

# Module 1: Working with genomics data

Andrew Gentles

Medicine (BMIR) and Biomedical Data Sciences

# Some selected topics in systems biology

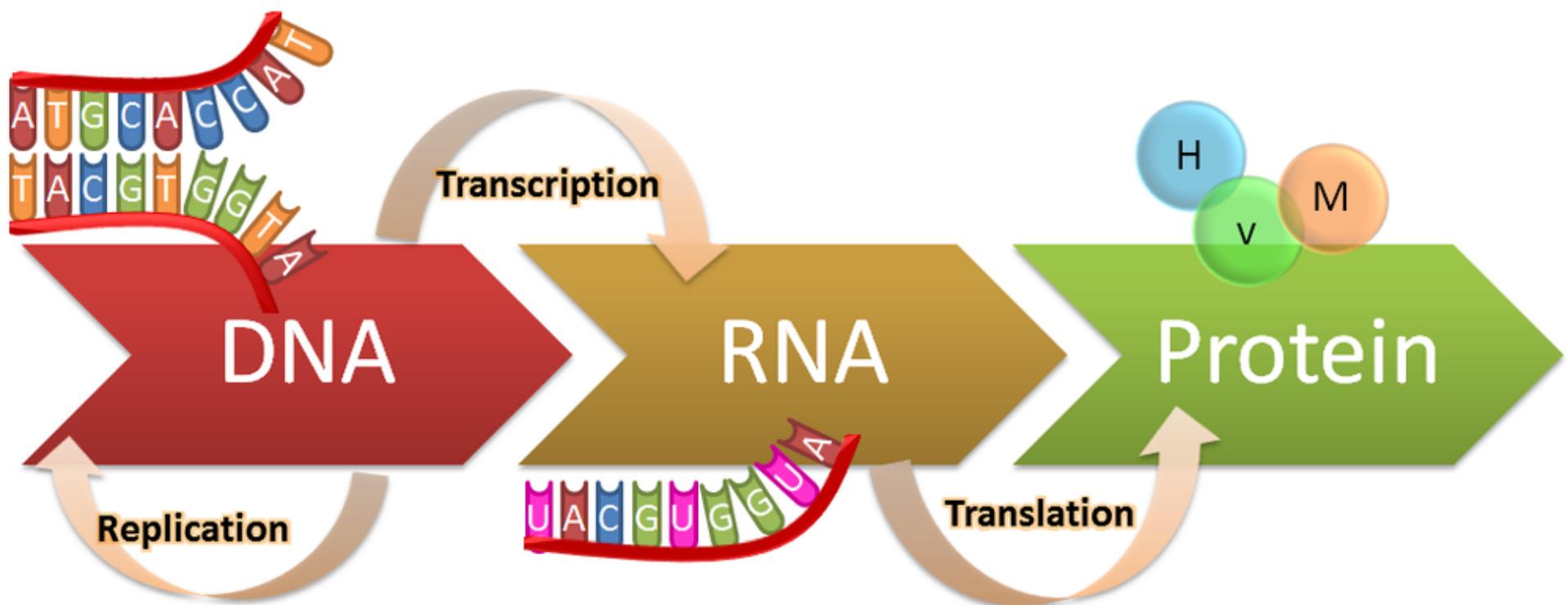
- Brief review of genomics data types
- Reconstructing regulatory networks
  - Transcriptional networks
- Connecting genomics with outcomes (survival)
- Dissecting tissues (e.g. tumors) at single cell resolution

# Cancer systems biology – one definition

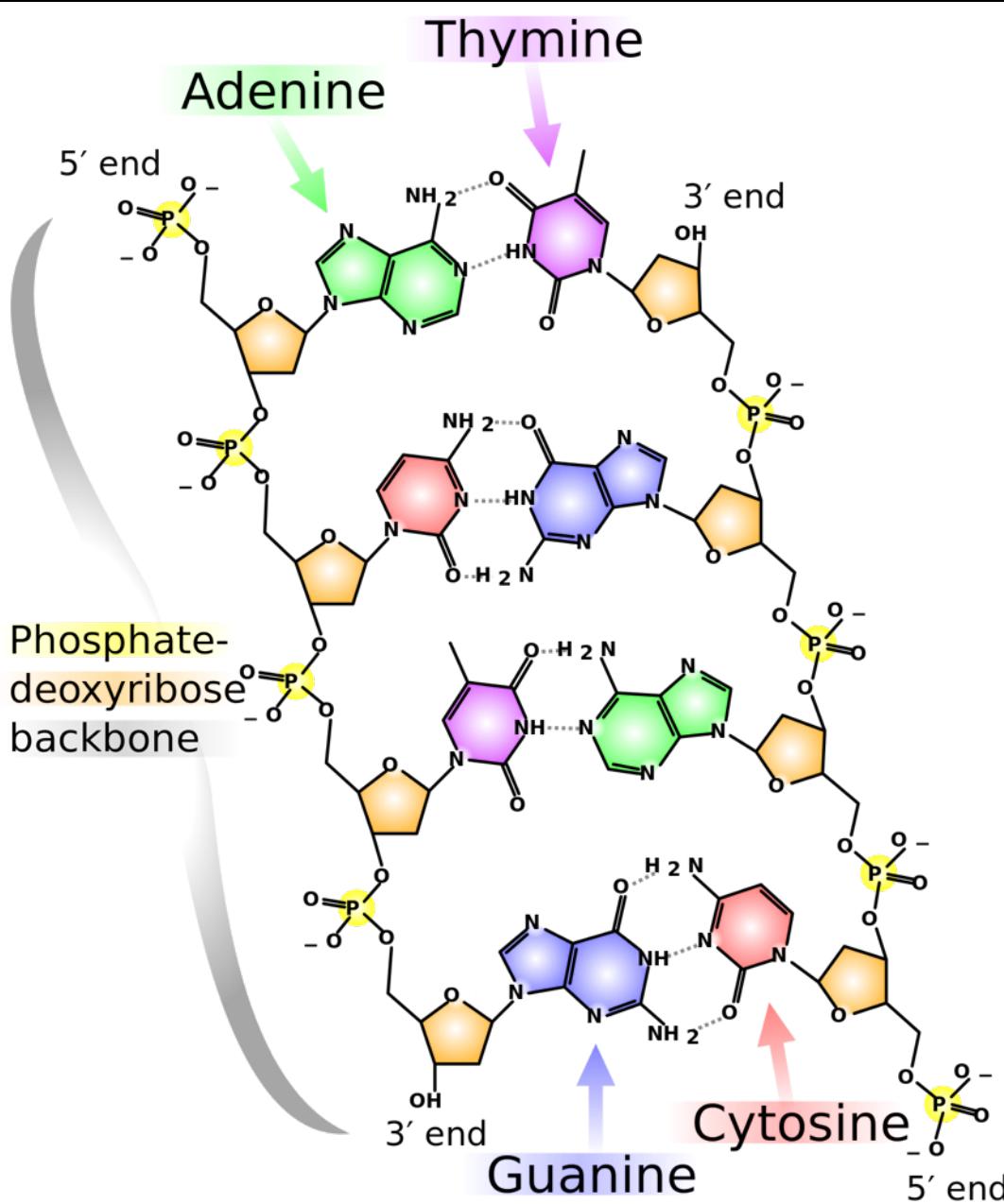
- Cancer systems biology adopts a holistic view of cancer aimed at integrating its many biological scales, including genetics, signaling networks, epigenetics, cellular behavior, histology, (pre)clinical manifestations and epidemiology.

[https://en.wikipedia.org/wiki/Cancer\\_systems\\_biology](https://en.wikipedia.org/wiki/Cancer_systems_biology)

# Central dogma of molecular biology

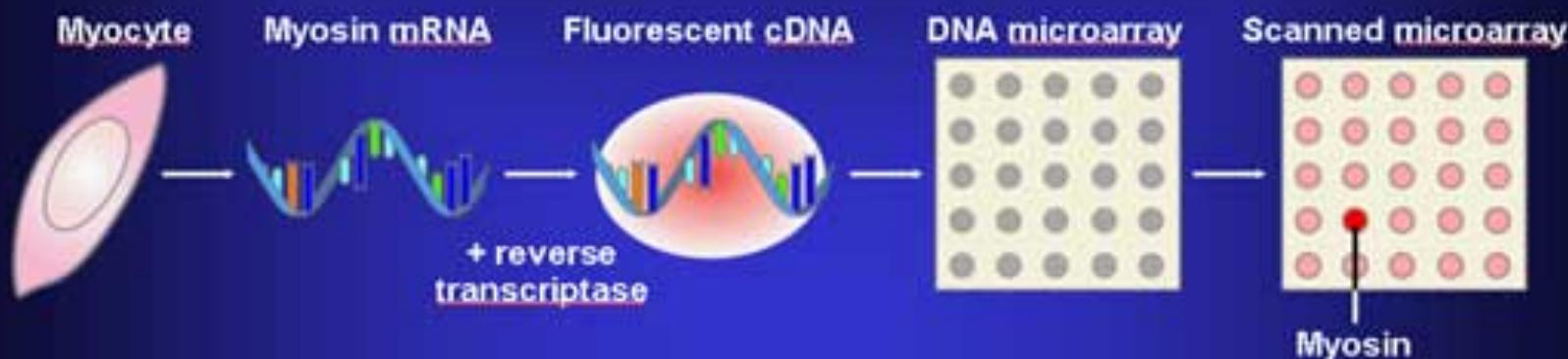


# DNA and RNA complementarity

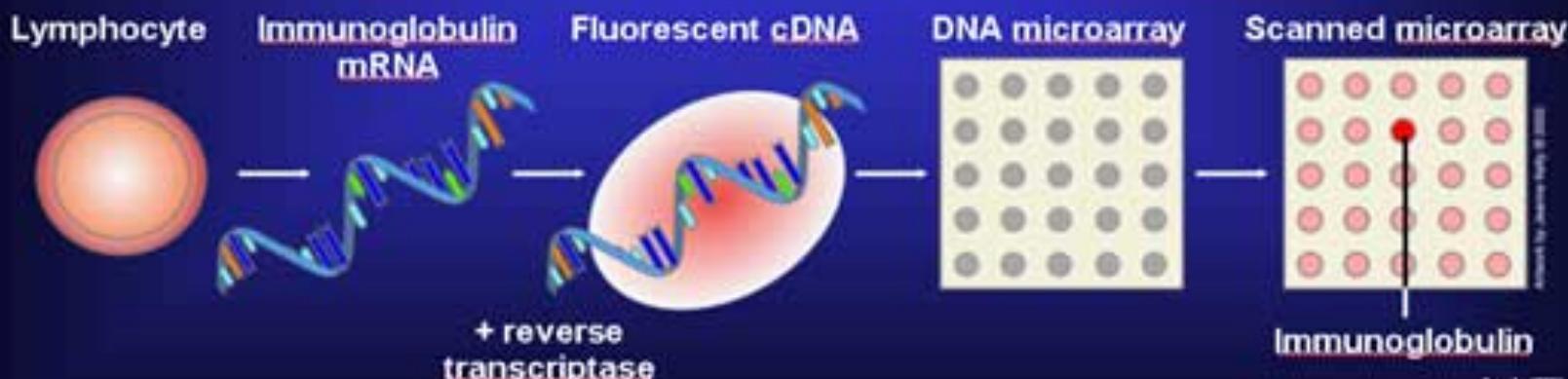


# Using DNA Microarrays to Measure Gene Expression

Fluorescent cDNA from muscle cell lights up myosin gene



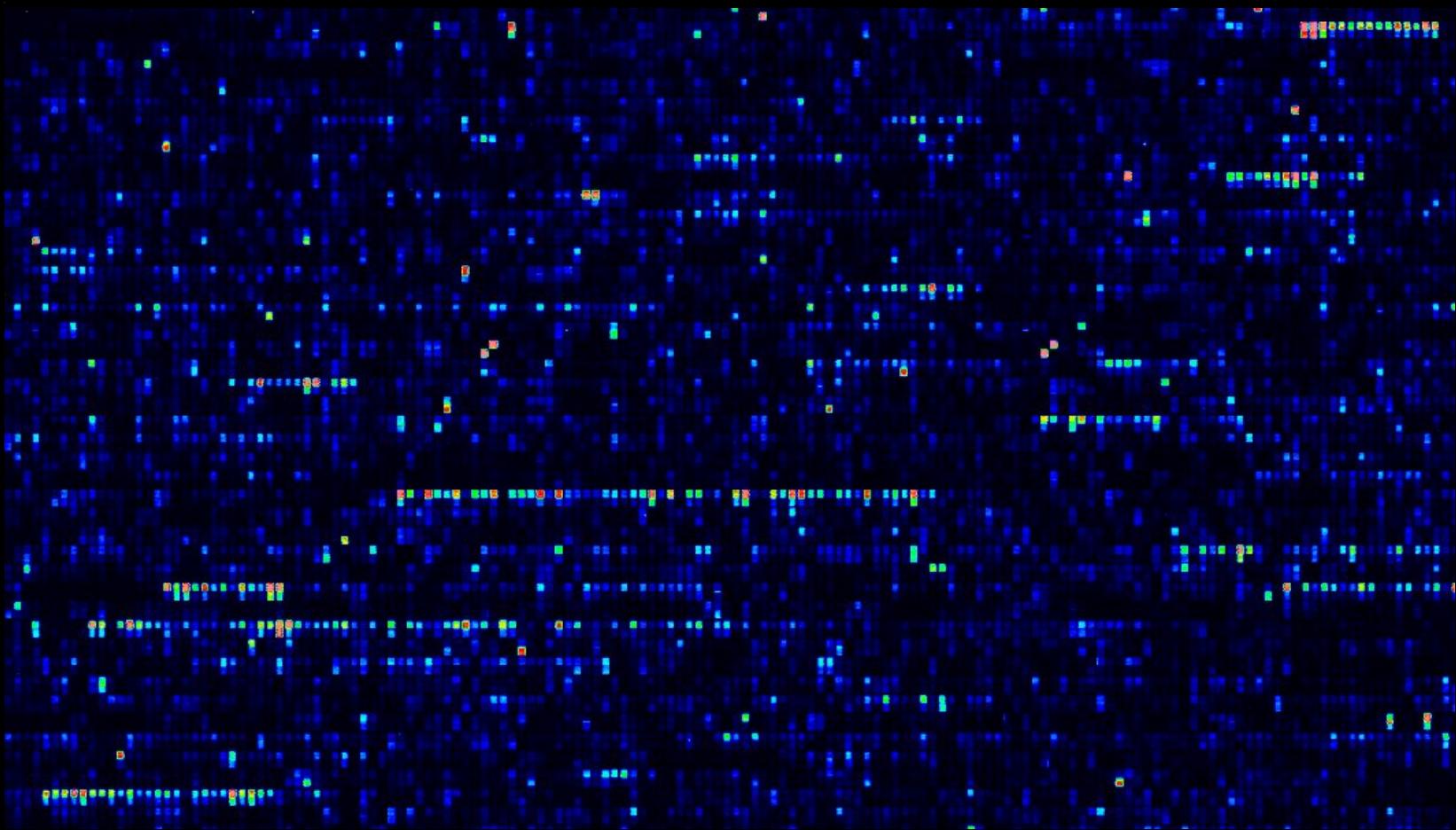
Fluorescent cDNA from lymphocyte lights up immunoglobulin gene



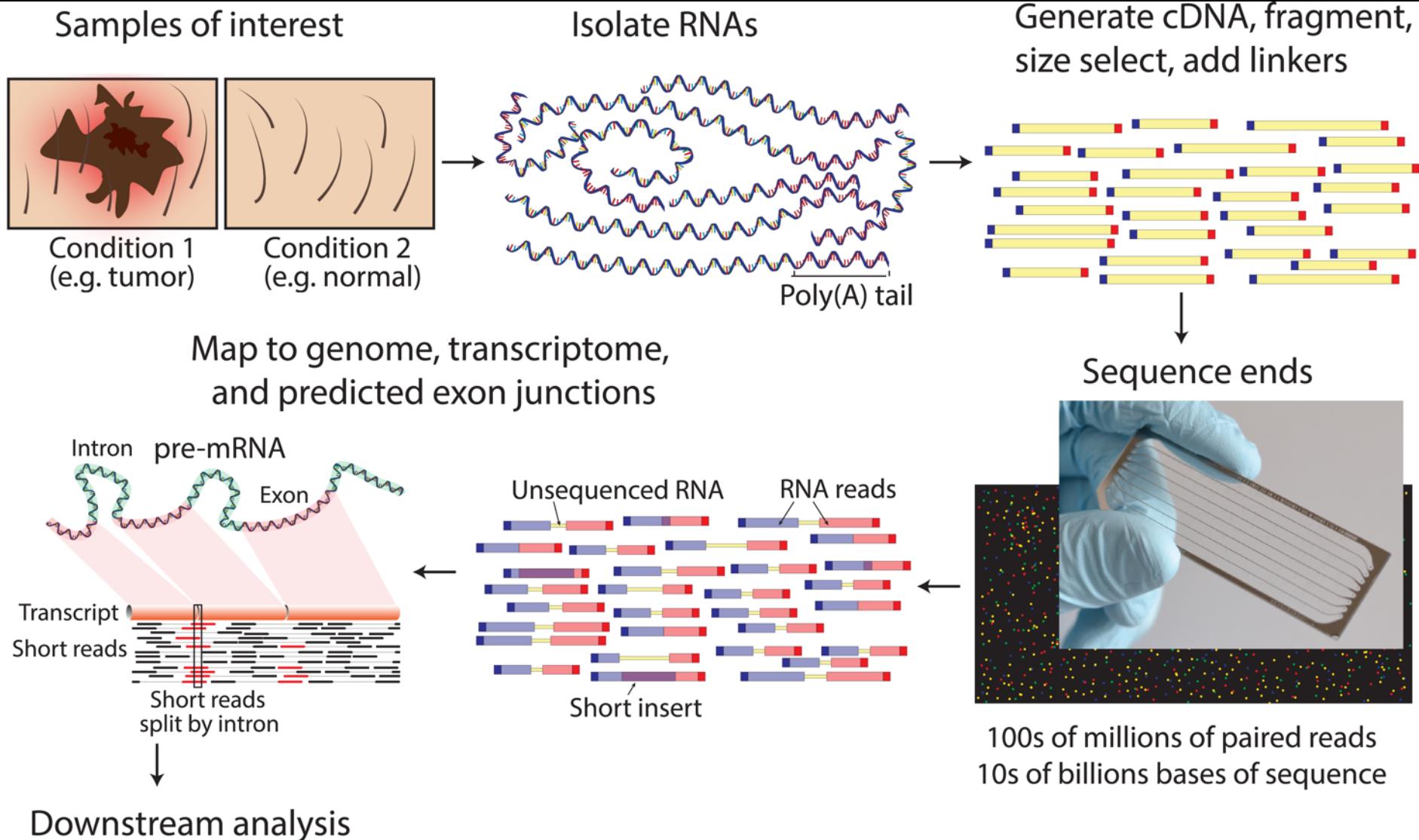
# We can assess activity of thousands of genes in a sample



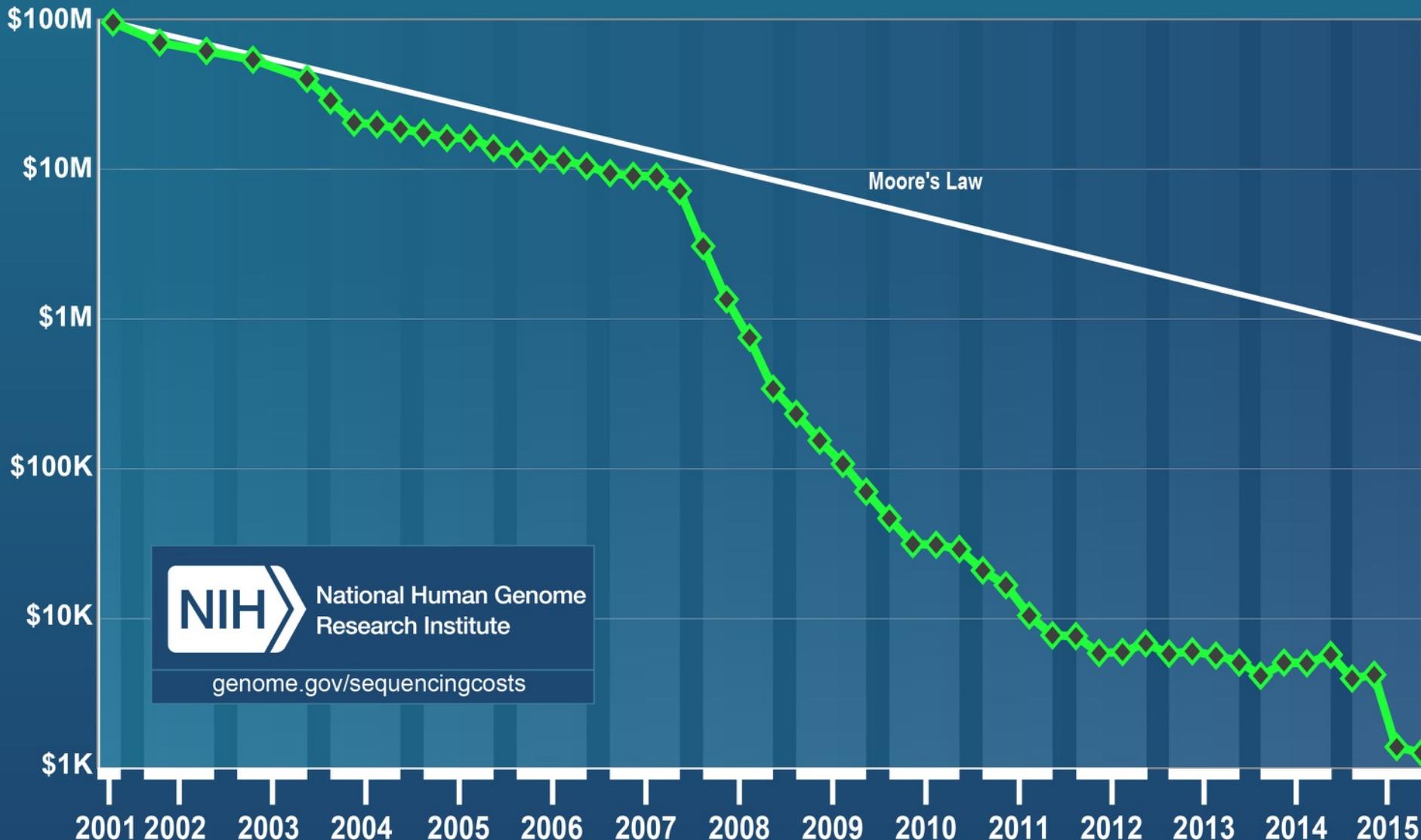
If a gene is expressed, it will bind to its complementary probe on the array



# Next generation sequencing



# *Cost per Genome*



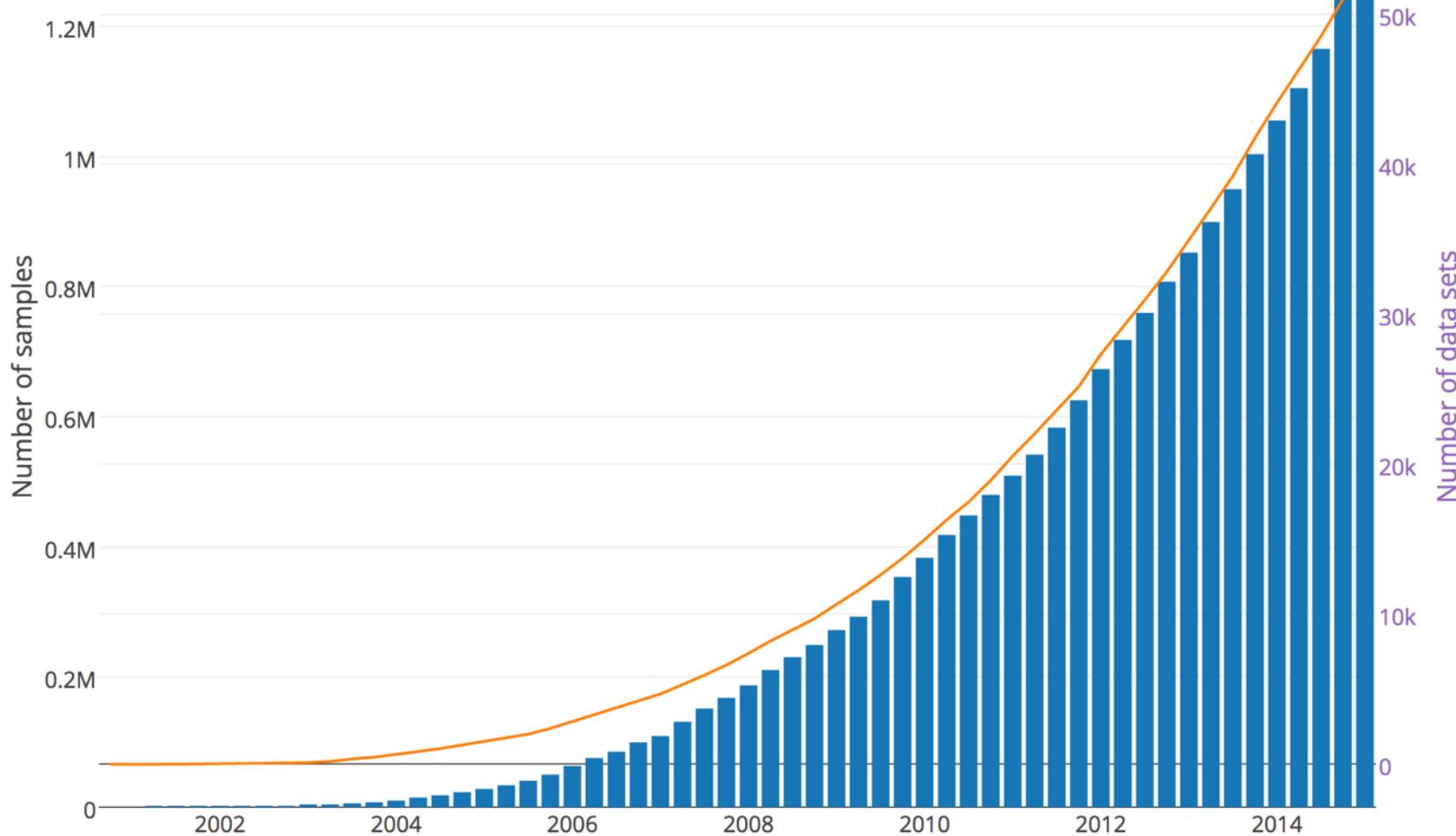
National Human Genome  
Research Institute

[genome.gov/sequencingcosts](http://genome.gov/sequencingcosts)

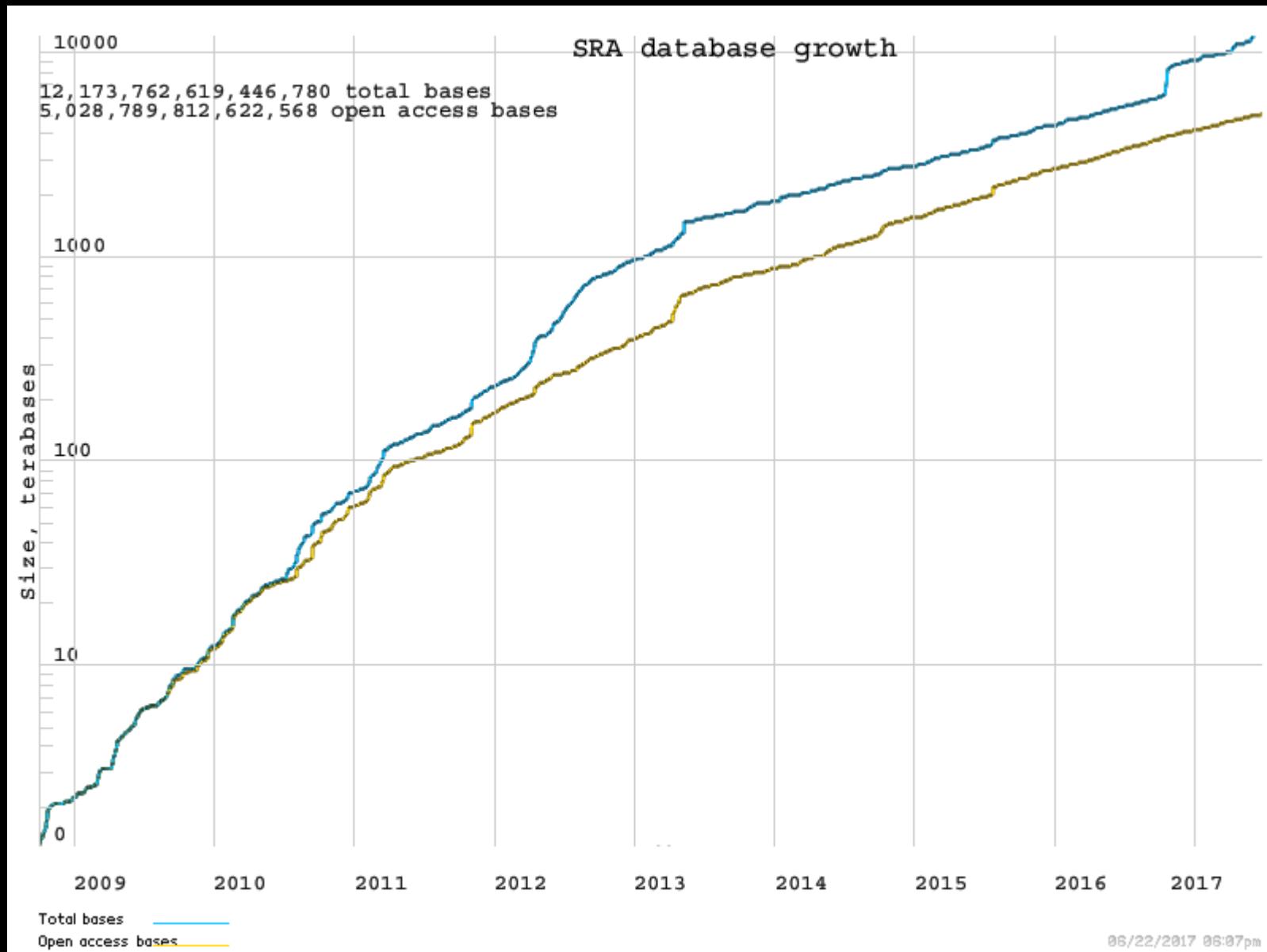
# Growth of gene expression omnibus at NCBI

<https://plot.ly/~sambucas/34/the-growth-of-the-gene-expression-omnibus/>

Samples  
Datasets



# Sequence data are accumulating rapidly



## Raw data – ballpark sizes

| Data type                             | Typical size ~    |
|---------------------------------------|-------------------|
| DNA microarray (Affymetrix cell file) | 10 Mb             |
| TCGA histology slide scan (SVS)       | 300 Mb            |
| Exome sequence fastq (100X)           | 10 Gb             |
| Single RNA-seq experiment             | 10 Gb             |
| Single whole genome                   | 600 Gb            |
|                                       |                   |
| Complete TCGA dataset (11,000 tumors) | 50 Pb (50,000 Tb) |

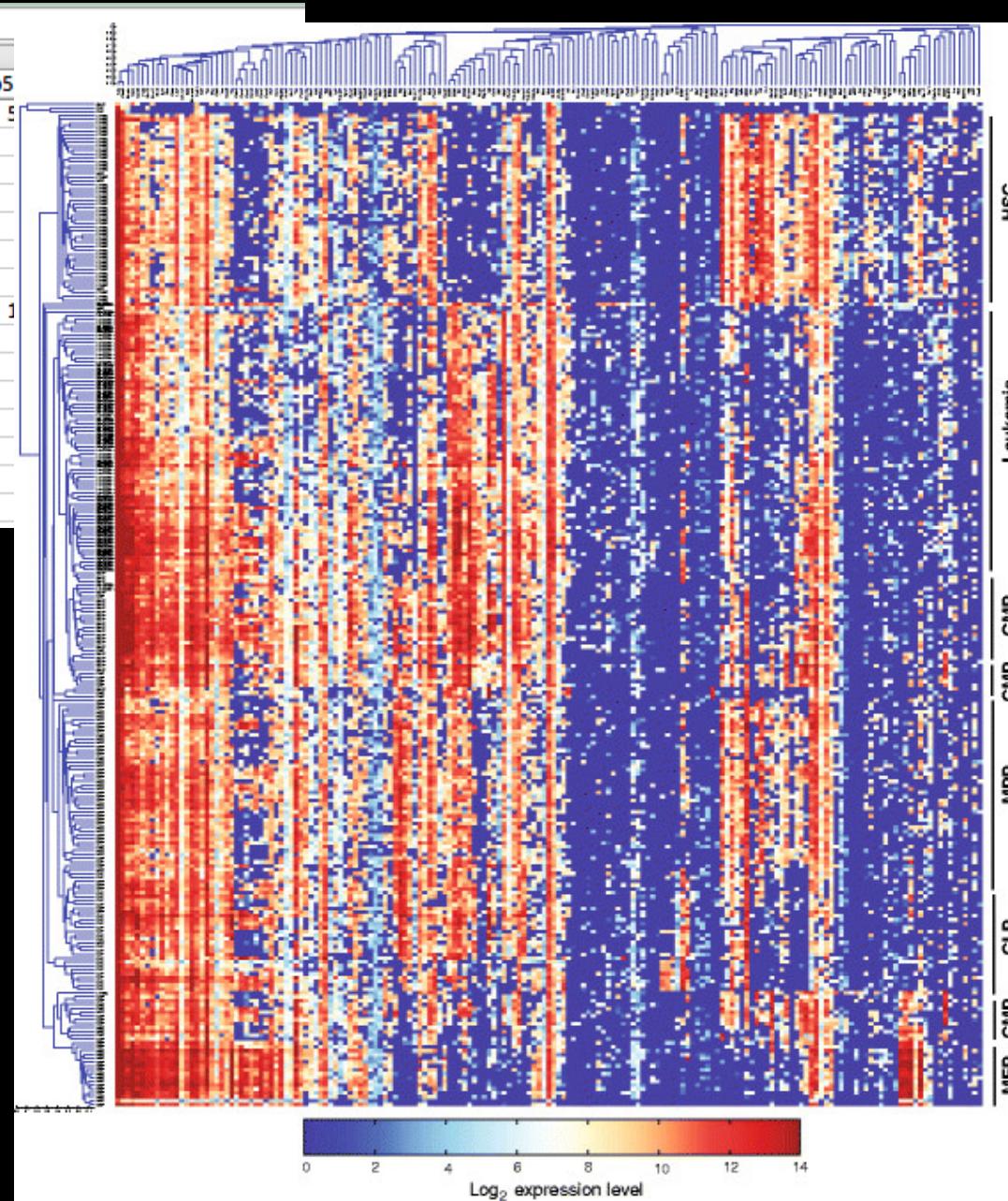
- Sequencing is cheap(ish)
- What about the analysis ?
  - Storage
  - Compute costs
  - Interpretation

# Endpoint is moderately large scale data

| J15 | A            | B        | C        | D        | E       | F    |
|-----|--------------|----------|----------|----------|---------|------|
| 1   | Acc ID       | Exp1     | Exp2     | Exp3     | Exp4    | Exp5 |
| 2   | NM_007818    | 67540.89 | 70924.09 | 80243.76 | 3501.2  | 5    |
| 3   | NM_001105160 | 811.93   | 801.36   | 740.71   | 128.67  | 5    |
| 4   | NM_028089    | 190.41   | 211.06   | 236.19   | 9.05    | 5    |
| 5   | NM_016696    | 66.77    | 57.56    | 101.09   | 750.9   | 5    |
| 6   | NM_013459    | 3.3      | 11.29    | 1.89     | 735.82  | 5    |
| 7   | NM_007809    | 45.34    | 36.12    | 51.02    | 245.27  | 5    |
| 8   | NM_009999    | 103.04   | 370.21   | 200.29   | 17.09   | 5    |
| 9   | NM_133960    | 7708.78  | 6976.38  | 6569.04  | 1731    | 1    |
| 10  | NM_027881    | 31.32    | 10.16    | 24.56    | 268.39  | 1    |
| 11  | NM_054053    | 31.32    | 24.83    | 19.84    | 323.68  | 1    |
| 12  | NM_007377    | 47.81    | 89.17    | 70.86    | 370.93  | 1    |
| 13  | NM_028064    | 703.95   | 689.62   | 662.29   | 214.11  | 1    |
| 14  | NM_008182    | 222.56   | 339.73   | 226.75   | 30.16   | 1    |
| 15  | NM_013661    | 12.36    | 11.29    | 8.5      | 97.51   | 1    |
| 16  | NM_007815    | 20613.09 | 25218.13 | 31540.46 | 5209.07 | 1    |

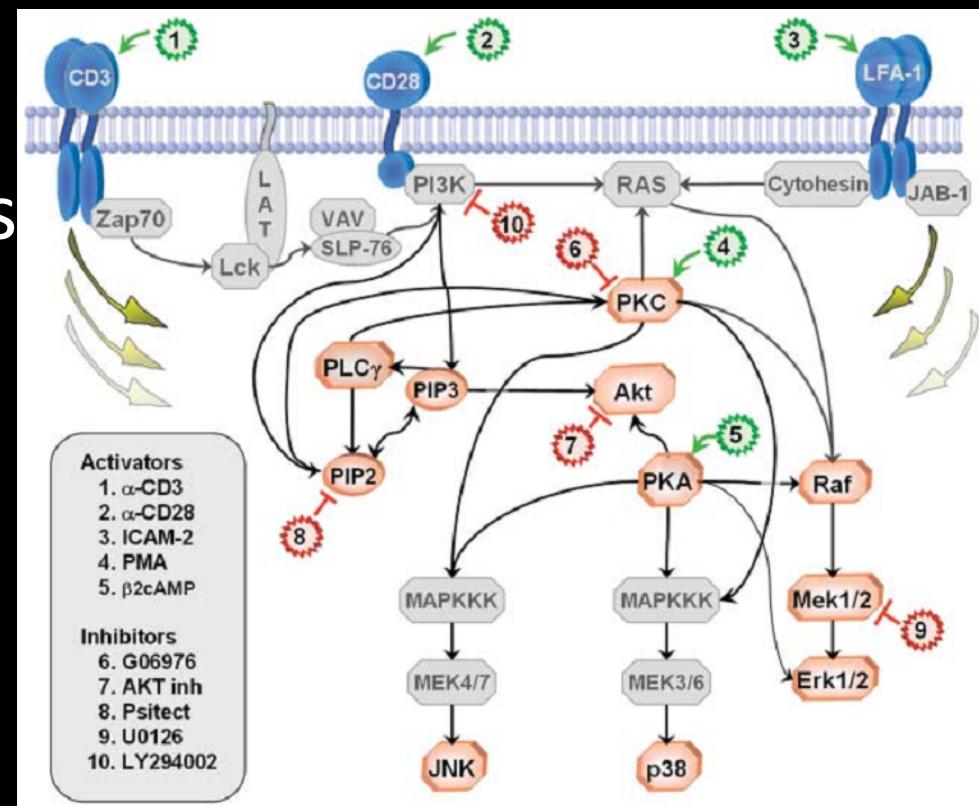
10-1000's columns (samples)

10,000's-100,000's rows (genes, transcripts)

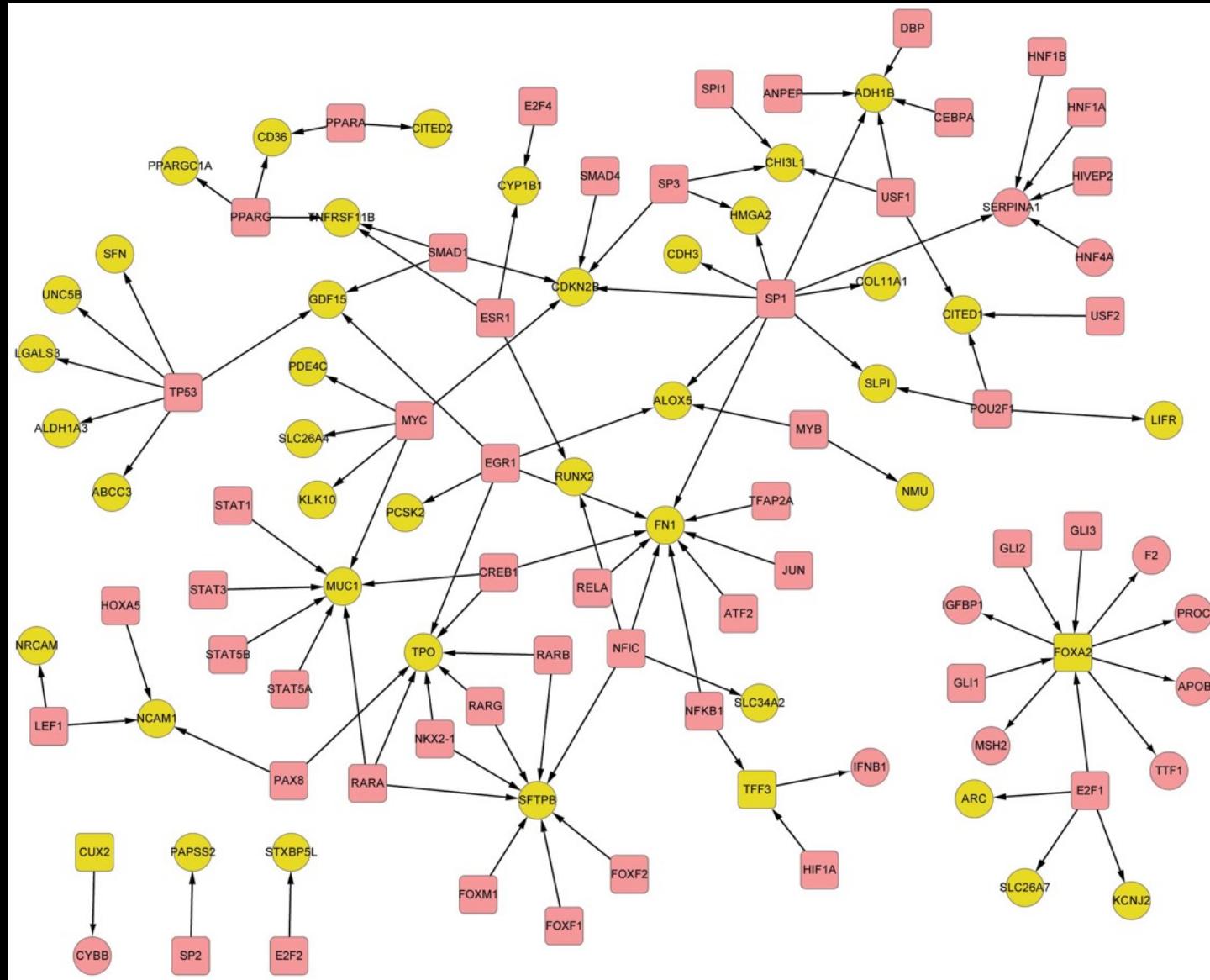


# Network types

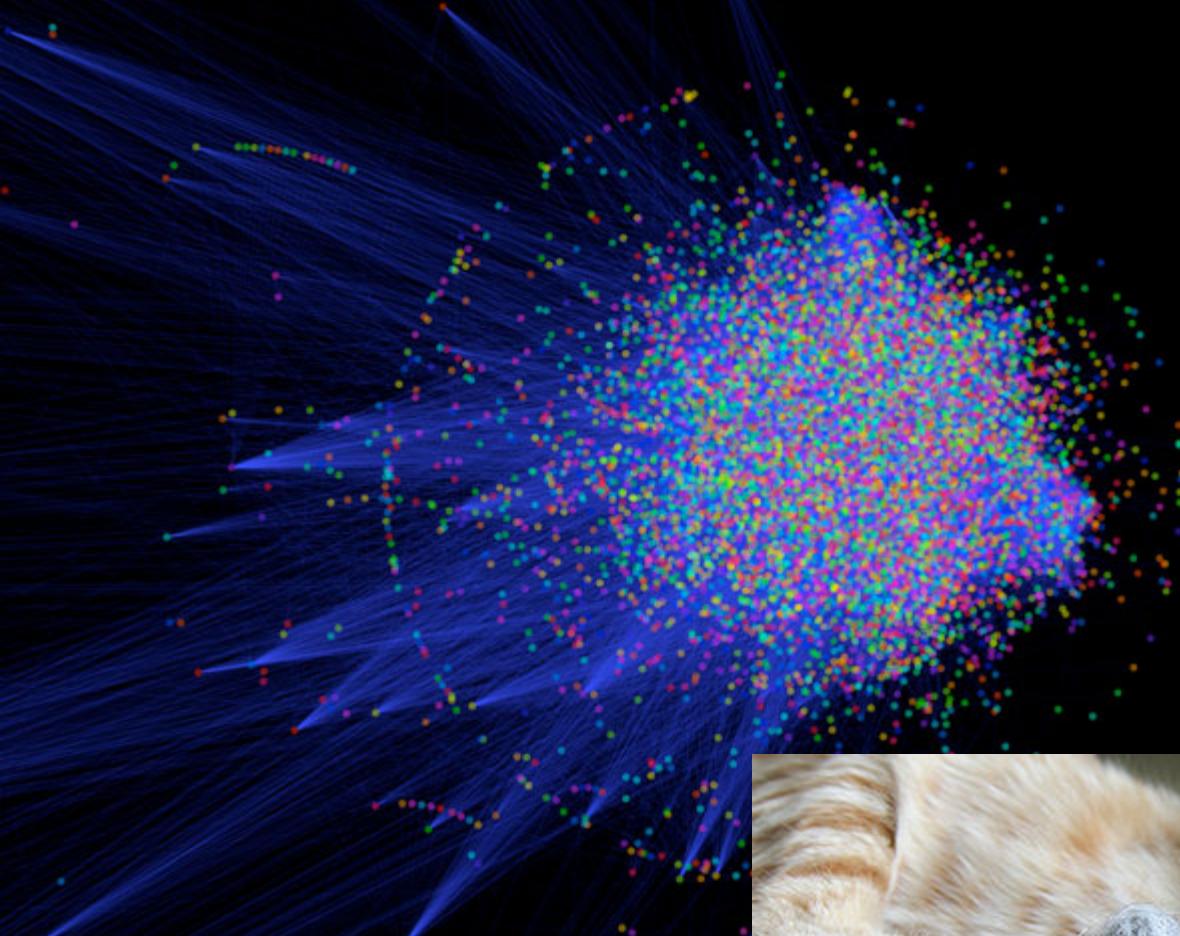
- Protein-protein
- Protein-DNA
- miRNA-RNA
- Transcriptional (expression) networks
- Signaling networks



# Gene regulatory networks



They are a “flowchart” of interactions, not a detailed model



Hard to interpret!

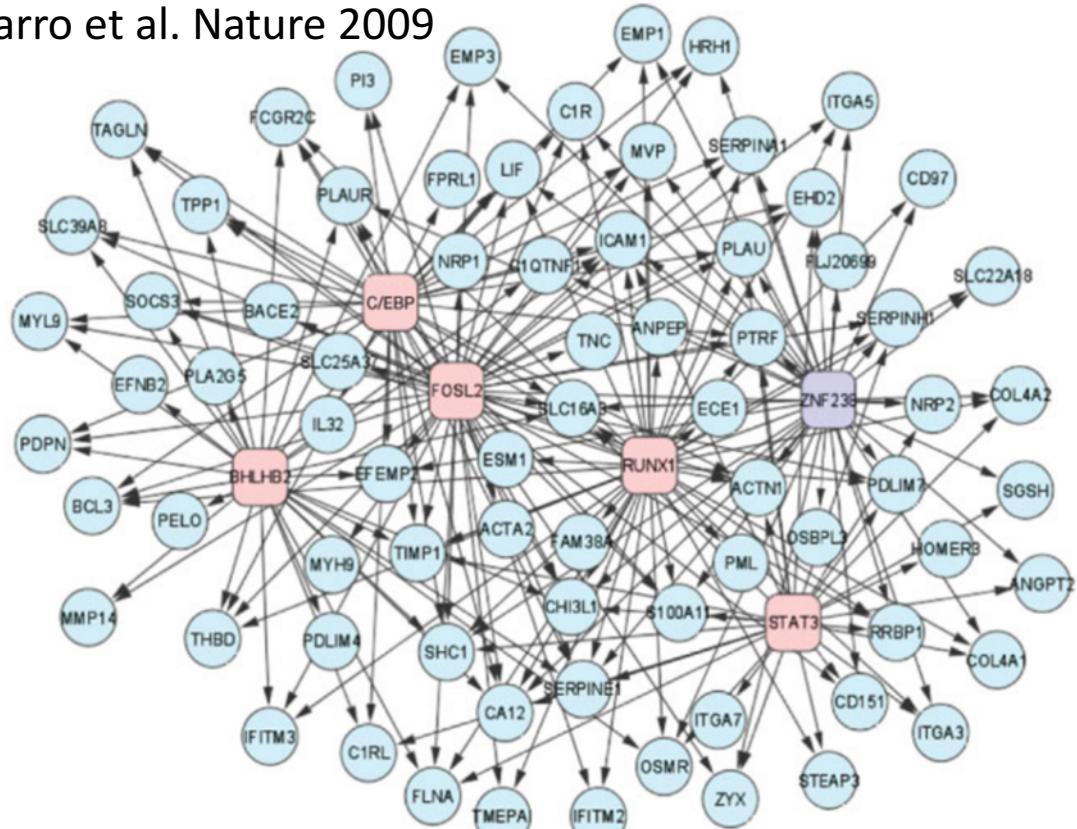


# Transcriptional regulatory networks

- Imperfect but powerful representation of the system
- Example: Glioblastoma subtypes
  - Hairball -> targets of transcription factors -> which are differentially expressed between GBM types

Actual causality is hard from observational data

Carro et al. Nature 2009



- How do these networks compare in size to ones in the physical world
  - E.g. electrical system, telecommunications
- What sort of qualitative differences are there?
- Do cats or dogs produce worse hairballs?

# Associating survival with expression levels

| Variable | Patient1 | Patient2 | Patient4 | Patient5 | Patient6 | Patient7 | ... |
|----------|----------|----------|----------|----------|----------|----------|-----|
| Time     | 12       | 3        | 8        | 35       | 14       | 22       | ... |
| Status   | 1        | 1        | 1        | 0        | 0        | 1        | ... |
| Gene1    | 1.45     | 0.15     | -0.59    | -1.88    | -0.83    | -0.26    | ... |
| Gene2    | 0.94     | -0.35    | 2.66     | -0.23    | 2.09     | -0.13    | ... |
| Gene3    | 0.91     | -0.32    | -0.82    | -0.35    | 0.86     | 0.32     | ... |
| ...      |          |          |          |          |          |          |     |

We could look at a fixed time e.g. alive/dead at 5 years – but this throws away information

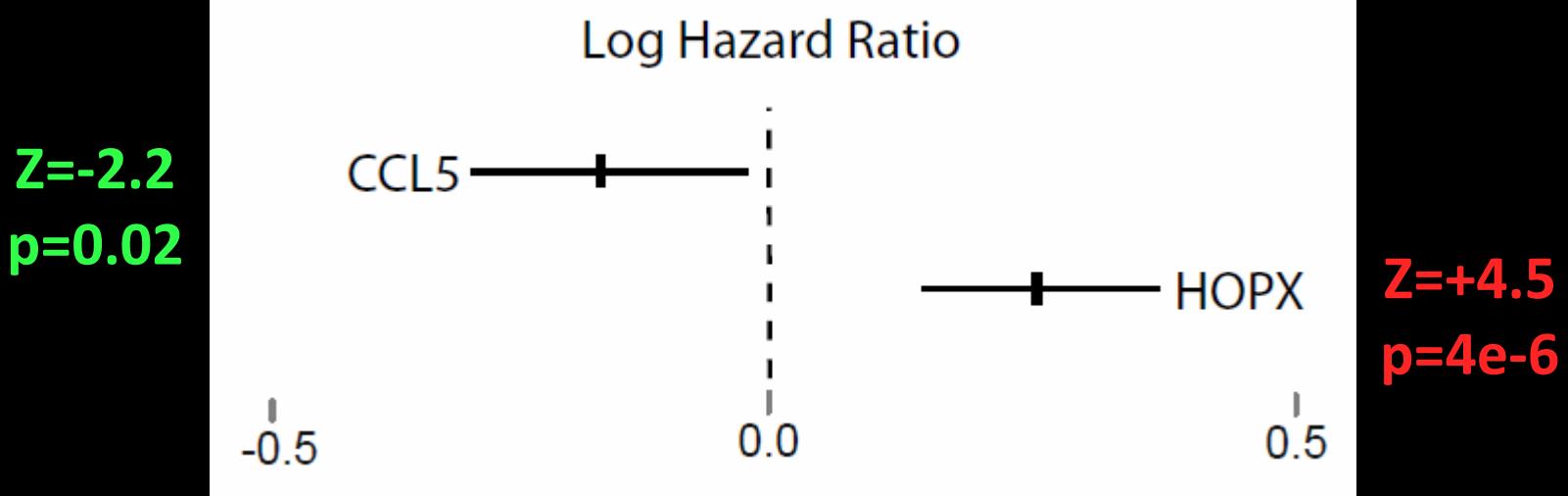
Survival often assessed by hazard ratio – the increase in risk of an event (e.g. death) for each unit increase in some variable (e.g. age, expression level of a gene)

*Extra detail:*

$$H(t, x_i) = h_0(t) \times \exp\{b_i x_i\}$$

$X_i$  = expression of gene  $i$  at sample collection

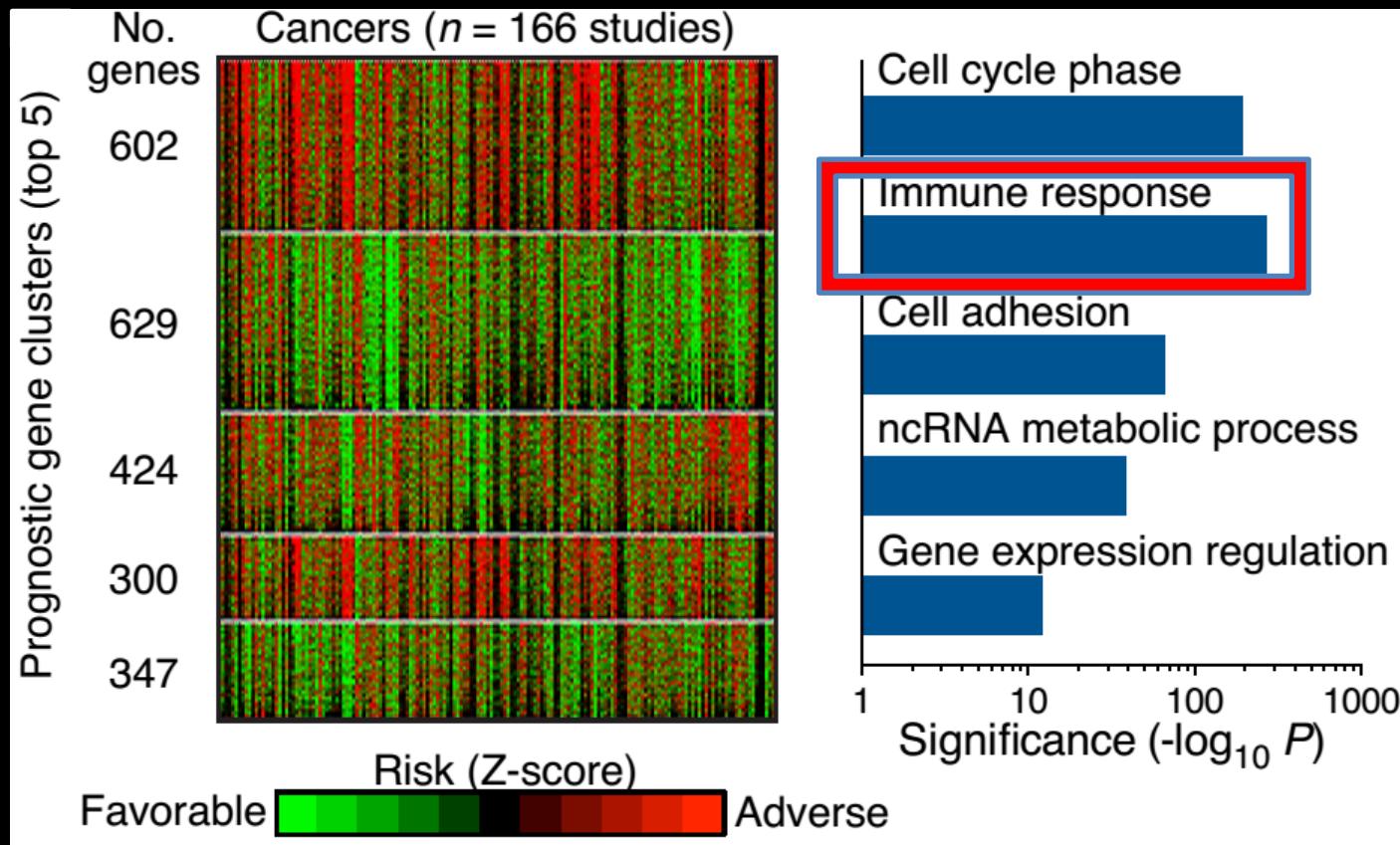
# Gene expression ~ survival



“Good” genes

“Bad” genes

# Clustering of outcome matrix -> biological processes



# Prognostic influence of immune infiltrates

| No. data sets | No. samples analyzed | B cells naive | B cells memory | Plasma cells | T cells CD8 | T cells CD4 naive | T cells CD4 memory RO unactivated | T cells CD4 memory RO activated | T cells follicular helper | T cells gamma delta | T cells regulatory (Tregs) | NK cells unstimulated | NK cells stimulated | Monocytes | Macrophages M0 | Macrophages M1 | Macrophages M2 | Dendritic cells unstimulated | Mast cells stimulated | Mast cells unstimulated | Eosinophils | Neutrophils               |                              |                     |
|---------------|----------------------|---------------|----------------|--------------|-------------|-------------------|-----------------------------------|---------------------------------|---------------------------|---------------------|----------------------------|-----------------------|---------------------|-----------|----------------|----------------|----------------|------------------------------|-----------------------|-------------------------|-------------|---------------------------|------------------------------|---------------------|
| 5             | 744                  | -1.85         | -1.53          | 1.53         | -2.62       | 0.65              | 2.67                              | -2.38                           | -1.61                     | 2.28                | -1.49                      | 1.98                  | 0.65                | 0.55      | 0.92           | -2.03          | -0.80          | -2.10                        | -0.50                 | 3.17                    | 0.93        | AML                       |                              |                     |
| 1             | 174                  | -1.61         | 1.44           | 0.60         | 1.10        | 0.65              | -2.79                             | 0.87                            | -1.16                     | 1.49                | 1.10                       | 1.94                  | -1.40               | -0.73     | 0.33           | -0.85          | 0.70           | 0.32                         | 0.70                  | 0.32                    | 0.32        | B-ALL                     |                              |                     |
| 1             | 107                  | 2.08          | 1.09           | 2.44         | -0.35       | -1.39             | -1.51                             | -2.06                           | 0.74                      | -1.83               | 1.07                       | -2.39                 | -0.54               | -2.60     | 0.54           | 2.34           | 2.10           | -0.84                        | -2.38                 | -0.65                   | 0.22        | CLL                       |                              |                     |
| 1             | 158                  | 2.07          |                | -1.16        |             | -1.07             | 1.38                              | -2.21                           | -1.35                     | -1.11               | 0.93                       | -0.75                 | 0.87                | -2.71     | -1.35          | -1.83          | 0.54           | -0.47                        | 1.04                  | 0.54                    | -0.57       | 0.32                      | Burkitt's lymphoma           |                     |
| 3             | 594                  | 2.28          |                | 1.30         | -0.63       | 0.68              | -2.47                             | 1.99                            | -3.33                     | -1.70               | -1.46                      | 0.91                  | 1.24                | 2.94      | -4.03          | 0.08           | 3.56           | -0.54                        | 1.61                  | -2.96                   | 1.55        | 0.08                      | DLBCL                        |                     |
| 1             | 180                  | -0.92         | 1.50           | 0.44         |             | -1.57             | -1.82                             | 1.49                            | 0.56                      |                     | 1.32                       |                       | 0.97                | 0.75      |                | -1.06          | 1.93           | 1.34                         |                       |                         |             | FL                        |                              |                     |
| 2             | 189                  | -1.11         | -0.44          | 0.22         | 0.63        | 0.83              | 1.12                              |                                 | 1.67                      |                     | 1.72                       | 1.82                  | -1.05               | 1.51      | 0.70           | -0.93          |                | -2.11                        | 1.84                  |                         |             | Multiple myeloma          |                              |                     |
| 3             | 70                   | 0.64          |                | 0.77         |             | -1.02             | -0.90                             | 1.75                            | 1.28                      | -0.80               | 1.05                       |                       | -0.84               | 1.03      | 0.96           | -0.64          |                | -1.07                        | 0.92                  | 2.02                    |             | Astrocytoma               |                              |                     |
| 6             | 283                  | -0.88         | -1.03          | -1.05        |             |                   | -1.71                             |                                 | 3.11                      | 1.37                | 1.59                       | 0.70                  |                     | -2.80     | 2.49           | 0.96           | -0.90          |                              | -1.62                 |                         | -1.16       | 0.32                      | Glioblastoma                 |                     |
| 1             | 30                   | -0.91         | 1.14           | -0.93        |             |                   | -0.75                             | 0.30                            |                           | 1.62                | -1.78                      |                       |                     | -1.20     | -0.42          | -0.99          | 2.15           | -0.70                        | 0.31                  | 1.14                    |             | Meningioma                |                              |                     |
| 1             | 15                   | -0.54         | 1.11           | 1.91         | -0.64       | 0.96              | -0.34                             | 0.94                            | -1.07                     | 0.50                |                            | -0.62                 |                     | 0.05      | 1.07           | 1.43           | -0.93          | -0.39                        | -0.60                 | 0.05                    | -1.63       |                           | Oligodendrogloma             |                     |
| 1             | 30                   | -1.16         | -1.62          |              | -0.97       | 2.20              |                                   | -0.89                           | -2.07                     | 0.66                | 1.79                       | 0.55                  |                     | 0.42      | 0.53           | 1.50           | -0.33          | 1.79                         | 0.55                  |                         |             | Bladder cancer            |                              |                     |
| 4             | 567                  | 0.72          | -1.62          | -0.78        | -0.76       | 1.34              | -0.52                             |                                 | -1.28                     | -0.60               | 1.34                       | -1.51                 | 1.37                | 1.05      | 1.17           | 0.07           | 0.55           | 1.31                         | -0.60                 | 0.55                    | 3.91        | Breast cancer             |                              |                     |
| 3             | 236                  | -1.02         | 2.49           | -1.74        | -0.68       | 1.75              | -0.91                             | -0.90                           |                           | -1.80               | 0.68                       | 0.33                  | 1.77                | 0.57      | -1.08          | 1.80           | 1.12           | 1.62                         | 1.36                  | 1.86                    |             | Colon cancer              |                              |                     |
| 1             | 20                   | 0.89          | 0.87           | -1.86        | -0.81       | 0.94              | 1.20                              |                                 | -1.28                     | -0.61               | -0.74                      | 0.22                  | -1.27               | 1.78      | 1.41           | -0.53          | 1.32           | -0.64                        |                       | -0.94                   | 0.97        | 2.58                      | Ewing sarcoma                |                     |
| 1             | 18                   | 0.90          | 1.90           | -0.87        |             |                   | 0.73                              | 0.88                            | 0.99                      | 0.21                |                            | -1.04                 | 2.03                |           | -1.54          | -1.72          | -0.49          | 1.33                         | 1.21                  | 0.50                    | 1.66        | -1.27                     | Gastric cancer               |                     |
| 2             | 96                   | -2.21         | 1.18           | -1.97        | -2.01       | -1.43             | 2.34                              | -1.69                           | -2.70                     | -1.40               | 1.72                       | 1.30                  | -0.30               | -1.48     | 1.28           | -0.86          | 2.05           | 2.27                         |                       | 3.27                    | 0.61        | 0.34                      | Germ cell tumors             |                     |
| 2             | 76                   | 0.83          |                | -1.18        | 0.76        | 0.57              | -1.39                             |                                 | -1.81                     | 0.36                |                            | -1.46                 |                     | 1.79      | -0.85          | 0.22           | -2.15          |                              | 1.72                  |                         | 0.92        | 0.34                      | Head and neck cancer         |                     |
| 9             | 902                  | -1.43         | -1.40          | -1.59        |             | -0.39             | -2.67                             | 3.00                            | -0.99                     | -0.54               | -0.73                      | 1.23                  |                     | -1.09     | 2.62           | 2.04           | 2.36           | -1.37                        | 3.64                  | -2.94                   | 2.26        | 0.08                      | 3.65                         | Lung adenocarcinoma |
| 7             | 408                  | 2.02          |                | -1.22        | 1.71        |                   | -1.75                             | -0.51                           | -1.23                     |                     | 1.97                       | -1.29                 | 0.74                |           | 0.91           |                | -0.54          | -1.92                        | -1.26                 | 1.19                    | 1.70        | 1.37                      | Lung squamous cell carcinoma |                     |
| 2             | 26                   | 0.68          | -1.25          | 2.71         | -0.52       | 0.59              | 0.78                              | -1.15                           |                           | -2.04               | 0.68                       | 0.21                  | 0.50                |           | 1.01           | -0.46          | -1.35          | -1.70                        | -0.45                 | 1.47                    |             | Lung large cell carcinoma |                              |                     |
| 1             | 19                   | -0.54         | -1.35          | 0.75         | 0.84        | 0.81              | -1.15                             | 1.86                            | -1.29                     | -1.32               | 1.01                       |                       | 0.46                |           | 1.19           |                |                | -0.71                        | 0.96                  |                         | 2.12        | Melanoma primary          |                              |                     |
| 2             | 62                   | -1.96         | 0.43           | -0.39        | -0.30       | -1.54             | 1.55                              | -0.53                           | -1.54                     | -0.79               | 0.59                       | -1.14                 | 0.91                | 0.86      | 0.43           | 1.17           | -0.68          |                              | 1.07                  | 1.17                    | -0.82       | 1.15                      | Melanoma metastasis          |                     |
| 1             | 33                   | 1.19          |                | 0.29         | -1.14       |                   | -1.56                             | 1.36                            | -0.82                     |                     | 0.71                       |                       | 0.48                | -1.36     | 0.80           | -0.67          |                | 0.05                         | 1.07                  | 1.41                    | 0.55        |                           | Osteosarcoma                 |                     |
| 6             | 745                  | 1.42          | -1.50          | -1.06        | 0.27        | 2.28              | -2.69                             | 0.73                            | 1.02                      |                     | -0.42                      | 0.84                  | 1.70                | -0.53     | 0.72           | 0.49           | -0.55          | 0.31                         | 1.68                  | 1.35                    | 0.05        |                           | Ovarian cancer               |                     |

# Heterogeneity of tissues (and tumors)



Bulk RNA-seq



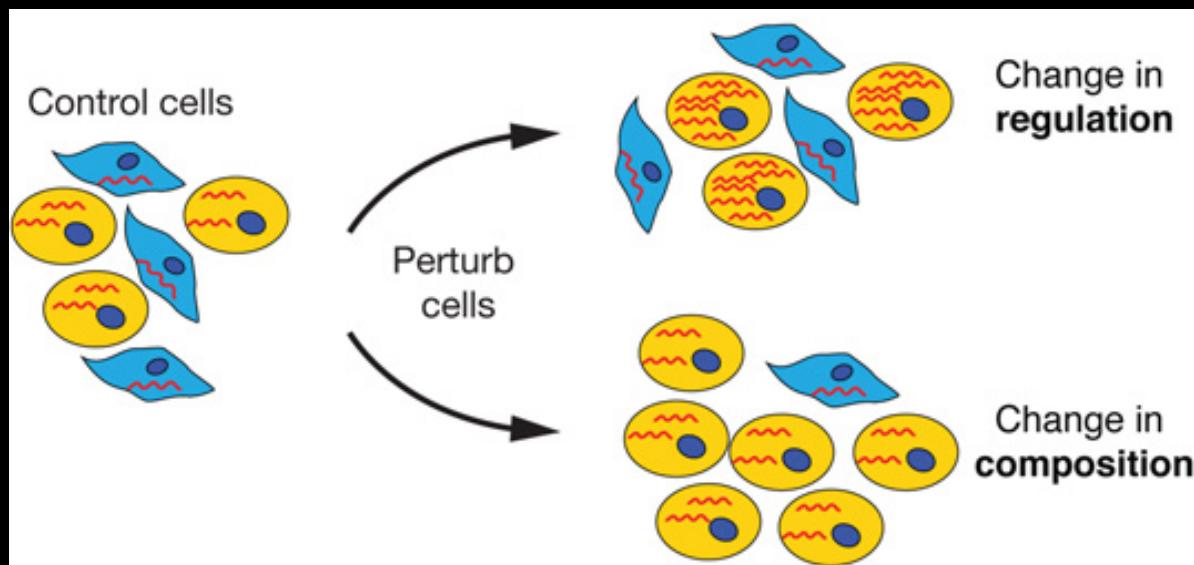
Single-cell RNA-seq



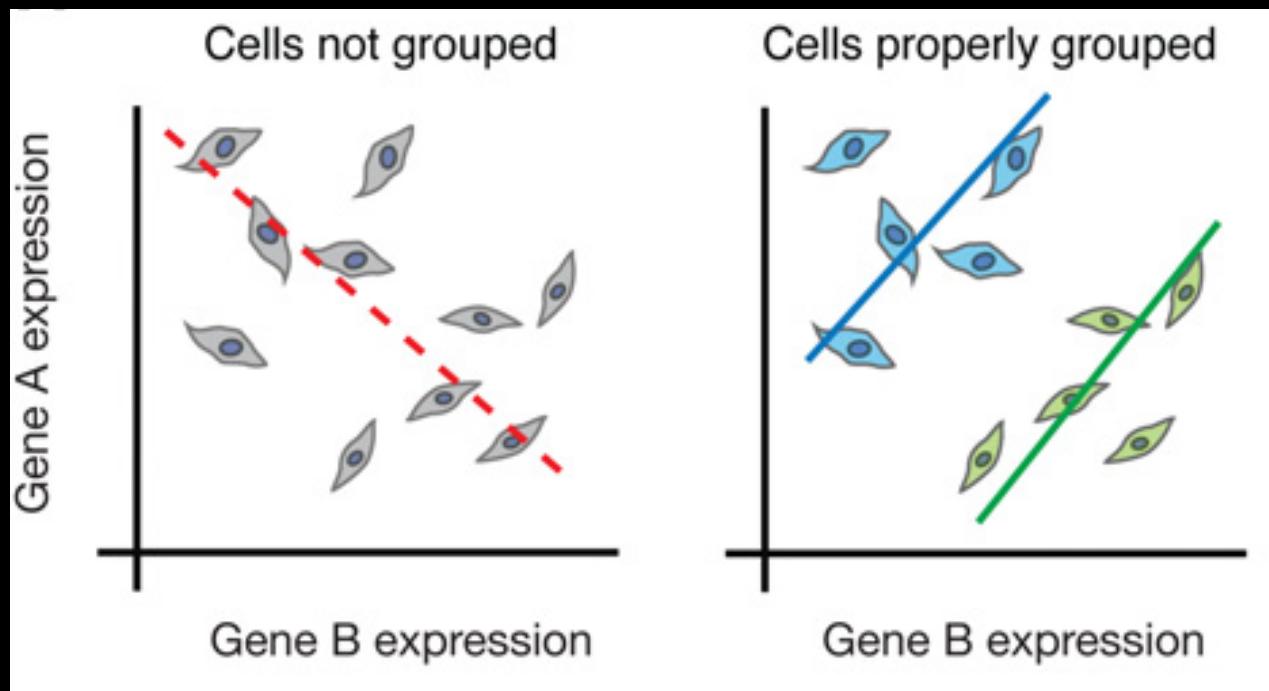
Profiling of sorted populations

- Can you tell what's in a smoothie by taste?
- How many different flavours?
- How small a proportion of the mix?
- Is it easier to detect presence or absence ?

## Averaging across cell types



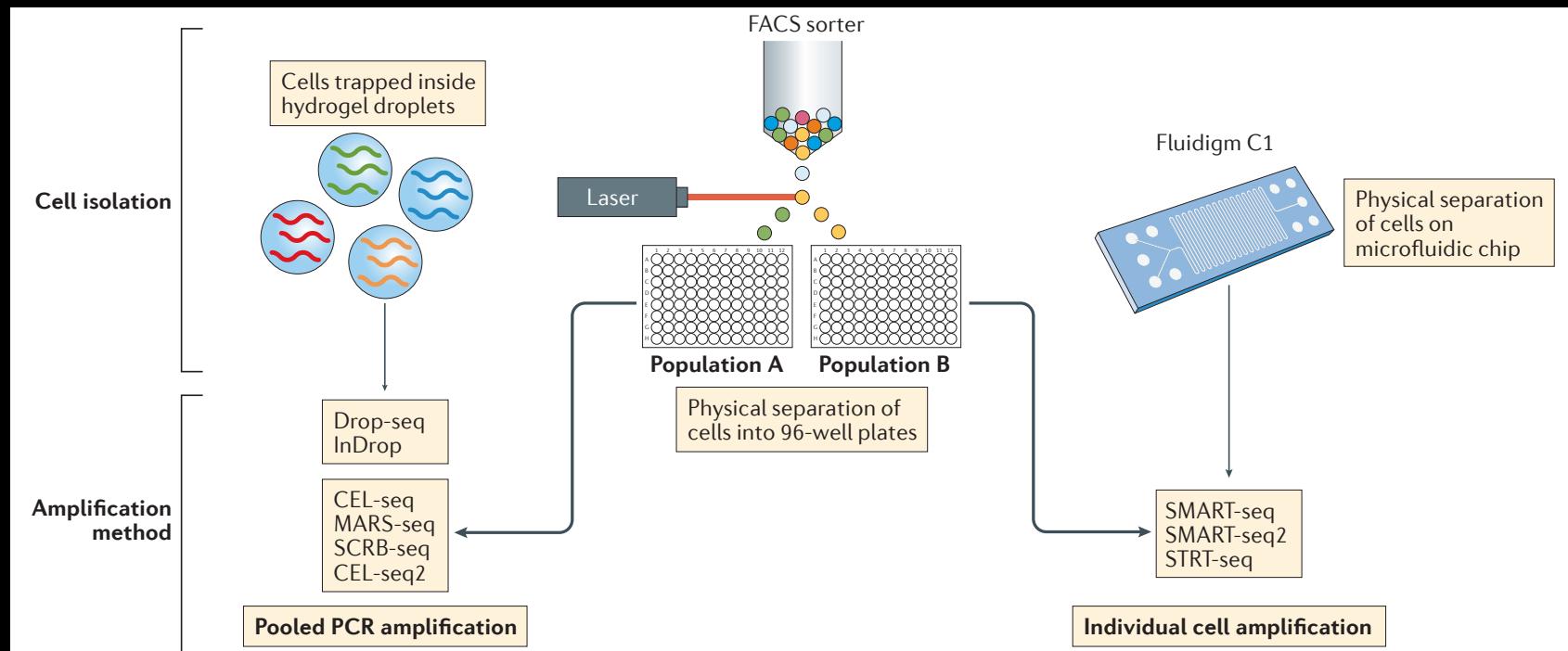
# Why single cell?



Simpson's paradox

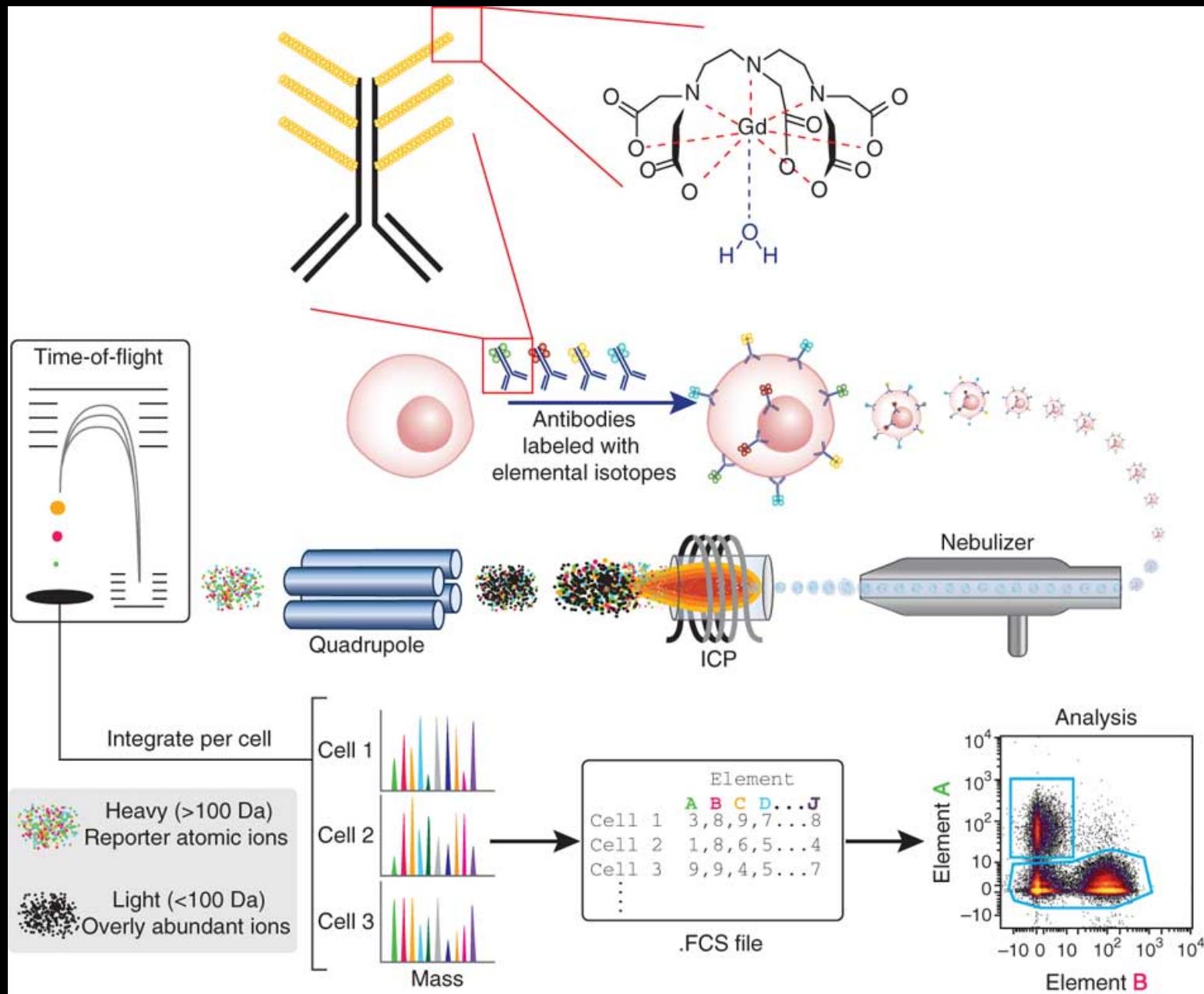
Trapnell et al. 2015 Genome Res. 25: 1491-1498

# Single cell RNA-seq approaches



Papalex & Satija Nat Rev Imm 2018

# Mass cytometry – single cell proteomics



# Single cell RNA-seq vs CyTOF

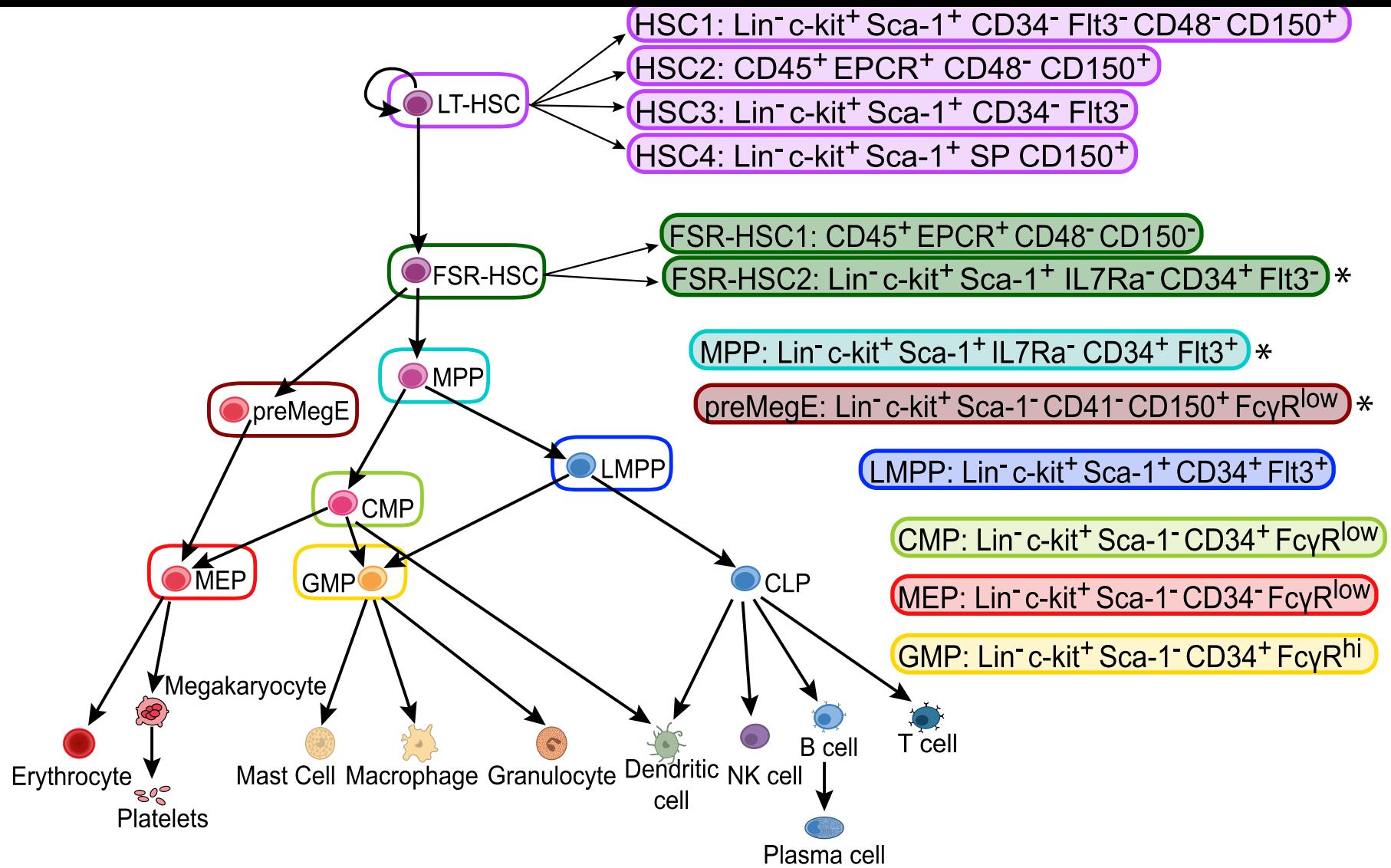
- Advantages and disadvantages to both
- scRNA-seq
  - 1000s of genes X 1000s of cells
  - Very sparse (up to 90% missing)
- CyTOF
  - ~40 proteins X up to millions of cells
  - “complete” matrix
  - Potential for spatial info (CODEX)

- Why would you choose RNA-seq vs CyTOF (or among different platforms)

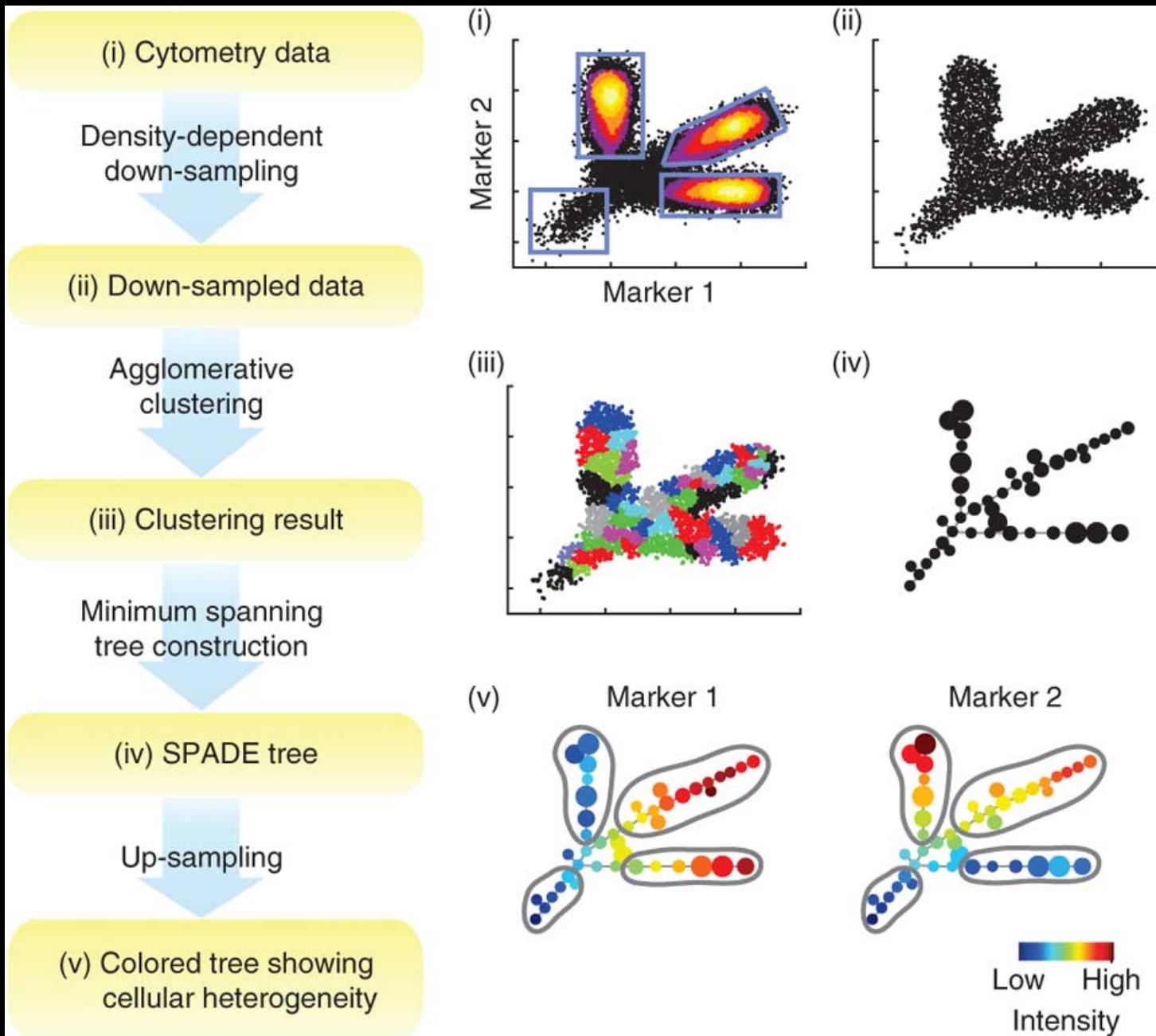
Hint: breadth vs depth

Which of these might be useful in the projects you are involved with this summer?

# Tissues form hierarchies

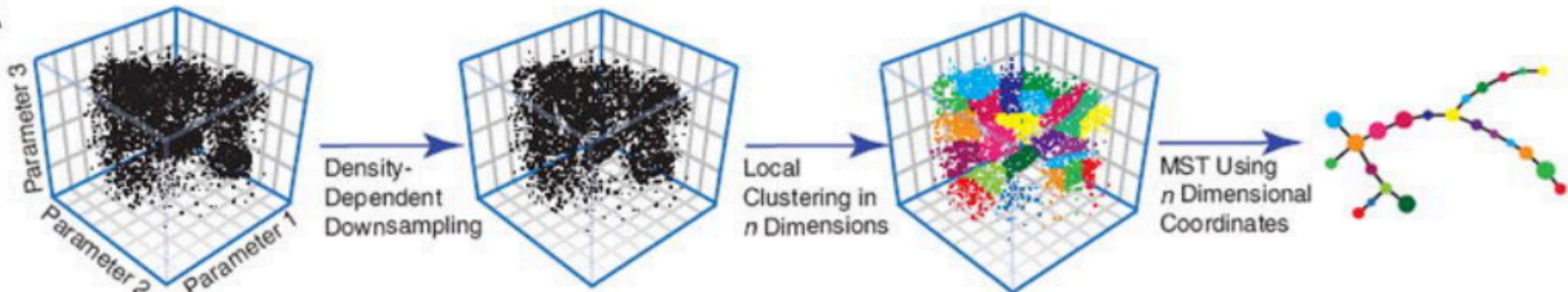


# SPADE analysis

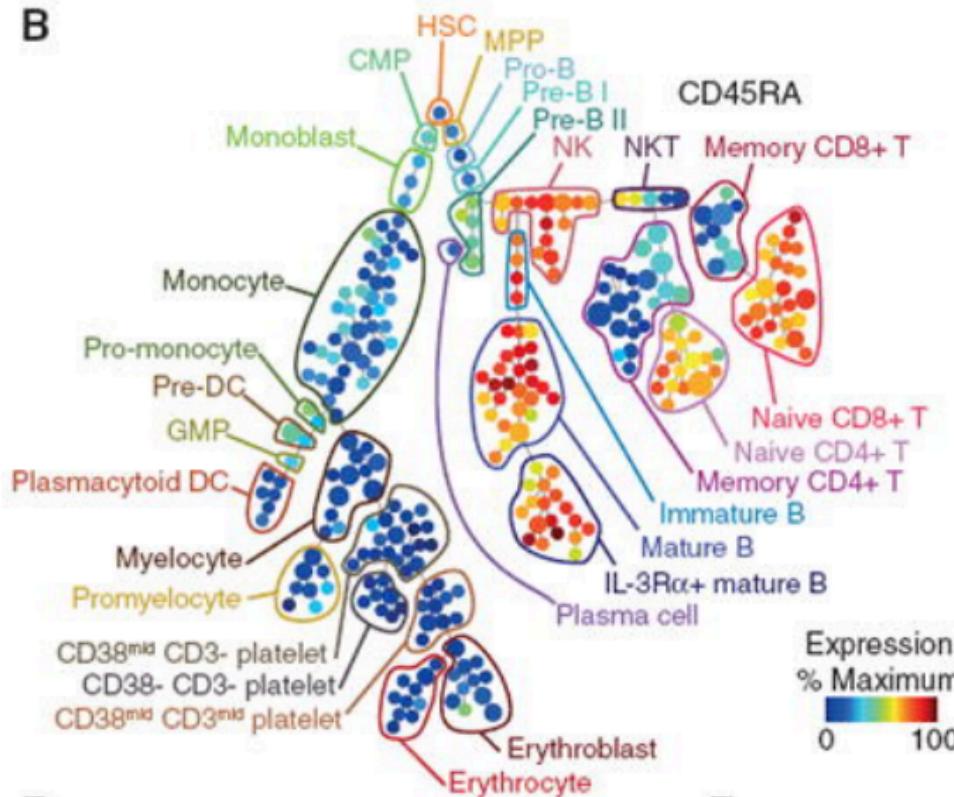


# SPADE analysis of single-cell blood data

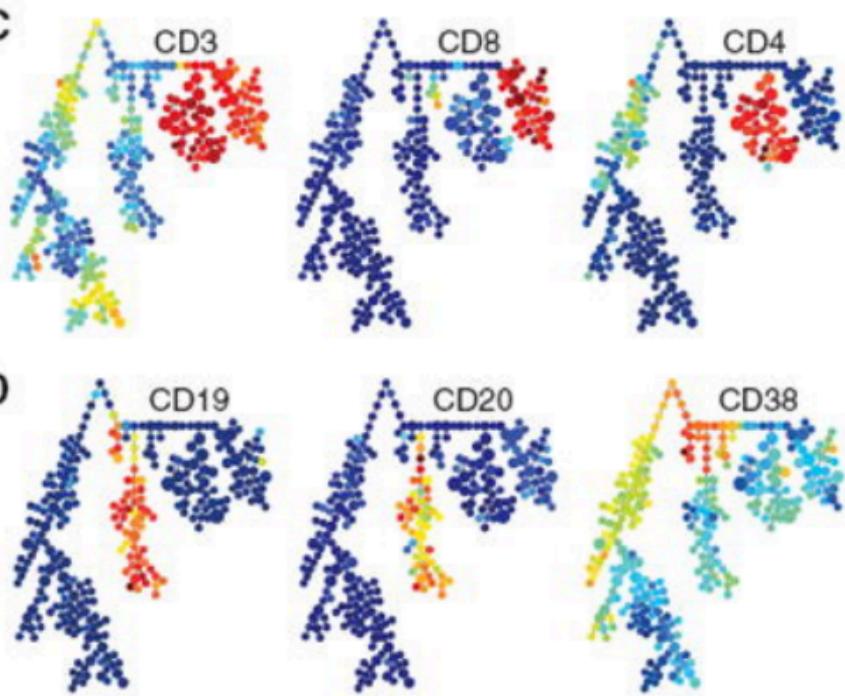
A



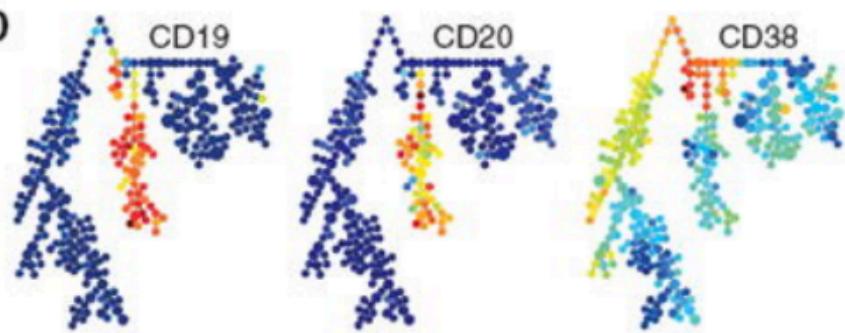
B



C



D



## Summing up

- Many high-throughput technologies generating huge amounts of data
- Skew towards RNA/DNA, because of sequence complementarity
- Key aims
  - Reconstruct networks underlying/driving disease
  - Identify biomarkers for prognosis, treatment etc
- Computational and statistical methods are key

## PS-ON cell line datasets

- Gene expression from various cell lines from various cancer types (breast, colon, etc)
- Physical measures on them:
  - Speed they move
  - Distance they move in certain time
  - Measured on various types of surface/medium

*More on the PS-ON resource here:*

<https://physics.cancer.gov/bioresources/>

## Activity for next time

- Inspect gene expression matrix and sample info for cell lines
- Connect the two together based on common identifiers
- Look at relationship between expression of genes and how fast cells can move
- Identify genes that might be related to metastasis