DSPT
DATA SCIENCE PORTUGAL

#71 - *Memes to Genes: Cracking the Human Code*

@ talktalkdesklisbon, 16th January 2020

What transcriptomic data tell us about disease

Nuno Barbosa Morais

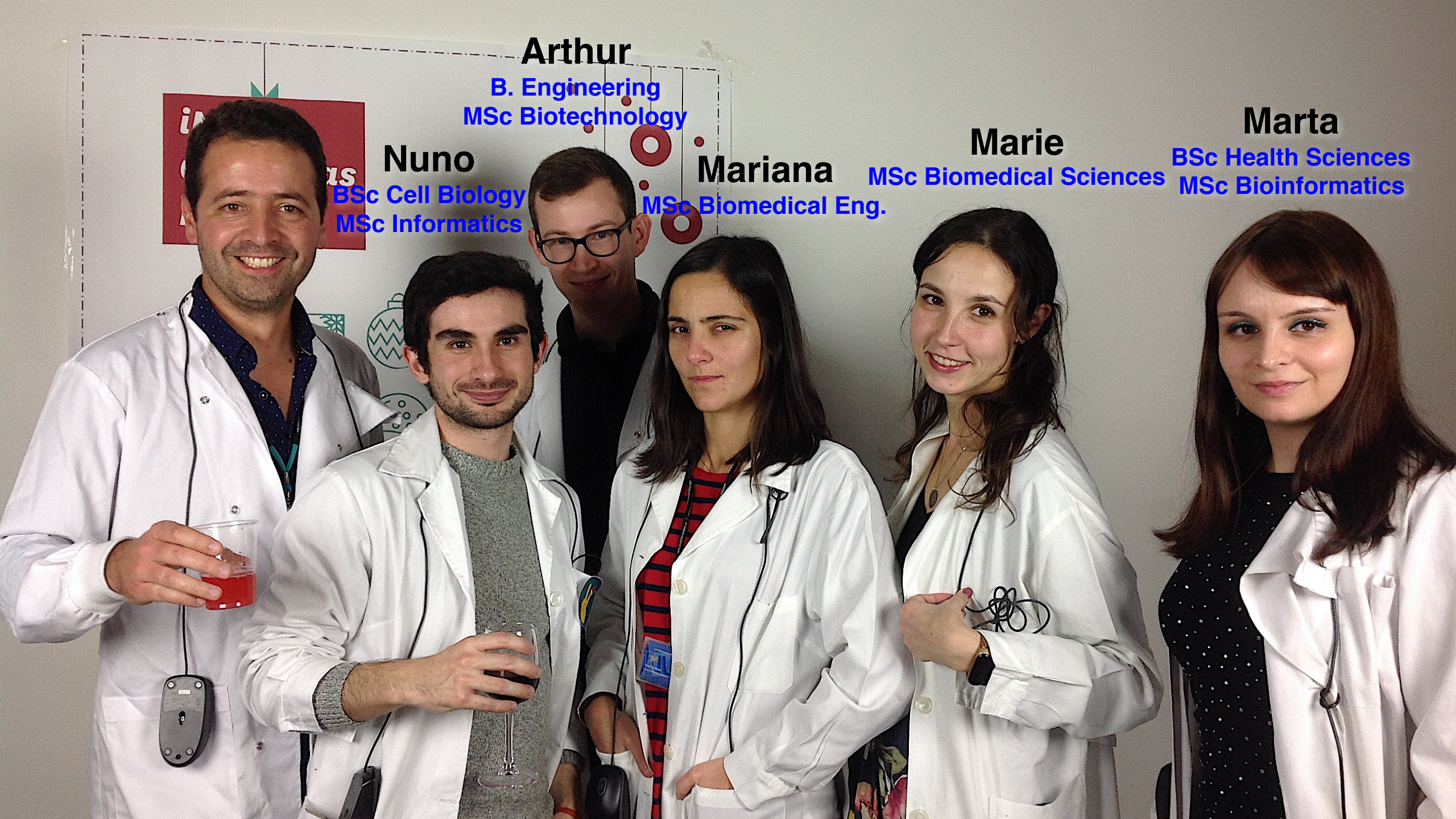
Instituto de Medicina Molecular | João Lobo Antunes

nmorais@medicina.ulisboa.pt

http://imm.medicina.ulisboa.pt/group/distrans/

https://twitter.com/NMoraisLab

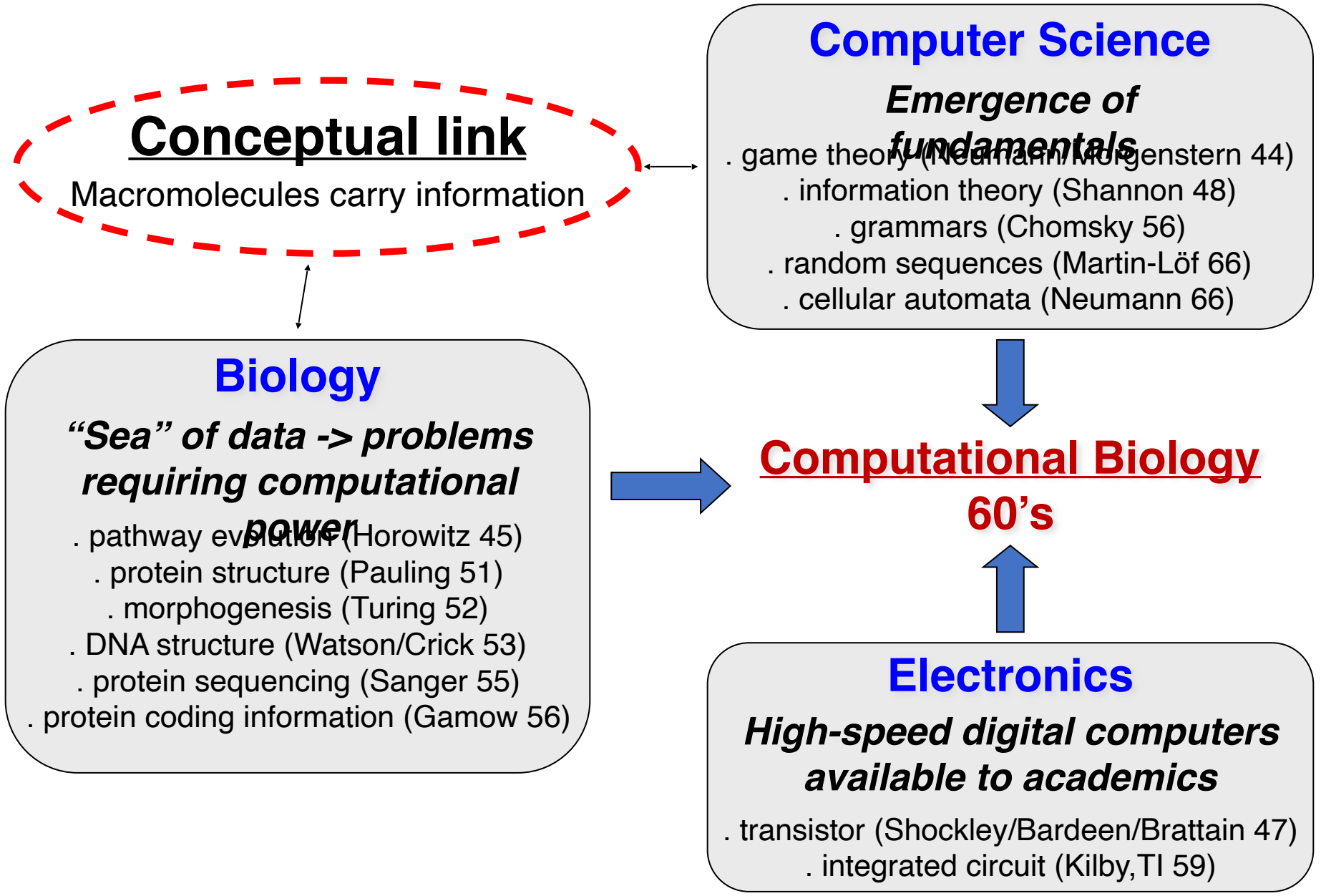

Instituto
de Medicina
Molecular
João
Lobo
Antunes


Arthur
B. Engineering
MSc Biotechnology
Nuno
BSc Cell Biology
MSc Informatics
Mariana
MSc Biomedical Eng.
Marie
MSc Biomedical Sciences
Marta
BSc Health Sciences
MSc Bioinformatics

Outline

1. Biological macromolecules as carriers of biomedically relevant data
2. What we do: Disease Transcriptomics
3. Why biology needs more data scientists

1. Biological macromolecules as carriers of biomedically relevant data

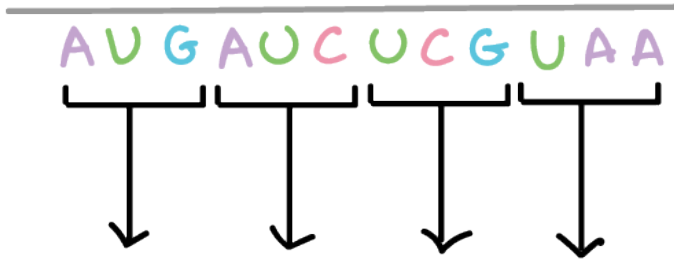

Conceptual link
Macromolecules carry information
Computer Science
Emergence of fundamentals
. game theory (Neumann/Morgenstern 44)
. information theory (Shannon 48)
. grammars (Chomsky 56)
. random sequences (Martin-Löf 66)
. cellular automata (Neumann 66)
Biology
"Sea" of data -> problems requiring computational power
. pathway evolution (Horowitz 45)
. protein structure (Pauling 51)
. morphogenesis (Turing 52)
. DNA structure (Watson/Crick 53)
. protein sequencing (Sanger 55)
. protein coding information (Gamow 56)
Computational Biology
60's
Electronics
High-speed digital computers available to academics
. transistor (Shockley/Bardeen/Brattain 47)
. integrated circuit (Kilby,TI 59)

THE CENTRAL DOGMA

DNA

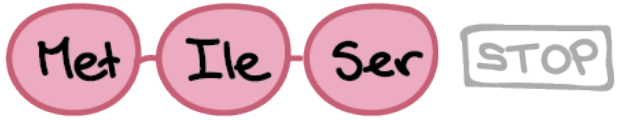
TRANSCRIPTION

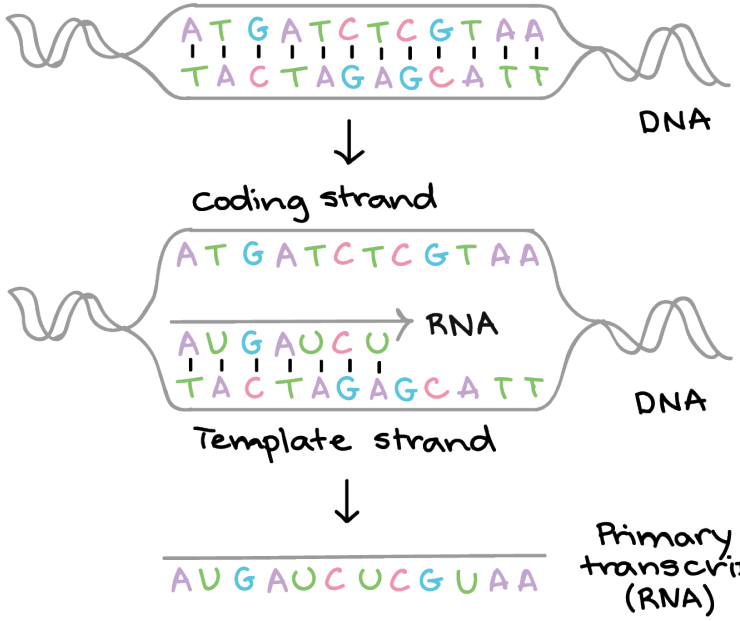
mRNA

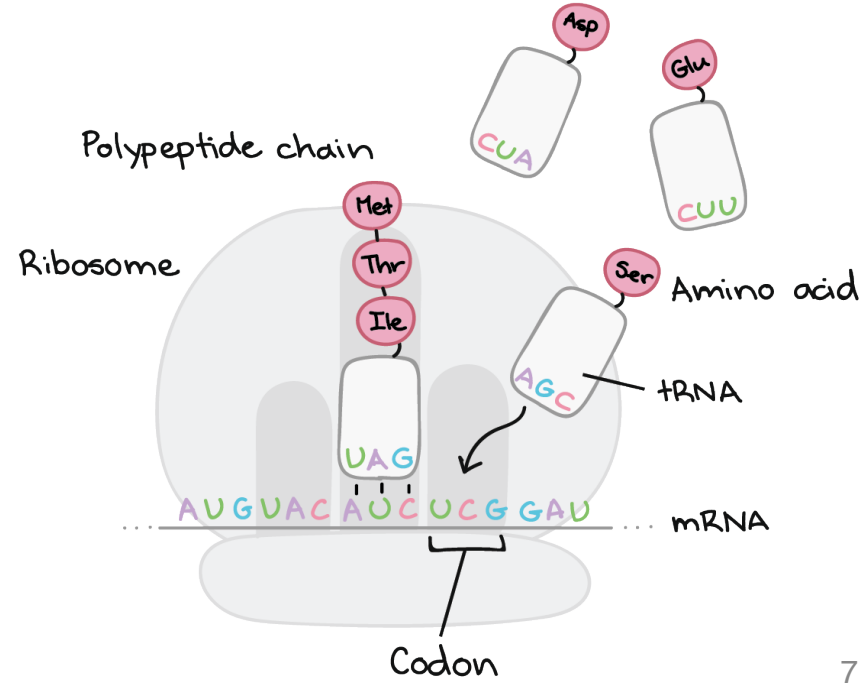


TRANSLATION

Polypeptide

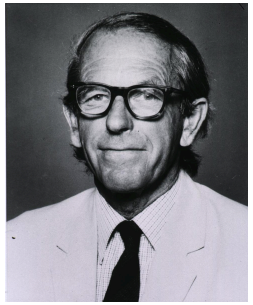






How to read DNA?

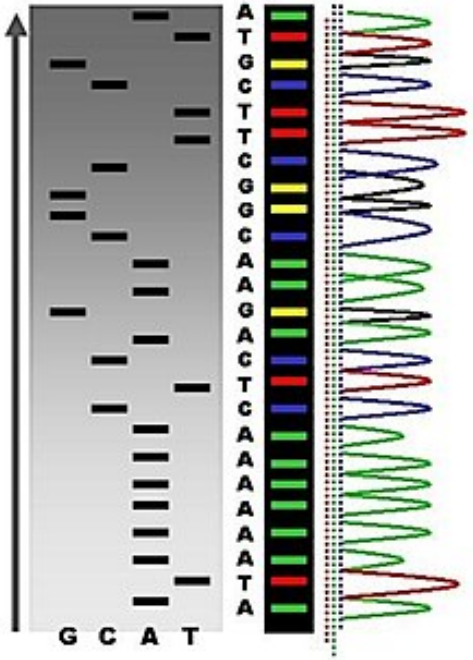
Fred Sanger
(1918-2013)



Chemistry 1958
Chemistry 1980

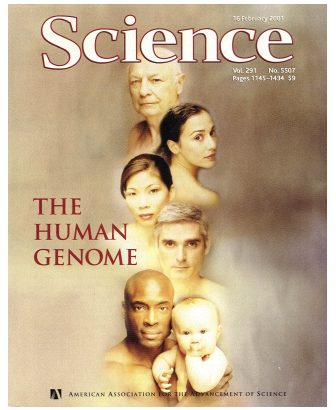
1977: Sanger sequencing

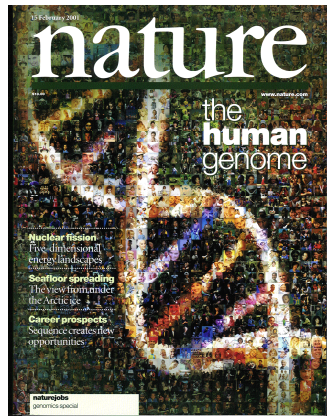
- *selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication*
- state-of-the-art for >30 years

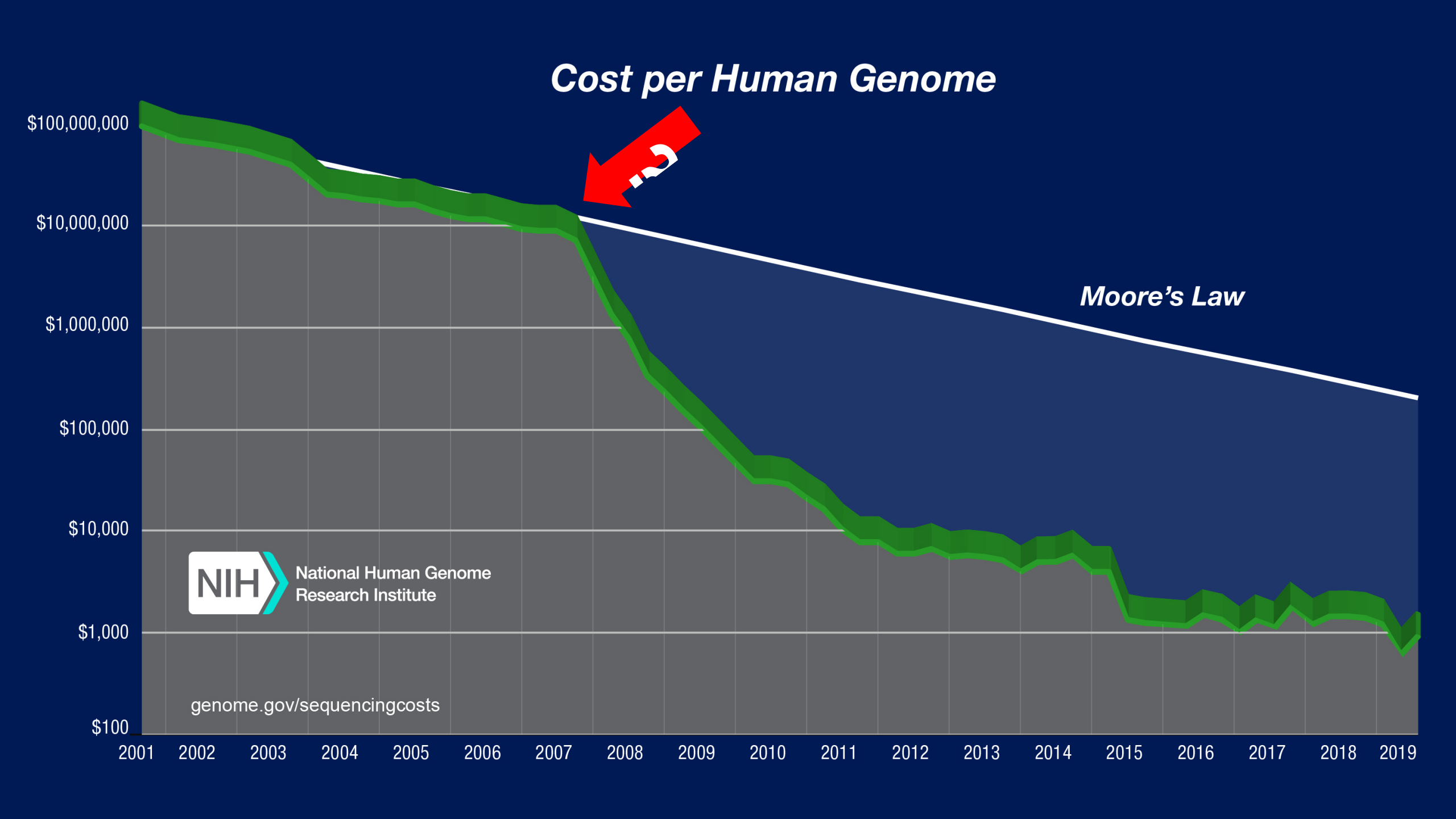


2001: first reference human genome

- ~$1.000.000.000
- 10 years

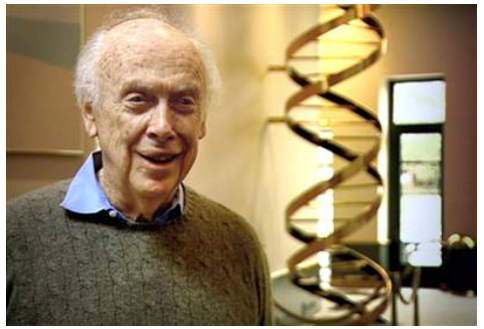





Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Individual genomics

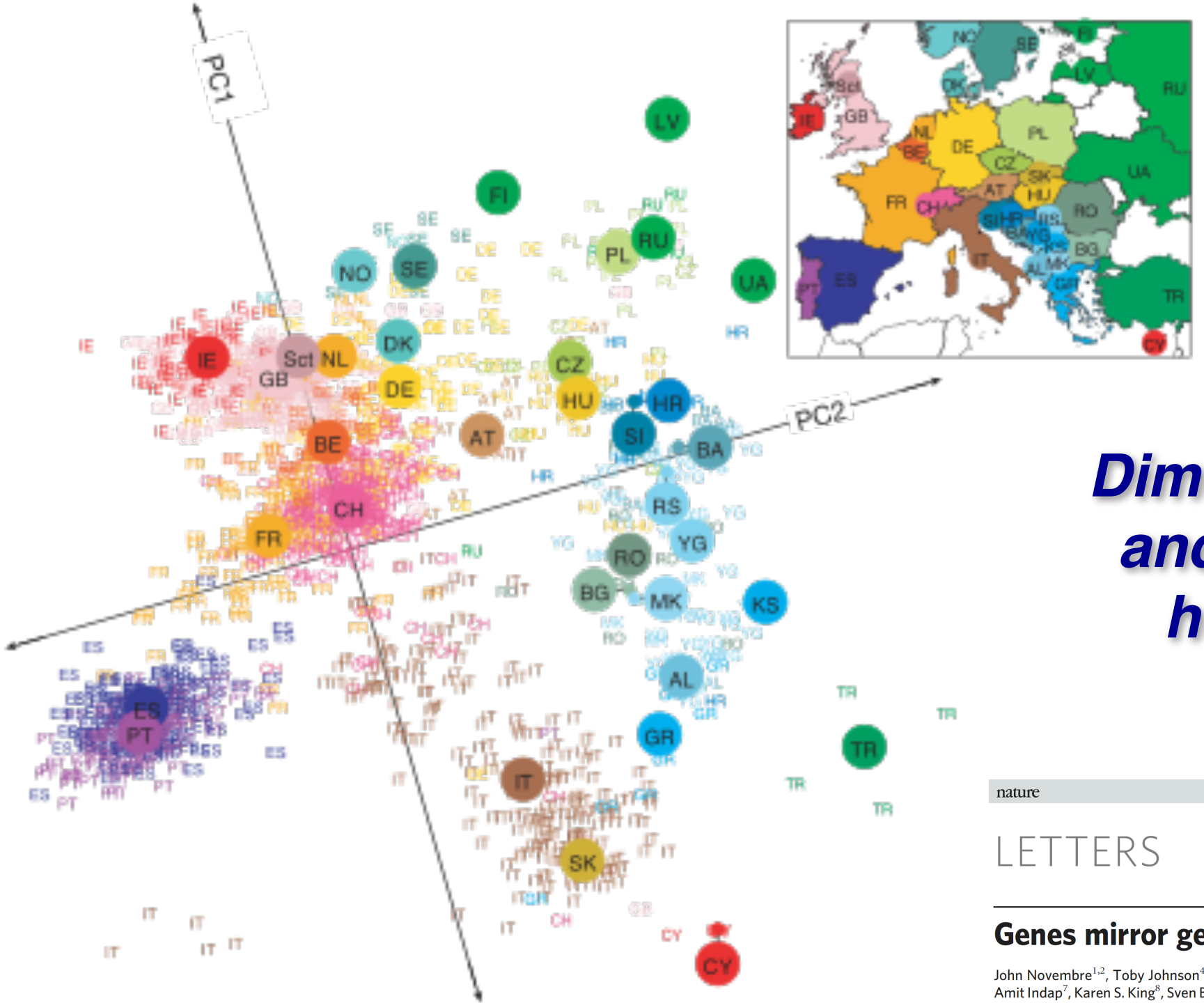
2007: Craig Venter's genome; ~10 M$ and 1 year with the
new high-throughput sequencing technologies



2008: James Watson's genome;
~$1.500.000 and 4 months

2018: BROAD INSTITUTE sequences its 100Kth human genome, global total approaching 1M

Currently:
> 1Tb in <4 days



Dimension reduction and smart plots can help us learn from genomic data!

nature Vol 456|6 November 2008|doi:10.1038/nature07331

LETTERS

Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

Challenge: how to interpret the human genome?

- Knowing the sequence of ~3x10^9 bases ("alphabet" of only 4: ACGT) is not enough to understand how cells work
- All the somatic cells of an individual have the same genome and are quite diverse (different organs)

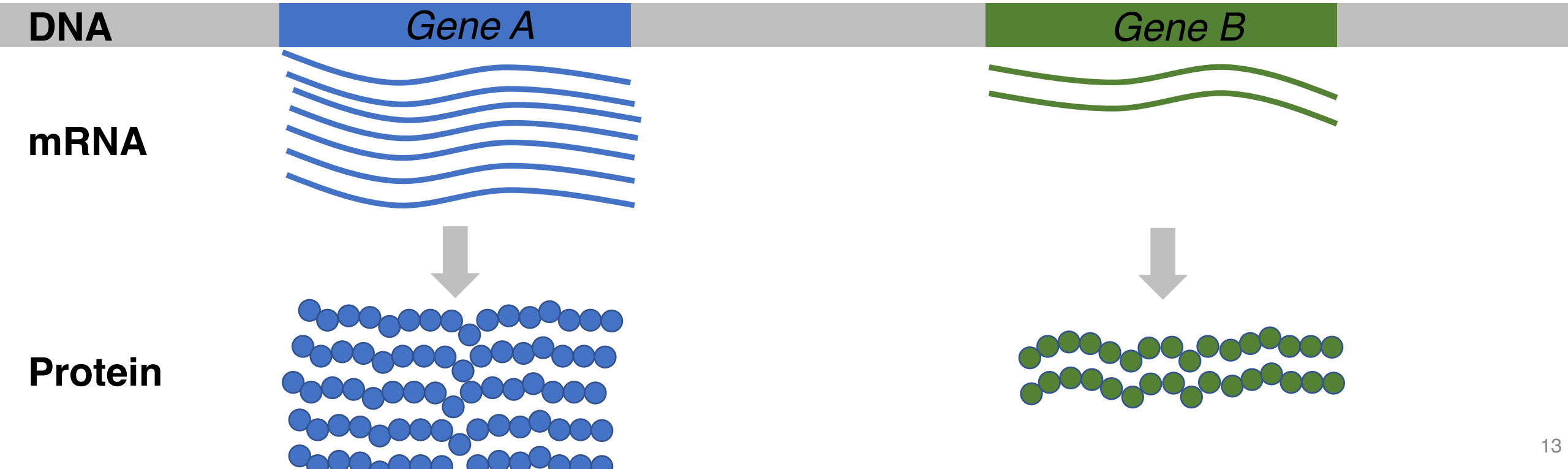
Post-genomic era: sequence → function

An approach: profiling the *expression* of all (1000s of) genes in parallel

Ultimate goal: measuring protein levels (possible but hard to do in a direct way)

Reasonable strategy: measuring levels of mRNA as surrogate for protein expression

sequence them (RNA-seq) and count th



How the data look like (breast cancer)

Gene expression (counts)

Biological samples (biopsies)

Genes

Gene	TCGA-A1-A0SB-01	TCGA-A1-A0SD-01	TCGA-A1-A0SE-01	TCGA-A1-A0SF-01	TCGA-A
A1BG	164	546	1341	836	
A1CF	0	0	0	1	
A2BP1	22	1	2	0	
A2LD1	127	331	498	526	
A2ML1	94	144	114	77	
A2M	102123	107181	101192	50316	
A4GALT	890	1409	1711	1142	
A4GNT	6	5	2	2	
AAA1	0	0	1	0	
AAAS	2139	2219	4294	3052	
AACSL	2930	3	4	1	
AACS	6533	3102	5271	1248	
AADACL2	0	0	1	0	
AADACL3	0	0	0	0	
AADACL4	1	0	0	0	
AADAC	42	4	13	4	
AADAT	694	226	200	67	
AAGAB	2913	9069	10465	3366	
AAK1	3811	5545	6963	2341	
AAMP	6514	9517	18397	8569	
AANAT	9	2	11	0	
AARS2	1764	2193	3338	1463	
AARSD1	2354	1846	3391	2169	
AARS	10855	9400	16774	6233	
AASDHPPT	3234	2936	3766	3134	
AASDH	1299	1589	2548	902	
AASS	1587	1852	2014	852	
AATF	2870	5656	9522	4625	
AATK	317	312	736	169	
ABAT	1422	3956	12660	3237	
ABCA10	143	177	319	146	

Sample annotation

Biomedical attributes

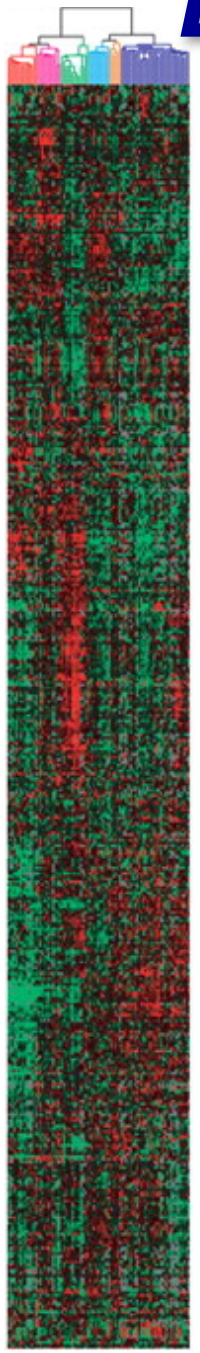
Biological samples

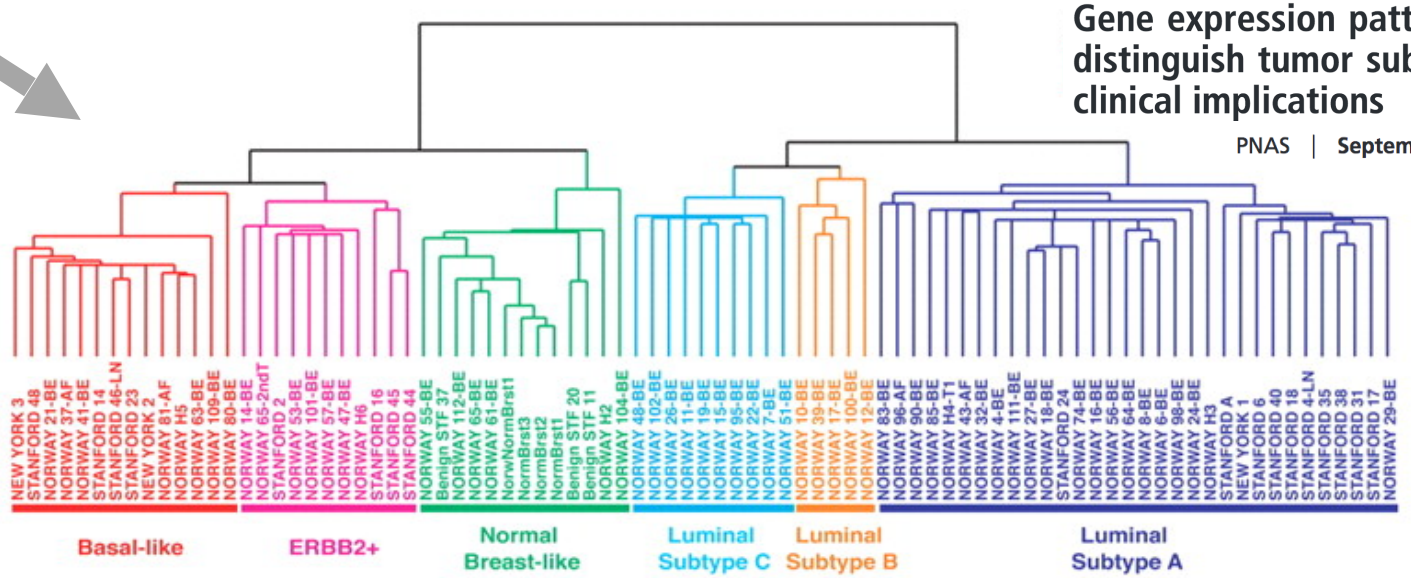
Patient ID	Gender	Age	Menopausal	Ethnicity	Year of diagnosis	Anatomic subdividic	Surgical proce
TCGA-A1-A0SB	female	70	post (prior bi	not hispanic or lat	2008	left	lumpectomy
TCGA-A1-A0SD	female	59	NA	not hispanic or lat	2005	left	lumpectomy
TCGA-A1-A0SE	female	56	pre (<6 mont	not hispanic or lat	2005	left upper outer qua	modified radi
TCGA-A1-A0SF	female	54	pre (<6 mont	not hispanic or lat	2006	left	modified radi
TCGA-A1-A0SG	female	61	post (prior bi	not hispanic or lat	2006	right	lumpectomy
TCGA-A1-A0SH	female	39	pre (<6 mont	not hispanic or lat	2006	left upper inner qua	lumpectomy
TCGA-A1-A0SI	female	52	NA	not hispanic or lat	2007	right	lumpectomy
TCGA-A1-A0SJ	female	39	NA	not hispanic or lat	2006	left	modified radi
TCGA-A1-A0SK	female	54	indeterminat	not hispanic or lat	2007	right upper outer qu	lumpectomy
TCGA-A1-A0SM	male	77	NA	not hispanic or lat	2007	left	modified radi
TCGA-A1-A0SN	female	50	post (prior bi	not hispanic or lat	2007	left	lumpectomy
TCGA-A1-A0SO	female	67	post (prior bi	not hispanic or lat	2007	right	modified radi
TCGA-A1-A0SP	female	40	NA	not hispanic or lat	2007	right	lumpectomy
TCGA-A1-A0SQ	female	45	pre (<6 mont	not hispanic or lat	2007	left	modified radi
TCGA-A2-A04N	female	66	post (prior bi	hispanic or latino	2002	right	lumpectomy
TCGA-A2-A04P	female	36	pre (<6 mont	NA	2003	left	lumpectomy
TCGA-A2-A04Q	female	48	post (prior bi	NA	2004	right upper outer qu	simple maste
TCGA-A2-A04R	female	36	pre (<6 mont	not hispanic or lat	2004	left upper inner qua	lumpectomy
TCGA-A2-A04T	female	62	post (prior bi	NA	2004	left upper inner qua	lumpectomy
TCGA-A2-A04U	female	47	peri (6-12 mo	not hispanic or lat	2004	right	simple maste
TCGA-A2-A04V	female	39	pre (<6 mont	not hispanic or lat	2005	right lower outer qu	simple maste
TCGA-A2-A04W	female	50	pre (<6 mont	not hispanic or lat	2005	right upper outer qu	modified radi
TCGA-A2-A04X	female	34	pre (<6 mont	NA	2006	left	simple maste
TCGA-A2-A04Y	female	53	post (prior bi	NA	2008	right upper outer qu	simple maste
TCGA-A2-A0CL	female	37	pre (<6 mont	not hispanic or lat	2006	right lower outer qu	modified radi
TCGA-A2-A0CM	female	40	pre (<6 mont	not hispanic or lat	2003	left	simple maste
TCGA-A2-A0CP	female	60	post (prior bi	not hispanic or lat	2003	right	lumpectomy
TCGA-A2-A0CQ	female	62	post (prior bi	not hispanic or lat	2003	right	lumpectomy
TCGA-A2-A0CS	female	73	post (prior bi	not hispanic or lat	2004	left	modified radi
TCGA-A2-A0CT	female	71	post (prior bi	not hispanic or lat	2005	right upper outer qu	simple maste
TCGA-A2-A0CU	female	73	post (prior bi	not hispanic or lat	2005	right lower outer qu	simple maste

Breast cancer clustering and classification

Samples

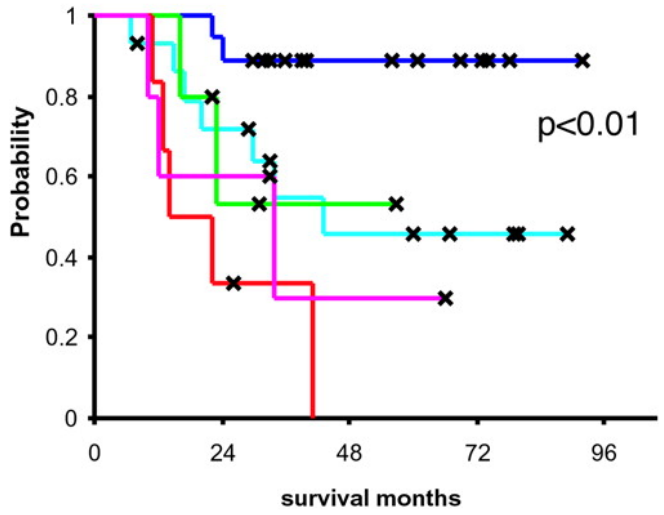
Genes

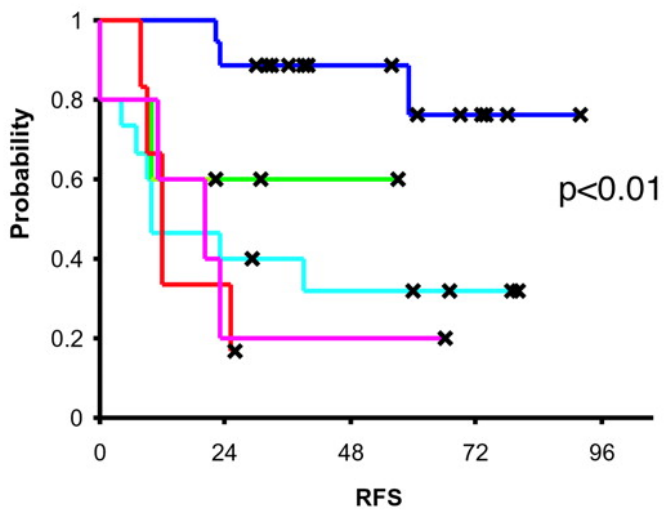




Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

PNAS | September 11, 2001 | vol. 98 | no. 19 | 10869–10874





X Censored, Lum A, Lum B+C, NorB-like, Basal, ERBB2+

Breast cancer molecular subtyping in the clinic

Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications

Therese Sørlie[a,b,c], Charles M. Perou[a,d], Robert Tibshirani[e], Turid Aas[f], Stephanie Geisler[g], Hilde Johnsen[b], Trevor Hastie[e], Michael B. Eisen[h], Matt van de Rijn[i], Stefanie S. Jeffrey[j], Thor Thorsen[k], Hanne Quist[l], John C. Matese[c], Patrick O. Brown[m], David Botstein[c], Per Eystein Lønning[g], and Anne-Lise Børresen-Dale[b,n]

PNAS | September 11, 2001 | vol. 98 | no. 19 | 10869–10874

Predictive Biomarkers and Personalized Medicine

Clinical Cancer Research

A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen

Stephen K. Chia[1], Vivien H. Bramwell[3], Dongsheng Tu[4], Lois E. Shepherd[4], Shan Jiang[4], Tammi Vickery[6], Elaine Mardis[6], Samuel Leung[2], Karen Ung[2], Kathleen I. Pritchard[5], Joel S. Parker[7], Philip S. Bernard[8], Charles M. Perou[7], Matthew J. Ellis[6], and Torsten O. Nielsen[2]

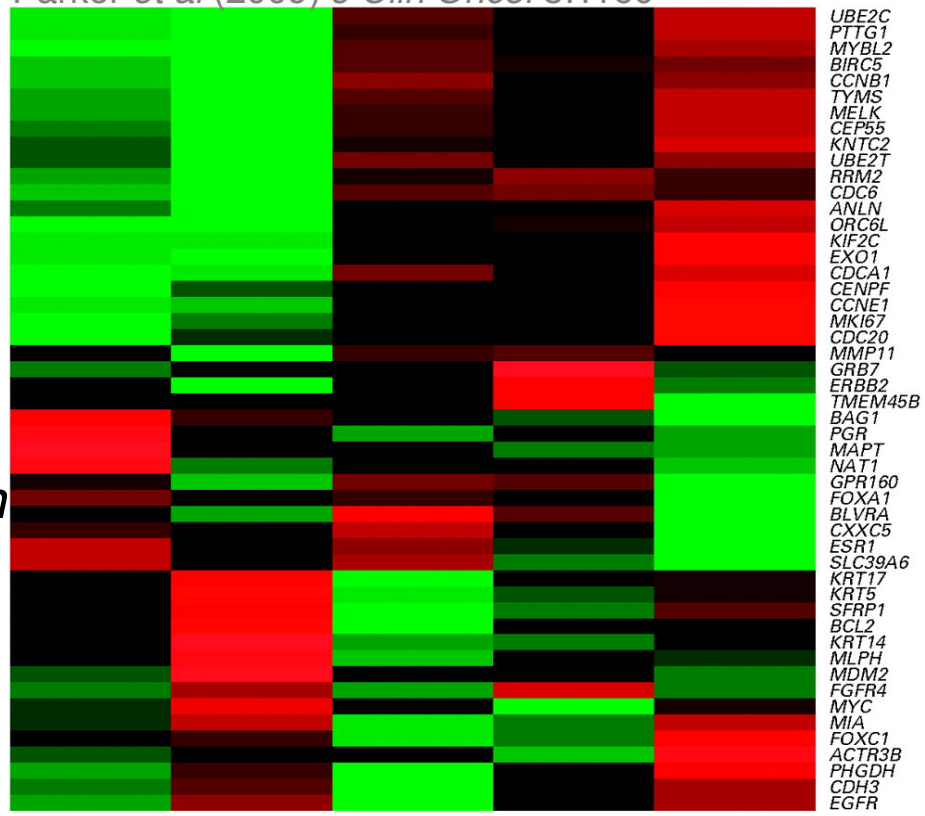
PAM50 (PROSIGNA)

PAM50 (Prosigna®) is a tumor profiling test that helps determine the benefit of using chemotherapy in addition to hormone therapy for some estrogen receptor-positive (ER-positive), HER2-negative breast cancers.

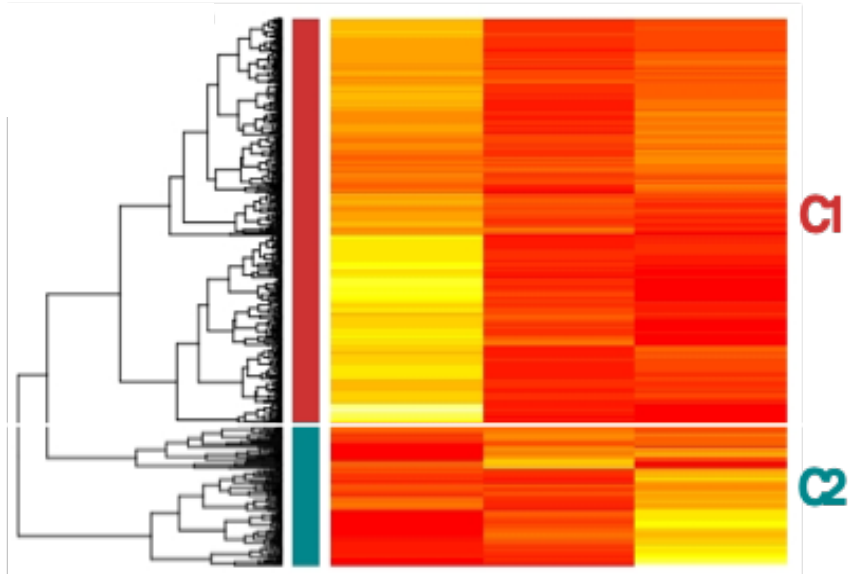
Patient

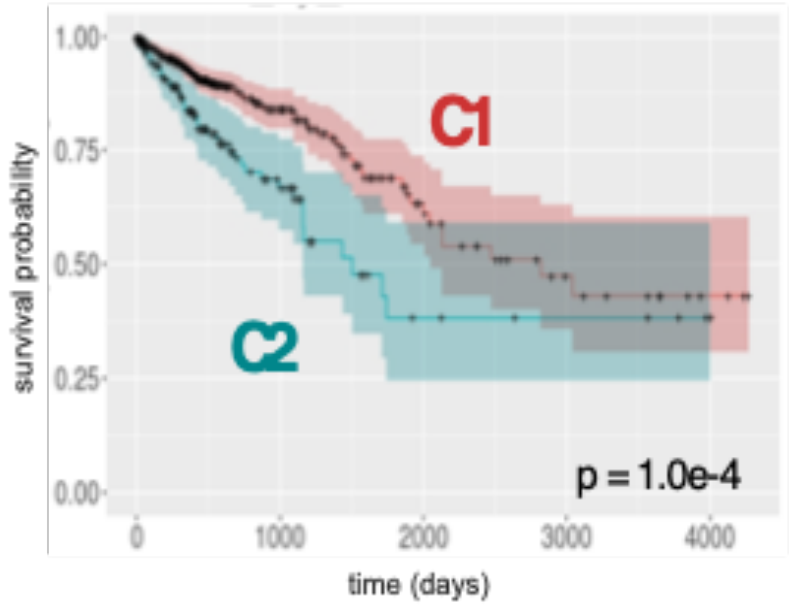
Classification

Parker *et al* (2009) *J Clin Oncol* 8:1160



2. What we do: Disease Transcriptomics



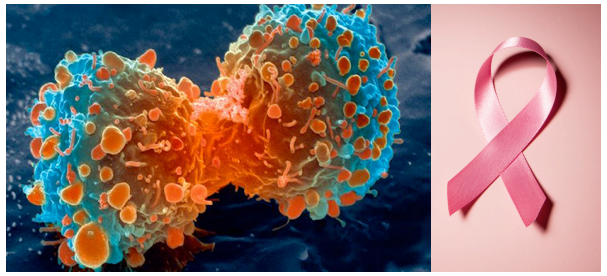


Goal

Understand how ageing-associated molecular (RNA) changes in human tissues increase proneness to disease

(Disease) models

Neurodegeneration



Cancer

Ageing

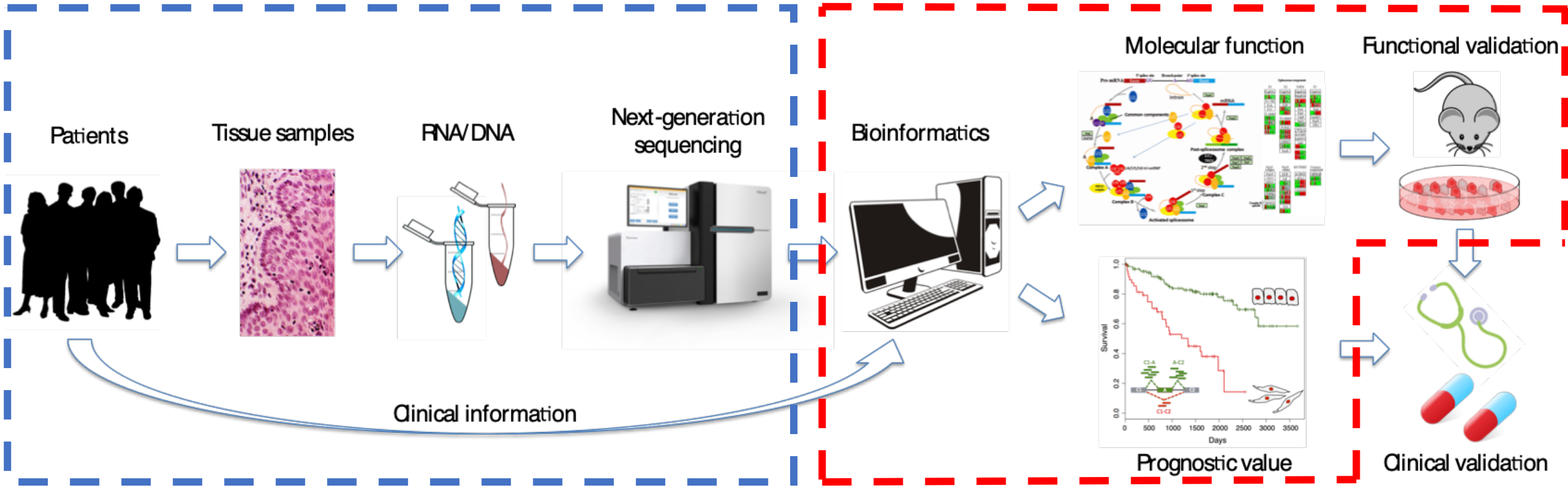
Why the transcriptome?

(the set of all RNA molecules in one cell or a population of cells)

- Early and accurately profileable measure of cells' response to stimuli

- Our expertise: computational biology + transcriptional regulation

- Timely: recent availability of data


Patients
Tissue samples
RNA/DNA
Next-generation sequencing
Bioinformatics
Molecular function
Functional validation
Clinical information
Survival
Days
Prognostic value
Clinical validation

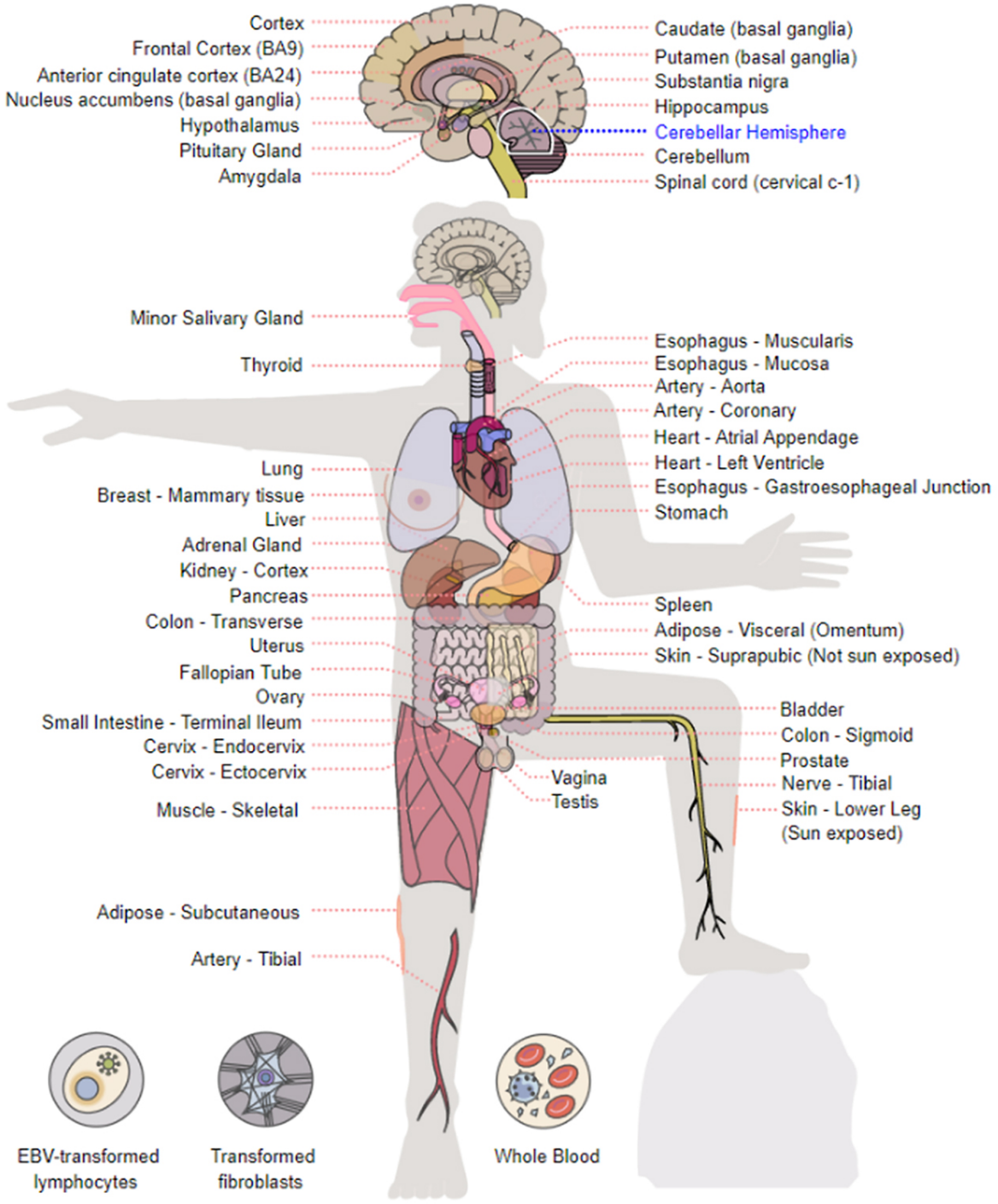
Data?

Big public data!

Genotype-Tissue Expression

RNA-seq from 53 human tissues

- >10.000 samples (100s per tissue)
- clinically annotated (age, gender, ethnicity, cause of death, diseases)



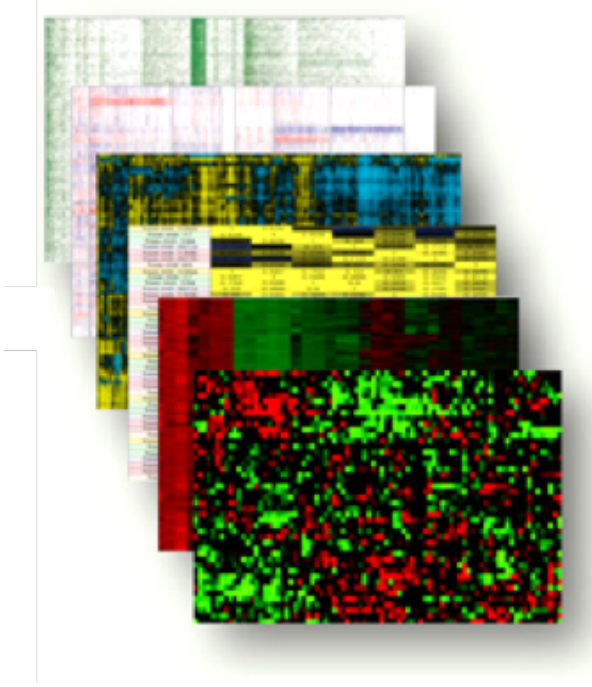
Casper *et al* (2018) *NAR*

TCGA

NIH THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute

Matched tumor & normal tissues from more than **11,000** patients, representing **33** cancer types.

- Clinical
- SNP6 CopyNum
- LowPass DNASeq CopyNum
- Mutation Annotation File
- methylation
- miR
- miRSeq
- mRNA
- mRNASeq
- raw Mutation Annotation File
- Reverse Phase Protein Array



Lamb *et al* (2006) *Science* 313:1929

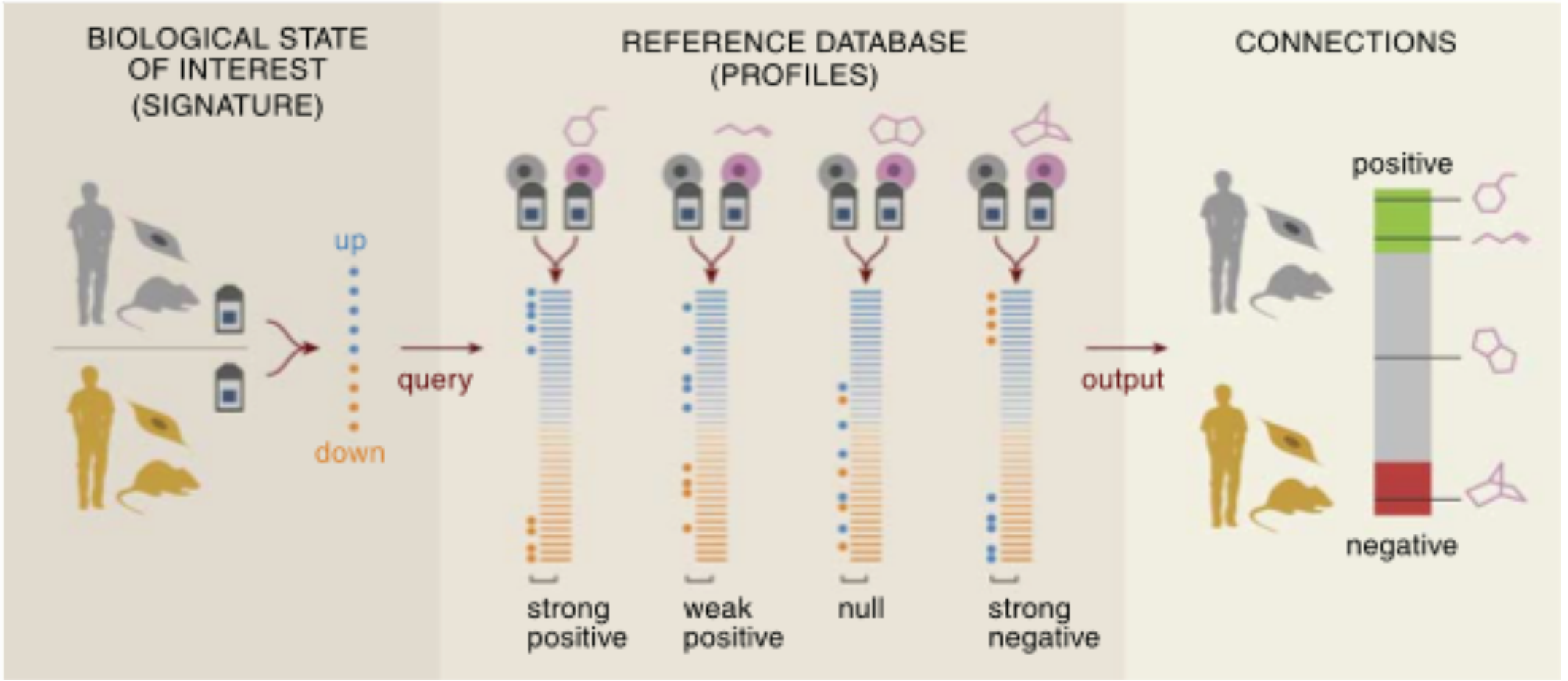
Subramanian *et al* (2017) *Cell* 171:1437

ConnectivityMap

BROAD INSTITUTE

Unravel biology with the world's largest perturbation-driven gene expression dataset.

Perturbations:
~20k small molecules
>5k genes



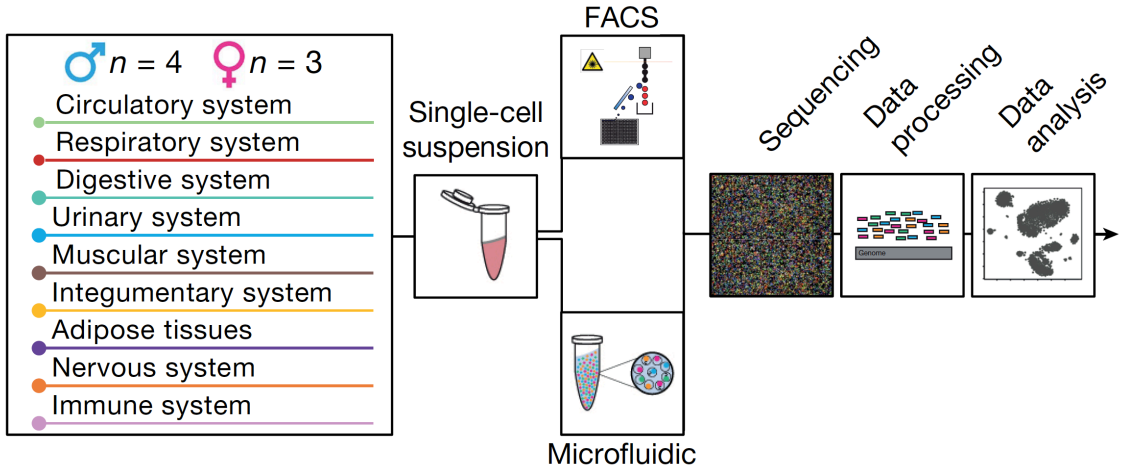
ARTICLE

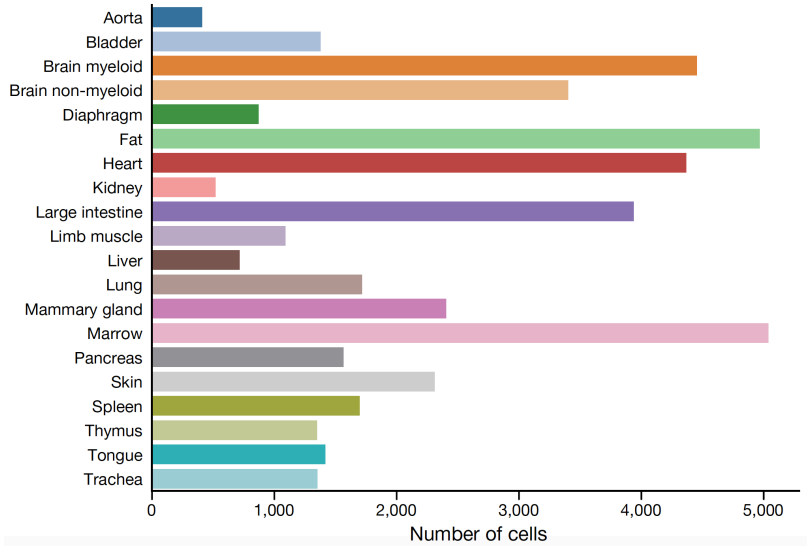
https://doi.org/10.1038/s41586-018-0590-4

18 OCTOBER 2018 | VOL 562 | NATURE | 367

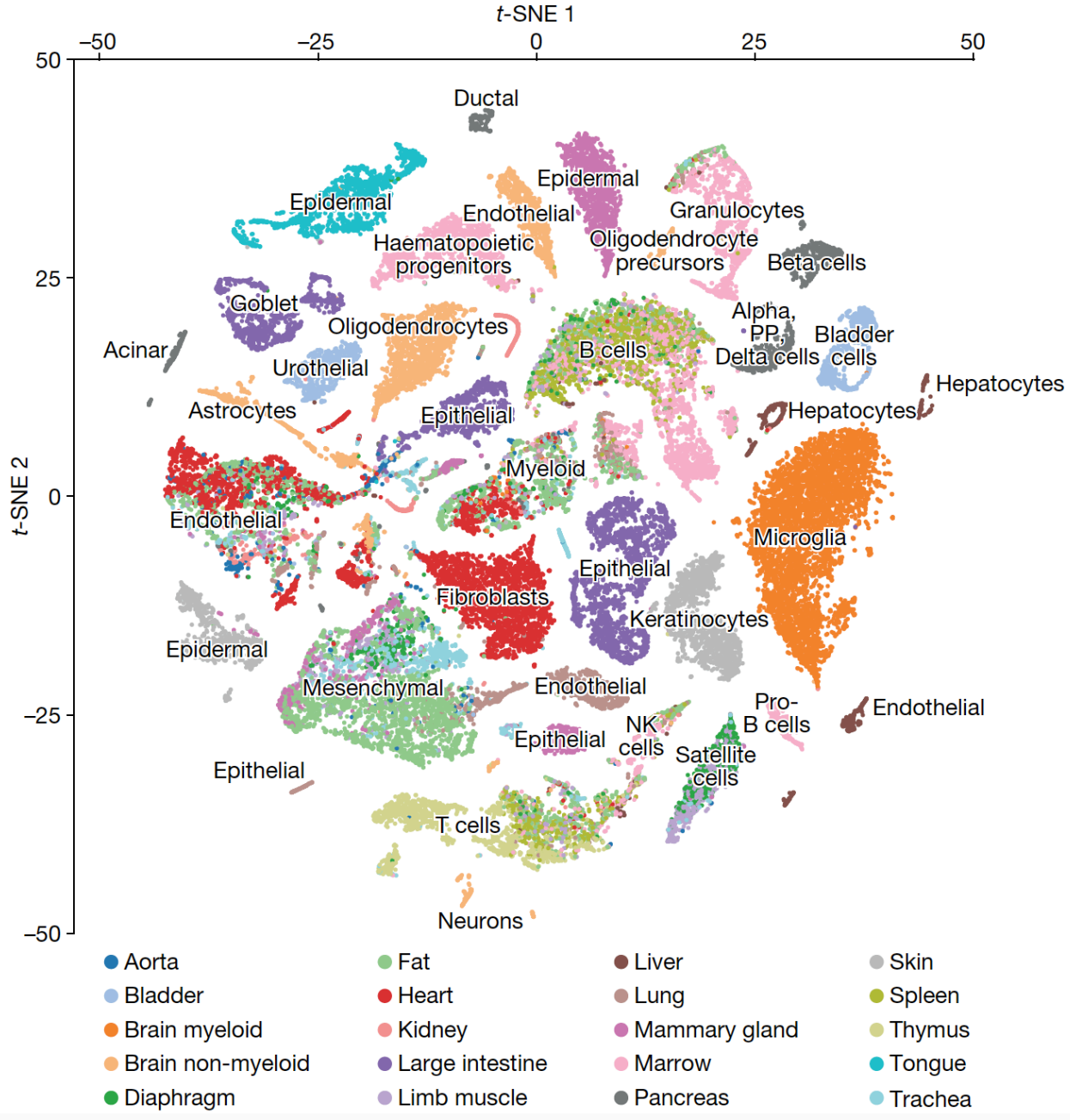
Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*

The Tabula Muris Consortium*

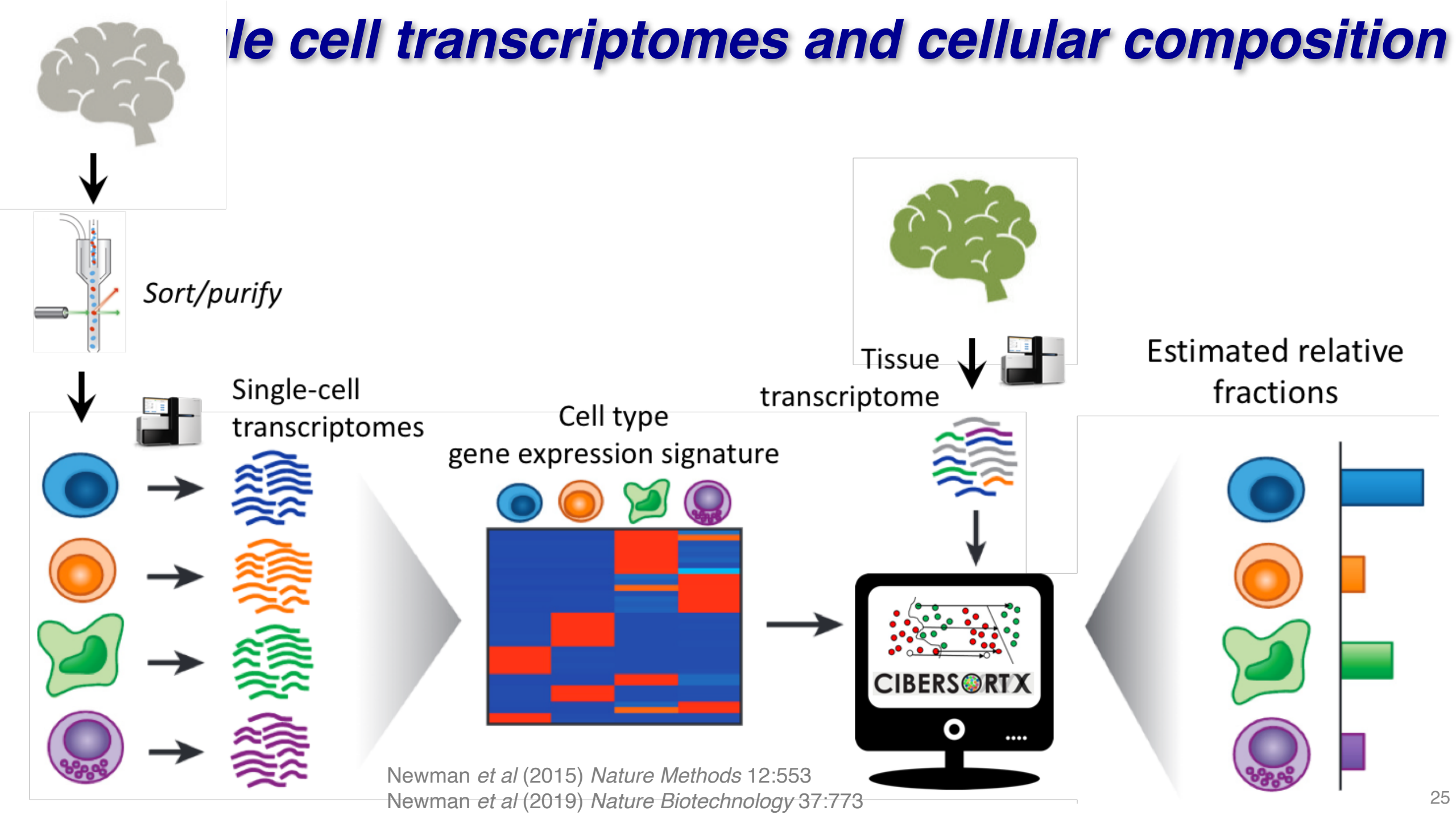




t-Distributed Stochastic Neighbor Embedding



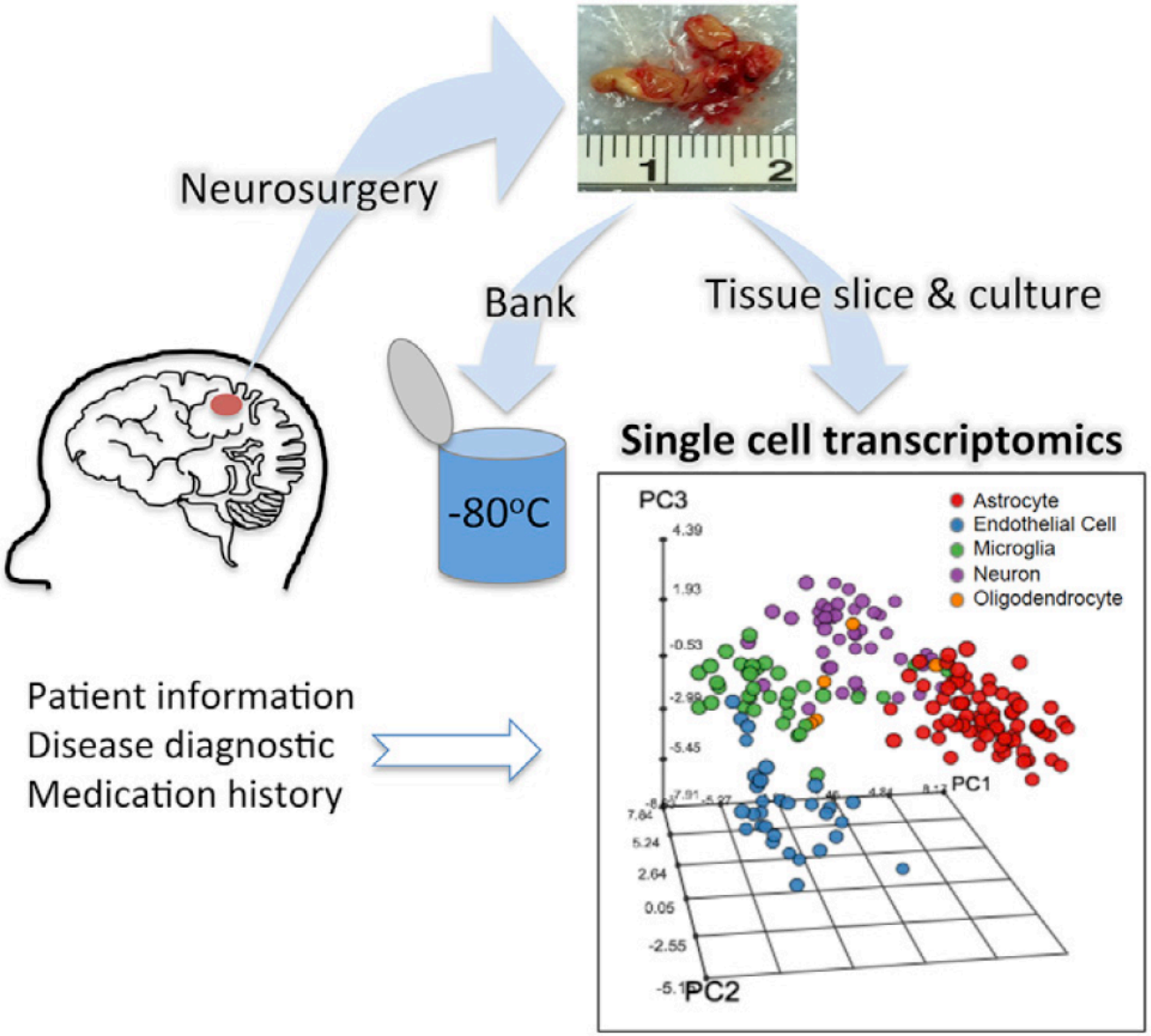
le cell transcriptomes and cellular composition



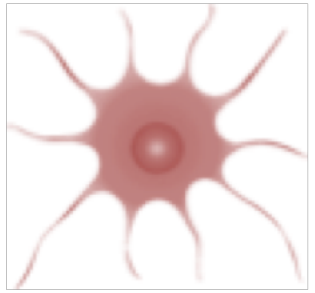
Newman *et al* (2015) *Nature Methods* 12:553
Newman *et al* (2019) *Nature Biotechnology* 37:773

Human brain cell transcriptomes

Darmanis *et al* (2015) *PNAS* 112:7285
Spaethling *et al* (2017) *Cell Reports* 18:791

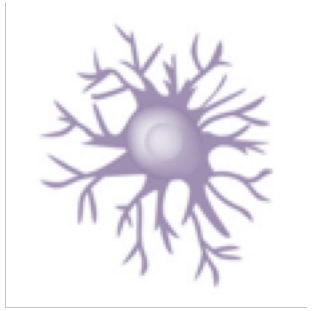


Neurons



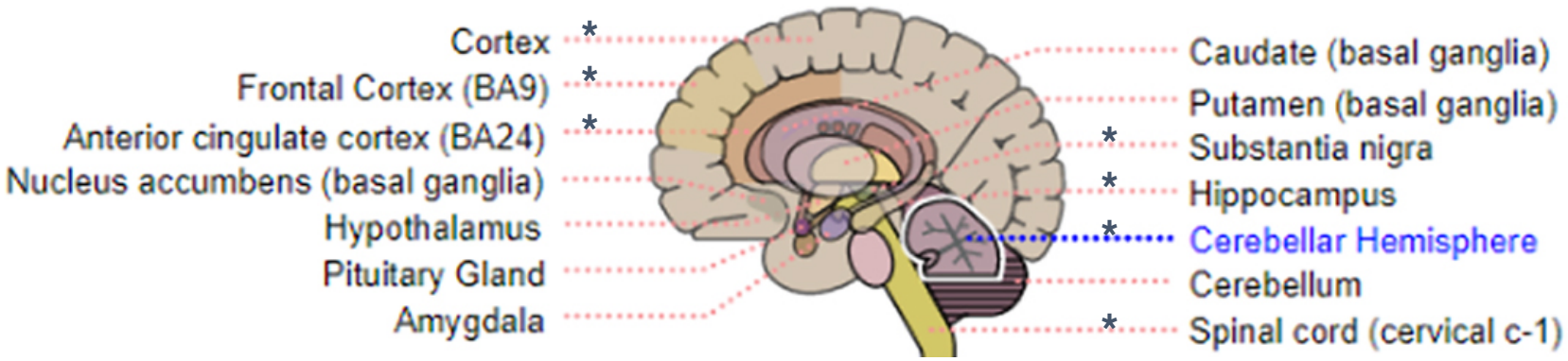
Astrocytes

Oligodendrocytes



Microglia

The ageing brain



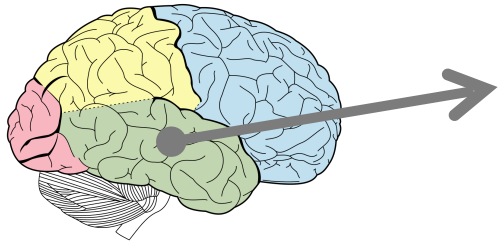
GTEx

	Cortex		Frontal cortex		Substantia nigra		Anterior cingulate cortex		Hippocampus		Cerebellar hemisphere		Spinal cord	
	rho	p-value	rho	p-value	rho	p-value	rho	p-value	rho	p-value	rho	p-value	rho	p-value
Neuron	-0.26	0.00087	-0.19	0.031	-0.23	0.032	-0.22	0.016	-0.33	0.00016	0.045	0.6	-0.29	0.0053
Astrocyte	0.24	0.0023	0.2	0.024	0.17	0.1	0.2	0.031	0.28	0.0014	0.0013	0.99	0.16	0.12
Oligodendrocyte	0.073	0.36	0.18	0.046	0.11	0.32	0.16	0.076	0.23	0.011	-0.31	0.00023	0.097	0.36
Microglia	0.11	0.18	0.082	0.36	0.091	0.4	0.25	0.0064	0.43	6.6E-07	0.03	0.73	0.3	0.004

positive correlation with age

negative correlation with age

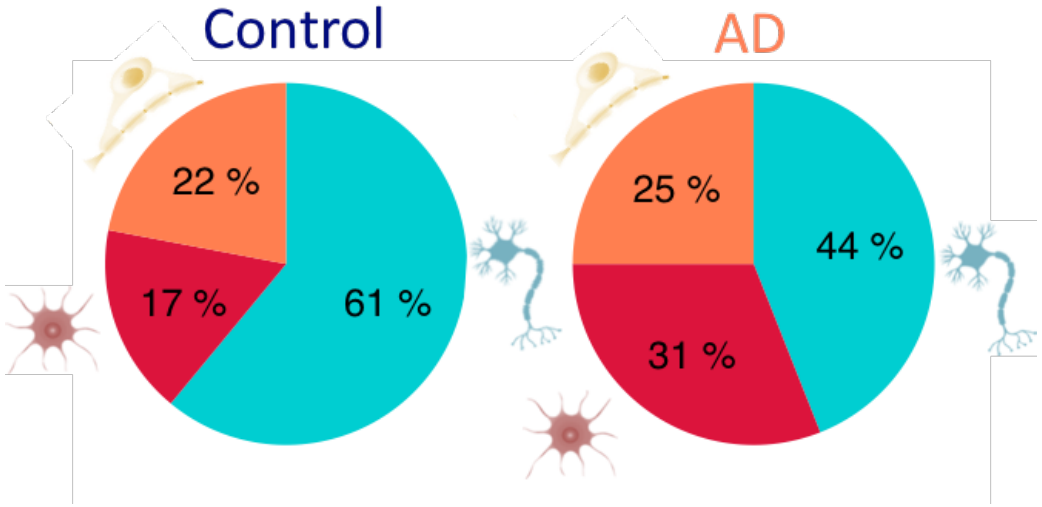
Loss of neurons in Alzheimer's disease (AD)

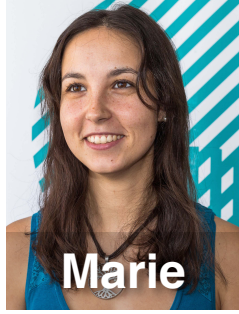


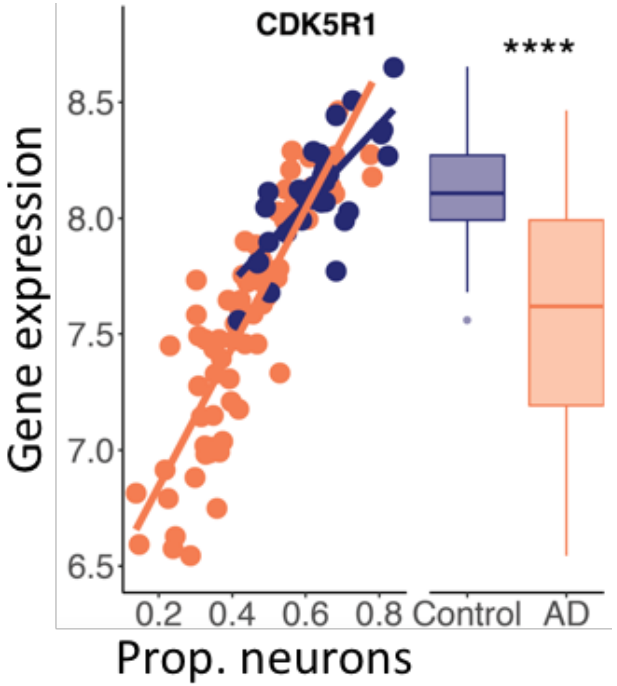
Temporal cortex RNA-seq
84 AD samples
54 Control samples

AMP-AD Knowledge Portal

Allen *et al* (2016) *Scientific Data* 3:160089



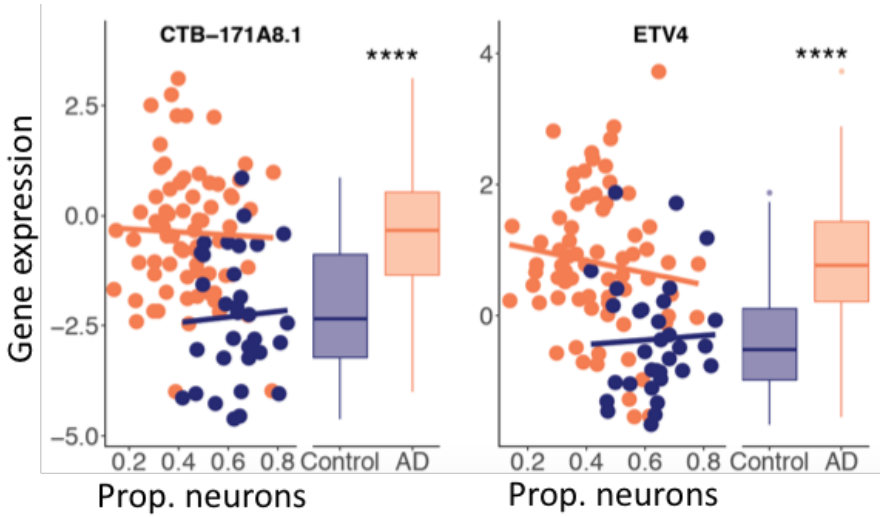




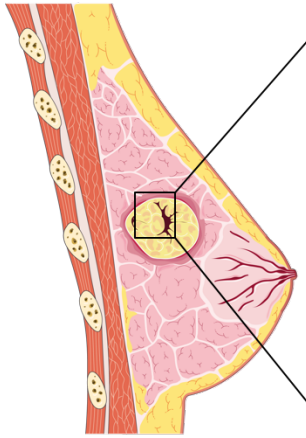
Mol Neurobiol
DOI 10.1007/s12035-016-0002-4

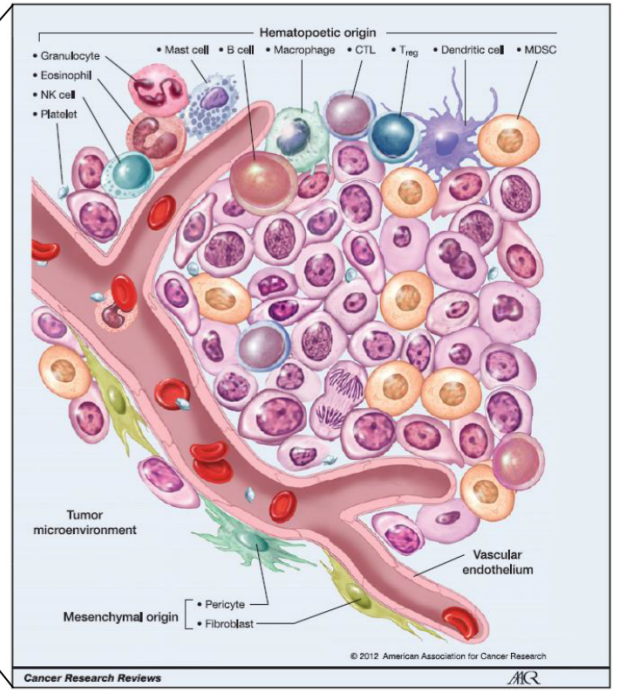
The miR-15/107 Family of microRNA Genes Regulates CDK5R1/p35 with Implications for Alzheimer's Disease Pathogenesis

Silvia Moncini[1] · Marta Lunghi[1] · Alice Valmadre[1] · Margherita Grasso[2] · Valerio Del Vescovo[2] · Paola Riva[1] · Michela Alessandra Denti[2,3] · Marco Venturin[1]



Immune infiltration in breast cancer





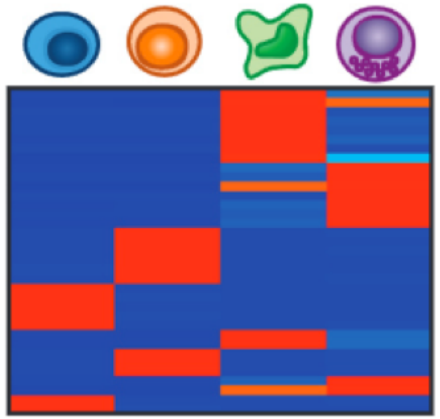
Kerkar and Restifo (2012) *Cancer Research 72:3125*

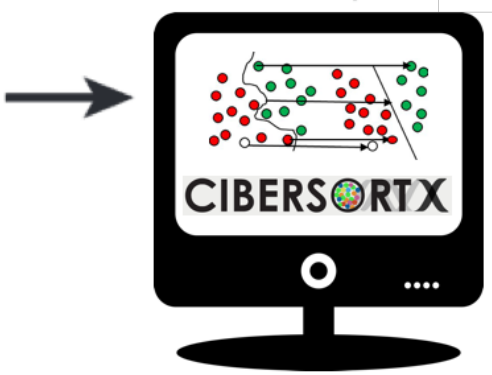
TCGA

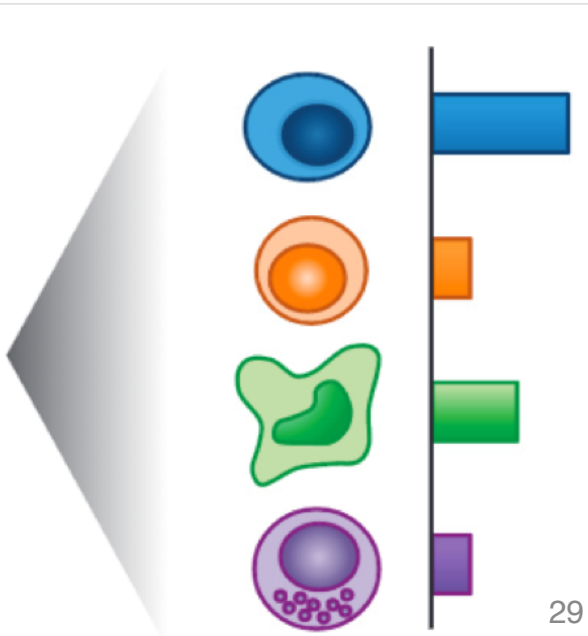
Breast tumour transcriptomes

Estimated relative fractions

LM22: leucocyte (immune cell) gene signature

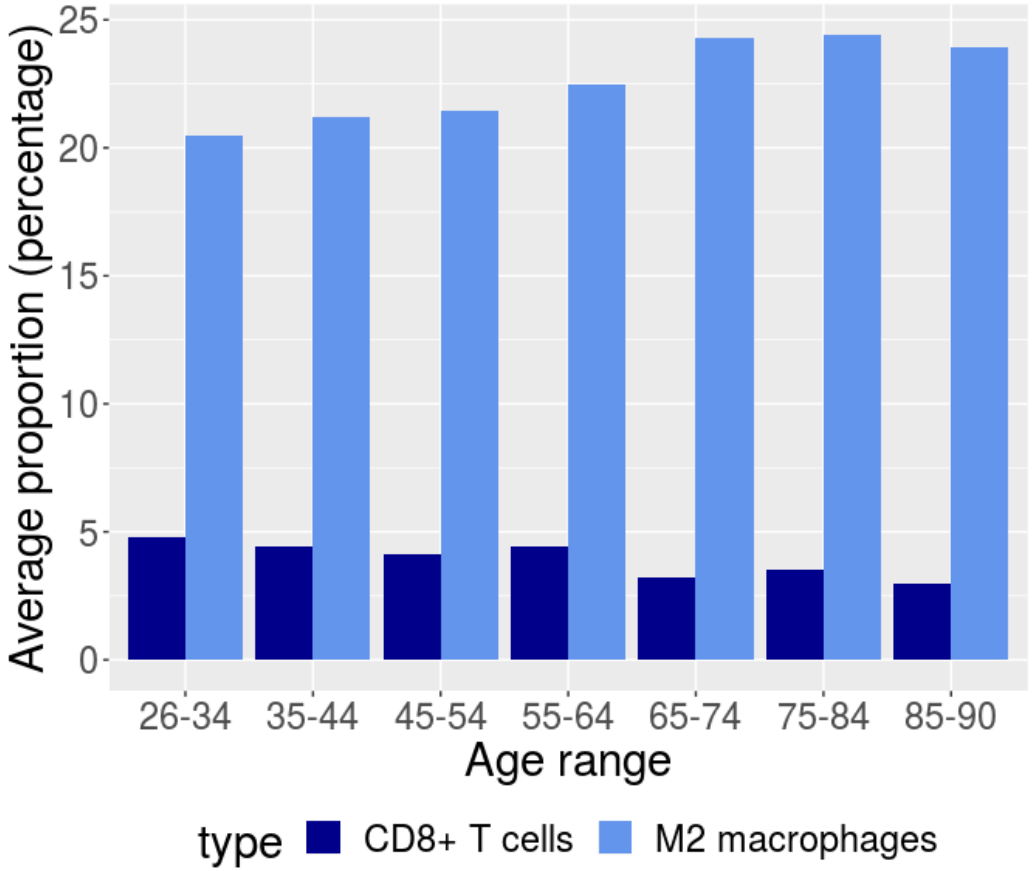




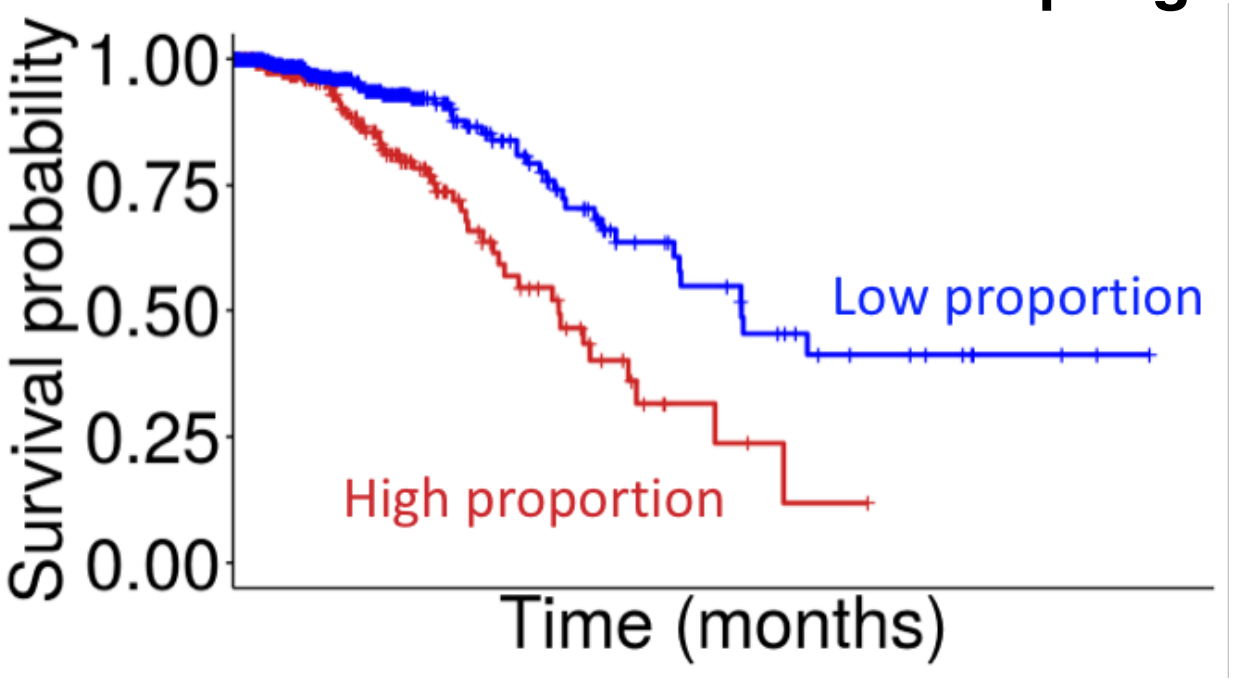


Immune infiltration and age in breast cancer

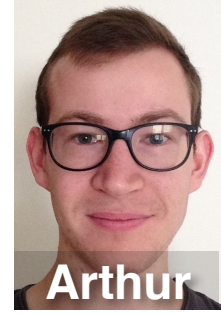
- **CD8+ T cells**: toxic to other cells, anti-tumoural
- **M2 macrophages**: tissue repair, pro-tumourigenic



M2 macrophages



How gene expression varies with age in human tissues

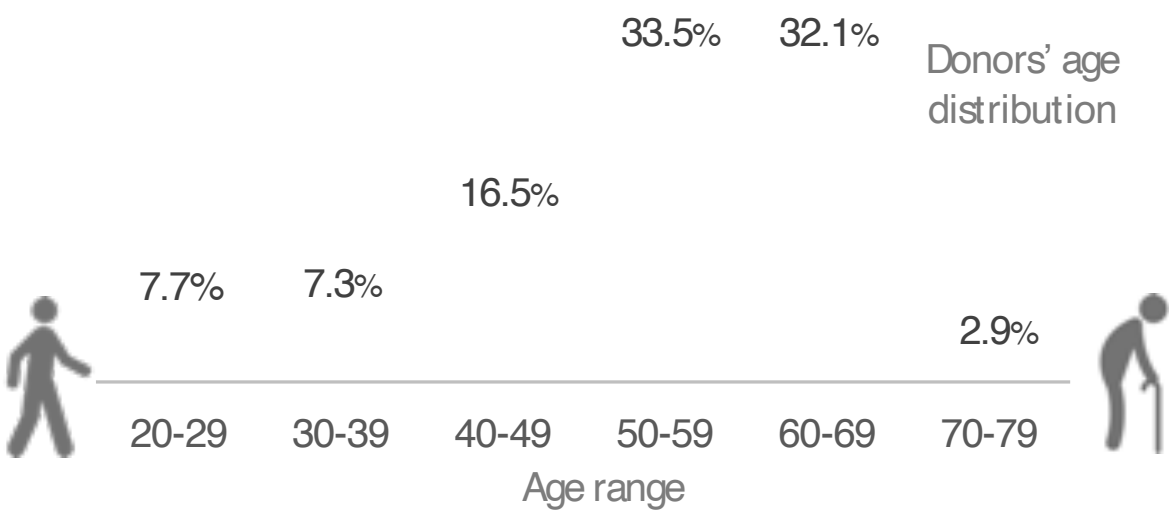


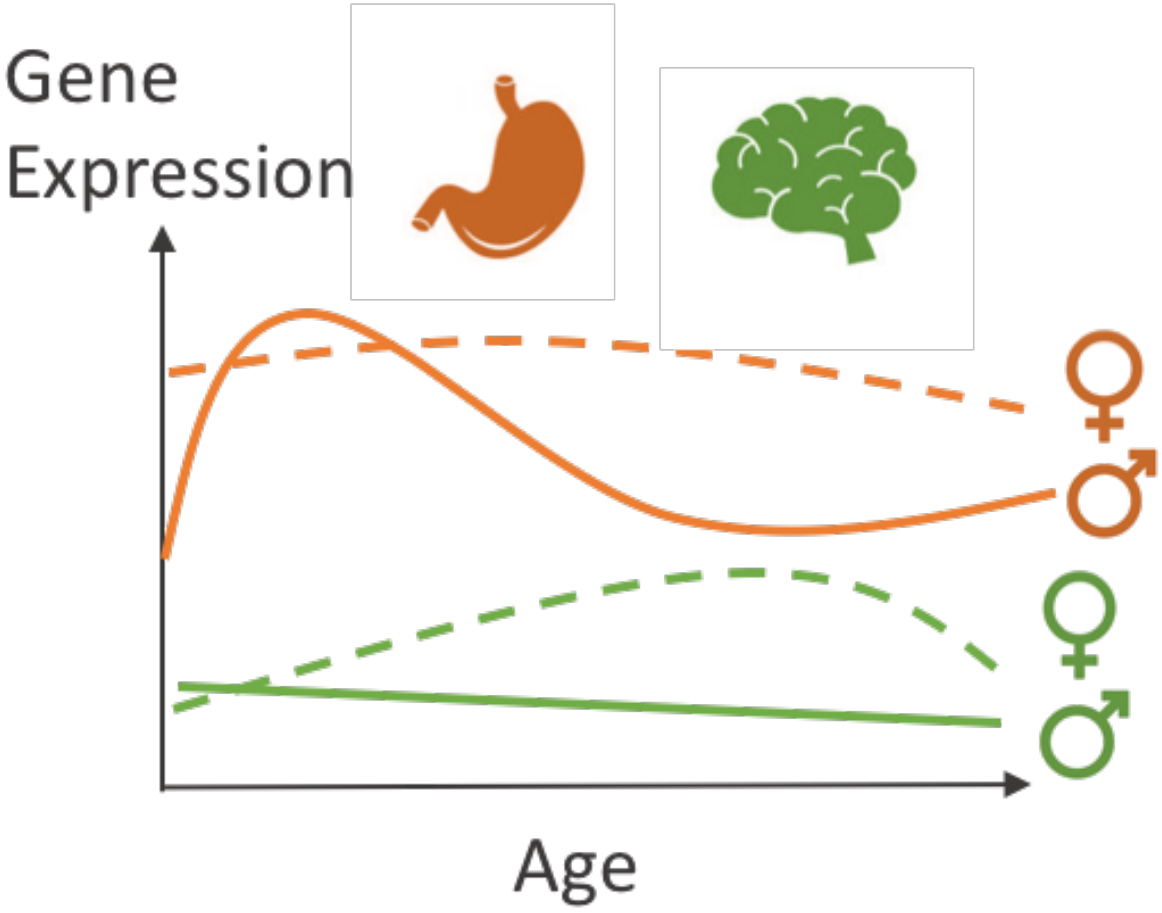
GTEx Portal

11 688 transcriptomes

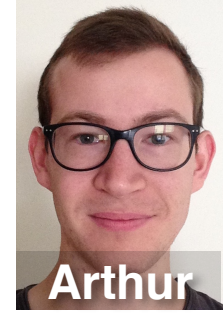
714 donors

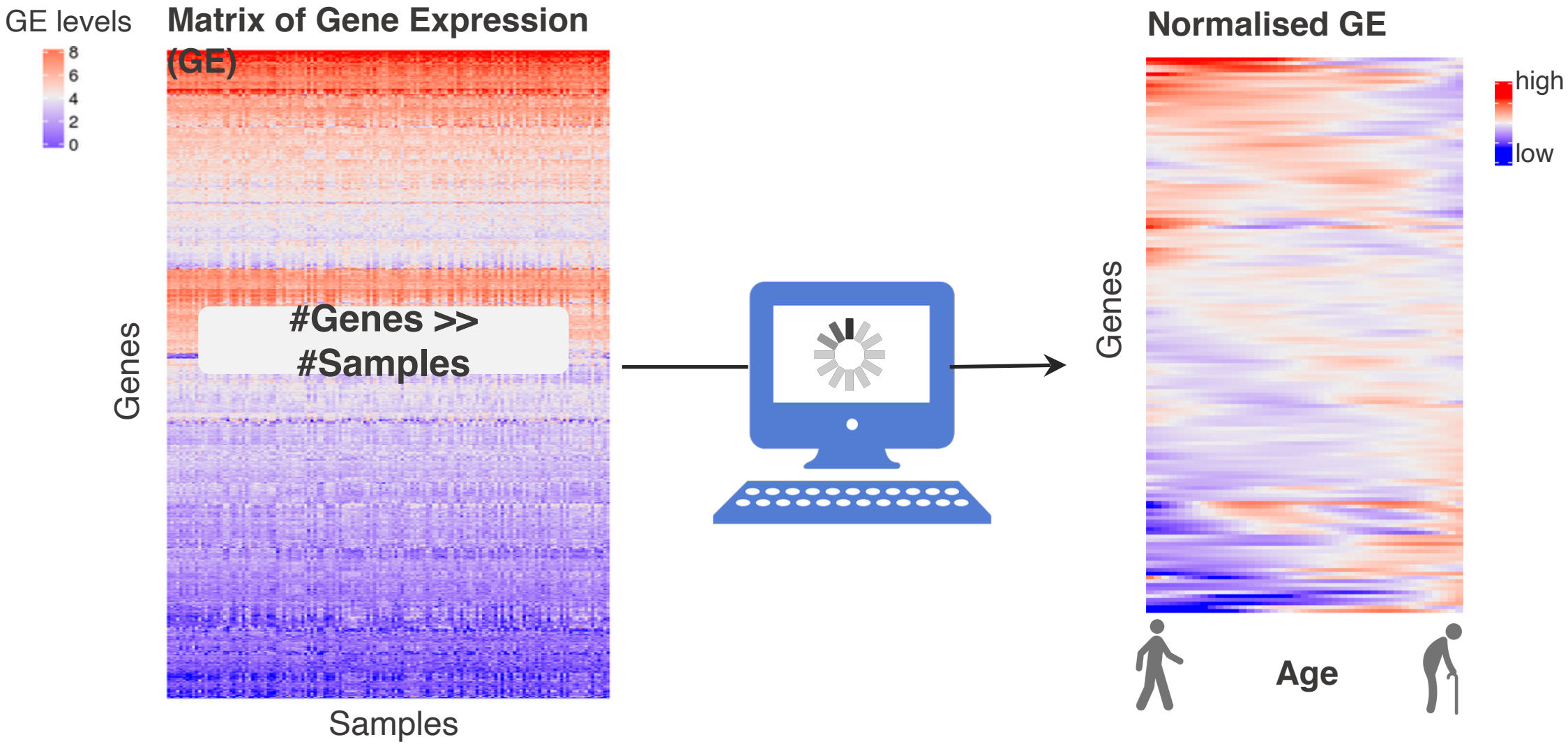
53 tissues



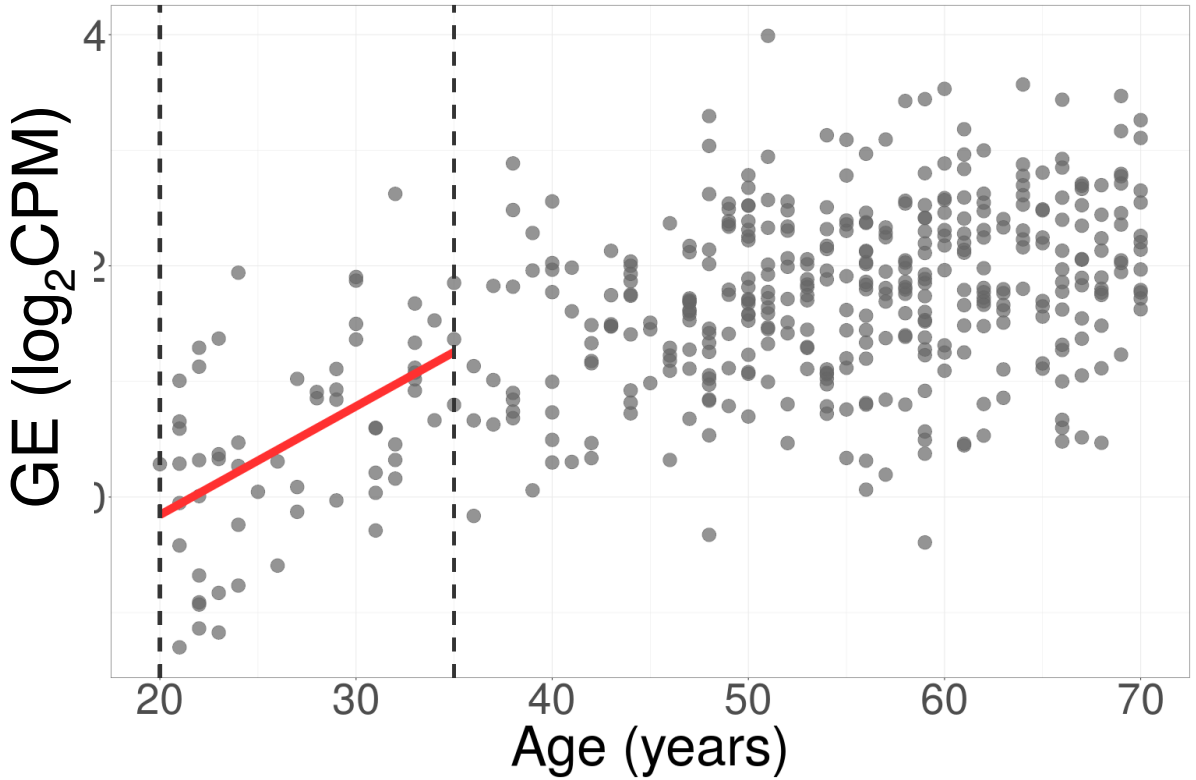


For each tissue/gender combination:

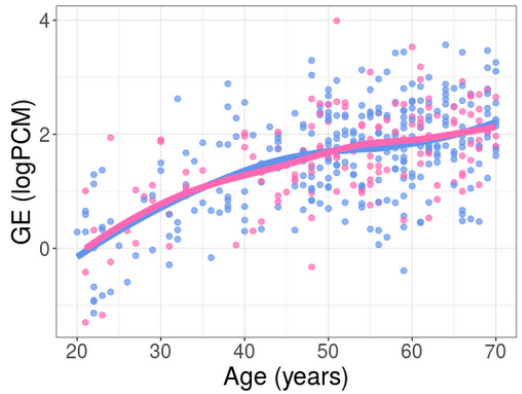




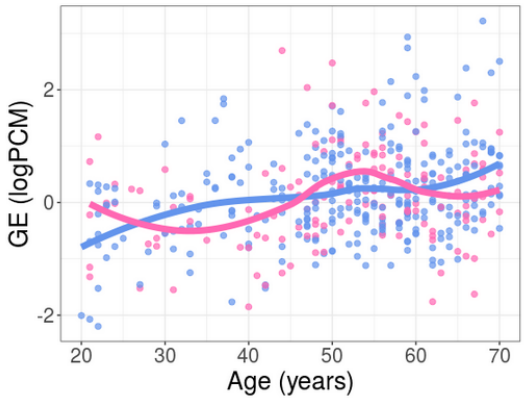
For each tissue/gender combination:



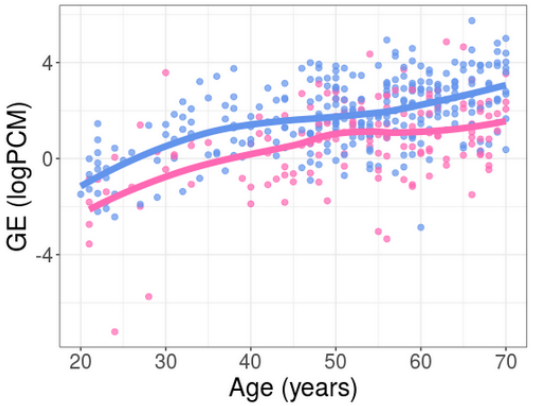
Gender differences?



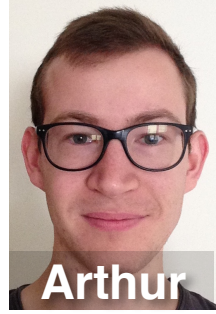
None



Trend

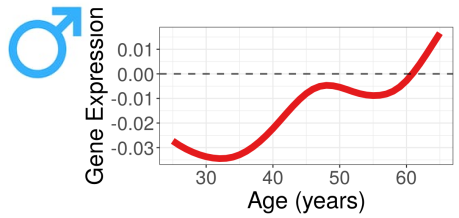


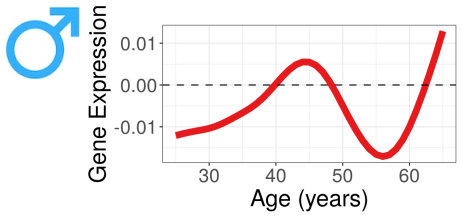
Magnitude

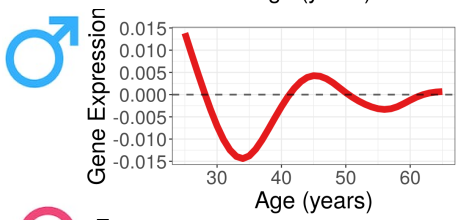


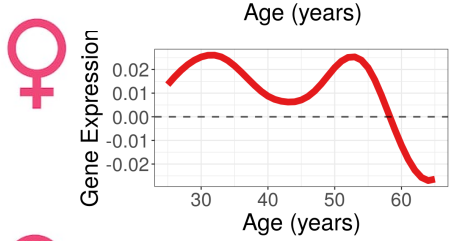
Tissue/gender-specific ageing

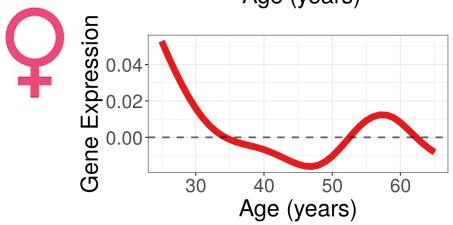
Heart
Atrial Appendage



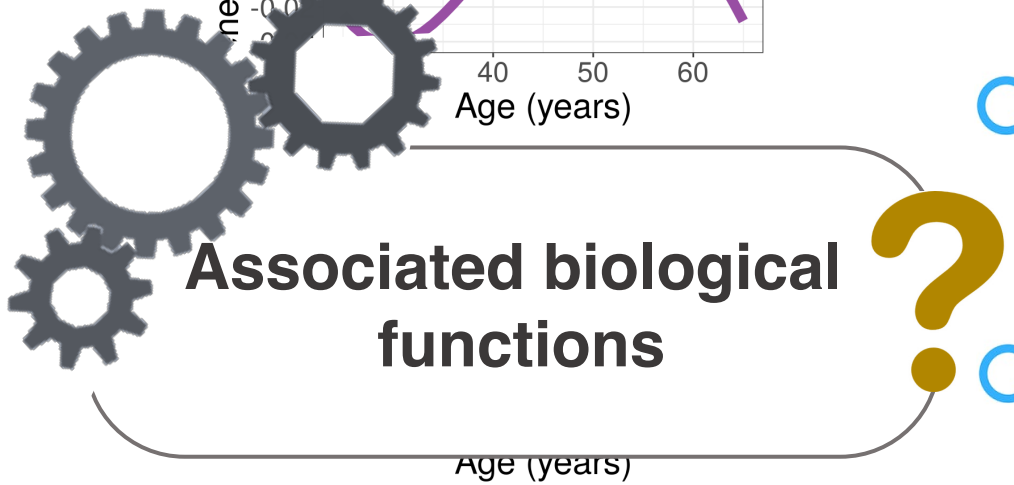


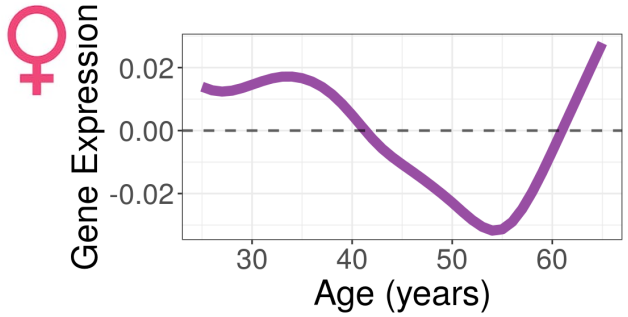


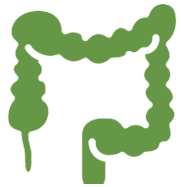




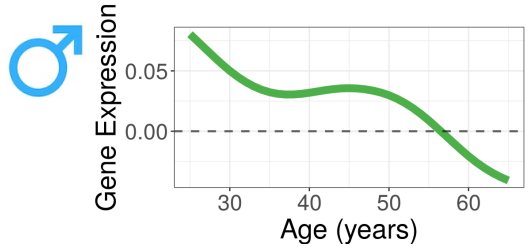
Stomach

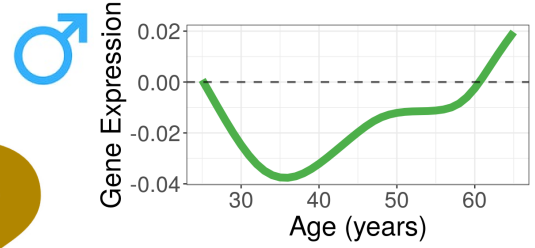


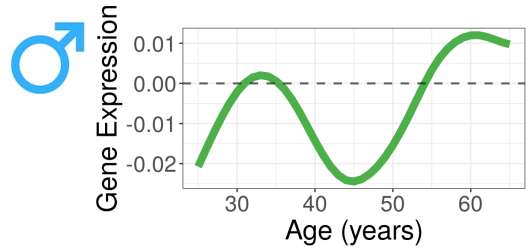


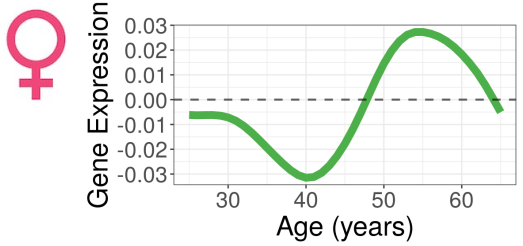


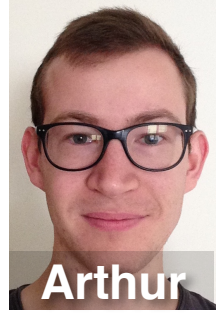
Colon
Sigmoid



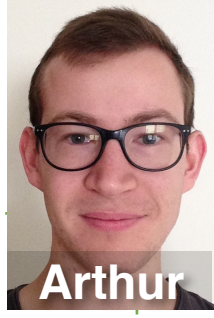








Tissue-specific ageing



Depleted in driver genes

Enriched in driver genes

Immune Response
Cell Proliferation
and Maintenance

Brain

ENDOCRINAL SYSTEM

Cell
Proliferation /
Degradation

Pituitary

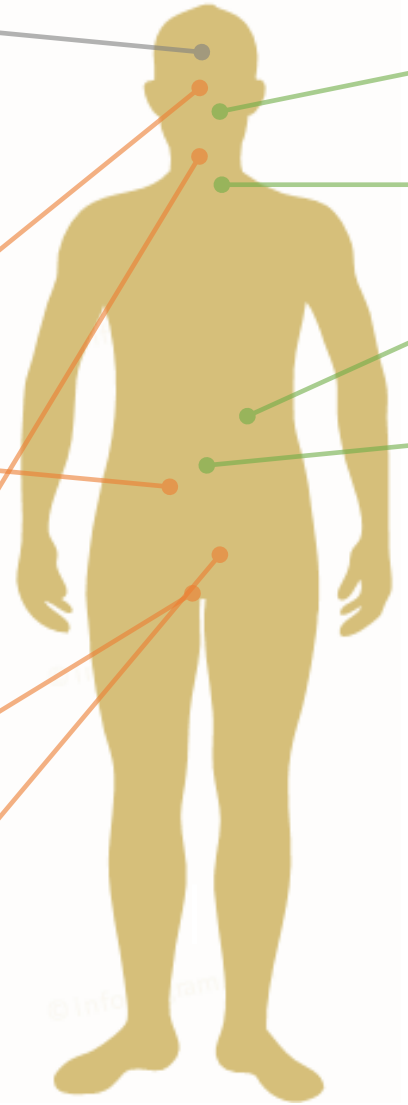
Adrenal Gland

Thyroid

Immune system

♂ Testis

♀ Ovary



METABOLIC FUNCTIONS

Minor Salivary Gland

Oesophagus

Colon

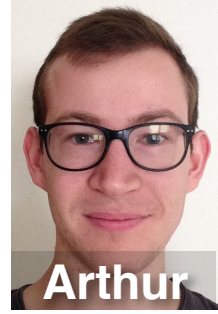
Stomach

Adipose

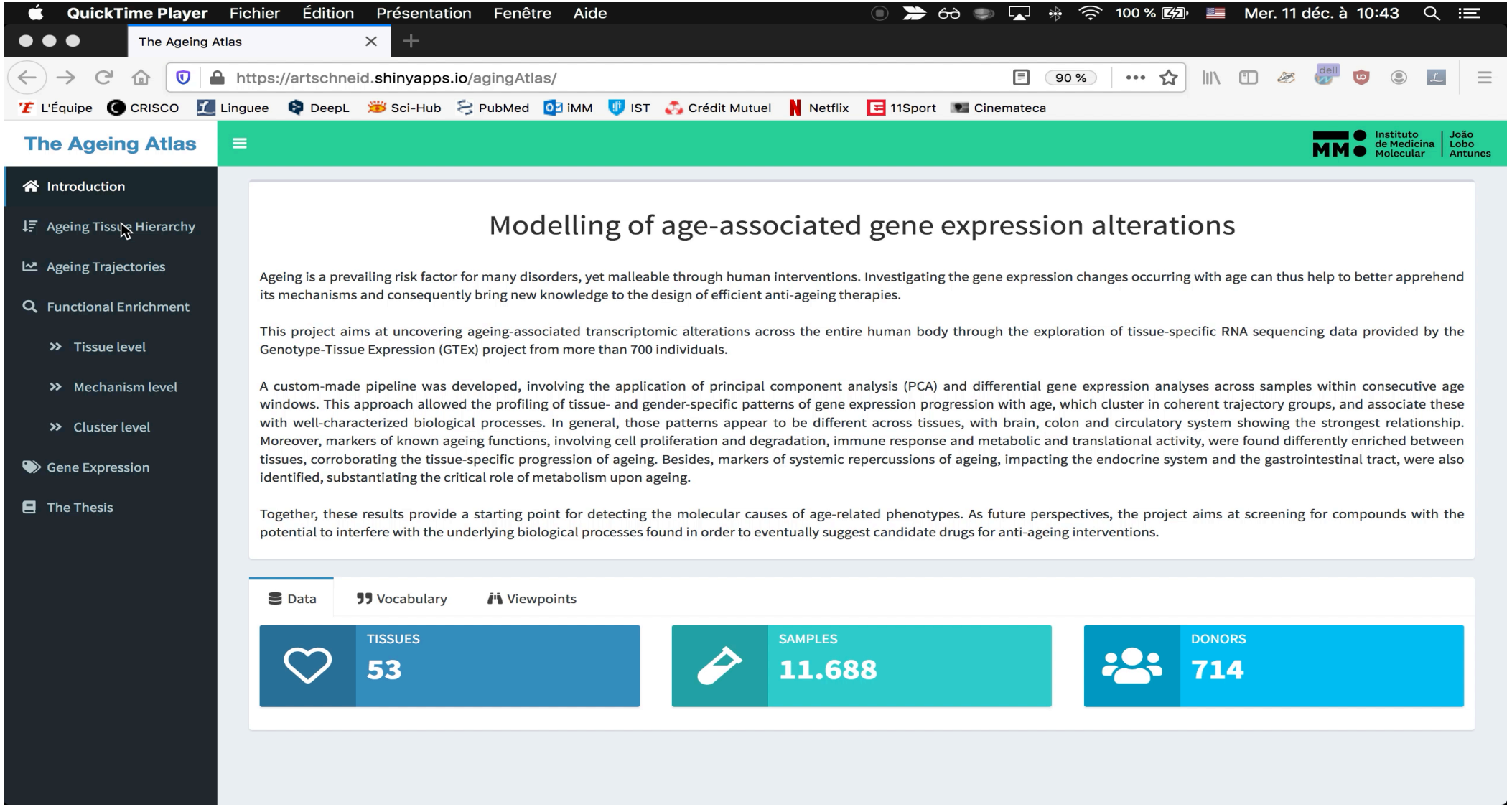
Cell Proliferation
and Maintenance

asynchronous inter-organ progression of human ageing

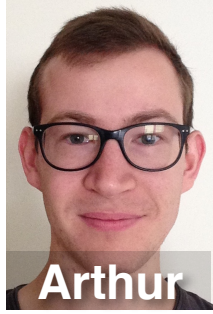
Future: the Ageing Atlas



https://artschneid.shinyapps.io/agingAtlas/



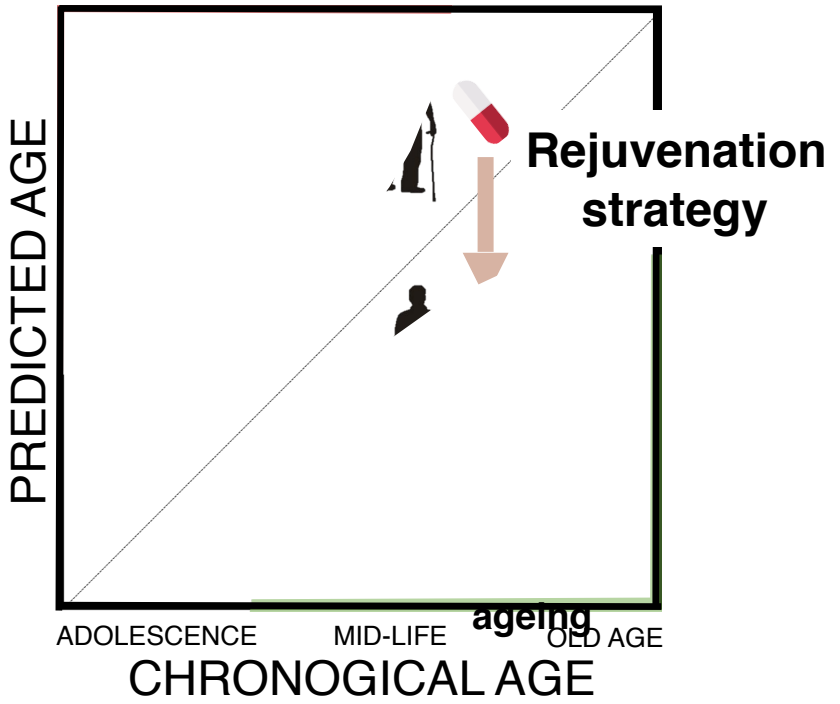
Future: translation



1. Characterise causal **genetic** or **cellular composition** alterations responsible for the profiled ageing patterns

Pinpoint **pharmacological** interventions with potential to revert them

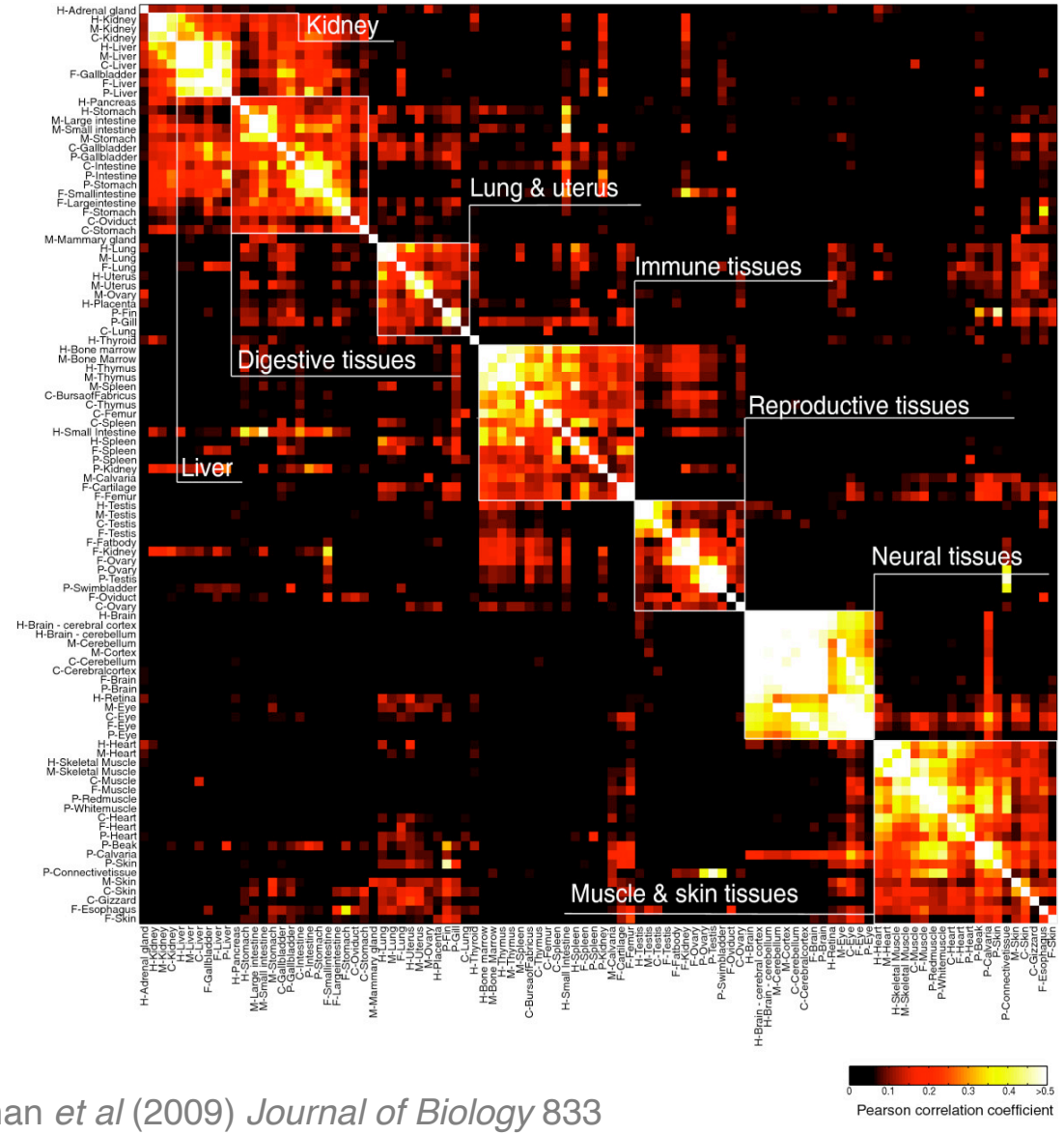
2. Develop machine learning model to **predict** the donor's age from transcriptomic data



3. Why biology needs more data scientists

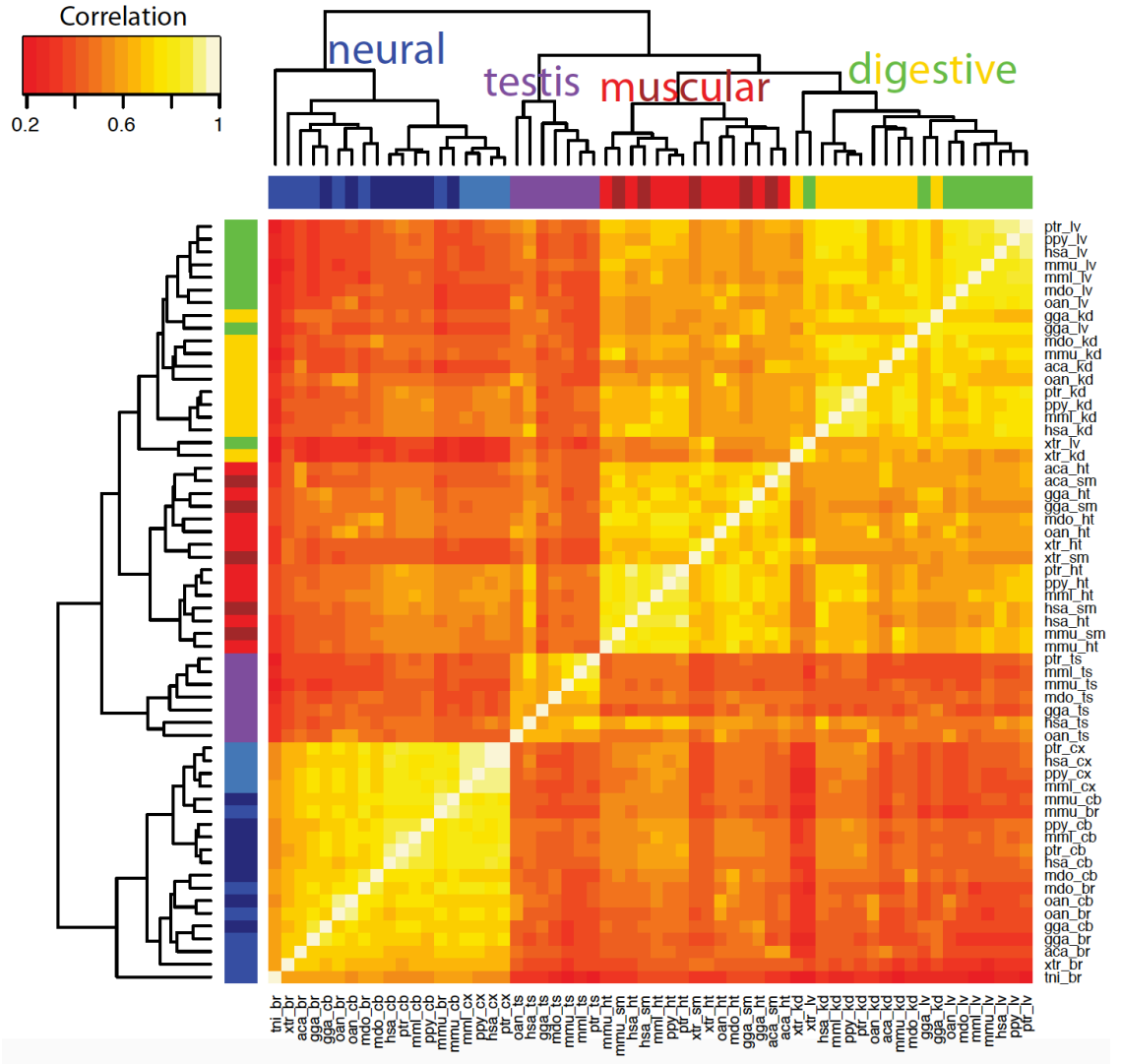
Tissue-specific gene expression is conserved in vertebrates

20 tissues/organs, 5 species, microarrays



Chan *et al* (2009) *Journal of Biology* 833

7 tissues/organs, 11 species, RNA-seq



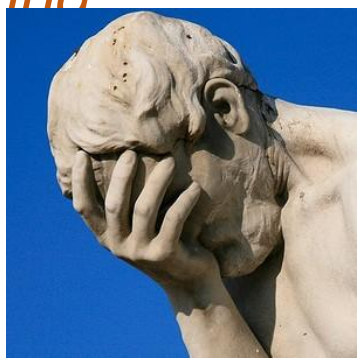
Barbosa-Morais *et al* (2012) *Science* 338:1587

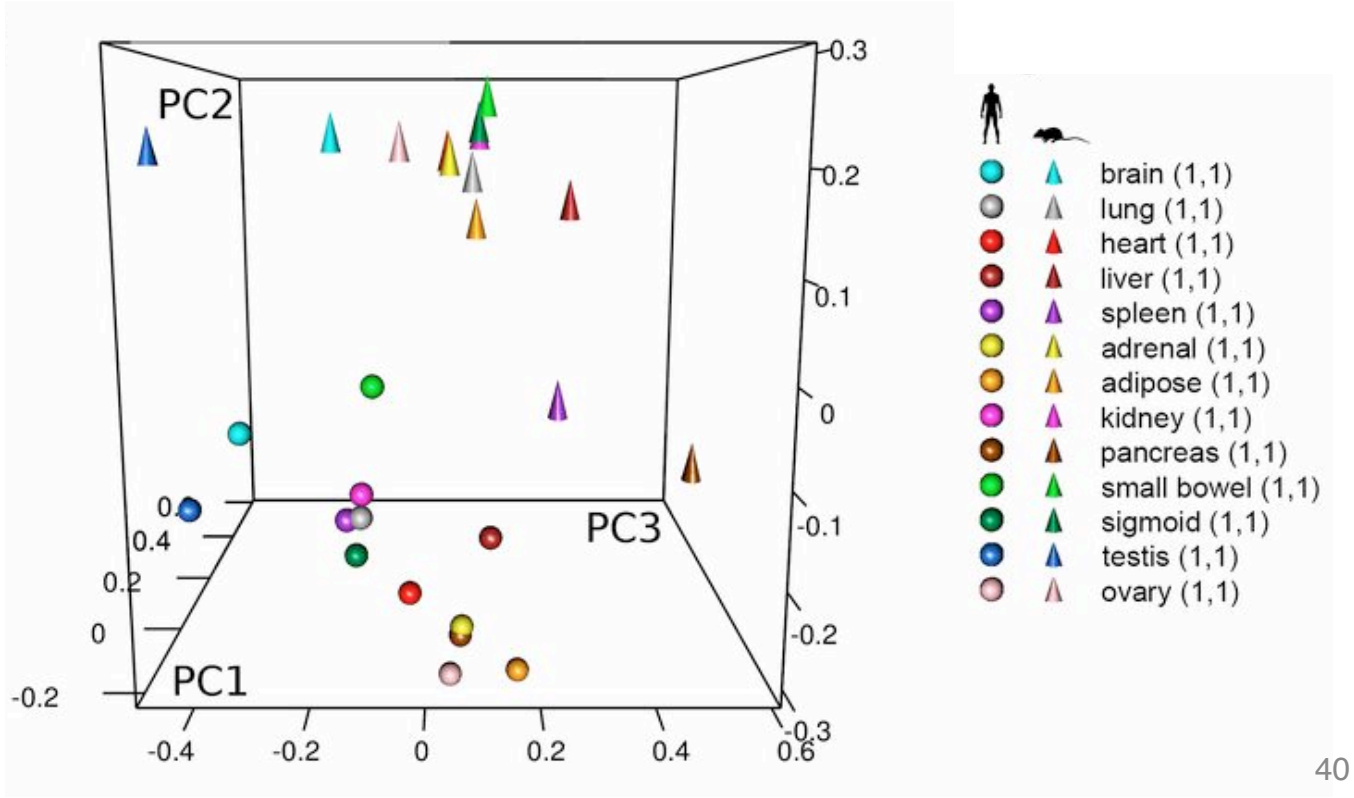
Gene Expression Is More Similar Among Tissues Within a Species Than Between Corresponding Tissues of the Two Species

Comparison of the transcriptional landscapes between human and mouse tissues

17224–17229 | PNAS | December 2, 2014 | vol. 111 | no. 48

analysis of only the 13 paired samples processed under one experimental protocol yielded the same species-specific clustering



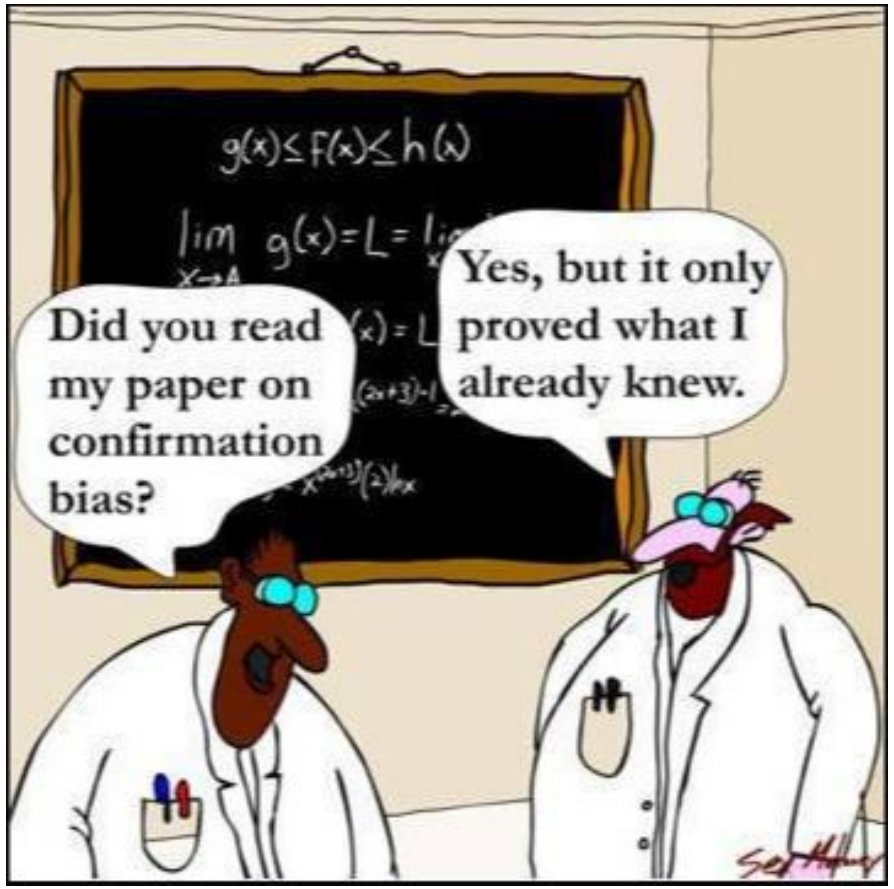


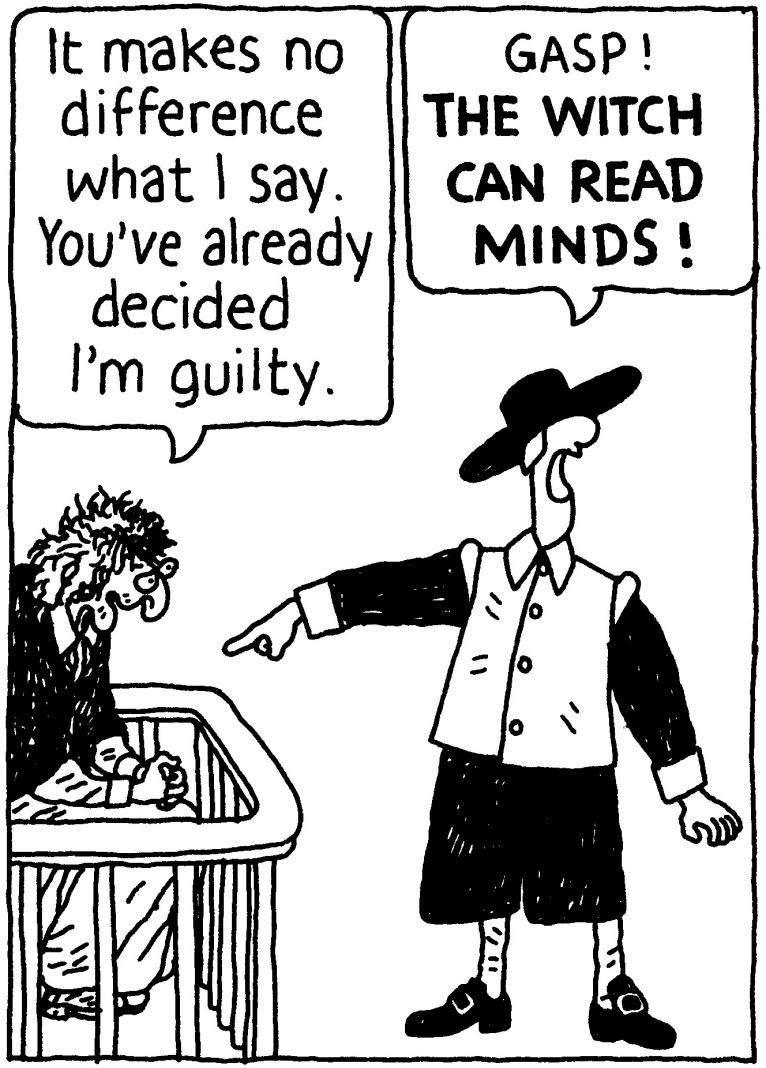
Apophenia

human tendency to perceive meaningful patterns within random data

Confirmation bias

tendency to interpret new evidence as confirmation of one's existing beliefs or theories

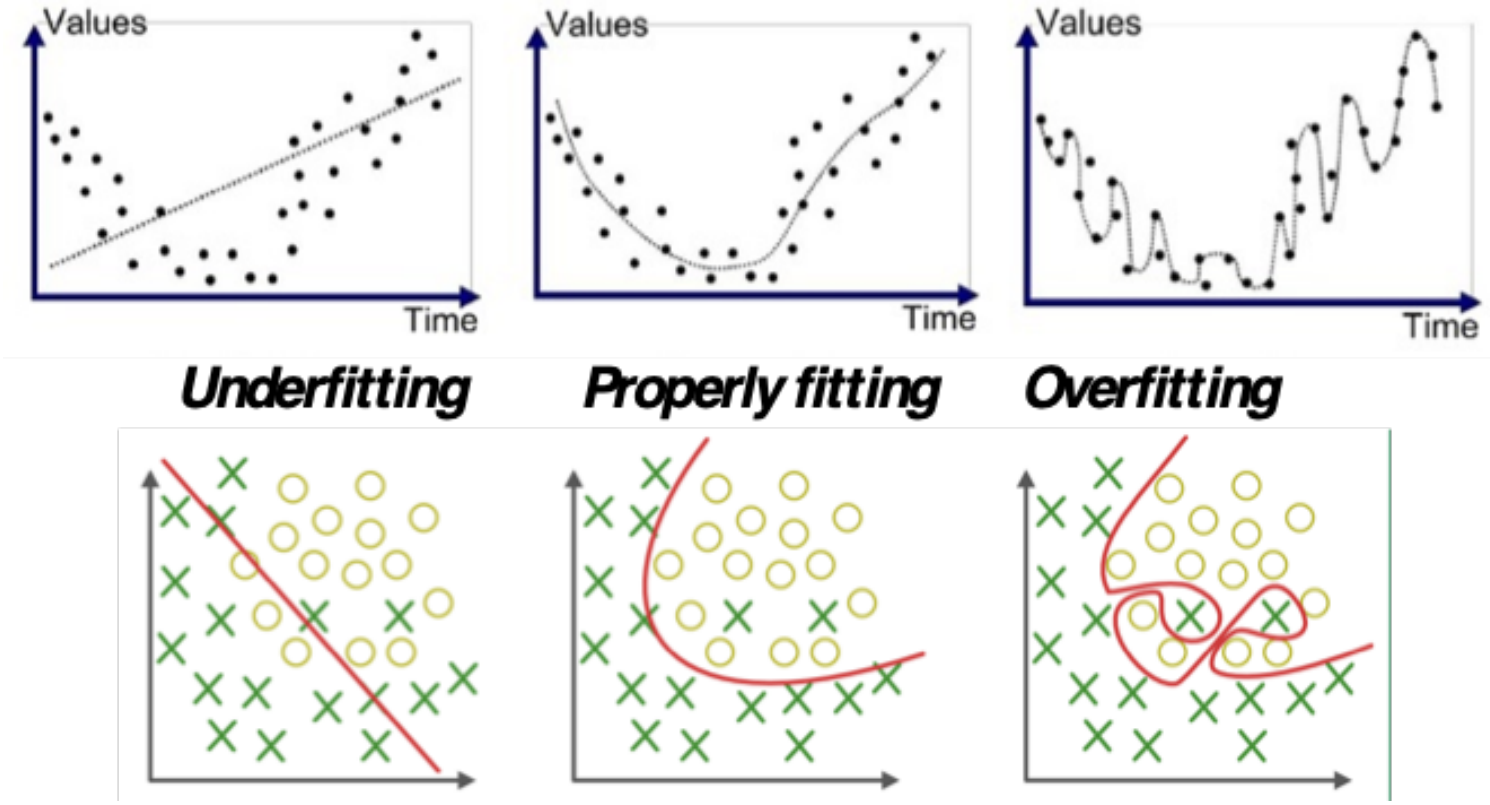




CartoonStock.com

Overfitting

the statistical model describes random error or noise instead of the underlying relationship



An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

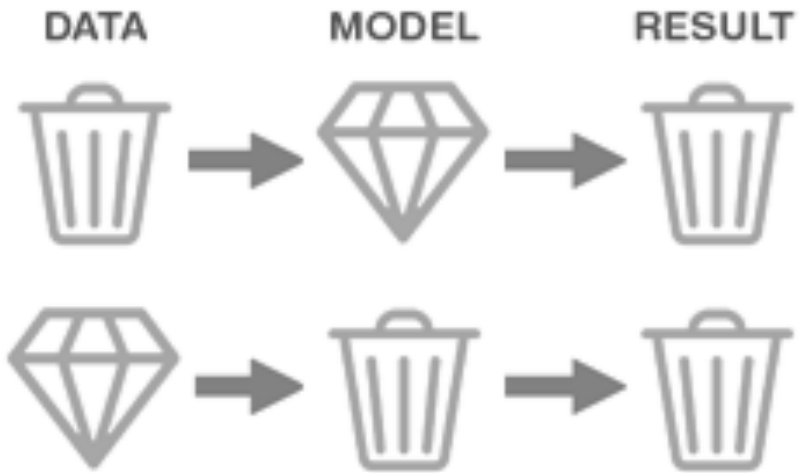
John Tukey (1915-2000) - Statistician

Distraction

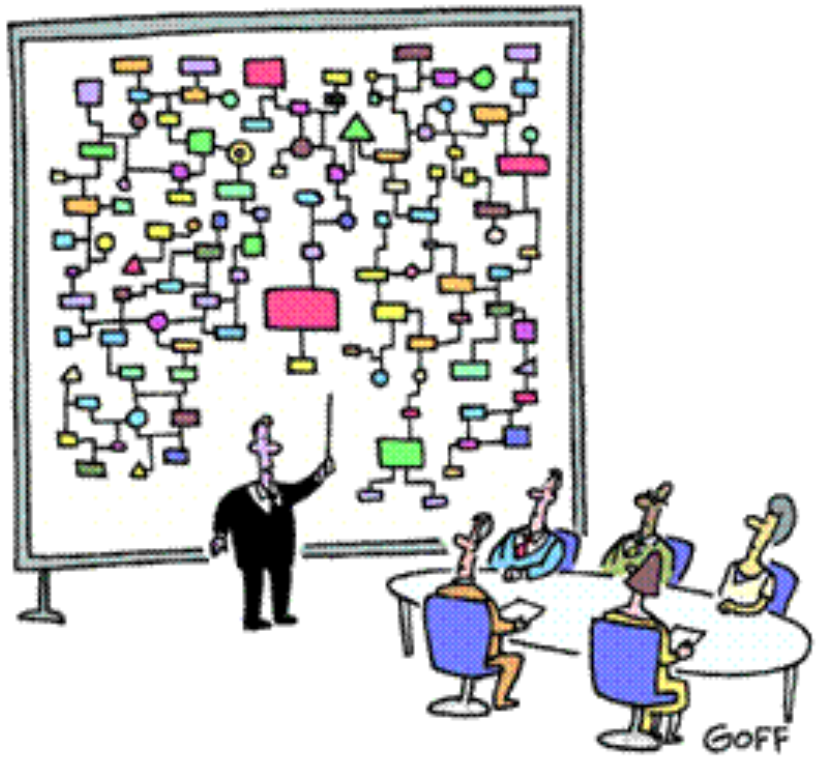
too much interesting information
(one always finds exciting biology, not necessarily related with the original question)

Complacency & illiteracy – the black box

- Not bothering with statistical concepts and assumptions
- Looking for easy **recipes**, blindly judging numbers (p-values) without "getting intimate" with data (e.g. plotting)
- Evaluation of fairness of method based on outcome



Misestimate of biological systems' complexity

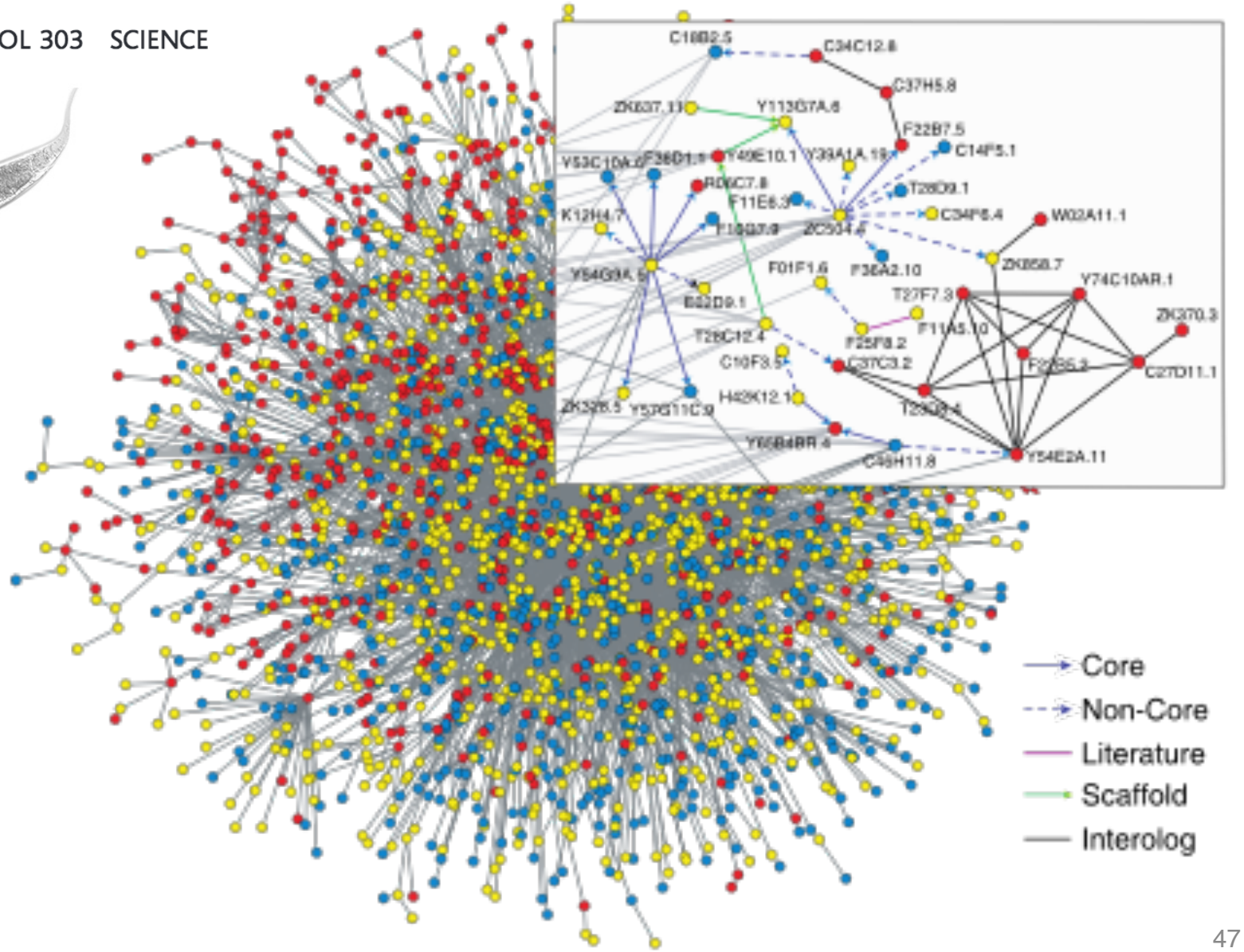


"And that's why we need a computer."

A Map of the Interactome Network of the Metazoan *C. elegans*

540 23 JANUARY 2004 VOL 303 SCIENCE

***Interactome**: the whole set of molecular interactions in a cell*



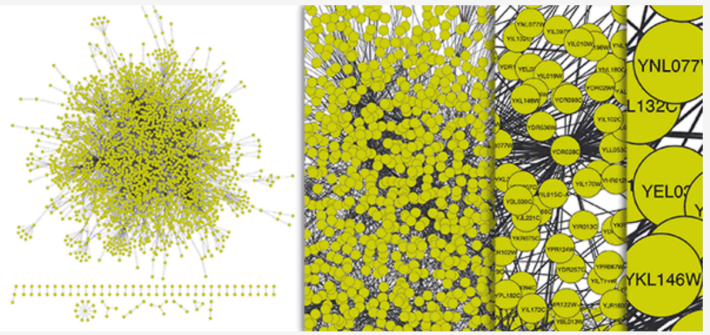
How informative?

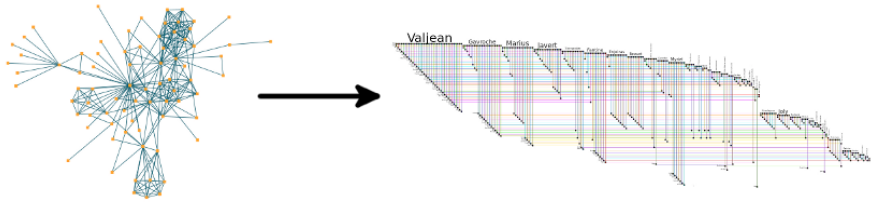
BioFabric

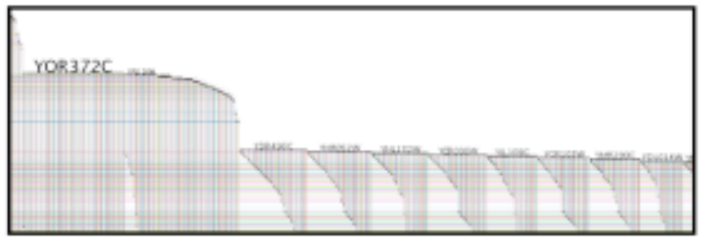
Institute for Systems Biology

News report:

Combing the Hairball

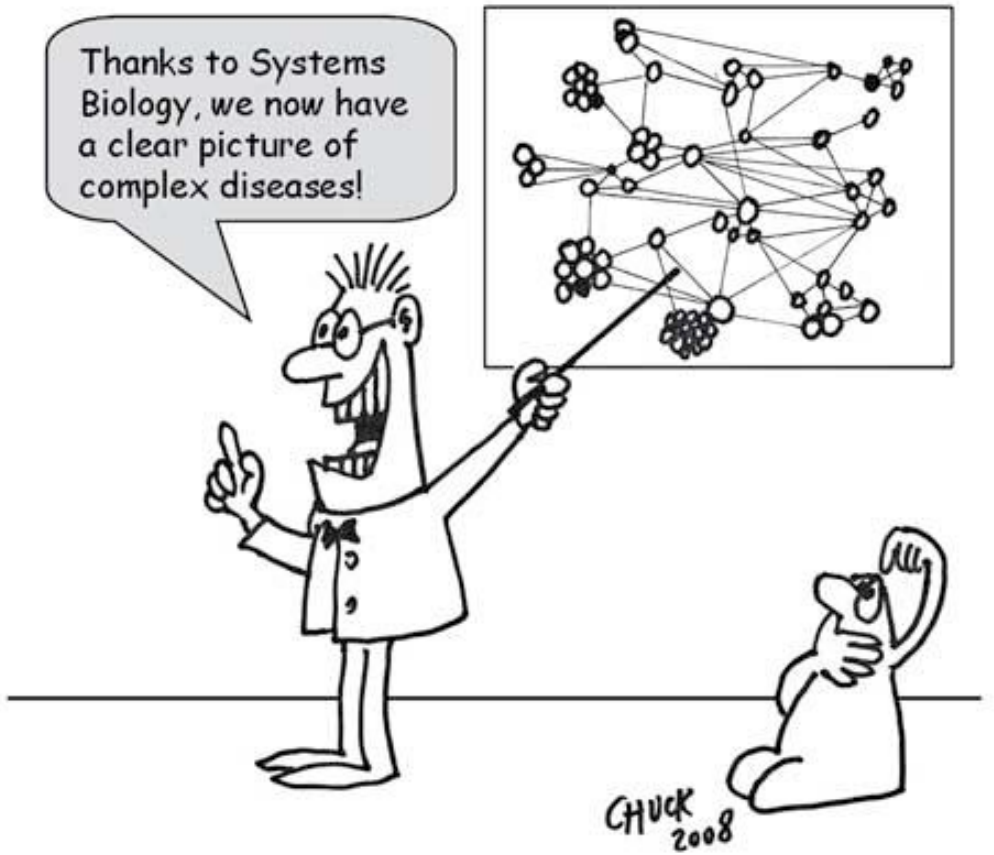






How informative?

Pride and prejudice



Biologists:

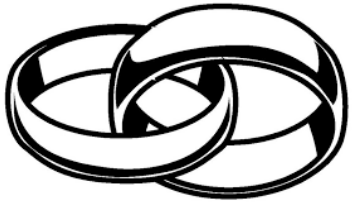
- See quantitative science as just a tool
- Claim biological significance with weak statistical support
- Underestimate their lack of (S)TEM

Quantitative scientists:

- Algorithm more important than the biological question
- Biological systems can be modeled by a set of simple equations (hard time dealing with biological variability)
- Find biology too descriptive and its conclusions far fetched

What makes good ~~computational biologists~~ biological data scientists

Drive to address
biological questions
(i.e. strong interest in biology)



Fluency in
data analytics
(e.g. algebra, statistics, algorithms)

Empathy!

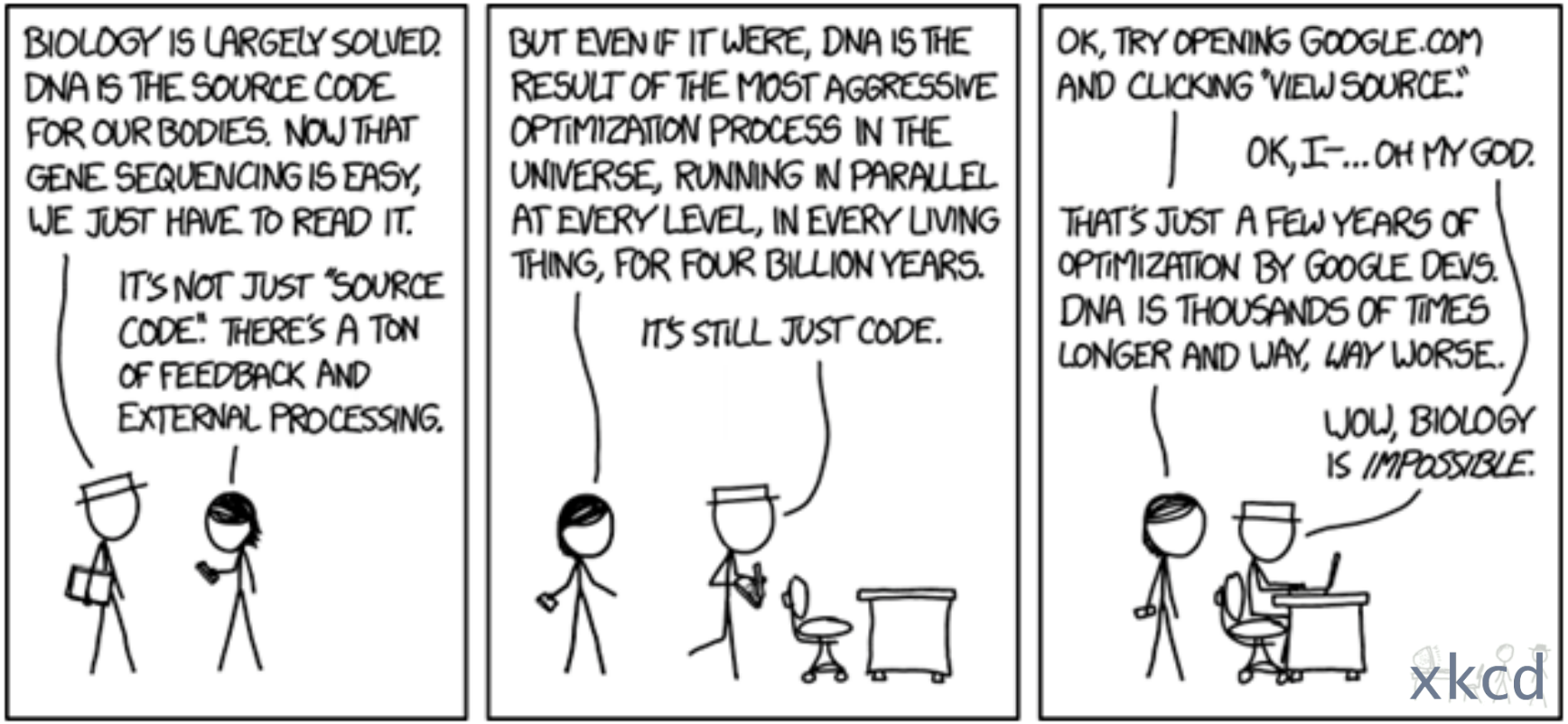
Genuine interest in the "other" side's questions and ability to understand how they think

Our anti-black box efforts

- Make our ways of analysing data intelligible for (and therefore open for scrutiny by) collaborators
- Develop tools for empowering colleagues to perform similar analyses while understanding decisions needed to be made at their every stage

Thank you!

Questions?



Genomic science is wonderful in that it brings together representatives of so many disciplines - clinicians, bench biologists, statisticians, bioinformatics scientists - all of whom tend to consider the others ***intellectual peasants****.*

Zak Cohane

Prof. of Biomedical Informatics @ Harvard