Why data science?

Harvard Business Review | Analytics And Data Science | Data Scientist... | Subscribe | Sign In

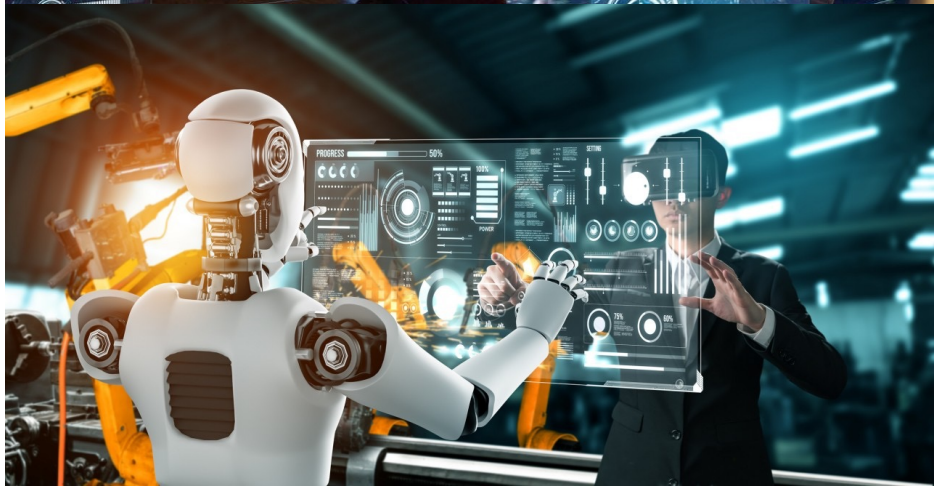
Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and DJ Patil

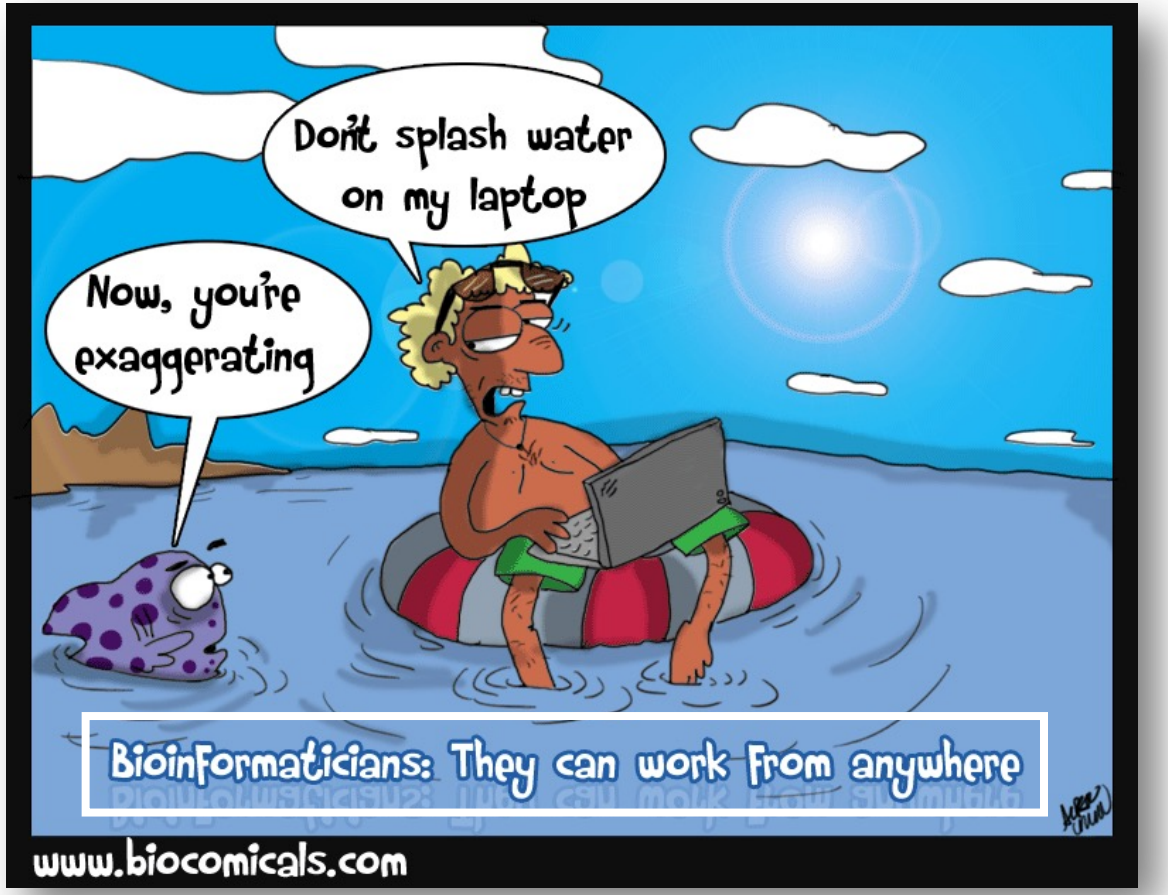
From the Magazine (October 2012)

It’s the future

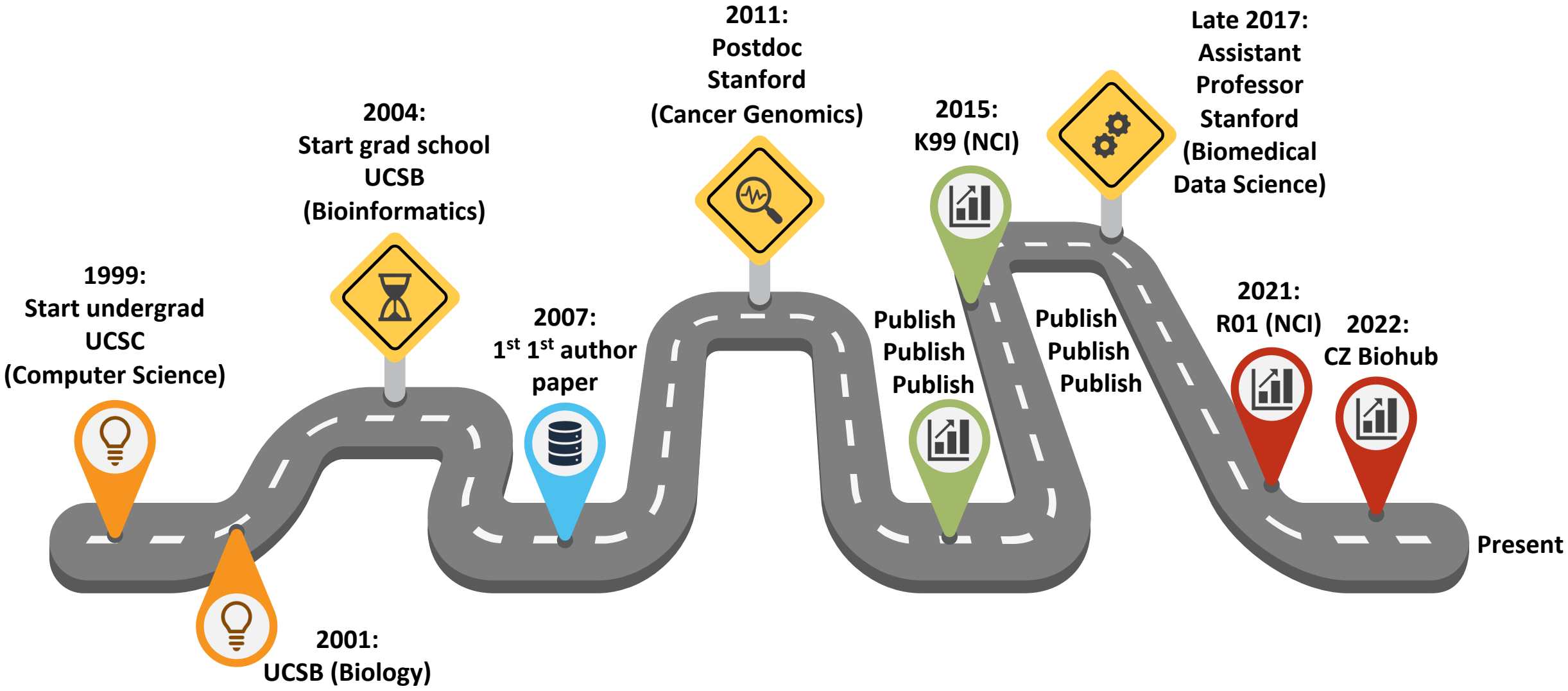


Why biomedical data science?

- Biology is an **information science**
  - *Massive resolution, complexity, and scale*
- Data science enables **analysis** of otherwise **impenetrable data**
  - *Data-driven, agnostic, systematic*
  - *See the forest (big picture), then select the most promising trees*
- Can **accelerate basic/clinical science** (by days, months, or years)
- **Robust** and **reproducible**
- Foundation of **precision medicine**

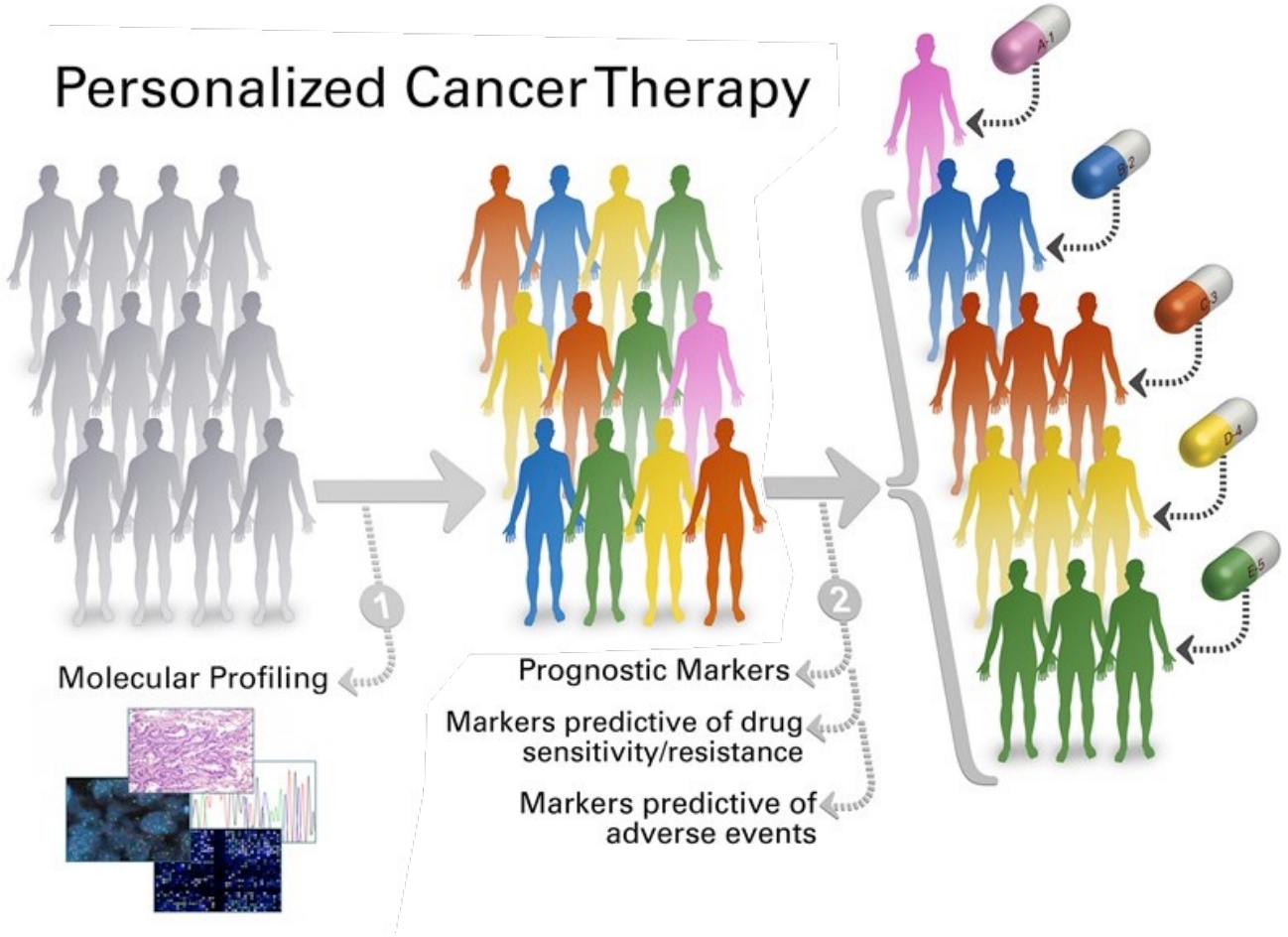


My path to running a data science research group



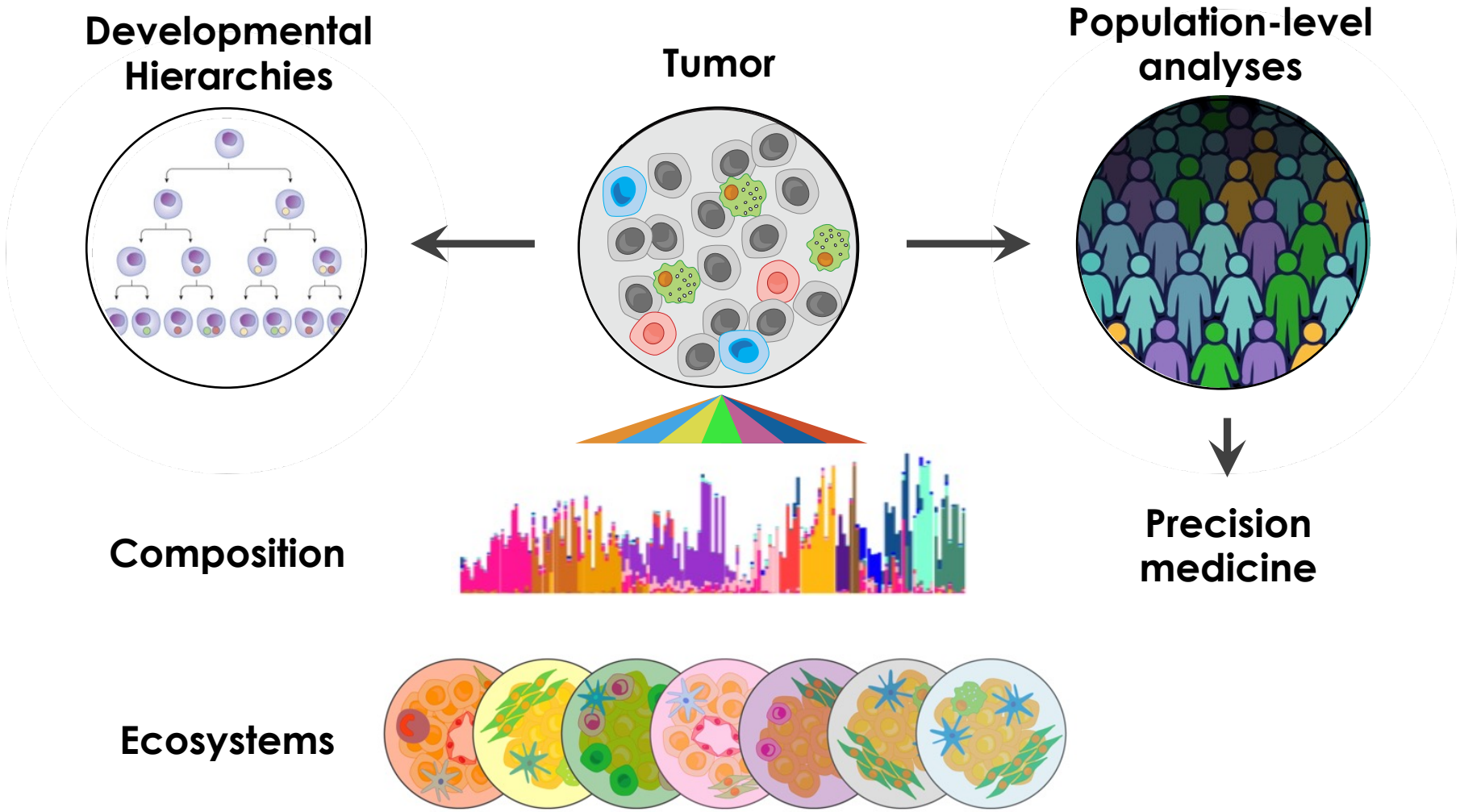
Cancer remains the 2nd leading cause of death

- Tumors are heterogenous on molecular, phenotypic, and spatial levels
- Every patient's cancer is unique
- Critical need for **precision oncology:** individualized diagnostics and treatments

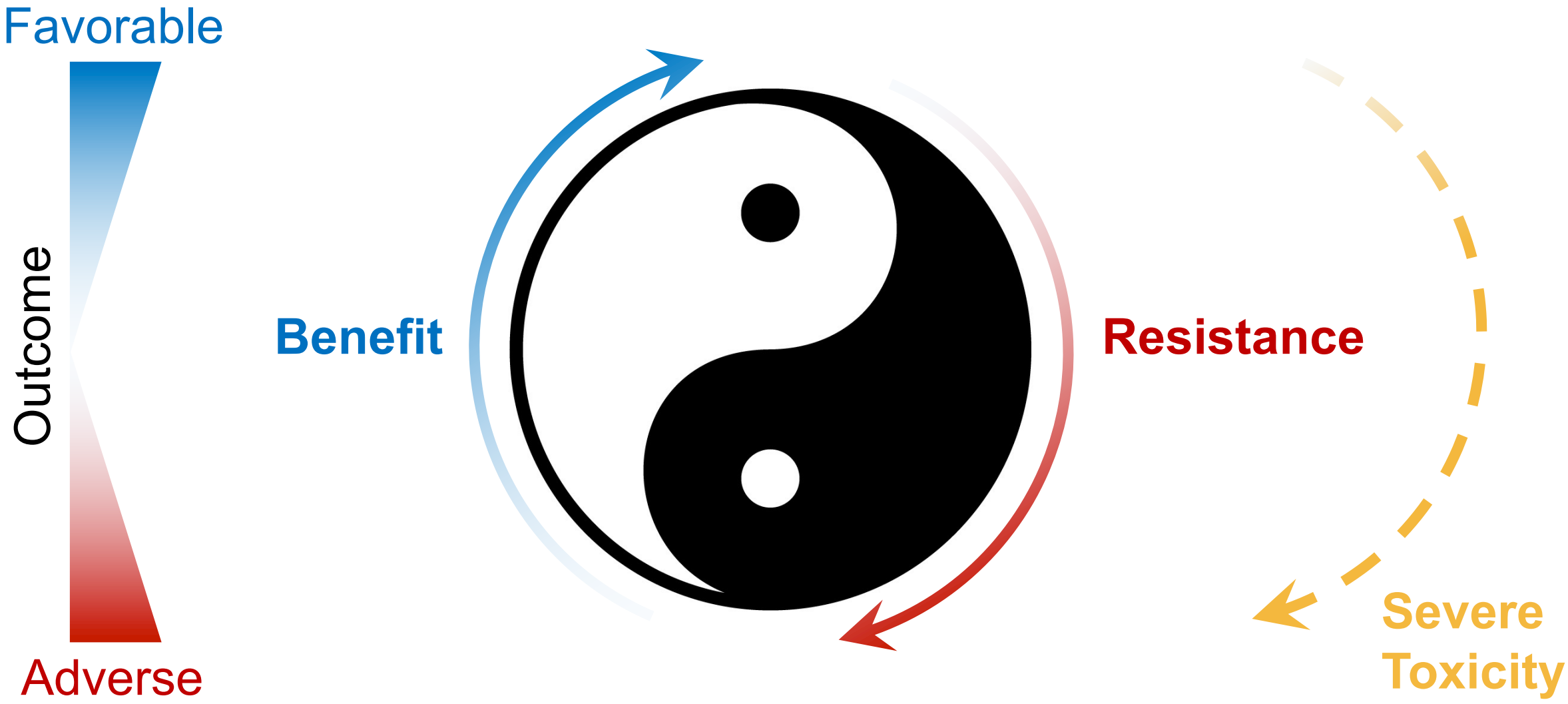


https://pct.mdanderson.org

What We Do


Developmental Hierarchies
Tumor
Population-level analyses
Composition
Precision medicine
Ecosystems

Today: Discovering cancer resistance mechanisms with data viz



Why data visualization (data viz)?

Pictures reveal hidden content –
Map of cell phone towers illuminates densely and sparsely populated areas (and their connections)

https://alpercinar.com/open-cell-id/

Why data visualization: *Seeing is believing*

Seeing is believing

Scientists can often make indirect measurements that tell us about things we can't actually see. For scientists who work on molecules, such as myself, this is especially true: Many of the small and large molecules that dance in my head are objects that I've never actually looked at. But for many outside of science, seeing is believing.

In my first administrative job at the University of North Carolina, I learned about this while running the campus planetarium. On clear nights, we would set up telescopes for public viewings. It was common for people to see Saturn through the telescope for the first time and then frantically look to see whether we had taped a cartoon of the ringed planet to the end of the telescope. They had assumed that Saturn didn't really look like the pictures in their grade-school classrooms.

While I was in that job almost 20 years ago, I was fortunate enough to convince the authors Will and Mary Pope Osborne to work with the university on a planetarium show based on their blockbuster children's book series *Magic Tree House*. At the most suspenseful part of the show, the protagonists Jack and Annie end up dangerously close to the event horizon of a black hole—the to talk about it today. (Spoiler alert: Jack and Annie are rescued from spaghettification at the event horizon by Mary Pope Osborne herself.)

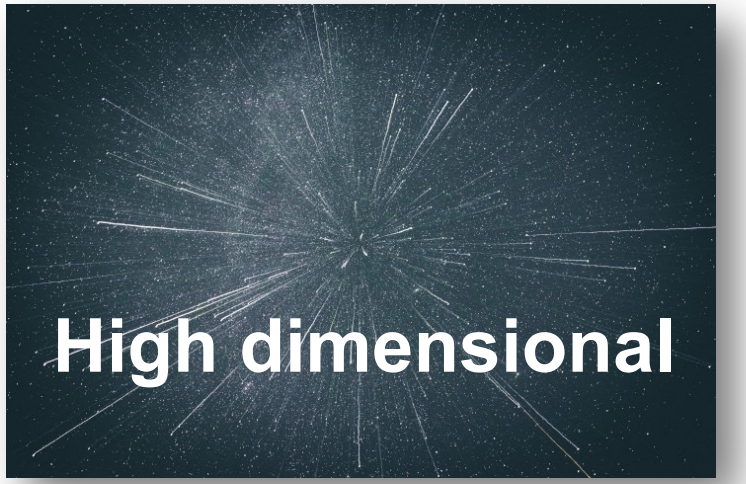
If we made the show today, we wouldn't have to guess at what the black hole looks like. The image of the event horizon of the supermassive black hole in the nearby galaxy Messier 87 was a magnificent technical achievement and a worthy Breakthrough of the Year. But it is more than that. For a skeptical public that often rolls their eyes when they hear scientists say that they know things exist even though they cannot be seen, this is one more important object that we can see. Given the influence of black holes on the evolution of galaxies, this is a remarkable milestone in every respect.

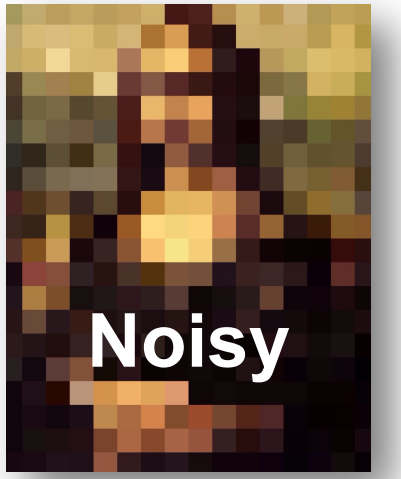
There were also some extraordinary runners-up this year. When I was at Washington University in St. Louis, I had the privilege of watching research progress on restoring the gut microbiome in malnourished children. It's intensely encouraging to know that there is a way to do this, and the companion papers that show how the microorganisms develop make it great science, as well. This has implications for public health in the developed world, too: Children need to start with excellent

H. Holden Thorp
Editor-in-Chief,
Science journals.
hthorp@aaas.org;
@hholdenthorp

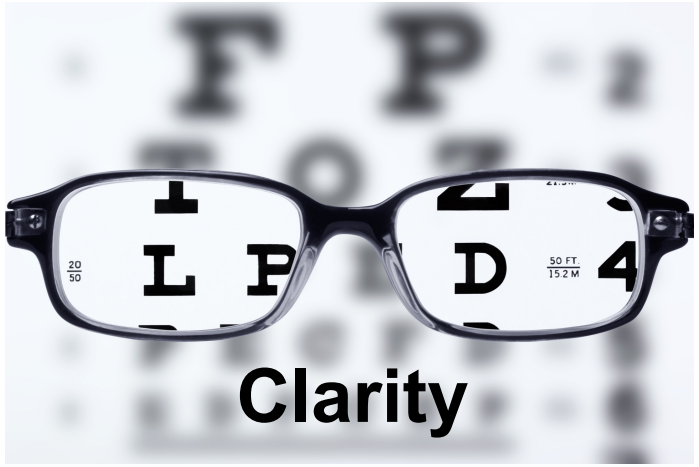
Why data visualization: *Omics data are otherwise impenetrable*

Omics data are …





Your goal is …



- Science without **effective communication** is content without delivery
- Figures are the **center** of attention
- **Data visualization** can facilitate scientific discovery

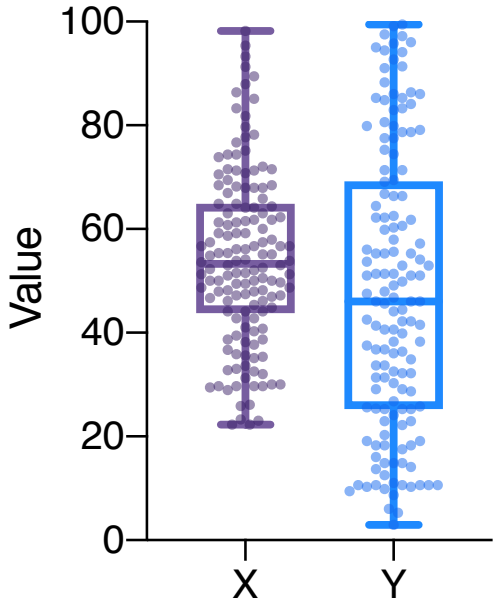
Exploratory visualization can be critical

Input data → **Now what?**

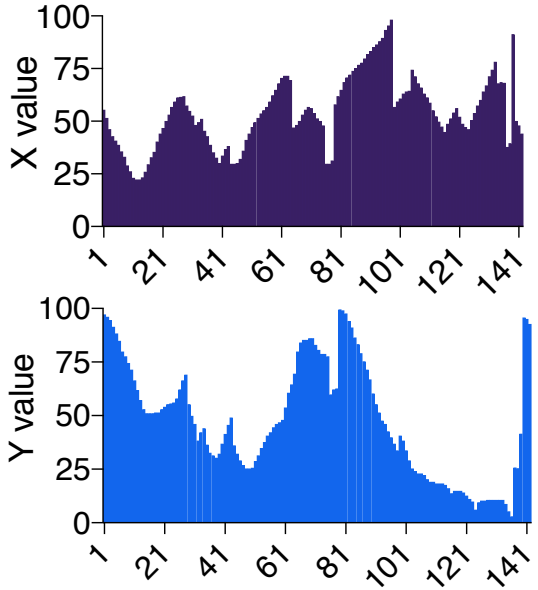
X	Y
55.3846	97.1795
51.5385	96.0256
46.1538	94.4872
42.8205	91.4103
40.7692	88.3333
38.7179	84.8718
35.641	79.8718
33.0769	77.5641
28.9744	74.4872
26.1538	71.4103
⋮	

142 rows

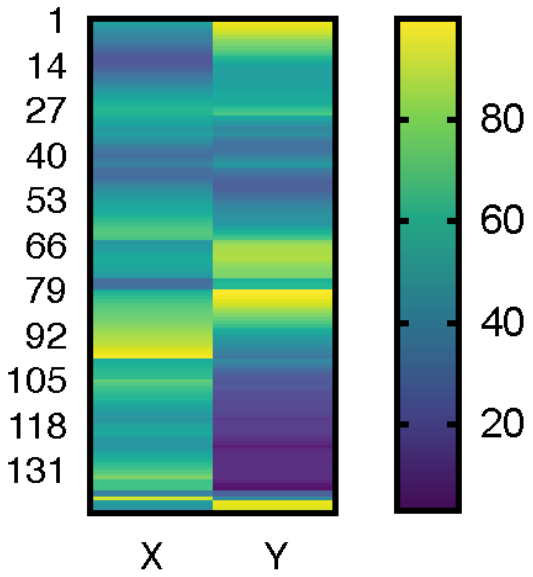
Box plot?



Bar plots?



Heat map?

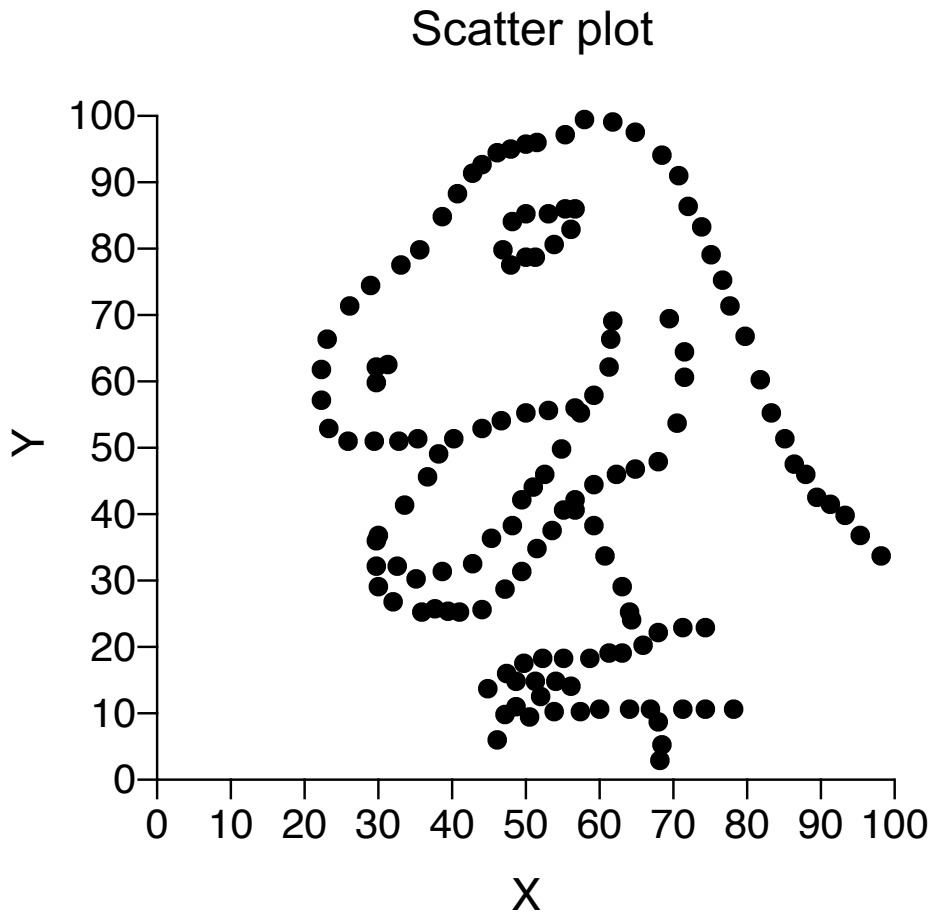


Know your data!

Input data

X	Y
55.3846	97.1795
51.5385	96.0256
46.1538	94.4872
42.8205	91.4103
40.7692	88.3333
38.7179	84.8718
35.641	79.8718
33.0769	77.5641
28.9744	74.4872
26.1538	71.4103
⋮	

142 rows



Today's agenda

- Introduction to data visualization
  - Common plots
  - How to make plots
- Application of data viz to immuno-oncology
- Caveats of data viz
- General tips and best practices
- Resources

Introduction to data visualization

- **Exploratory analysis**
  - Quality control
    - E.g., Outlier identification
  - Discovery
- **Figure generation**
  - Ten Simple Rules for Better Figures (Rougier et al., *PLOS Comp Biol* 2014): https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
- **Fundamentals of data viz**
  - https://clauswilke.com/dataviz/
  - https://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

How to do data visualization?

R and Python

- Preferred for reproducible figures and exploratory plots

Microsoft Excel/PowerPoint

- Useful for exploratory charts and simple schematics; from PowerPoint, can export as PDF or directly copy and paste into Illustrator for polishing

GraphPad Prism

- Publication quality, but limited to smaller datasets
- Tips: use ½ point line/axis thickness, remove bolding, use Helvetica font

Adobe Illustrator

- Schematics; polishing of figures made elsewhere; arranging multi-panel figures for publication

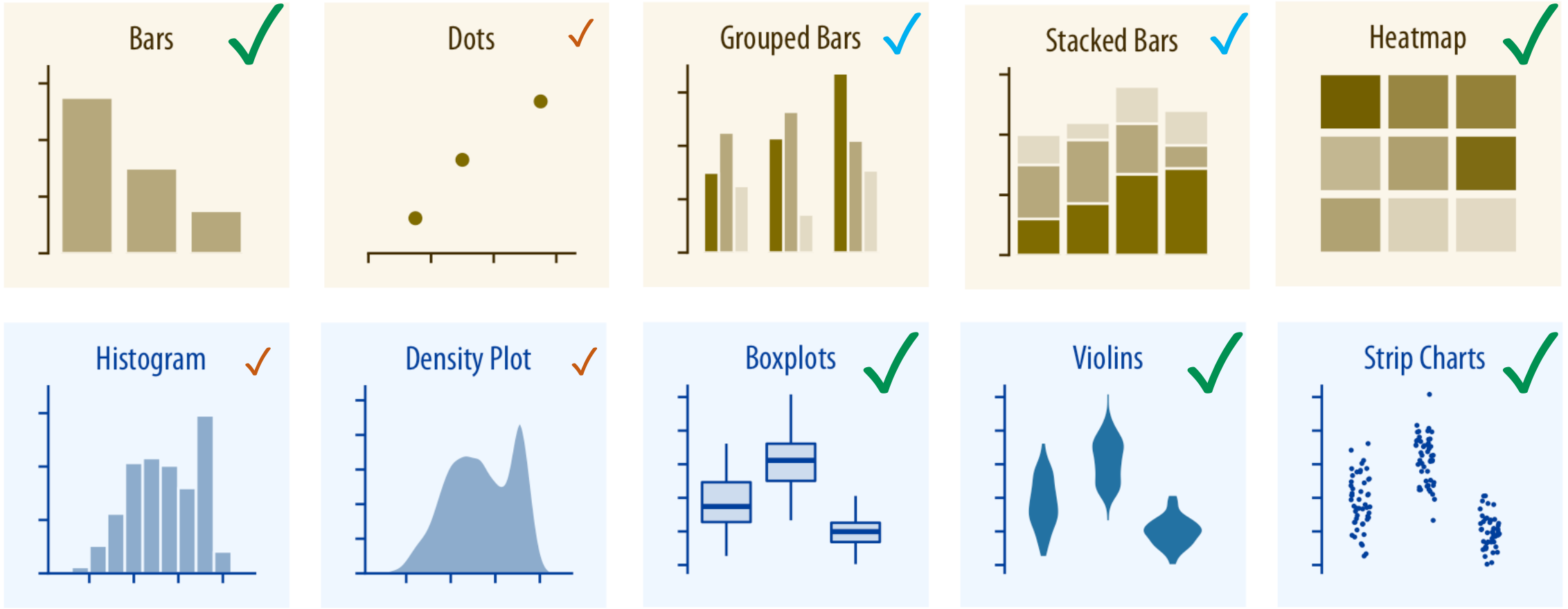
BioRender for schematics focused on biological sciences

Key graphical packages in R and Python

- Core libraries
- ggplot2
- ComplexHeatmap (Bioconductor)

- Plotly
- Matplotlib (fully customizable)
- Seaborn
  - Higher level version of Matplotlib with less options but easier to use

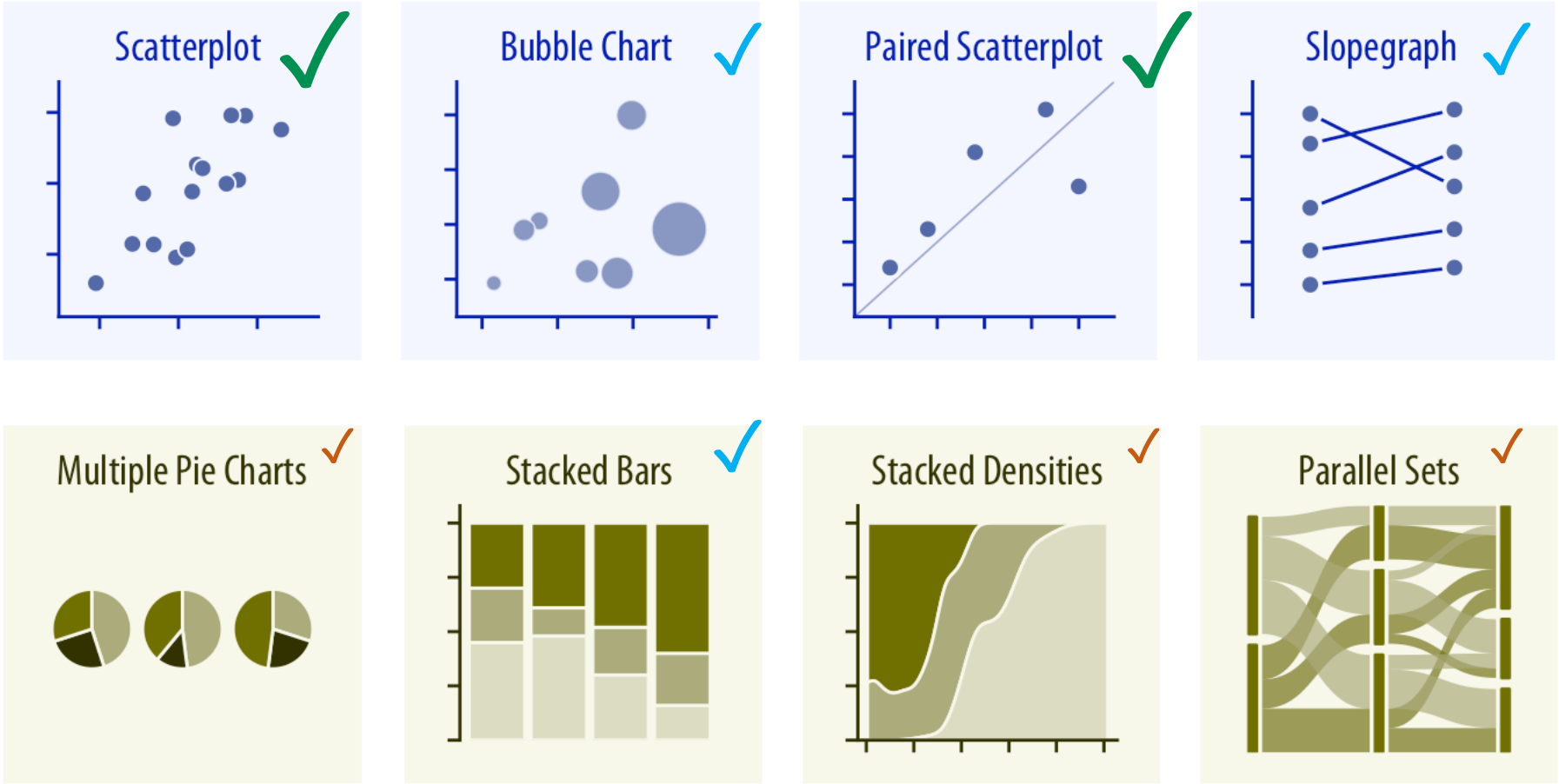
Basic plots



✓ Indispensable ✓ Often used ✓ Less common

https://clauswilke.com/dataviz/

Basic plots (cont.)



✓ Indispensable ✓ Often used ✓ Less common

https://clauswilke.com/dataviz/

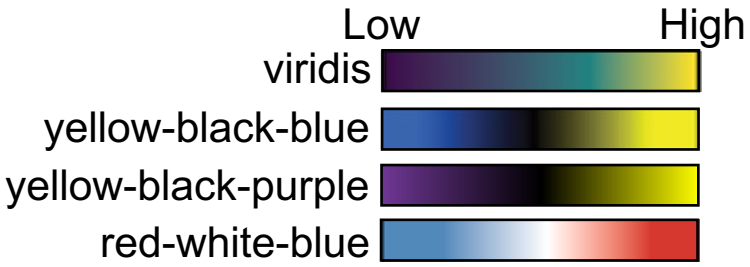
Heat maps

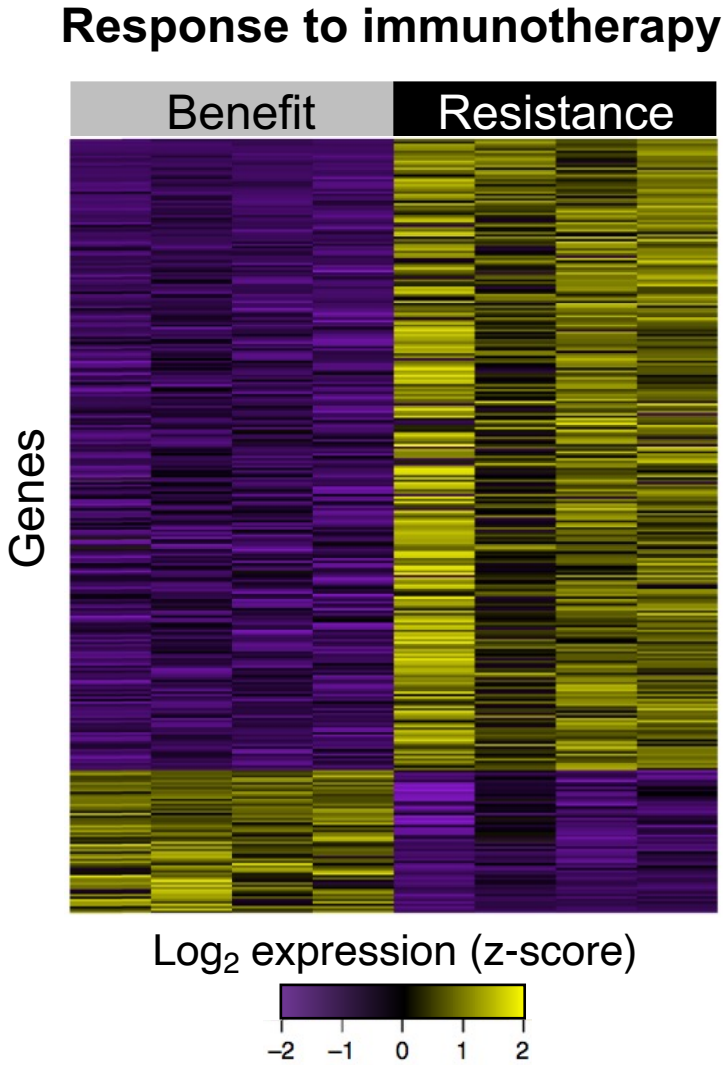
Visualize patterns in high-dimensional data, e.g., coordinately expressed genes, accessible chromatin, ChIP-seq peaks, etc.

Scaling and normalization are critical

Rows often expressed as z-scores

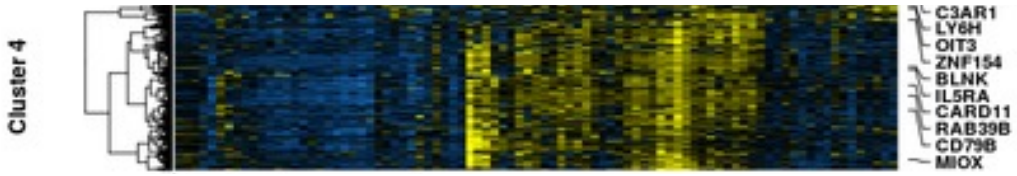
Popular color schemes (red-green color-blind accessible)

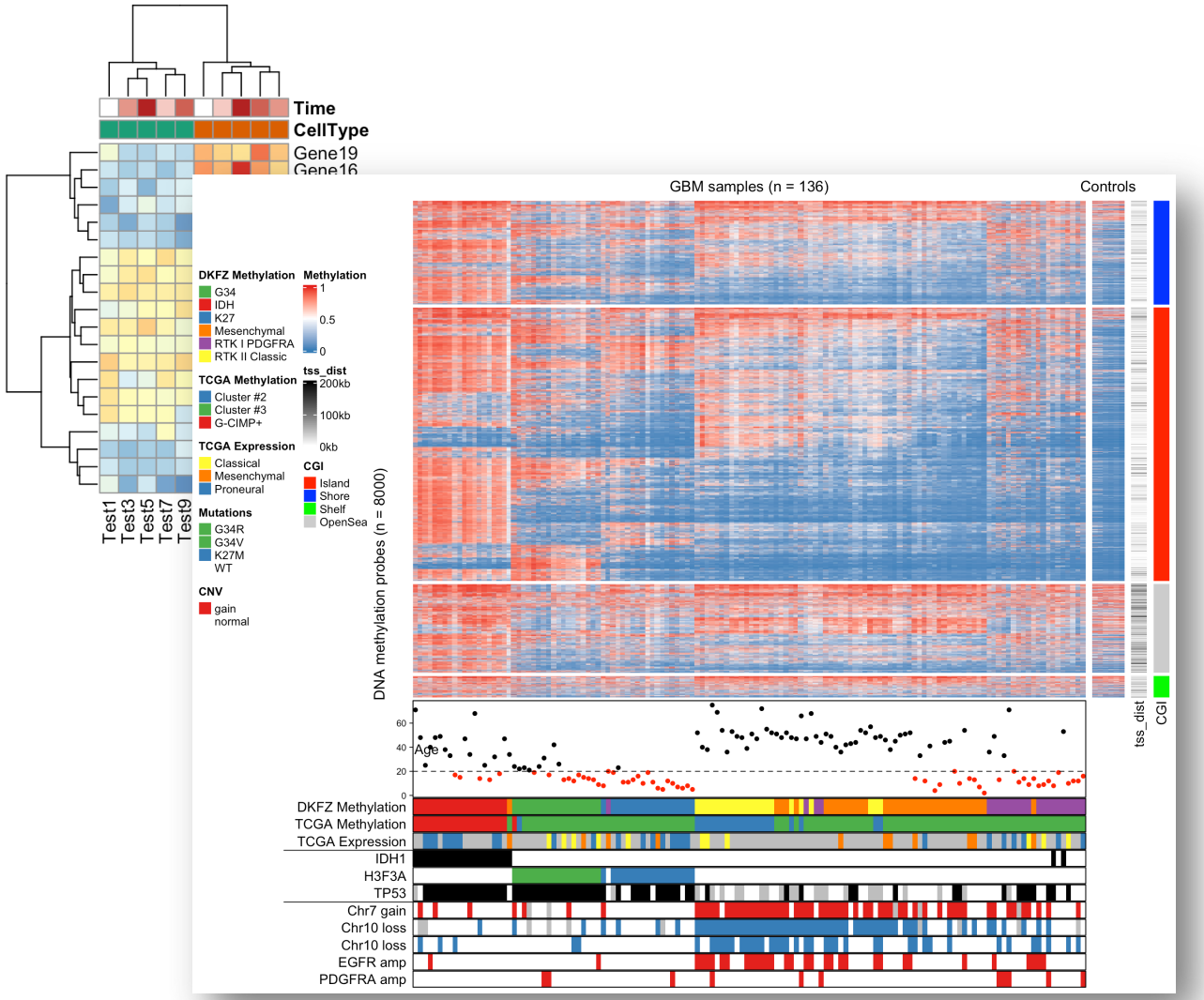




ComplexHeatmap (R)

- Bioconductor package for layering meta-data and other plots with one or more heat maps
- Highly customizable
- Supports multiple omic-style visualizations
- Addresses label crowding

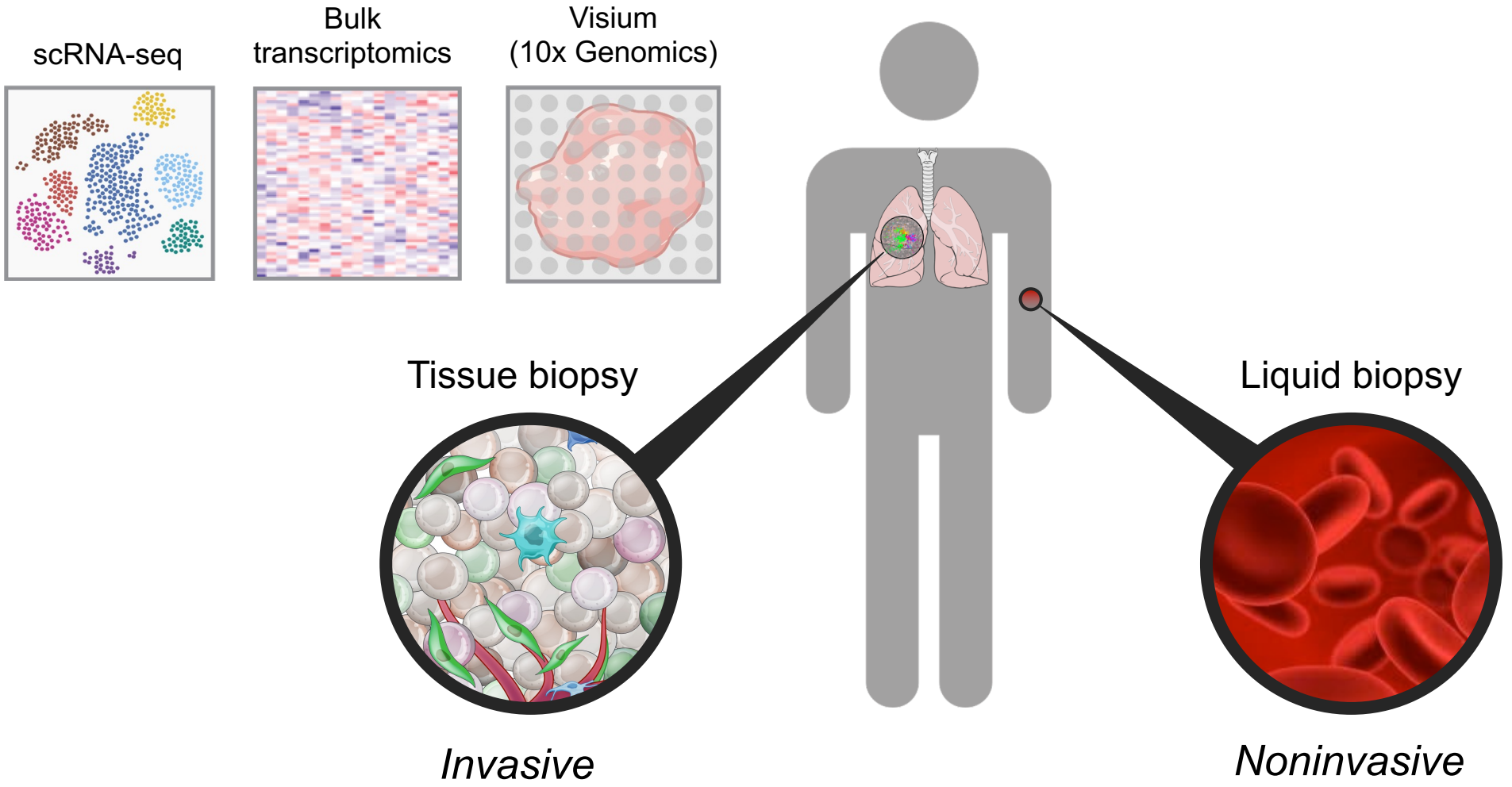




https://jokergoo.github.io/ComplexHeatmap-reference/book/

Tutorial with data from a publication: https://github.com/kevinblighe/E-MTAB-6141

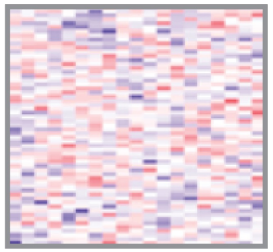
Data visualization for immuno-oncology



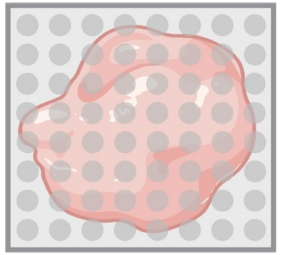
Data visualization for immuno-oncology

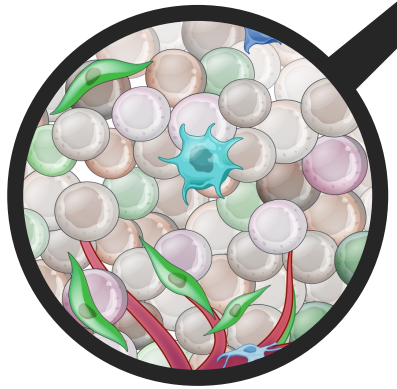
scRNA-seq

Bulk transcriptomics



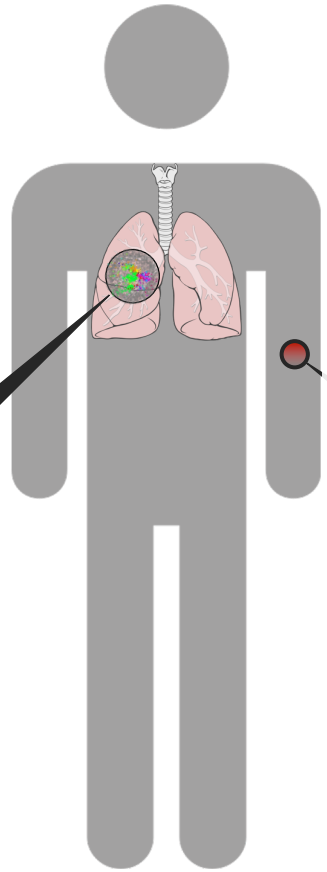
Visium (10x Genomics)





Tissue biopsy

Invasive



Liquid biopsy

Noninvasive

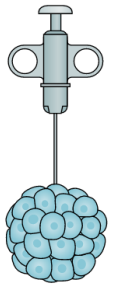
Data viz for **DNA sequencing data**

- Extensively covered elsewhere; see, e.g., https://genviz.org/

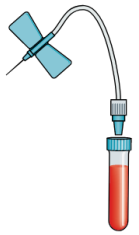
Combining single-cell and bulk assays for immunotherapy profiling

Discovery
- Few patients
- Extensive biospecimen collection
- Single-cell, high-dimensional analysis
- Technically challenging
- High cost relative to validation

Tumour biospy

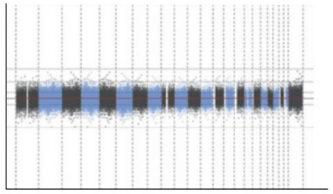


Peripheral blood (for PBMCs)

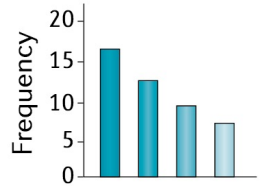


scRNA-seq/scTCR-seq/ Hi-Dim cytology (cryopreserved specimens)

Validation
- Many patients
- Limited biospecimen collection
- Conventional, low-dimensional analysis
- Technically straightforward
- Low cost relative to discovery



Whole-exome and bulk RNA sequencing (frozen/OCT)

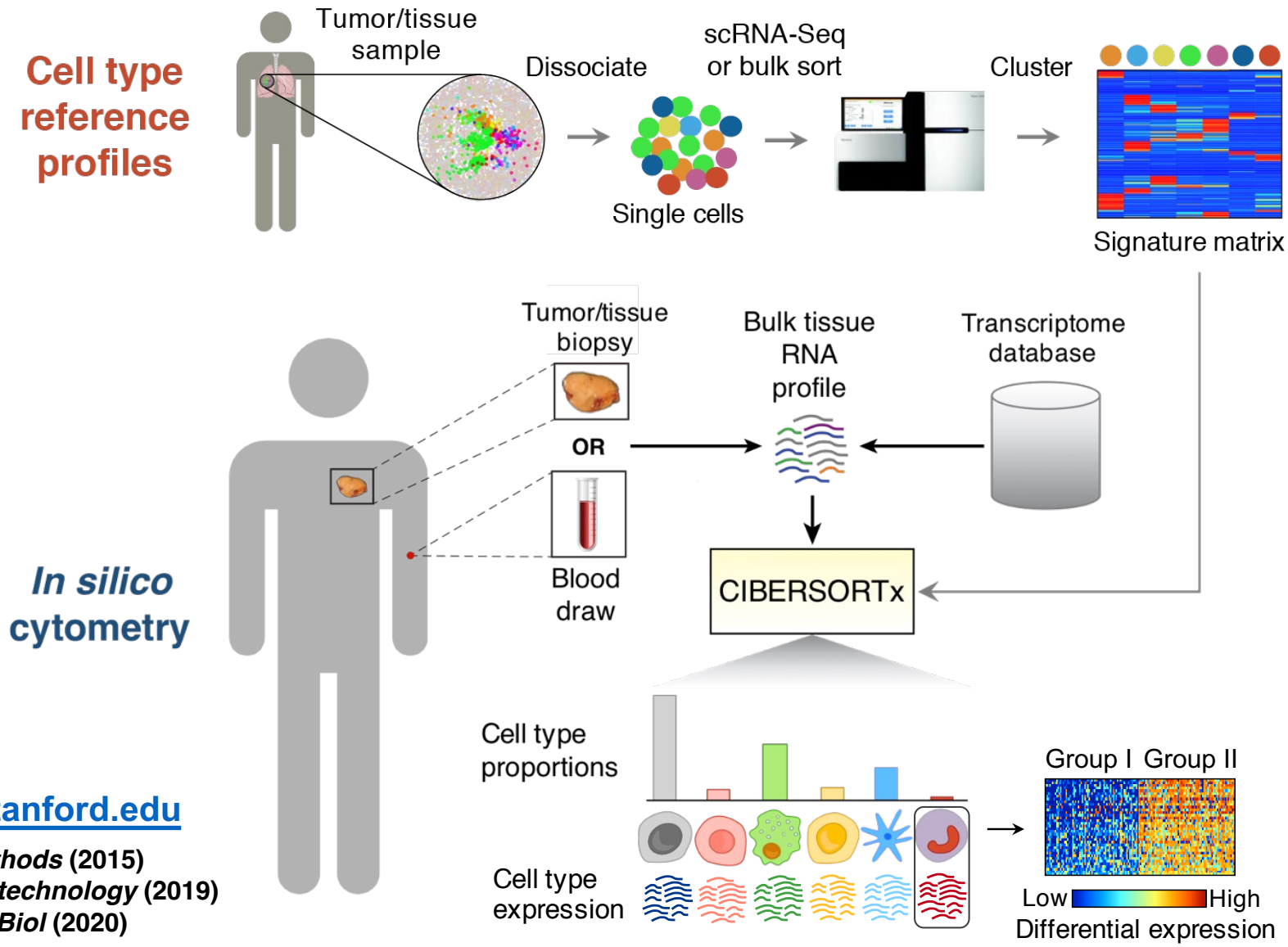


Bulk TCR repertoire analysis (cryopreserved PBMCs, frozen tumour)

Use discovery single-cell data and apply analytical tools (e.g. CIBERSORTx) to impute cell fractions/states from bulk data

(Gohil et al., ***Nat Rev Clin Oncol*** 2021 18:244-256)

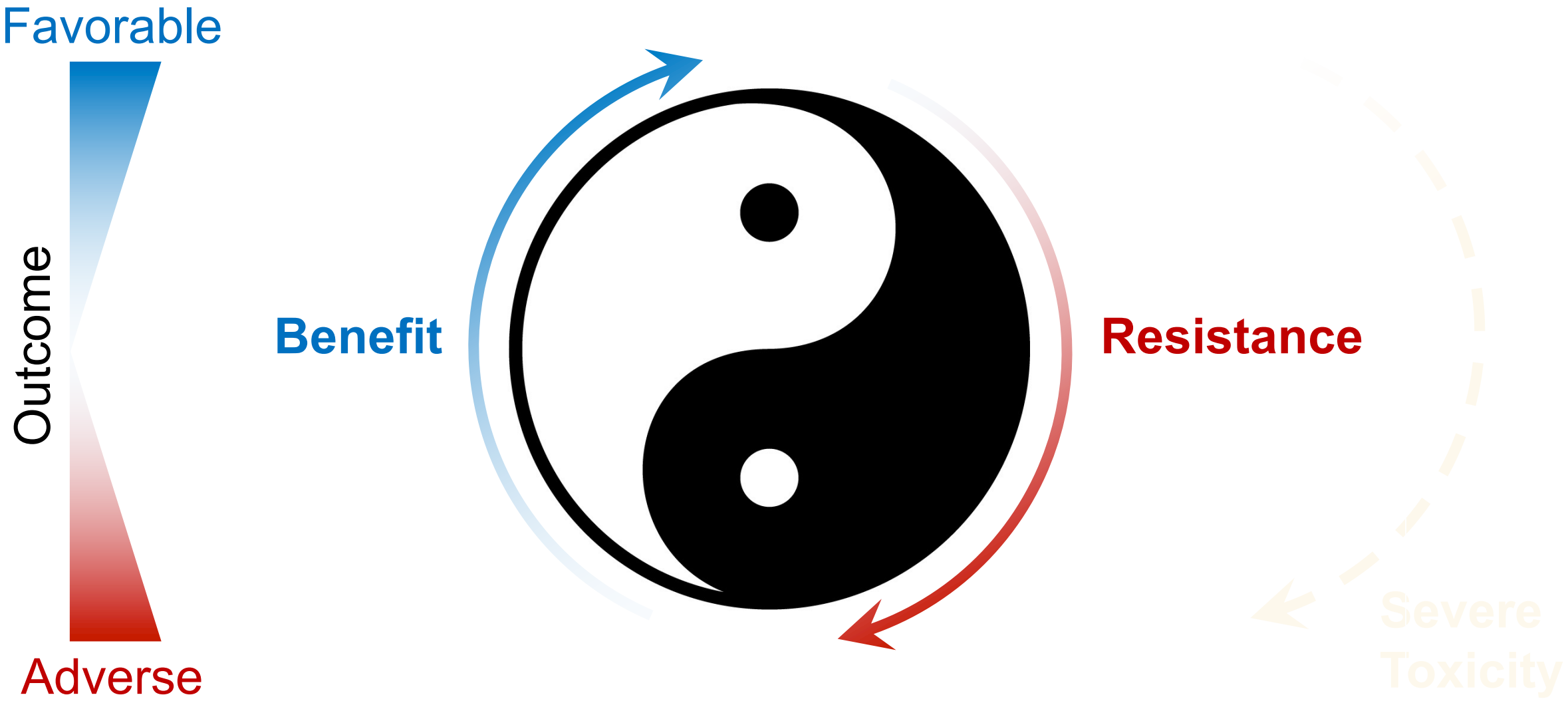
Digital cytometry with CIBERSORTx



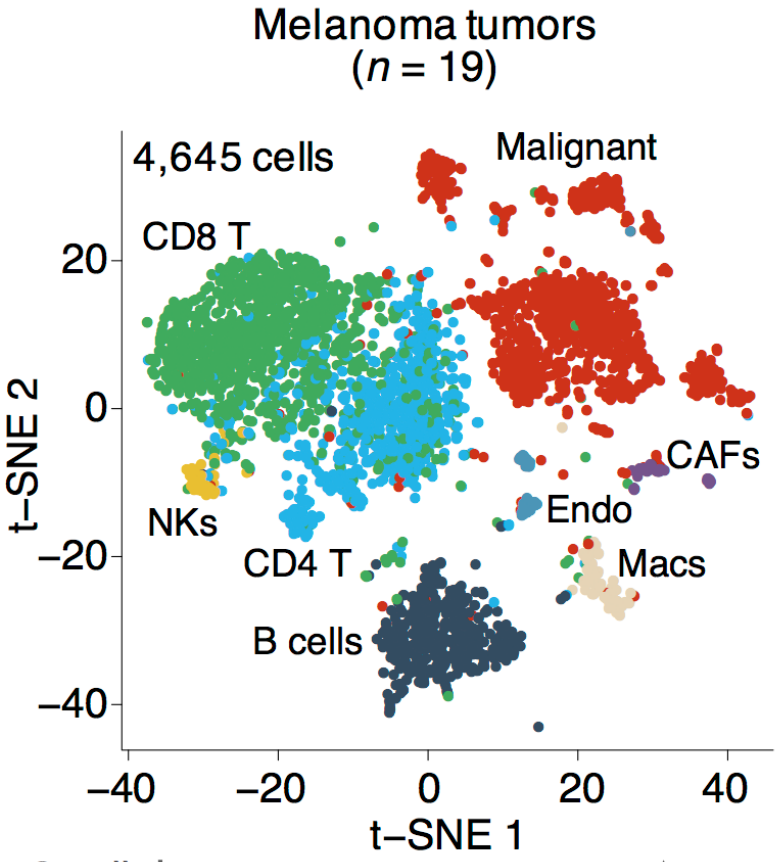
https://cibersortx.stanford.edu

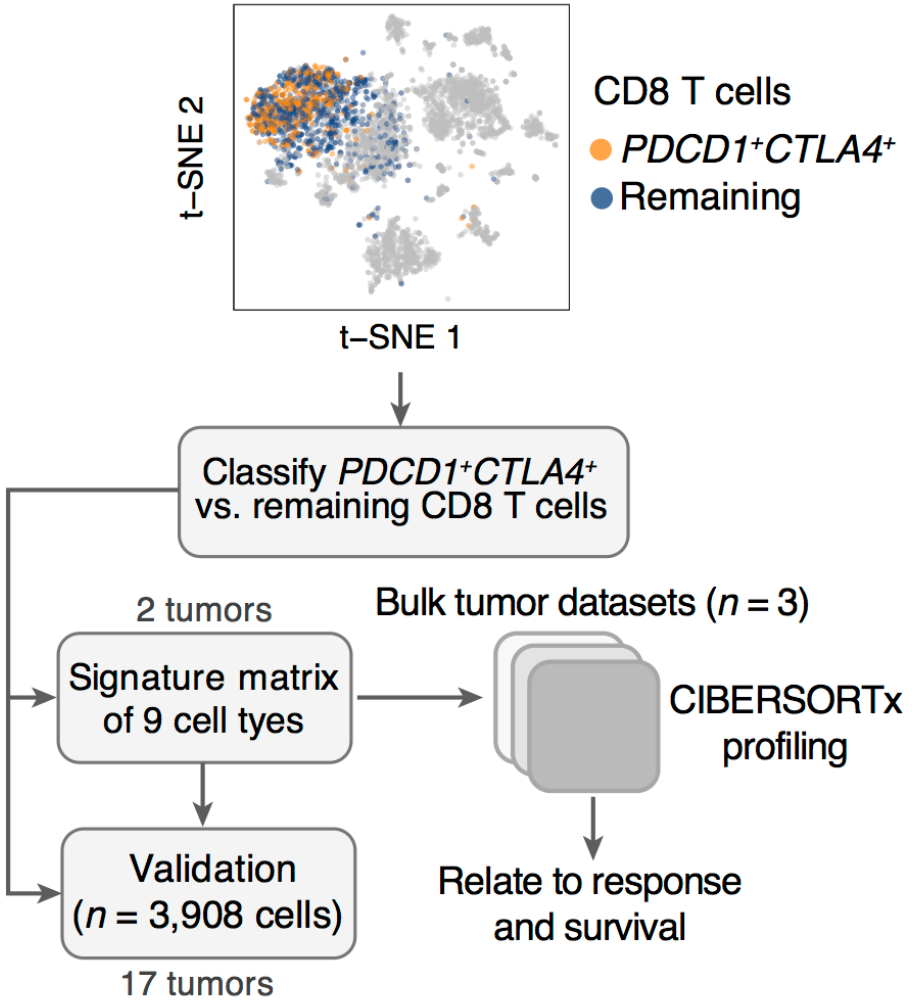
Newman et al., *Nature Methods* (2015)
Newman et al., *Nature Biotechnology* (2019)
Steen et al., *Methods Mol Biol* (2020)

Discovering cancer resistance mechanisms



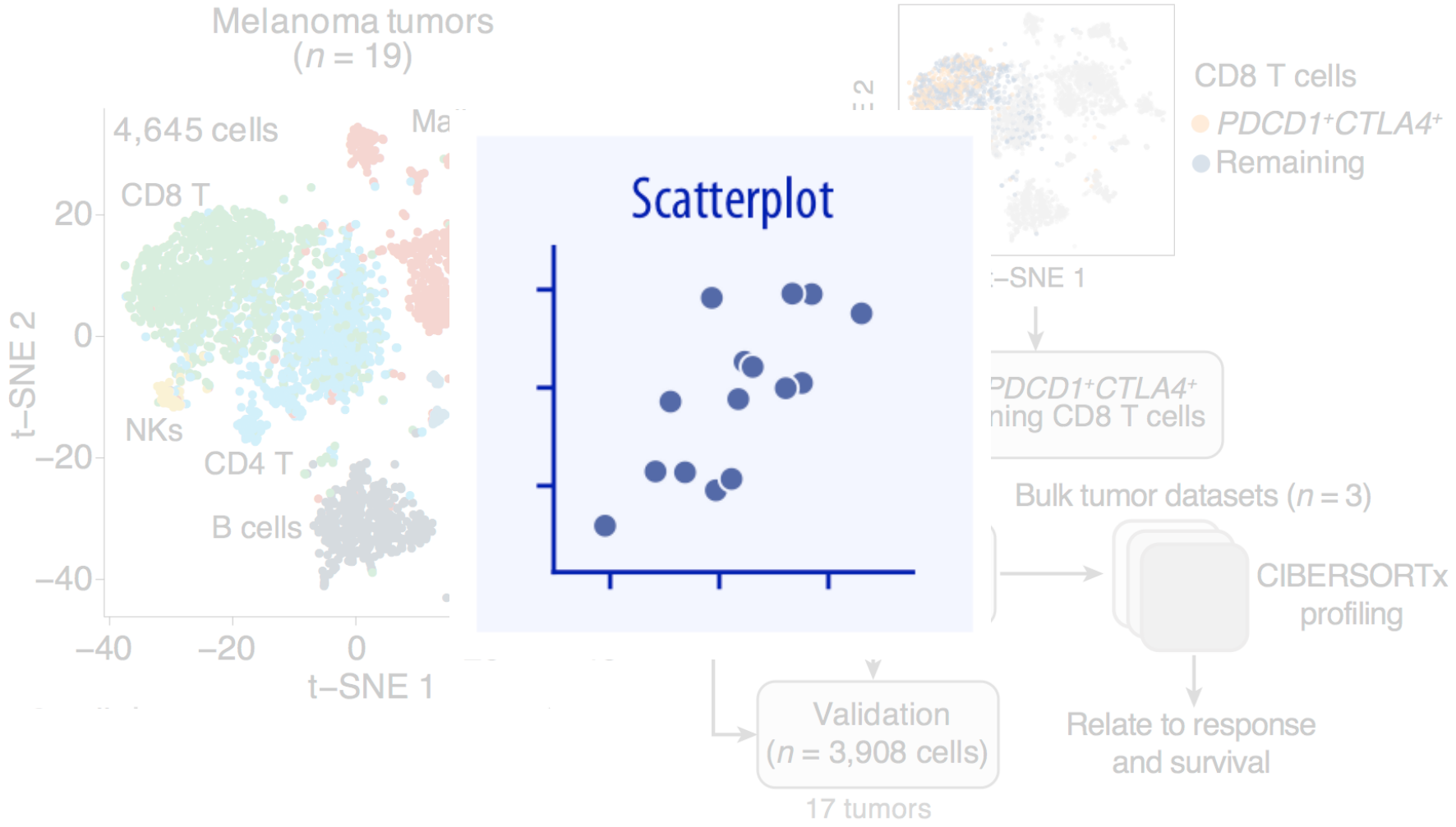
Single-cell reference maps for cellular biomarker discovery





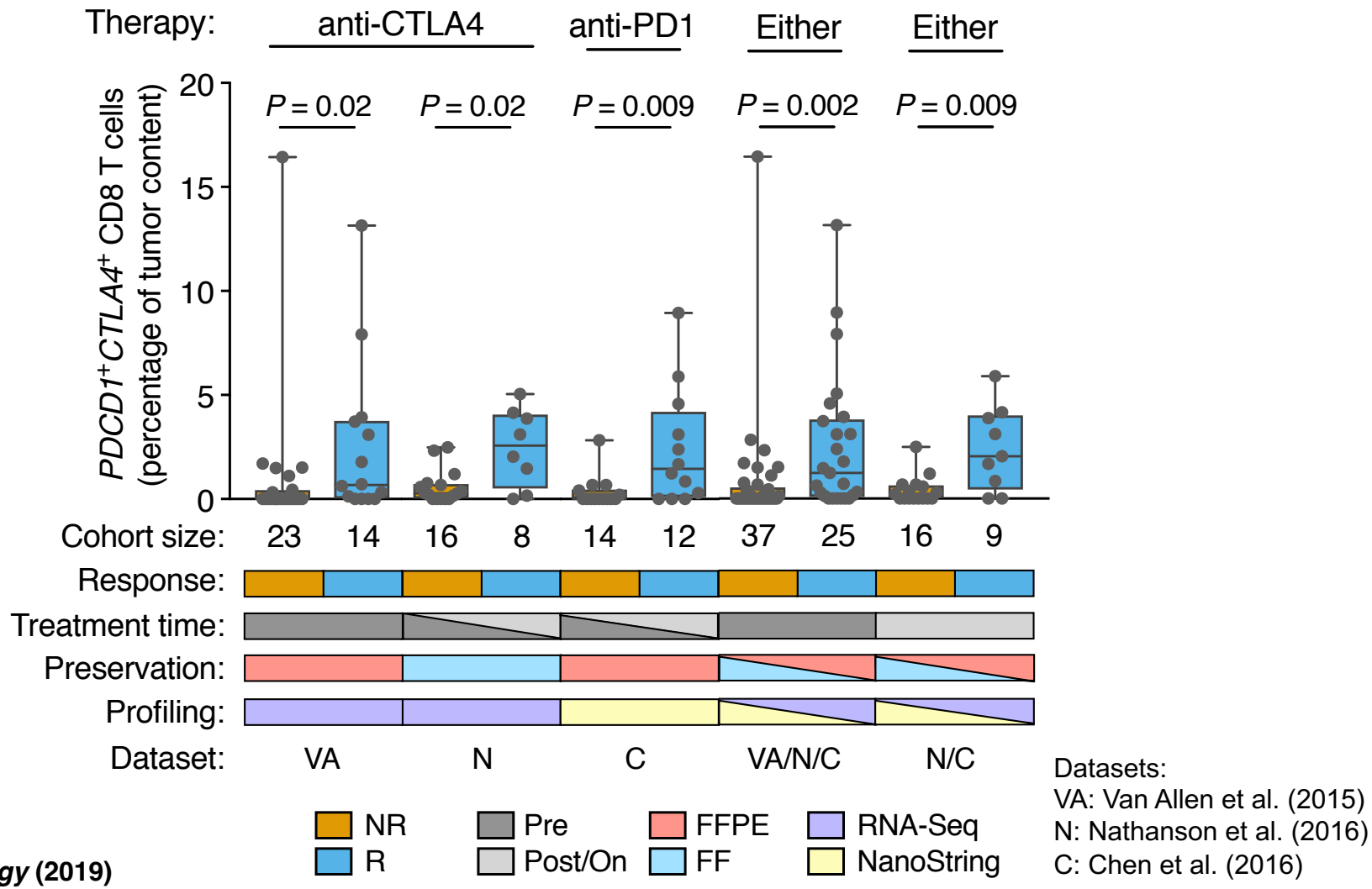
Newman et al., *Nature Biotechnology* (2019)

Single-cell reference maps for cellular biomarker discovery



Newman et al., *Nature Biotechnology* (2019)

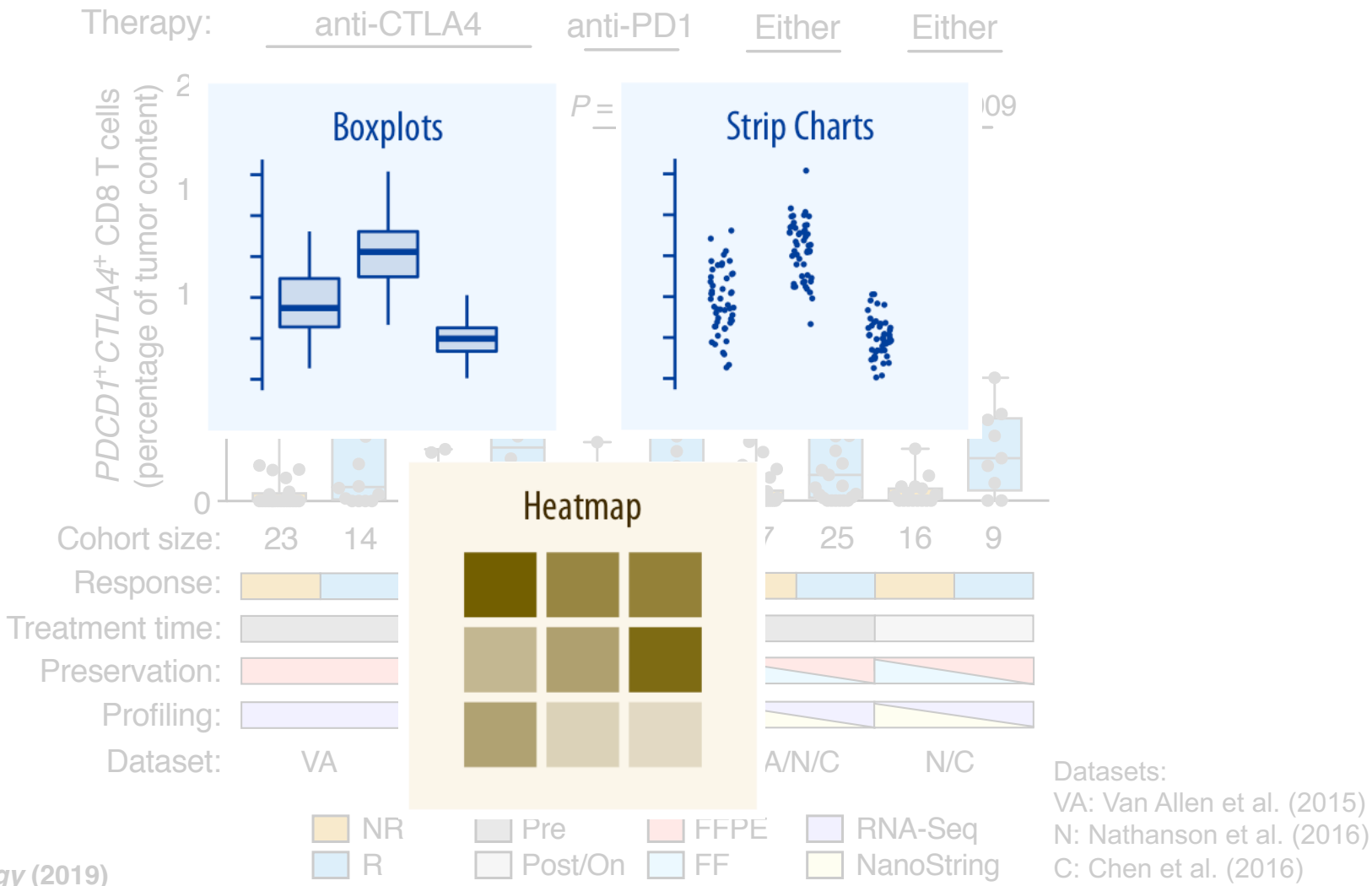
Association of *PDCD1*$^{+}$/*CTLA4*$^{+}$ CD8 TILs with response to immune checkpoint blockade in patients with melanoma



Datasets:
VA: Van Allen et al. (2015)
N: Nathanson et al. (2016)
C: Chen et al. (2016)

Newman et al., *Nature Biotechnology* (2019)

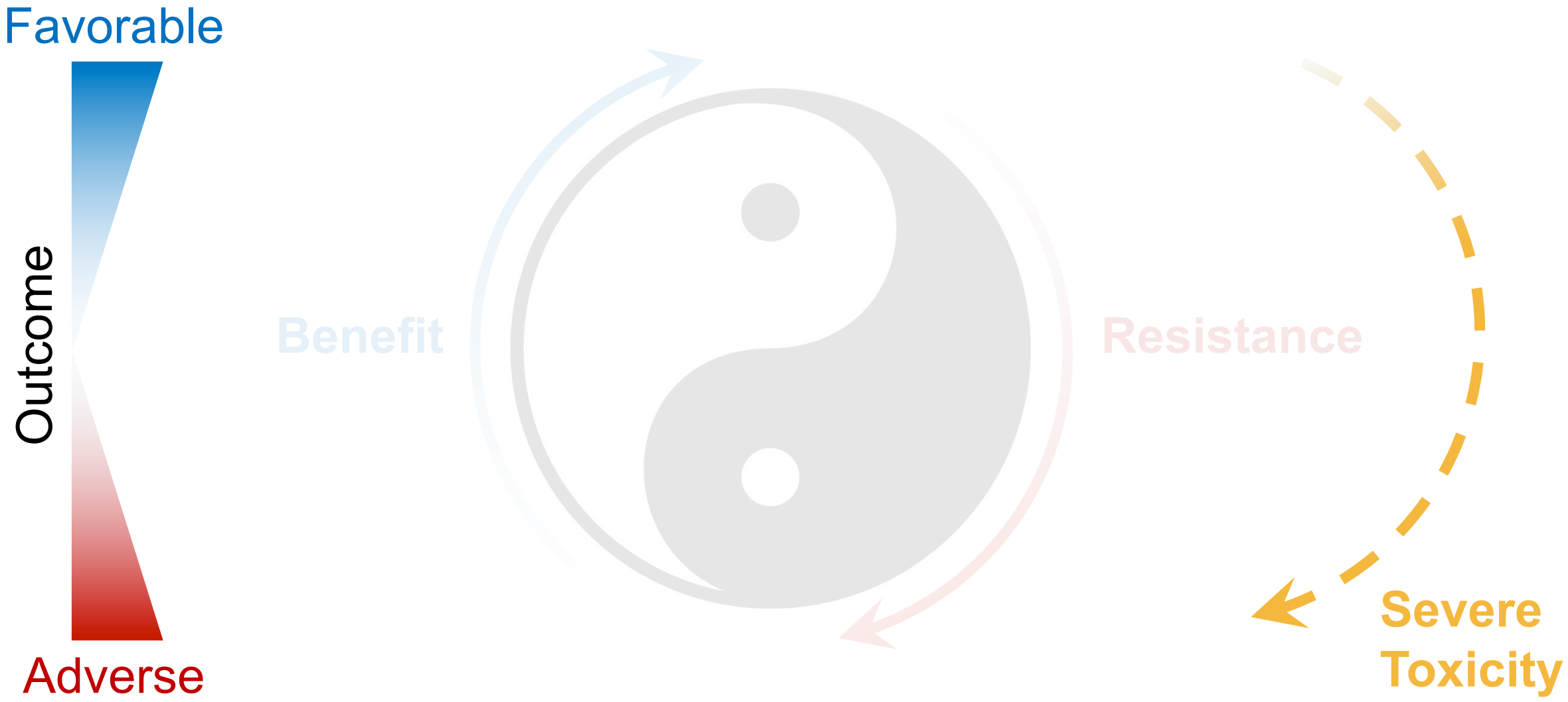
Association of *PDCD1*$^{+}$/*CTLA4*$^{+}$ CD8 TILs with response to immune checkpoint blockade in patients with melanoma



Datasets:
VA: Van Allen et al. (2015)
N: Nathanson et al. (2016)
C: Chen et al. (2016)

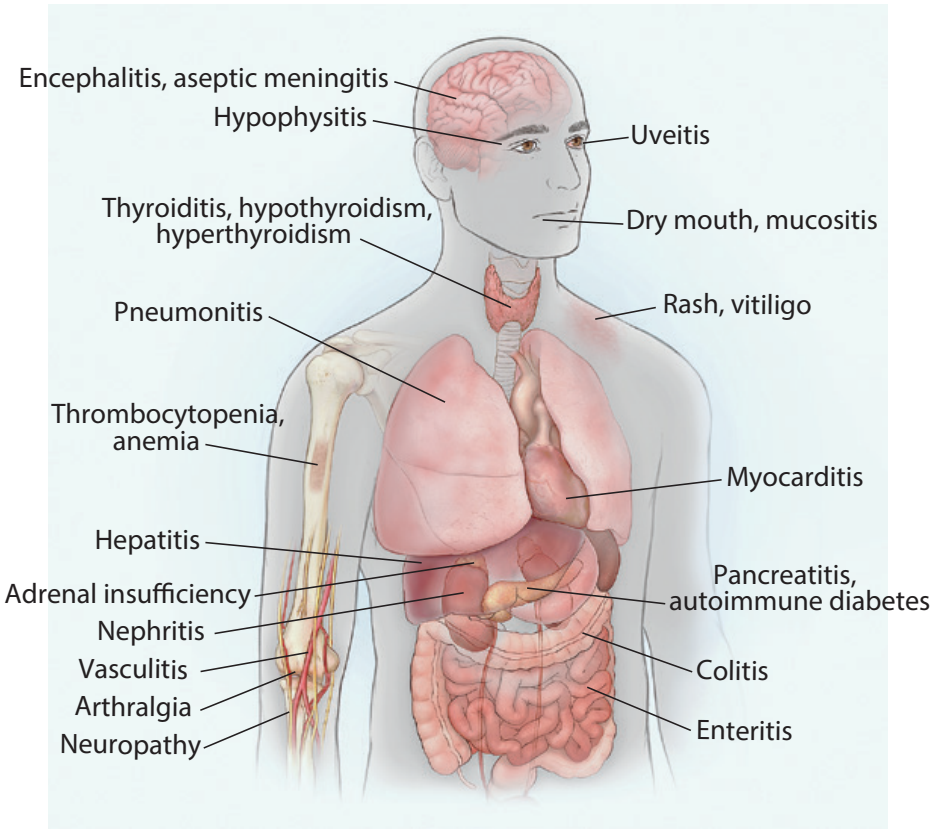
Newman et al., *Nature Biotechnology* (2019)

Discovering toxicity mechanisms



Toxicity is ”the dark side” of cancer immunotherapy

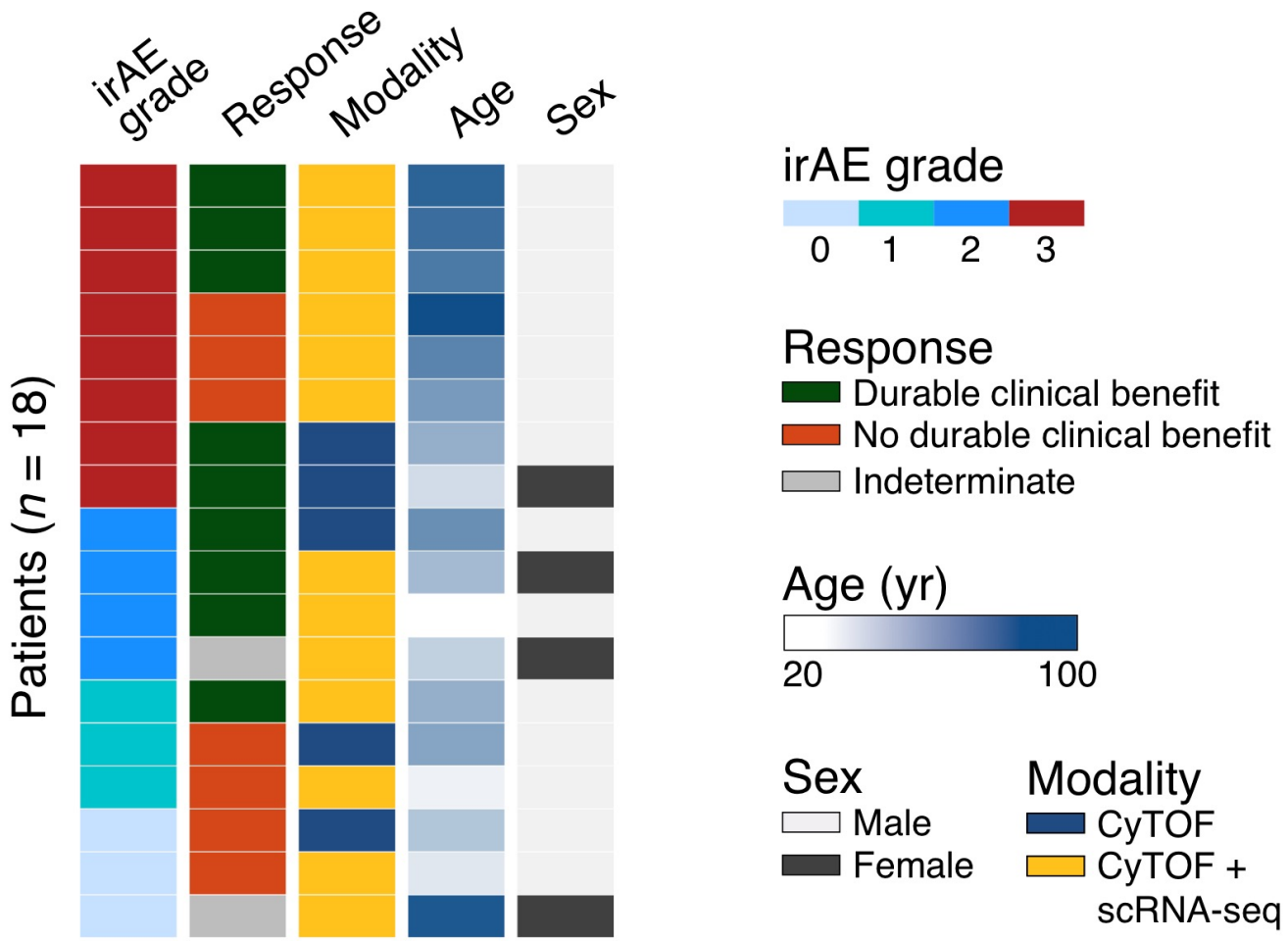
adapted from **Postow et al. *NEJM* (2019)**



- **Immune-related adverse events (irAEs)** can occur in **any** organ system
- Immune checkpoint inhibitor (ICI) efficacy hindered by irAEs (Wolchok et al., *NEJM* 2017): Checkmate 067
  - 59% experienced **grade 3 to 4 irAEs** with combination ICIs
  - 39% experienced irAEs that led to **treatment discontinuation**
- Pathogenesis remains **unclear**

No toxicity	Asymptomatic	Symptomatic	Severe	Life threatening	Death
Grade 0	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**	**Grade 5**

Single-cell discovery cohort



Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

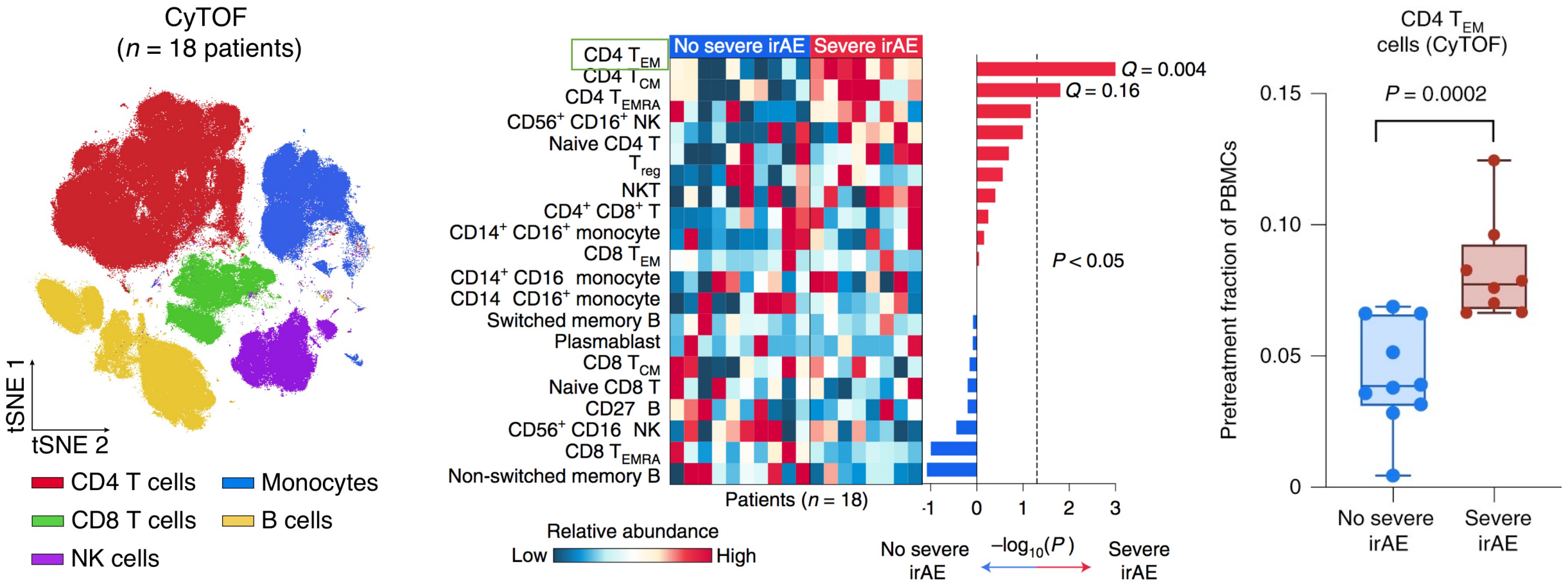
Single-cell discovery cohort

Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

CyTOF

Determinants of severe irAEs from pretreatment blood

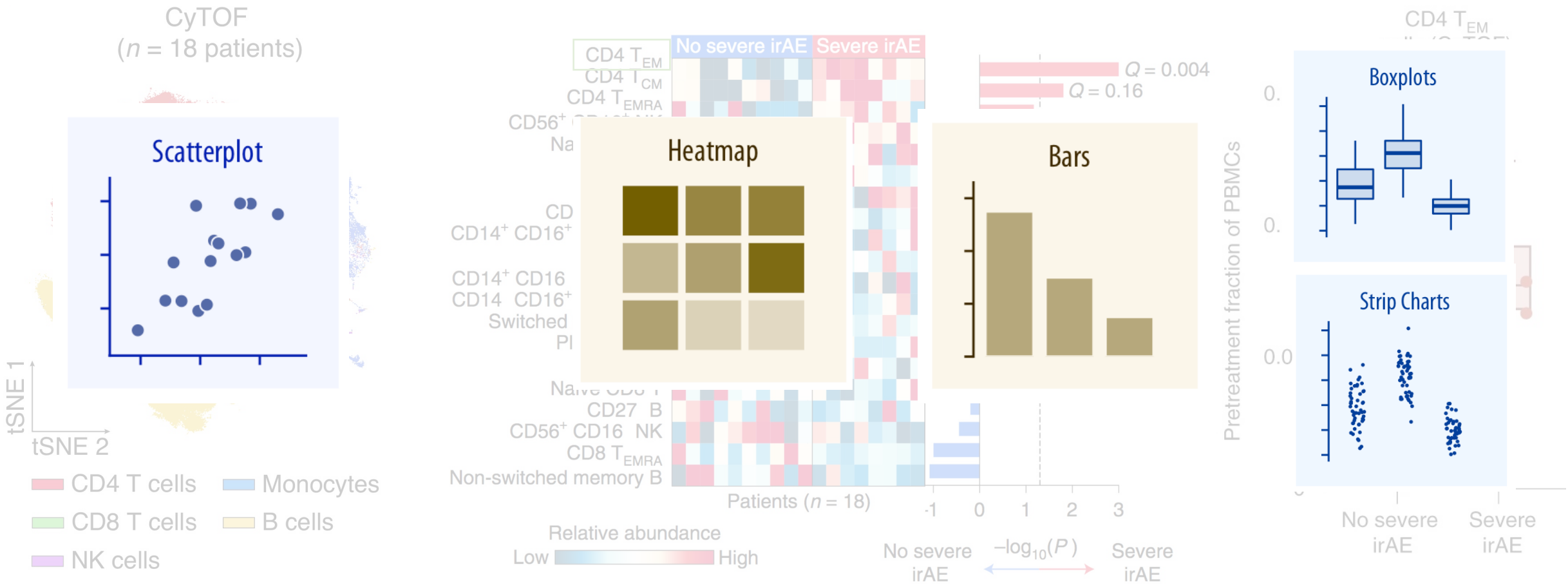
Elevated CD4 T_{EM} cells in pretreatment blood significantly associated with severe irAEs



Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

Determinants of severe irAEs from pretreatment blood

Elevated CD4 T_{EM} cells in pretreatment blood significantly associated with severe irAEs

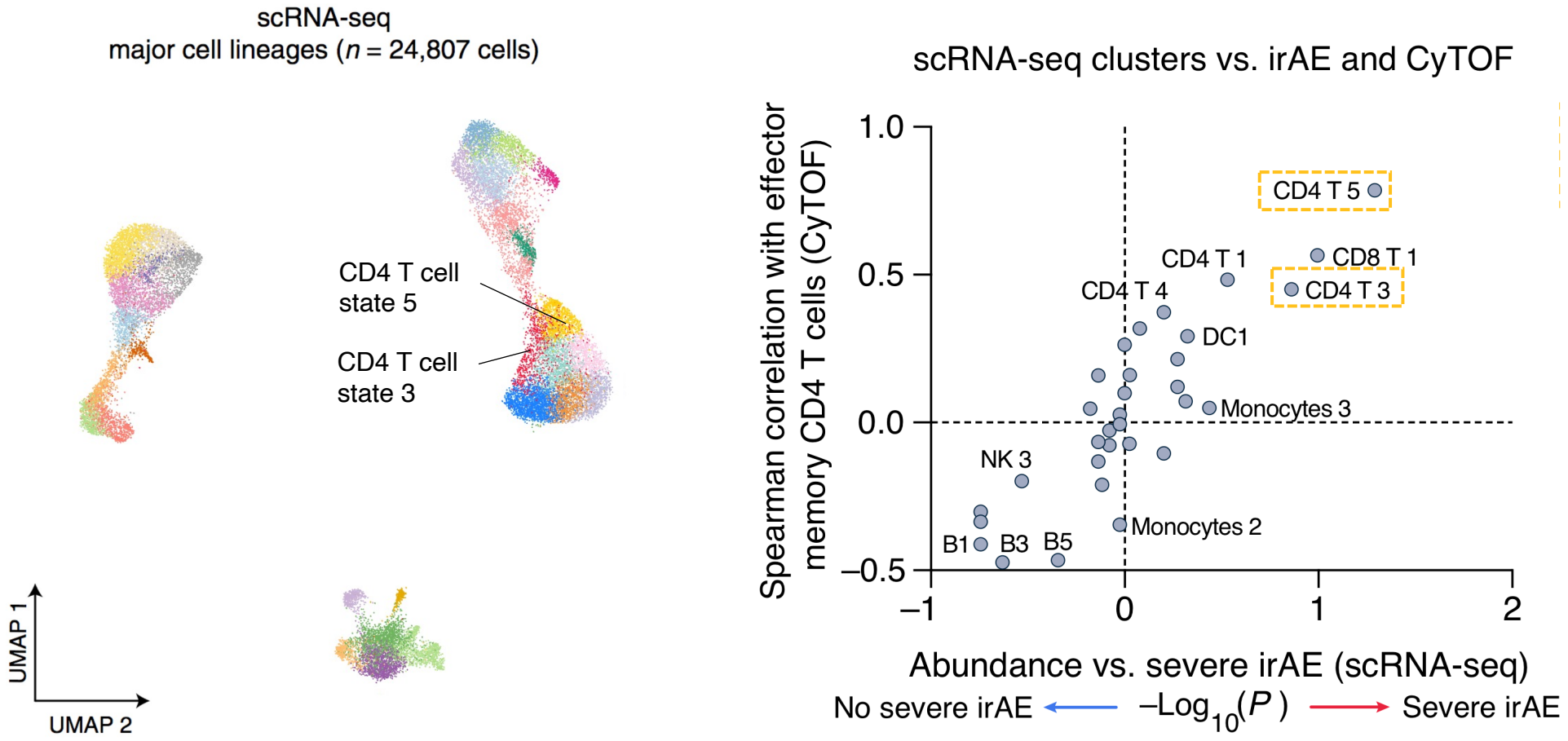


Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

scRNA-seq

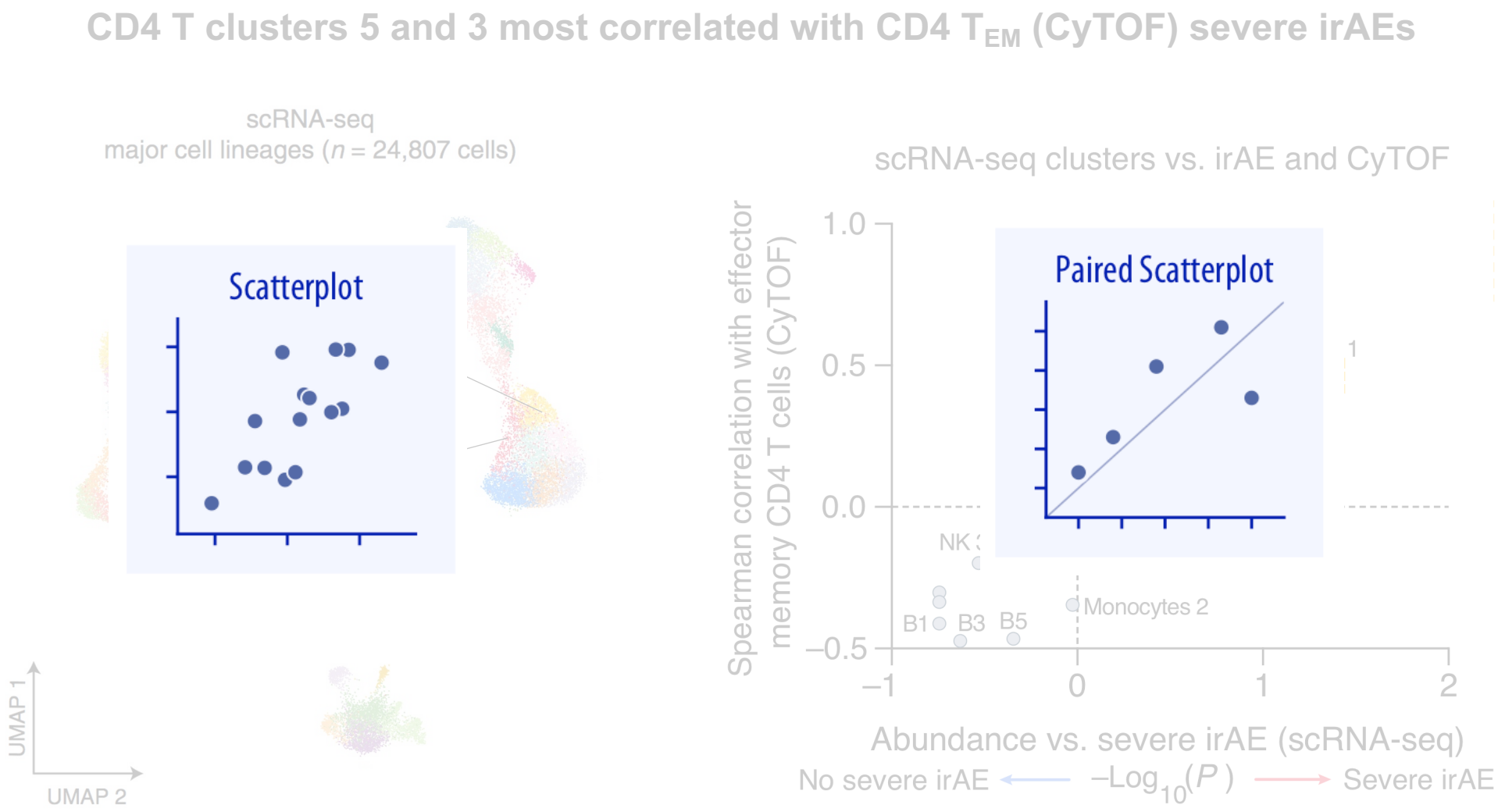
Paired analysis of 13 patients by scRNA-seq

CD4 T clusters 5 and 3 most correlated with CD4 T_{EM} (CyTOF) severe irAEs



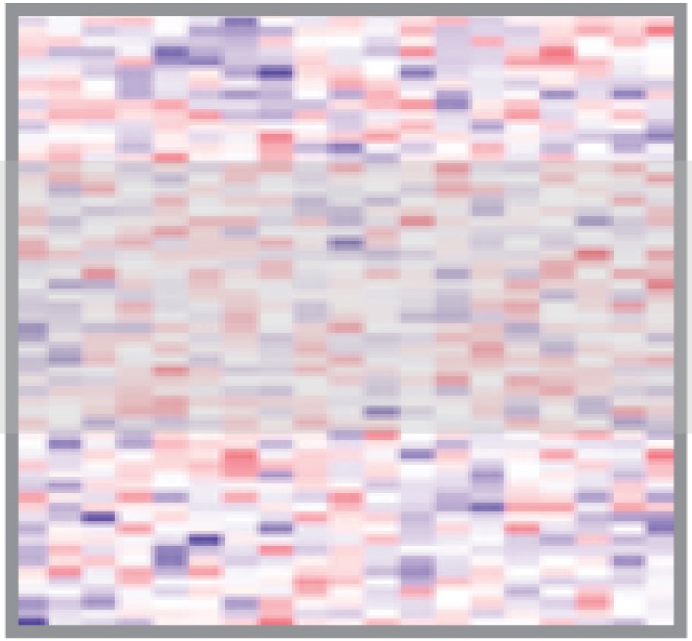
Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

Paired analysis of 13 patients by scRNA-seq

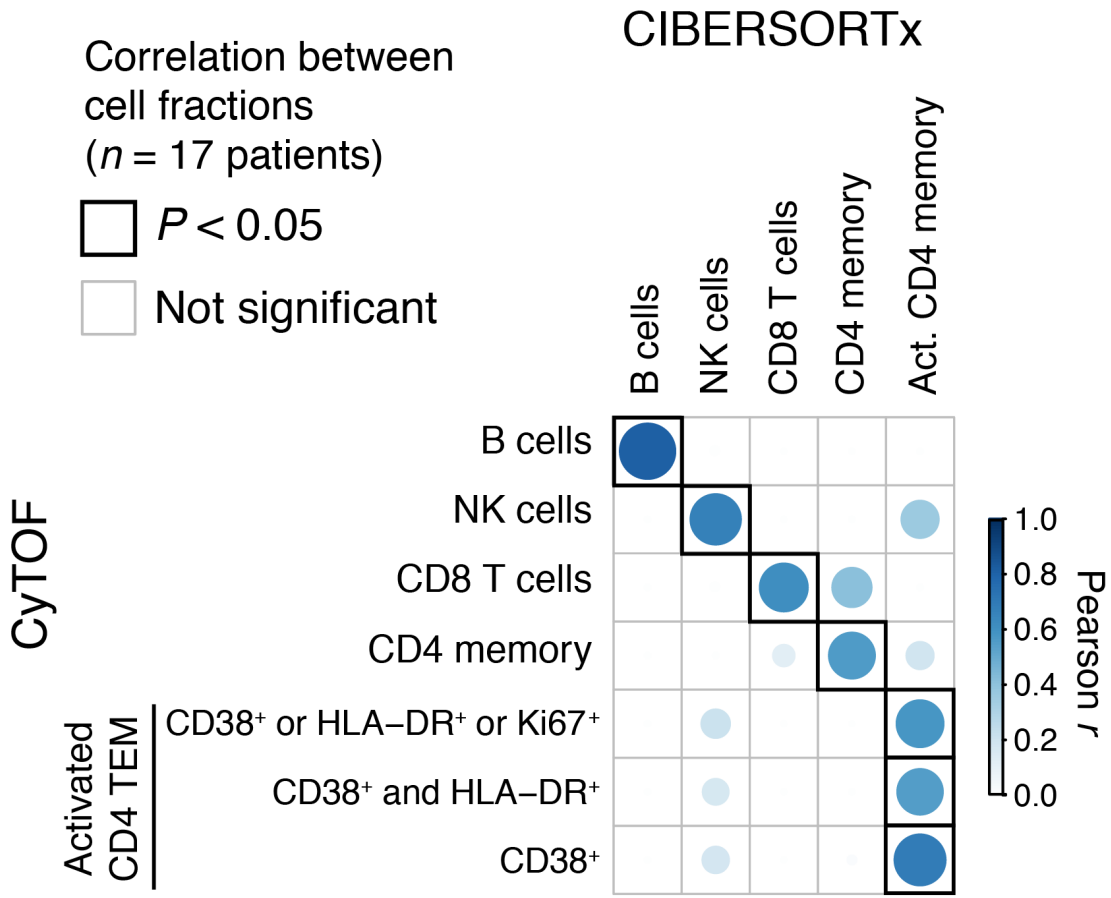


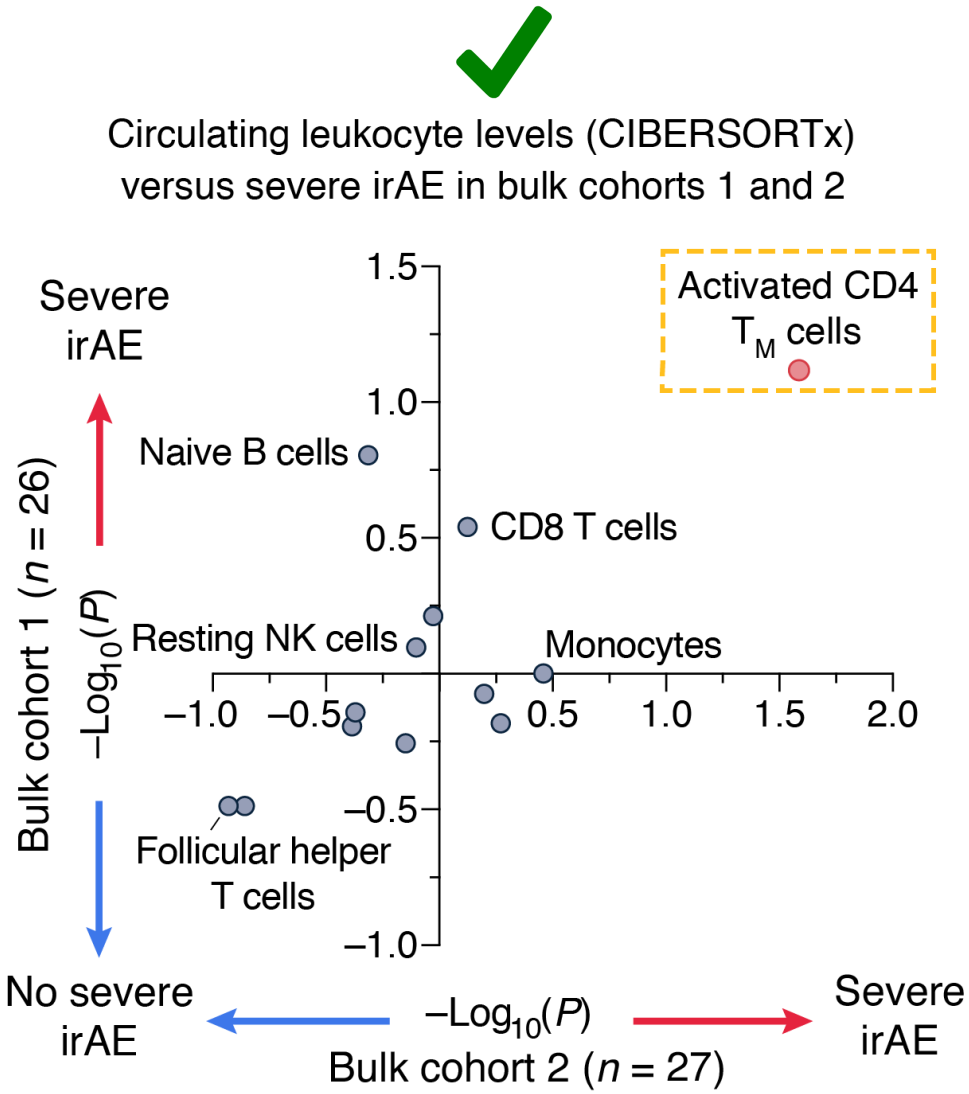
Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

Bulk RNA-seq



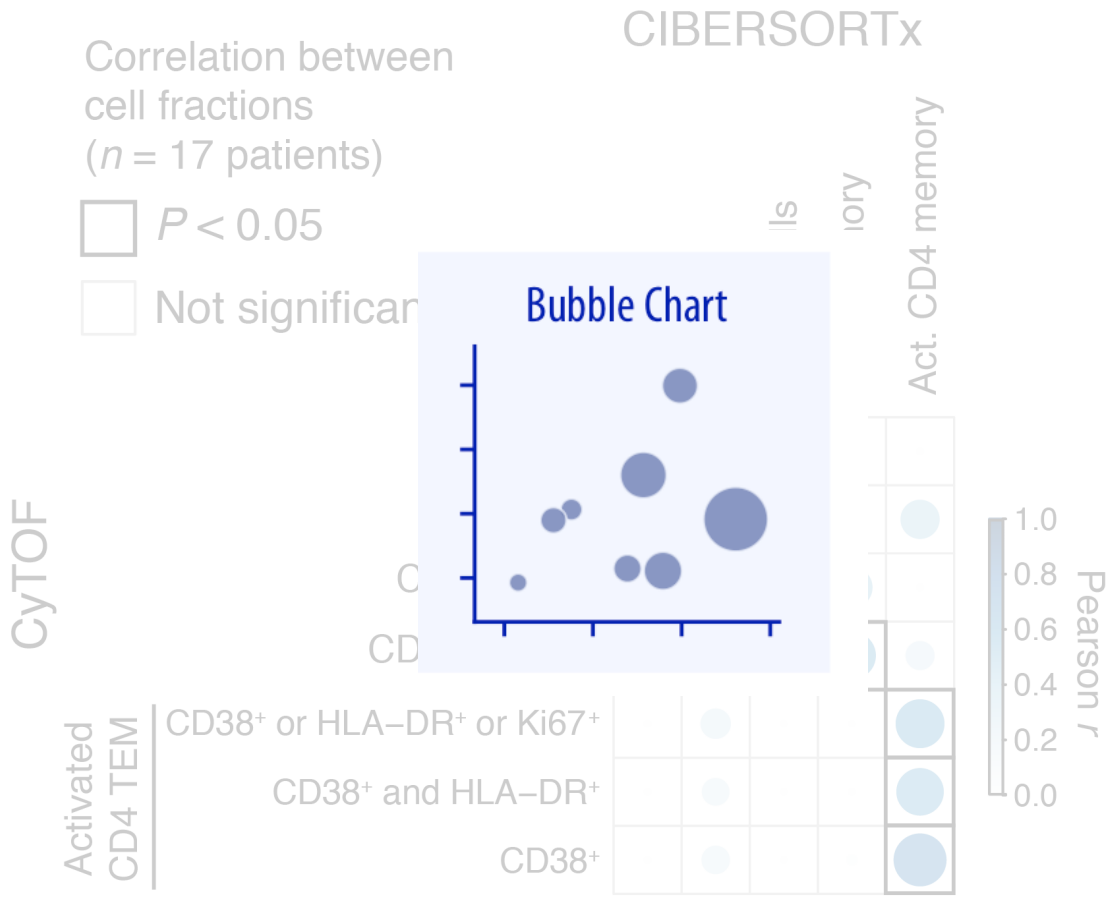
Does bulk RNA-seq agree with single-cell data?

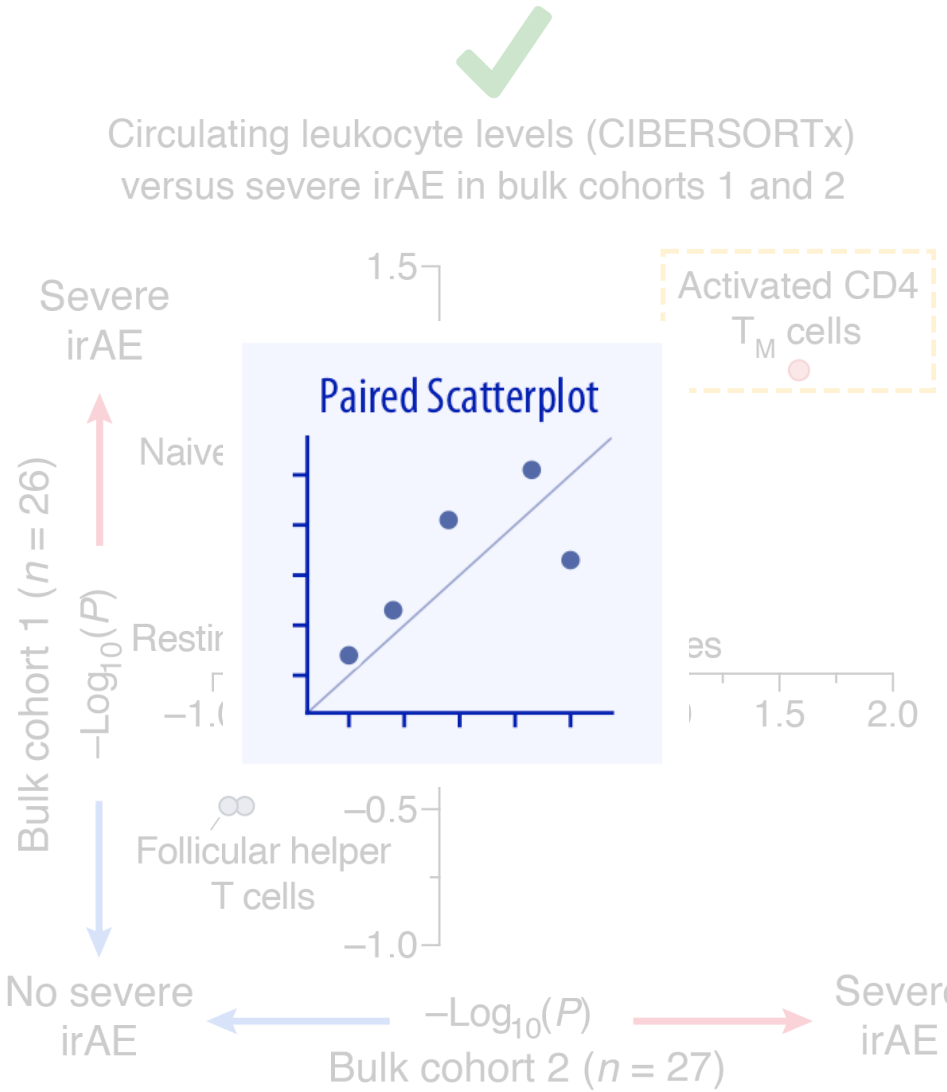




Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

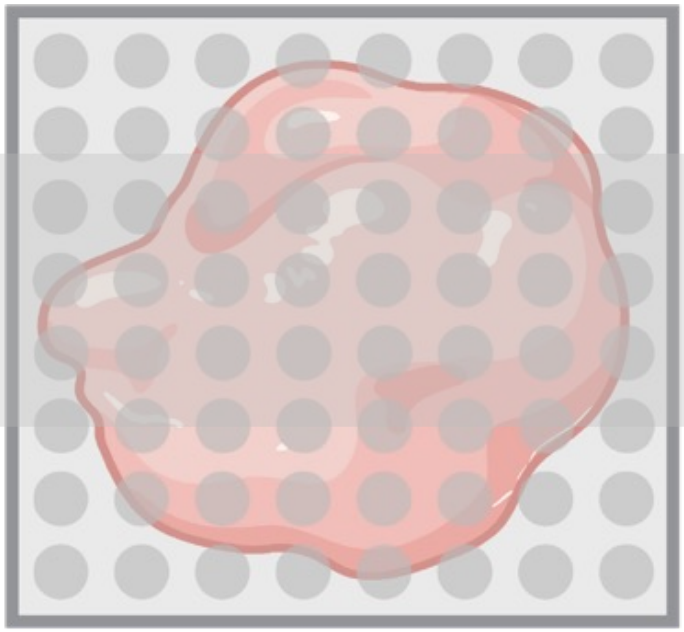
Does bulk RNA-seq agree with single-cell data?





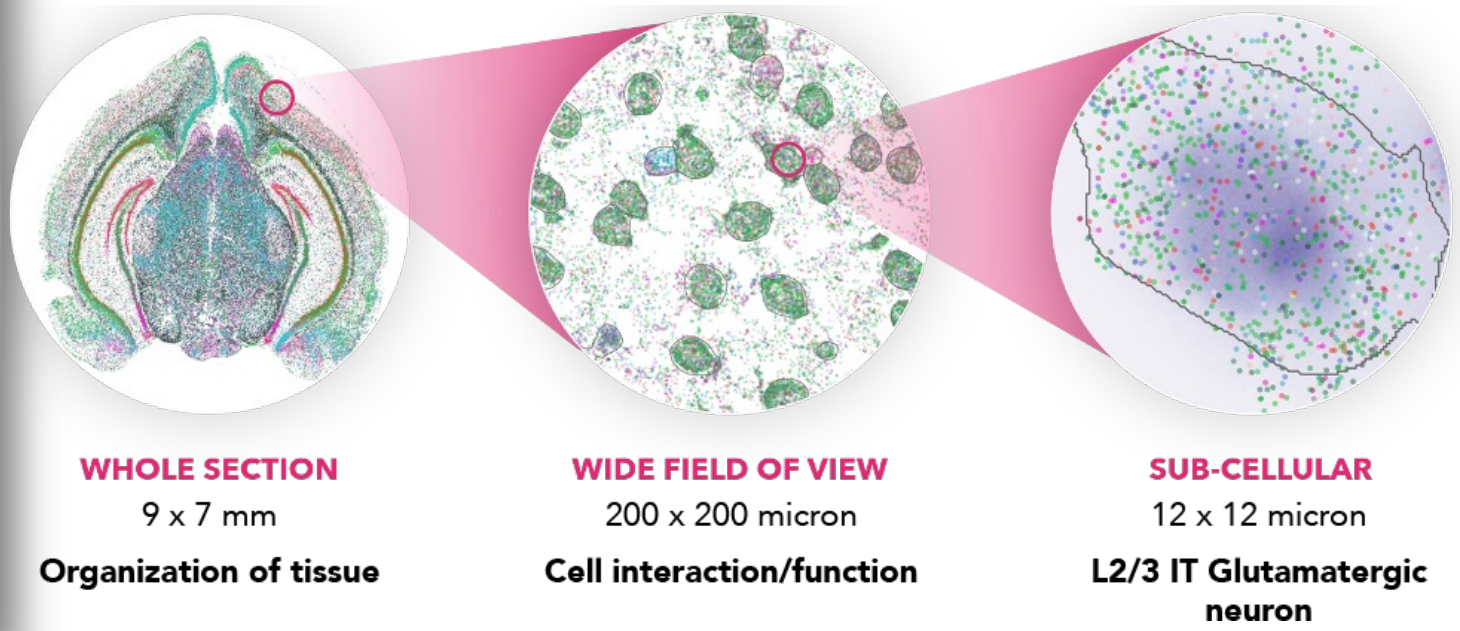
Lozano*/Chaudhuri*/Nene* et al., *Nature Medicine* (2022)

Spatial transcriptomics



Rapid advances in spatial assay development

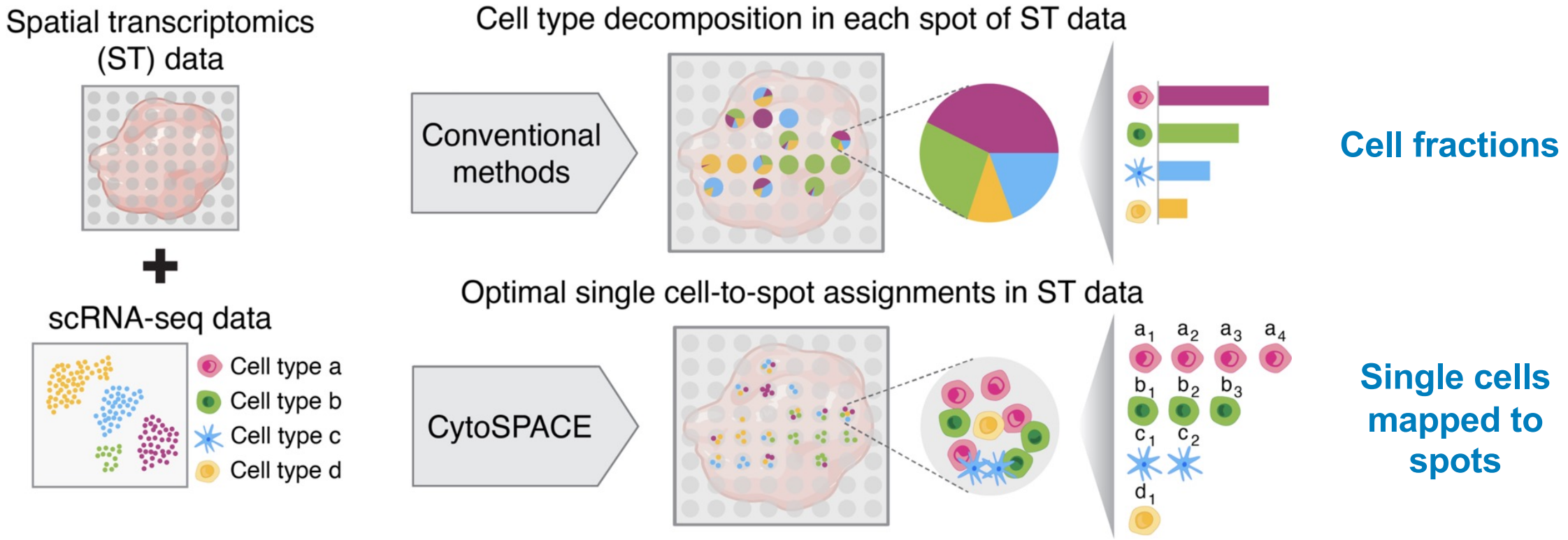
MERSCOPE (Vizgen) for single-cell spatial profiling of 500 genes



https://vizgen.com/products/

Single-cell profiling with spatial transcriptomics

- Current spatial transcriptomics (ST) platforms are **low spatial resolution** or have **low gene recovery**
- Most deconvolution methods impute cell type fractions

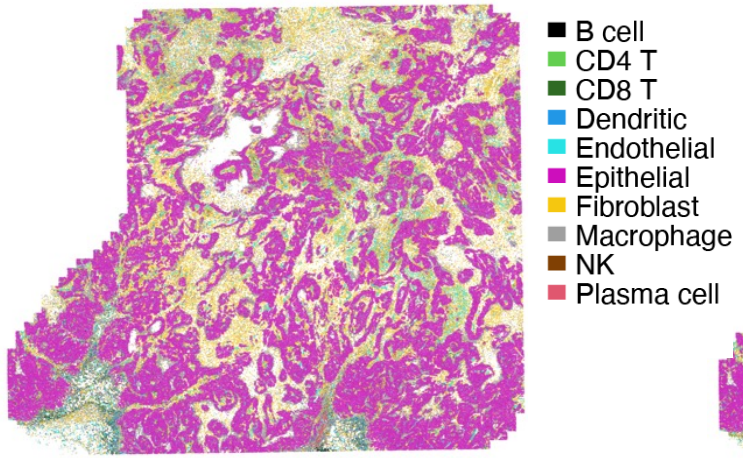


Vahid*, Brown*, Steen* et al., *Nature Biotechnology* (2023)

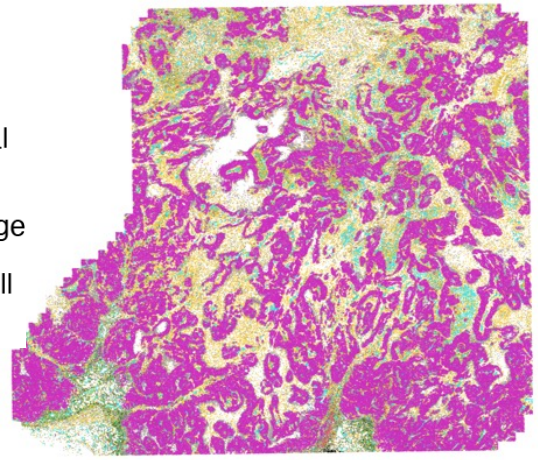
https://github.com/digitalcytometry/cytospace

Enhanced gene recovery in single-cell spatial transcriptomic data

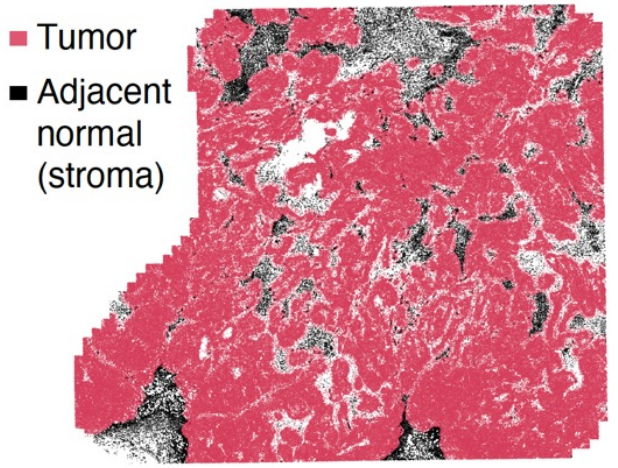
Breast tumor specimen (MERSCOPE, *n* = 500 genes)



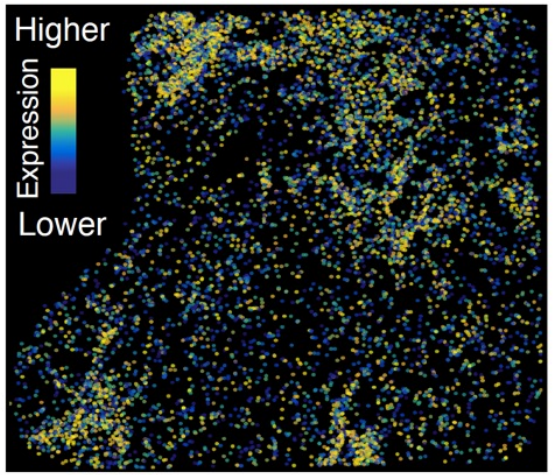
scRNA-seq atlas (Wu et al.) mapped to MERSCOPE with CytoSPACE



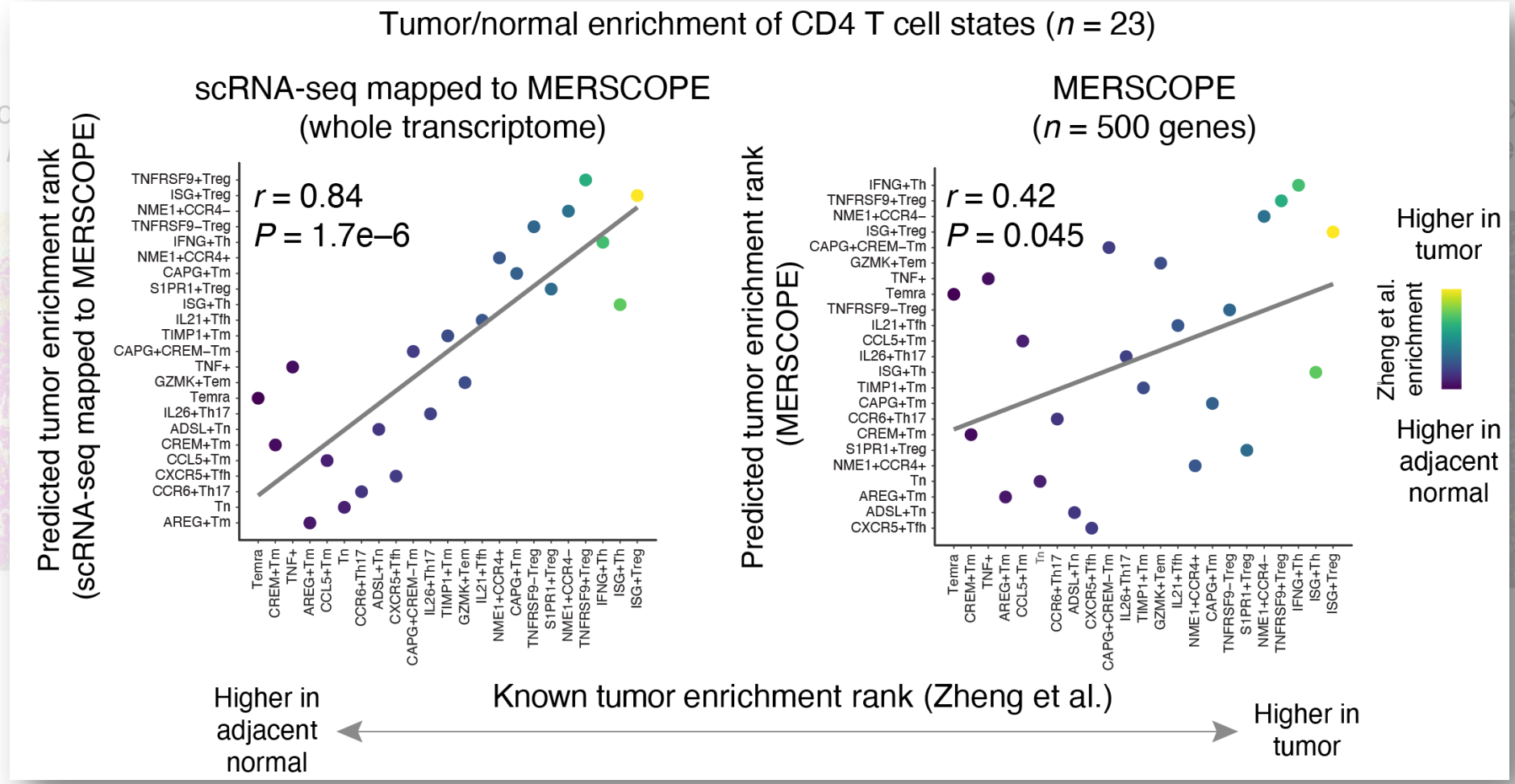
Tumor/normal regions



FOLR2 expression in macrophages (scRNA-seq)



Enhanced gene recovery in single-cell spatial transcriptomic data



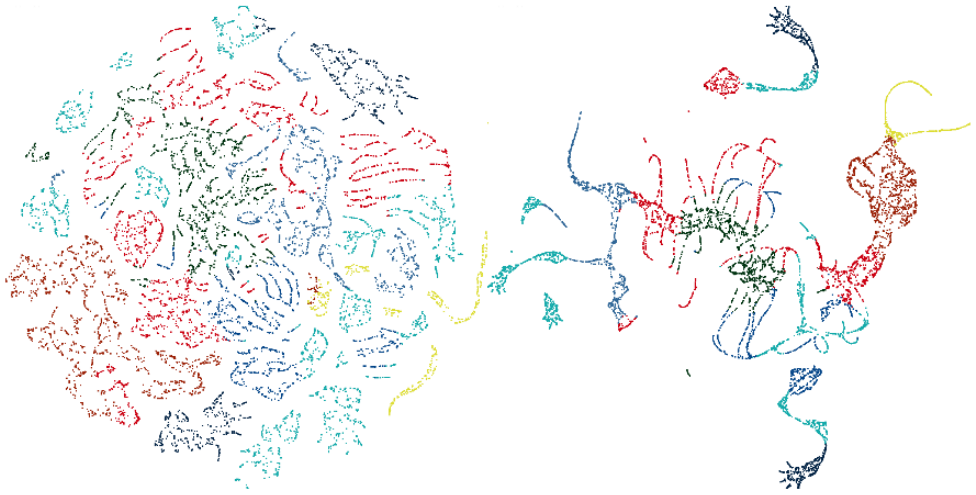
Caveats of data visualization

- Be mindful of the limitations of different visualization techniques, especially those that perform **dimensionality reduction** (e.g., PCA, t-SNE, UMAP)
- What you see is **not always complete or accurate**

Common visualization workflow for scRNA-seq

1. Perform QC
2. Filter for most variable genes
3. Do PCA to extract most informative signal (top 10-40 PCs)
4. Do UMAP (or t-SNE) in 2D – *no technique is always better

t-SNE **UMAP**



3D mammoth skeleton projected into 2D

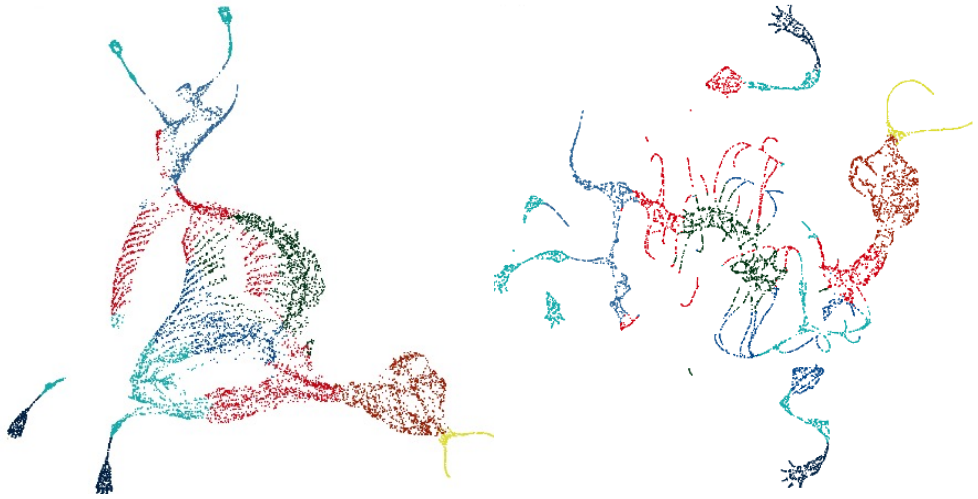
t-SNE	Perplexity	50	13min
UMAP	Nneigh	50	2min

https://pair-code.github.io/understanding-umap/

Common visualization workflow for scRNA-seq

1. Perform QC
2. Filter for most variable genes
3. Do PCA to extract most informative signal (top 10-40 PCs)
4. Do UMAP (or t-SNE) in 2D – *no technique is always better

t-SNE **UMAP**

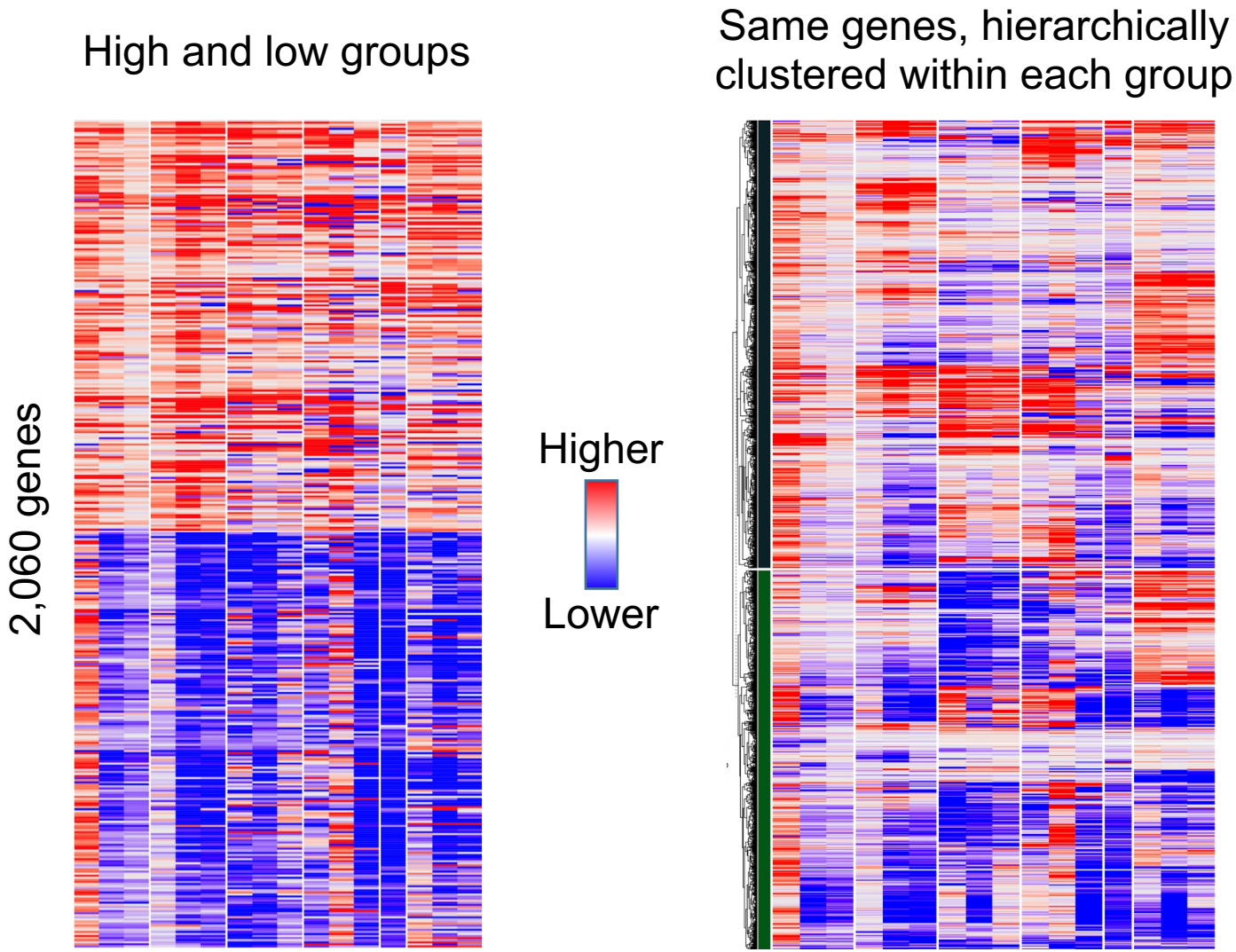


3D mammoth skeleton projected into 2D

t-SNE	Perplexity	2000	2hr
UMAP	Nneigh	50	2min

https://pair-code.github.io/understanding-umap/

Without statistics, not all patterns are meaningful



General tips and best practices

- Make text BIGGER!

As a rule of thumb, stand back from your monitor **at least 3 feet**

If the text isn't legible, **enlarge it**

- Make figures self-contained (minimize reliance on captions)
- Use **consistent font size** for all text except panel letters
- Use color and/or shapes to distinguish categories or to brighten up the figure

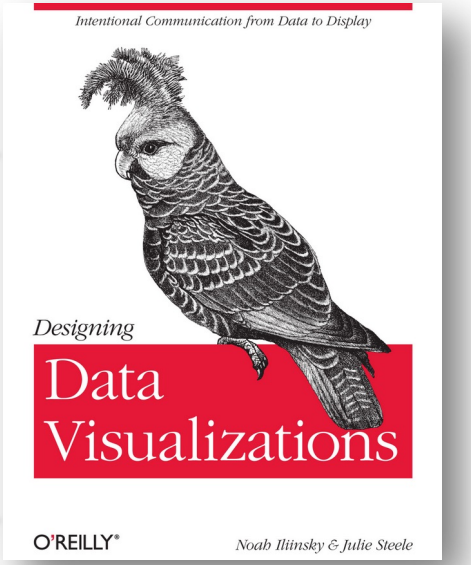
General tips and best practices

Know your audience

Optimize your data presentation for simplicity, impact, and cosmetic appeal

*"Sometimes a designer will make the visualization **more complicated than it needs to be**, not because he is trying to make the data look bad, but for precisely the opposite reason: he wants the data to look as good as possible. This is an equally bad mistake."*

*"Your data is important and meaningful all on its own; **you don't have to make it special by trying to get fancy**. Every dot, line and word should serve a communicative purpose: if it is extraneous or outside the scope of the visualization's goals, it must go. **Edit ruthlessly**. Don't decorate your data."*



Resources

- **Figure generation**
  - Ten Simple Rules for Better Figures (Rougier et al., *PLOS Comp Biol* 2014): https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003833
- **Fundamentals of data viz**
  - https://clauswilke.com/dataviz/
  - https://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html
- **Newman Lab software tools**
  - https://anlab.stanford.edu/software
- **Immunotherapy expression datasets**
  - http://tide.dfci.harvard.edu/download/