Big Data in Genomics and Life Sciences

Alvis Brāzma
European Bioinformatics Institute
European Molecular Biology Laboratory

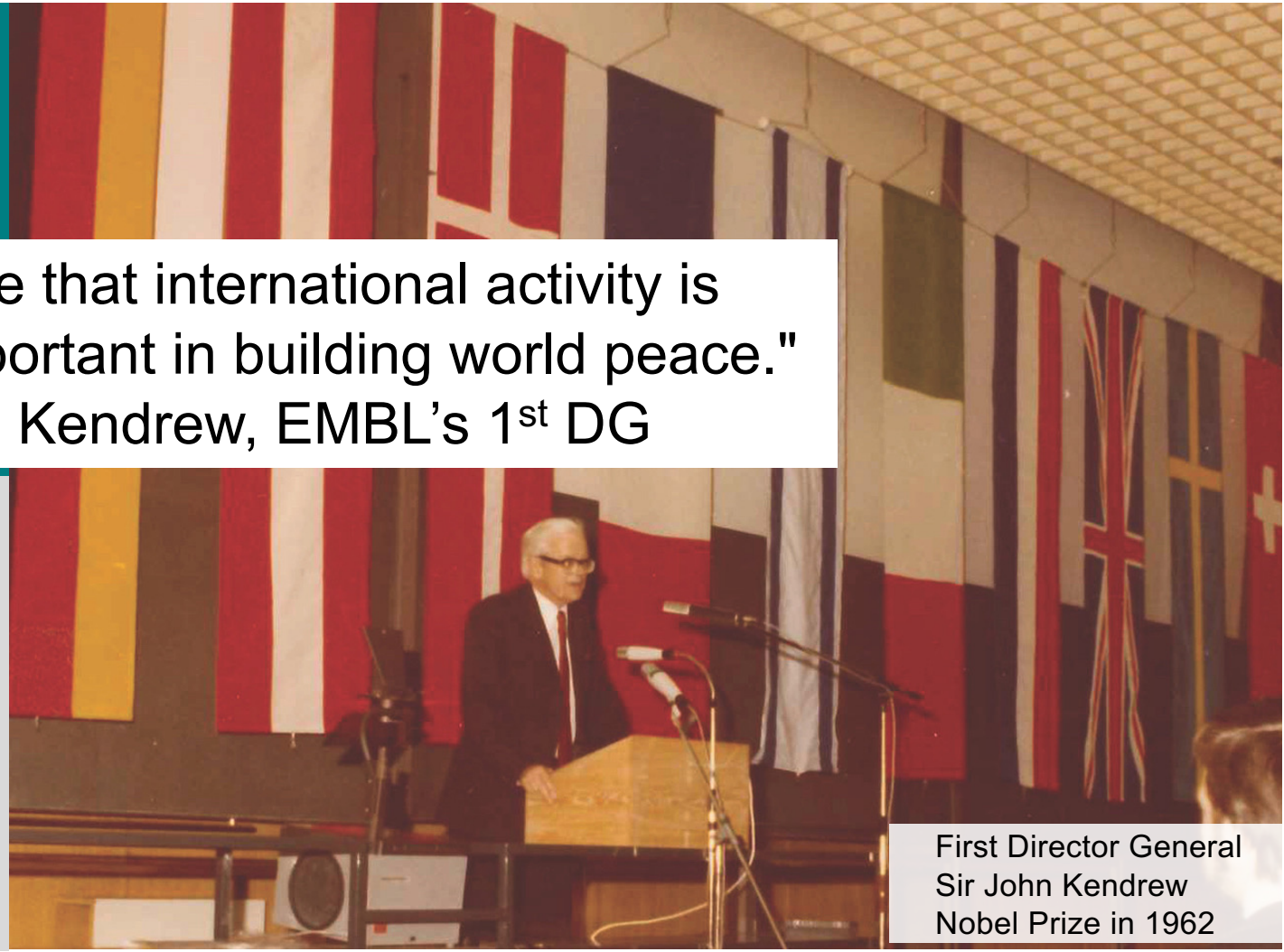
Rīga, July 2022

EMBL

European Molecular Biology Laboratory - EMBL

Europe's centre of excellence in *life sciences* research, services and training

"I believe that international activity is very important in building world peace."
Sir John Kendrew, EMBL's 1st DG



First Director General
Sir John Kendrew
Nobel Prize in 1962

Founded in 1974 by 10 states as an intergovernmental organisation to promote the molecular life sciences in Europe and beyond

EMBL member states in 2022

Member states (27)

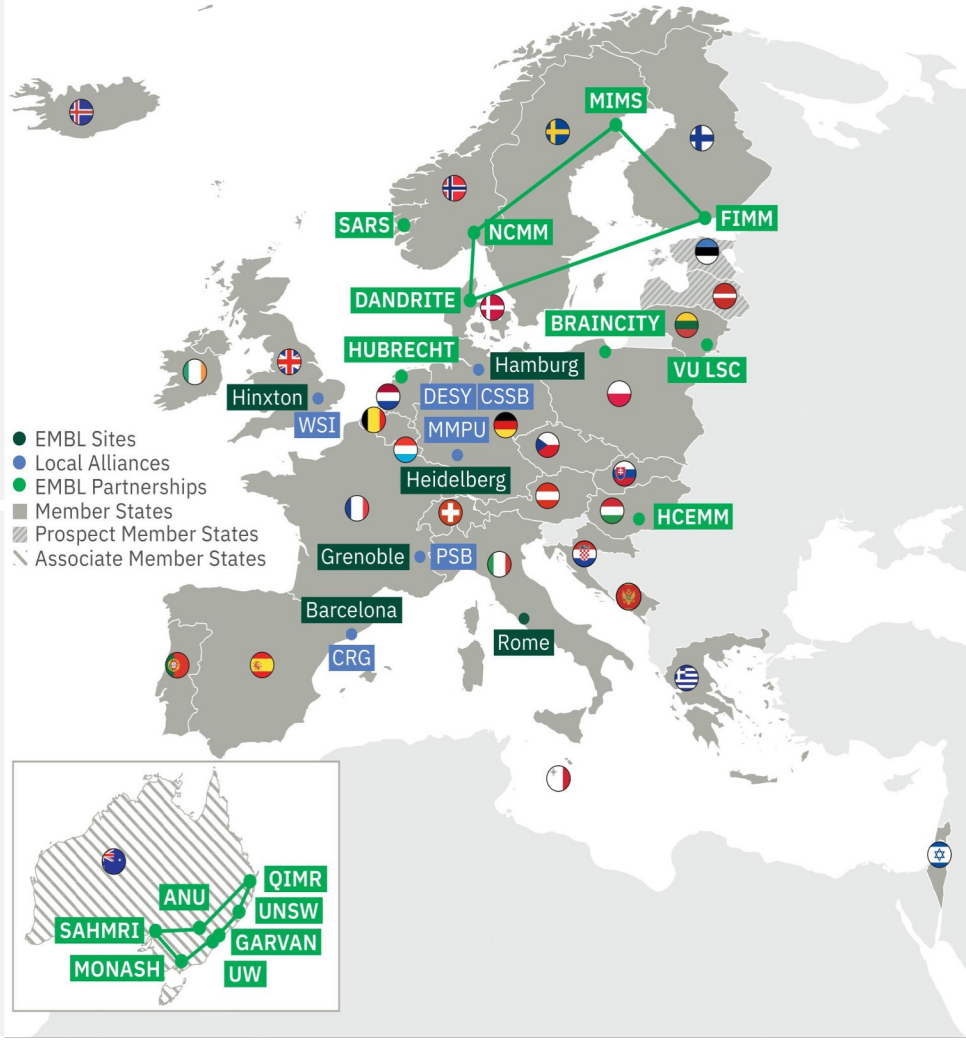
Austria 1974
Denmark 1974
France 1974
Germany 1974
Israel 1974
Italy 1974
Netherlands 1974
Sweden 1974
Switzerland 1974
United Kingdom 1974
Finland 1984
Greece 1984
Norway 1985
Spain 1986
Belgium 1990
Portugal 1998
Ireland 2003
Iceland 2005
Croatia 2006
Luxembourg 2007
Czech Republic 2014
Malta 2016
Hungary 2017
Slovakia 2018
Montenegro 2018
Lithuania 2019
Poland 2019

Associate member states

Australia 2008

Prospect member states

Estonia
Latvia



Six sites with almost 1900 people and >90 nationalities

EMBL-EBI

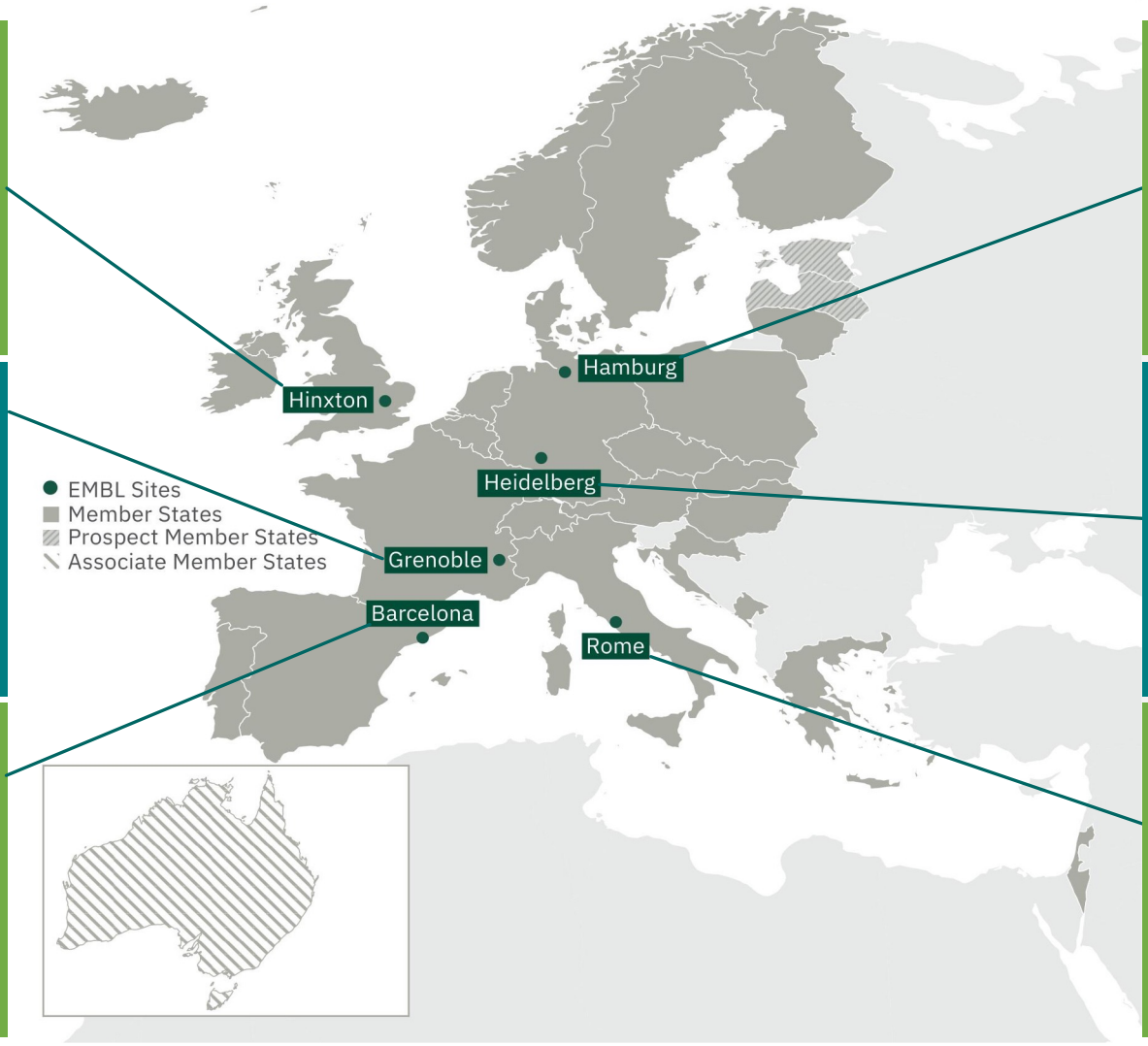
Bioinformatics

Grenoble

Structural biology

Barcelona

Tissue biology and disease modelling



Hamburg

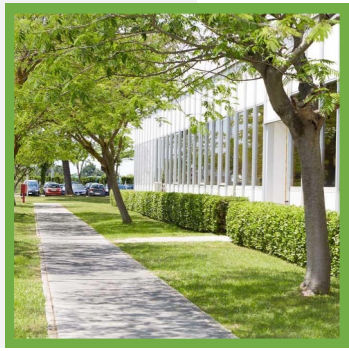
Structural biology

Heidelberg

Life sciences

Rome

Epigenetics and neurobiology



EMBL

Wellcome Genome Campus in Hinxton, near Cambridge, UK

EMBL

What is EMBL-EBI?

- World leading source of public biomolecular data
- Cutting edge research in bioinformatics and computational biology
- Part of the European Molecular Biology Laboratory (EMBL), Europe's flagship laboratory for the life sciences.

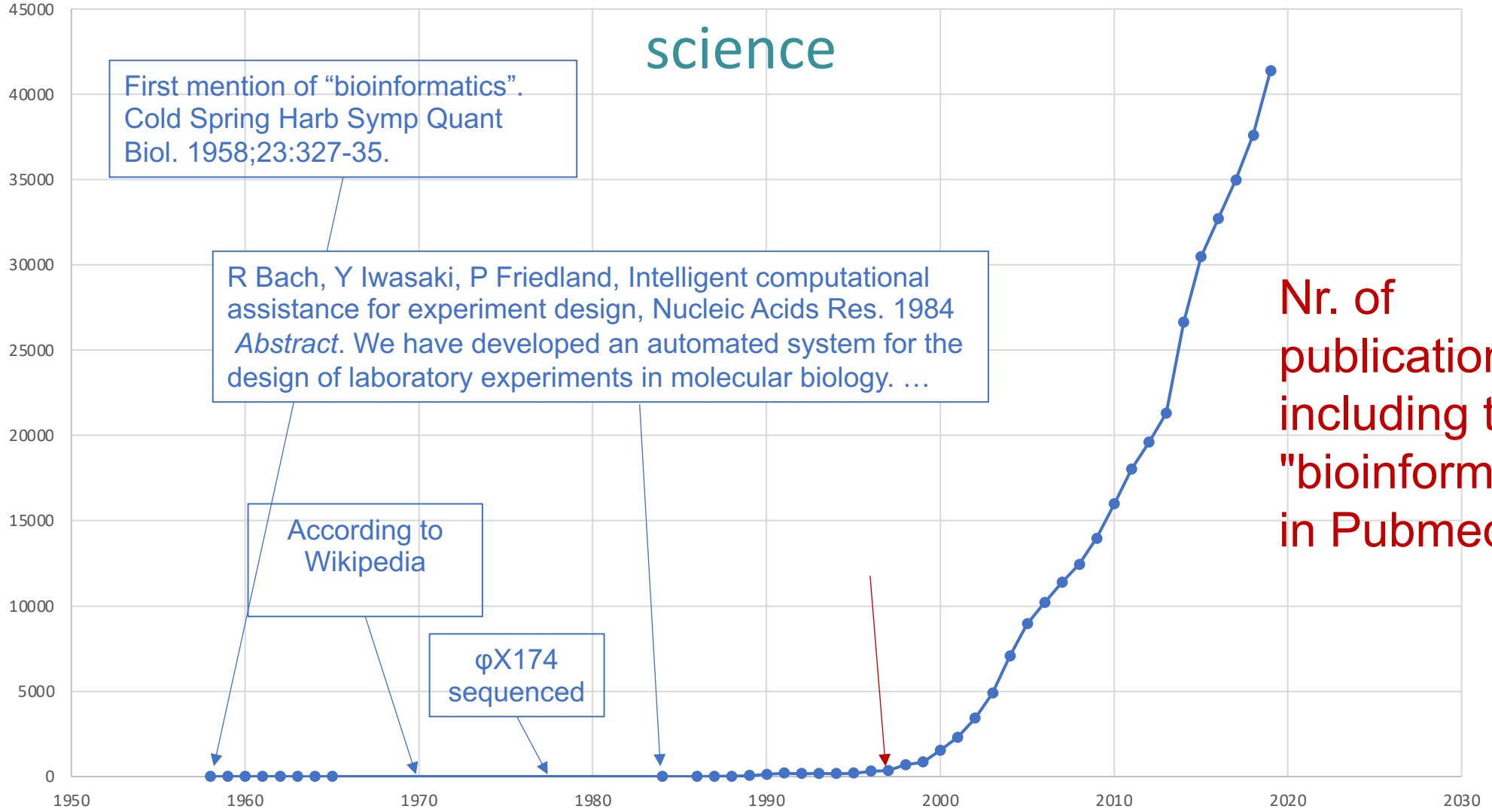
EMBL

Bioinformatics is interdisciplinary field that develops methods and software tools for understanding biological data (Wikipedia 2022)

Data science in biology

EMBL

Bioinformatics: from fringe to mainstream science



Sputnink 1957

Moon-landing 1968

EMBL 1974

PCR 1983

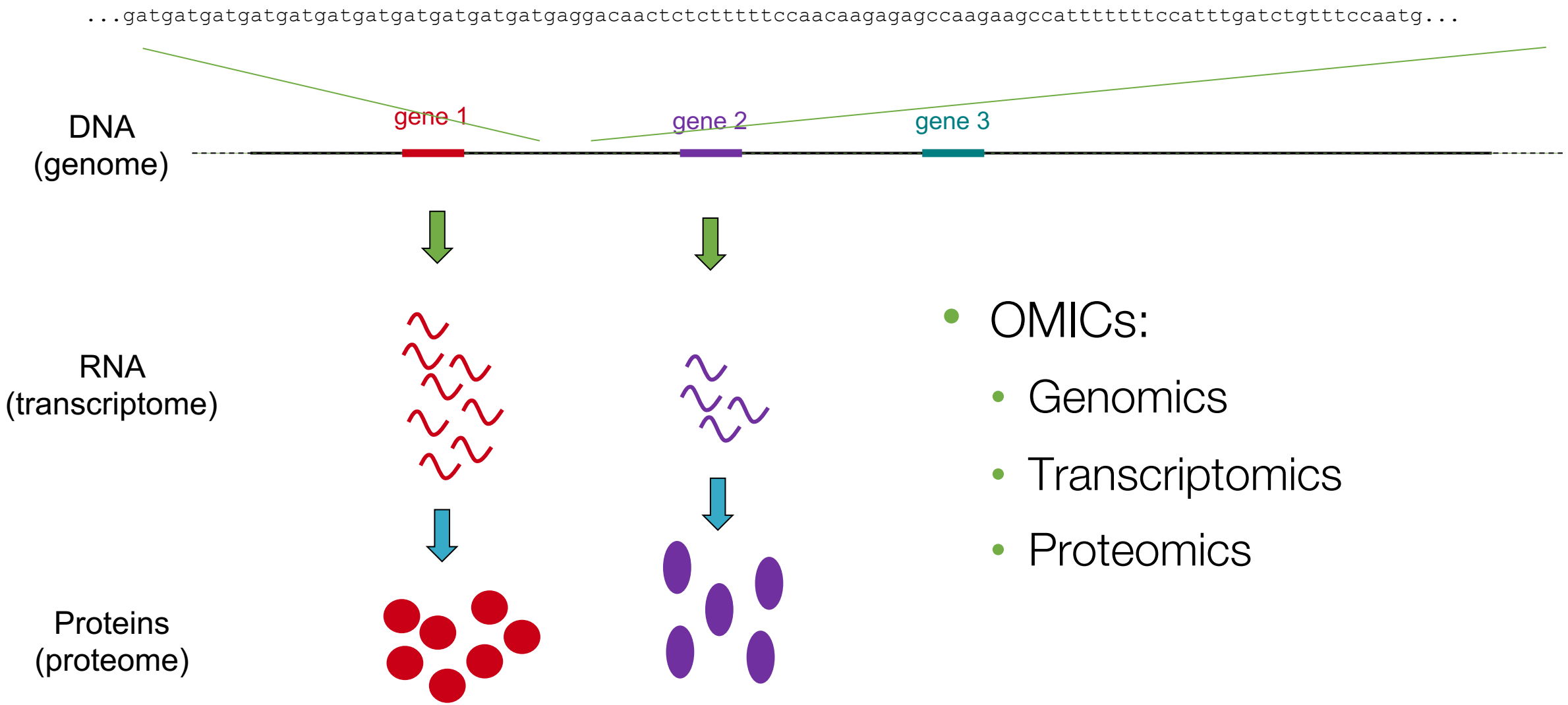
EMBL-EBI 1994

Dot-com crash

RNA vaccines

A computer wins Nobel Prize for developing a new drug

Central dogma of molecular biology and OMICs



- OMICs:
  - Genomics
  - Transcriptomics
  - Proteomics

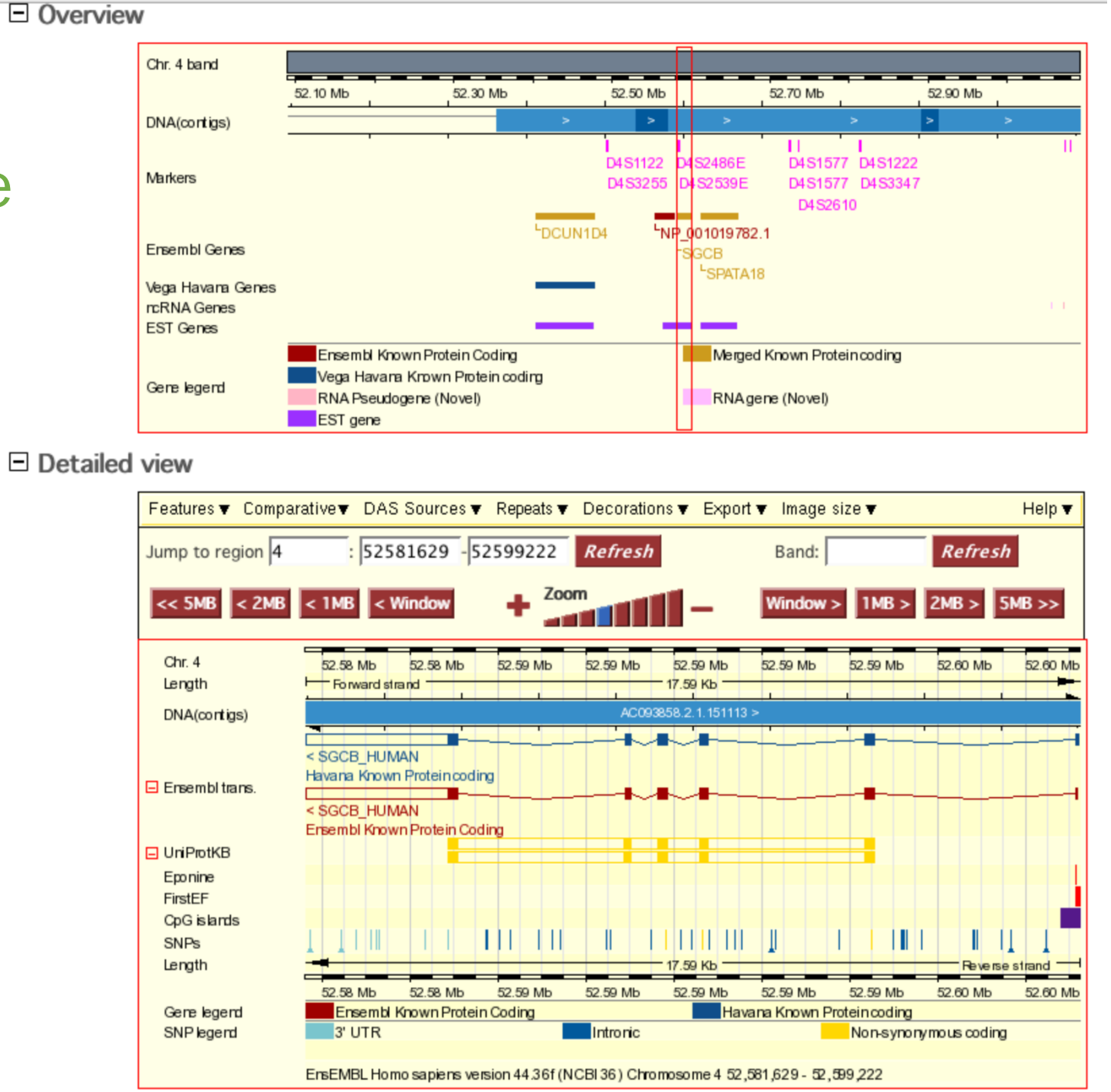
Human Genome project was an epic

- 1990 – 2003 about 85% of the **3 billion letters** of human genome sequenced
  - The first ”gapless” human genome sequence was achieved only in 2022
- HGP budget over the 13 years was $ 3 billion, though the actual sequencing costs were considerably less
- Some lessons from HGP
  - There are 20,000 protein coding genes and some other genes (most shared with other mammals)
  - Less than 2% of the human genome codes for proteins (other parts are interesting too)
  - Over a half of human genome consist of repeated sequences of different length apparently not having any function (junk genome)
- Sequencing one human genome in 2022 is < $1000
- Genomes of two unrelated individuals differ in 1 ”letter” per 3000 on average

EMBL

Ensembl Genome Browser at EBI

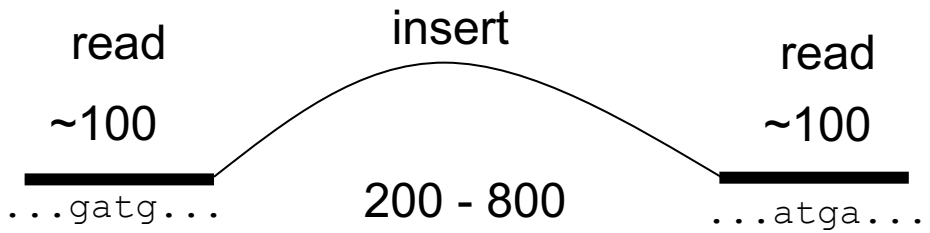
www.ensemble.org



EMBL

Personalized genomics - assembling a new genome vs. mapping sequencing “reads” to a reference genome

- The output of a DNA sequencing machine is a set of many ~100 letter long sequences that jointly cover (most of the genome) 10 to 50 times on average
  - Current sequencing technologies output pairs of such ~100 letter sequences with approximately known distance between them

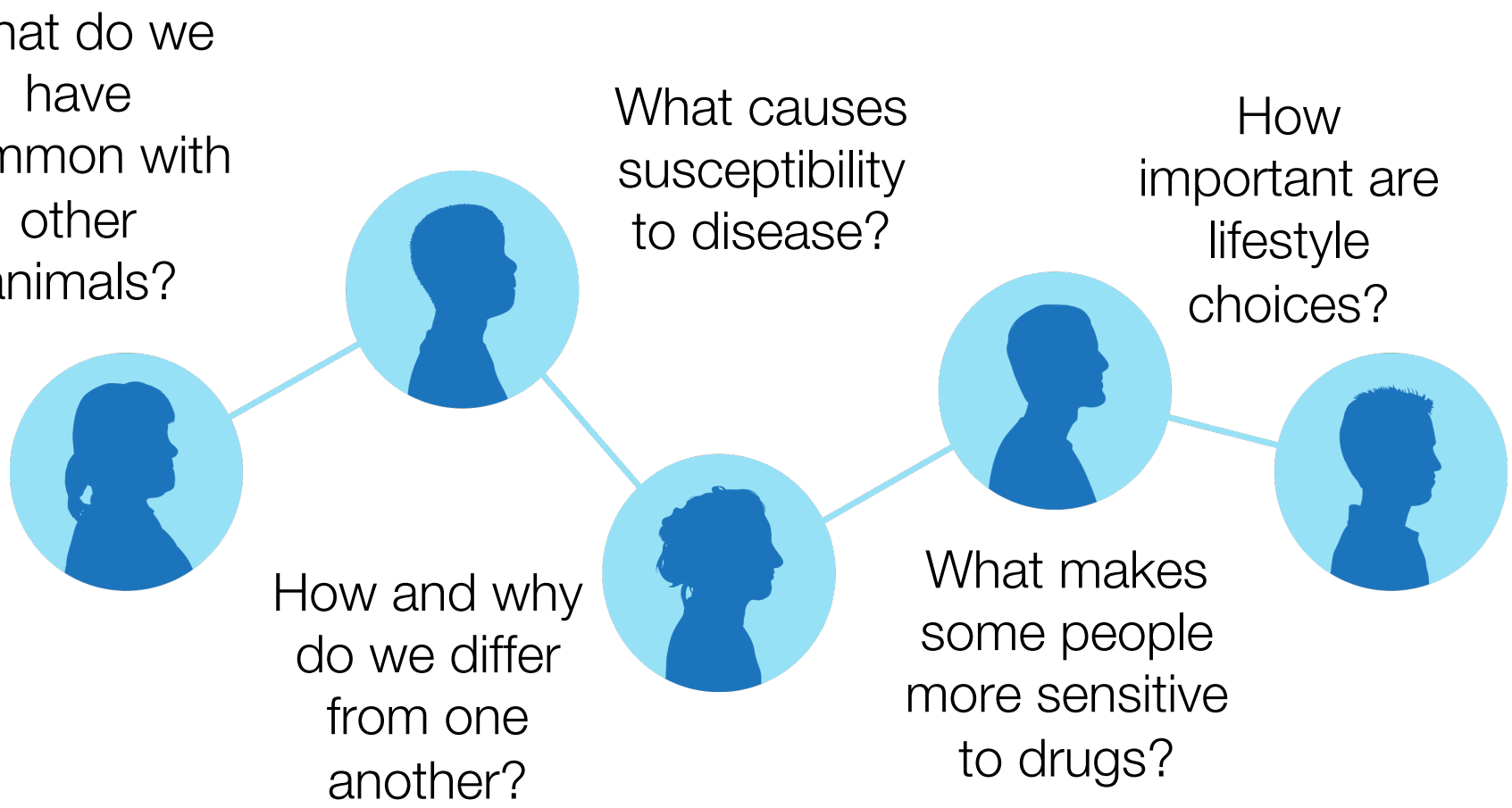


The “insert” may help to deal with repeated sequences in the genome

- Once we have one – reference – genome, to get another we can “map” these “reads” onto the reference to obtain the new one
- Mismatches of letters in the “read” can be genome variation or errors
- If we can ”trust” the reference? Many references for different populations

EMBL

Personal genomics: what do genomes tell us?



Rare diseases

Genomics promises a leap forward for rare disease diagnosis

Faster and cheaper DNA sequencing brings new hope to patients

Jessica suffers from a rare condition that was diagnosed through DNA analysis

“Kate Palmer and Simon Wright were in despair. Their four-year-old daughter Jessica was suffering from epilepsy, poorly co-ordinated movement and slow mental development, but doctors had been unable to pinpoint the rare disorder causing these problems.”

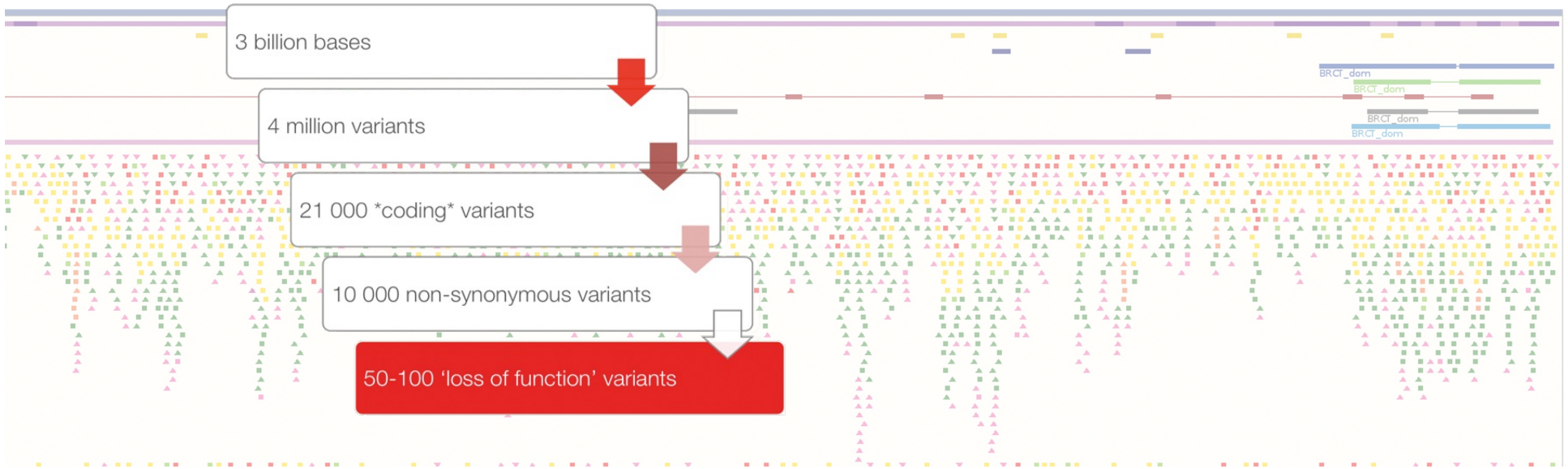
- Jessica was enrolled in the UK 100,000 genome project
- By analysing Jessica’s genome, a mutation in a gene called SLC2A1 was found starving her brain of the sugar
- This condition is extremely rare, but there is a treatment in the form of a diet that enables the brain to maximise glucose production
- After a month on the new diet, Jessica’s parents “noticed a big increase in her speech, energy levels and general steadiness”

Sharing personal genome data

- While reference are an average of many genomes, individual genomes are unique and can be linked to person and other information
- Implication to privacy, health insurance, etc
- Pseudonymisation (or even anonymization) is not sufficient to be sure that one's genome identity is not revealed – linking to health records, etc
- Consent and law
- Controlled access by bone-fide researchers
- In many countries personal genome data are not allowed to leave the country by law
- Solutions: federation of data, common standards, cloud computing

Know your genome

- Figuring out which regions are involved in disease – and what they do – is a major challenge.



European Genome-Phenome Archive

EUROPEAN GENOME-PHENOME ARCHIVE
10TH ANNIVERSARY

- Resource for permanent secure archiving and sharing of all types of potentially identifiable human genetic and phenotypic data
- Distributed data access model
- Access to bona-fide researchers controlled by Data Access Committees

How the EGA is managed

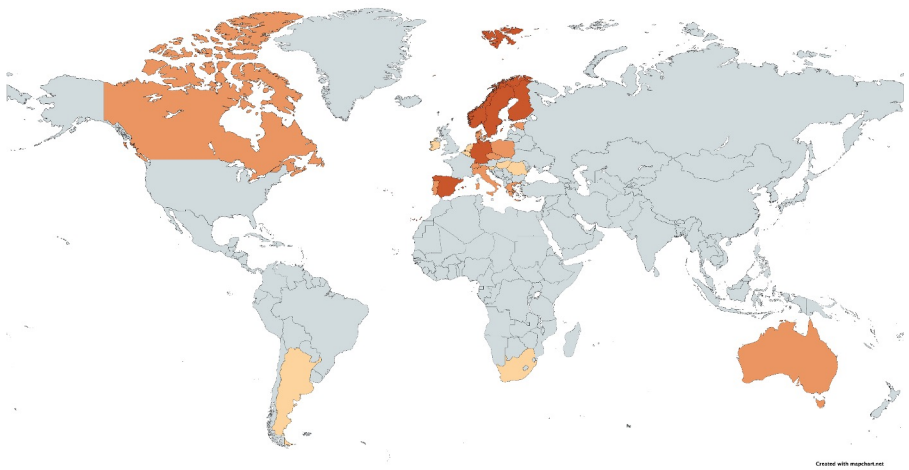
The EGA was launched in 2008 by the EBI

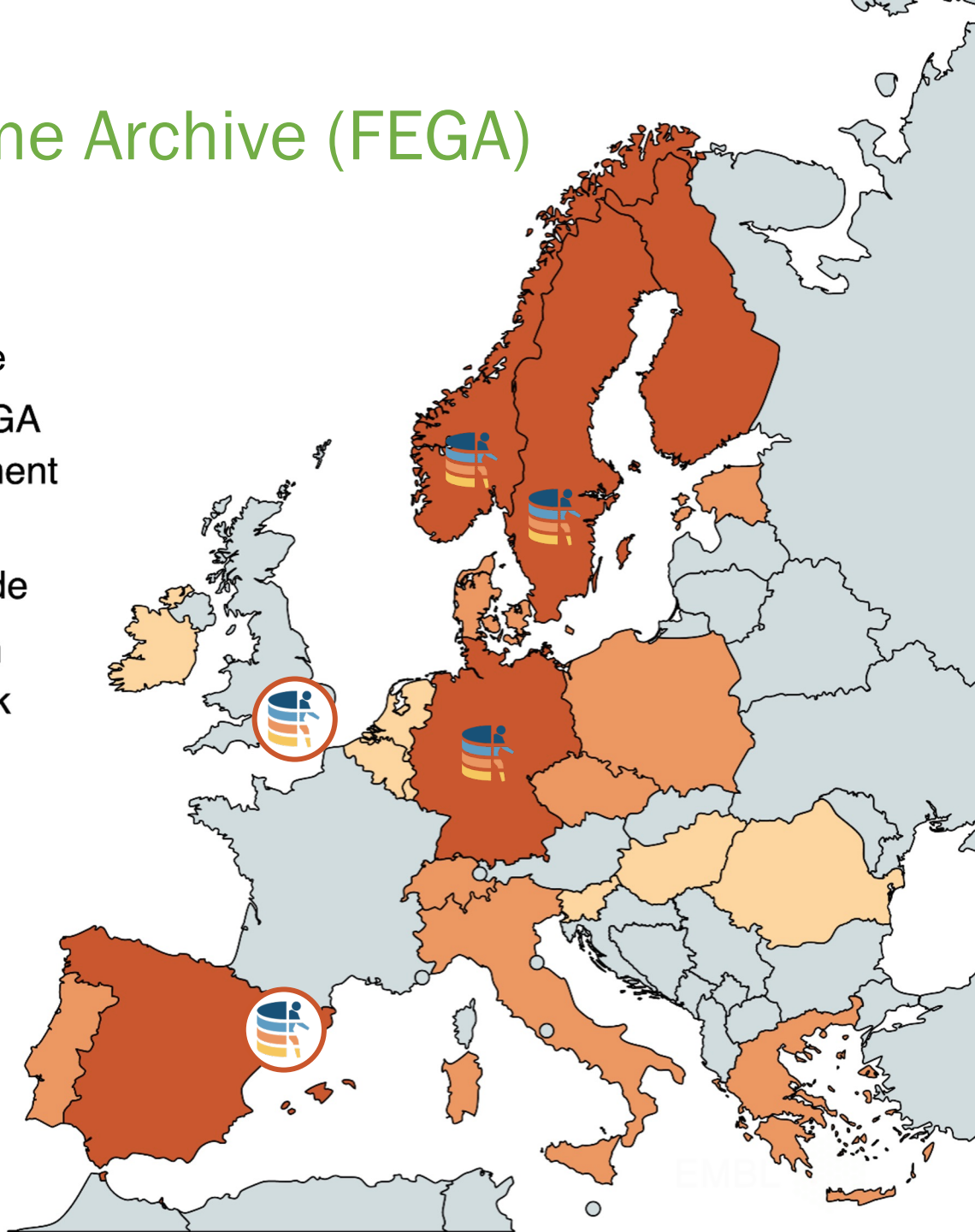
In 2012, EBI and CRG started working together to establish EGA as a joint venture, which has grown in the context of ELIXIR

Federated European Genome-phenome Archive (FEGA)

>20 countries / national initiatives and counting!

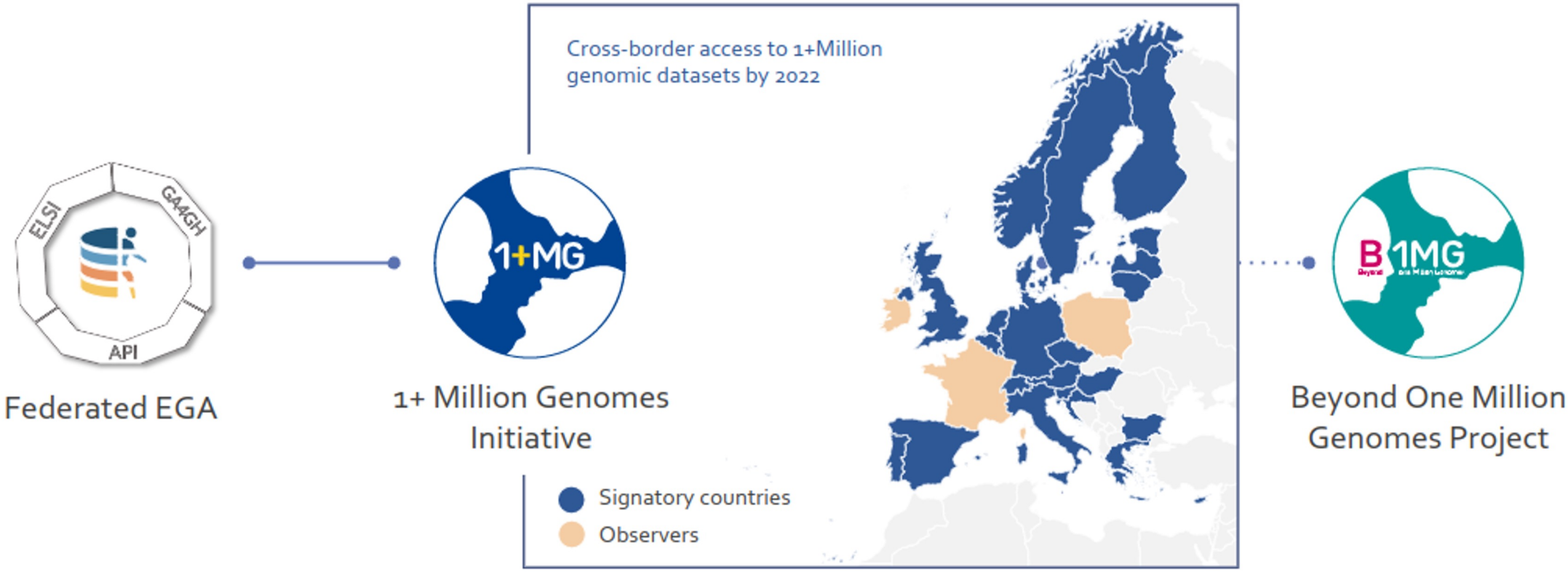
- Central EGA node
- Federated EGA node
- Preparing to sign FEGA Collaboration Agreement
- Engaging in work to establish a FEGA node
- Expressing interest in joining FEGA Network





Beyond 1 Million Genomes Project B1MG

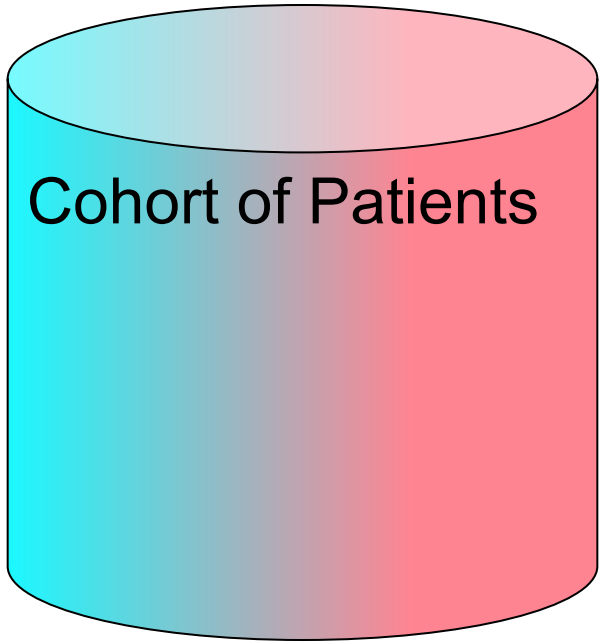
Building towards genomic data infrastructure



EMBL

Data generated for research project vs. data generated as a part of healthcare

Research data:
Disease Associations
Molecular biology resource



Healthcare genomics data and electronic health records

Feedback to patients
Actionable variants

Iceland; Denmark; Faroe; Finland;
Dundee; UK BioBank; (others)

Global Alliance
for Genomics & Health

The Global Alliance for Genomics and Health (GA4GH) is an international, nonprofit alliance formed in 2013 to accelerate the potential of research and medicine to advance human health. **Bringing together** 600+ leading organizations working in healthcare, research, patient advocacy, life science, and **information technology**, the GA4GH community is working together **to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data.**

https://www.ga4gh.org/

EMBL

Not only humans

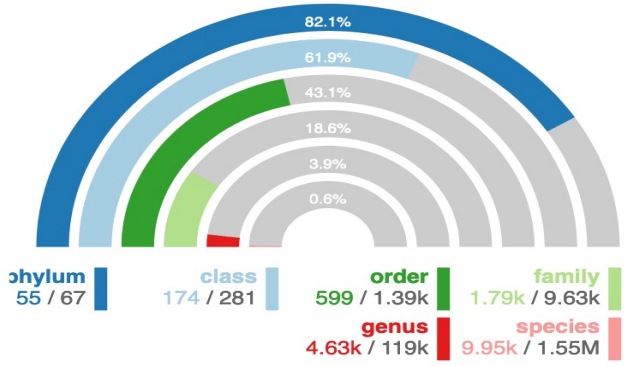
To sequence a 1.5 million known eukaryotic species in 10 years

Earth Biogenome Project (EBP)

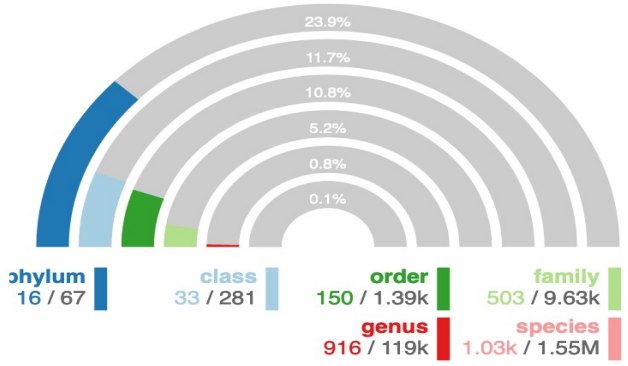
The Earth BioGenome Project (EBP), a moonshot for biology, aims to sequence, catalog and characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years. More information can be found in the EBP official web portal at https://www.earthbiogenome.org.

EBP Contribution to Eukaryotic Genome Sequencing

Progress of Eukaryotic Genome Sequencing by Taxon Rank: all assemblies in INSDC and those submitted under the EBP umbrella (BioProject PRJNA533106)



All INSDC taxa - Taxa with assemblies out of all Eukaryotic taxa in INSDC

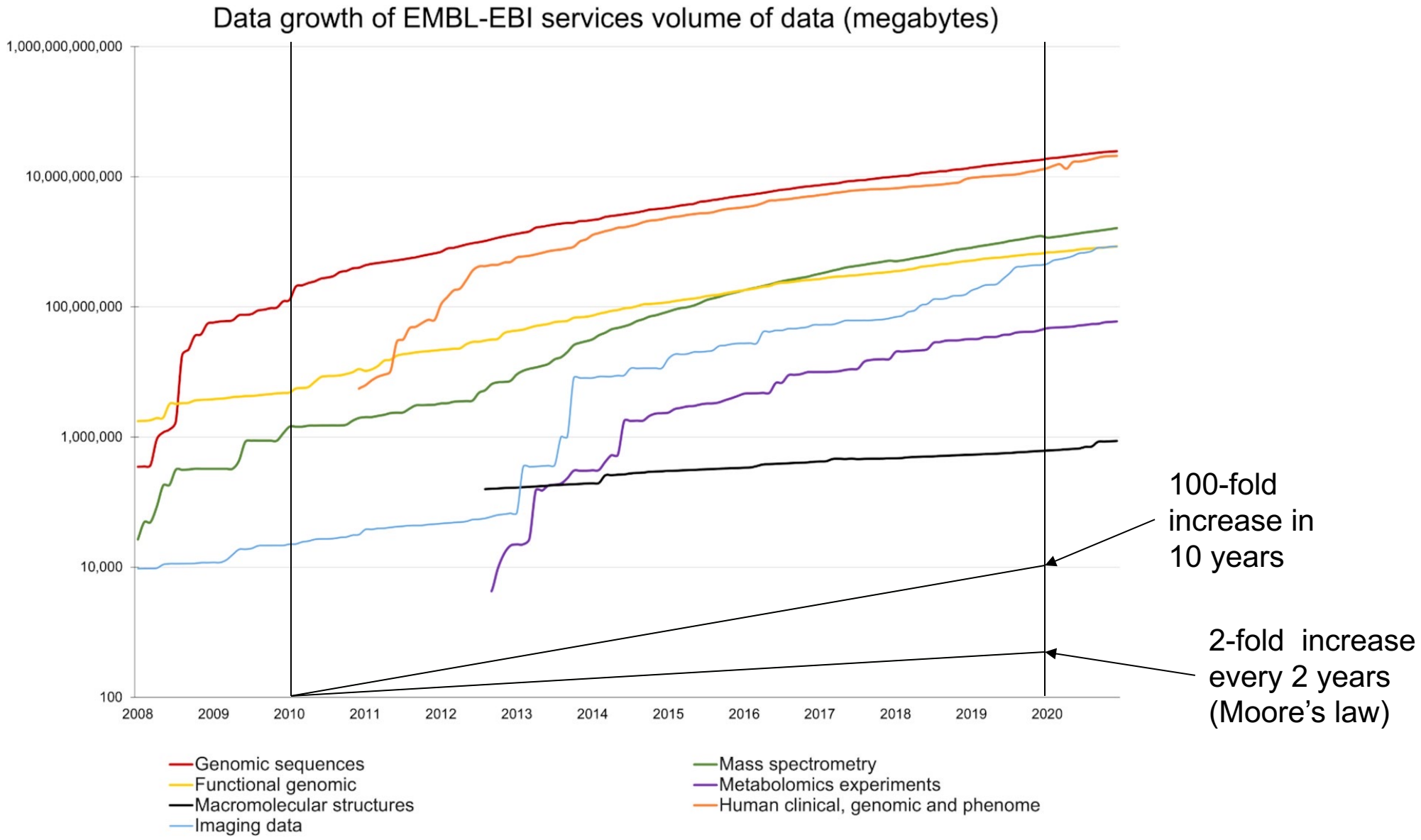


EBP Umbrella - EBP taxa with assemblies out of all Eukaryotic taxa in INSDC

1	superkingdoms
3	kingdoms
16	phyla
33	classes
150	orders
503	families
916	genera
1,025	species
19	subspecies

Counts of taxa sequenced under the EBP Umbrella. Contribution of Projects under EBP Umbrella Project Id PRJNA533106, corresponding to 17 out of the 49 affiliated projects in the EBP Network

Click on report to see and download individually


Data growth of EMBL-EBI services volume of data (megabytes)
1,000,000,000,000
10,000,000,000
100,000,000
1,000,000
10,000
100
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Genomic sequences
Functional genomic
Macromolecular structures
Imaging data
Mass spectrometry
Metabolomics experiments
Human clinical, genomic and phenome
100-fold increase in 10 years
2-fold increase every 2 years (Moore's law)

Cancer Genomics

- Cancer is a genome disease
- Human has ~10^{13} cells, many growing and dividing, there are 2-3 mutations per every cell division
- As important genes, such as DNA-repair genes, get mutated, more mutations accumulate, and genomes get increasingly changed
- Roughly 300 cancer genes are known, mutation in which can lead to cancer
- Some cells escape control and start proliferate – tumors
- Some go on to migrate – metastasis

Vol 464|15 April 2010|doi:10.1038/nature08987 nature

PERSPECTIVES

International network of cancer genome projects

The International Cancer Genome Consortium (ICGC) was launched to coordinate large-scale cancer genome studies in tumours from 50 different cancer types and/or subtypes that are of clinical and societal importance across the globe. Systematic studies of more than 25,000 cancer genomes at the genomic, epigenomic and transcriptomic levels will reveal the repertoire of oncogenic mutations, uncover traces of the mutagenic influences, define clinically relevant subtypes for prognosis and therapeutic management, and enable the development of new cancer therapies

sites at CpG dinucleotides.

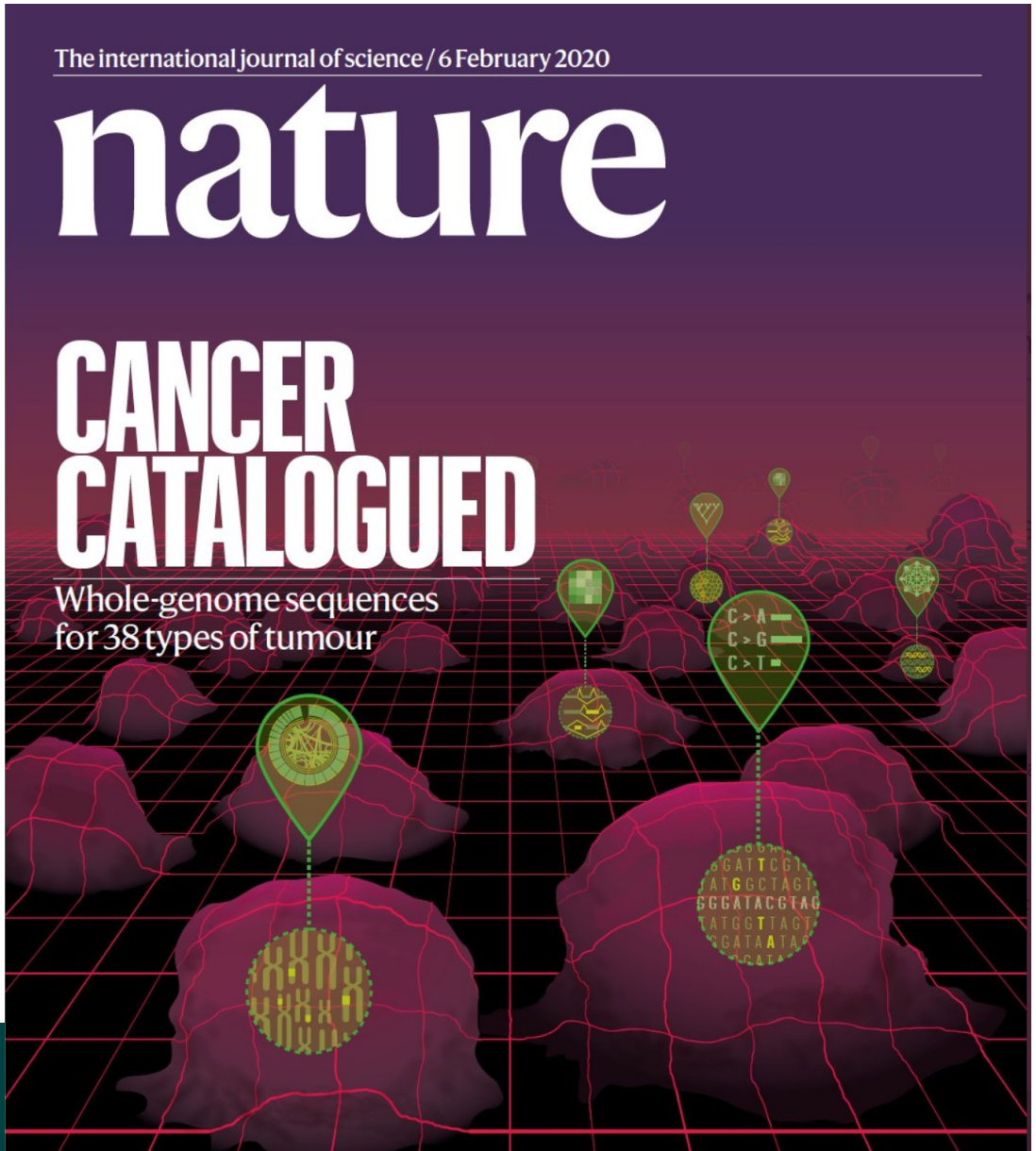
A subset of the somatic mutations in cancer cells confers oncogenic properties such as growth advantage, tissue invasion and metastasis, angiogenesis, and evasion of apoptosis[2]. These are termed 'driver' mutations. The identification of driver mutations will provide insights into cancer biology and highlight new drug targets and diagnostic tests. Knowledge of cancer mutations has already led to the development of specific therapies, such as trastuzumab for *HER2* (also known

the consortium (Box 1) were released in April 2008 (http://www.icgc.org/files/ICGC_April_29_2008.pdf). Since then, working groups and initial member projects have further refined the policies and plans for international collaboration.

Bioethical framework

ICGC members agreed to a core set of bioethical elements for consent as a precondition of membership (Box 2). The Ethics and Policy

Pan-Cancer Analysis of Whole Genomes of ICGC completed 2020



Article

Genomic basis for RNA alterations in cancer

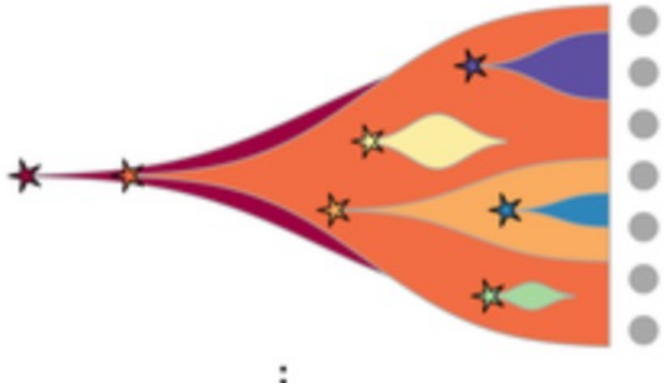
https://doi.org/10.1038/s41586-020-1970-0
Received: 29 March 2018
Accepted: 11 December 2019
Published online: 5 February 2020
Open access

PCAWG Transcriptome Core Group[1,35], Claudia Calabrese[2,35], Natalie R. Davidson[3,4,5,6,7,35], Deniz Demircioğlu[8,9,35], Nuno A. Fonseca[2,35], Yao He[10,35], André Kahles[3,4,6,7,35], Kjong-Van Lehmann[3,4,6,7,35], Fenglin Liu[10,35], Yuichi Shiraishi[11,35], Cameron M. Soulette[12,35], Lara Urban[2,35], Liliana Greger[2], Siliang Li[13,14], Dongbing Liu[13,14], Marc D. Perry[15,16], Qian Xiang[15], Fan Zhang[10], Junjun Zhang[15], Peter Bailey[17], Serap Erkek[18], Katherine A. Hoadley[19], Yong Hou[13,14], Matthew R. Huska[20], Helena Kilpinen[21], Jan O. Korbel[18], Maximillian G. Marin[12], Julia Markowski[20], Tannistha Nandi[9], Qiang Pan-Hammarström[13,22], Chandra Sekhar Pedamallu[23,28,29], Reiner Siebert[24], Stefan G. Stark[3,4,6,7], Hong Su[13,14], Patrick Tan[9,25], Sebastian M. Waszak[18], Christina Yung[15], Shida Zhu[13,14], Philip Awadalla[15,26], Chad J. Creighton[27], Matthew Meyerson[23,28,29], B. F. Francis Ouellette[30], Kui Wu[13,14], Huanming Yang[13], PCAWG Transcriptome Working Group[1], Alvis Brazma[2,36*], Angela N. Brooks[12,23,28,36*], Jonathan Göke[9,31,36], Gunnar Rätsch[3,4,5,6,7,36*], Roland F. Schwarz[2,20,32,33,36], Oliver Stegle[2,18,33,36], Zemin Zhang[10,36] & PCAWG Consortium[34]

Transcript alterations often result from somatic changes in cancer genomes[1]. Various forms of RNA alterations have been described in cancer, including overexpression[2], altered splicing[3] and gene fusions[4]; however, it is difficult to attribute these to underlying genomic changes owing to heterogeneity among patients and tumour types, and the relatively small cohorts of patients for whom samples have been analysed by both transcriptome and whole-genome sequencing. Here we present, to our knowledge, the most comprehensive catalogue of cancer-associated gene alterations to date, obtained by characterizing tumour transcriptomes from 1,188 donors of the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)[5]. Using matched whole-genome sequencing data, we associated several categories of RNA alterations with germline and somatic DNA alterations, and identified probable genetic mechanisms. Somatic copy-number alterations were the

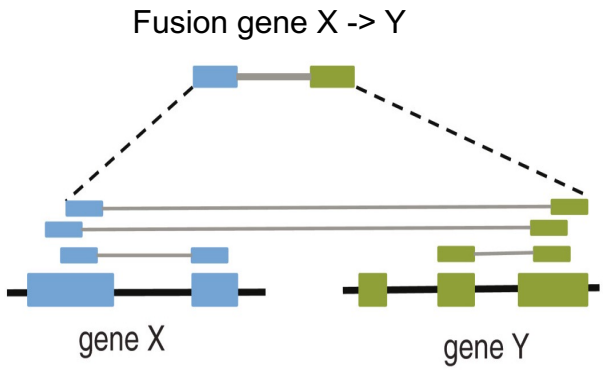
Cancer Genomics

- Cancer genomics is complicated
  - Because of clonality there is not longer one genome
  - To find what is changed in cancer cells we compare normal genome (e.g., from blood cells) to tumor genomes – personalized genome becomes the rererence



Cancer related genome rearrangements cause genes to fuse creating ”cancer genes”

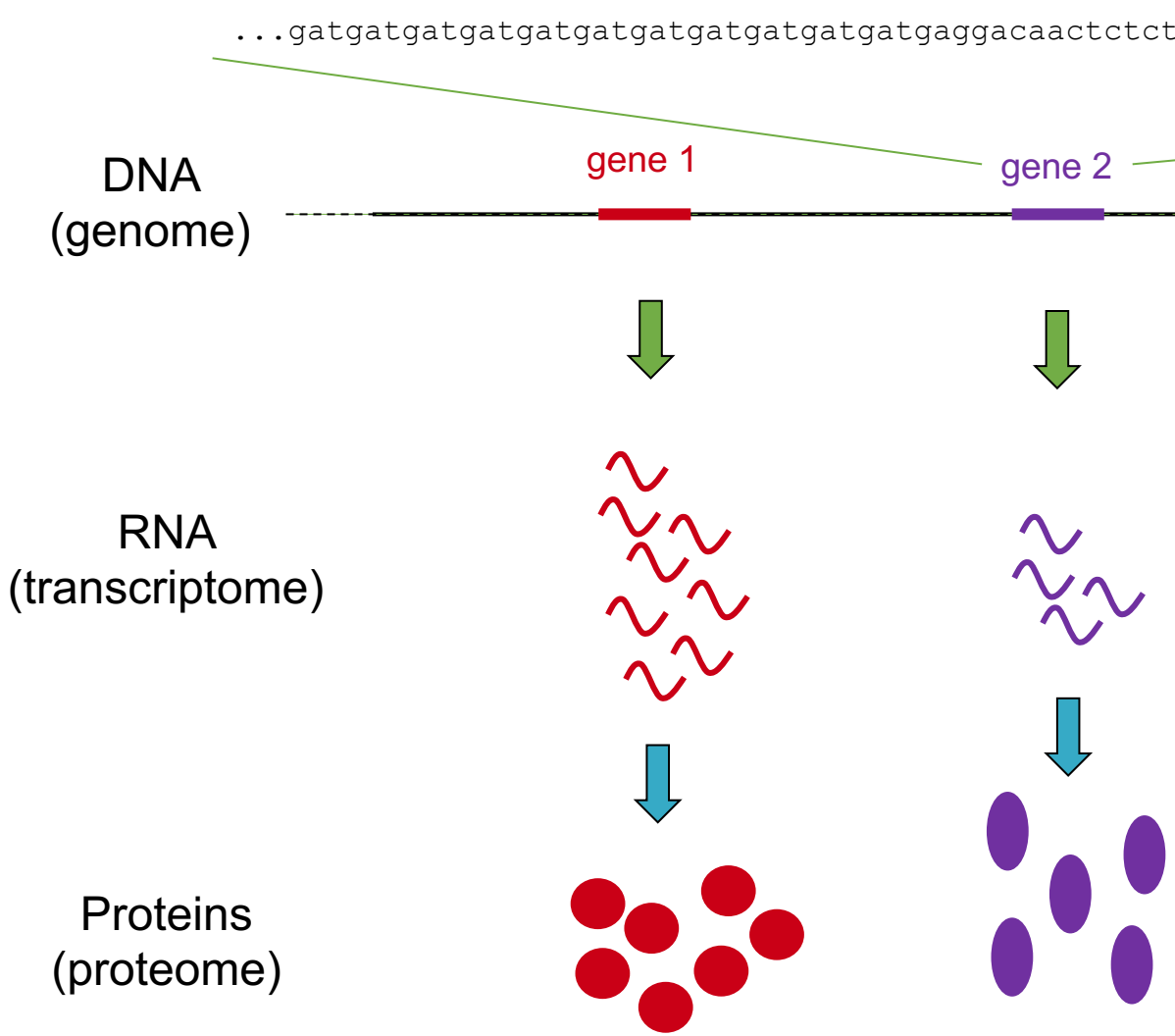
- Hybrid genes formed from two previously separate genes as a result from of genomic rearrangements, read-through transcription and trans-splicing.



  - Fusion genes are tumor-specific and therefore important targets for therapy.

Figure adapted from Edgren et al, Genome Biol. 2011

Central dogma of molecular biology - transcriptomics



- Human genome
- **Transcriptome**
  - different genes are *expressed* in different cells under different conditions (disease vs normal)
  - RNA abundance levels can be assessed by sequencing (RNA-seq)
  - There is at least 1000-fold dynamic range
- Proteome – there is no straight-forward

Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

To Single Cell Expression Atlas

Search across **65 species**, **4,052 studies**, **134,900 assays**

Ensembl 99, Ensembl Genomes 46, WormBase ParaSite 14, EFO 3.10.0

Search | Gene set enrichment

Gene / Gene properties

REG1B ×

Examples: REG1B, zinc finger, O14777 (UniProt), GO:0010468 (regulation of gene expression)

Species

Homo sapiens

Biological conditions

Enter condition query...

Examples: lung, leaf, valproic acid, cancer

Search | Clear

Animals | Plants | Fungi

Homo sapiens

1449 experiments

Baseline: 59

Differential: 1390

Mus musculus

1153 experiments

Baseline: 46

Differential: 1107

Rattus norvegicus

152 experiments

Baseline: 3

Differential: 149

Drosophila melanogaster

140 experiments

Baseline: 4

Differential: 136

Gallus gallus

36 experiments

Baseline: 3

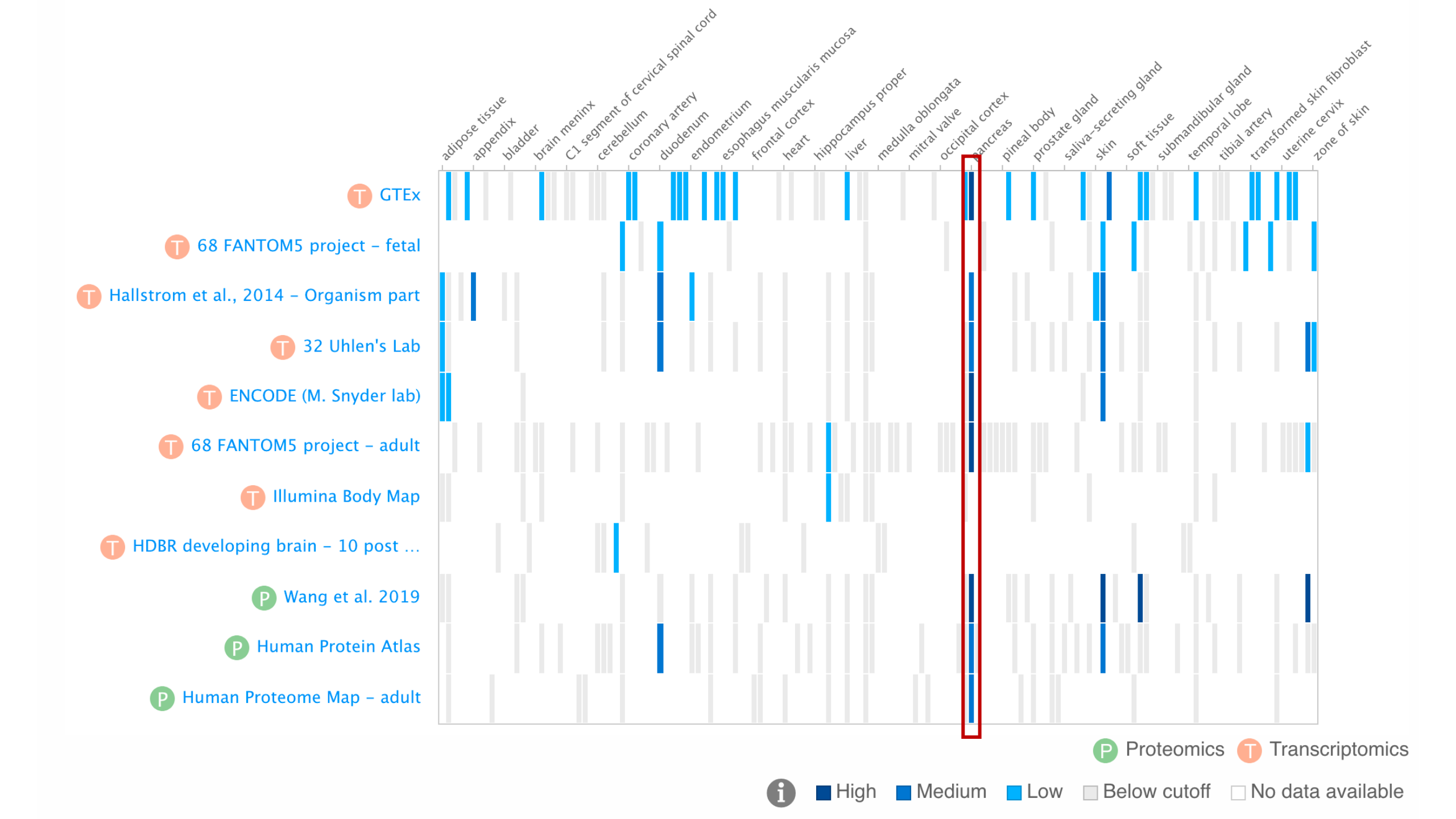
Differential: 33

Caenorhabditis elegans

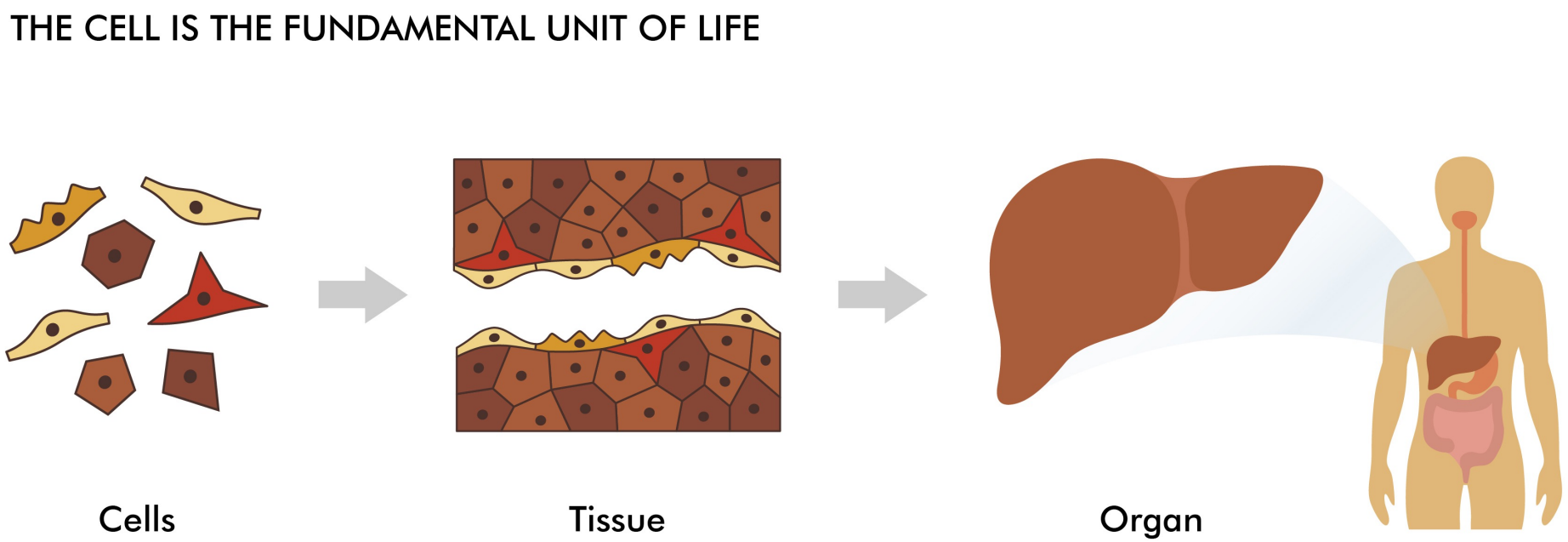
29 experiments

Baseline: 1

Differential: 28

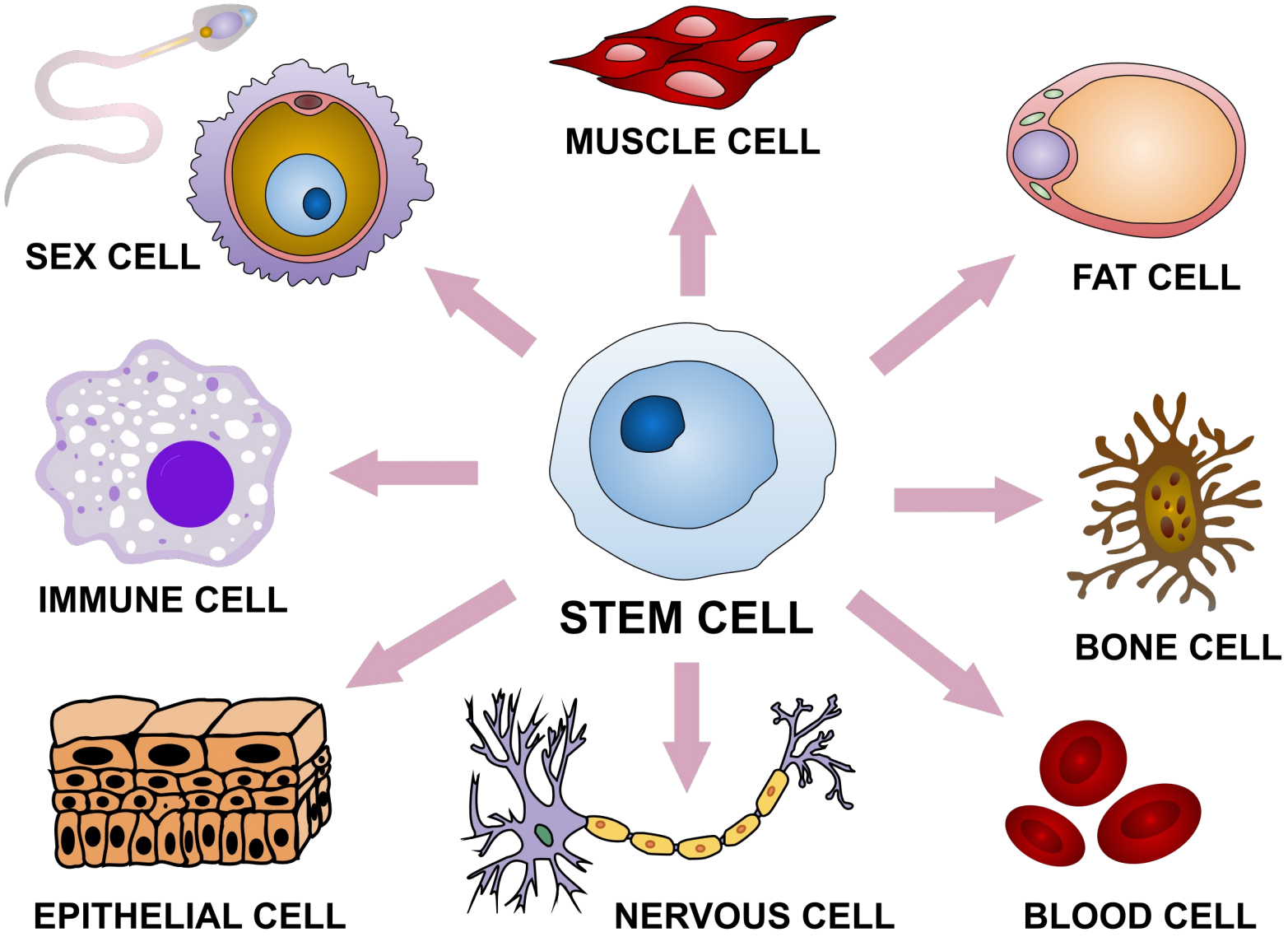

adipose tissue
appendix
bladder
brain meninx
C1 segment of cervical spinal cord
cerebellum
coronary artery
duodenum
endometrium
esophagus muscularis mucosa
frontal cortex
heart
hippocampus proper
liver
medulla oblongata
mitral valve
occipital cortex
pancreas
pineal body
prostate gland
saliva-secreting gland
skin
soft tissue
submandibular gland
temporal lobe
tibial artery
transformed skin fibroblast
uterine cervix
zone of skin
GTEx
68 FANTOM5 project - fetal
Hallstrom et al., 2014 - Organism part
32 Uhlen's Lab
ENCODE (M. Snyder lab)
68 FANTOM5 project - adult
Illumina Body Map
HDBR developing brain - 10 post ...
Wang et al. 2019
Human Protein Atlas
Human Proteome Map - adult
Proteomics
Transcriptomics
High
Medium
Low
Below cutoff
No data available

Tissues consists of cells of different types



It is often said that human has ~200 different cell types

Human body contains >10^{13} cells of 200 different types



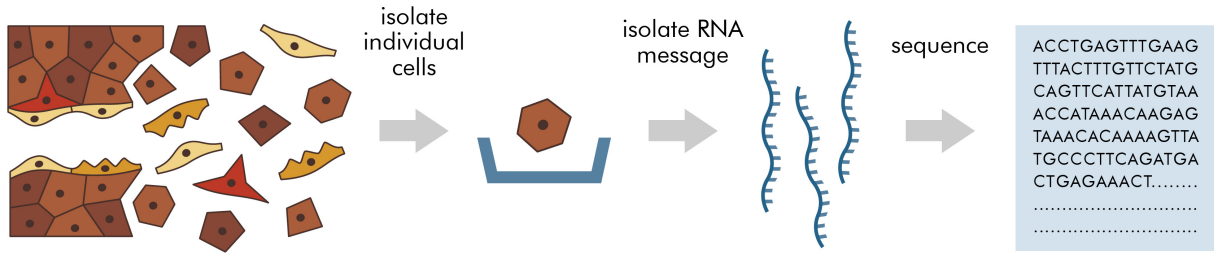
EMBL

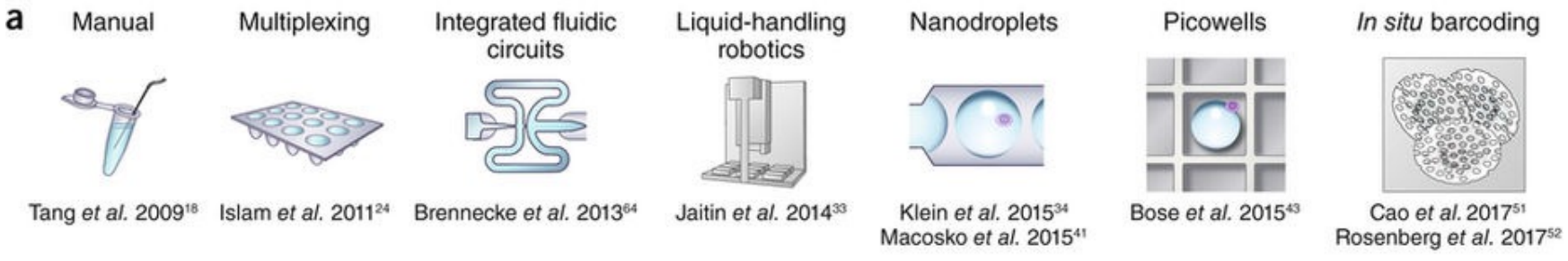
From looking at gene expression in tissues to individual cells

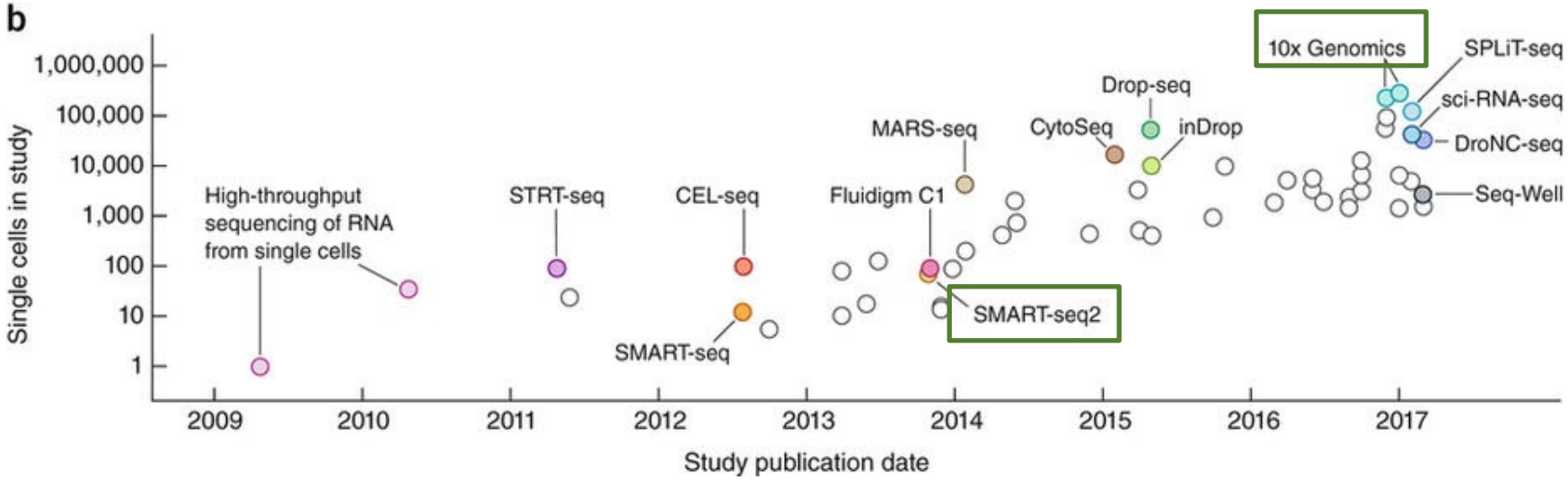
Single-cell genomics

EMBL

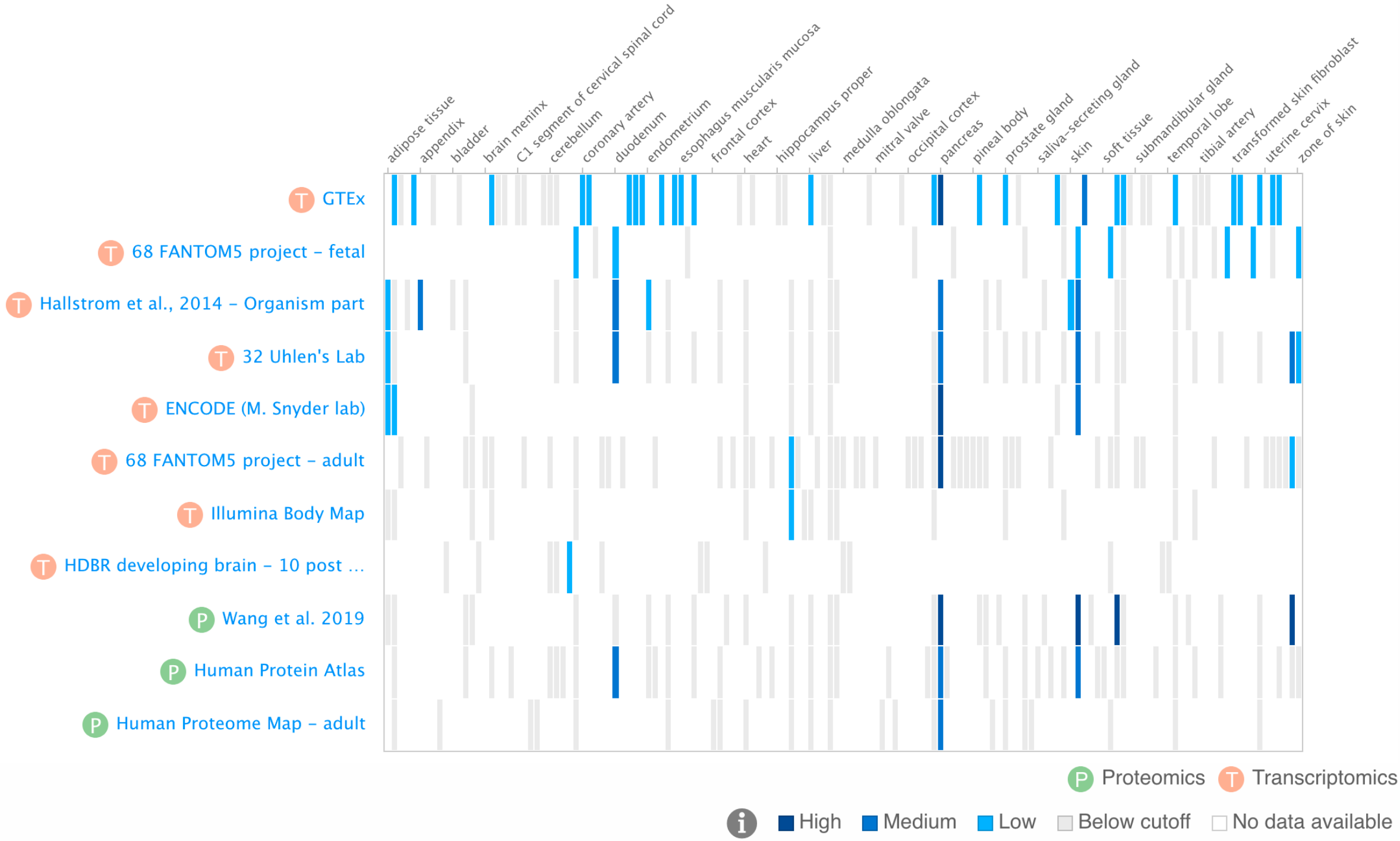
Single Cell RNA sequencing







EMBL


adipose tissue
appendix
bladder
brain meninx
C1 segment of cervical spinal cord
cerebellum
coronary artery
duodenum
endometrium
esophagus muscularis mucosa
frontal cortex
heart
hippocampus proper
liver
medulla oblongata
mitral valve
occipital cortex
pancreas
pineal body
prostate gland
saliva-secreting gland
skin
soft tissue
submandibular gland
temporal lobe
tibial artery
transformed skin fibroblast
uterine cervix
zone of skin
GTEx
68 FANTOM5 project - fetal
Hallstrom et al., 2014 - Organism part
32 Uhlen's Lab
ENCODE (M. Snyder lab)
68 FANTOM5 project - adult
Illumina Body Map
HDBR developing brain - 10 post ...
Wang et al. 2019
Human Protein Atlas
Human Proteome Map - adult
Proteomics
Transcriptomics
High
Medium
Low
Below cutoff
No data available

Single Cell Expression Atlas

Single cell transcriptome analysis of human pancreas

Single-cell RNA-Seq mRNA baseline

Organism: *Homo sapiens*

Publication:

- Enge M, Arda HE, Mignardi M, Beausang J, Bottino R et al. (2017) *Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns.*

REG1B expression in human pancreas

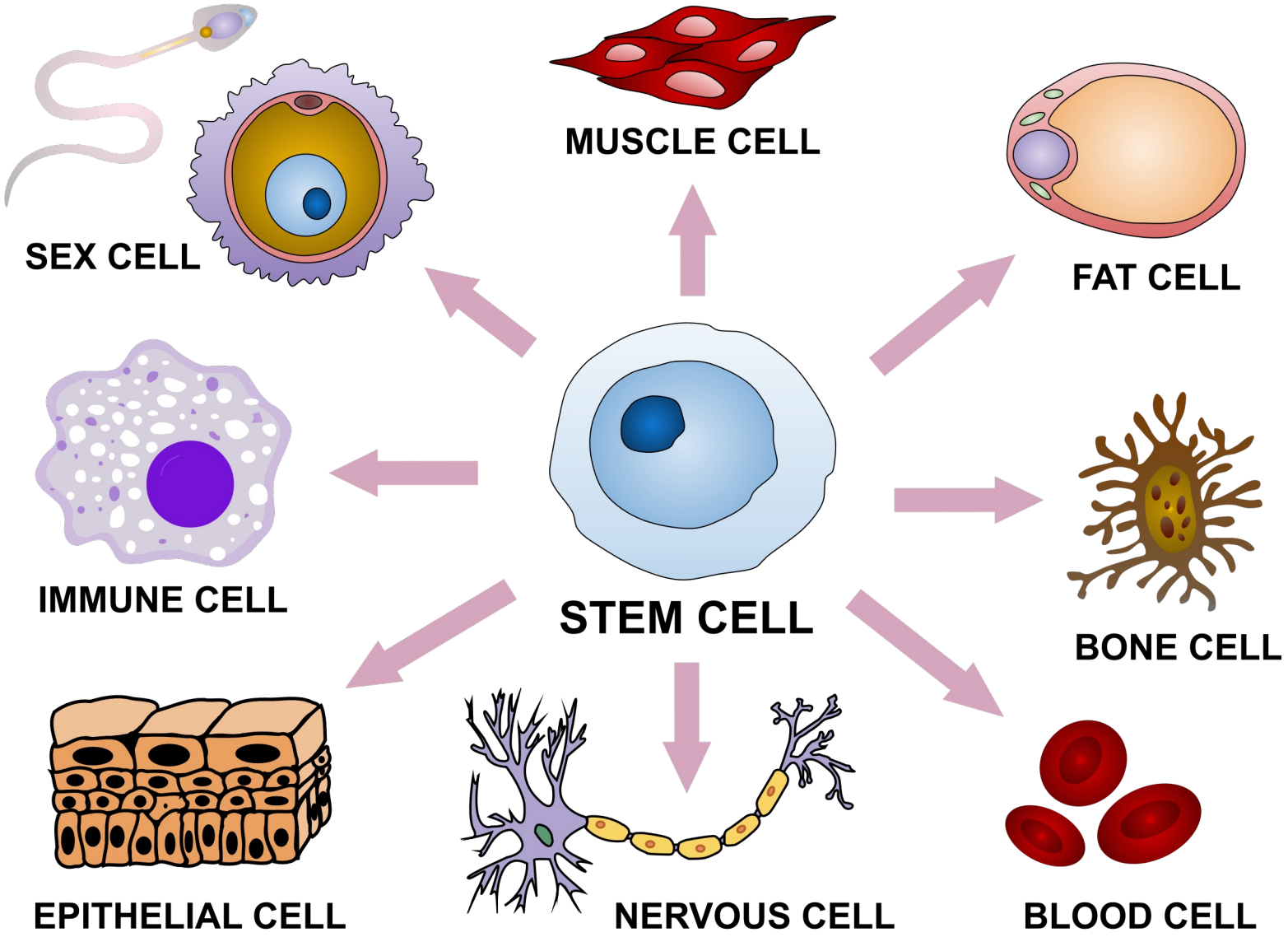
Gene expression

Acinar cells

TSN plots (non-liner PCR)

EMBL

Human body contains >10^{13} cells of 200 different types

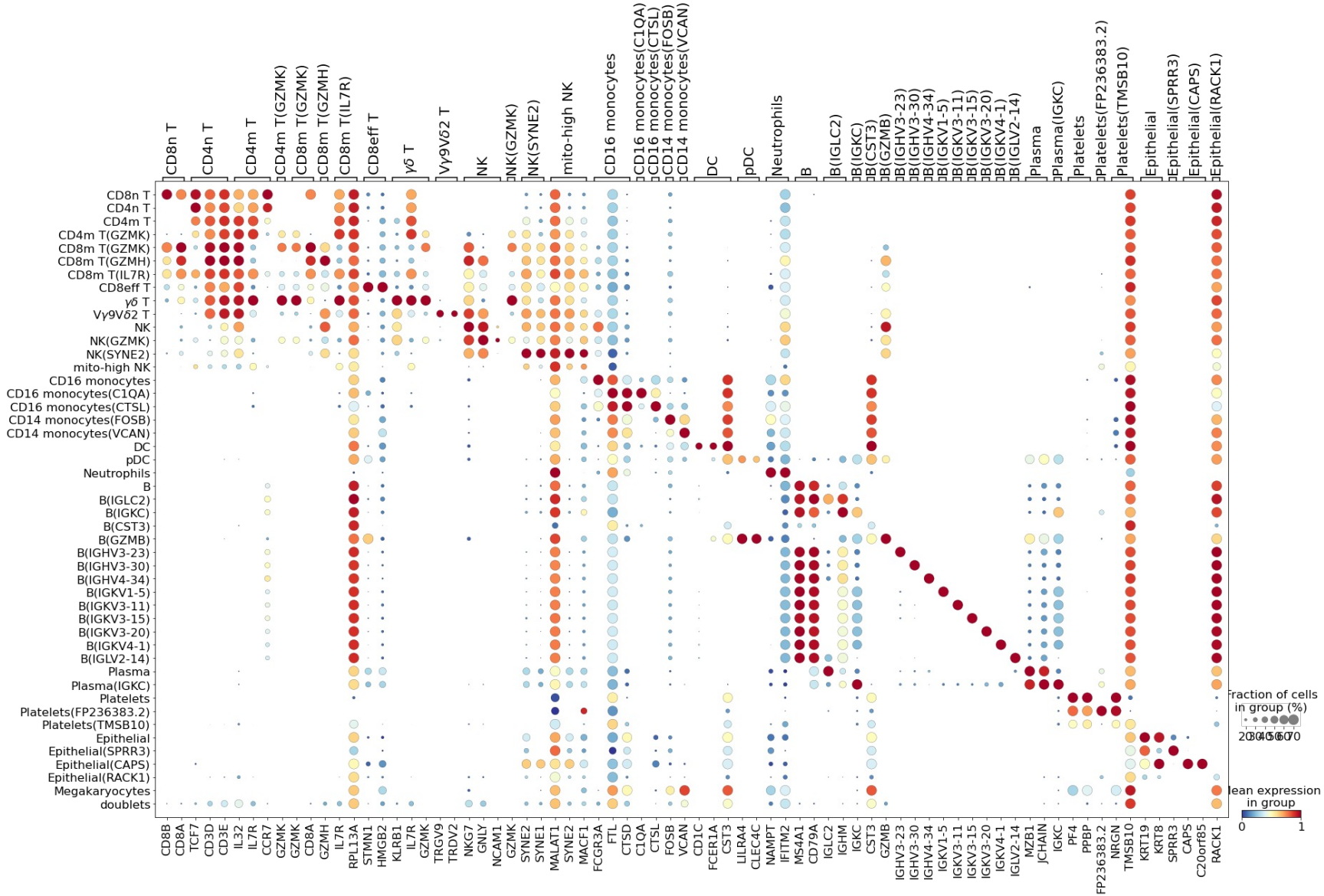


EMBL

What is a cell type?

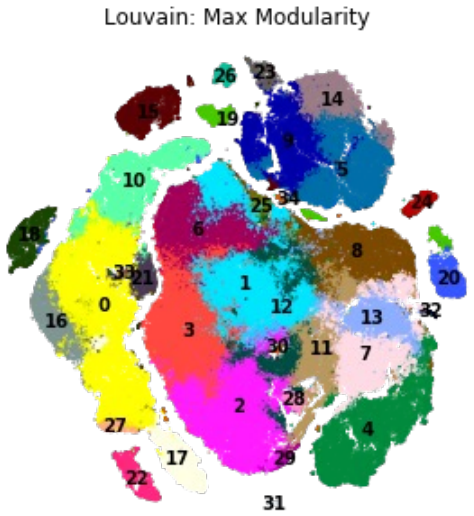
- A cell type is a classification traditionally used to distinguish between morphologically or phenotypically distinct cell forms within a species
- Nowadays cell types are typically defined through sets of genes specifically expressed there, called ***marker genes***
- Both definitions are only proxies and there is no generally agreed definition of this concept

Marker genes for immune cells



EMBL

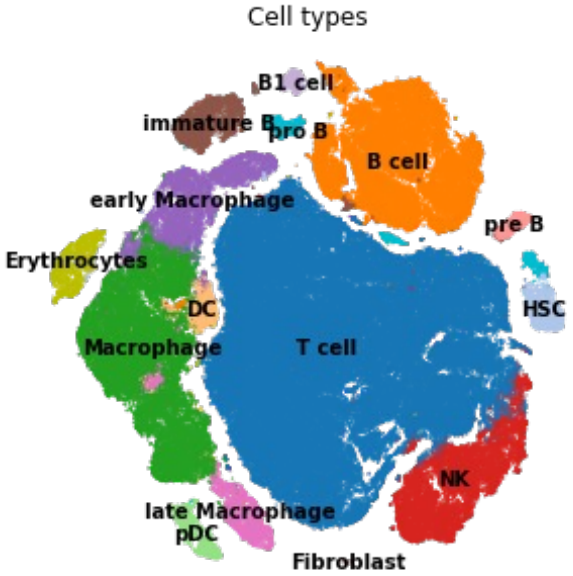
Automated approaches using known marker genes



Clustering

Look for marker genes in each cluster

Merge clusters of same type



These methods do not allow for discovering new cell “types” of states

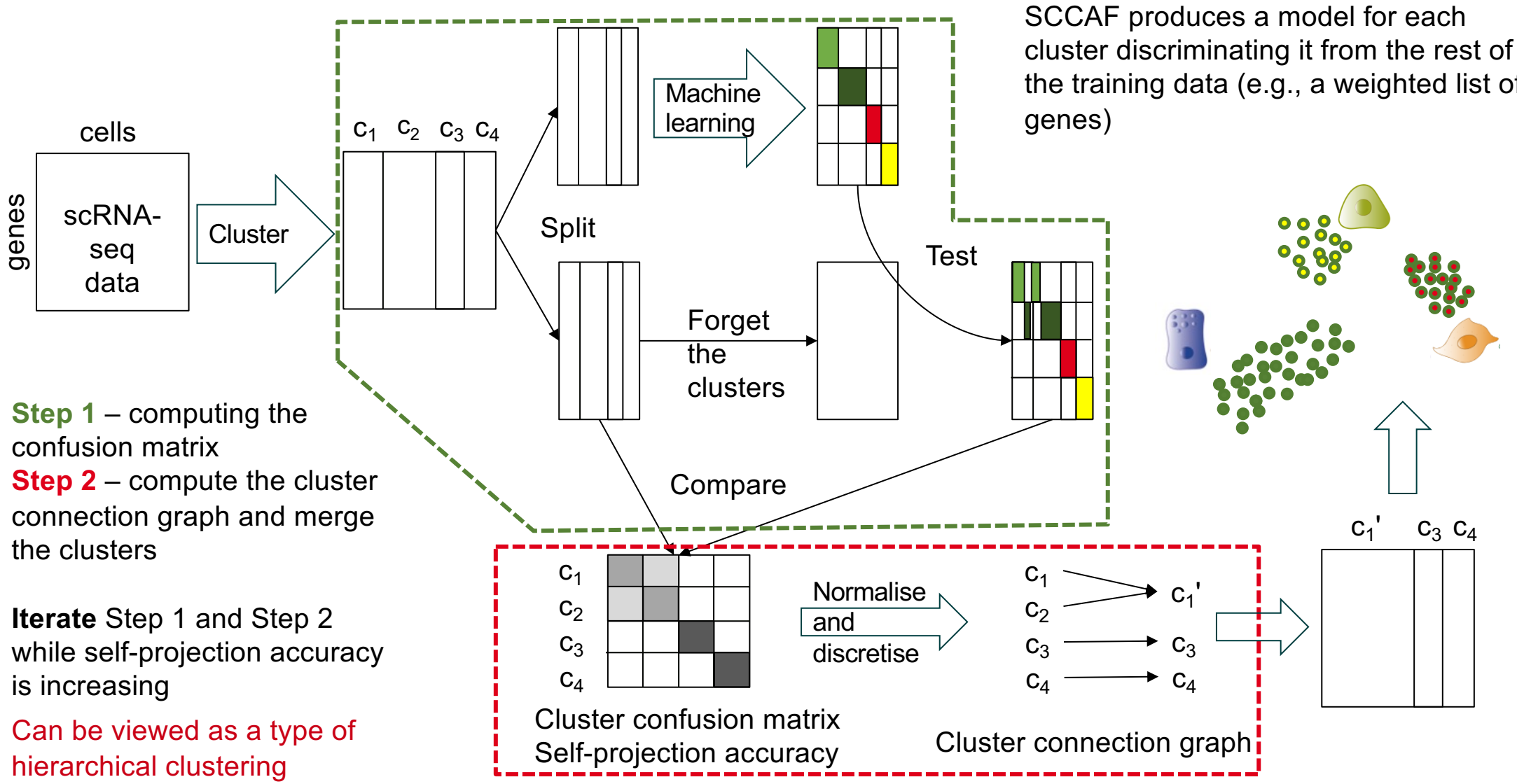
EMBL

We wanted to automate the cell type identification process in a way that allows for **discovery of new cell "types"** as well as new marker genes automatically

- Problem of finding the cell clusters corresponding to biologically meaningful cell groups (cell "types")
- Building gene expression "models" for each group – which are the genes that are differentially expressed for the group?

- Single Cell Clustering Assessment Framework - SCCAF

SCCAF in a nutshell:



Step 1 – computing the confusion matrix

Step 2 – compute the cluster connection graph and merge the clusters

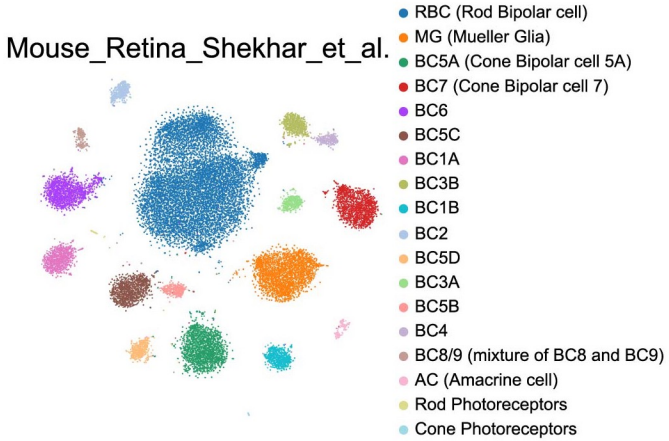
Iterate Step 1 and Step 2 while self-projection accuracy is increasing

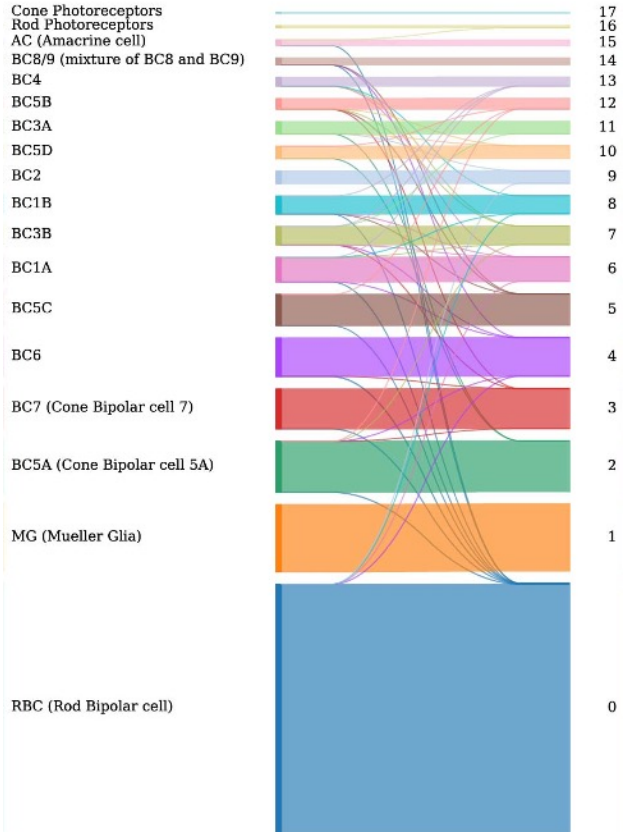
Can be viewed as a type of hierarchical clustering

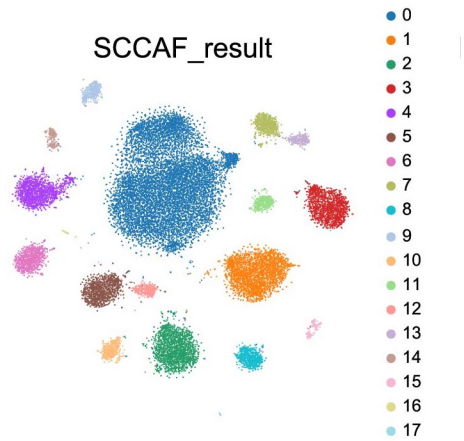
Miao et al, *Nature Methods*, 2020

EMBL

Comparison with expert annotation of the mouse retina dataset from Shekhar et al. 2016



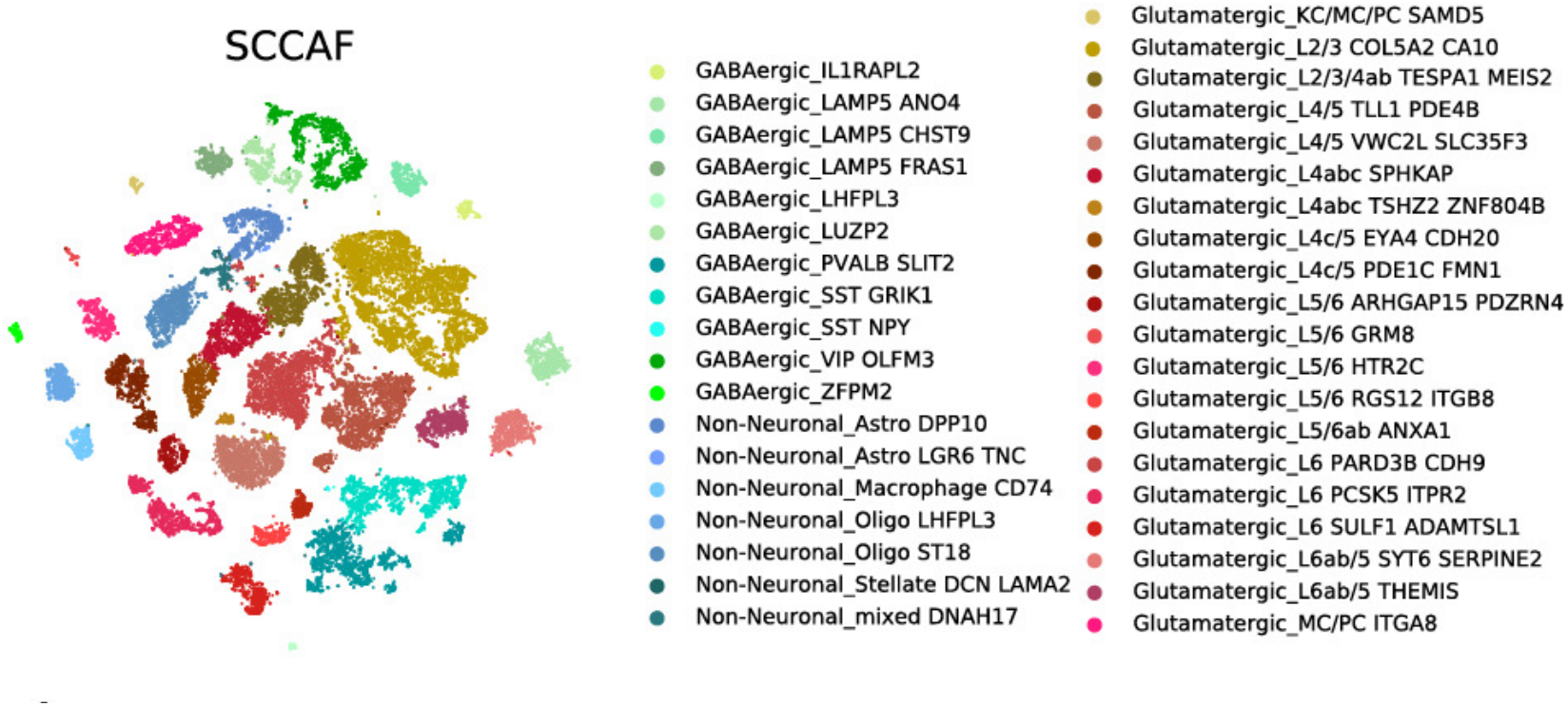




Adjusted Rand index >0.99.

EMBL

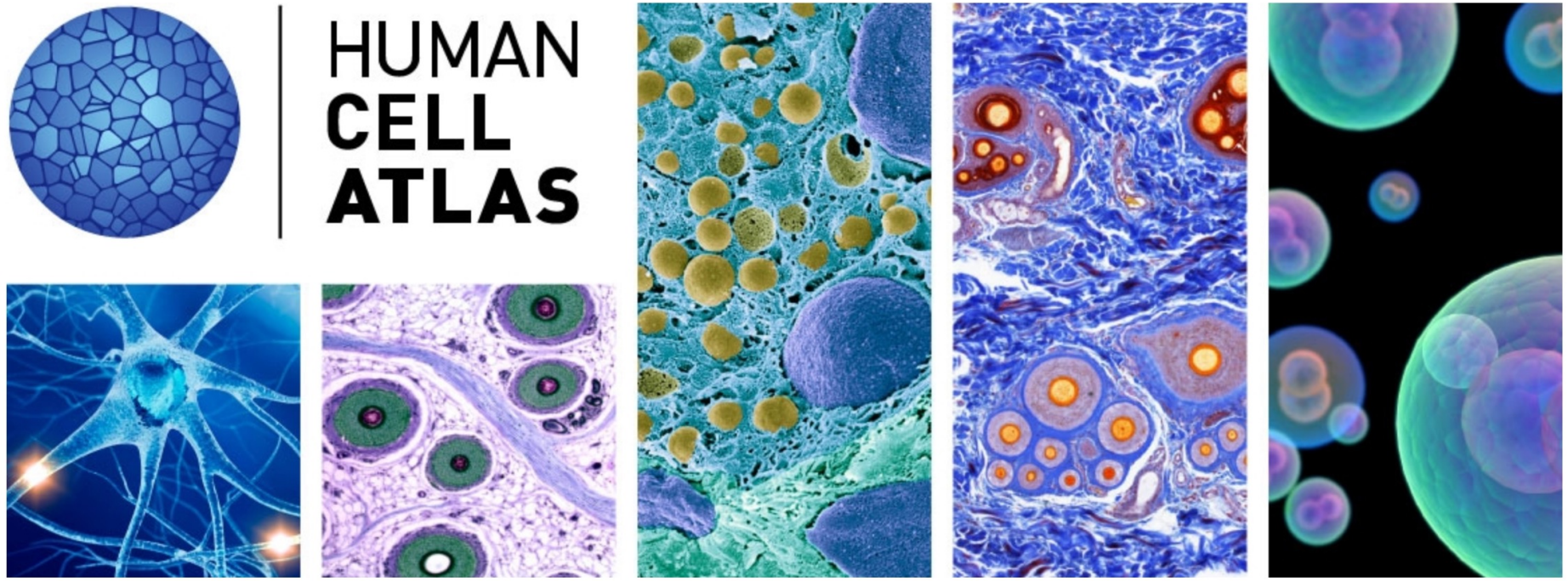
Annotating human brain data based on model trained on mouse (Allen Brain Atlas)



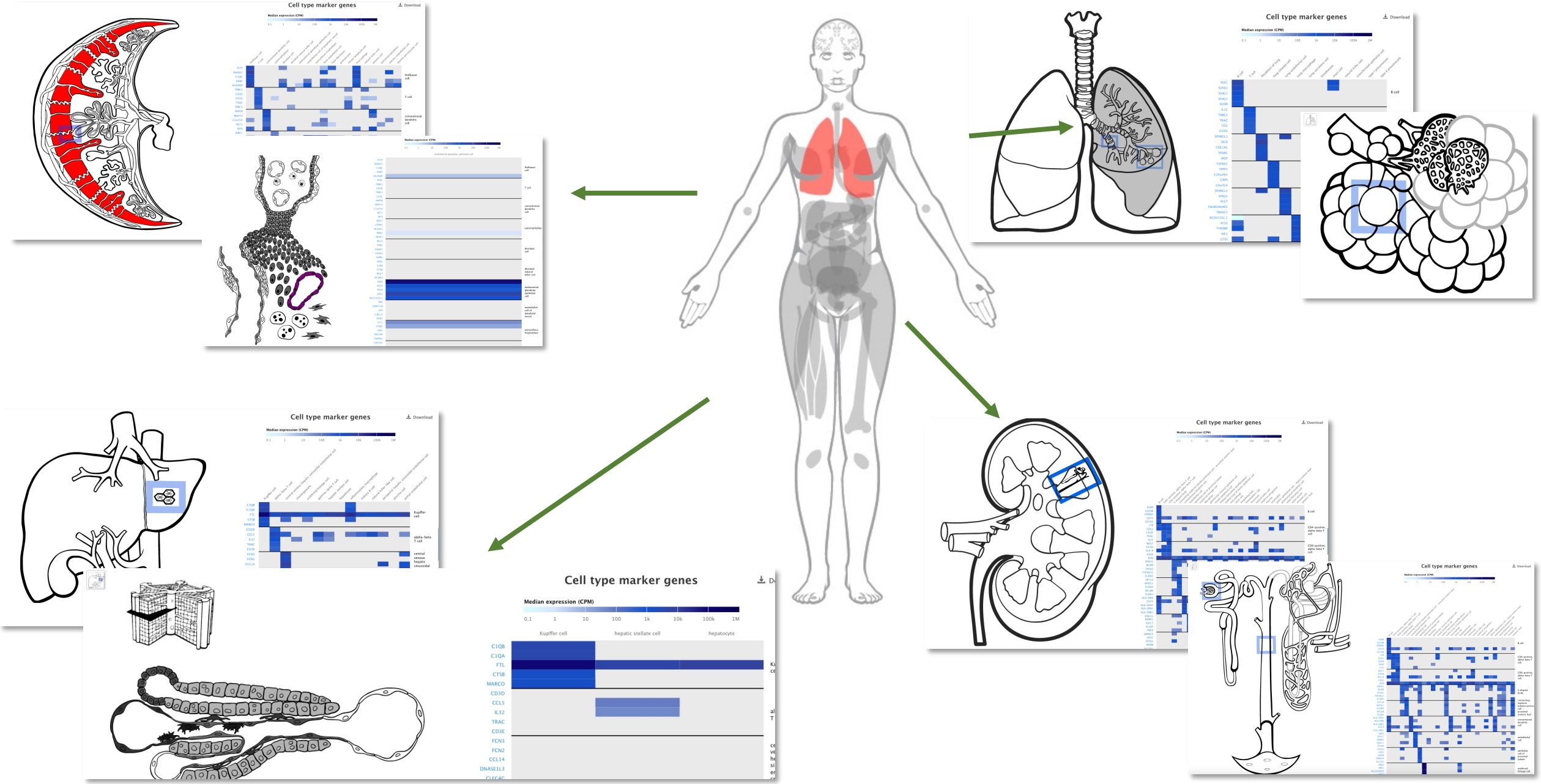
SCCAF has been implemented in the Expression Atlas Galaxy pipeline

Allen Cell Types Database

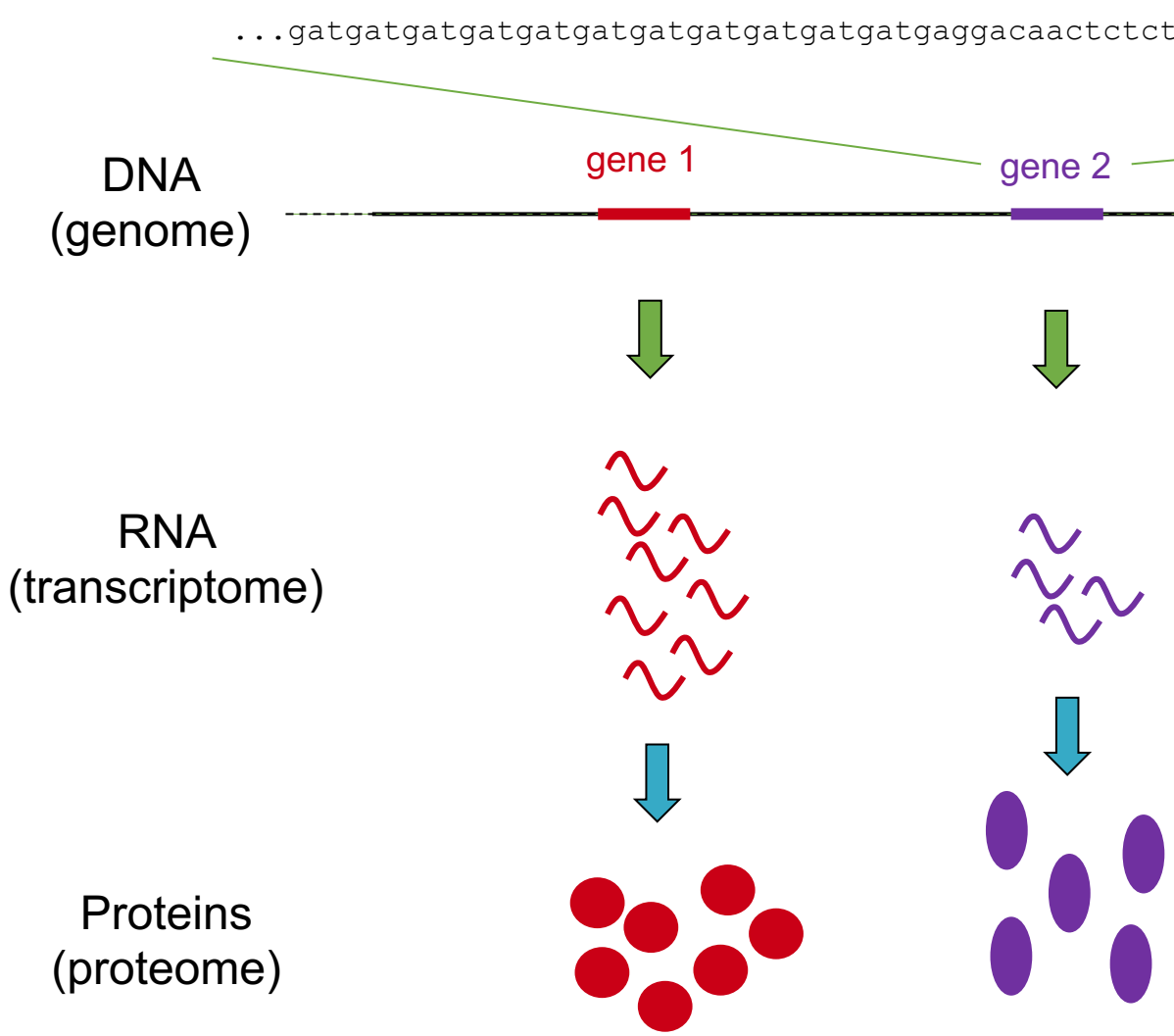
EMBL



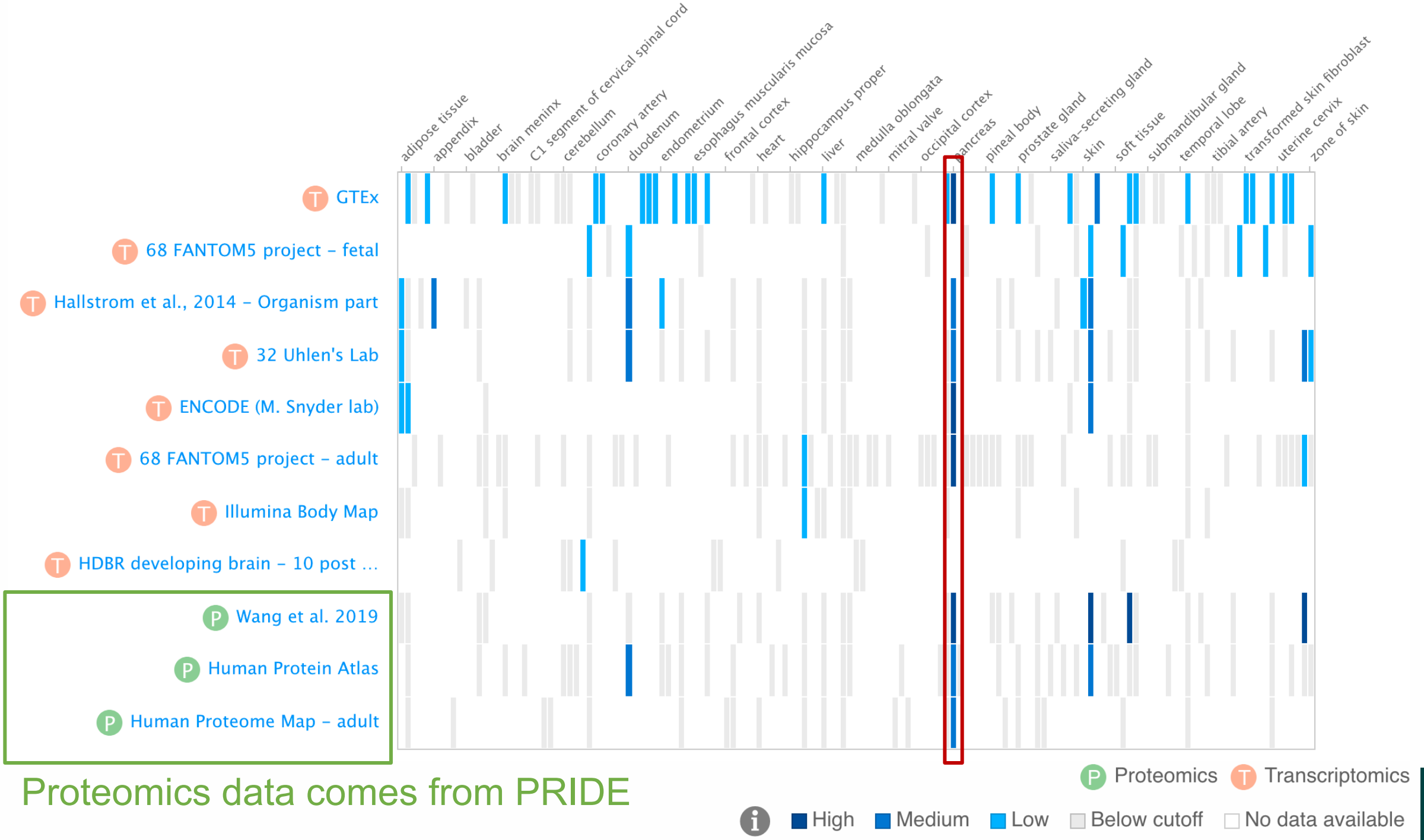
The Human Cell Atlas (HCA) is a global partnership of scientists who are actively working to create an exhaustive guidebook of the types and properties of all human cells.


Cell type marker genes
Download
Median expression (CPM)
0.1
1
10
100
1k
10k
100k
1M
Kupffer cell
hepatic stellate cell
hepatocyte
C1QB
C1QA
FTL
CTSB
MARCO
CD3D
CCL5
IL32
TRAC
CD3E
FCN3
FCN2
CCL14
DNASE1L3
B cell
T cell
alpha-beta T cell
central venous hepatic sinusoidal

Central dogma of molecular biology

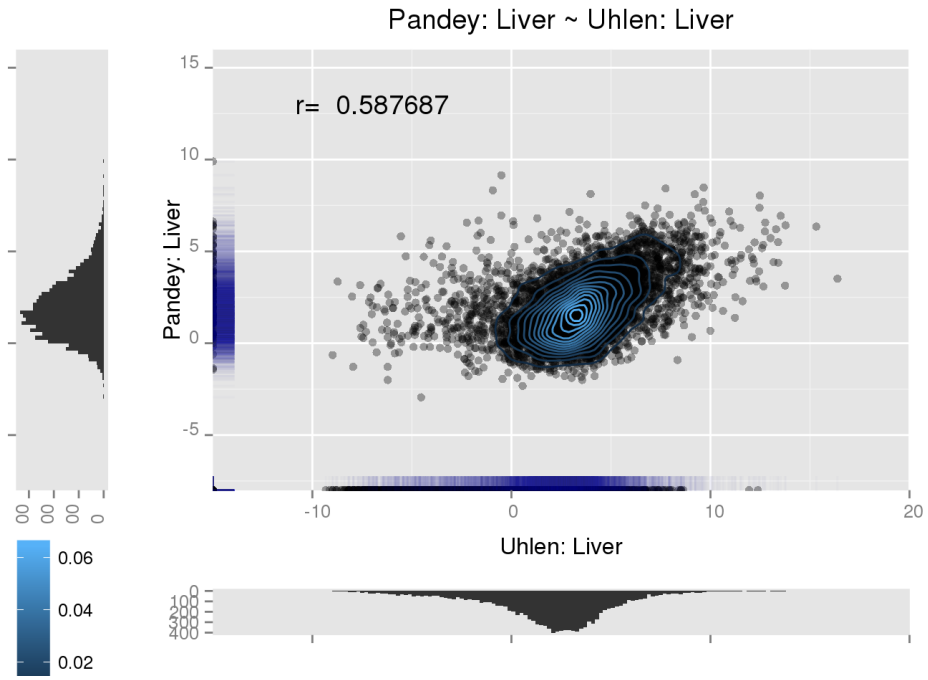


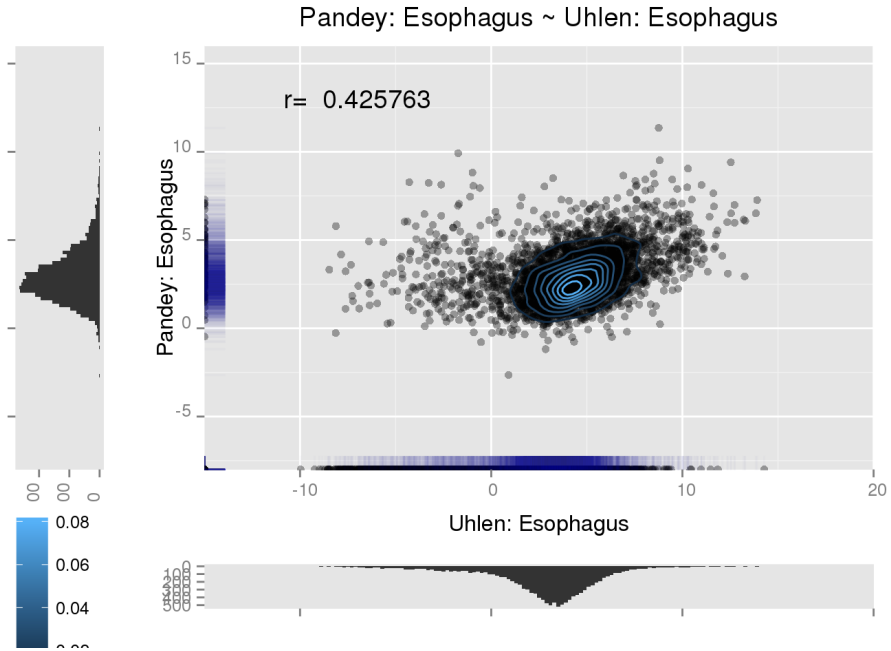
- Human genome
- Transcriptome
- Proteome
  - Protein abundance is what really matters but it is difficult to assess
  - The dynamic range is in 1,000,000s
  - There is no straight-forward correlation between RNA and protein abundances
  - Can we predict the proteome computationally?



Proteomics data comes from PRIDE

What can we say about protein expression from RNA?





Correlation between RNA and protein abundances are not all that high but can we predict protein abundances from RNA abundances? Moreover, can we predict proteins that are difficult to measure diretly?

Proteomics
Proteomics and Systems Biology

Research Article | Open Access

Using Deep Learning to Extrapolate Protein Expression Measurements

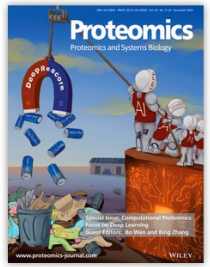
Mitra Parissa Barzine, Karlis Freivalds, James C. Wright, Mārtiņš Opmanis, Darta Rituma, Fatemeh Zamanzad Ghavidel, Andrew F. Jarnuczak, Edgars Celms, Kārlis Čerāns, Inge Jonassen, Lelde Lace, Juan Antonio Vizcaíno, Jyoti Sharma Choudhary ✉, Alvis Brazma ✉, Juris Viksna ✉ ... See fewer authors

First published: 16 September 2020 | https://doi.org/10.1002/pmic.202000009 | Citations: 1

Abstract

Mass spectrometry (MS)-based quantitative proteomics experiments typically assay a subset of up to 60% of the ≈20 000 human protein coding genes. Computational methods for imputing the missing values using RNA expression data usually allow only for imputations of proteins measured in at least some of the samples. In silico methods for comprehensively estimating abundances across all proteins are still missing.

Here, a novel method is proposed using deep learning to extrapolate the observed protein expression values in label-free MS experiments to all proteins, leveraging gene functional annotations and RNA measurements as key predictive attributes. This method is tested on four datasets, including human cell lines and human and mouse tissues. This method predicts the protein expression values with average R^2 scores between 0.46 and 0.54, which is significantly better than predictions based on correlations using the RNA expression data alone. Moreover, it is demonstrated that the derived models can be "transferred" across experiments and species. For instance, the model derived from human tissues gave a $R^2 = 0.51$ when applied to mouse tissue data. It is concluded that



Volume 20, Issue 21-22
Special Issue: Computational Proteomics: Focus on Deep Learning
November 2020
2000009

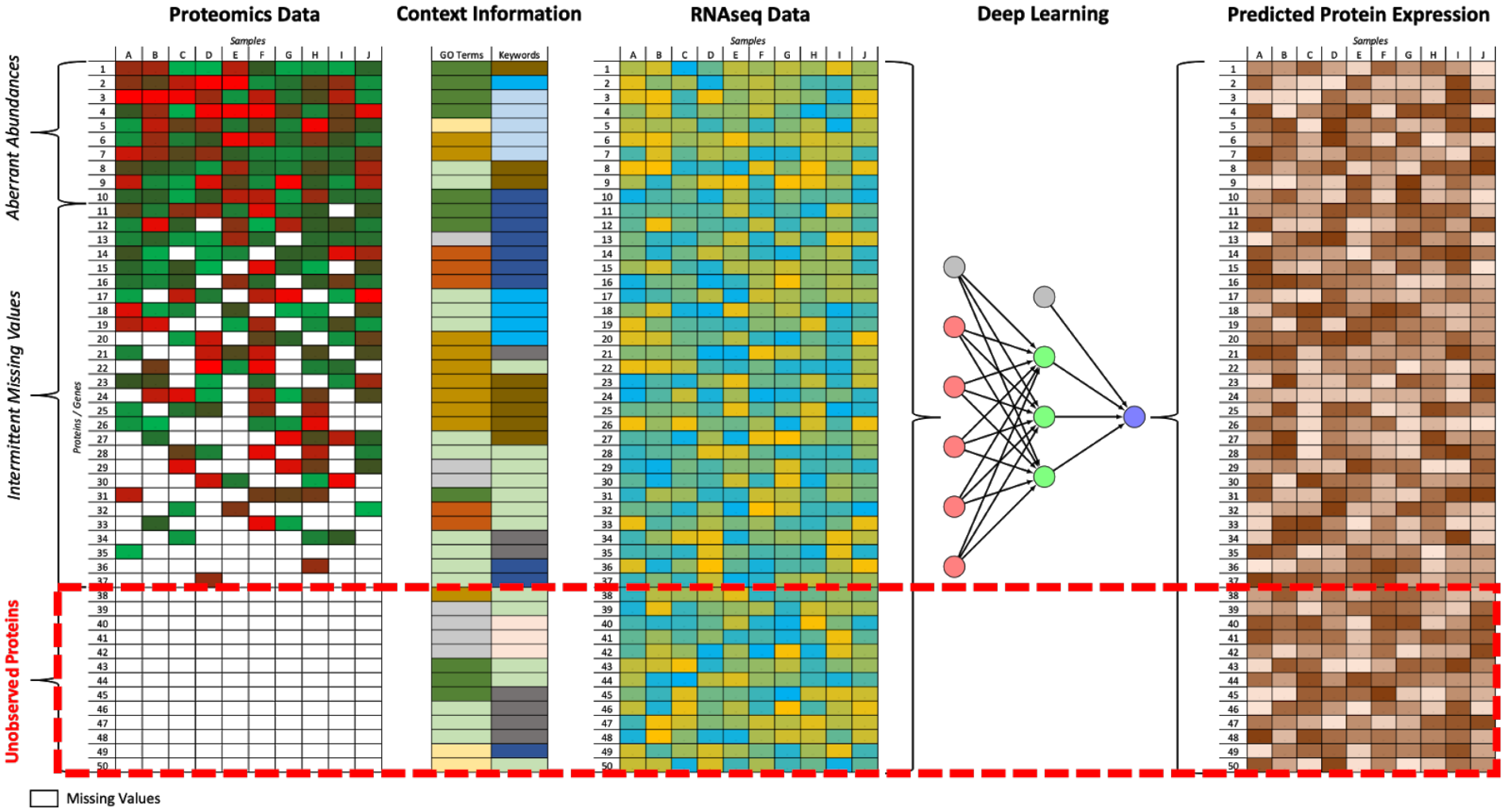
Metrics

Citations: 1

Am score 4

Details

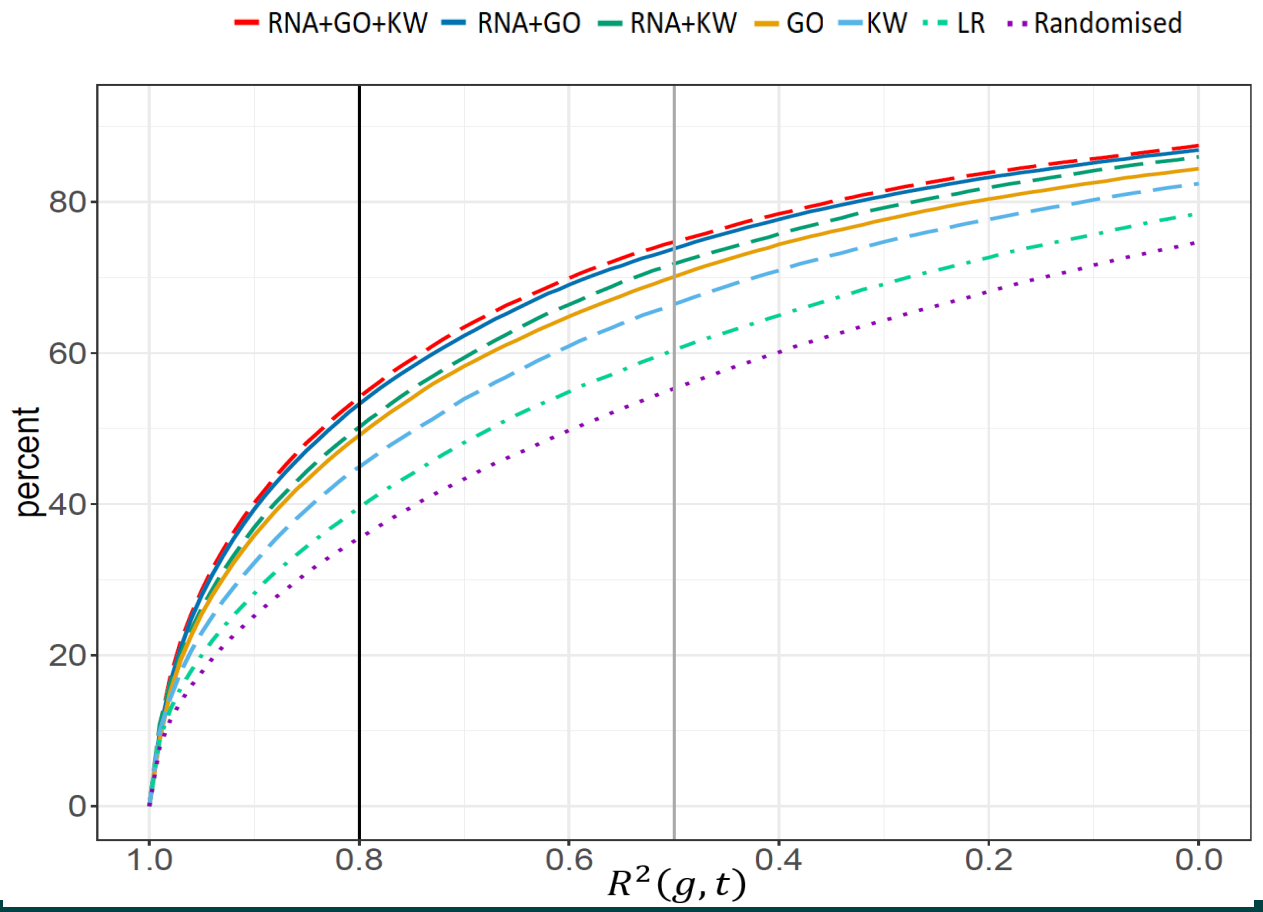
But what about predicting protein expression from RNA and other data?



Barzine et al, *Proteomics*, Nov 2020

EMBL

Deep learning allows to make good predictions



$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

Conclusions

- Biology is increasingly becoming a data science (like in most human activities, there is a shift from material processing to data processing)
- Biological and health data are growing faster than Moore's law creating challenges and opportunities for computer and data scientists
- Growing importance of AI and ML to biological data analysis
- Personalized genomics is having an impact on healthcare now