified cell types with can shift measures for your known cell types

Issues – Microenvironment effects

- Reference sets are often derived from purified **non-tumor** cells
- Do pure cell populations accurately reflect the gene expression patterns of cells in a tumor?
- Cell state versus cell identities - microenvironment affects cell state

Issues - Cell size biases

- Cells are not all the same size
- Methods may assume that each cell contributes an equal amount of RNA to total pool
- BUT bigger cells can have more RNA



Issues – limited reference sets

- Uneven background dataset availability
  - Not all cell types available for all methods
  - Not all species available

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)

Issues – practical problems

Method may :

- Require raw data availability
- Need all samples be run at same time
- Not have good or accessible software
    - CIBERSORT, XCell, TIMER = webtools limit bulk use
    - EPIC = R package and webtool available
    - quanTIseq = Bash command line package (available as Docker image)
    - Immundeconv = R package containing all major methods

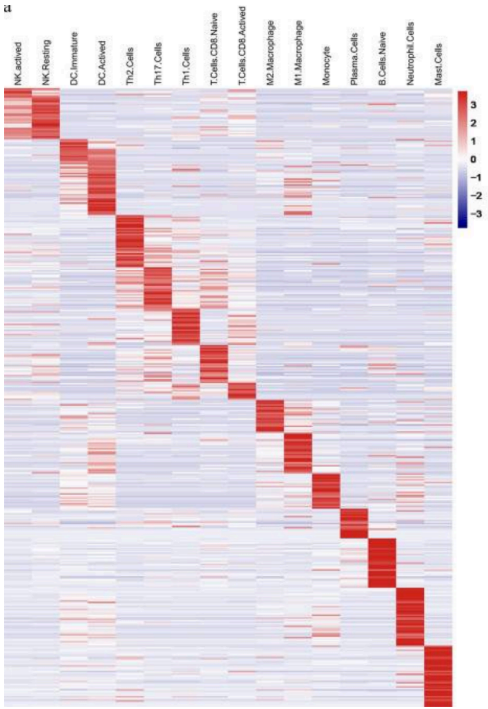
Marker base methods – Keeping it simple

Which has the most strawberries?

(trick question, these are all the same)

Marker based methods

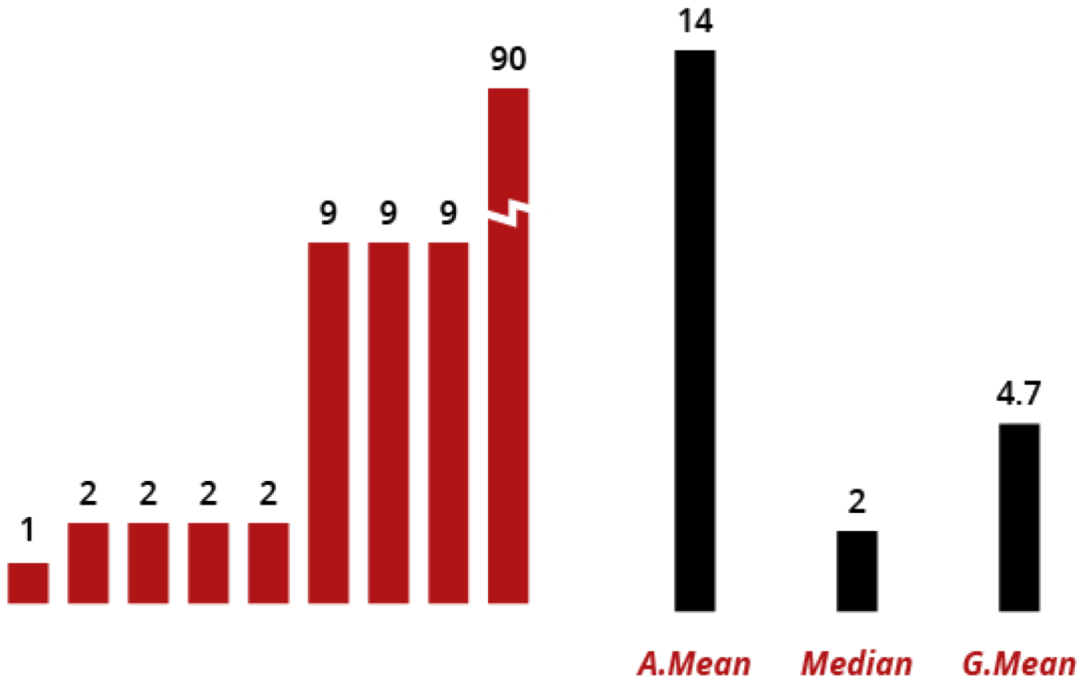
- Using lists of genes that are characteristic for a cell type
    - Derived from targeted transcriptomics or literature studies
- Semi-quantitiative
    - Can compare between samples but not between cell types



Kassambara, A. *et al.* GenomicScape: an easy-to-use web tool for gene expression data analysis. Application to investigate the molecular events in the differentiation of B cells into plasma cells. *PLoS Comput. Biol.* **11**, e1004077 (2015).

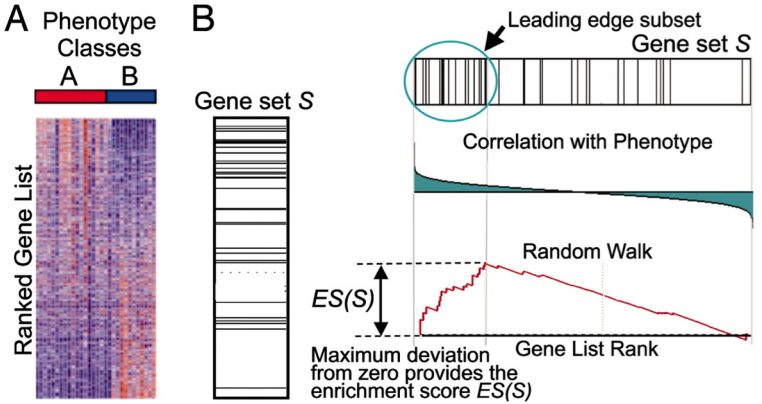
Marker methods

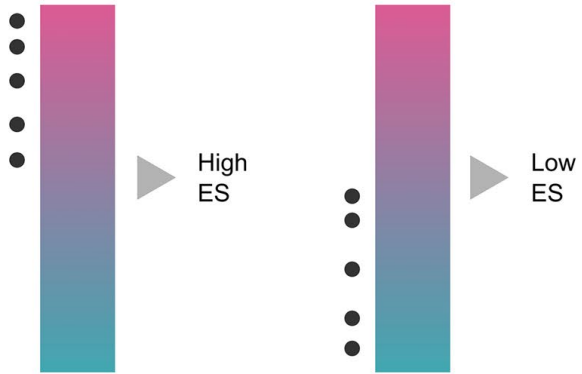
- Can use simple ”robust” summaries



Marker methods

- Can use more robust GSEA methods
  - Gene Set Enrichment Analysis
  - Rank based



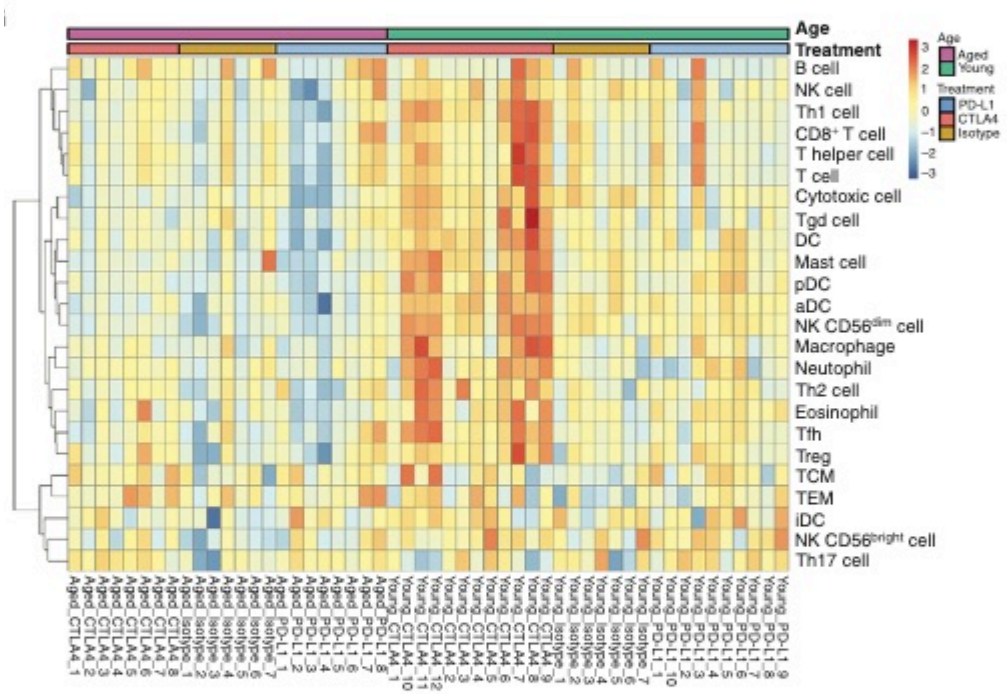


Marker methods – an example

- We had mouse data which precluded most published methods
- Had to get creative!

- Used the Nanostring Mouse PanCancer Immune Profiling Panel genes as cell type markers
- Used the geometric mean expression of the marker sets in each sample

- Were able to compare immune signatures across samples (but not across cell types)
- GOOD ENOUGH



Sceneay, J. *et al.* Interferon Signaling is Diminished with Age and is Associated with Immune Checkpoint Blockade Efficacy in Triple-Negative Breast Cancer. *Cancer Discov.* CD–18 (2019).

Marker methods – a 2nd example

- Wanted to look at levels of pro-metastatic immunsuppressive neutrophils in two different biological conditions

- Had mouse RNA-seq data
- no reference data for the cell types

Couldn't do any of the populat deconvolutions methods!

What did we have to work with?

- Differential expression of KEP cells compared to controls from Coffelt 2015
  (Coffelt, S. B. *et al.* IL-17-producing \gamma\delta T cells and neutrophils conspire to promote breast cancer metastasis. *Nature* **522**, 345–348 (2015).)

  1. Genes upregulated in KEP cells (pro-metastatic immunosuppressive neutrophil markers- Signature 1)

  2. Genes downregulated in KEP cells (control neutrophil markers - Signature 2)

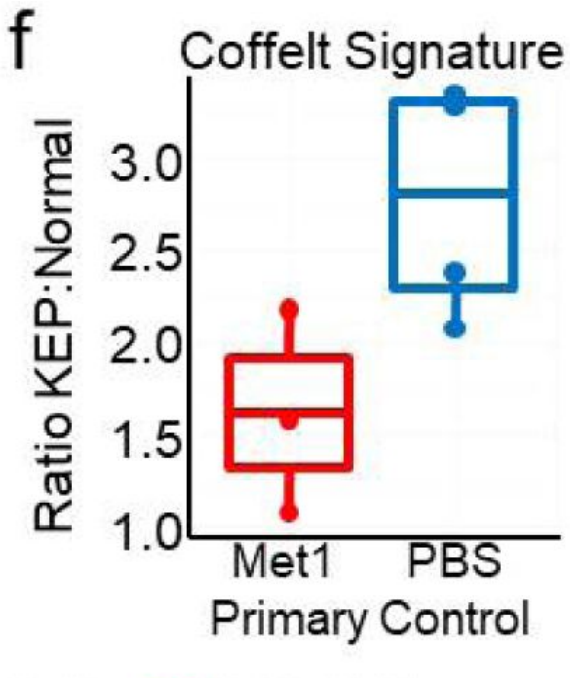
- Samples with more pro-metastatic immunosuppressive neutrophils should have HIGHER expression of genes upregulated in KEP cells (Signature 1) and LOWER expression of genes downregulated in KEP cells (Signature 2)
- So a ratio of Signature1:Signature2 will be higher in samples with more pro-metastatic immunosuppressive neutrophils

Marker methods – a 2nd example

- Ratio of genes expressed by pro-metastatic immunosuppressive neutrophils from *K14cre;Cdh1F/F;Trp53F/F* (KEP) mice to control neutrophils from wild type littermates (KEP:Normal)

Blue – control lungs
Red – lungs from primary tumour-bearing animals

- Higher ratios indicate higher pro-metastatic KEP signatures.



Take home messages

- Carefully consider your options and what you need from the experiment
  - Tradeoffs with any method
  - Is your data appropriate for the method?
  - Avoid deconvolution if you can
    - While not perfect, marker based methods are simple and less prone to assumptions
- Validate, validate, validate

Future

- Addressing spillover, technological biases and limited reference sets
    - Better references and marker sets – single cell RNA-Seq
- Microenvironment and unknown cell types issues
    - Single cell RNAseq analysis of exemplar samples

“…we believe that the improvements made to signature matrices outweigh potential algorithmic improvements”

Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Recent publications that used single cell to improve deconvolution

1. Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).

2. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk Tissue Cell Type Deconvolution with Multi-Subject Single-Cell Expression Reference. *bioRxiv* 354944 (2018). doi:10.1101/354944

3. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).

4. Menden, K., Marouf, M., Dalmia, A., Heutink, P. & Bonn, S. Deep-learning-based cell composition analysis from tissue expression profiles. *bioRxiv* 659227 (2019). doi:10.1101/659227

Future

Tumor Deconvolution DREAM Challenge

https://www.synapse.org/#!Synapse:syn15589870/wiki/582446

The goal of this Challenge is to evaluate the ability of computational methods to deconvolve bulk expression data, reflecting a mixture of cell types, into individual immune components.

Methods will be assessed based on *in vitro* and *in silico* admixtures specifically generated for this Challenge.

Acknowledgements

Nanostring immune panel work

Victor Barrera

McAllister lab

Sandy McAllister

Zafira Castano

Jaclyn Sceneay

The Harvard Chan Bioinformatics Core

Funding

HSCI HARVARD STEM CELL INSTITUTE®

HARVARD CATALYST | THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Contact us anytime at bioinformatics@hsph.harvard.edu or via our website at bioinformatics.sph.Harvard.edu

Methods – “bakeoff”

Datasets

- integrated scRNA-seq dataset of more than 11 000 single cancer, stromal and immune cells from 23 melanoma and ovarian cancer patients
  - simulate bulk RNAseq and validate results
  - individually retrieved and aggregated 500 random immune- and non-immune cells
- three independent datasets that have been profiled with FACS
  - PBMCs
  - Ovarian cancer
  - Melanoma

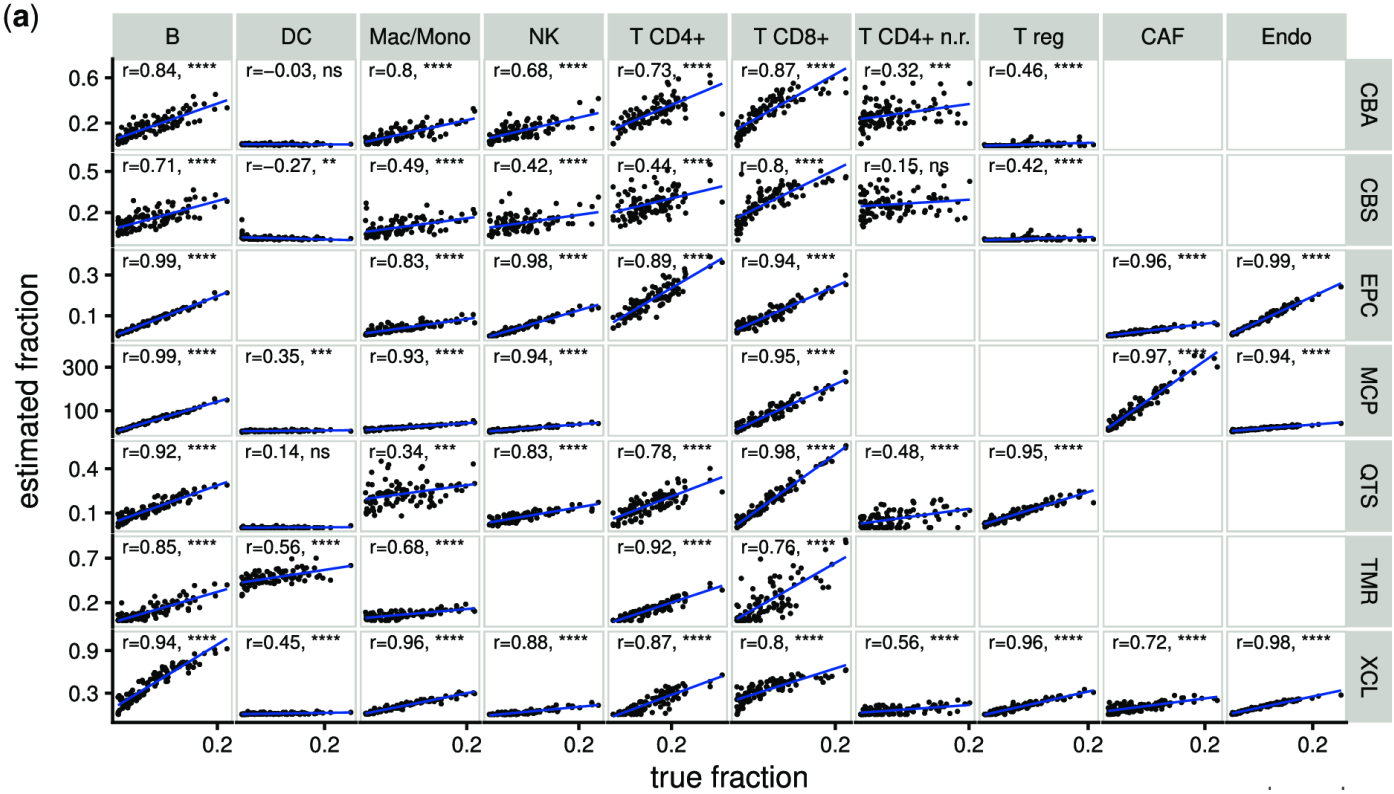
Methods

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)

Comparing methods - correlations

- simulated data sets drawn from scRNA-seq data



Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

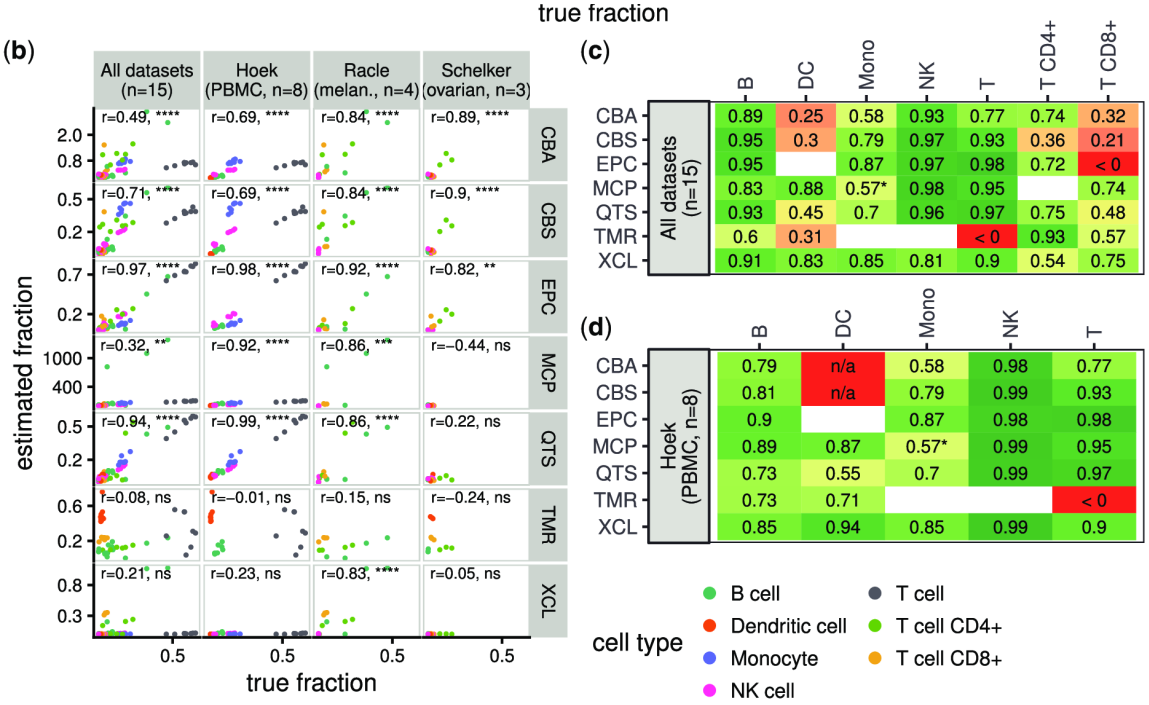
Comparing methods – correlations

Real data

Hoek = PBMCs

Racle = Melanoma

Schelker = Ovarian



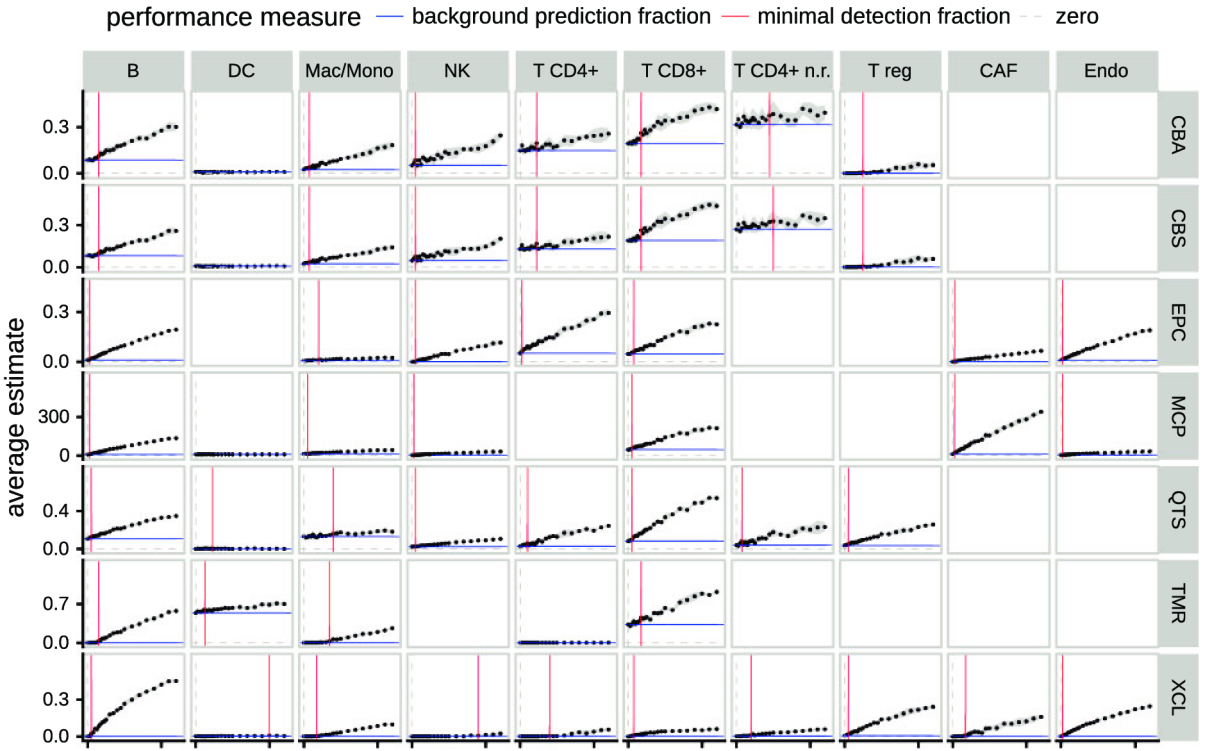
Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Comparing methods – detection limits

- simulated bulk RNA-seq samples with an increasing amount of the cell type of interest (x-axis)
- background of 1000 cells randomly sampled from the other cell types

Figure explanation

- dots = the mean predicted score across five independently simulated samples for each fraction of spike-in cells
- red line = minimal detection fraction, i.e. the minimal fraction needed for a method to detect its abundance as different from background
- blue line = background prediction level, i.e. average estimate of a method while the cell type is absent



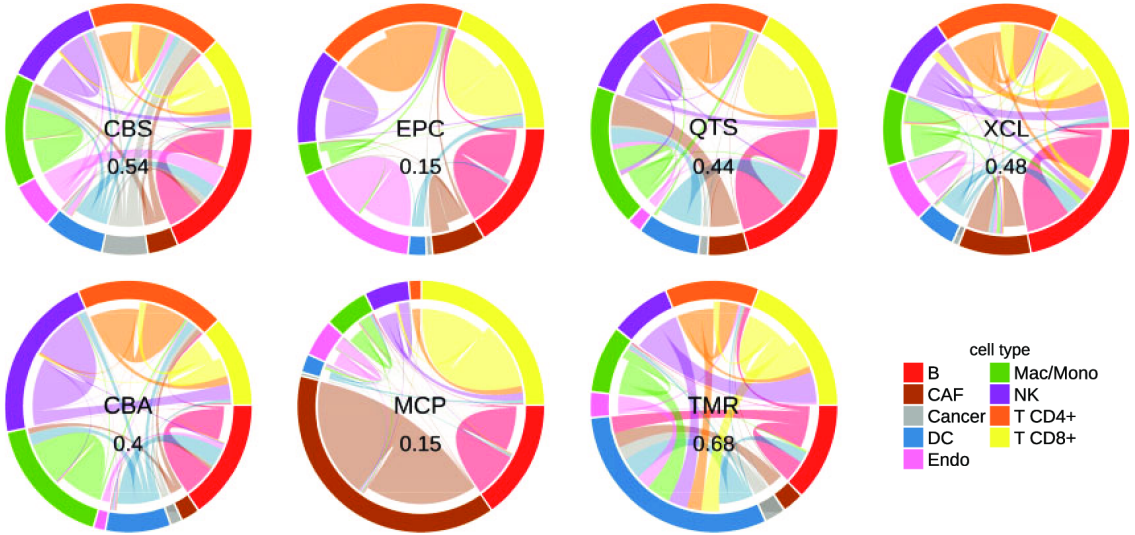
Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Comparing methods – spillover

- simulated bulk RNA-seq samples containing only cells of one of the nine immune and non-immune cell types

Figure Explanation

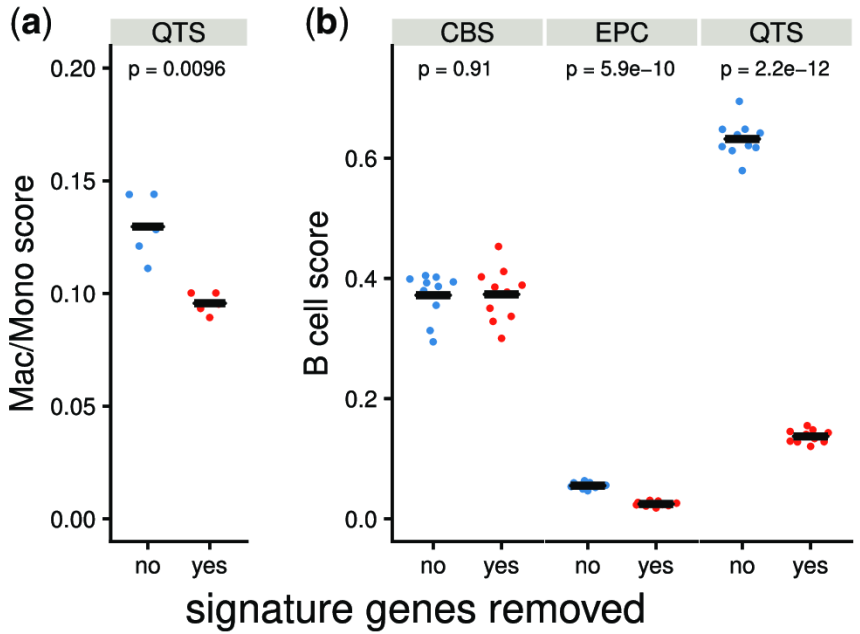
- outer circle indicates different samples
- interior connections refer to method predictions
- size of a border segment reflects the predicted score for that cell type (
- connection leading to border segment of same color indicates a correctly predicted cell type fraction
- connection leading to a different color indicates spillover



Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Comparing methods – spillover improvements

- Spillover can be improved with more specific signatures



Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Comparing methods – recommendations

- No “one-size-fits-all” method

1. General purpose deconvolution
   - EPIC and quanTIseq
2. absolute levels not needed (inferring changes between treatment and control groups)
   - MCP-counter
     - low spillover
3. presence/absence of a cell type
   - xCell
     - best results when cells actually absent

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.			

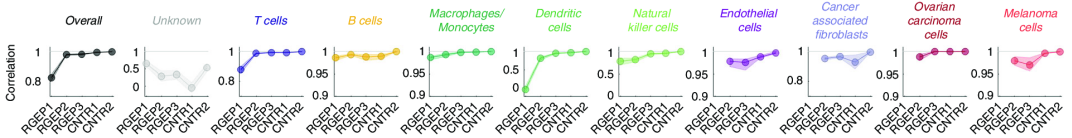
Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).

Combination methods – Using single cell data

Worked witih single cell samples from multiple sites:

1. PBMCs
2. Ascites
3. Melanoma

and multiple patients



REGP1 - PBMC only derived signatures (equivalent to current signatures)

REGP2 - Consensus signatures from all single cell samples (PBMCs plus melanoma and ascites)

REGP3 - Indication specific signatures from single cell

REGP4 - Patient malignant, consensus non-malignant signatures

REGP5 - Patient specific all cell types signatures

Schelker, M. *et al.* Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* **8**, 2032 (2017).