BIO334 Practical Bioinformatics

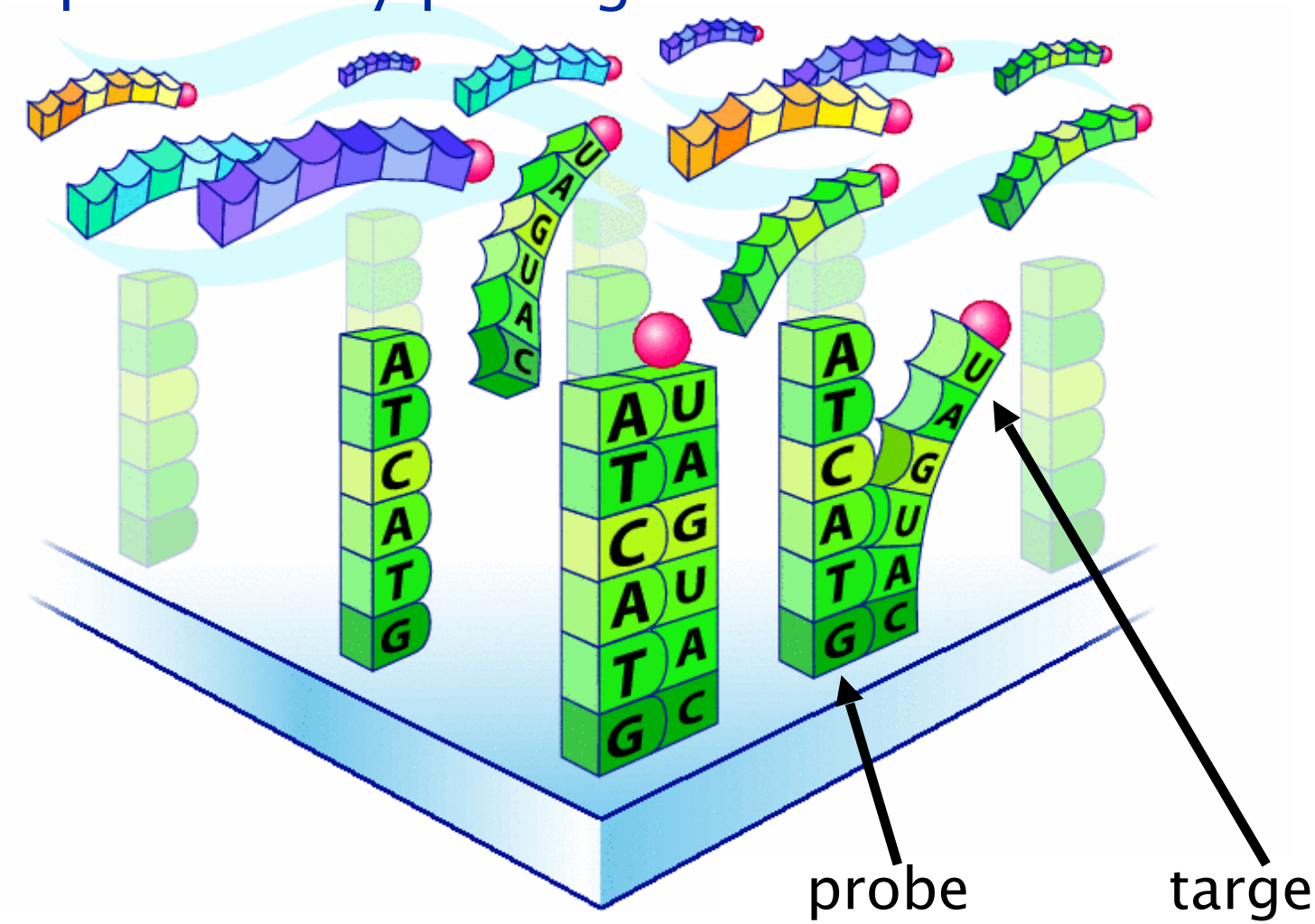
Technologies, Transcriptomic/Proteomic Data Analyses (non-exhaustive)

Technologies
in our research area

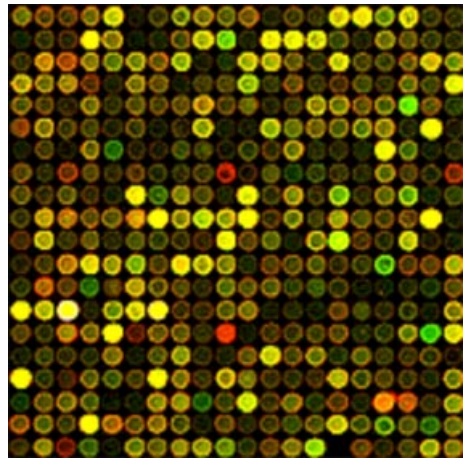
high-throughput sequencing, single cell, cytometry, etc.

“it’s just data”

DNA microarray: **parallel** northern blots; Nature gives a complementary pairing

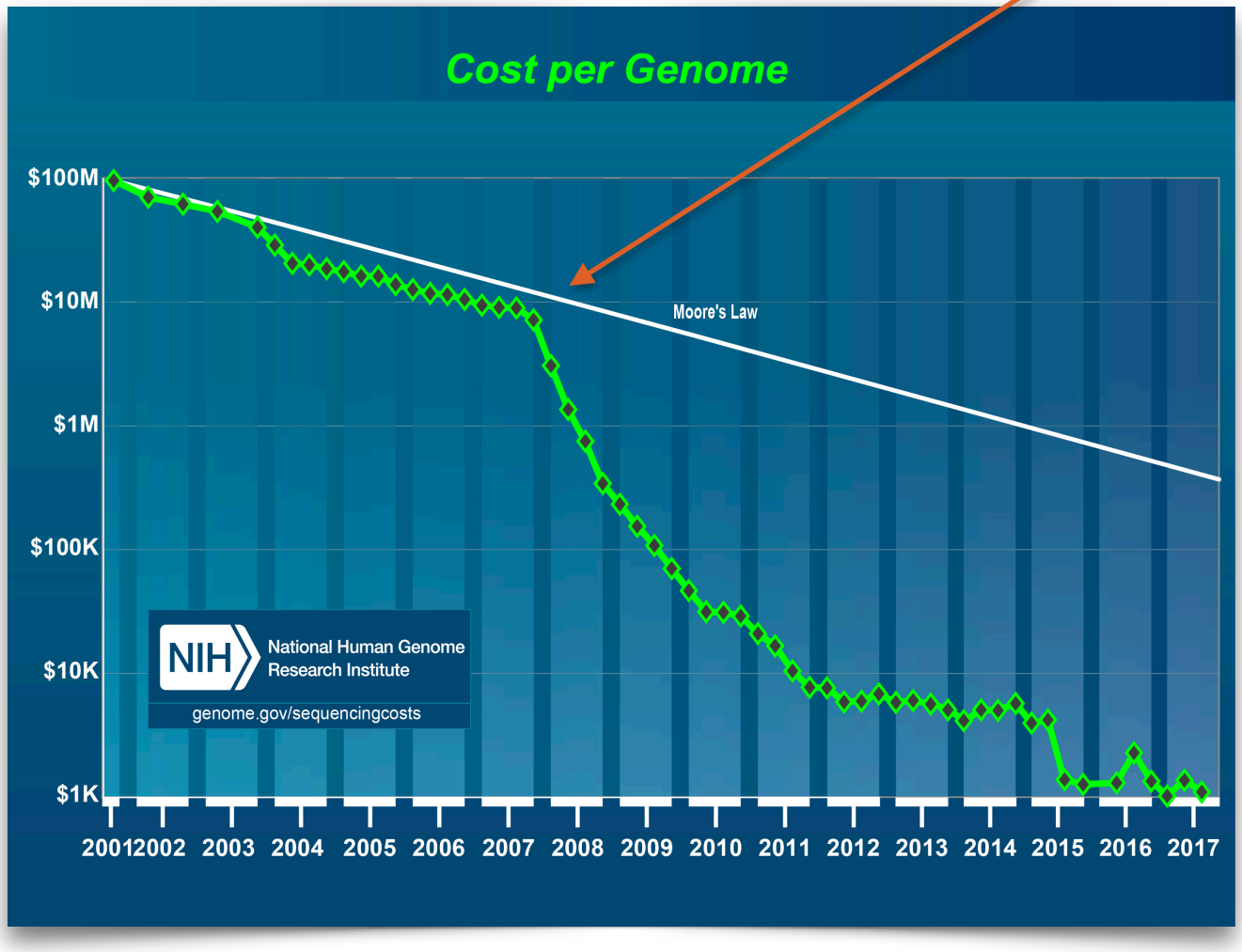


Abundance (of complementary DNA species) measured by flouresence intensity



High-throughput sequencing

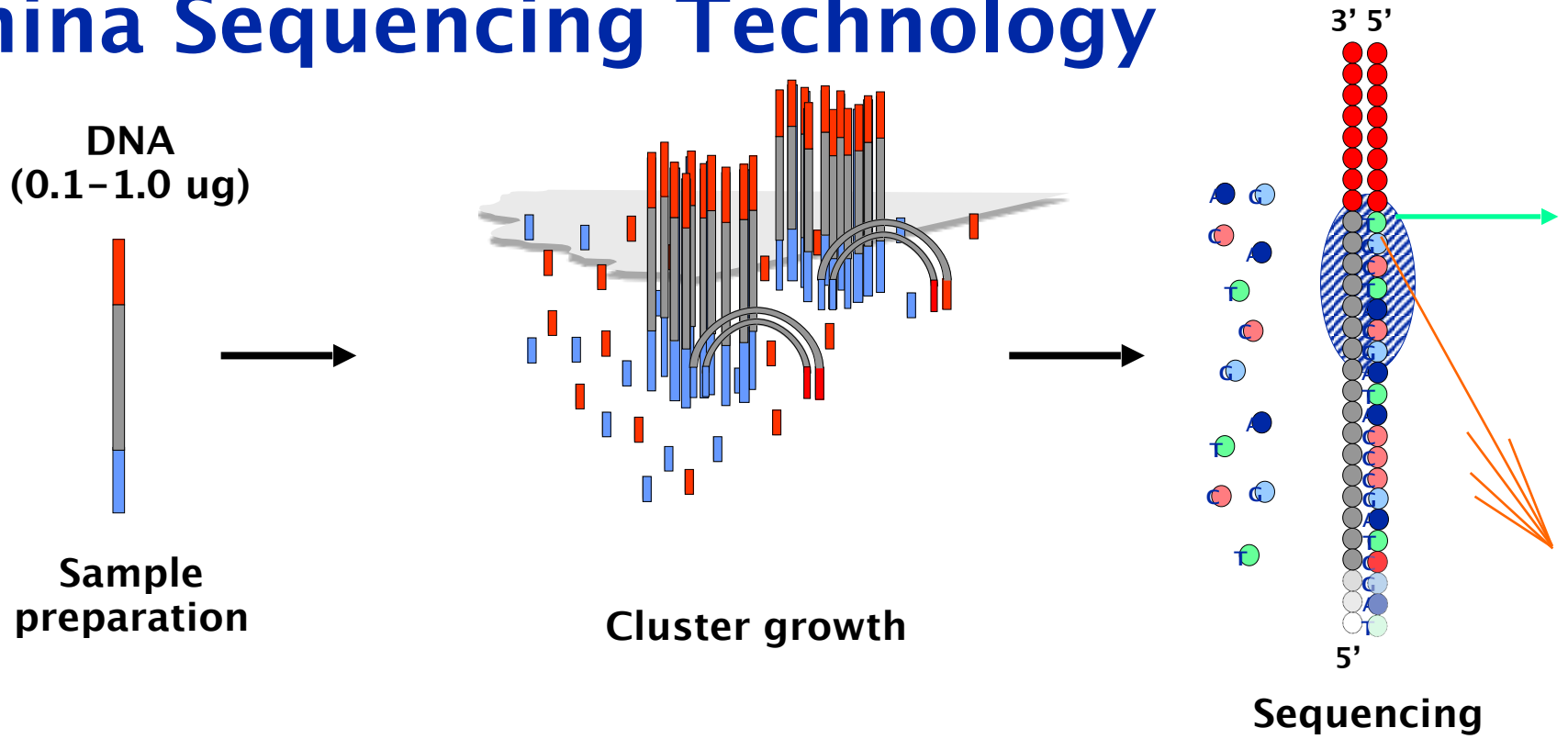
(Solexa) Illumina



https://www.statnews.com/2017/01/09/illumina-ushering-in-the-100-genome/

Many assays based on sequencing DNA:
https://liorpachter.wordpress.com/seq/

Illumina Sequencing Technology



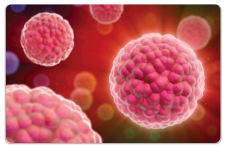
This slide courtesy of Gary Schroth, Illumina

Applications of high-throughput sequencing

Common Sequencing Applications

Cancer Research

NGS-based sequencing enables cancer researchers to detect rare somatic variants, tumor subclones, and circulating DNA fragments. Learn more about sequencing for cancer research.

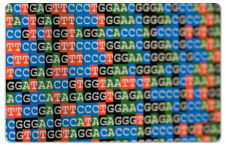


Microbiology Research

From environmental metagenomics studies to infectious disease surveillance and more, NGS-based sequencing can help researchers gain genetic insight into bacteria and viruses. Learn more about microbial genomics.

Complex Disease Research

Illumina sequencing is introducing new avenues for understanding immunological, neurological, and other complex disorders on a molecular level. Learn more about complex disease genomics.



Reproductive and Genetic Health

Illumina sequencing and array technologies deliver fast, accurate information that can guide choices along the reproductive and genetic health journey. Find reproductive and genetic health solutions.

ETH zürich | University of Zurich[UZH]

Student portal | Alumni association | B-Fabric | Login | Contact | en

Keyword or person | Departments

Functional Genomics Center Zurich

About us | Working with us | OMICS areas | Education | Research & Publications | FAQ | News & Events

ETH Zurich > UZH > FGCZ

- **Services**
  - Proteomics/Protein analysis services
  - **Genomics/Transcriptomics services**
  - Metabolomics/Biophysics services
- User Lab Access
- Collaboration
- FGCZ Policies
- Job Offers
- FGCZ Terms and Conditions

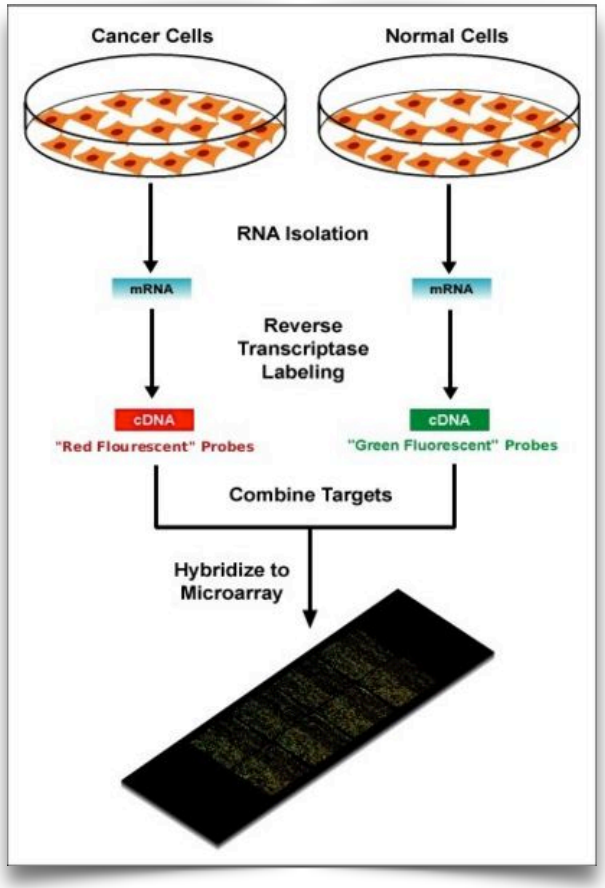
Genomics/Transcriptomics services

All services in Genomics/Transcripomics require a project submission via B-Fabric, our project management system.

If you have specific questions about our Genomics/Transcripomics services please refer to our FAQ → section; alternatively, or in case you would like to request a quote,please do not hesitate to get in touch with our sequencig team at sequencing@fgcz.ethz.ch

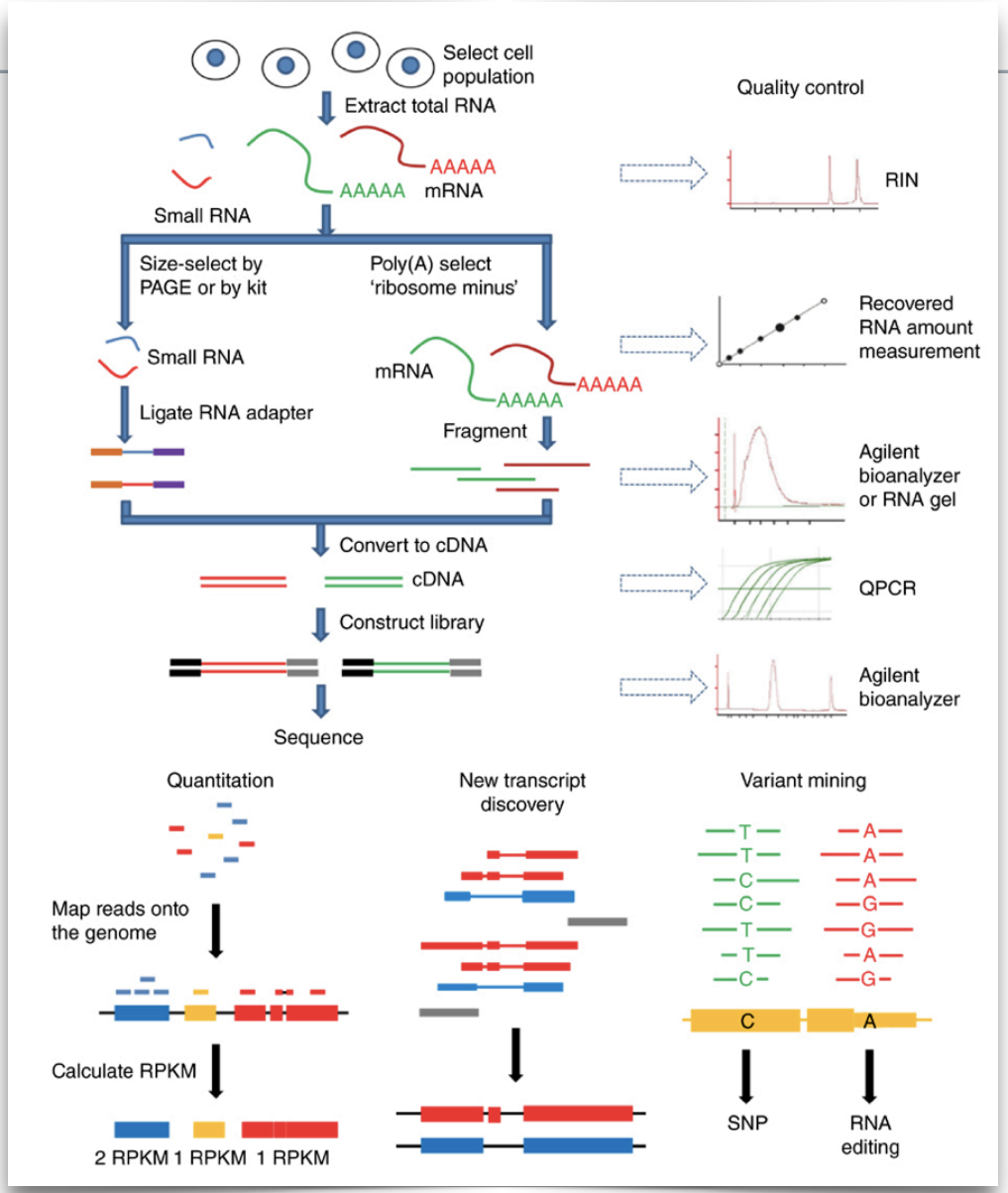
Application Group	Application	Order via B-Fabric
DNA sequencing	Whole Exome Sequencing	Project
DNA sequencing	Methylation Profiling	Project
DNA sequencing	ChIP-Seq	Project
DNA sequencing	Targeted Sequencing and Metagenomics	Project
DNA sequencing	*De novo* Genome Assembly	Project
DNA sequencing	Whole Genome Resequencing	Project
RNA sequencing	Transcriptome Profiling	Project
RNA sequencing	Small RNA Profiling	Project
RNA sequencing	De novo Transcriptome Assembly	Project

Abundance by Fluorescence Intensity (DNA microarray)



http://en.wikipedia.org/wiki/DNA_microarray

Abundance by Counting (RNA-seq)



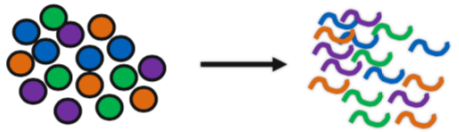
Zeng & Mortazavi, Nature Immunology, 2012

Gene Expression Profiling: questions of interest

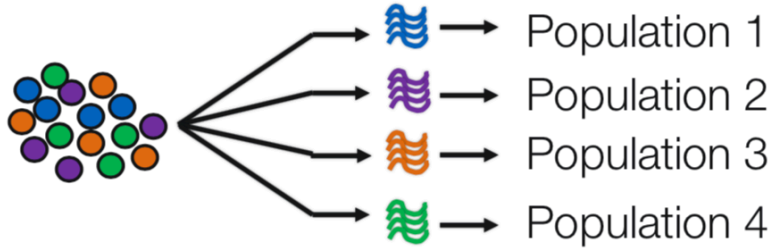
- What genes have changed in expression? (e.g. between disease/normal, affected by treatment)
  Gene discovery, differential expression
- Is a specified group of genes all up-regulated in a particular condition?
  Gene **set** differential expression
- Can the expression profile predict outcome?
  Class prediction, classification
- Are there tumour sub-types not previously identified? Do my genes group into previously undiscovered pathways?
  Class discovery, clustering
- (with single cell gene expression): What changes in cell type composition are observed? What genes have changed in expression in a given subtype of cells?

https://hbctraining.github.io/scRNA-seq/slides/Single_Cell_2_27_20.pdf

Bulk vs Single Cell RNA Sequencing (scRNA-seq)

Bulk RNA-seq  average expression level

- comparative transcriptomics
- disease biomarker
- homogenous systems

scRNA-seq 

- define heterogeneity
- identify rare cell population
- cell population dynamics

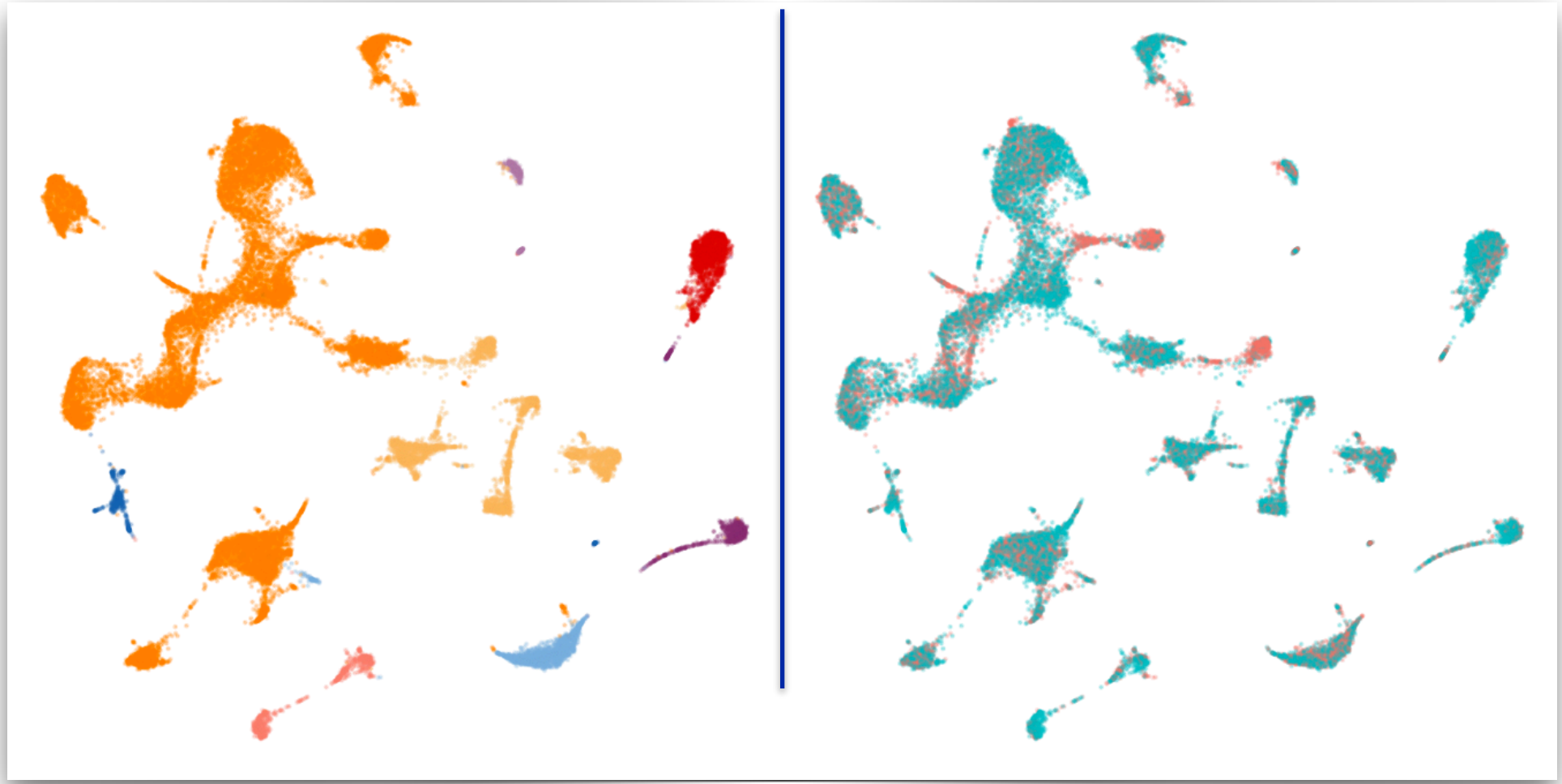
Motivation: Single-cell RNA-seq: finding cell subpopulation-specific changes in state

frontal cortex

single nuclei RNA-seq (10x)

Data from:
4 mice vehicle treated
4 mice LPS treated

Each dot is one cell



Flow cytometry

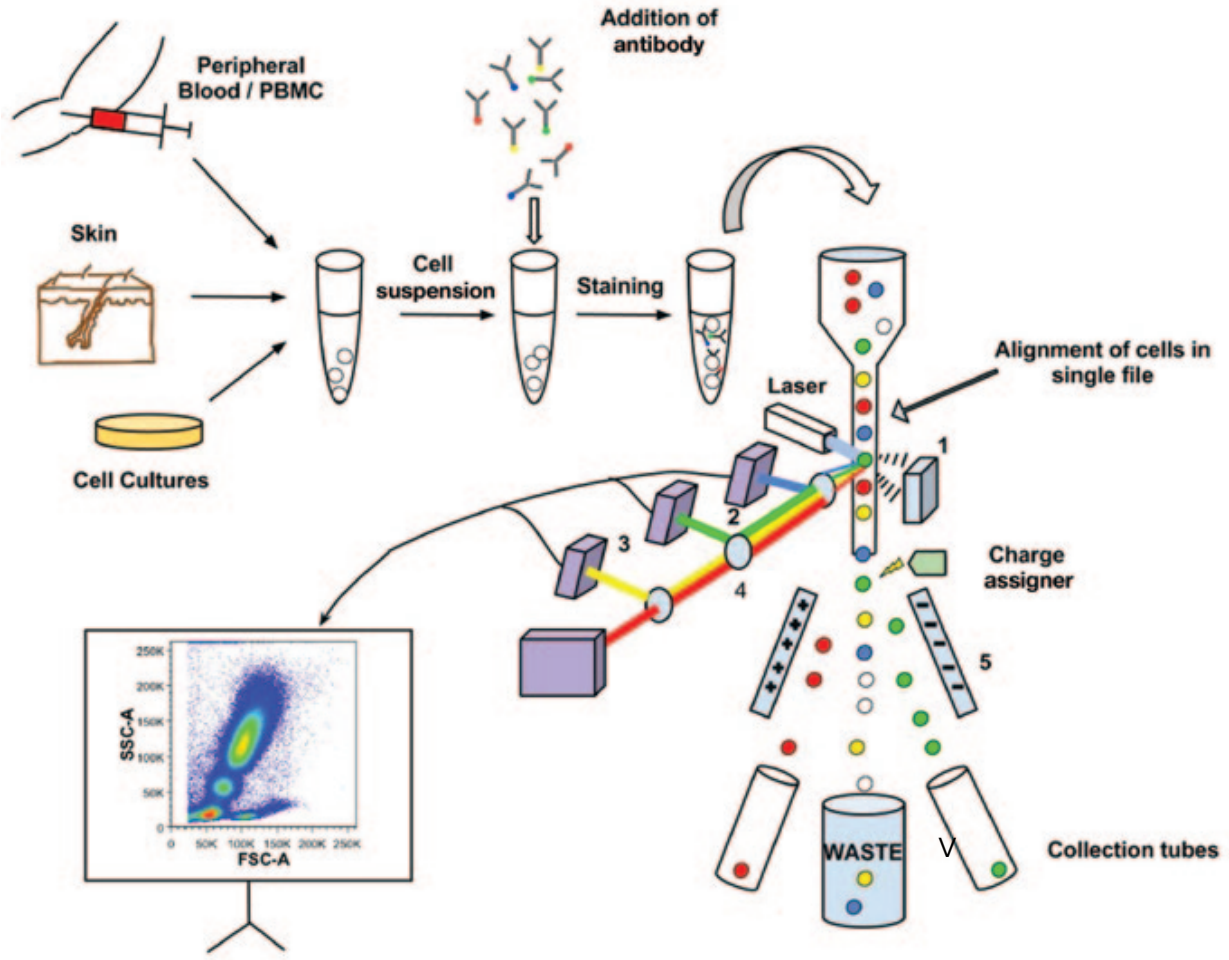
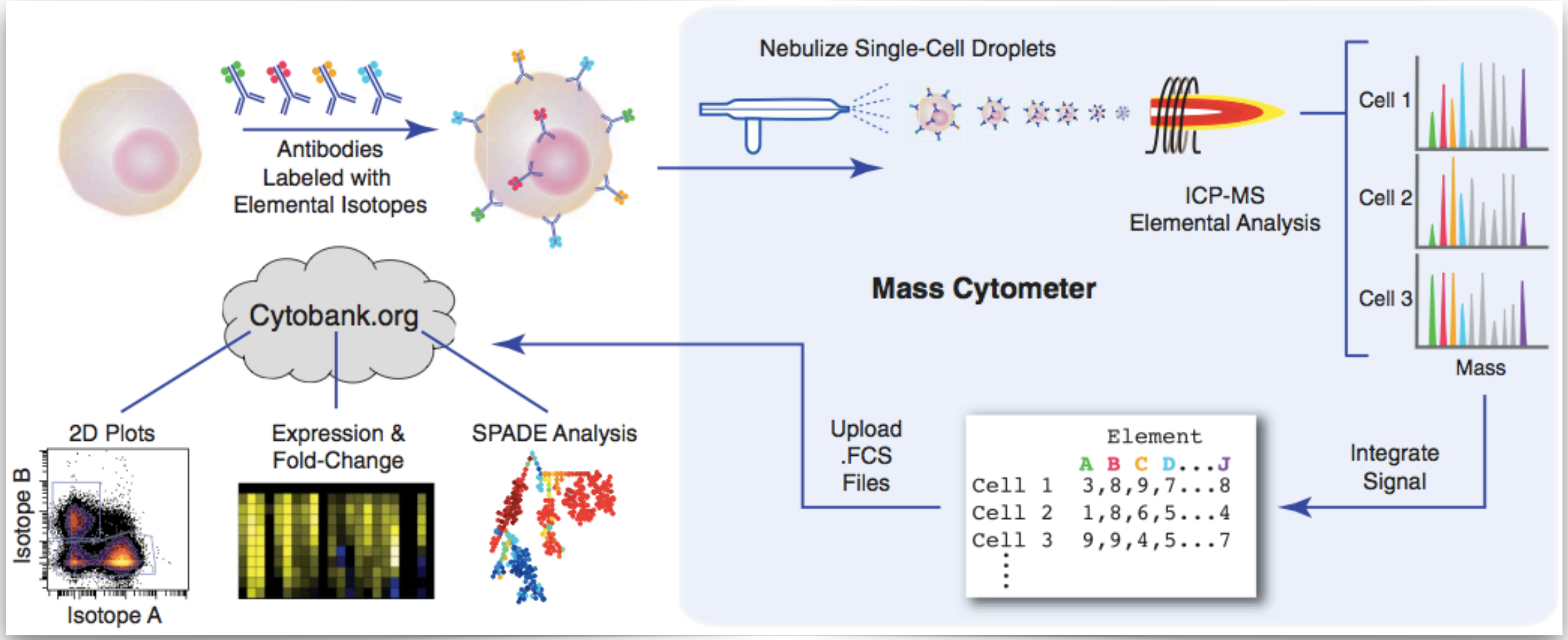


Figure 1. Schematic representation of a flow cytometer. For details please see text. (1) Forward-scatter detector, (2) side-scatter detector, (3) fluorescence detector, (4) filters and mirrors, and (5) charged deflection plates.

Jahan-Tigh et al., Journal of Investigative Dermatology, 2012

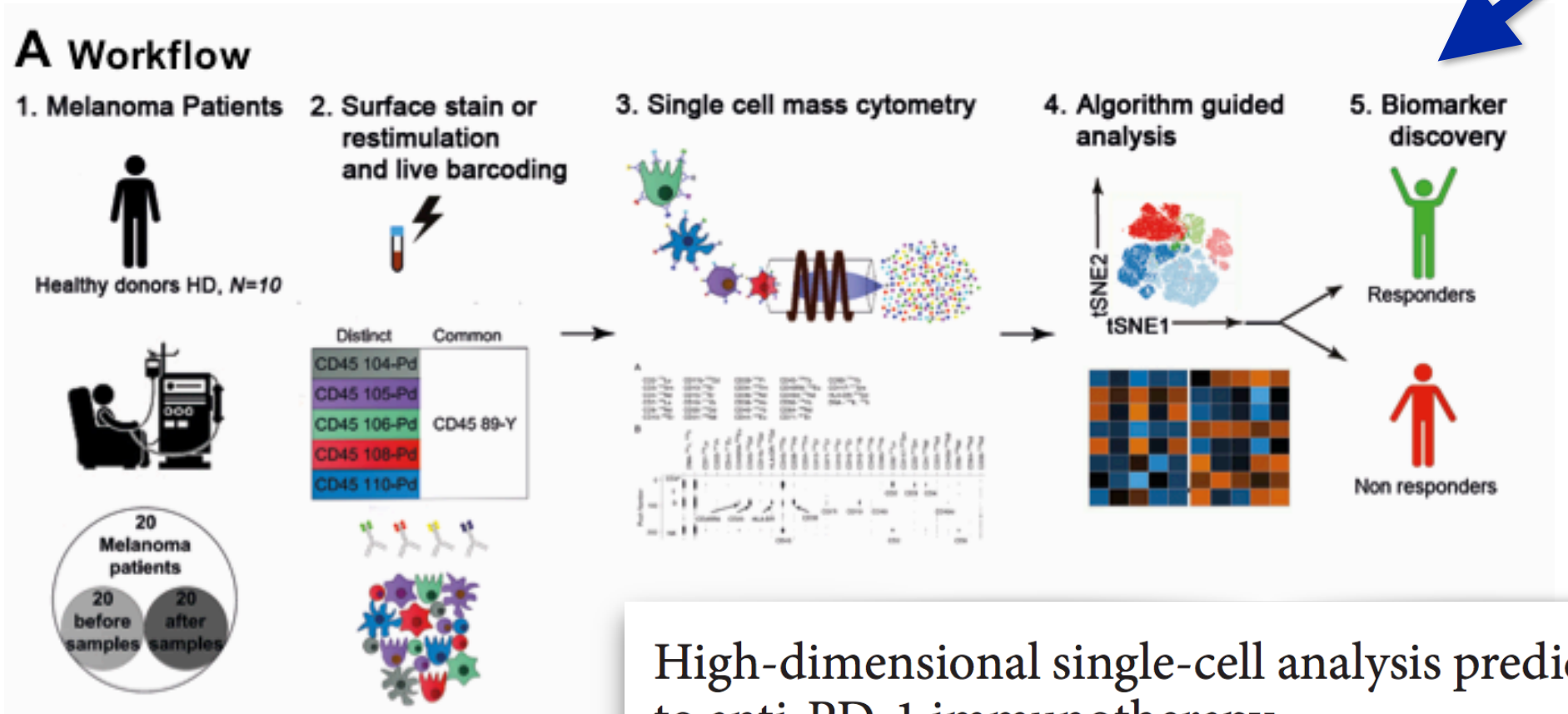
Mass cytometry



Bendall et al. (2011), Fig. 1A

Algorithm guided analysis = statistics

Finding molecular biomarkers associated with drug response

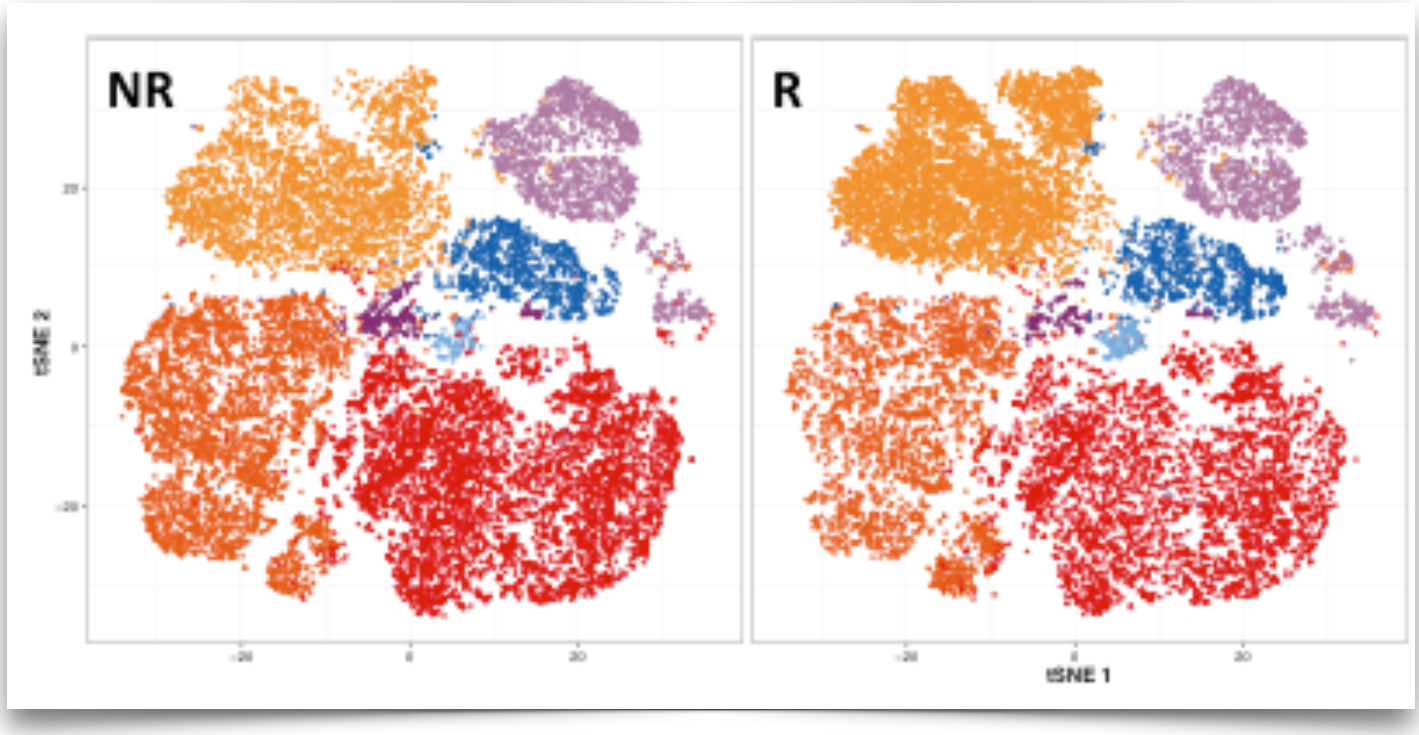


High-dimensional single-cell analysis predicts response to anti-PD-1 immunotherapy

Carsten Krieg[1,6], Malgorzata Nowicka[2,3], Silvia Guglietta[4], Sabrina Schindler[5], Felix J Hartmann[1], Lukas M Weber[2,3], Reinhard Dummer[5], Mark D Robinson[2,3], Mitchell P Levesque[5,7] & Burkhard Becher[1,7]

Differential abundance of cell populations

tSNE projection
(each dot = cell,
cells from multiple
patients)



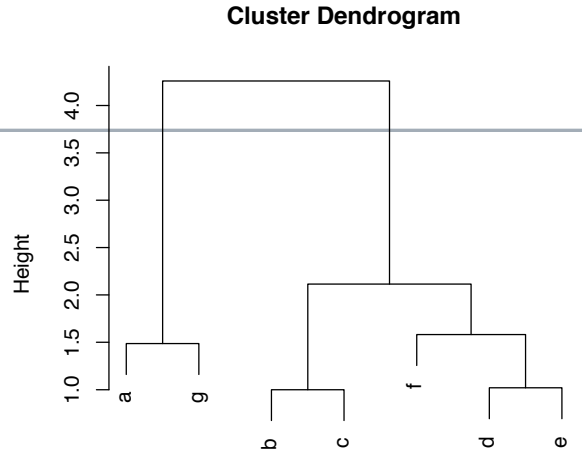
NR: non-responders
R: responders

Under the hood: Generalized linear mixed model to assess the change in relative abundance of subpopulations.

Hierarchical Clustering

http://en.wikipedia.org/wiki/Hierarchical_clustering

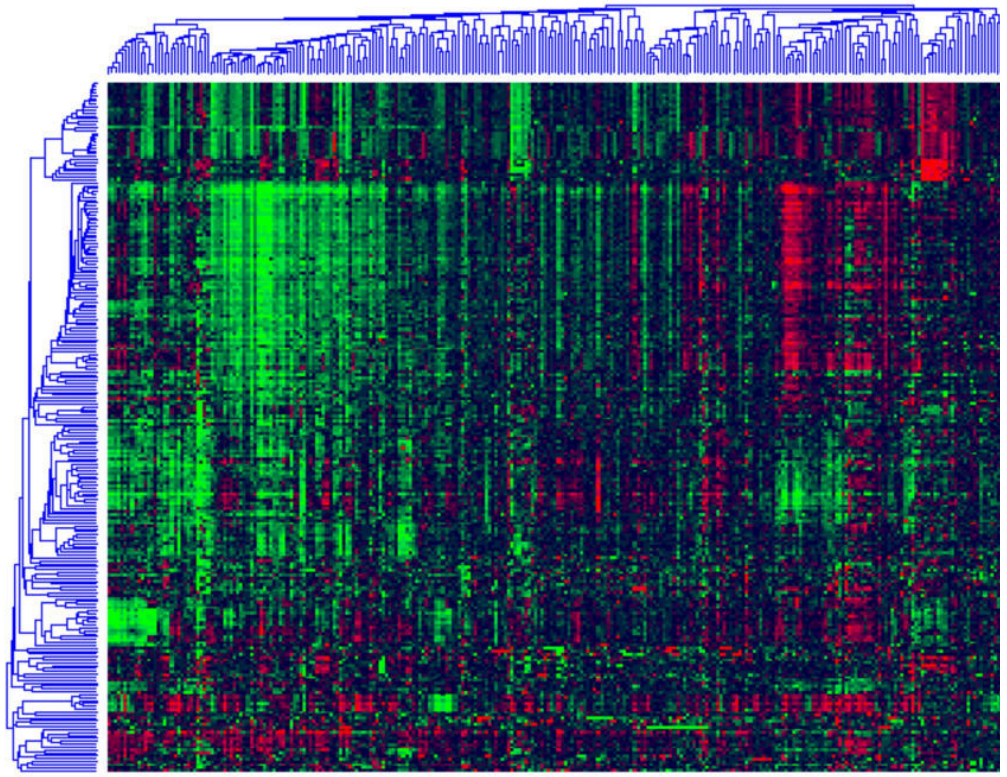
Hierarchical (Agglomerative) Clustering



Divisive (all features start as 1 cluster, then subsequently split) versus Agglomerative (every feature is it's own cluster, then subsequently merged)

Metric: to define how similar any two vectors are.

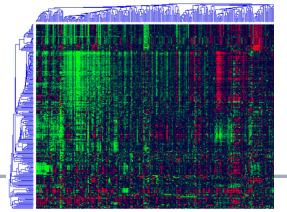
Linkage: determines how clusters are merged into a tree



http://en.wikipedia.org/wiki/Euclidean_distance

Euclidean distance:

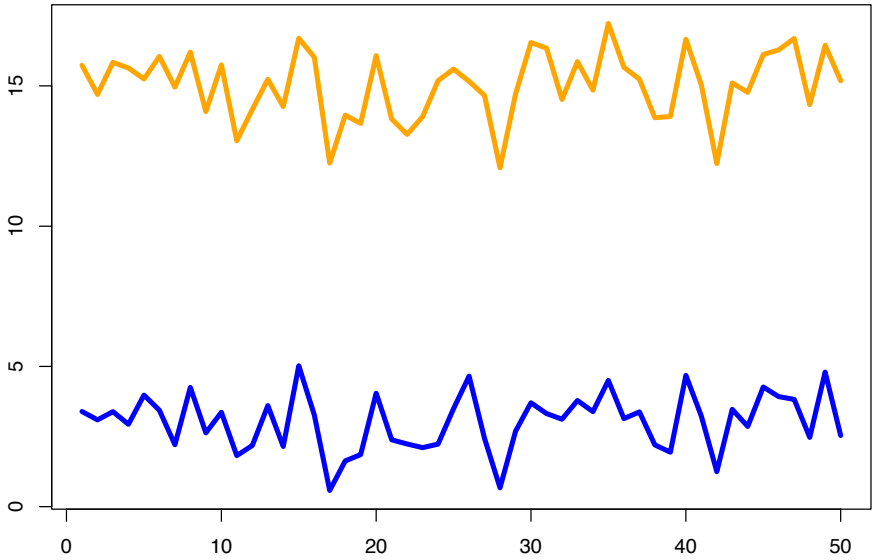
$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

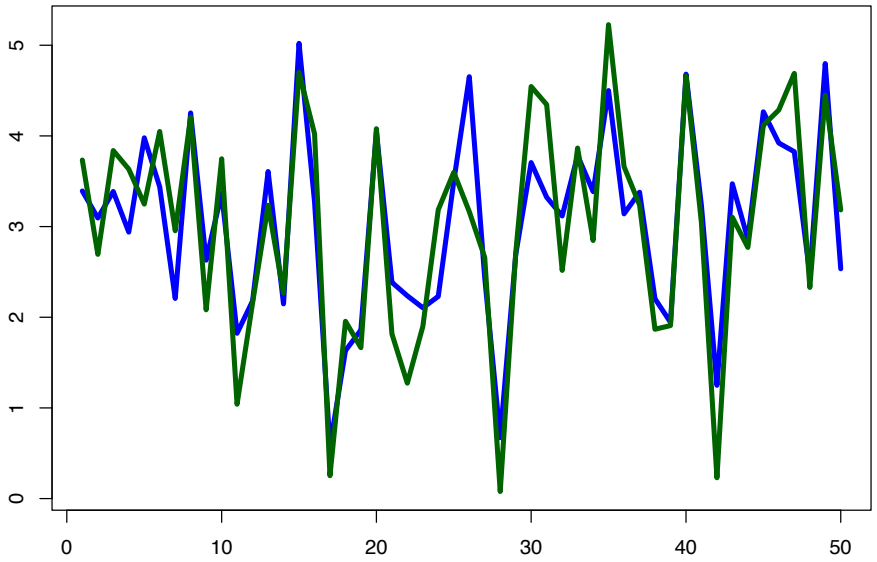


Are these “vectors” similar ?

```
> sqrt(sum((x-(y-12))^2))
[1] 3.926007
> sqrt(sum((x-y)^2))
[1] 84.84028
```

It depends how you define similar.

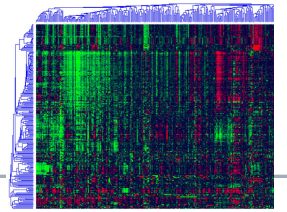




Euclidean distance: 84.84 3.92

http://en.wikipedia.org/wiki/Correlation_and_dependence

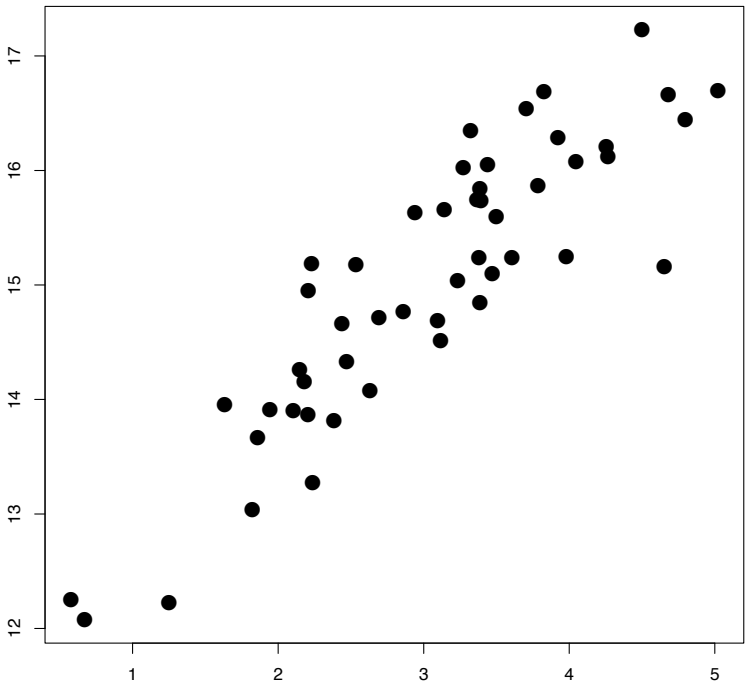
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

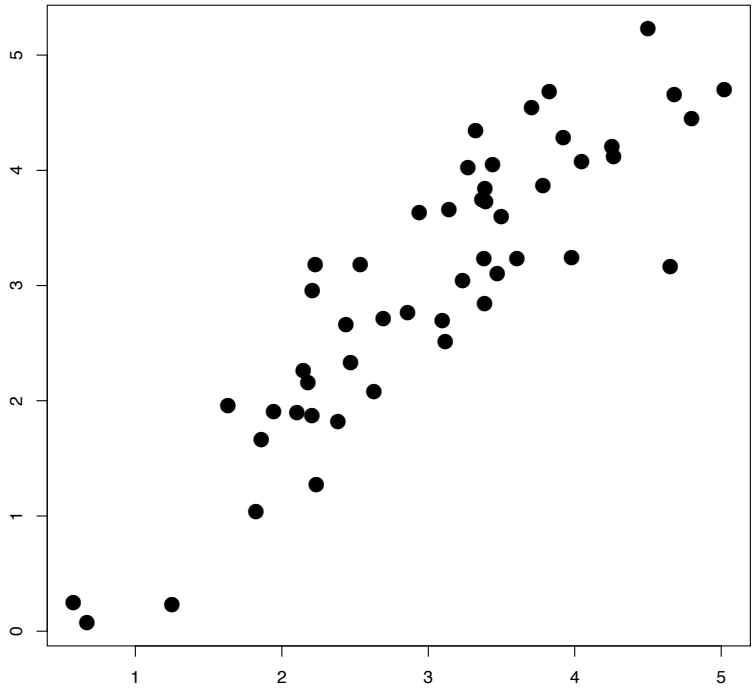


Are these “vectors” similar ?

```
> cor(x,y)
[1] 0.8901139
> cor(x,y-12)
[1] 0.8901139
```

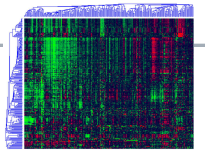
It depends how you define similar.





Correlation: 0.89 0.89

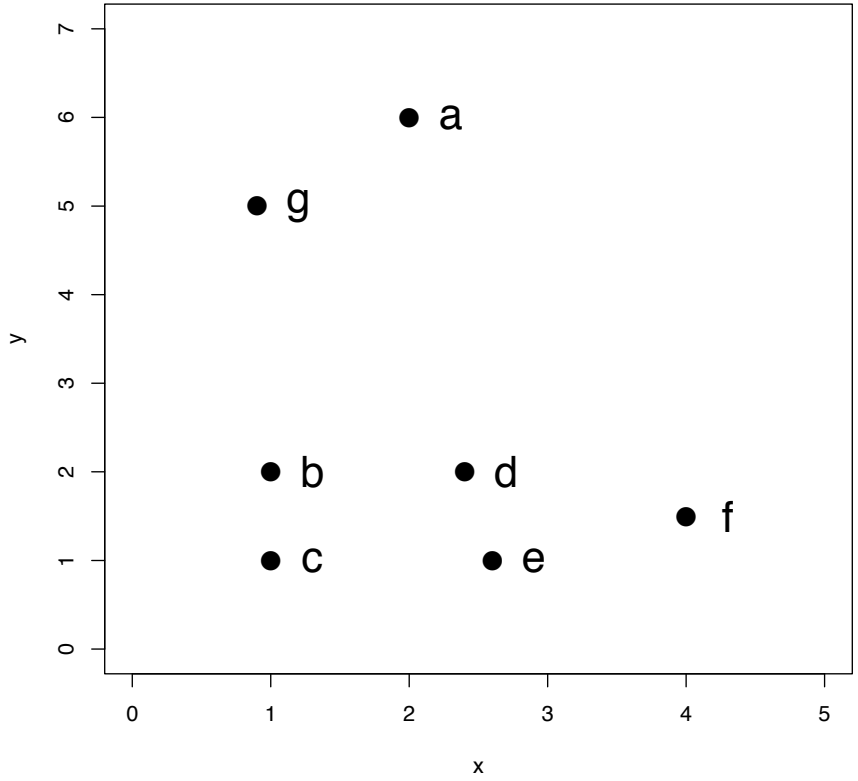
http://en.wikipedia.org/wiki/Hierarchical_clustering



Hierarchical (Agglomerative) Clustering

Start with distances.

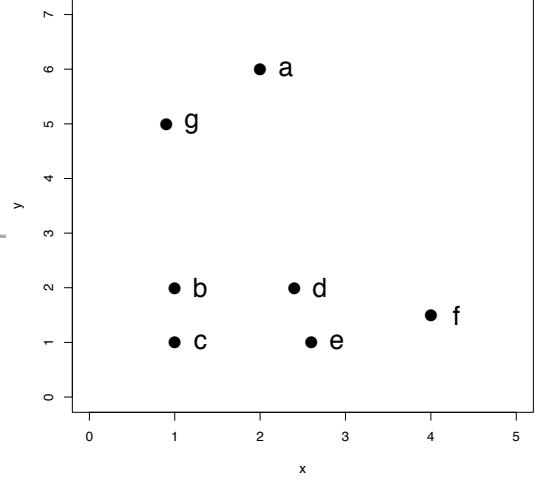
Linkage: determines how clusters are merged into a tree.

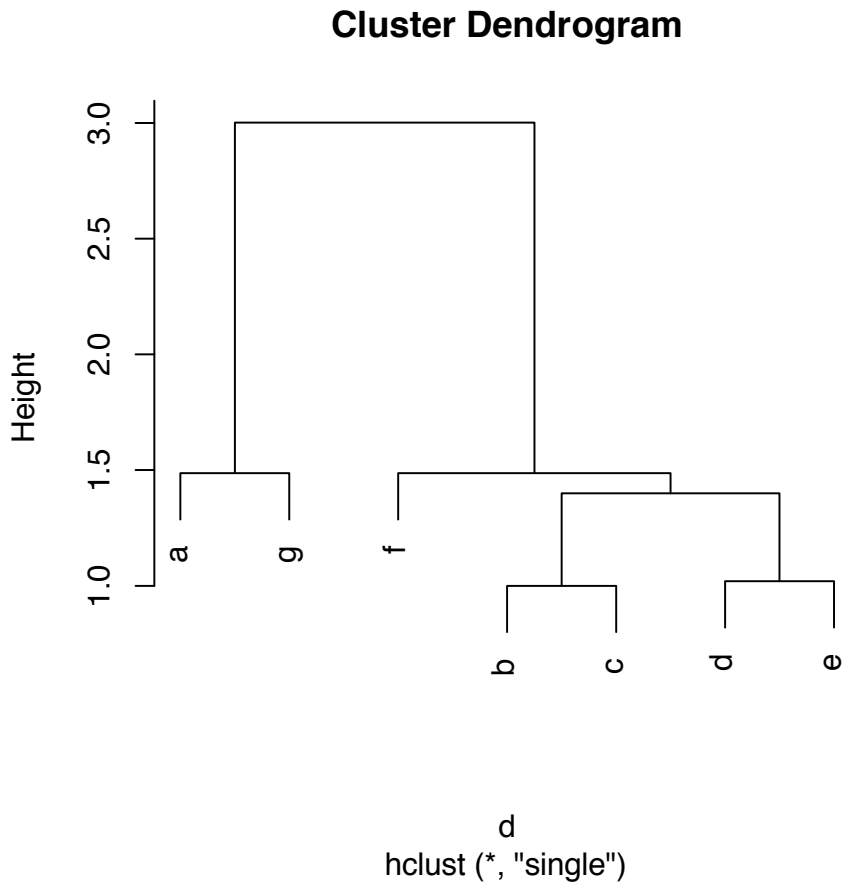


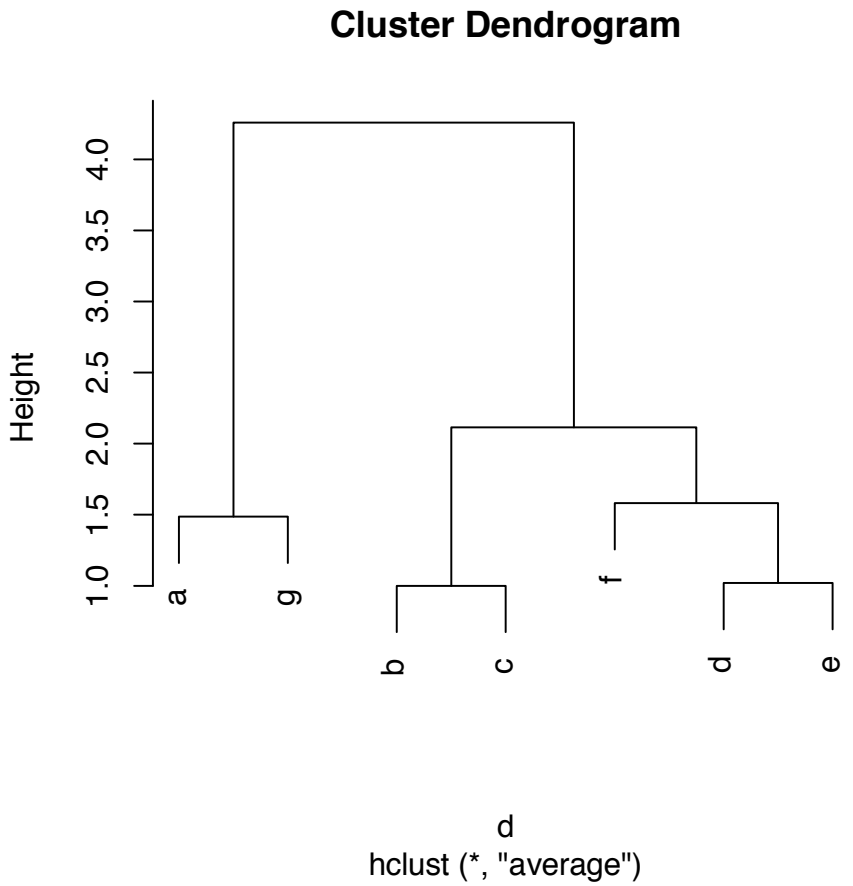
From eyeballing, here is a likely set of merges:

b,c
d,e
a,g,
(d,e),f
(b,c),((d,e),f)
ALL

Different linkages







dimension reduction
(exploratory data analysis)

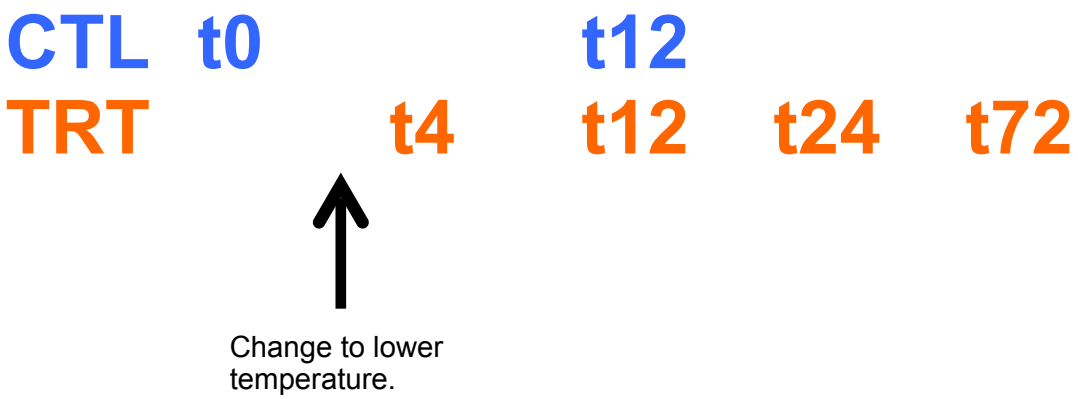
"To consult the statistician after an experiment is finished is often merely to ask [them] to conduct a post mortem examination. [They] can perhaps say what the experiment died of." R. A. Fisher

Motivation for exploratory data analysis: Case Study

(from Stefano, a former M.Sc. student in my Institute)

He is studying gene expression in fruitfly and is interested in transcriptional responses following "heat shock".

Basic schematic of experiment:

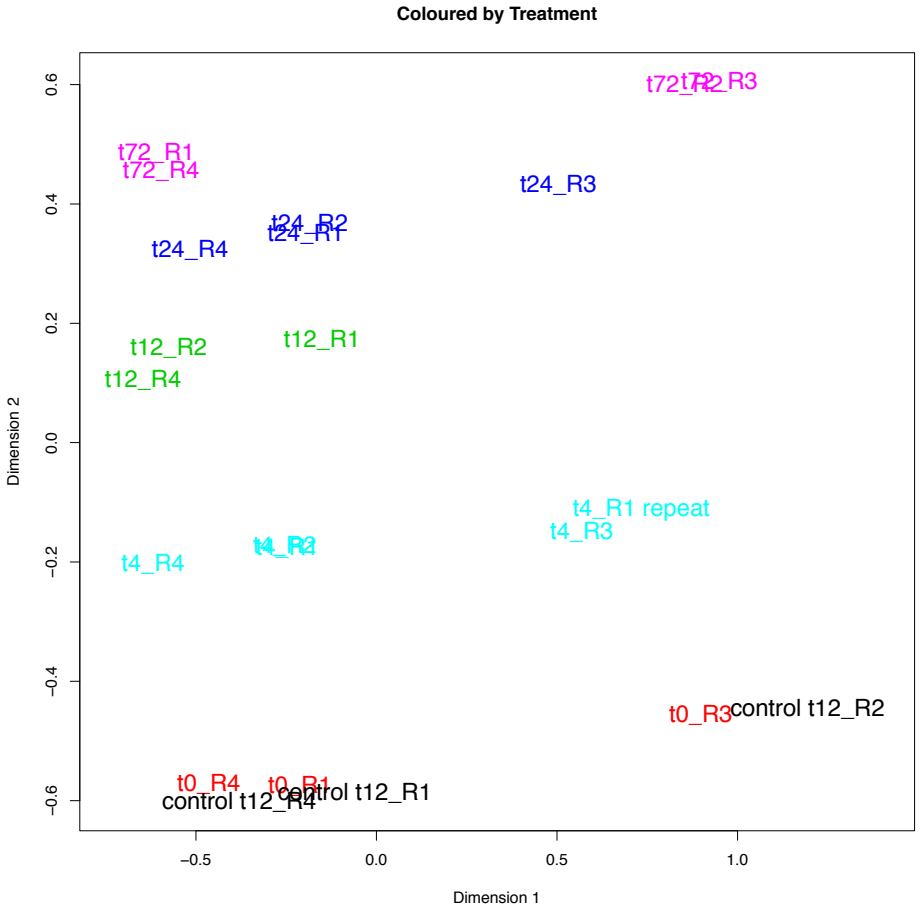


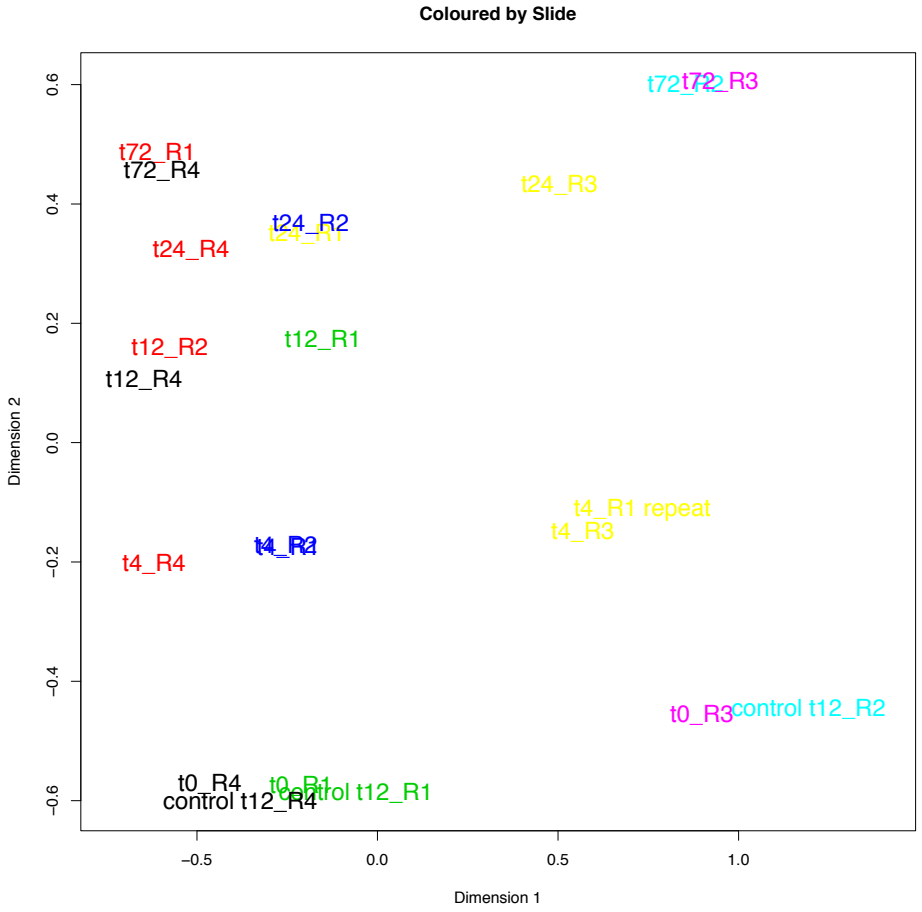
~4 replicates for each condition

http://en.wikipedia.org/wiki/Multidimensional_scaling

```
library(limma)
plotMDS(d)  # 'd' is a matrix
"Plot samples on a two-dimensional scatterplot so that
distances on the plot approximate the typical log2 fold
changes between the samples."
```

Take a close look at where the 24 samples are to each other relative to the X- and Y-axes





22 samples x ~20,000 genes

reduced to 22 samples x 2 dimensions

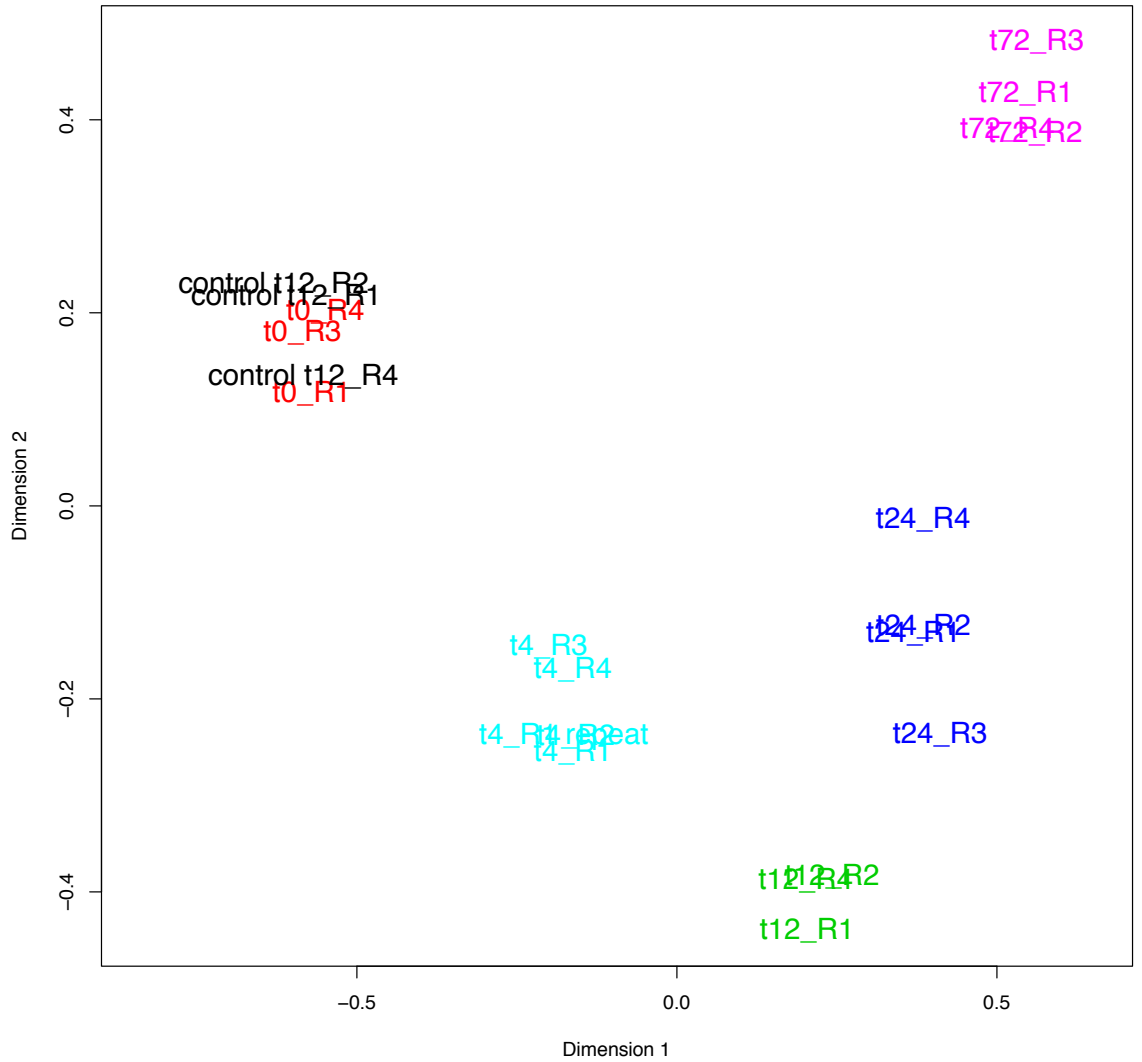
Magic: Surrogate variable analysis to detect and "remove" batch effects

OPEN ACCESS Freely available online PLOS GENETICS

Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

Jeffrey T. Leek[1], **John D. Storey**[1,2*]

1 Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

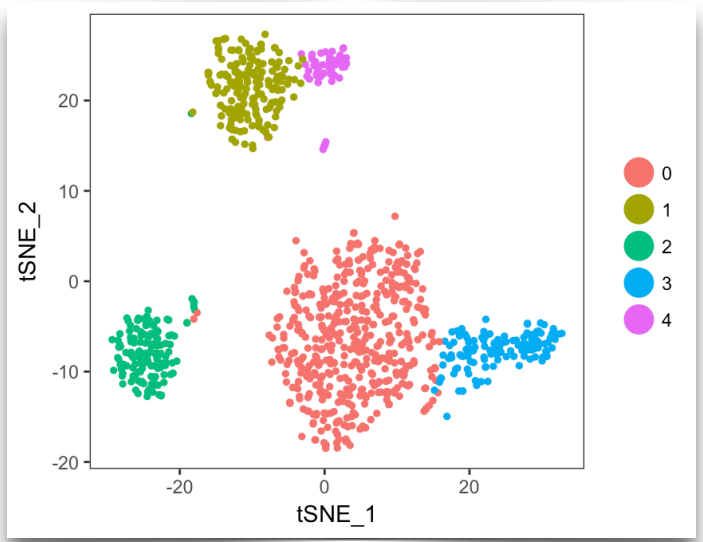


Dimension reduction: general introduction

- Many types of data come as a matrix of N samples (e.g., cells, patients) x G features (e.g., genes, proteins)
- Each sample is a point in G-dimensional space
- Goal: represent the data in 2-3 dimensions, but preserve **structure** as best as possible (i.e., points that are **close** in G dimensions should be close in 2 or 3 dimensions)

Dimension reduction is versatile

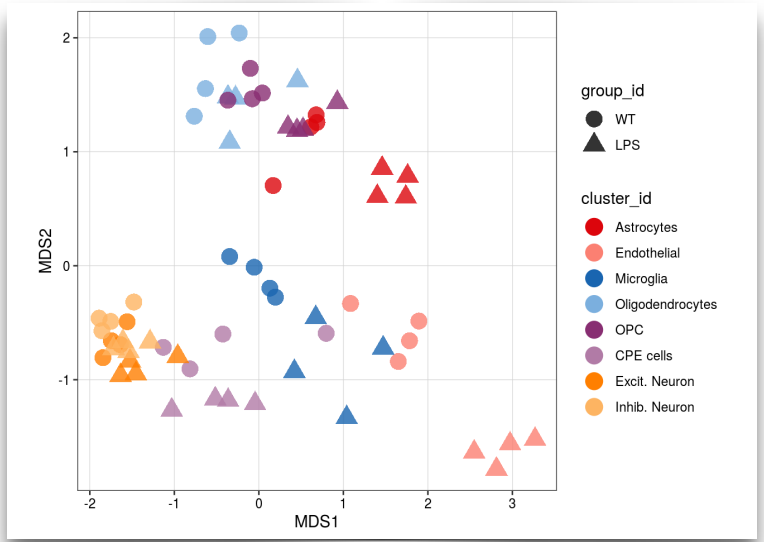
K features x N cells —> 2 dimensions x N cells



Each point = **single cell**

(10x PBMC)

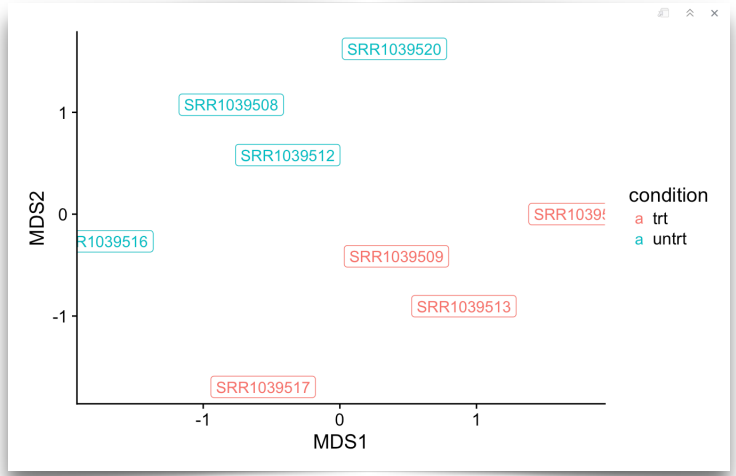
N cells x K features —> N cell subpopulations x 2 dimensions



Each point = **subpopulation from a single sample**

(LPS mouse cortex)

P samples x K features —> P samples x 2 dimensions



Each point = **sample**

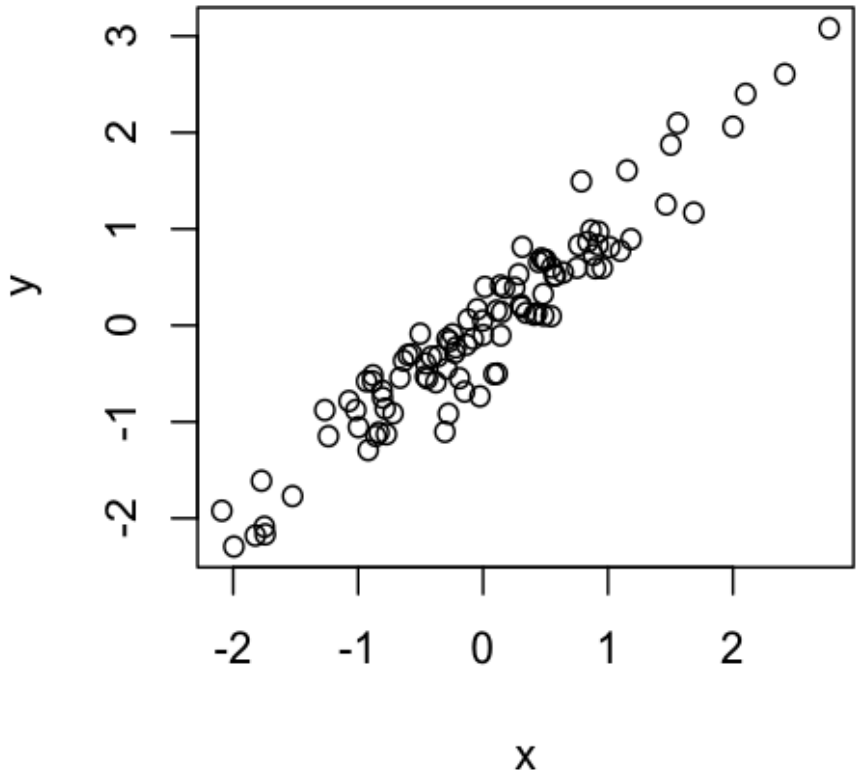
(airway)

Introduction to dimension reduction: PCA (principal components analysis)

- Form successive *linear* combinations of the features that are: orthogonal, ordered by variance

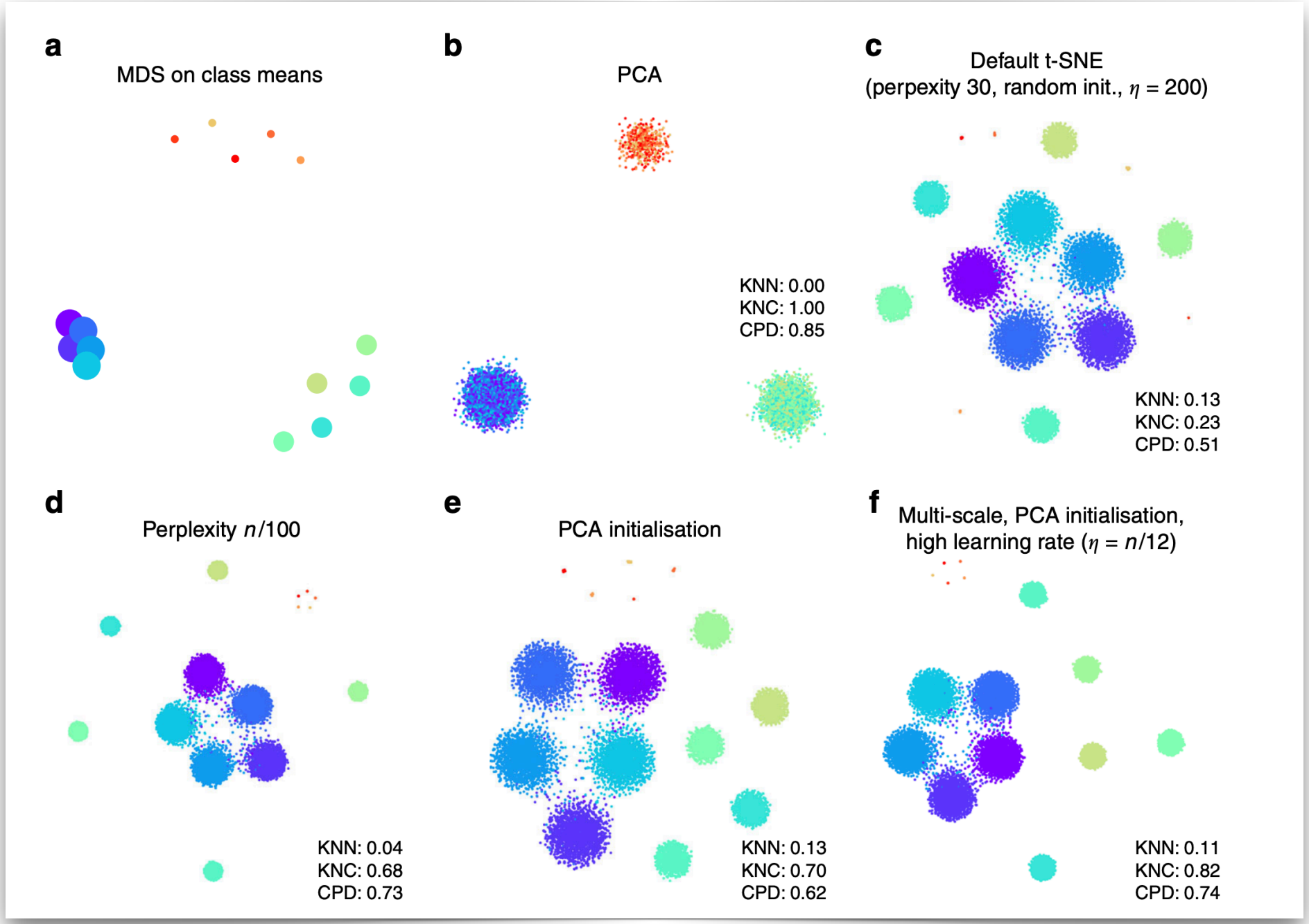
$$Y = XA$$
$$Y_{rk} = a_{1k}x_{r1} + a_{2k}x_{r2} + \cdots + a_{pk}x_{rp}$$

- A is the loadings matrix
- Typically, first 2-3 columns ('principal components') of Y are retained for visualisation; often top P PCs are retained for other analyses (e.g., clustering)



https://strata.uga.edu/software/pdf/pcaTutorial.pdf

Many variations (linear/non-linear), many notions of distance, many ways to "compress"



Kobak et al. 2019