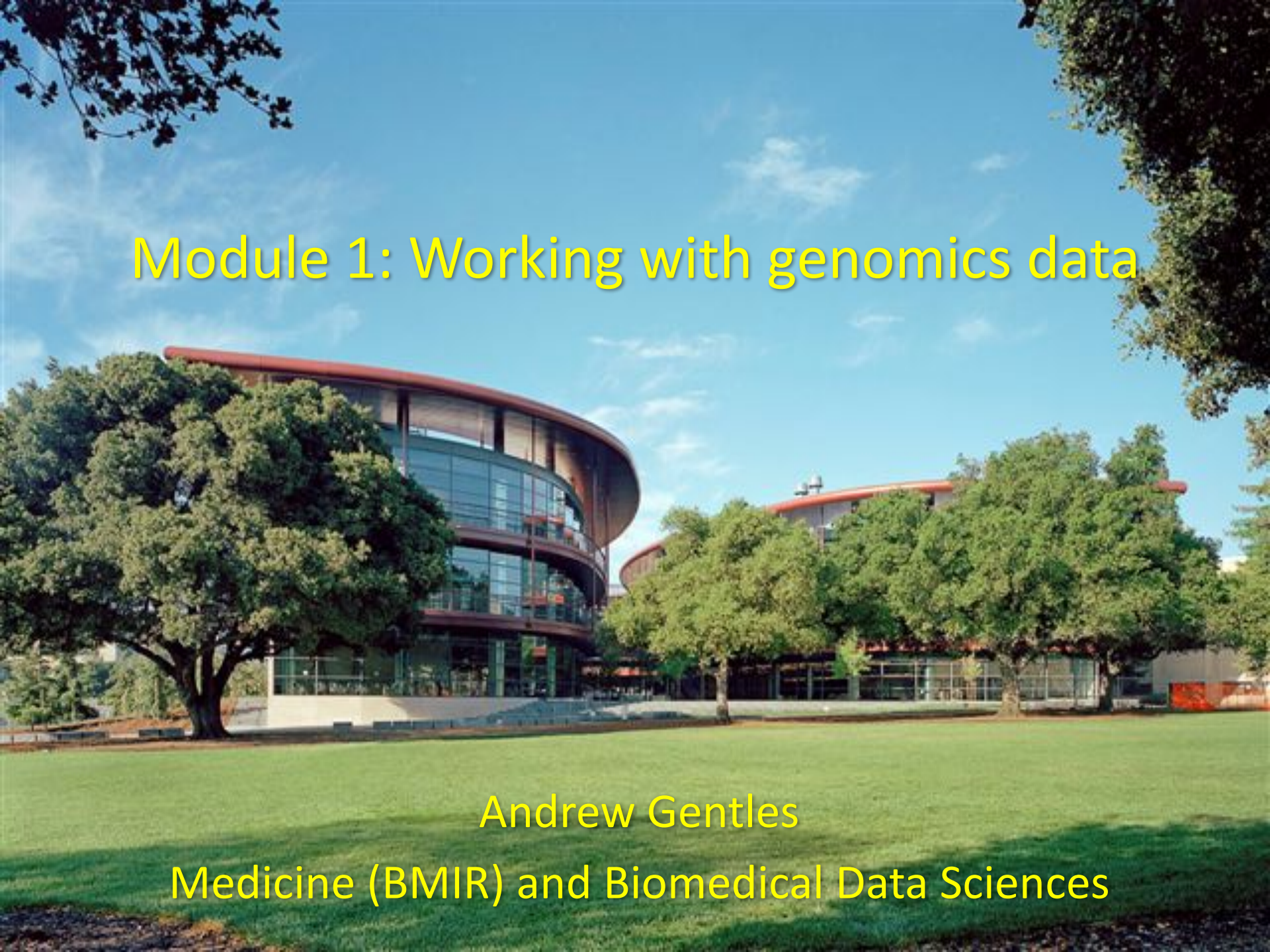


Module 1: Working with genomics data

Andrew Gentles

Medicine (BMIR) and Biomedical Data Sciences

Some selected topics in systems biology

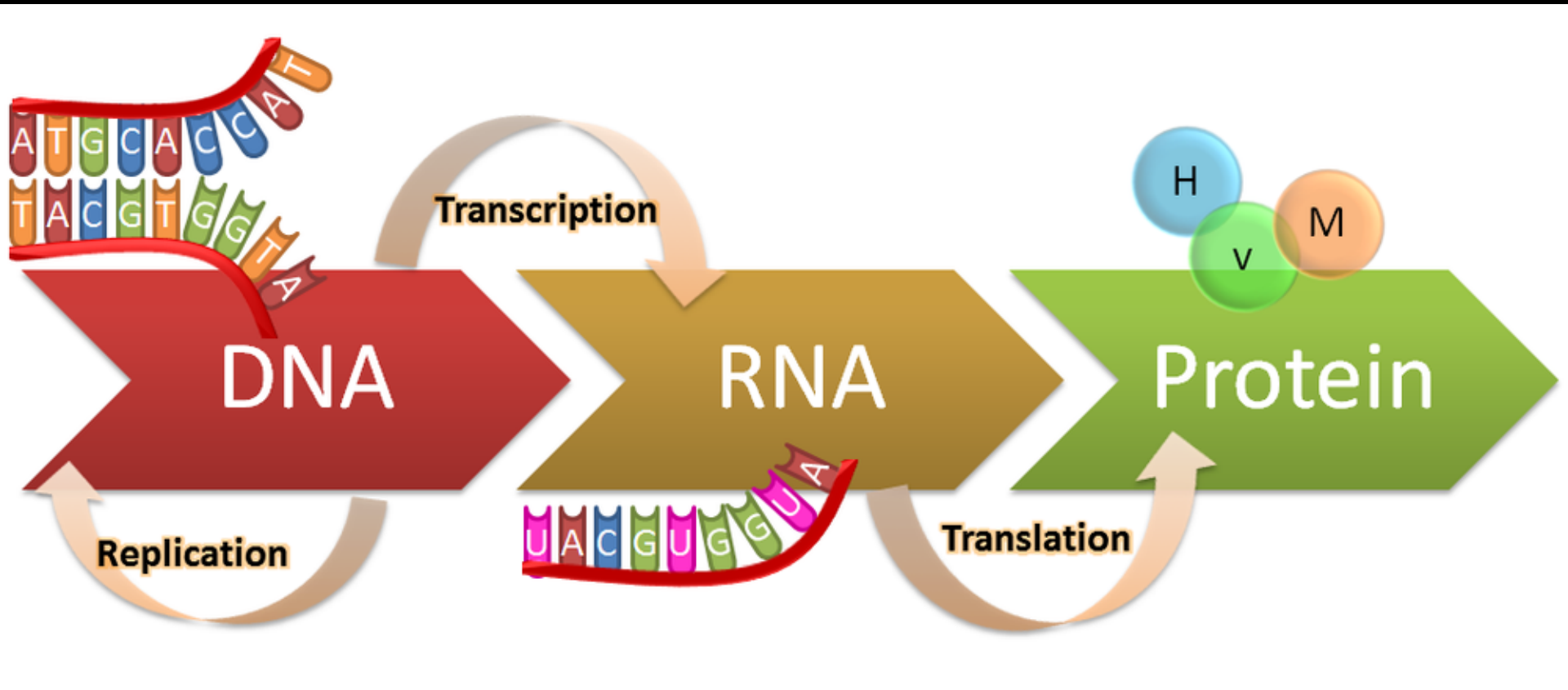
- Brief review of genomics data types
- Reconstructing regulatory networks
  - Transcriptional networks
- Connecting genomics with outcomes (survival)
- Dissecting tissues (e.g. tumors) at single cell resolution

Cancer systems biology – one definition

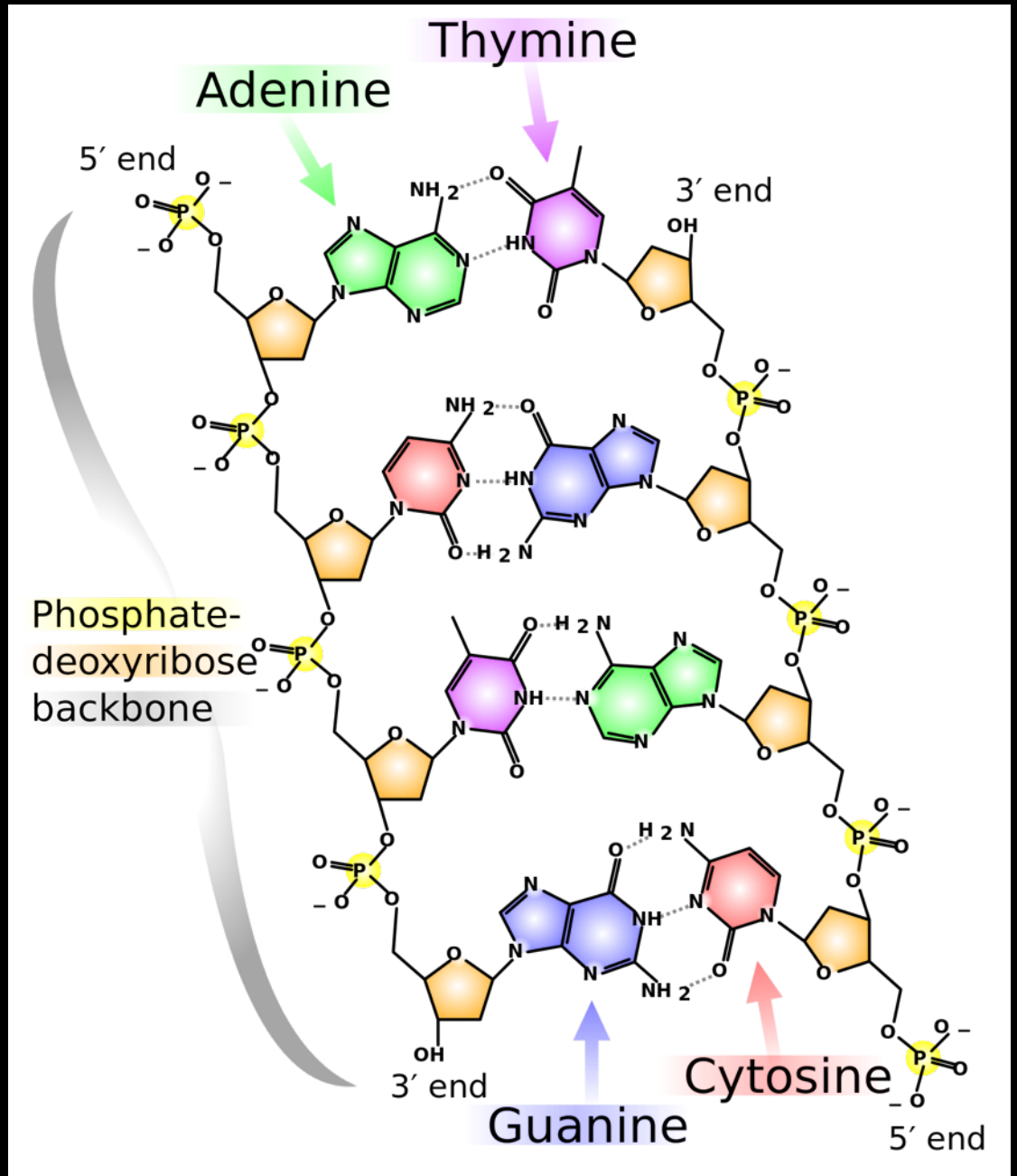
- Cancer systems biology adopts a holistic view of cancer aimed at integrating its many biological scales, including genetics, signaling networks, epigenetics, cellular behavior, histology, (pre)clinical manifestations and epidemiology.

https://en.wikipedia.org/wiki/Cancer_systems_biology

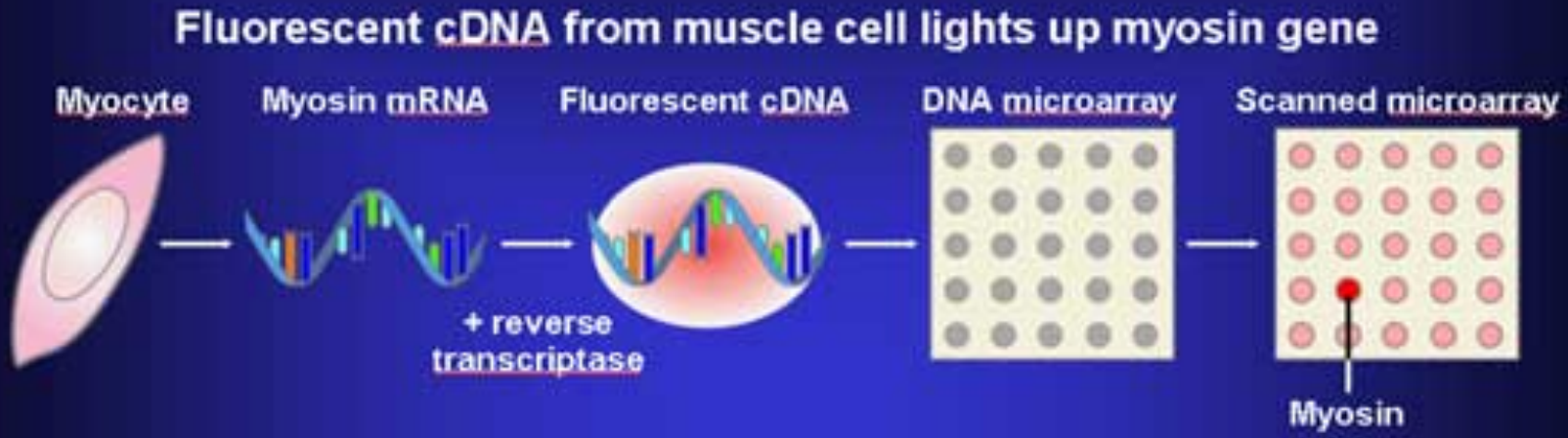
Central dogma of molecular biology

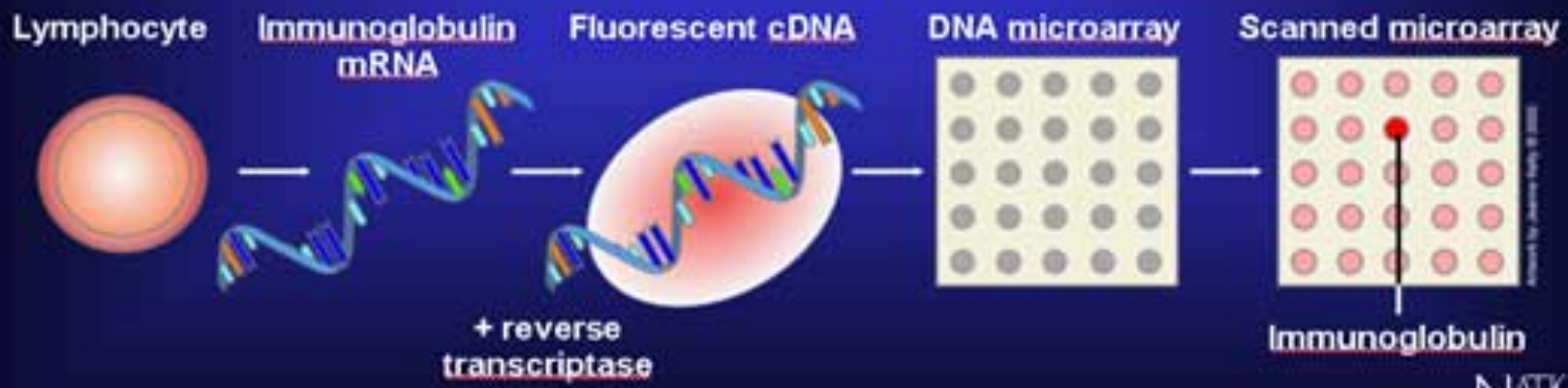


DNA and RNA complementarity



Using DNA Microarrays to Measure Gene Expression

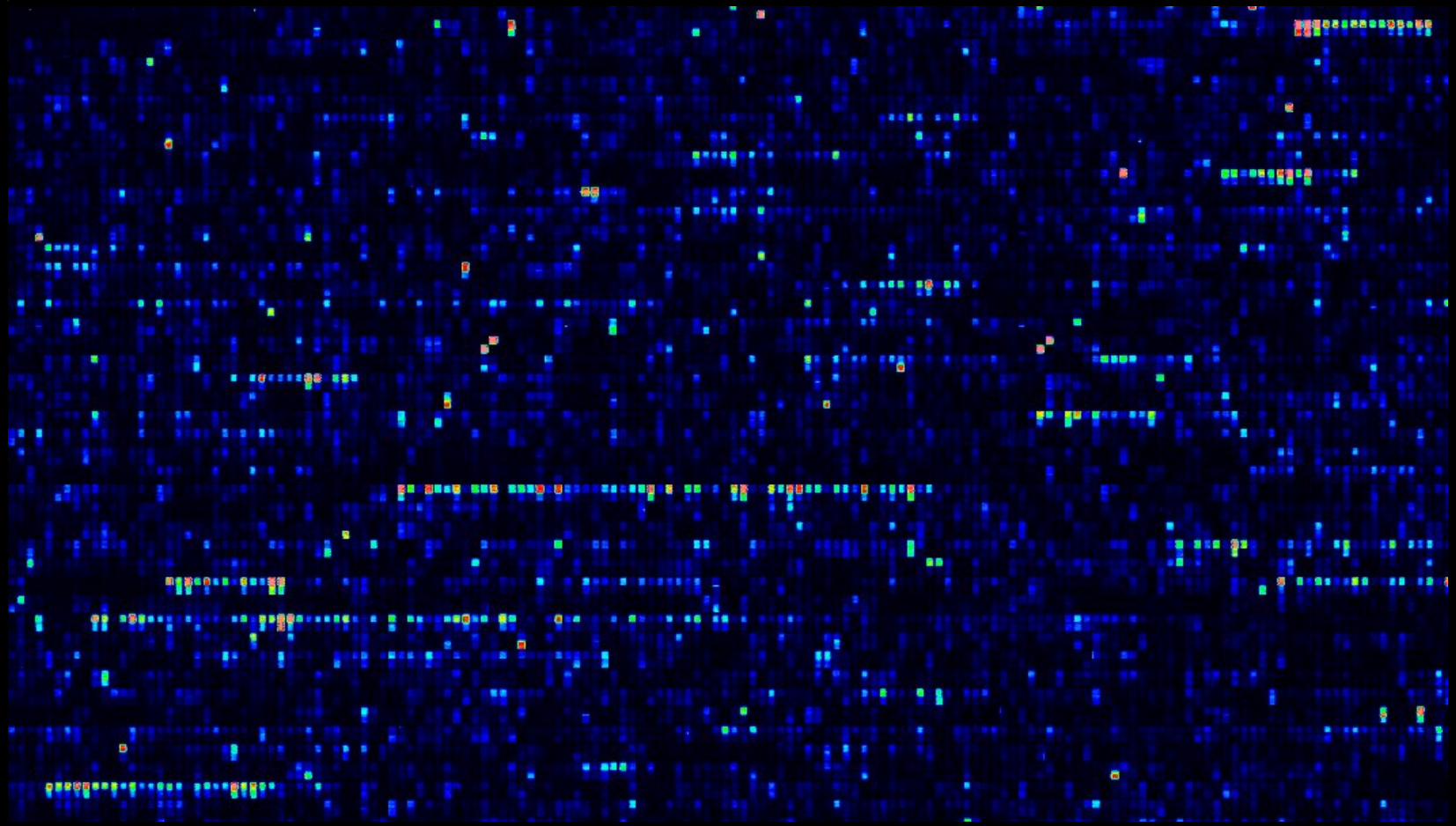




NATIONAL CANCER INSTITUTE

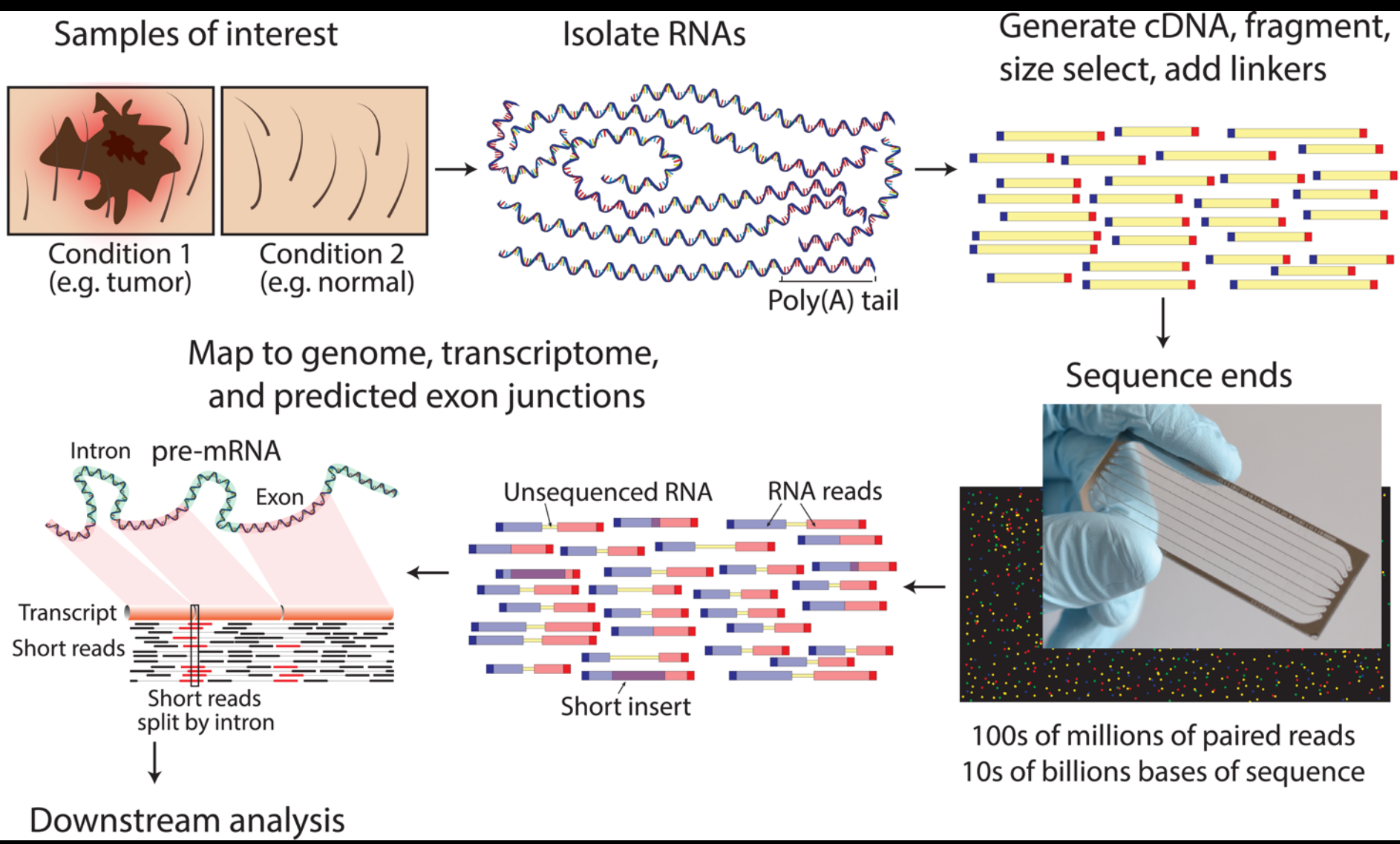
We can assess activity of thousands of genes in a sample

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



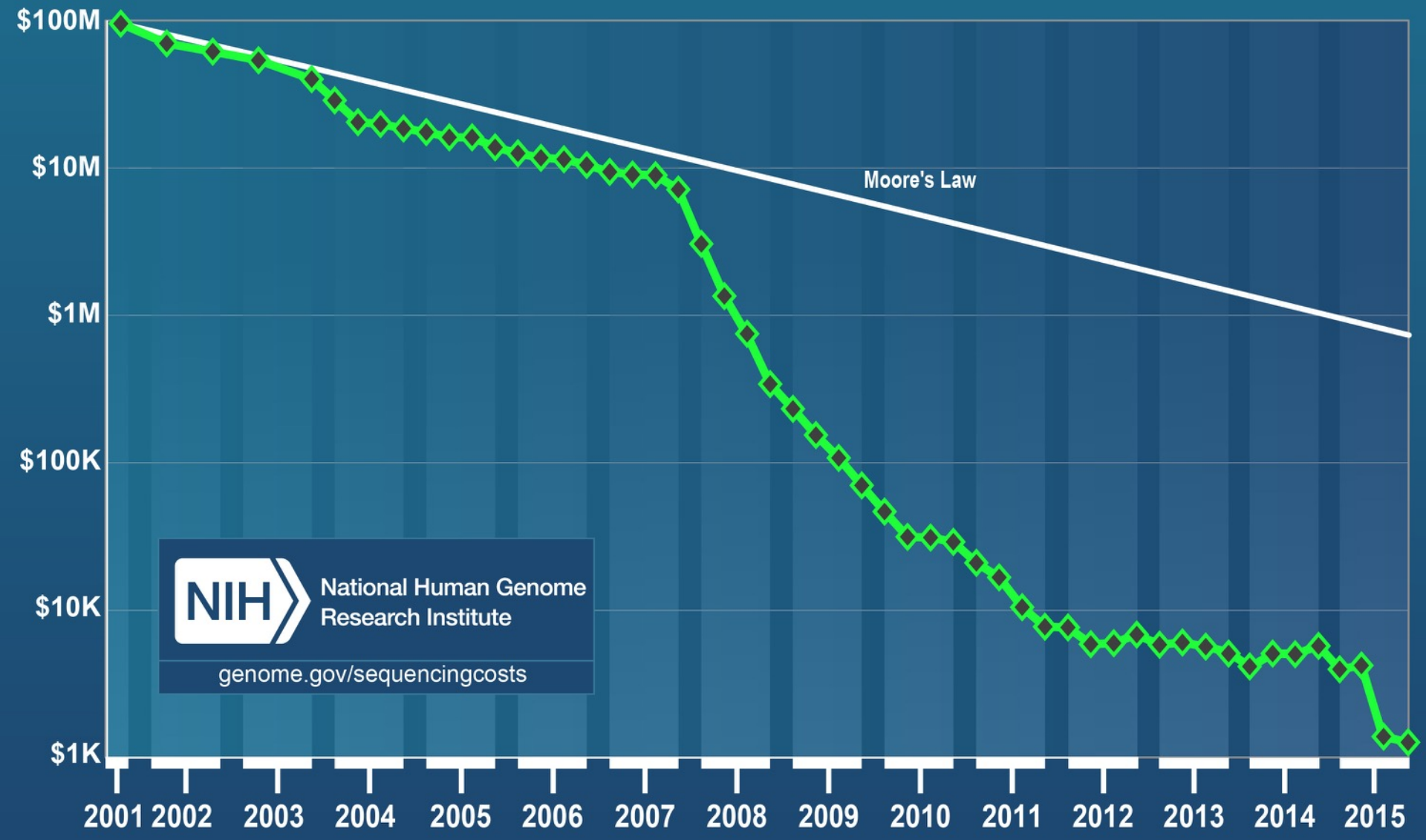
http://learn.genetics.utah.edu/content/labs/microarray/

Next generation sequencing



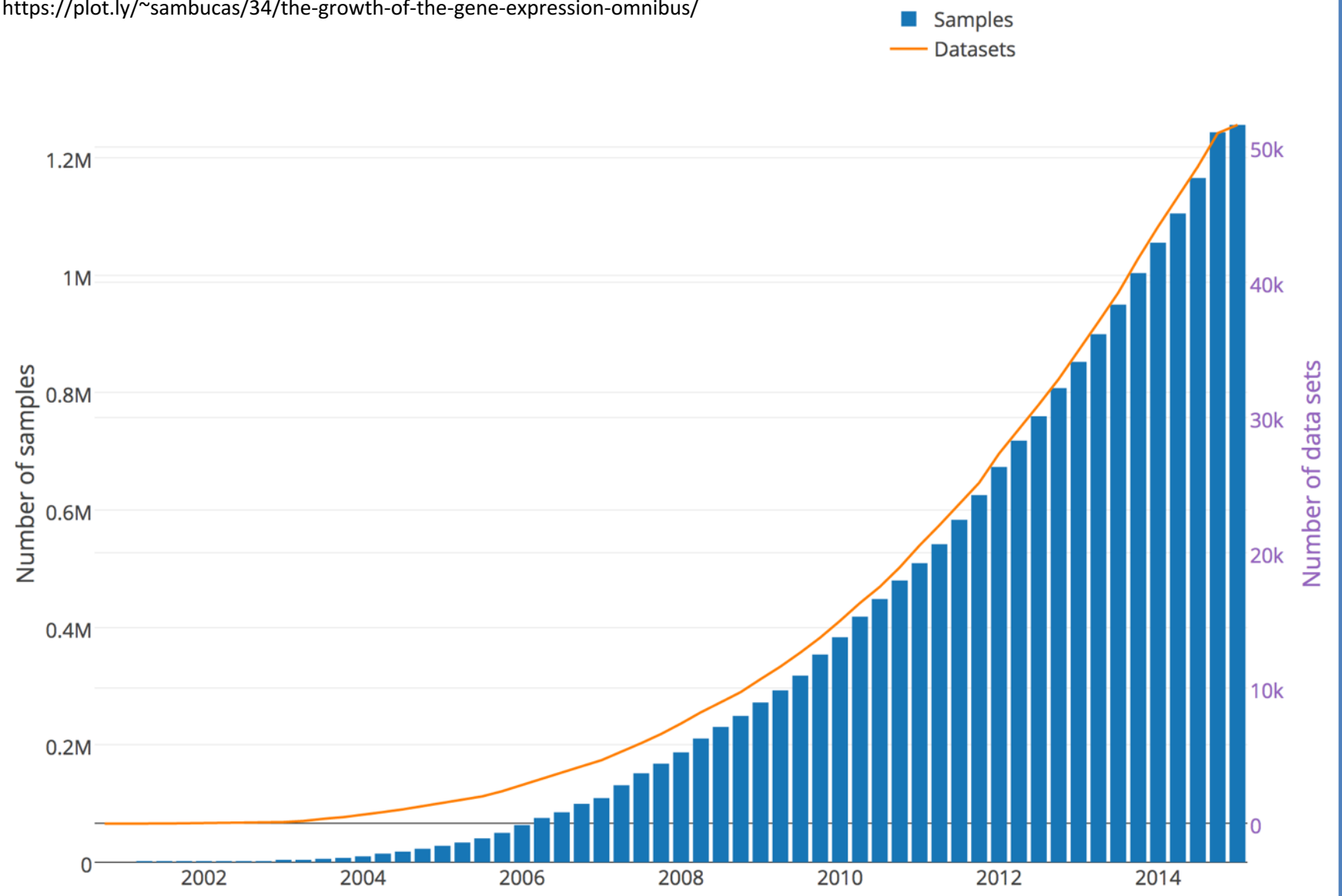
https://en.wikipedia.org/wiki/RNA-Seq

Cost per Genome

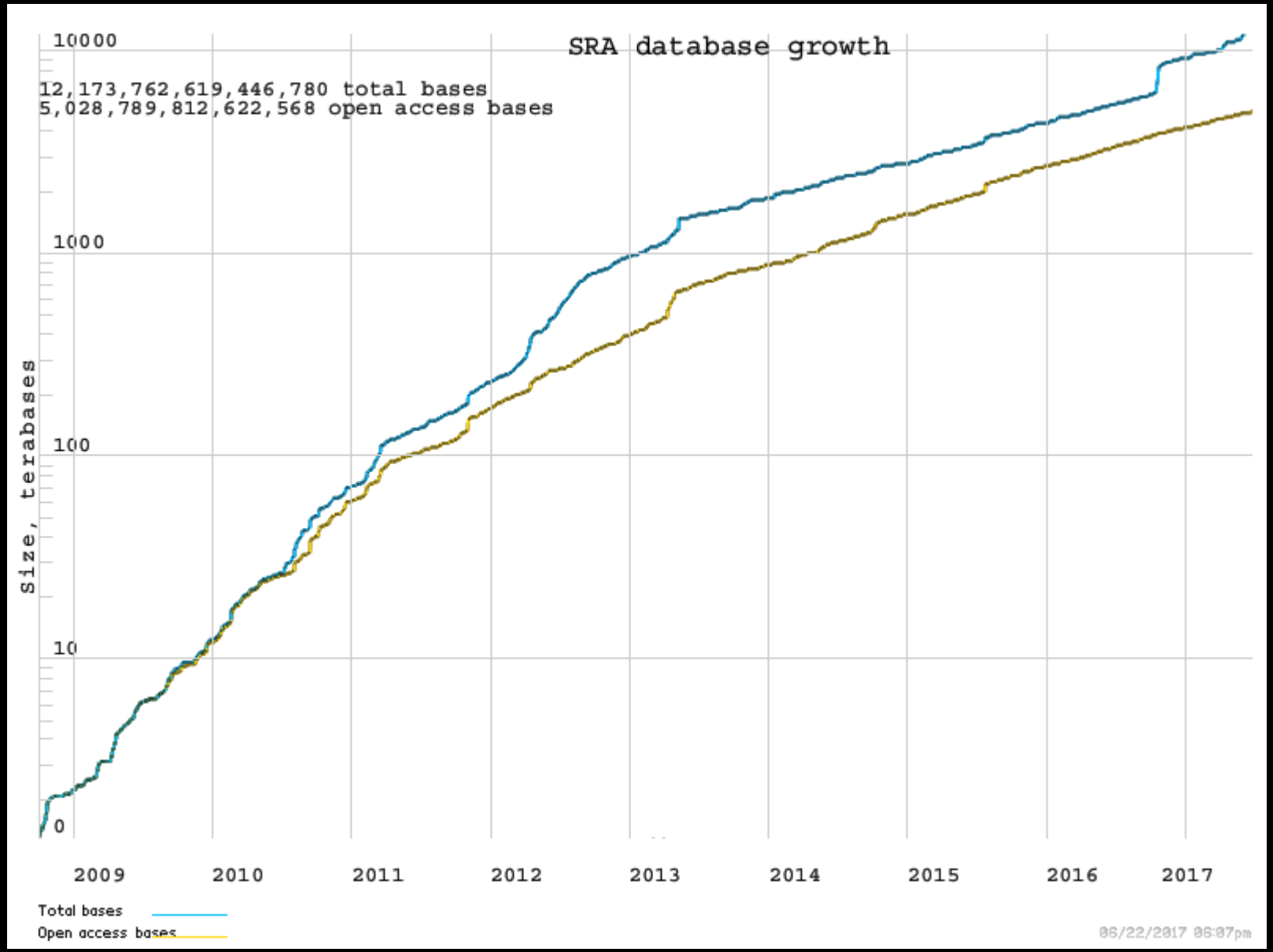


Growth of gene expression omnibus at NCBI

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



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

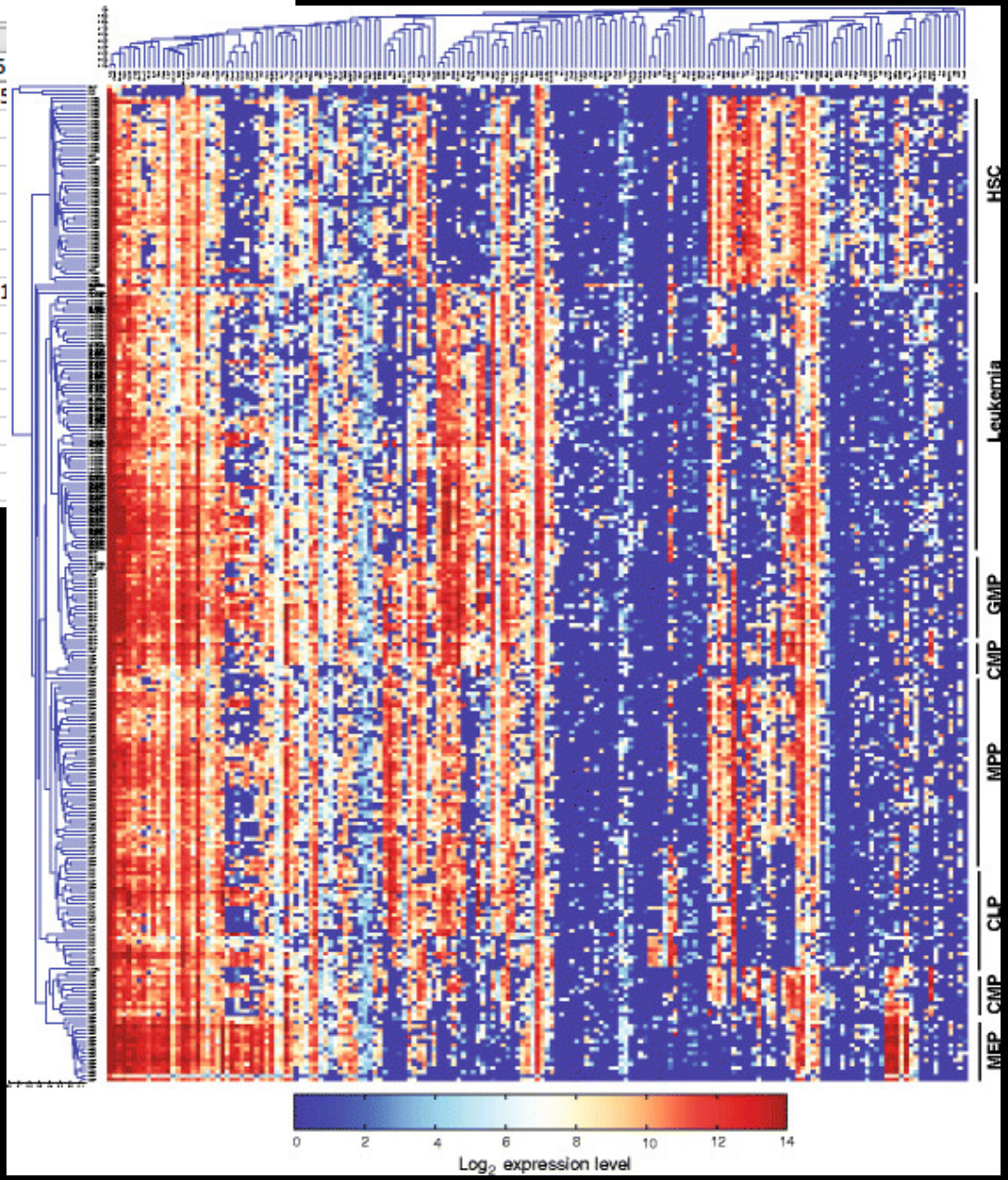
Endpoint is moderately large scale data

J15

	A	B	C	D	E	
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	
4	NM_028089	190.41	211.06	236.19	9.05	
5	NM_016696	66.77	57.56	101.09	750.9	
6	NM_013459	3.3	11.29	1.89	735.82	
7	NM_007809	45.34	36.12	51.02	245.27	
8	NM_009999	103.04	370.21	200.29	17.09	
9	NM_133960	7708.78	6976.38	6569.04	1731	1
10	NM_027881	31.32	10.16	24.56	268.39	
11	NM_054053	31.32	24.83	19.84	323.68	
12	NM_007377	47.81	89.17	70.86	370.93	
13	NM_028064	703.95	689.62	662.29	214.11	
14	NM_008182	222.56	339.73	226.75	30.16	
15	NM_013661	12.36	11.29	8.5	97.51	
16	NM_007815	20613.09	25218.13	31540.46	5209.07	

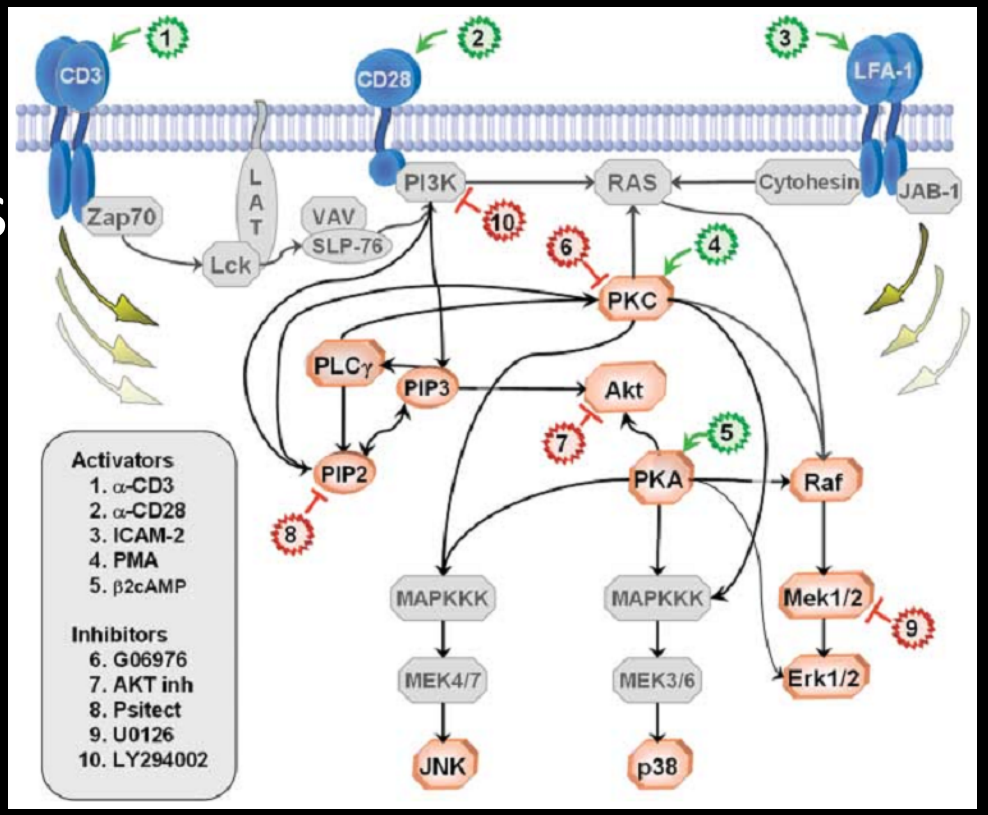
10-1000's columns (samples)

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



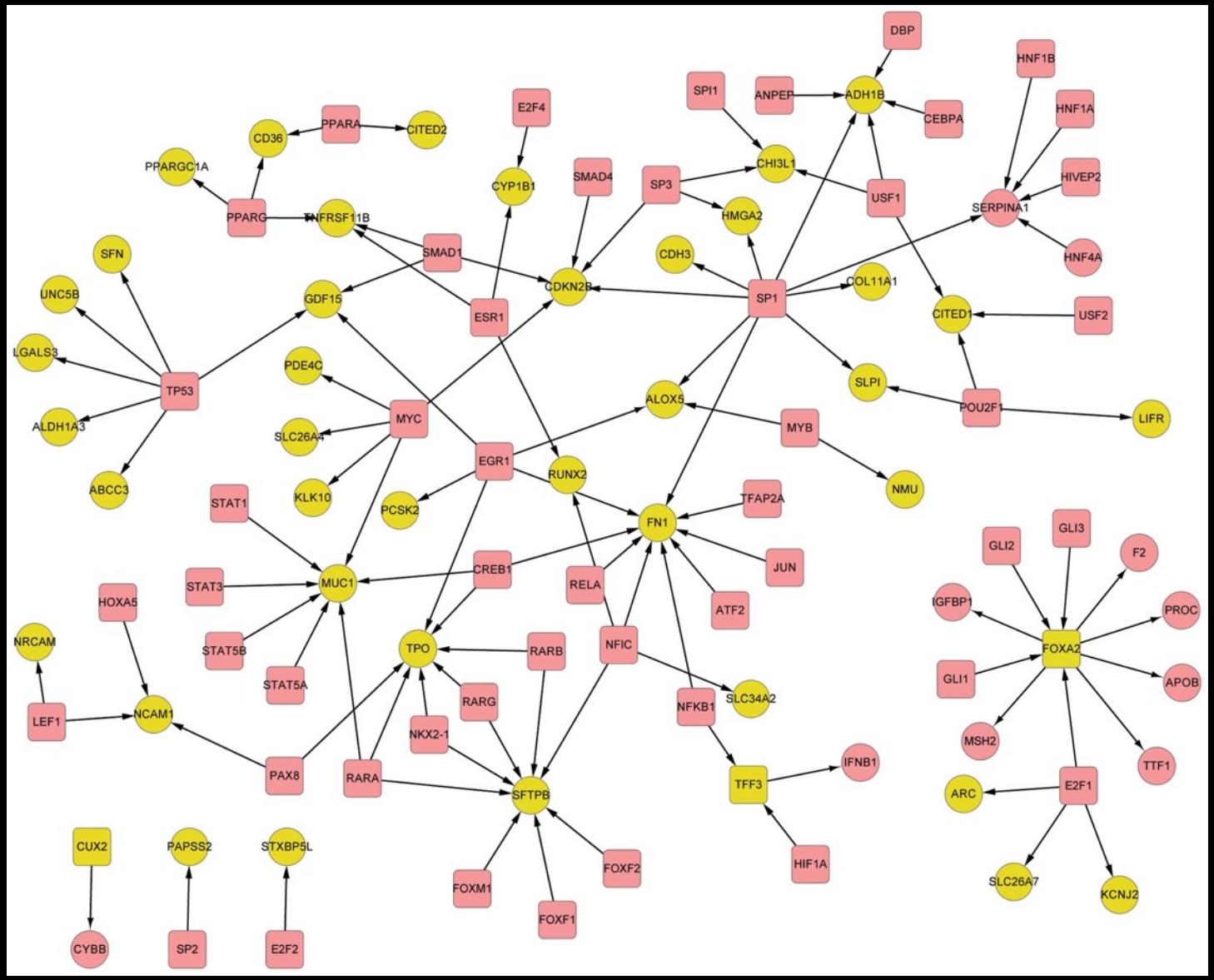
Network types

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

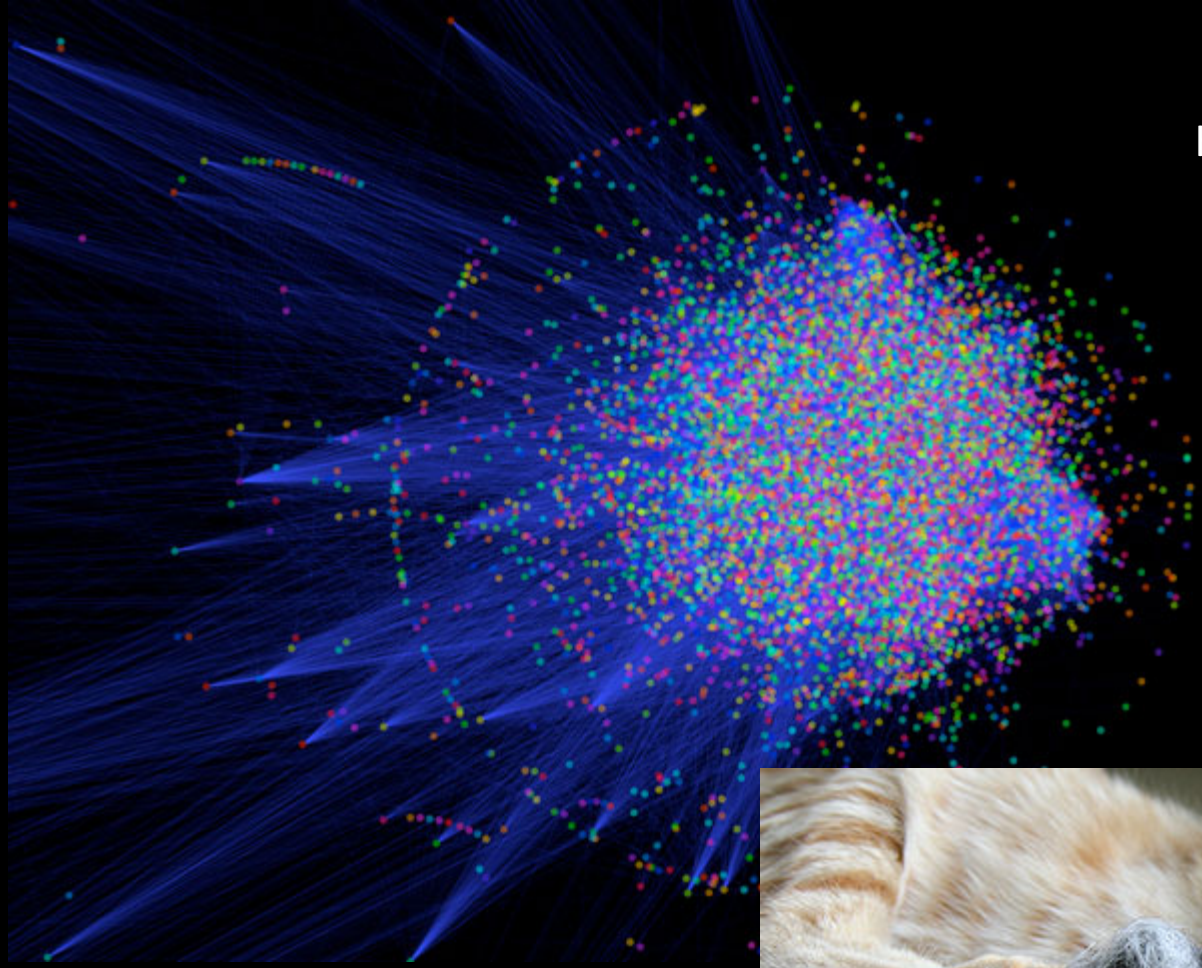


Sachs et al. http://www.sciencemag.org/content/308/5721/523.full

Gene regulatory networks



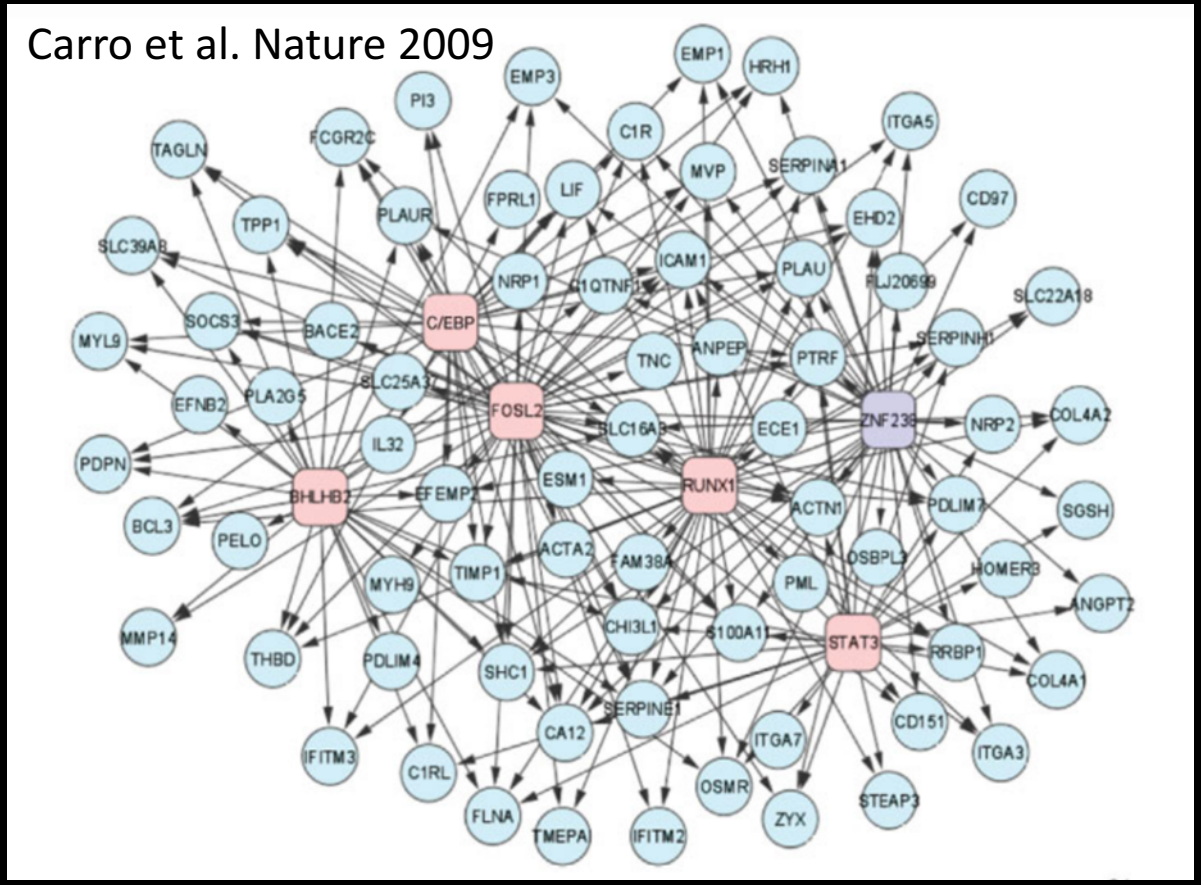
They are a “flowchart” of interactions, not a detailed model



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

- 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)

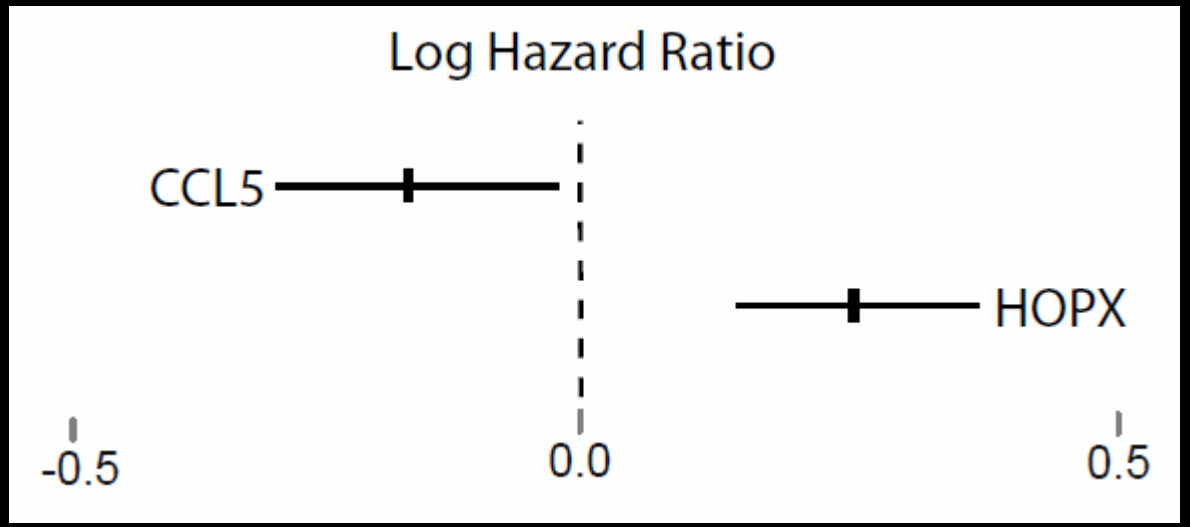
Extra detail:

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

X_i = expression of gene *i* at sample collection

Gene expression ~ survival

Z=-2.2
p=0.02

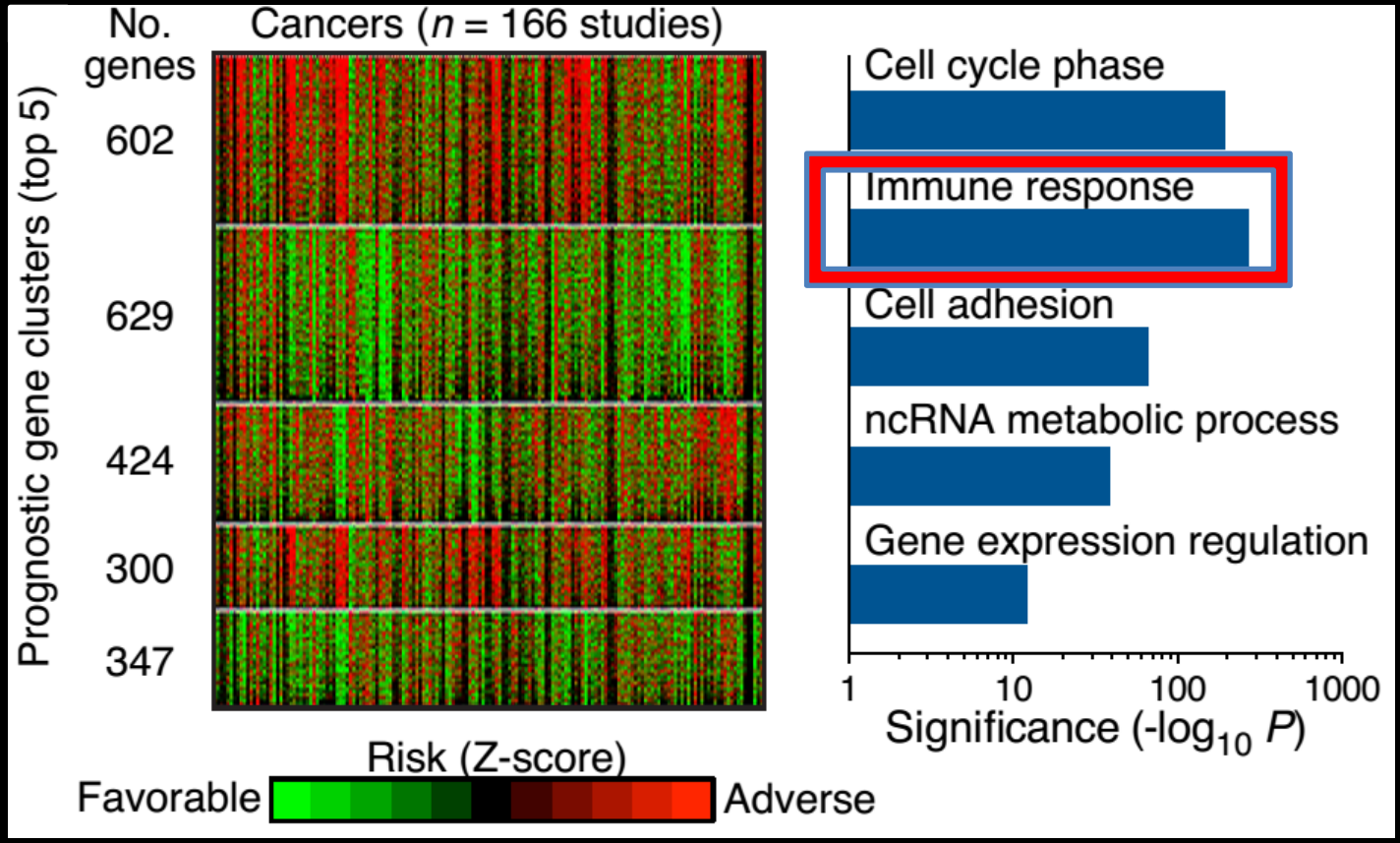


Z=+4.5
p=4e-6

"Good" genes

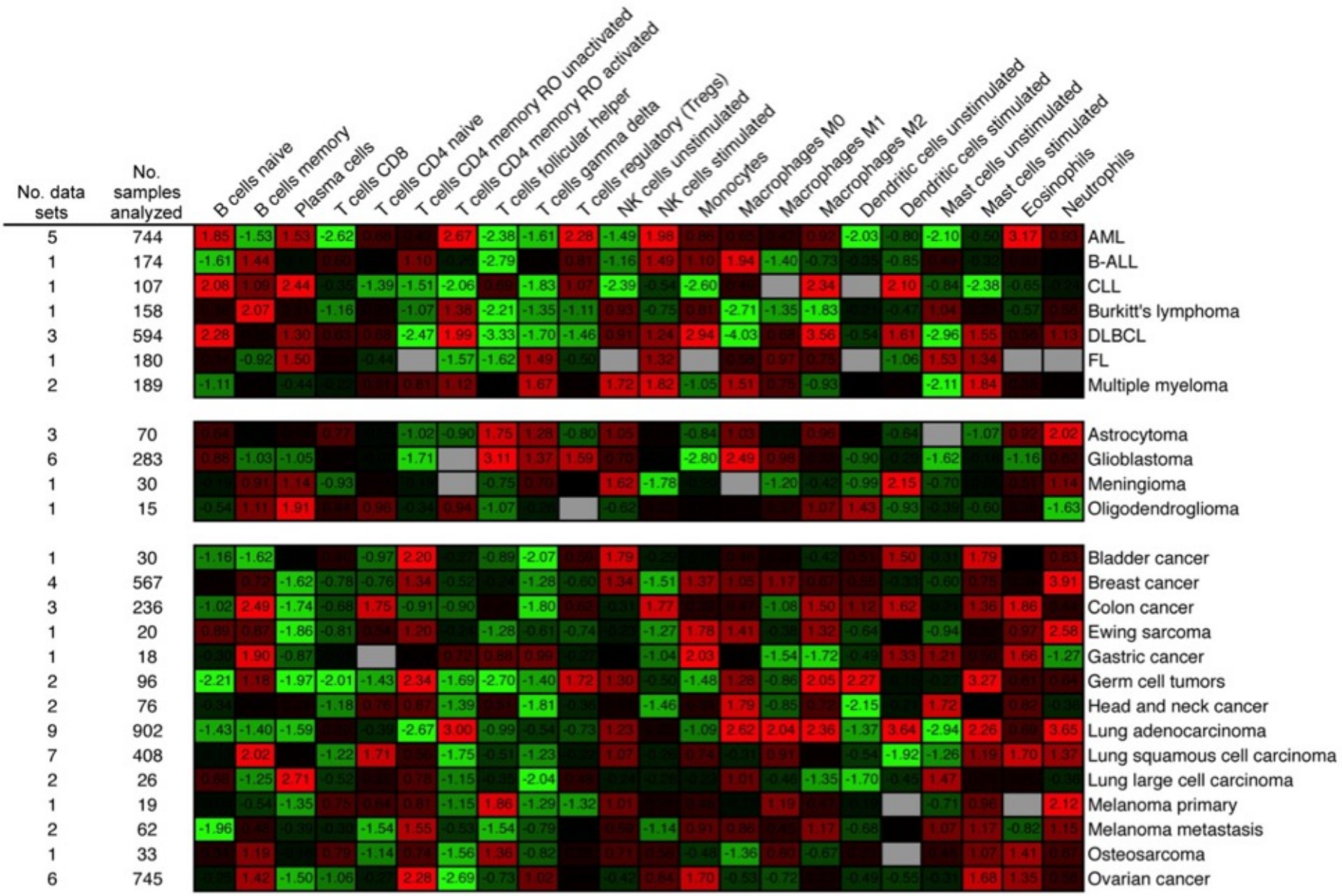
"Bad" genes

Clustering of outcome matrix -> biological processes



http://precog.stanford.edu

Prognostic influence of immune infiltrates



Heterogeneity of tissues (and tumors)

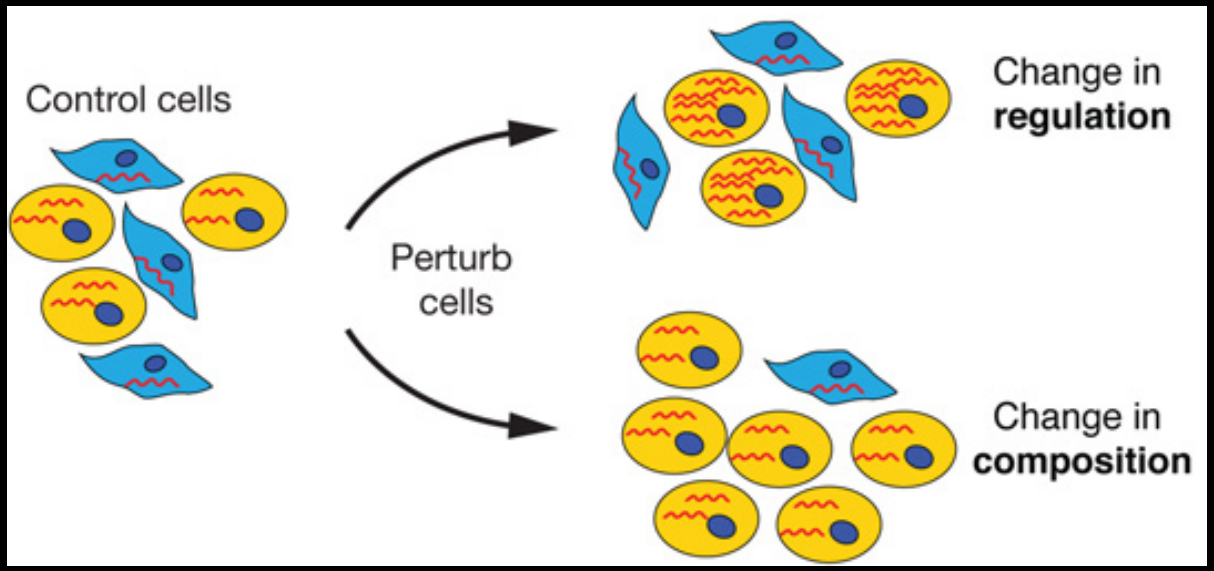
Bulk RNA-seq

Single-cell RNA-seq

Profiling of sorted populations

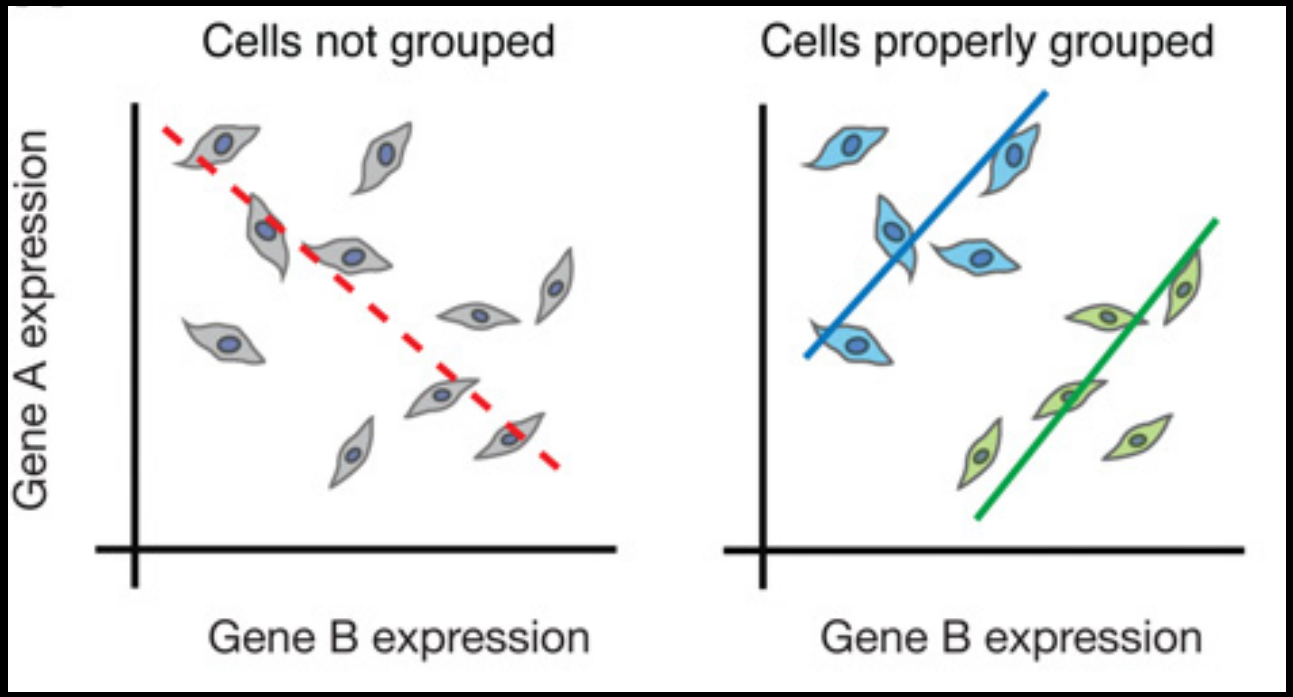
- Can you tell what’s in a smoothie by taste?
- How many different flavours?
- How small a proportion of the mix?
- Is it easier to detect presence or absence ?

Averaging across cell types



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

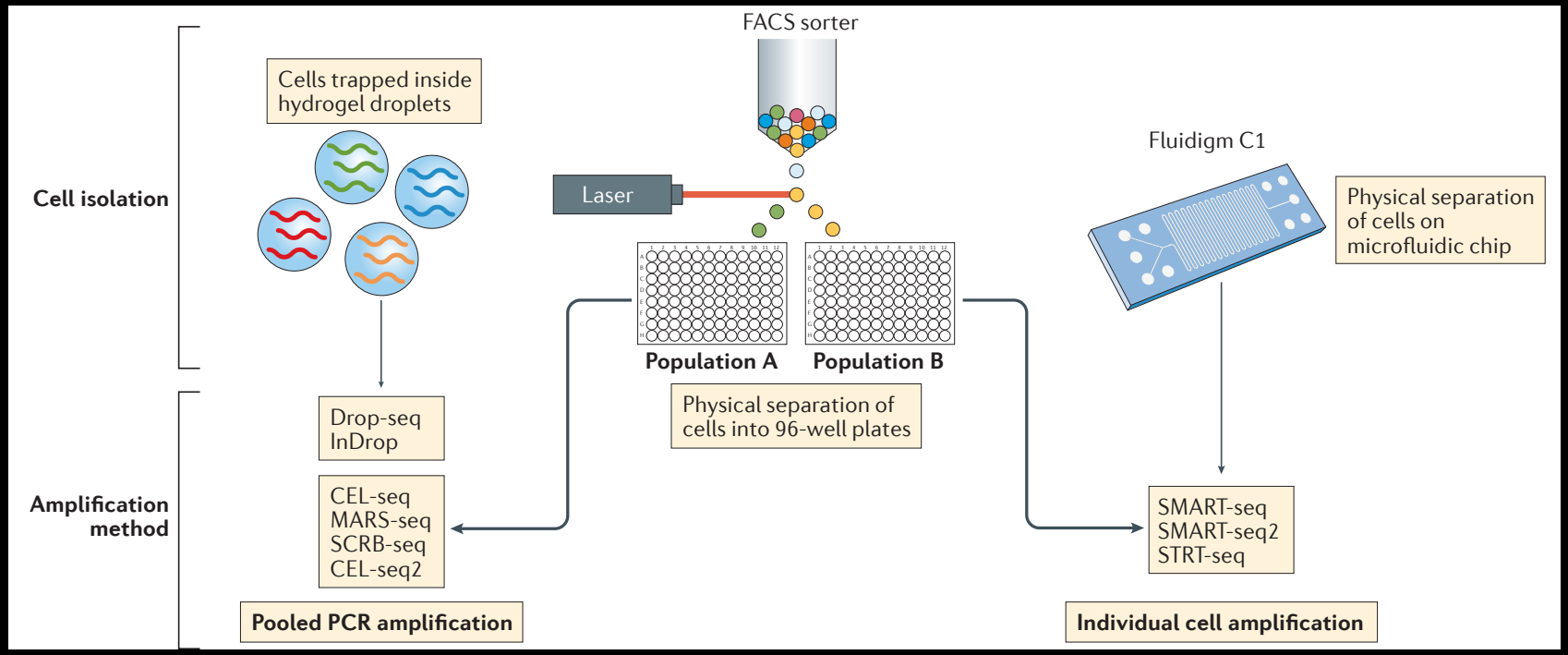
Why single cell?



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

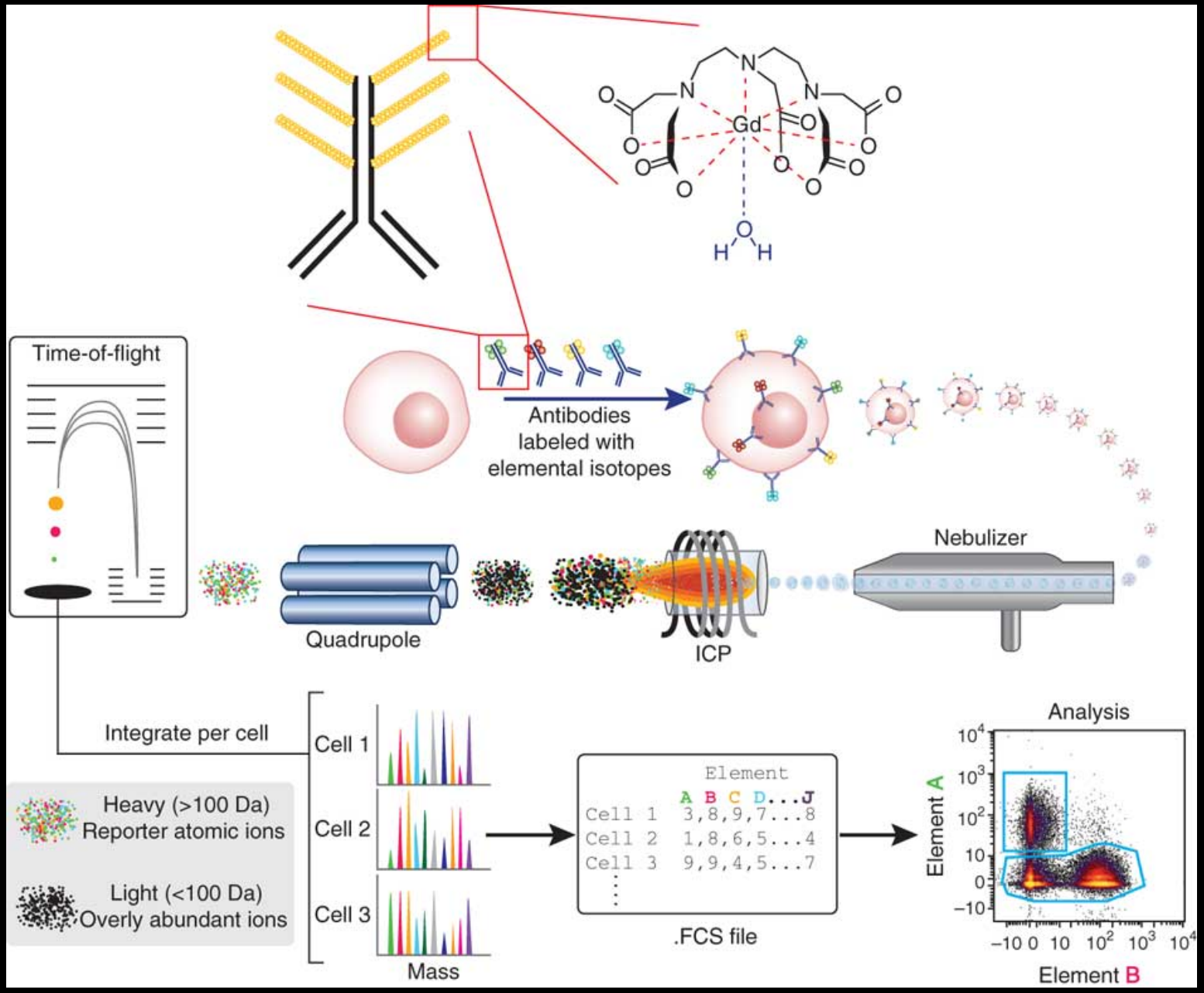
Simpson's paradox

Single cell RNA-seq approaches



Papalexi & Satuja 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)

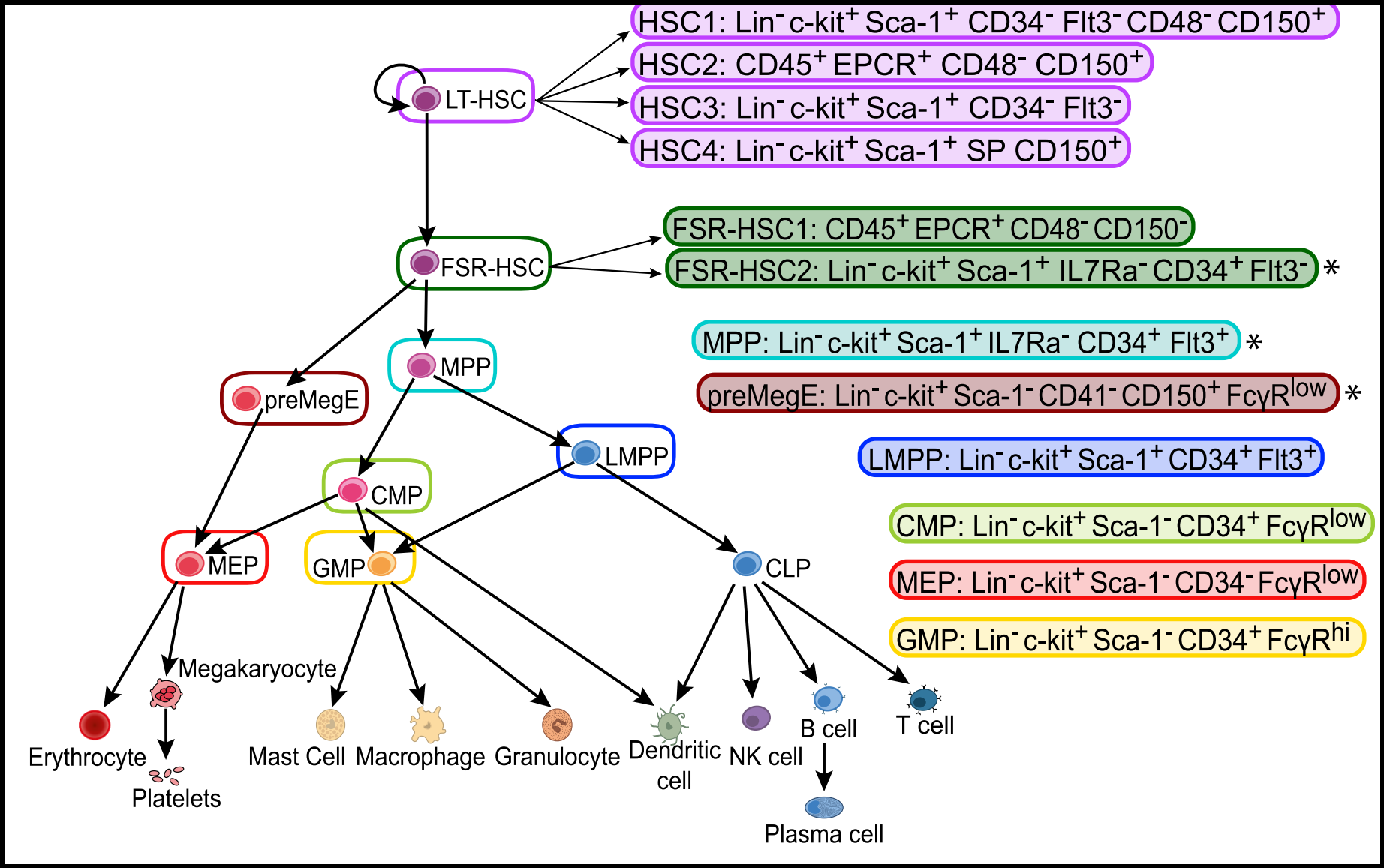
- CyTOF
  - ~40 proteins X up to millions of cells
  - “complete” matrix
  - Potential for spatial info (CODEX)

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

Hint: breadth vs depth

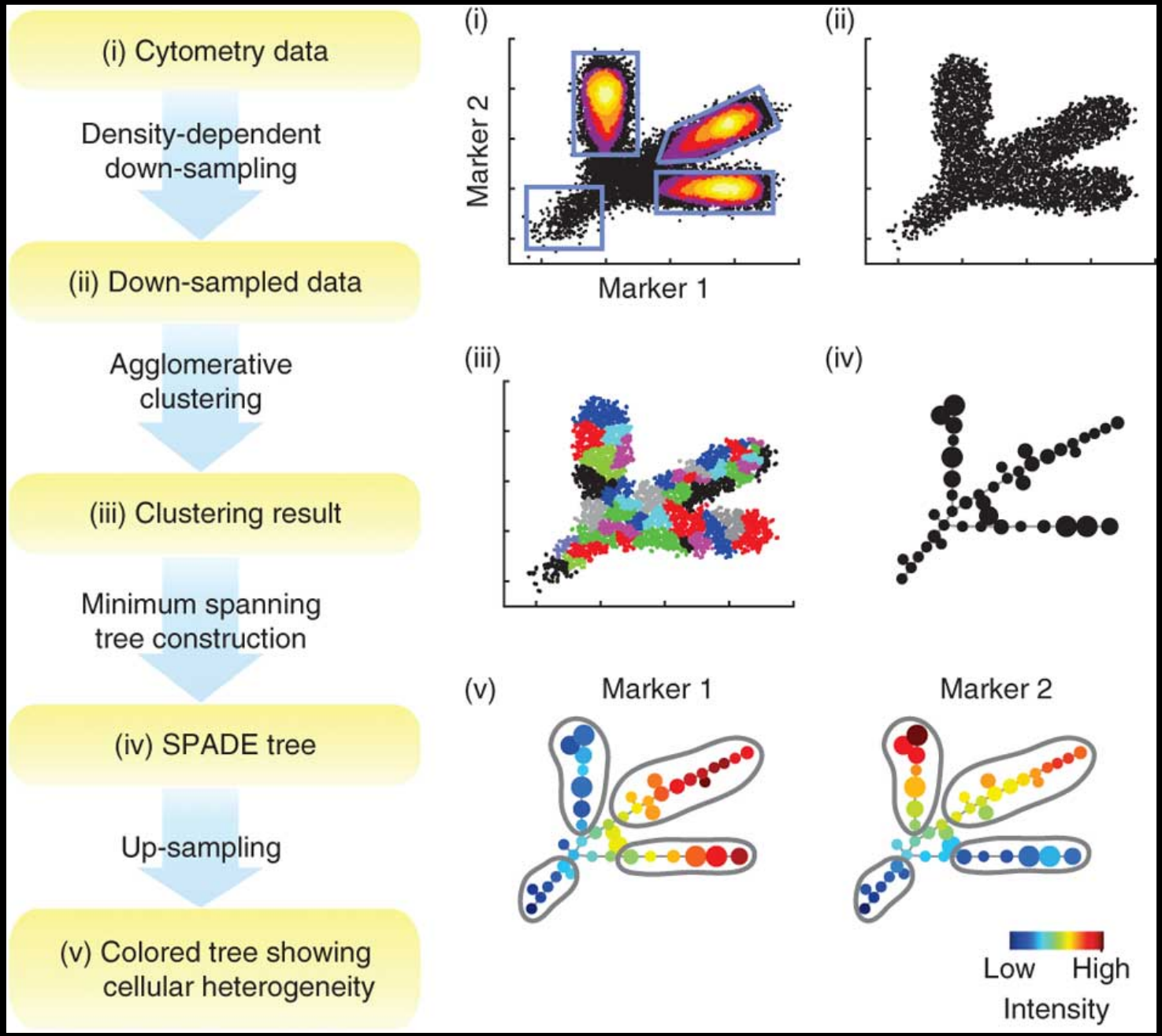
Which of these might be useful in the projects you are involved with this summer?

Tissues form hierarchies



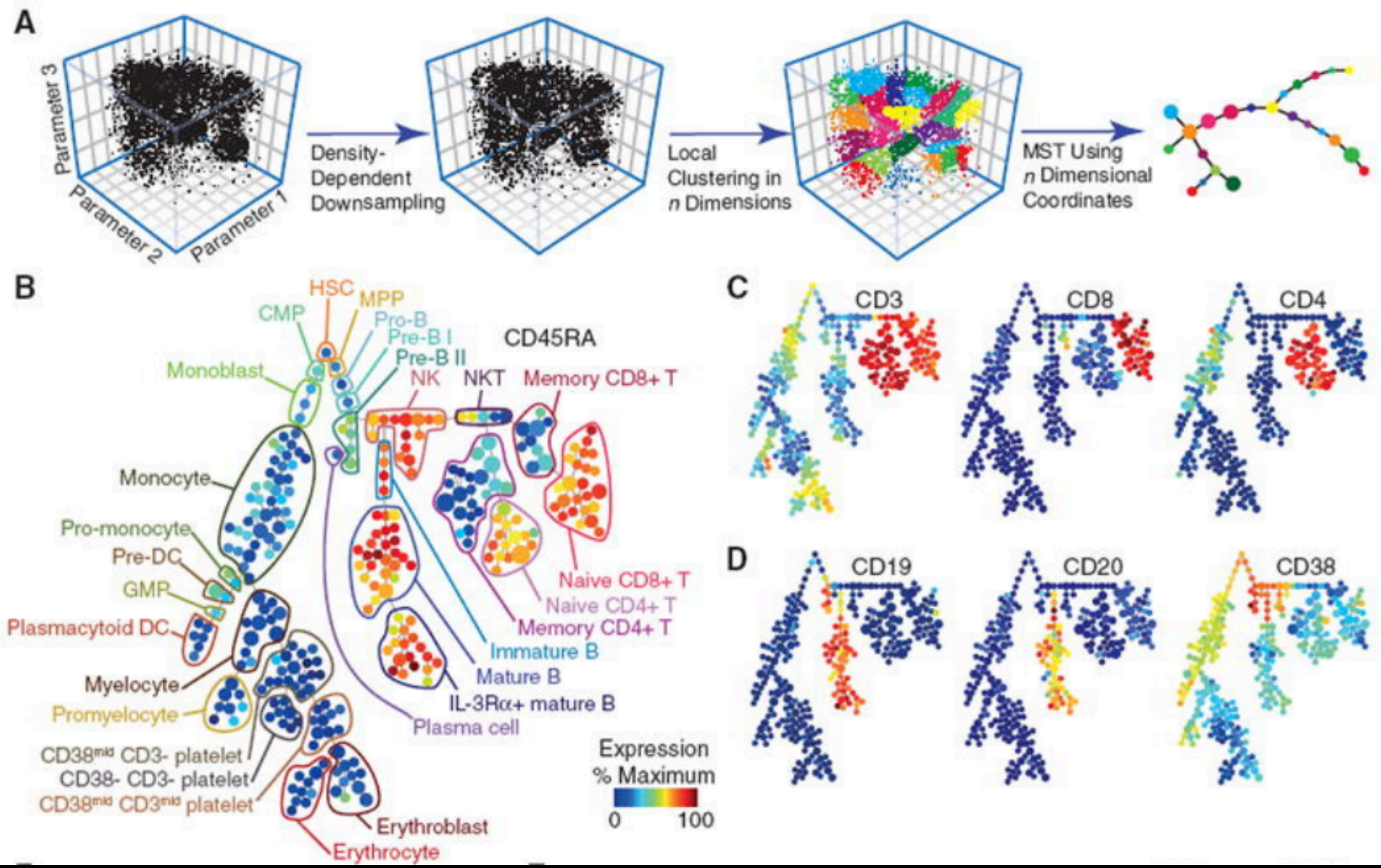
Hamey et al PNAS 2017

SPADE analysis



Qiu et al Nature Methods 2011

SPADE analysis of single-cell blood data



Bendall et al Science 2011

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