Working with genomics data

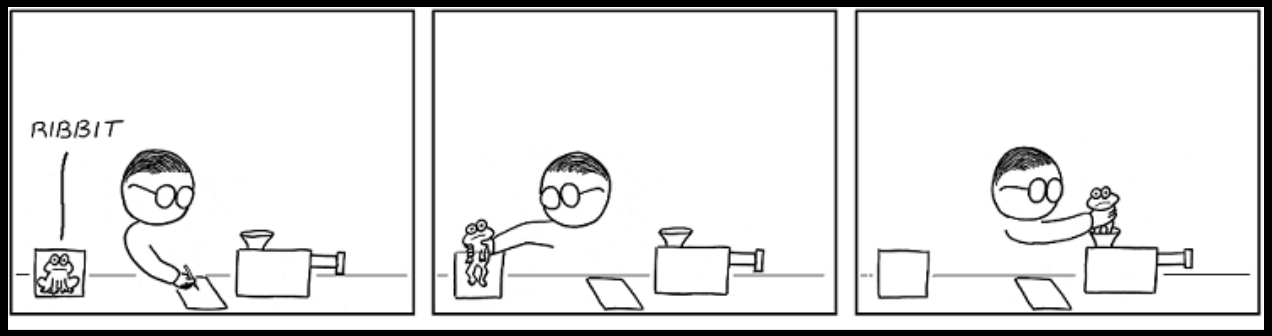
Andrew Gentles

Medicine (BMIR) and Biomedical Data Sciences

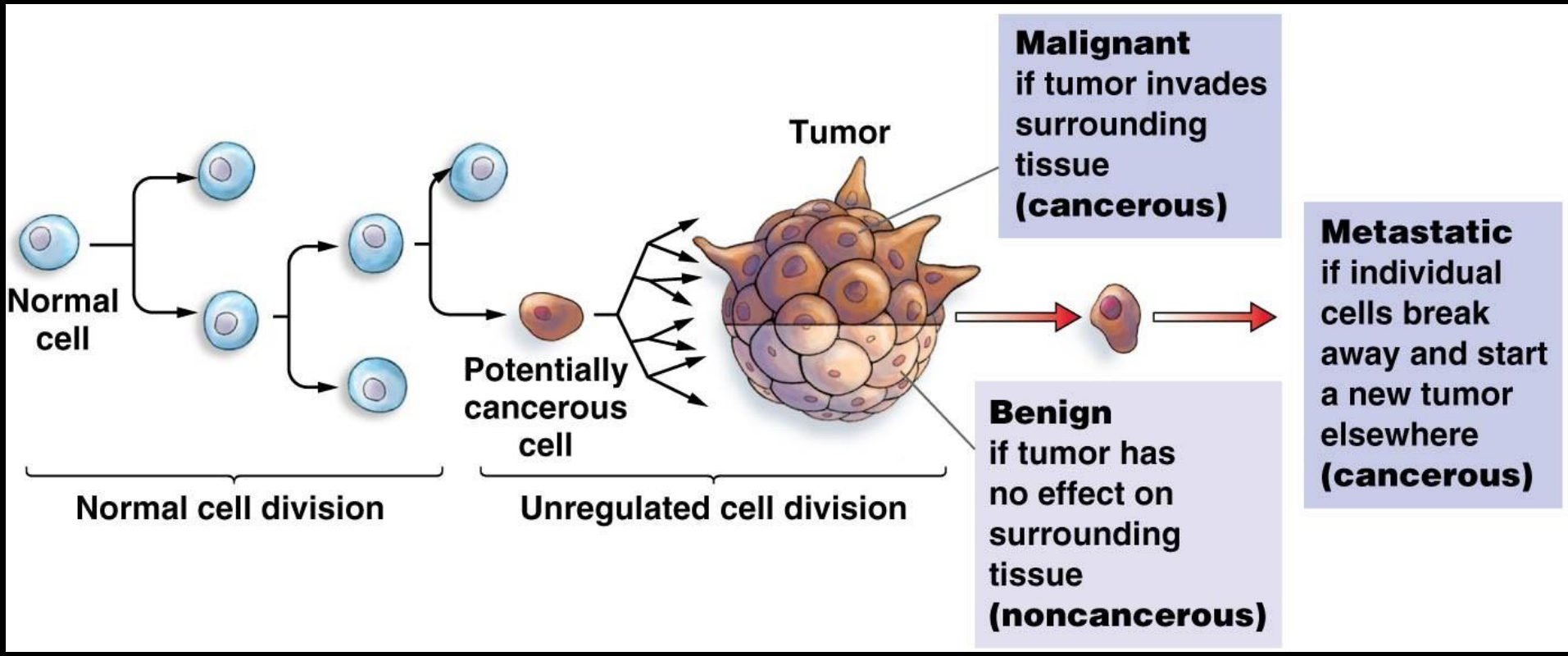

United Kingdom
Country Boundary
Road
River
Sovereign State Capital
Country Capital
City or Town
London is the capital of both the country of England and the sovereign state of the U.K.
0 50 100 KM
0 50 100 Miles
© 2018 Geology.com
Norwegian Sea
SHETLAND ISLANDS
North Sea
Lerwick
NORTH ATLANTIC OCEAN
1° 30' W
ORKNEY ISLANDS
Kirkwall
Thurso
Stornoway
Lewis with Harris
The Minch
Ullapool
North Uist
South Uist
Skye
Loch Ness
Inverness
Aberdeen
Rum
HEBRIDES
SCOTLAND
Fort William
Dundee
Perth
Saint Andrews
Mull
Oban
Kirkcaldy
Glasgow
Edinburgh
Islay
Tweed
Ayr
Arran
NORTH ATLANTIC OCEAN
Dumfries
Stranraer
Newcastle upon Tyne
Sunderland
Carlisle
Londonderry
Belfast
NORTHERN IRELAND
North Channel
Workington
Middlesbrough
Scarborough
Lough Neagh
Sligo
Douglas
Barrow-in-Furness
ENGLAND
Lough Allen
Dundalk
Isle of Man (Crown Dependency)
York
Kingston upon Hull
Westport
Blackpool
Bradford
Leeds
Preston
Drogheda
Irish Sea
Manchester
Grimsby
Lough Ree
Liverpool
Sheffield
Trent
Galway
Dublin
Holyhead
Anglesey
Chester
Stoke-on-Trent
Shannon
IRELAND
Caernarfon
Derby
Nottingham
King's Lynn
Norwich
Lough Derg
Shrewsbury
Wolverhampton
Leicester
Peterborough
Limerick
Birmingham
Severn
Tralee
Aberystwyth
Coventry
Cambridge
Waterford
Wexford
Wye
Northampton
Ipswich
Fishguard
Hereford
Stratford-upon-Avon
Cork
WALES
Saint George's Channel
Gloucester
Luton
Oxford
Swansea
Newport
London
Southend-on-Sea
Thames
Bristol
Reading
Canterbury
Cardiff
Bath
Dover
Bristol Channel
Southampton
Brighton
Hastings
Strait of Dover
Calais
Exeter
Bournemouth
Portsmouth
Weymouth
Isle of Wight
Plymouth
Torbay
English Channel
Penzance
ISLES OF SCILLY
Alderney (Crown Dependency)
FRANCE


ALICE
CMS
ATLAS

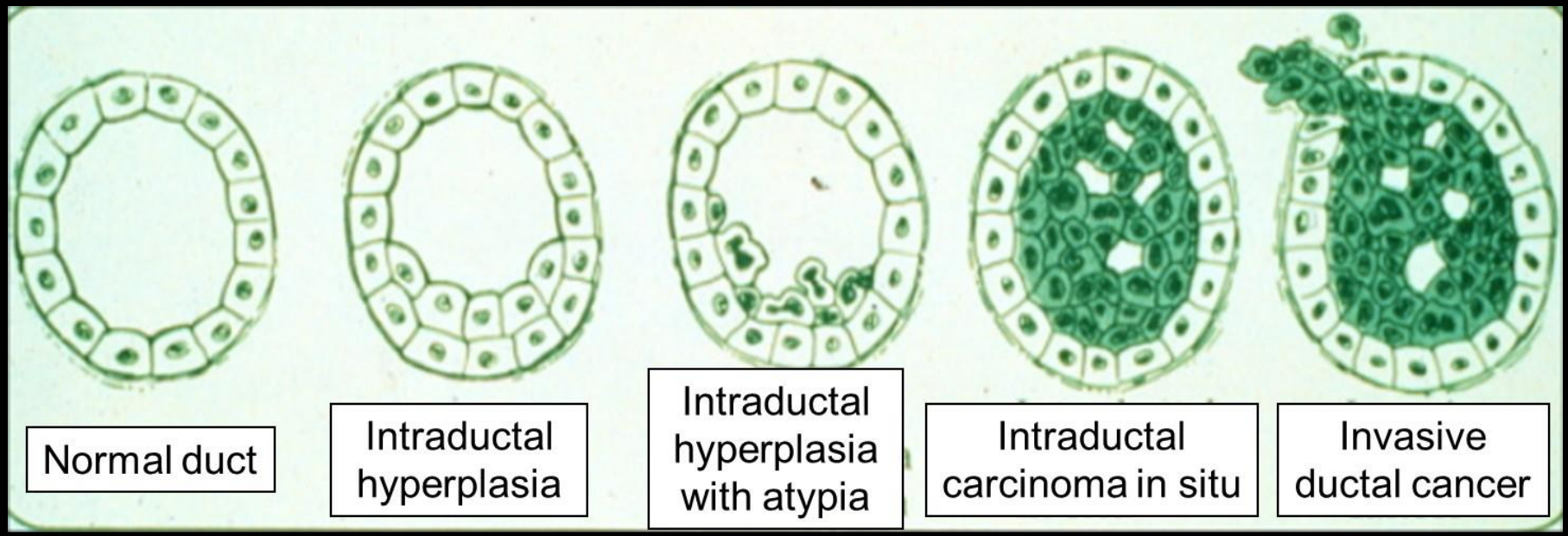
Physics -> Biology



How do tumors develop?



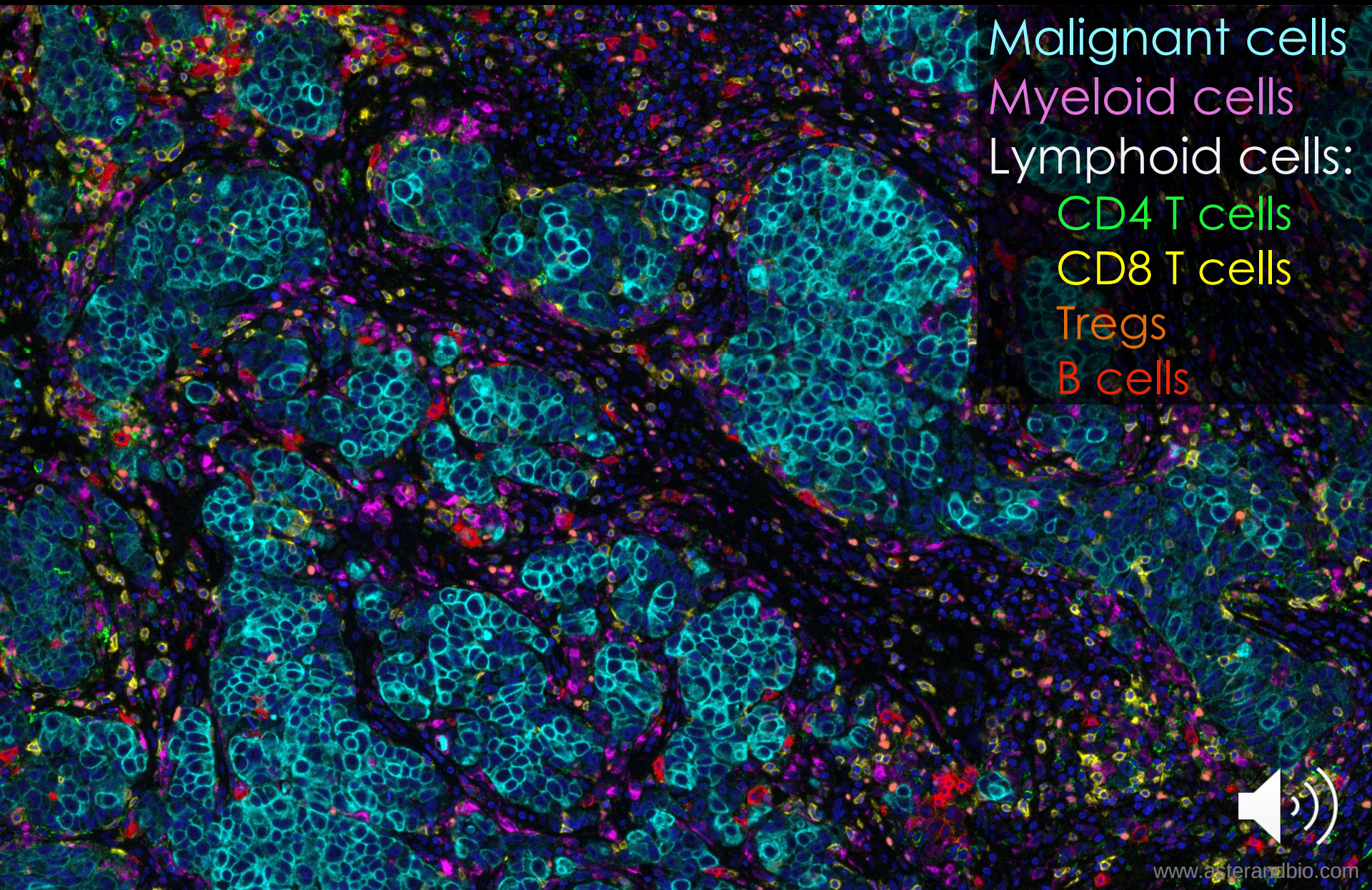
Breast Cancer Development



5 year survival:
Stage 1: 95-100%
Stage 2: 86%
Stage 3: 57%
Stage 4: 20%

BUMC.org

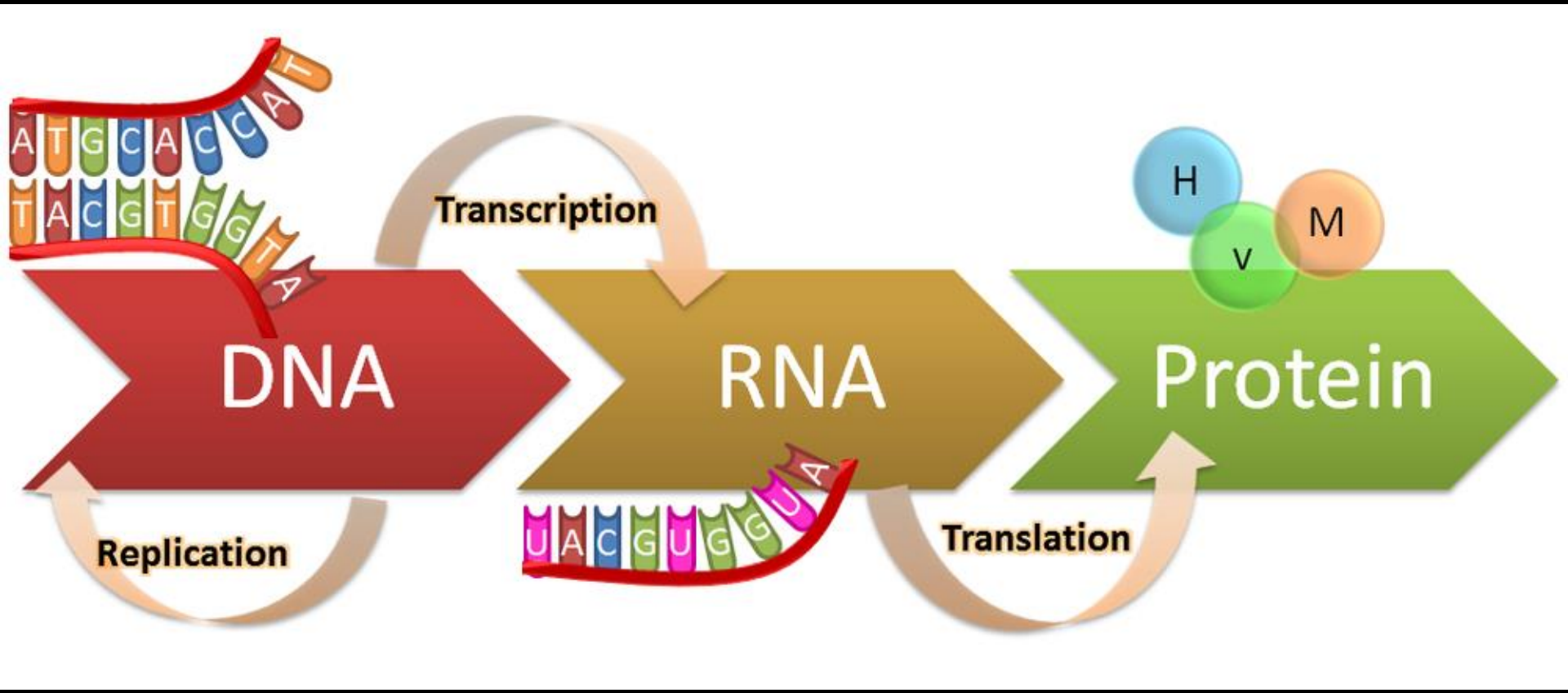
Tumors are complex cellular ecosystems



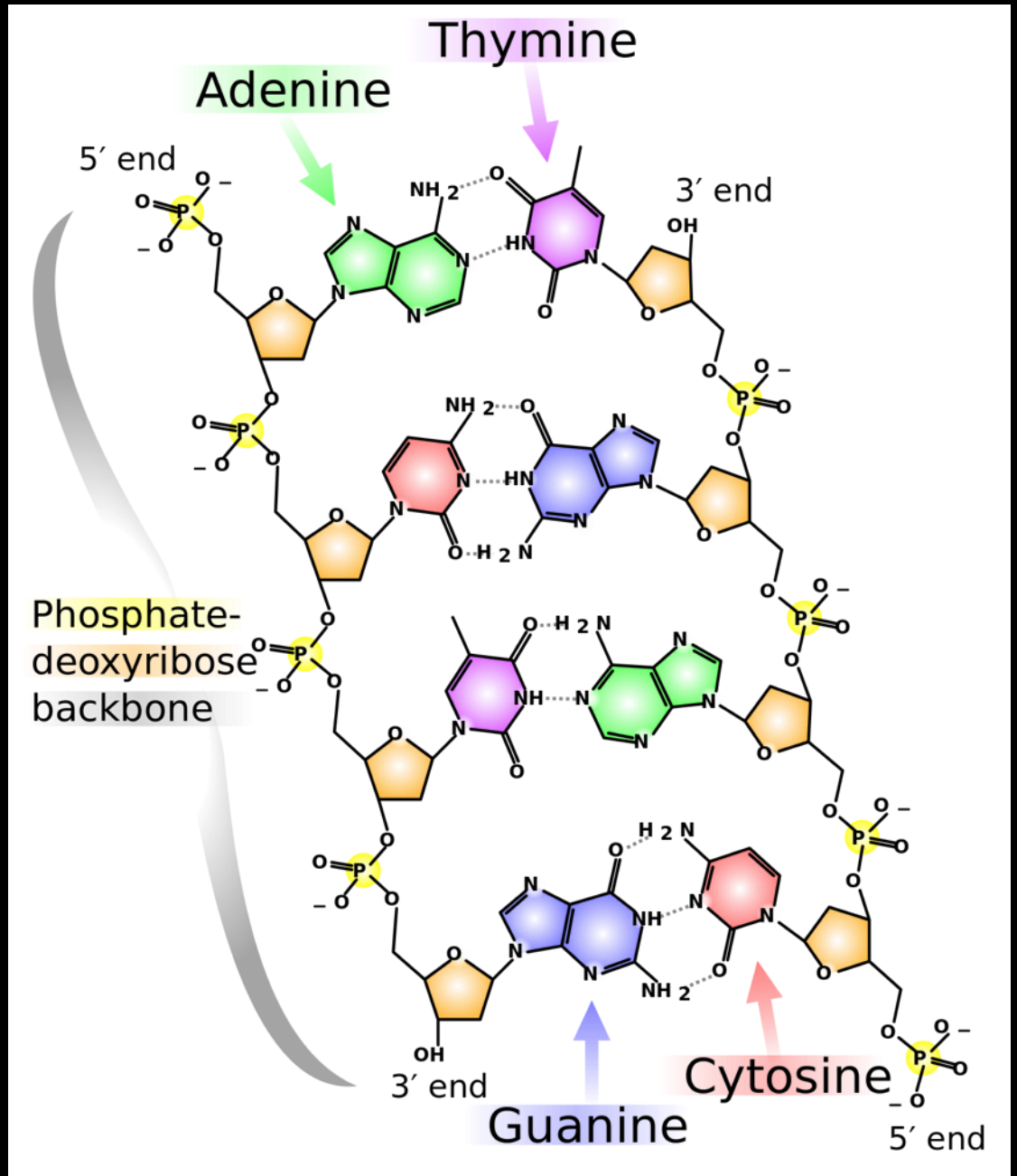
Working with data - systems biology

- What is genomic data?
- Regulatory networks
  - Transcriptional networks
- Connecting genomics with outcomes (survival)
- Dissecting tissues (e.g. tumors) at single cell resolution

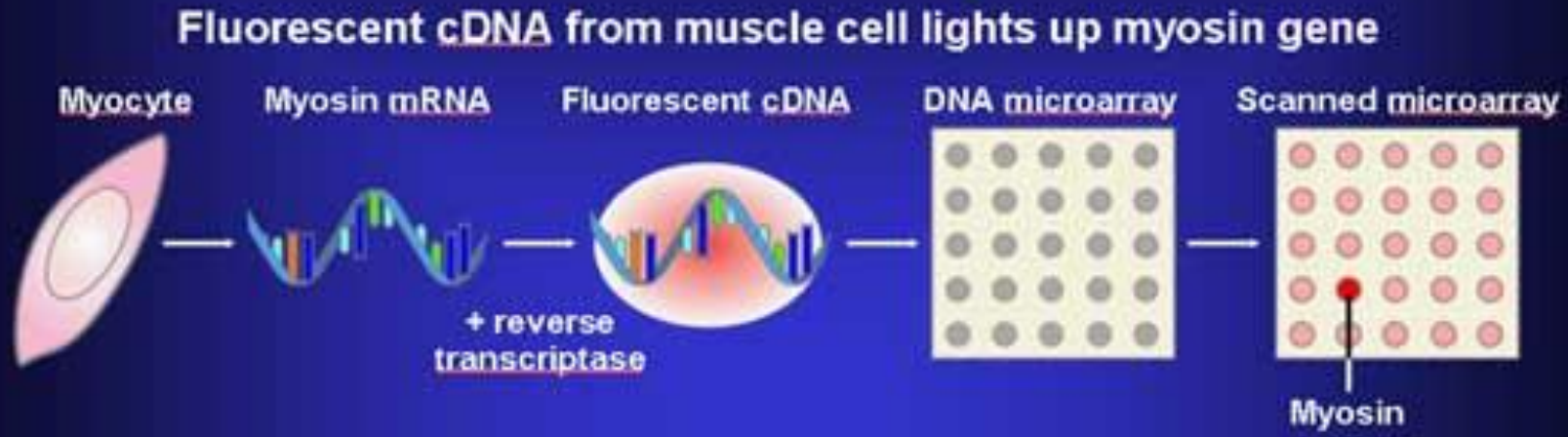
Central dogma of molecular biology

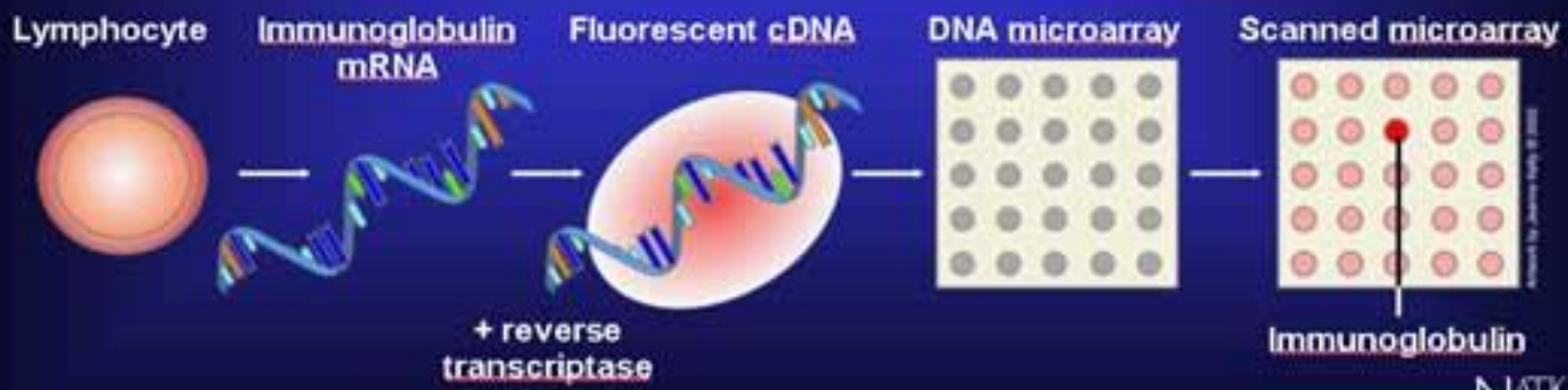


DNA and RNA complementarity



Using DNA Microarrays to Measure Gene Expression

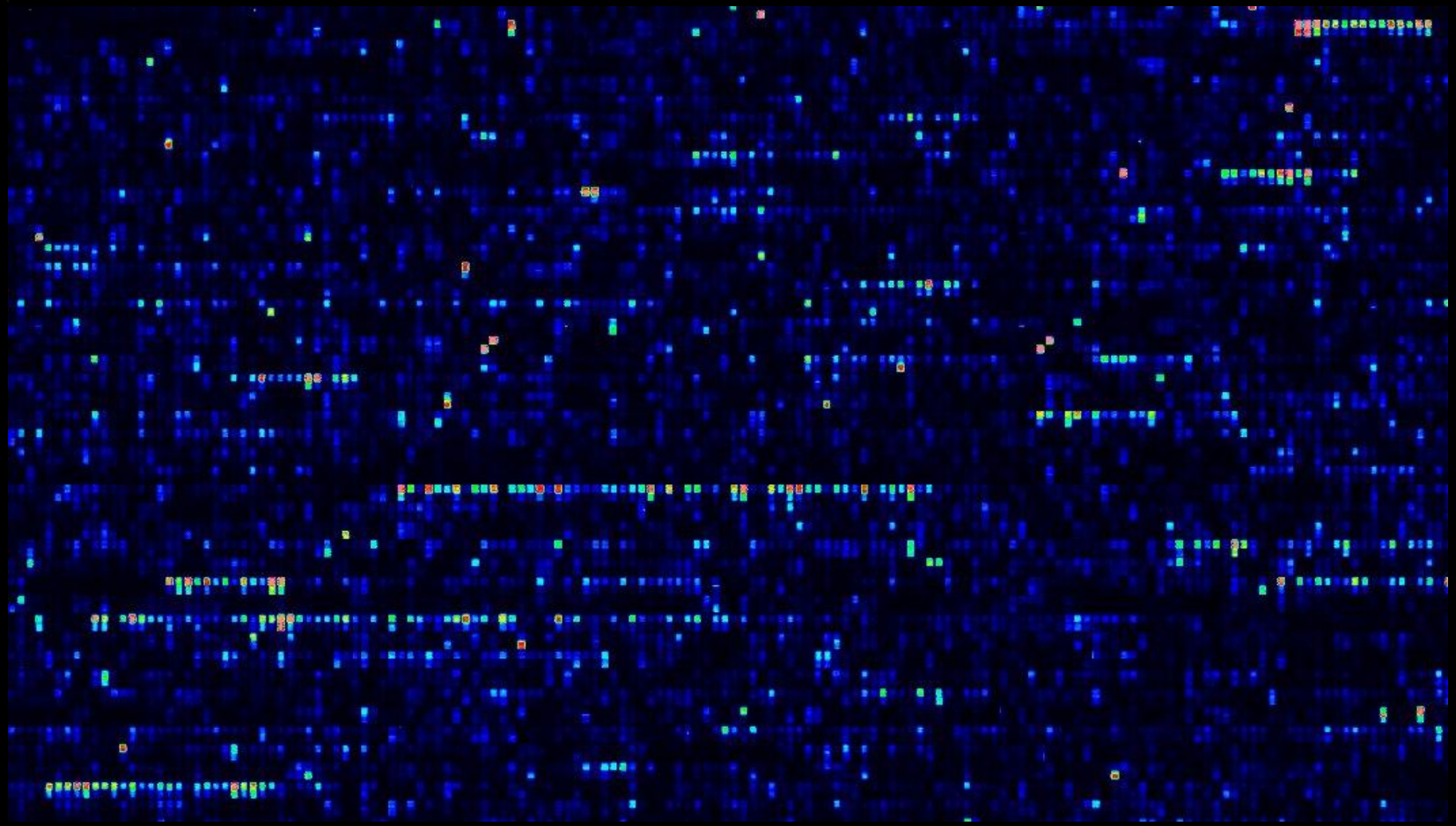




NATIONAL CANCER INSTITUTE

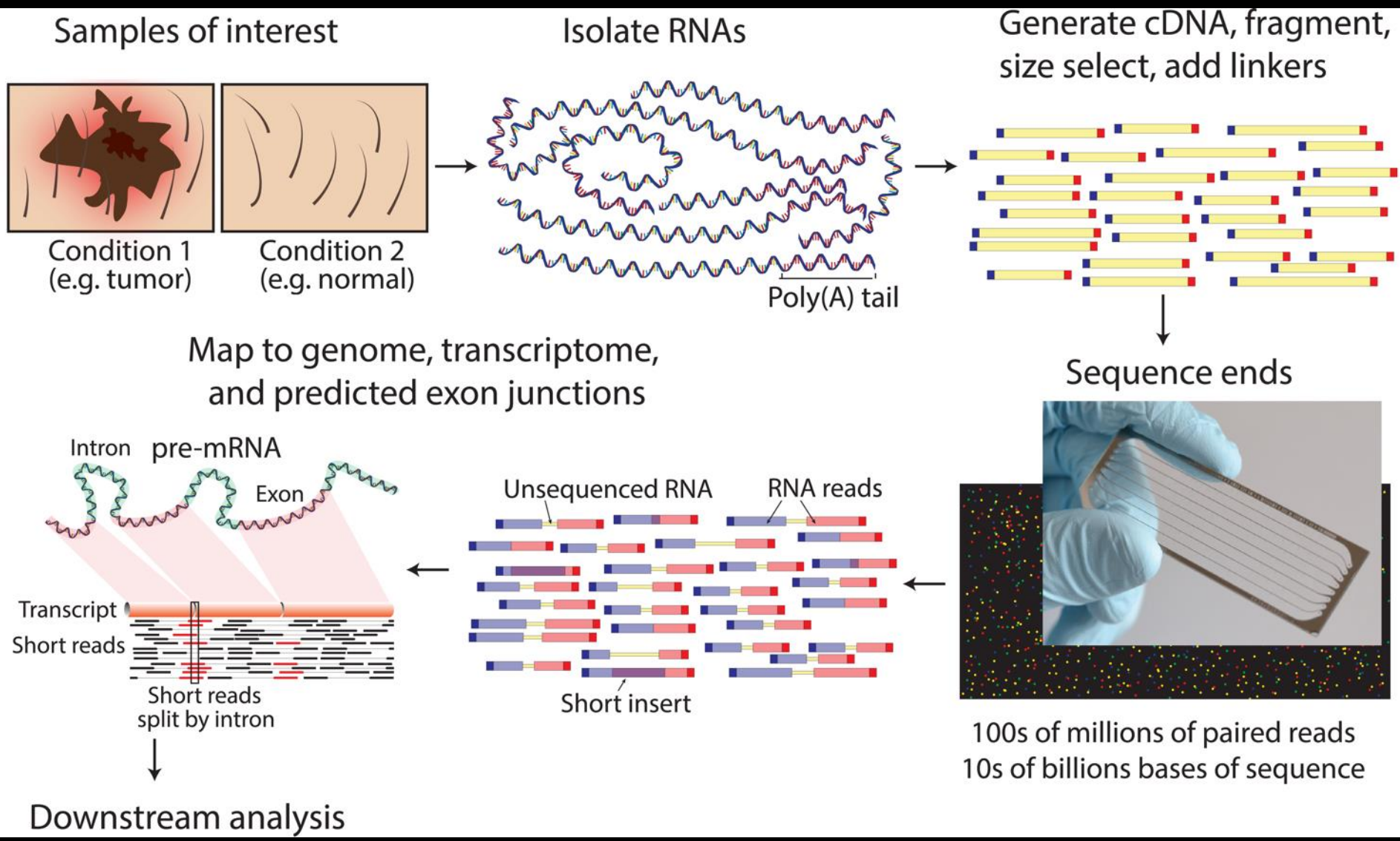
We can assess activity of thousands of genes in a sample

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



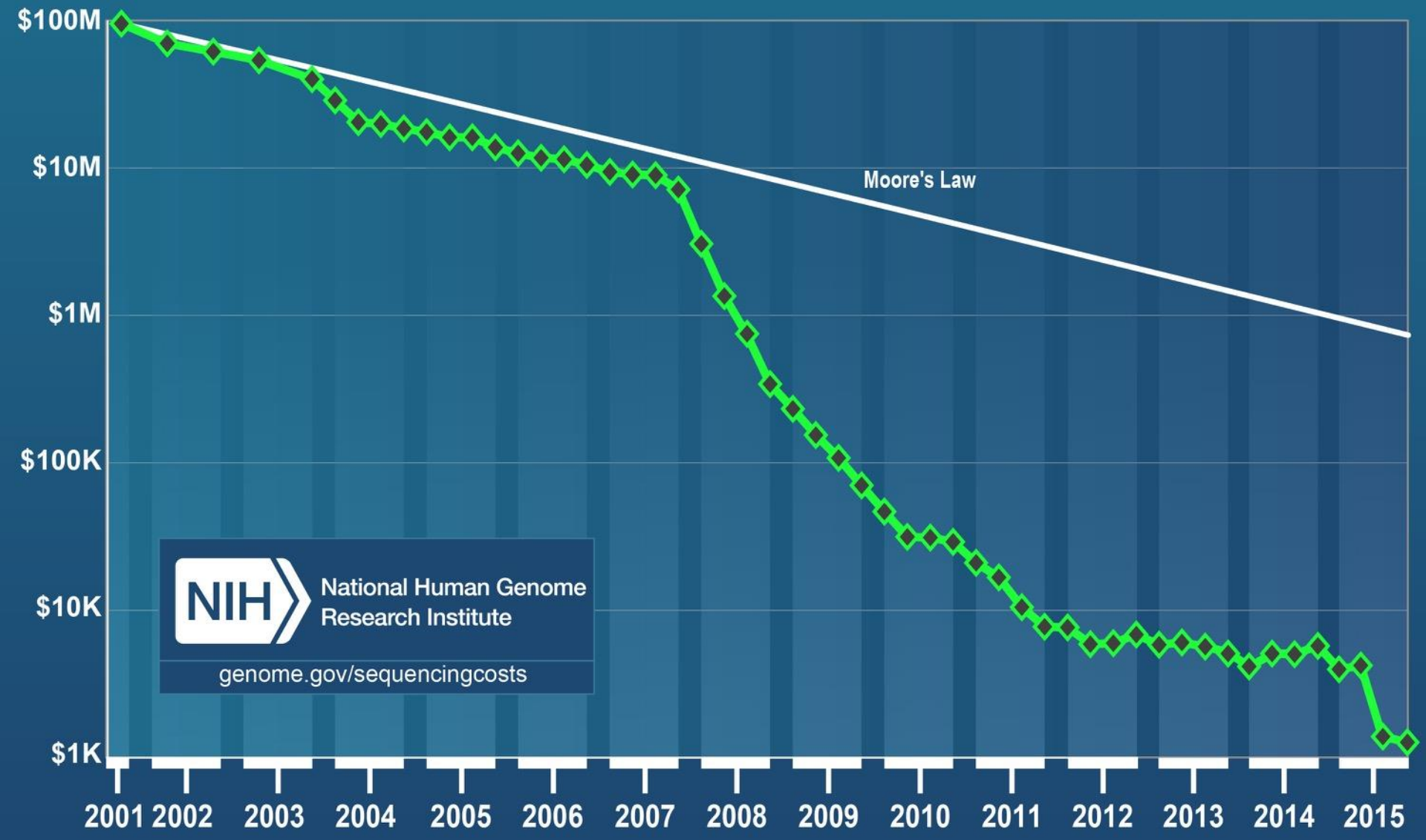
http://learn.genetics.utah.edu/content/labs/microarray/

Next generation sequencing



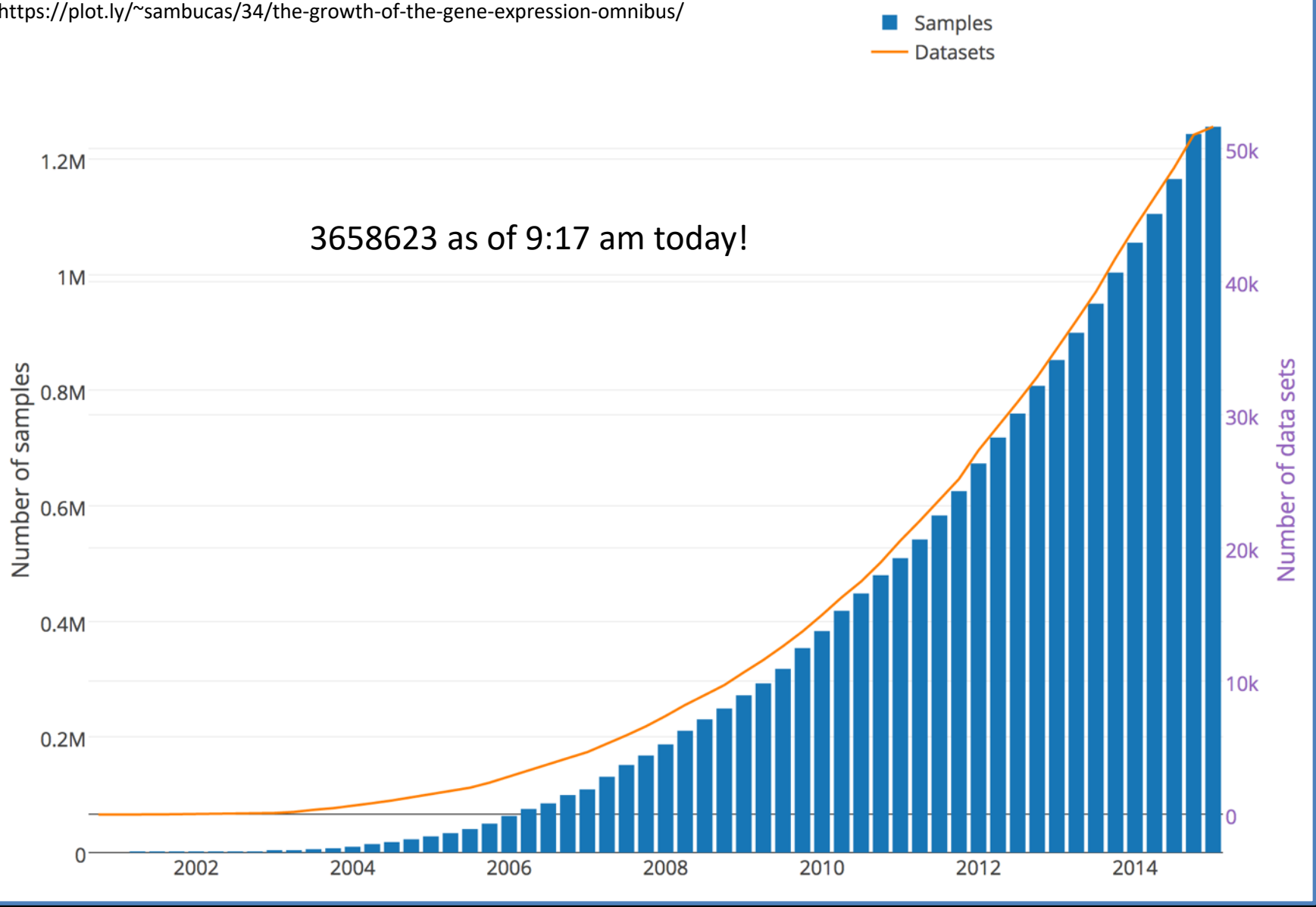
https://en.wikipedia.org/wiki/RNA-Seq

Cost per Genome



Growth of gene expression omnibus at NCBI

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



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)

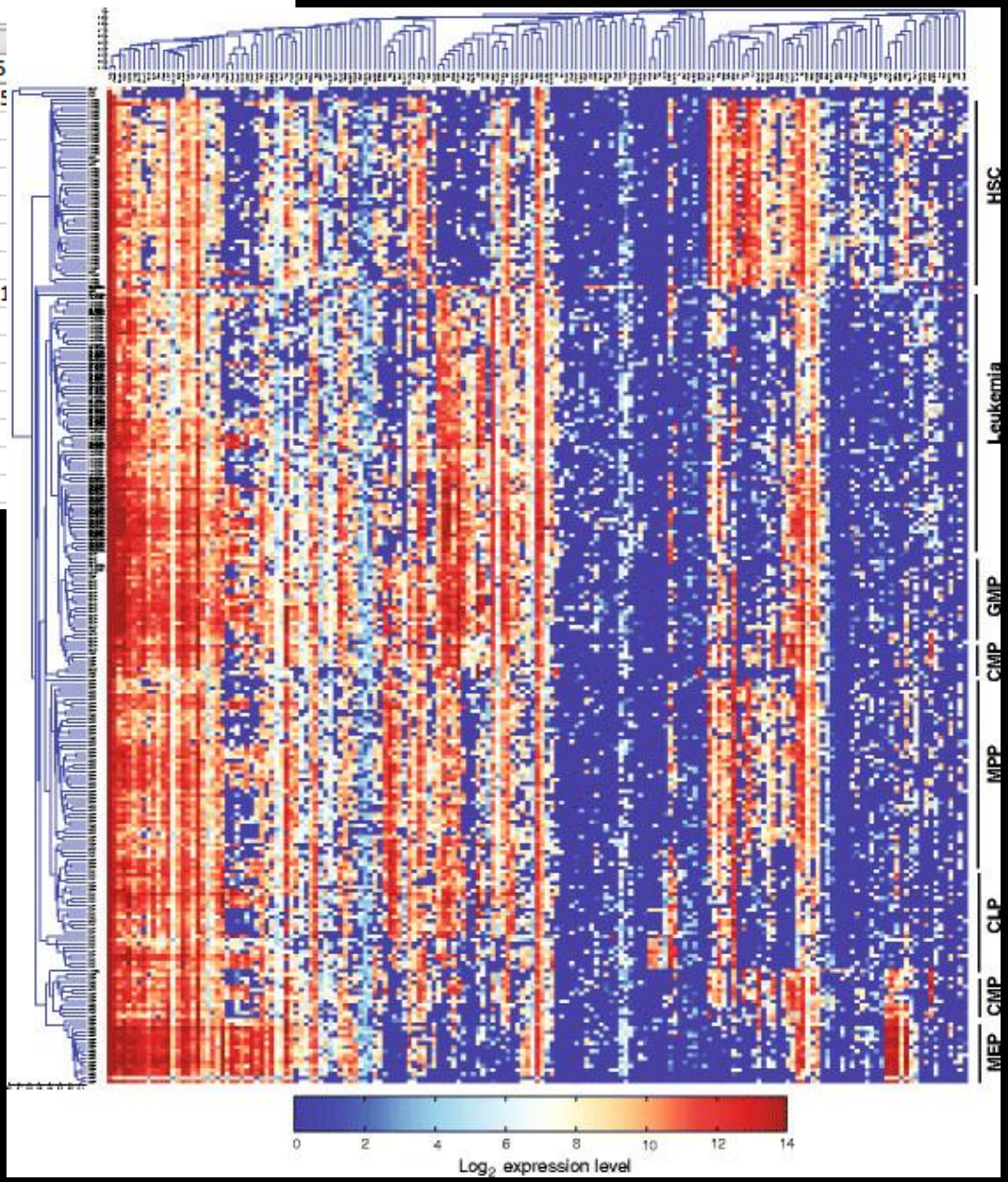
- What about the analysis ?
  - Storage
  - Compute costs
  - Interpretation

- *It will be a long time before computational biologists are out of a job*

Moderately large scale data

J15

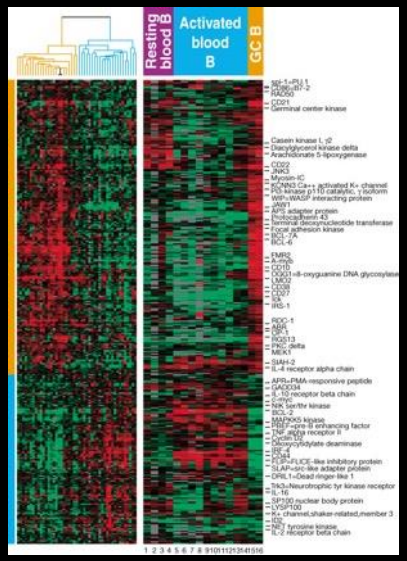
	A	B	C	D	E	
1	Acc ID	Exp1	Exp2	Exp3	Exp4	Exp5
2	NM_007818	67540.89	70924.09	80243.76	3501.2	5
3	NM_001105160	811.93	801.36	740.71	128.67	
4	NM_028089	190.41	211.06	236.19	9.05	
5	NM_016696	66.77	57.56	101.09	750.9	
6	NM_013459	3.3	11.29	1.89	735.82	
7	NM_007809	45.34	36.12	51.02	245.27	
8	NM_009999	103.04	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	

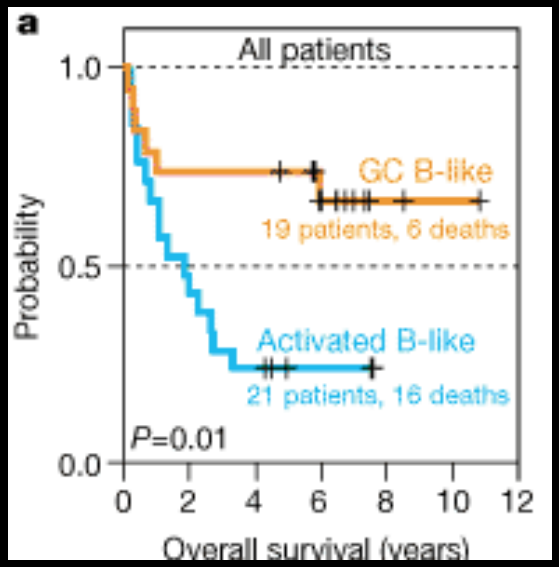


10-1000's columns (samples)

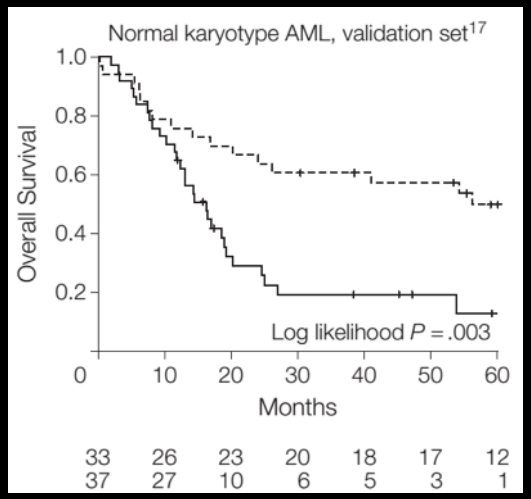
10,000's-100,000's rows (genes, transcripts)

Genomic profiles and clinical outcome

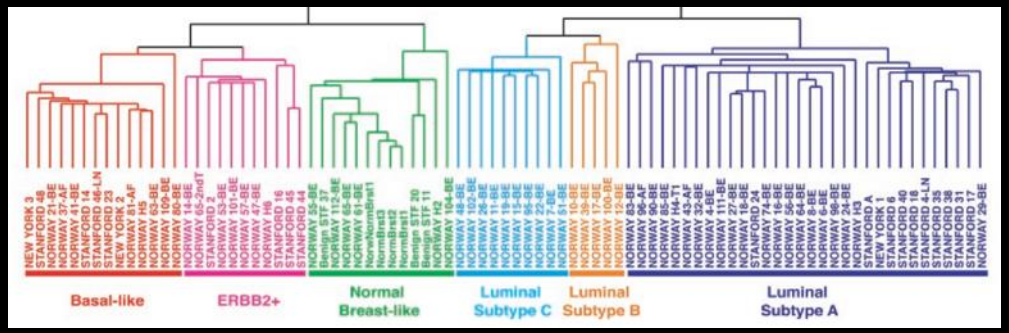




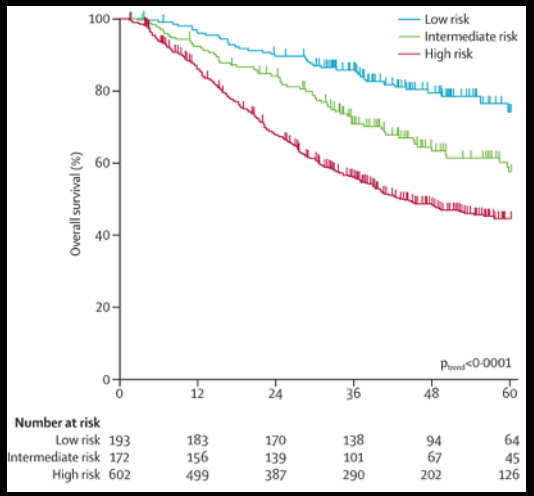
Alizadeh et al Nature 2000. **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**



Gentles et al *JAMA* 2010



Sorlie et al *PNAS* 2001



Kratz et al. *Lancet* 2012

Associating survival with expression levels

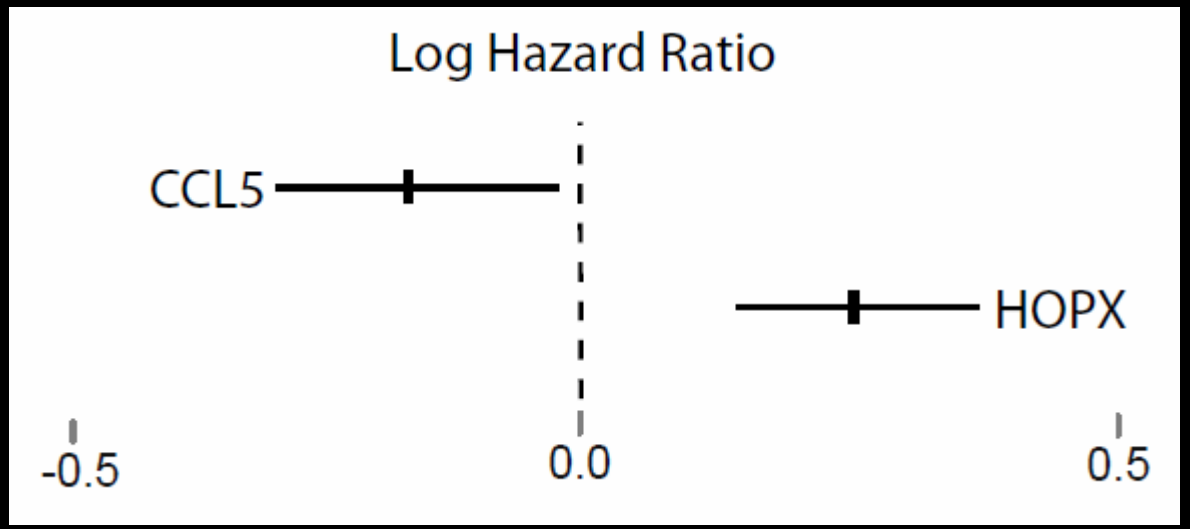
Variable	Patient1	Patient2	Patient4	Patient5	Patient6	Patient7	…
Time	12	3	8	35	14	22	…
Status	1	1	1	0	0	1	…
Gene1	1.45	0.15	-0.59	-1.88	-0.83	-0.26	…
Gene2	0.94	-0.35	2.66	-0.23	2.09	-0.13	…
Gene3	0.91	-0.32	-0.82	-0.35	0.86	0.32	…
…							

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)

Gene expression ~ survival

Z=-2.2
p=0.02

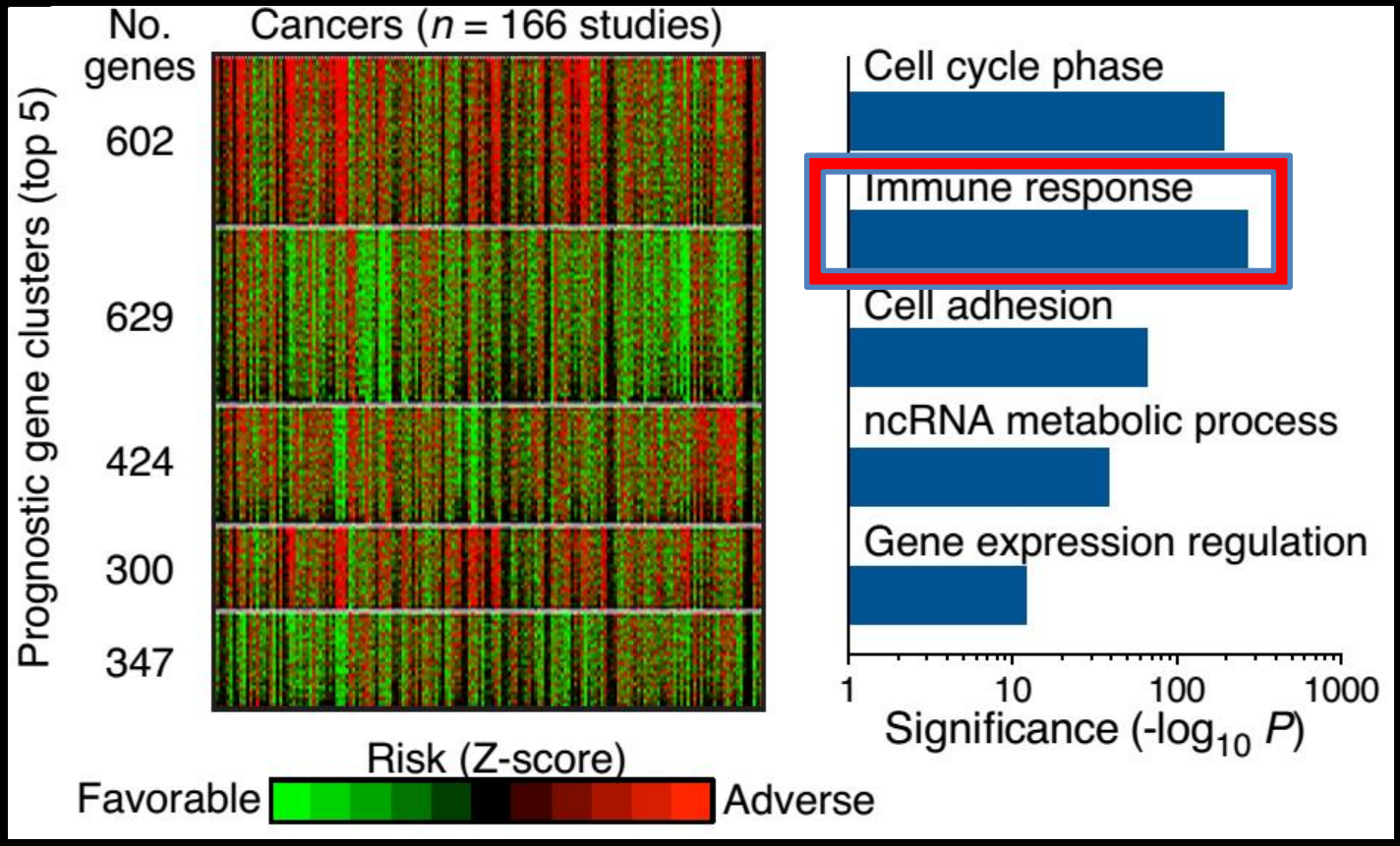


Z=+4.5
p=4e-6

"Good" genes

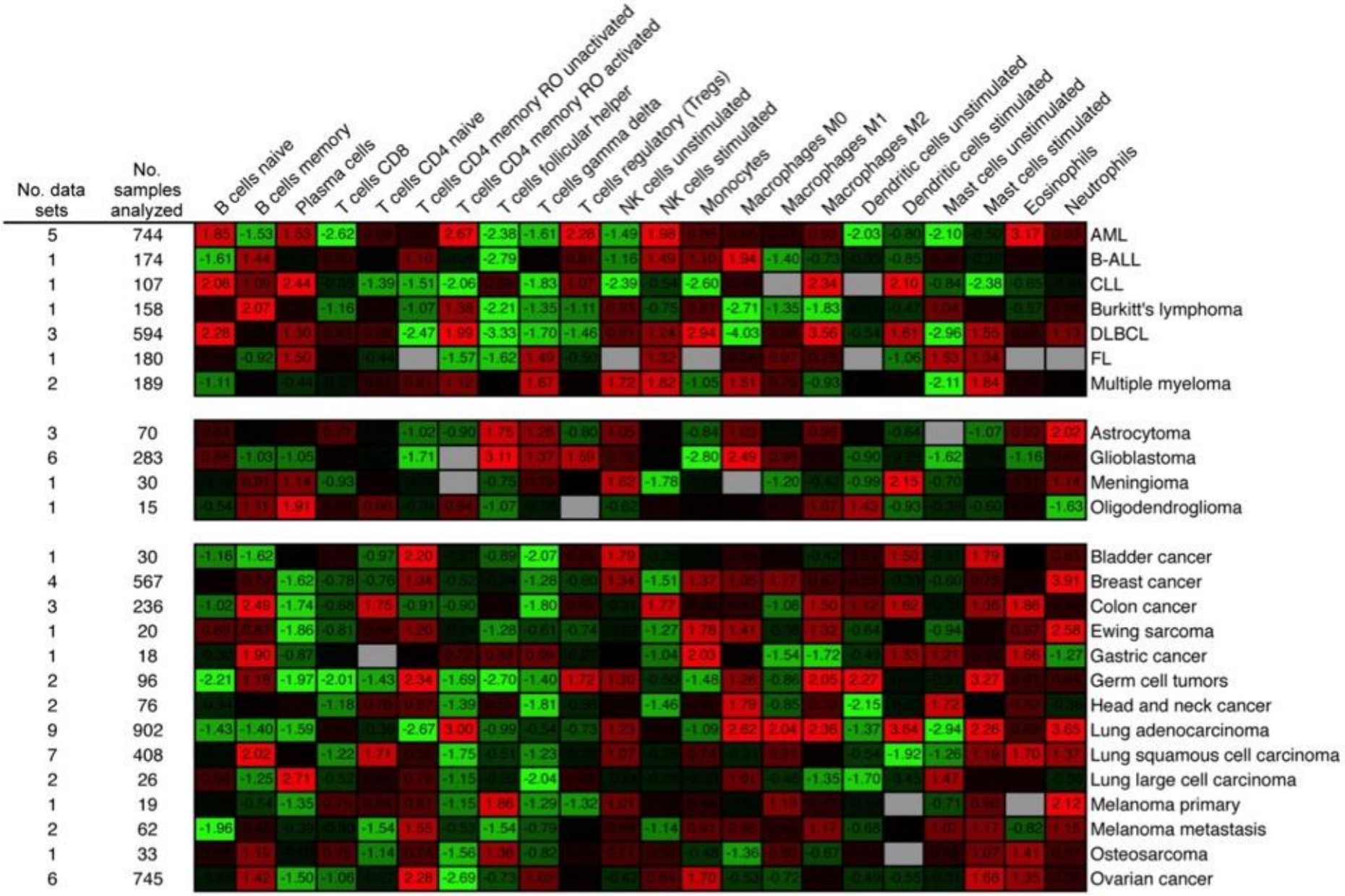
"Bad" genes

Clustering of outcome matrix -> biological processes



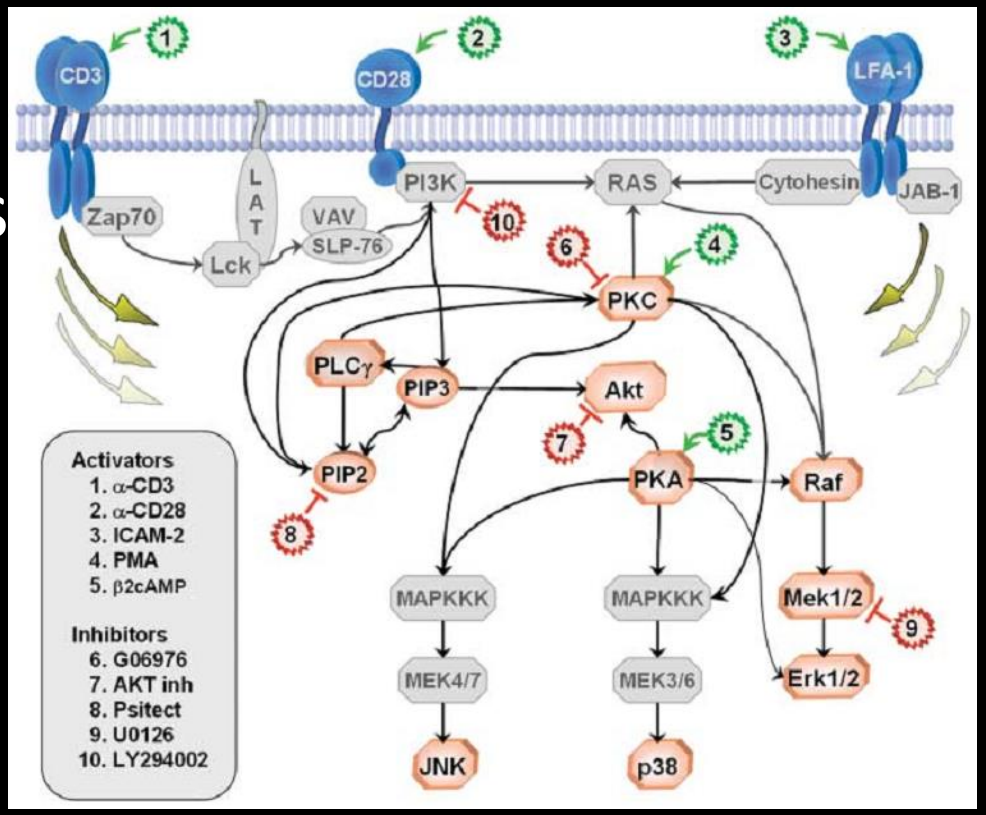
http://precog.stanford.edu

Prognostic influence of immune infiltrates



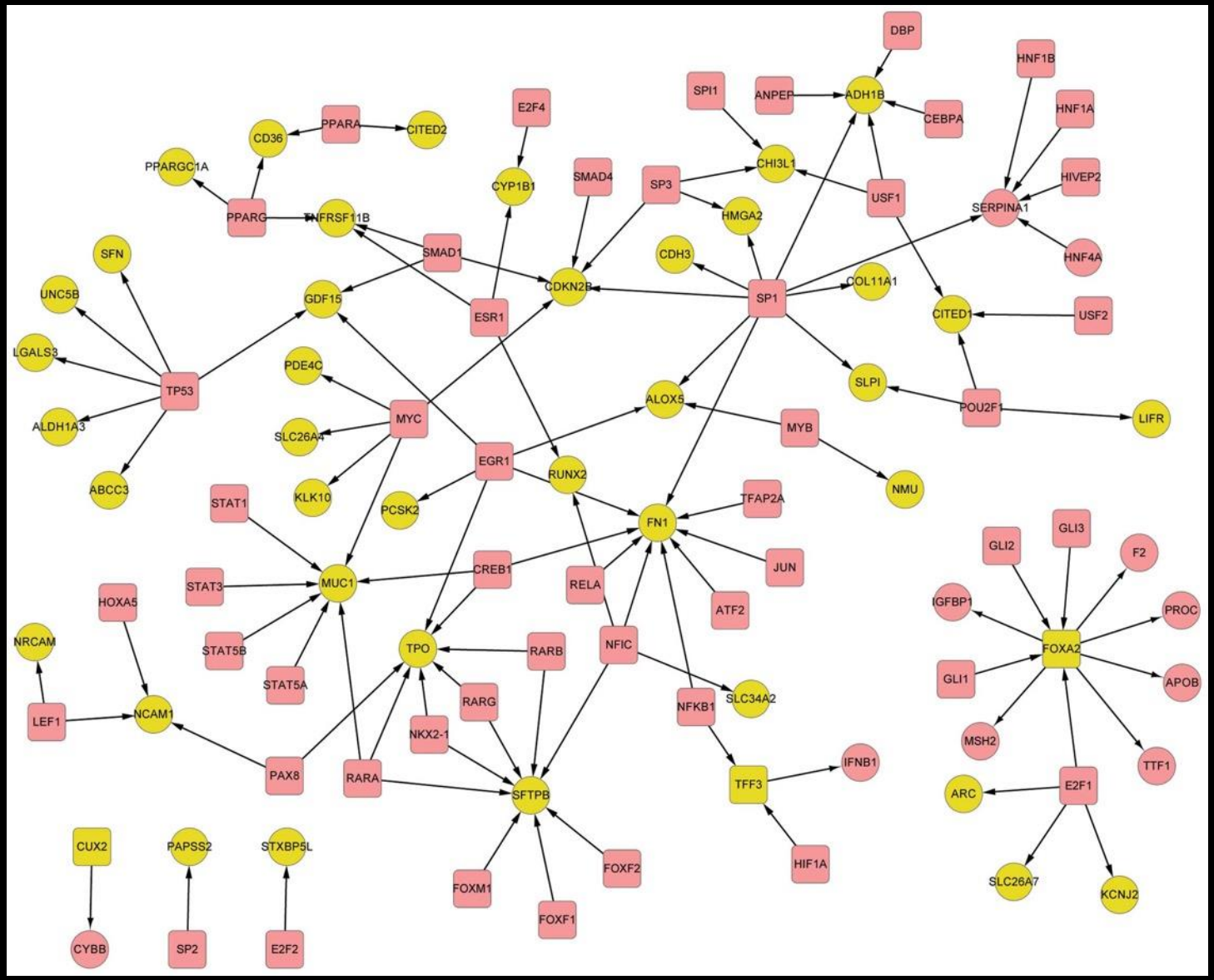
Network types

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

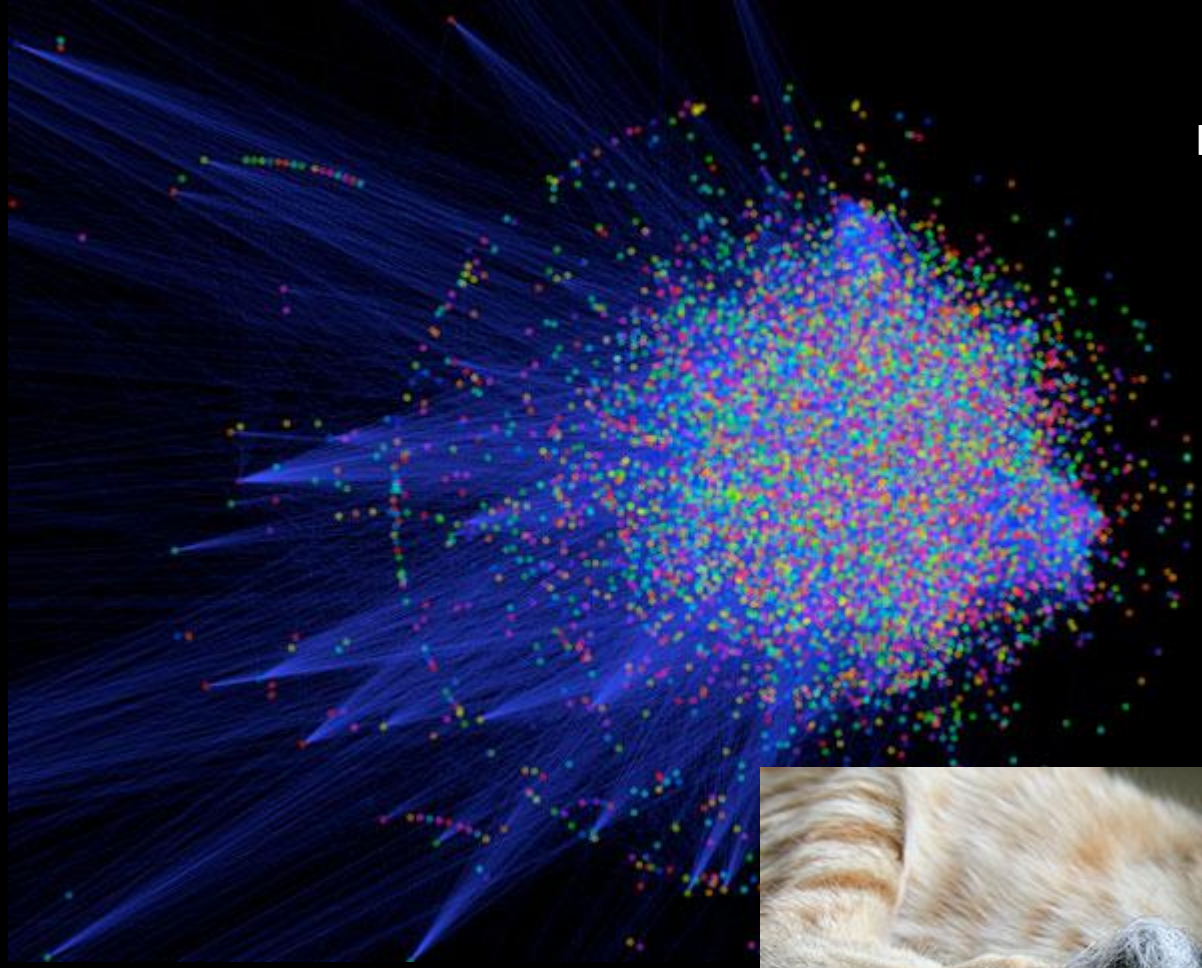


Sachs et al. http://www.sciencemag.org/content/308/5721/523.full

Gene regulatory networks



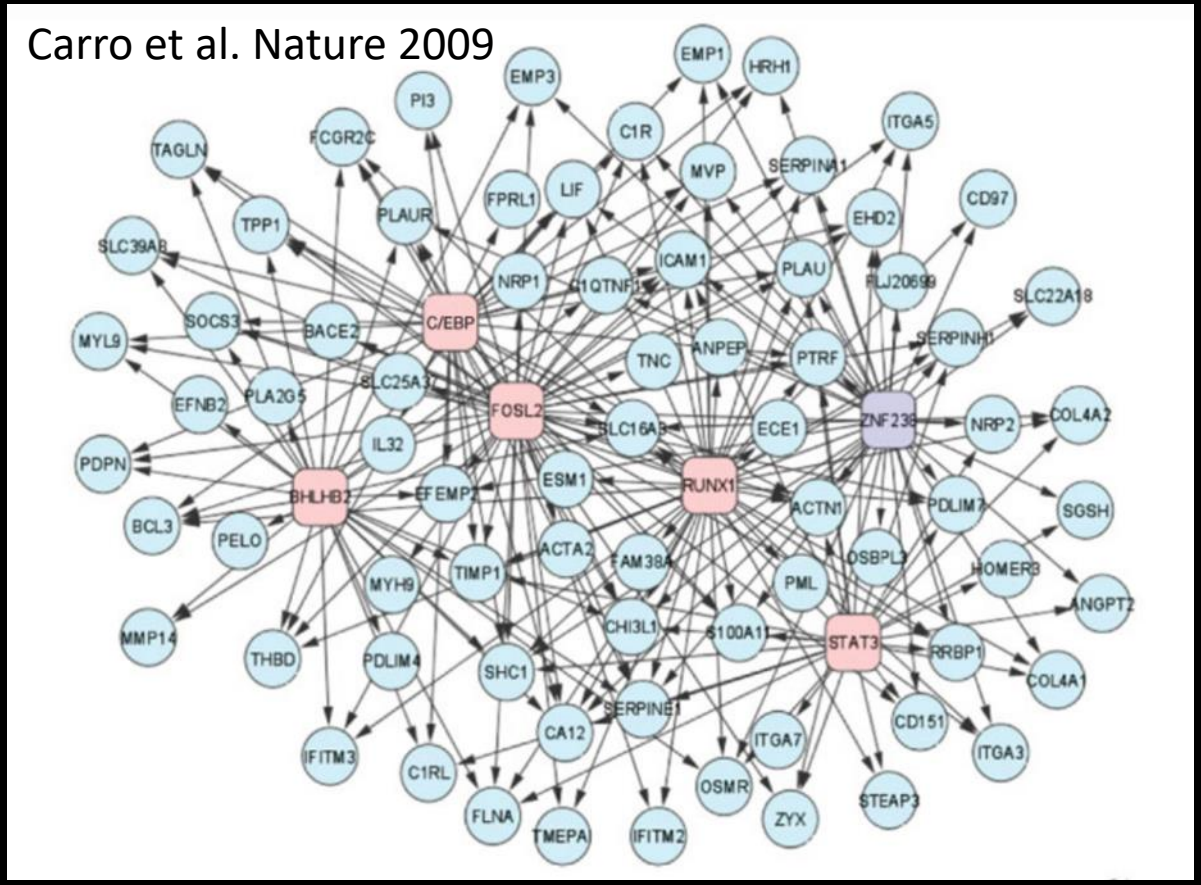
They are a “flowchart” of interactions, not a detailed model



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?

Heterogeneity of tissues (and tumors)

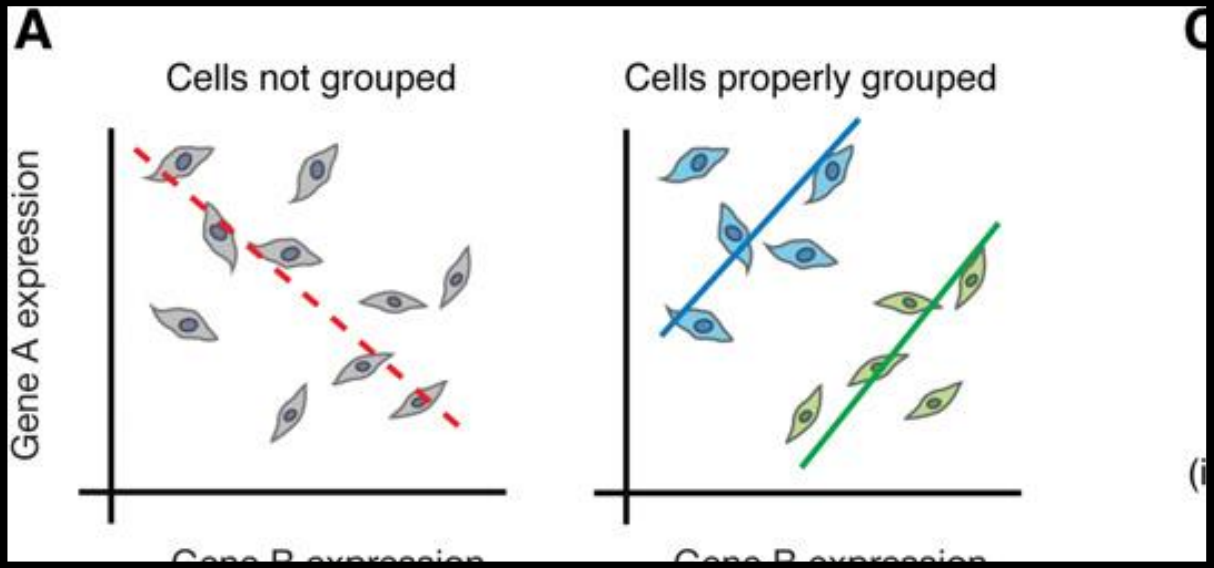
Bulk RNA-seq

Single-cell RNA-seq

Profiling of sorted populations

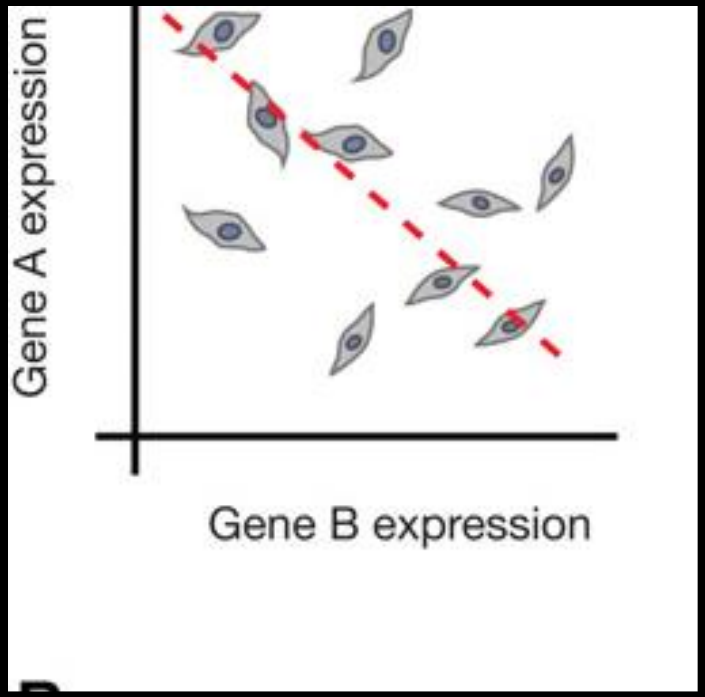
- Can you tell what’s in a smoothie by taste?
- How many different flavours?
- How small a proportion of the mix?
- Is it easier to detect presence or absence ?

Averaging across cell types



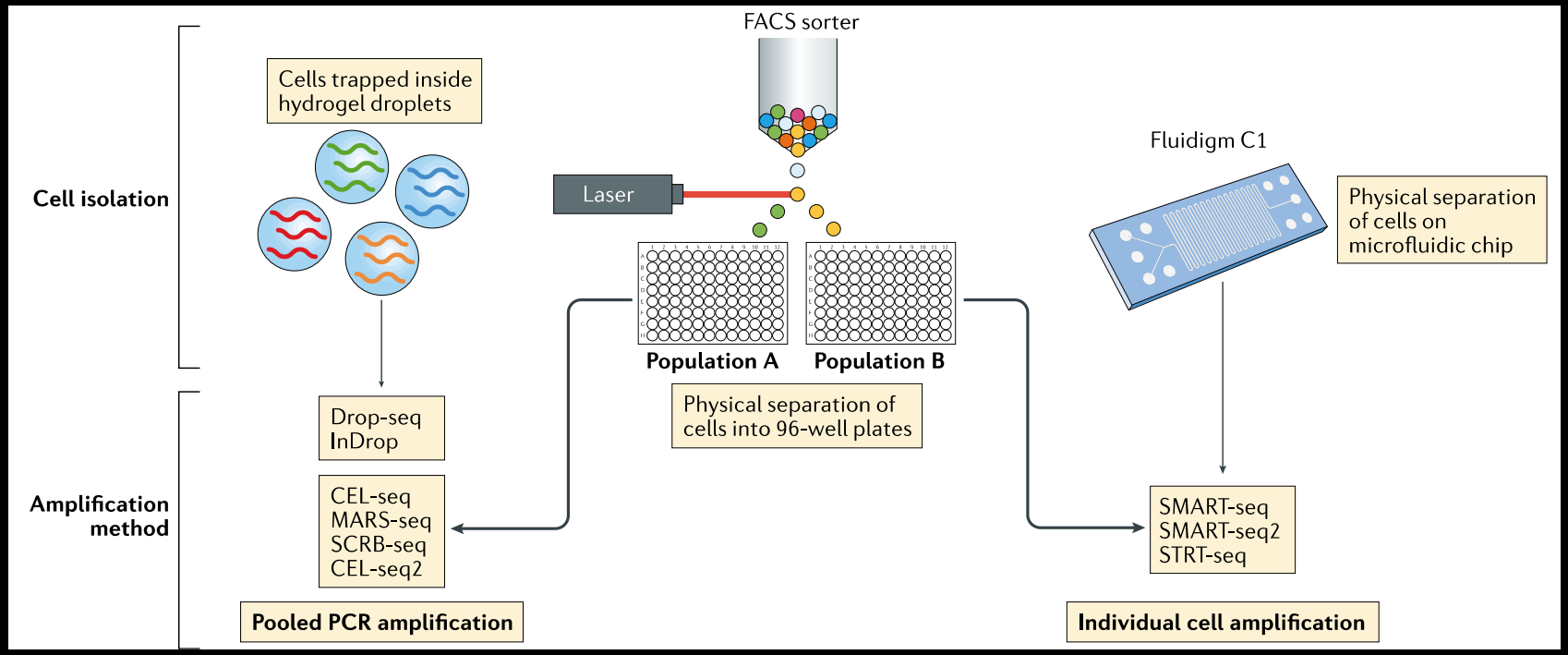
Trapnell et al. 2015 Genome Res. 25: 1491-1498

Why single cell?



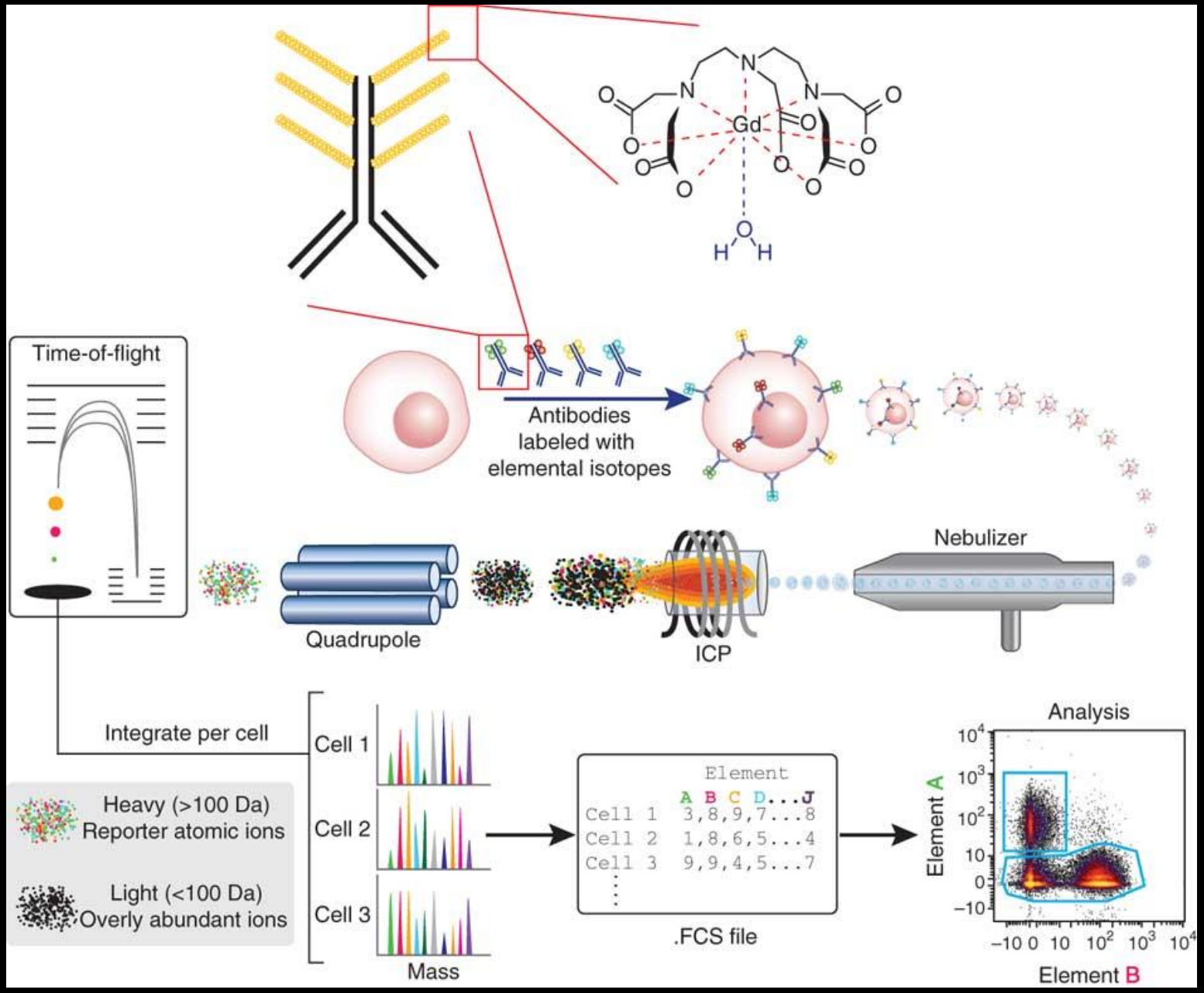
Trapnell et al. 2015 Genome Res. 25: 1491-1498

Single cell RNA-seq approaches



Papalexi & Satuja Nat Rev Imm 2018

Mass cytometry – single cell proteomics



Summary

- We have technologies to measure many things at the same time on large numbers of samples
- We can test which of thousands of genes are connected with things of importance such as how long patients survive
- Which genes can tell us about why some patients live longer
- The immune system is very important – almost doesn’t matter what treatment
- Possible treatment options