Introduction to Systems Biology
Class 01

Katia de Paiva Lopes, PhD

Rush Alzheimer's Disease Center (RADC)

Instituto de Assistência Médica ao Servidor Público Estadual de São Paulo (IAMSPE)

Universidade Federal do Paraná (UFPR)

What is this course about + goals

Big picture!

High-level understanding of biological systems with a focus on Networks for Medicine and Health Sciences

Major goal: Hands-on practice!

Systems Biology as Defined by NIH

"Ask five different astrophysicists to define a black hole, the saying goes, and you'll get five different answers. But ask five biomedical researchers to **define systems biology**, and you'll get **10 different answers** . . . or maybe more."

Christopher Wanjek

"Some people think of it as **bioinformatics**, taking an enormous amount of information and processing it. The other school of thought thinks of it as **computational biology**, computing on how the systems work. You need both of these parts."

Ron Germain, chief of NIAID's Laboratory of Systems Biology

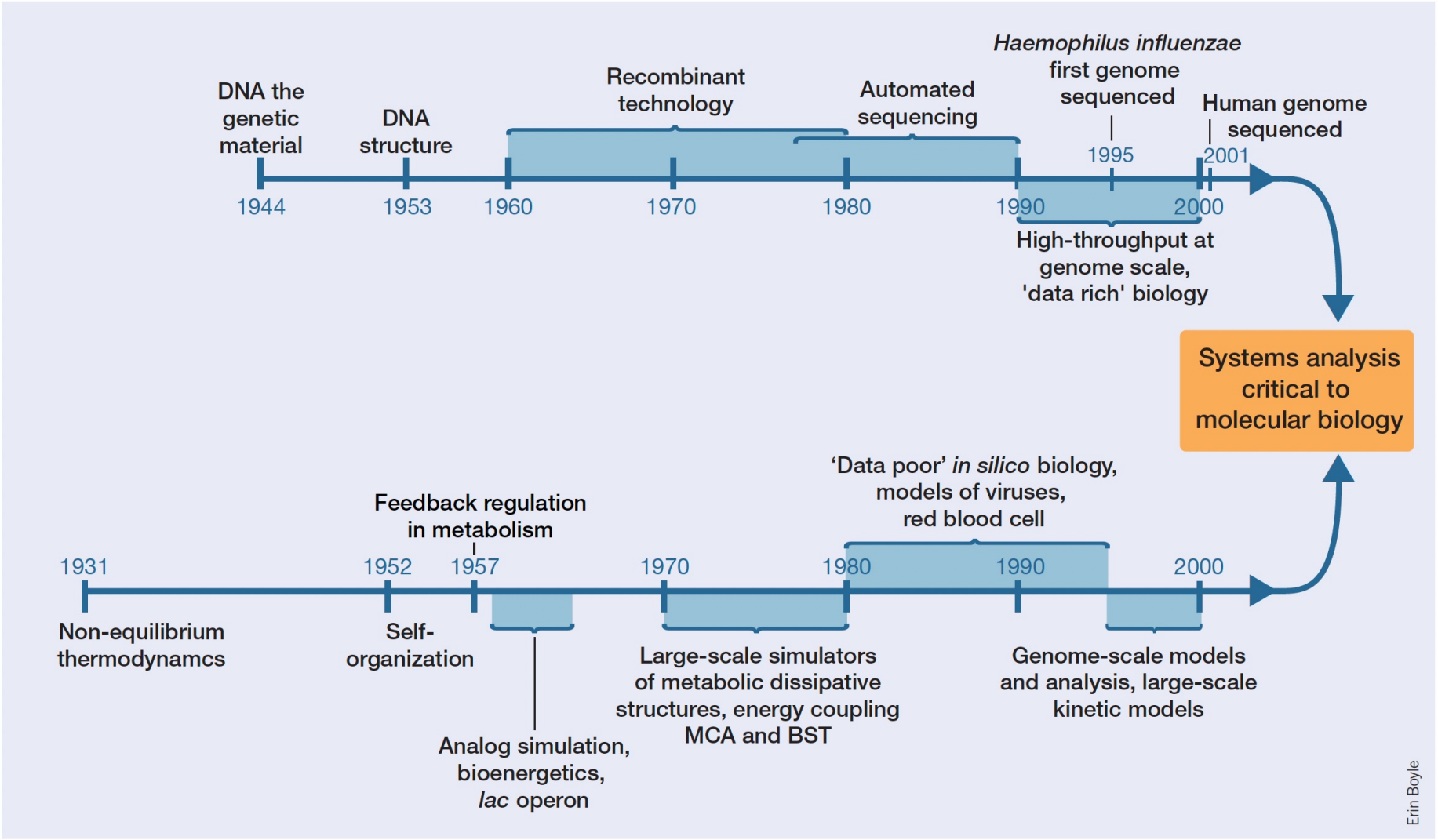
https://irp.nih.gov/catalyst/19/6/systems-biology-as-defined-by-nih

I want to listen to you

Define Systems Biology

Systems Biology – the origins

Systems Biology – brief timeline

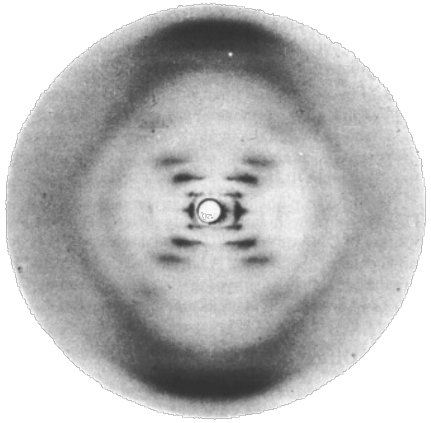


Westerhoff & Palsson, Nat Biotechnology, 2004

Hallmarks

1953 – Elucidation of DNA structure

"Molecular structure of nucleic acids"
Watson and Crick, Nature 1953 .



X-ray crystallography
By Rosalind Franklyn

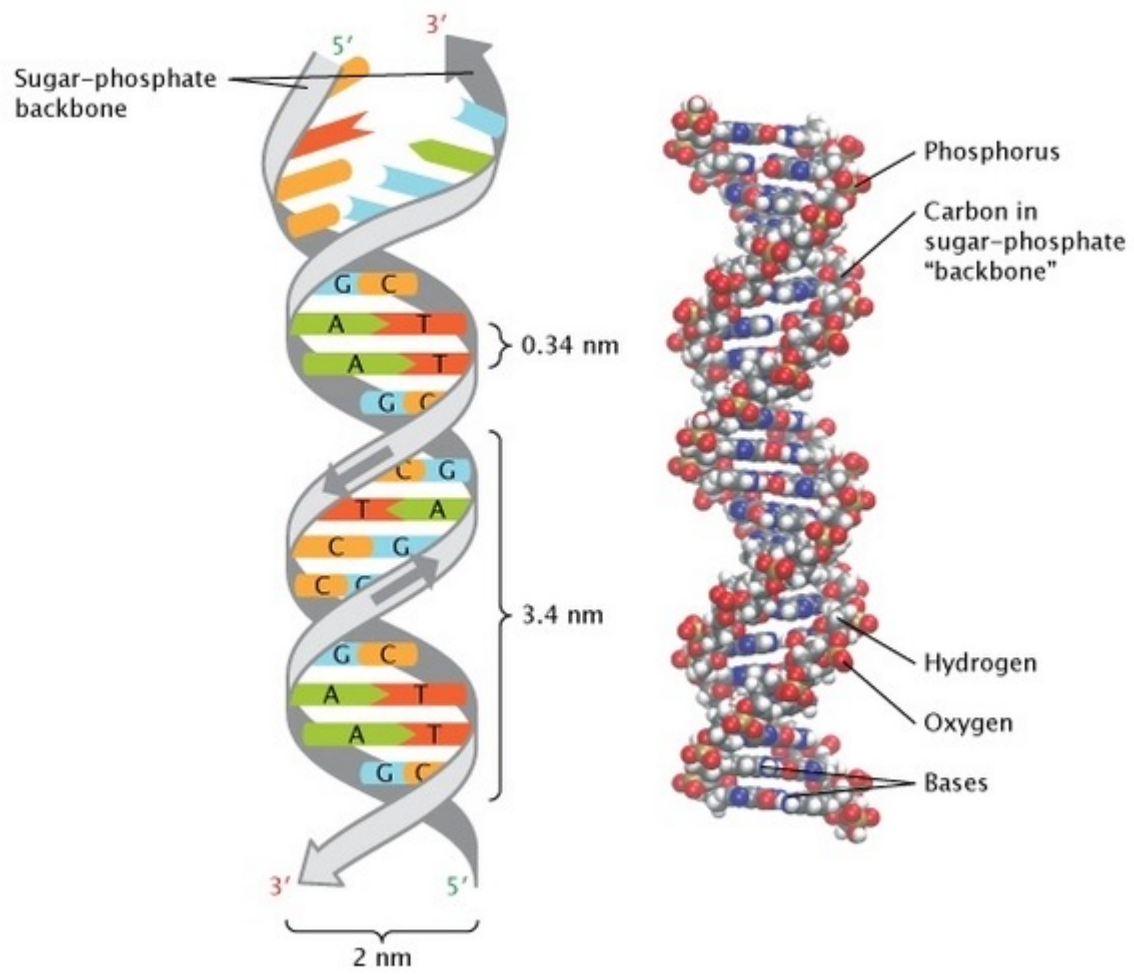
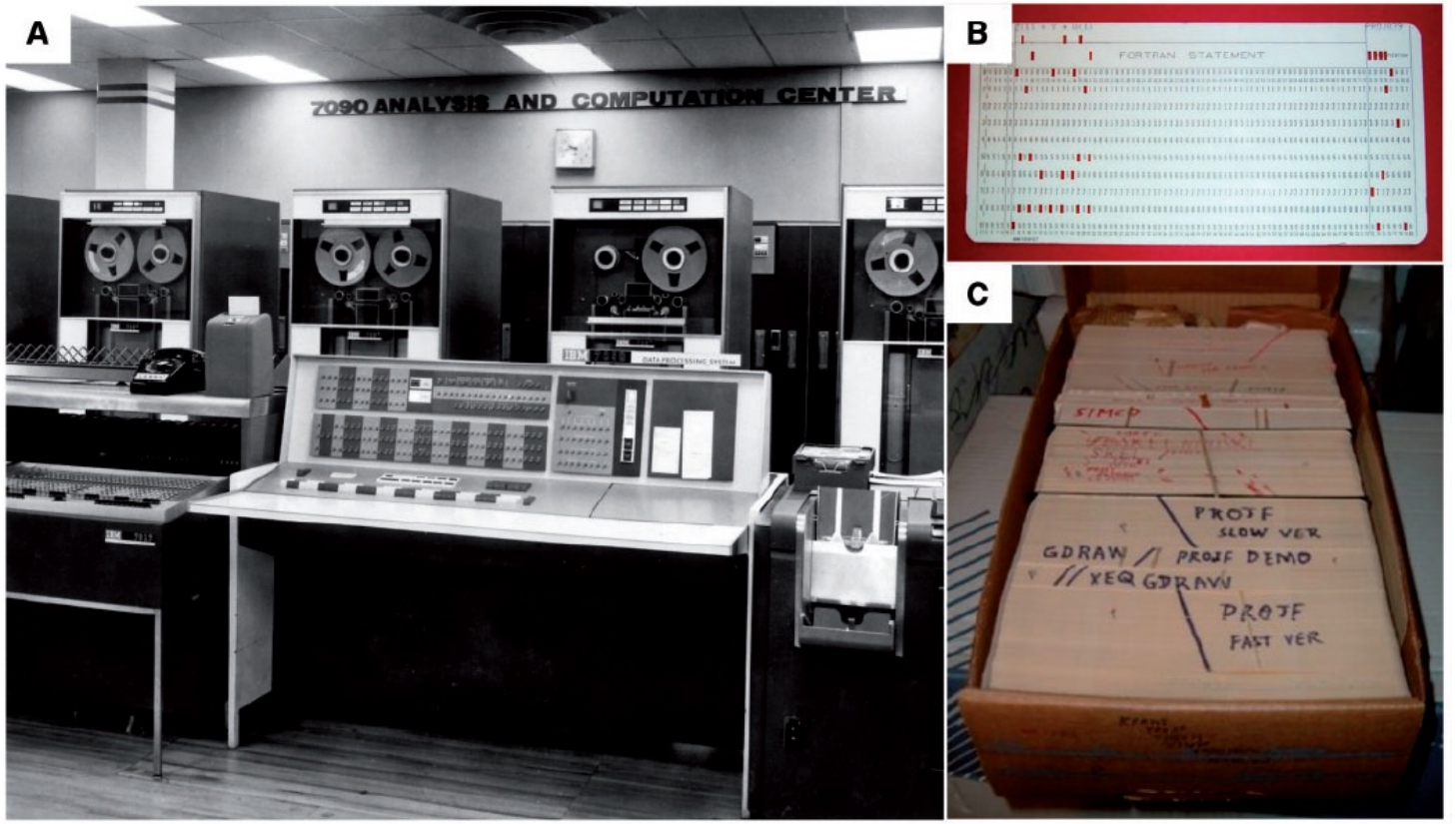


Figure: https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397/

Hallmarks

The first bioinformatician

Margareth Dayhoff, PhD (1925 – 1983)



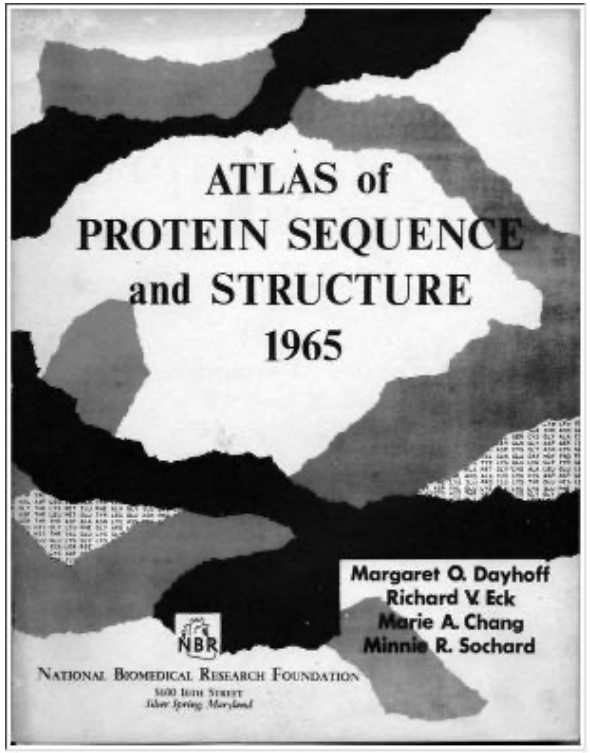
Gauthier et al, Briefings in Bioinformatics, 2018.

Margaret Dayhoff (1925 - 1983)

Born: Philadelphia, Pennsylvania, United States. Dayhoff is known as the founder of bioinformatics. This she did by pioneering the application of mathematics and computational techniques to the sequencing of proteins and nucleic acids and establishing the first publicly available database for research in the area. (Photo credit: Ruth E Dayhoff, National Library of Medicine).

Margaret Dayhoff: timeline of key events

Date	Event
11 Mar 1925	Margaret Dayhoff was born in Philadelphia, Pennsylvania, USA
1965	Atlas of Protein Sequence and Structure published
September 1980	First DNA sequence database created
1980	Largest nucleic acid sequence database in the world made available free over telephone network
5 Feb 1983	Margaret Dayhoff died in Silver Spring, Maryland, USA



209,159 structures
PDB – Sep 4th, 2023

http://www.whatisbiotechnology.org/people/Dayhoff

Hallmarks

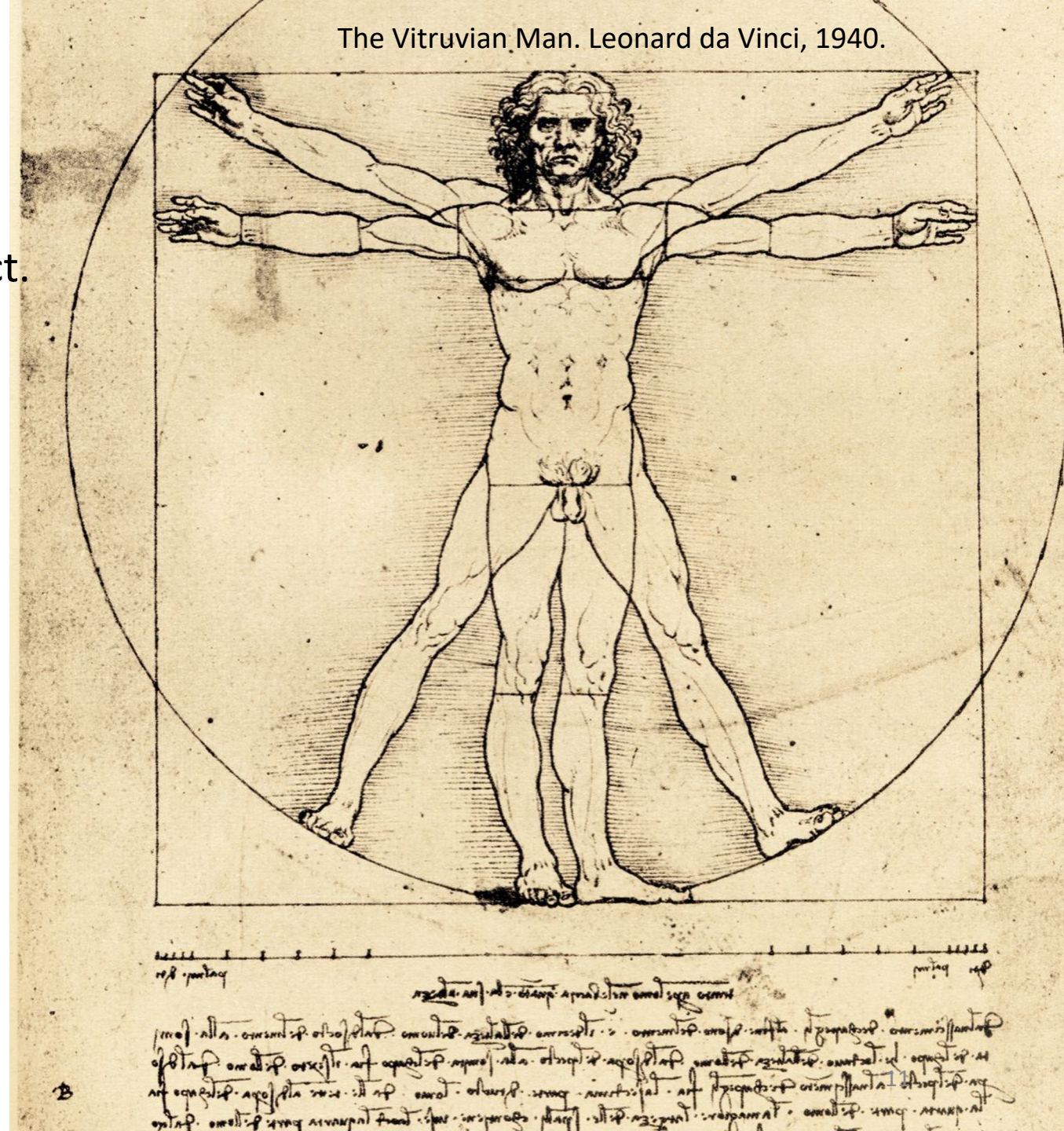
- 1990 – Launch of The Human Genome Project.

Private vs Public

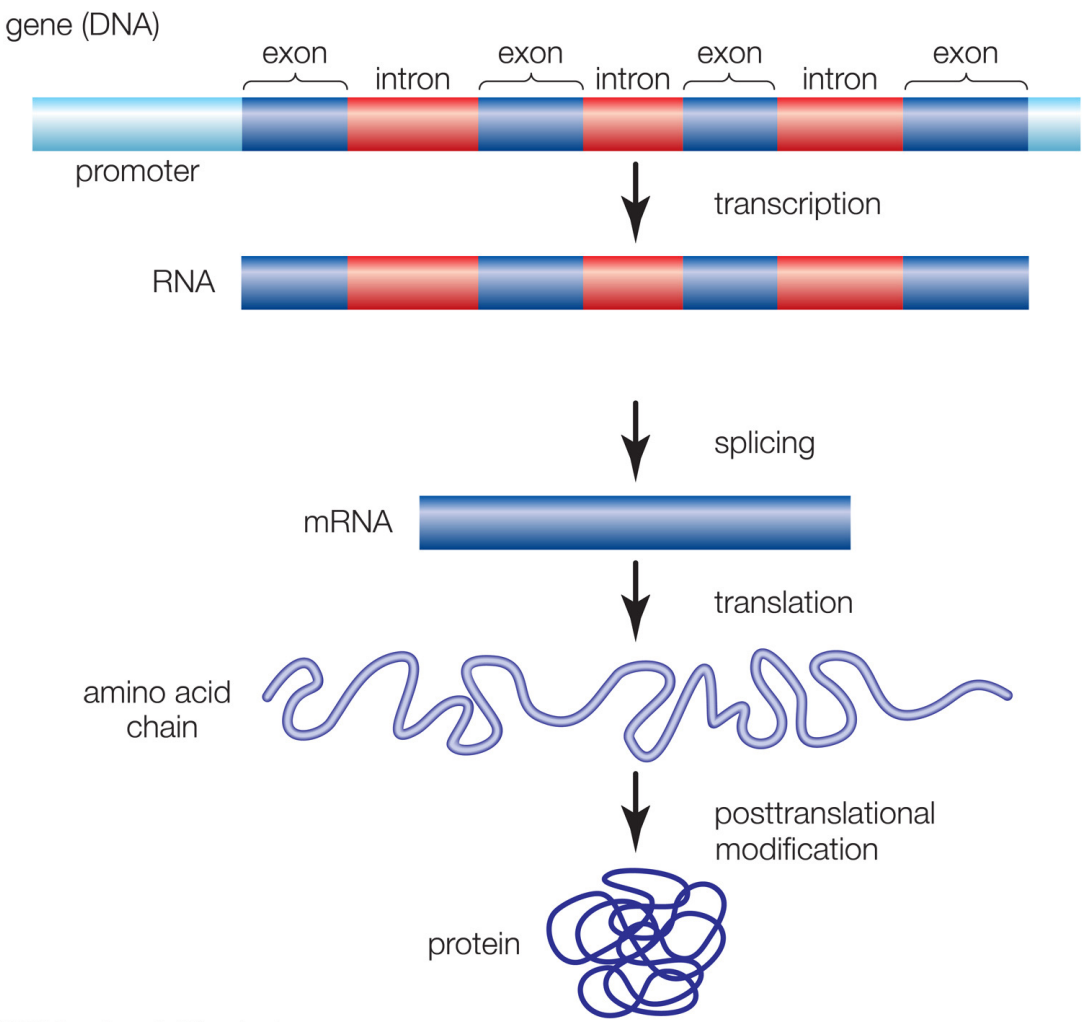
Craig Venter, Ari Patrinos, and Francis Collins

https://web.ornl.gov/sci/techresources/Human_Genome/project/privatesector.shtml

The Vitruvian Man. Leonard da Vinci, 1940.



Gene fragment sequencing – Venter strategy



This……..is..the……..Struc….ture….,,,…..of…your….(…genome…)…..

Words = genes
Ellipsis = introns
Punctuation = regulatory regions

Struct, of your genome

Source: “The gene”, book from Siddhartha Mukherjee

Genome sequencing – Venter strategy

1992 – Venter left INS and started his own Institute:

IGOR: Institute for Genome Research

TIGR: The Institute for Genome Research

1993 – Expanded his strategy of sequencing genes fragments to start sequencing complete genes and genomes.

Collaboration: Bert Volgestein, Kenz Kinzler (cancer research)

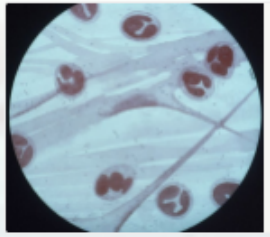
Hamilton Smith (bacteriology research)

Source: “The gene”, book from Siddhartha Mukherjee

1995 - 1st genome sequenced

Craig Venter, Hamilton Smith, Claire Fraser and collaborators

1995



Haemophilus influenza

- **Common name:** Haemophilus influenza

- **What is it?** A non-moving rod-shaped bacterium that causes meningitis.
- **Why was it sequenced?** This was the first bacteria to be sequenced.
- **Who sequenced it?** Craig Venter and his team at The Institute for Genomic Research in Rockville, Maryland, USA.
- **How many bases?** 1.8 million
- **How many chromosomes?** 1 circular chromosome

Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Fleischmann, et al. *Science, 1995.*

Human genome

1993 – Watson left the Human Genome Project

Francis Collins was nominated as the new PI of the Project

Private sector

Shotgun strategy

Struct, truct, structu, cture

Structure

Public sector

Organize the genome fragments in a map

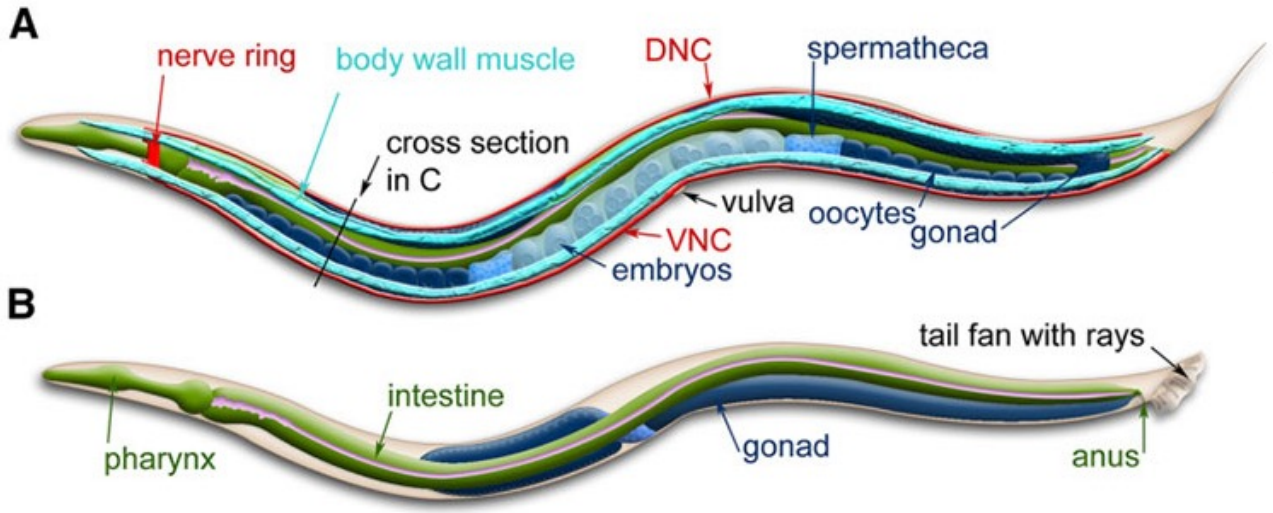
"Who is next to who?"

Source: "The gene", book from Siddhartha Mukherjee

Caenorhabditis elegans genome

1998 – Launch of the Celera company = Accelerate.

Dec 1998 - Public sector publish the *Caenorhabditis elegans* genome.

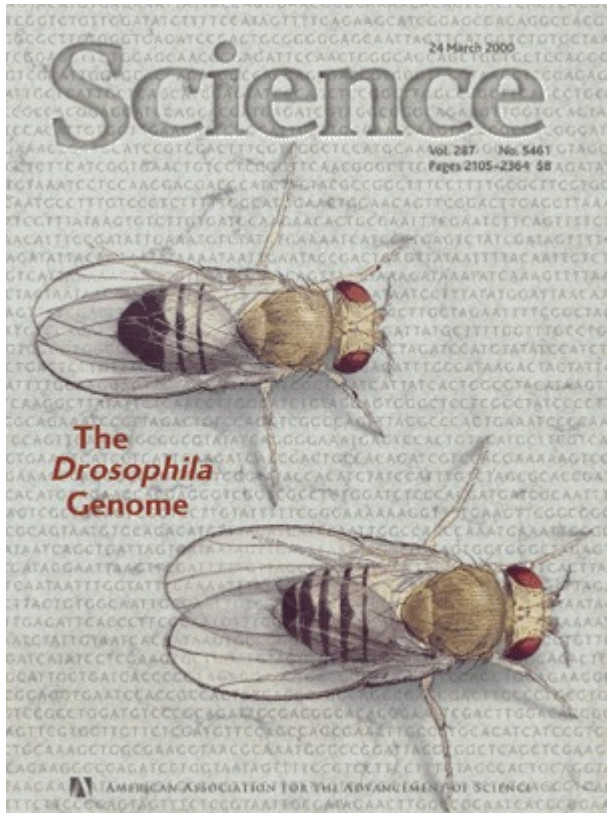


18,891 genes
Discovery: The large number of non-coding regions

Fig: https://andor.oxinst.com/learning/view/article/advantages-of-using-caenorhabditis-elegans-as-a-model-organism

Drosophila melanogaster genome

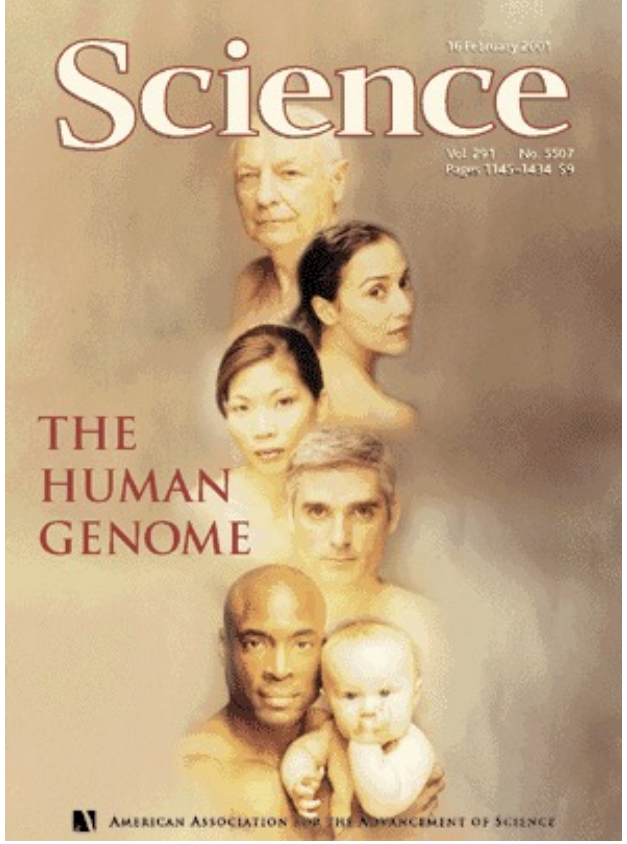
Sept 1999 – Celera announced the work on *Drosophila melanogaster* genome.



13,601 genes
5,000 less than the nematoid
Published March 2020

By this time, people were convinced by the shotgun method.

Source: “The gene”, book from Siddhartha Mukherjee

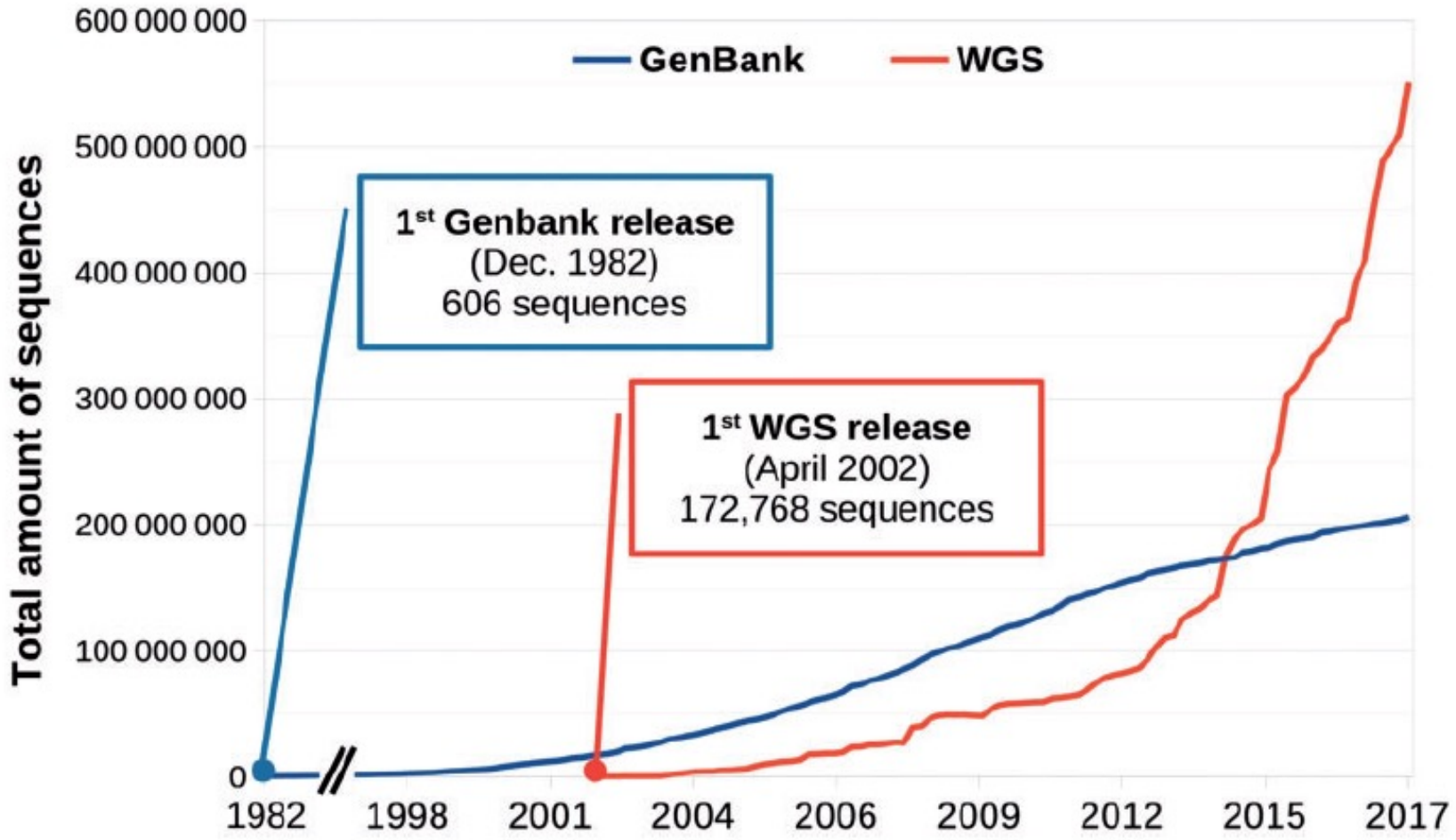


> 20 years!

- Human Genome sequenced
- 2001


PHASE TWO: INTERPRETATION
SHERMAN Yellow Ledger
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

2000 – 2010: High-throughput bioinformatics



Gauthier et al, 2018

https://www.ncbi.nlm.nih.gov/genbank/statistics/

Next Generation Sequencing - NGS

Platform	Technology	Model	Length of read	Throughput (per day)	Company
Automated Sanger sequenc	Capillary electrophoresis, ye®-terminator istry	3730 xl	Up to 900 bp	<3 Mb	Applied Biosystems
454	Pyrosequencing	GS FLX+	500–700 bp	700 Mb	454 Life Sciences (Roche)
SOLiD™	Sequencing by ligation	5500 xl	75 bp	30 Gb	Life Technologies (ABI)
Illumina	Clonal single molecule array	HiSeq2000	50–150 bp	Up to 55 Gb	Illumina, Inc.
Complete genomics	DNA nanoball array, ligation-based sequencing	N/A [2]	70 bp	8.8 Gb	Complete Genomics
Ion Torrent	Hydrogen ion semiconductor	Ion 316 Chip	100 bp	100 Mb [3]	Life Technologies (ABI)
HeliScope™	Imaging single nucleotide incorporation	Single Molecule Sequencer	35 bp	1 Gb	Helicos
PacBio	SMRT™ technology	PacBio *RS*	>1,000 bp	500 Mb [4]	Pacific Biosciences

2004

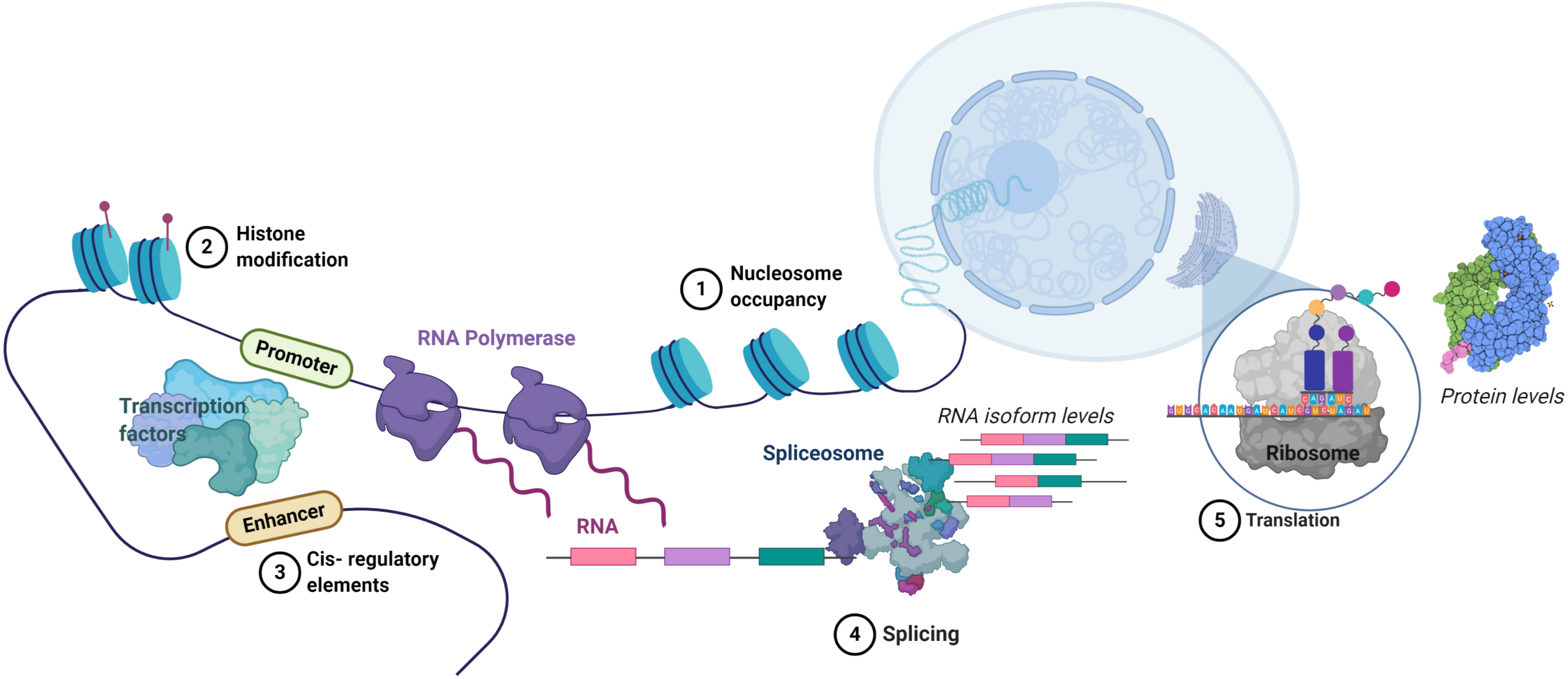
Li et. al. Review Cancers, 2011.

Tasks

- Q1: What do we mean when we say that the **Human genome** sequencing brought more **questions** than expected?
- Q2: What are the **steps after** sequencing an organism?
- Q3: **How many** Mammals are deposited on the NCBI? https://www.ncbi.nlm.nih.gov/genome/browse/ -> Filters -> Subgroup

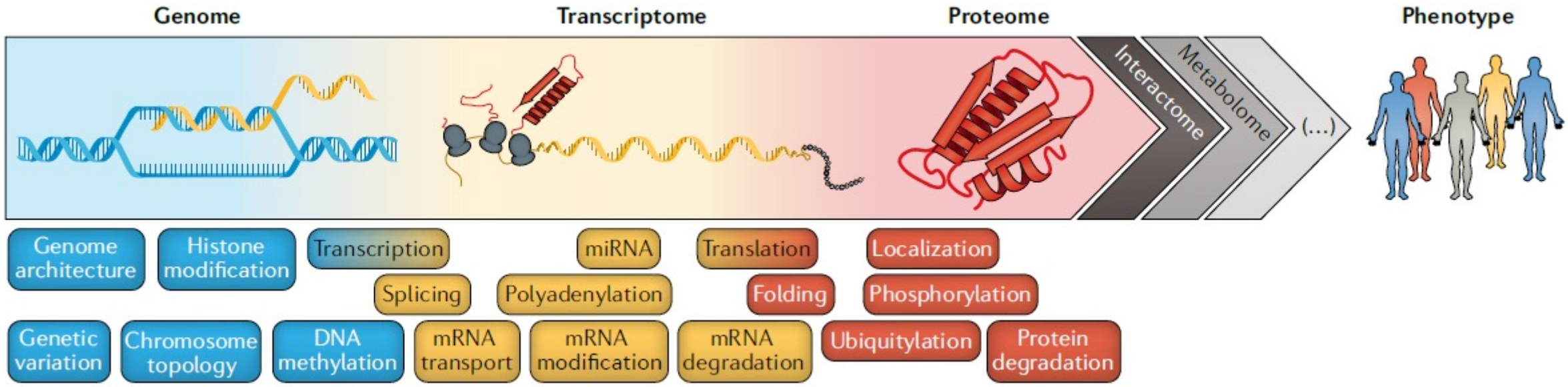
Distinct types of biological data

Gene expression regulation



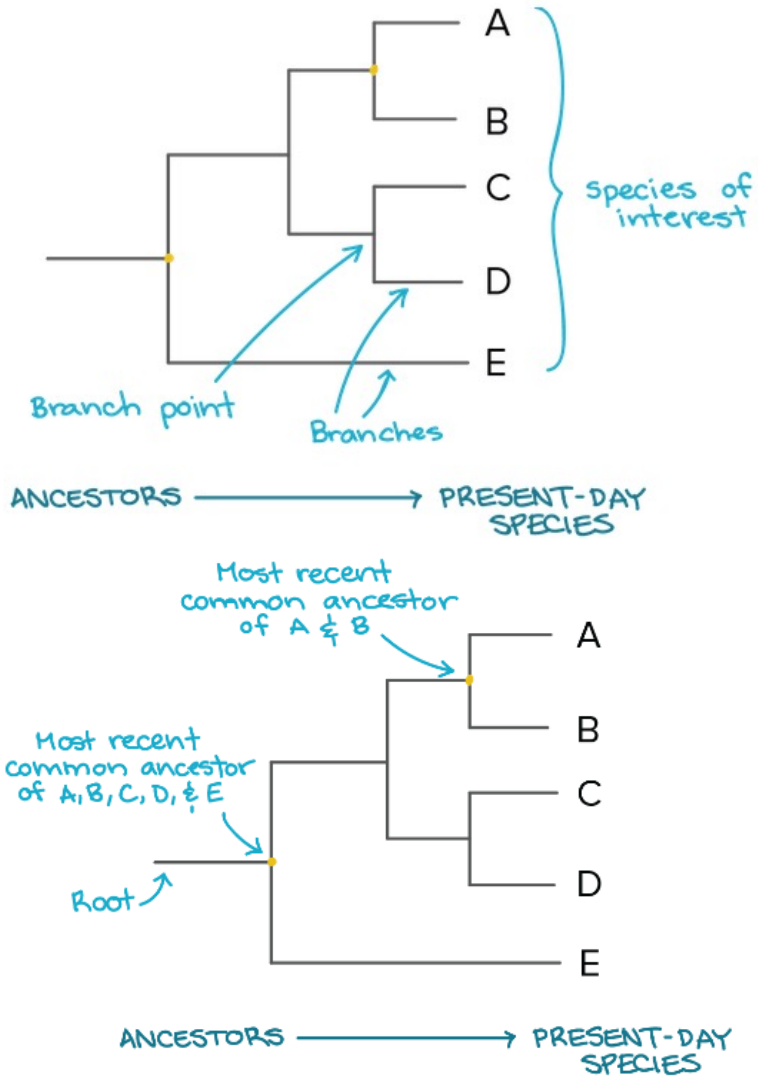
Slide: Ricardo Vialle, 2020.

OMIC's data

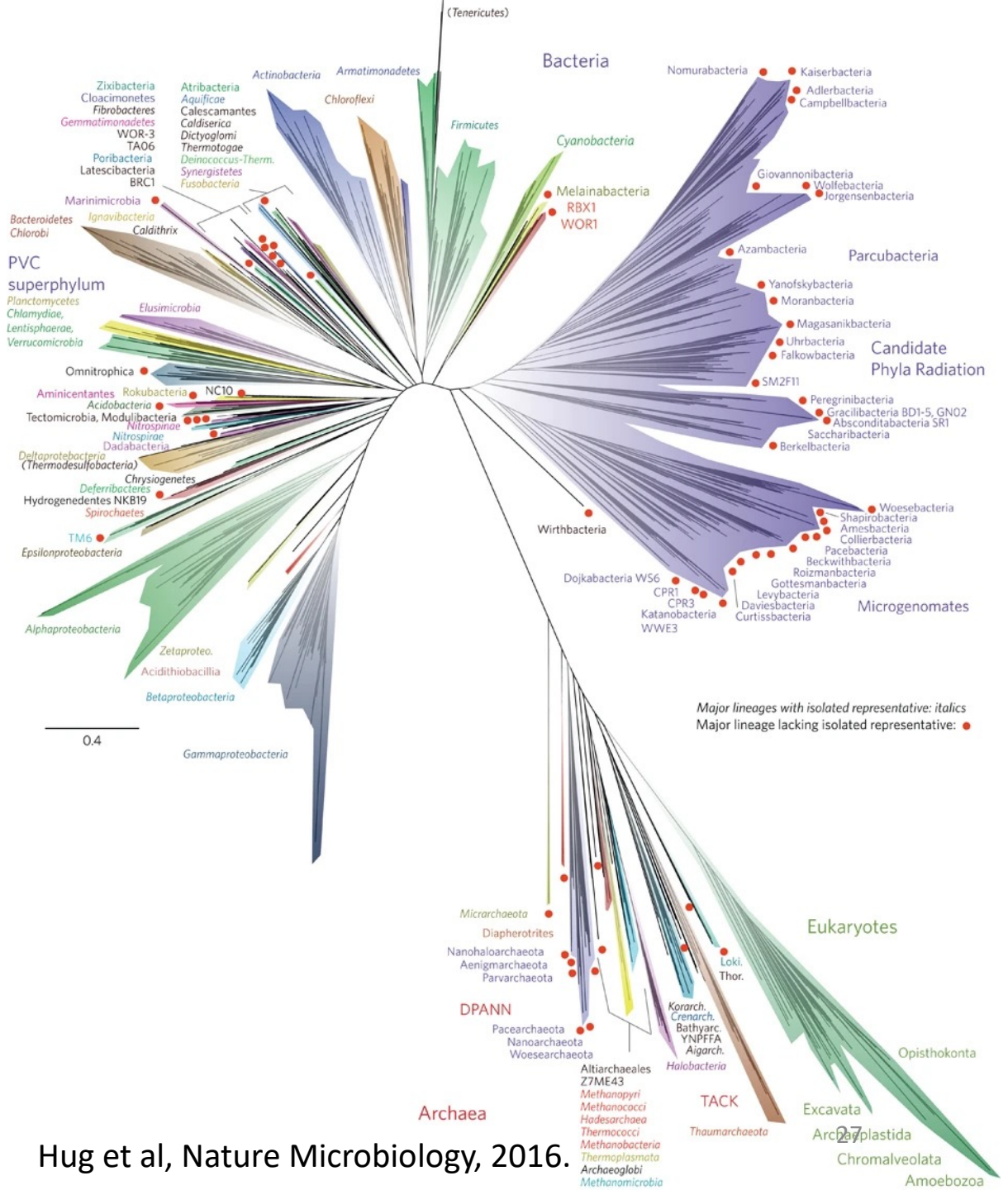


Christopher Buccitelli and Matthias Selbach, Nature Review Genetics, 2020.

Phylogenetic inference

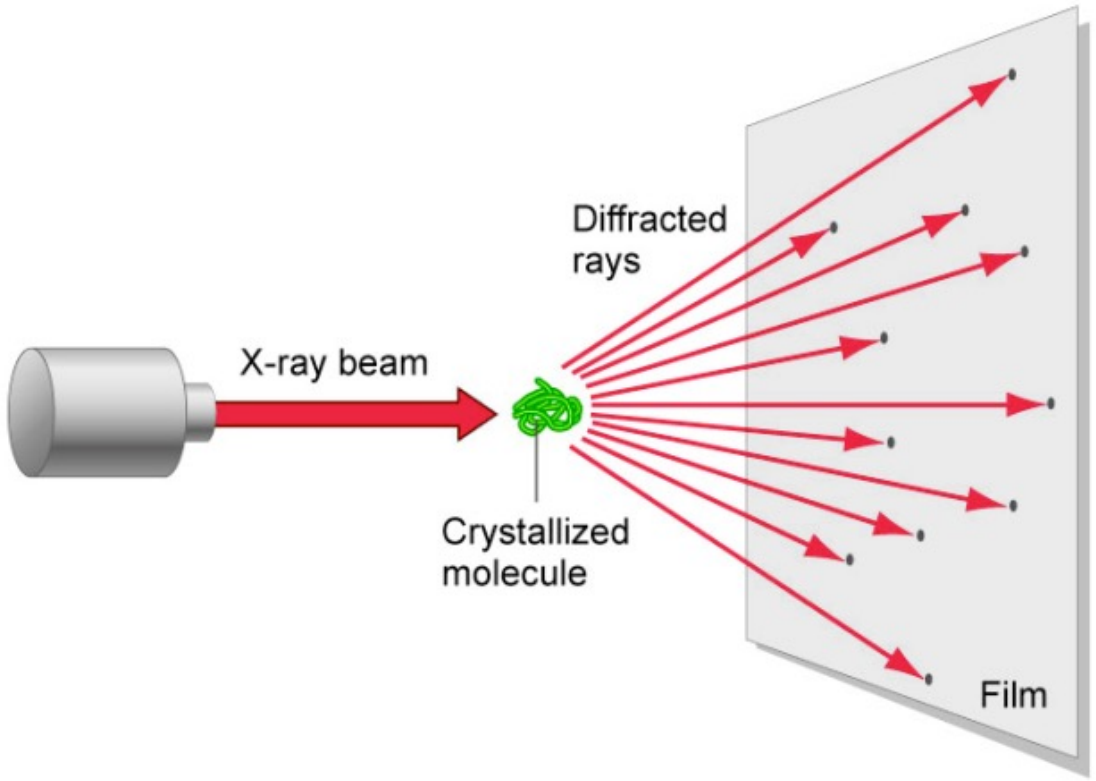


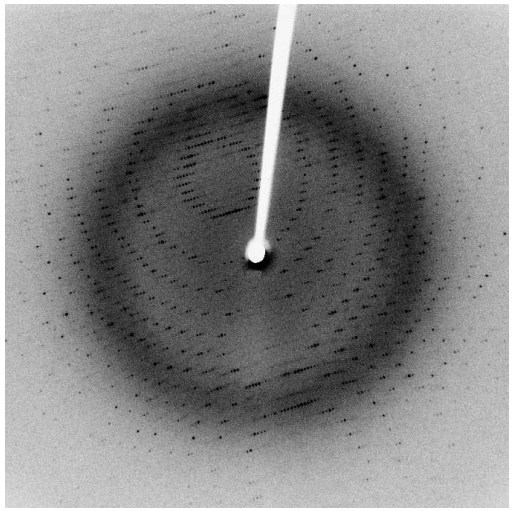
https://www.khanacademy.org/science/high-school-biology/hs-evolution/hs-phylogeny/a/phylogenetic-trees

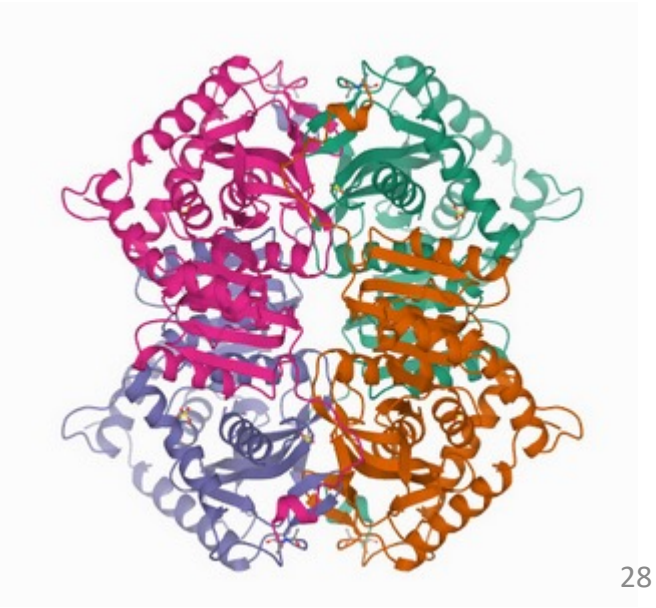


Hug et al, Nature Microbiology, 2016.

Protein structure - examples







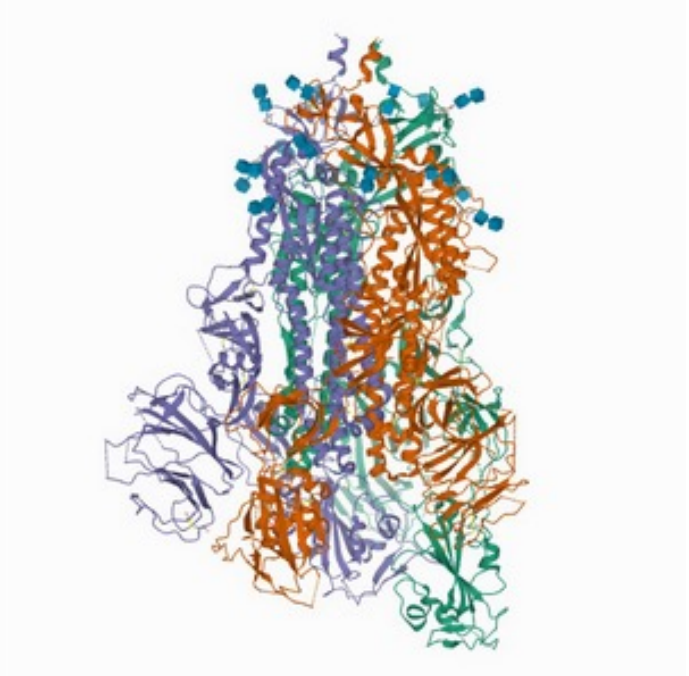
https://www.britannica.com/science/X-ray-diffraction#/media/1/650434/152944

RCSB PDB
PROTEIN DATA BANK

174507 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB file
26,186 lines

```
ATOM   4094  C   GLU A 702     198.955 208.031 189.462  1.00 92.70           C
ATOM   4095  O   GLU A 702     197.882 207.525 189.126  1.00 92.70           O
ATOM   4096  CB  GLU A 702     199.082 206.838 191.673  1.00 92.70           C
ATOM   4097  CG  GLU A 702     199.697 205.635 190.963  1.00 92.70           C
ATOM   4098  CD  GLU A 702     199.988 204.476 191.893  1.00 92.70           C
ATOM   4099  OE1 GLU A 702     201.053 203.843 191.734  1.00 92.70           O
ATOM   4100  OE2 GLU A 702     199.156 204.197 192.780  1.00 92.70           O
ATOM   4101  N   ASN A 703     199.831 208.510 188.586  1.00 84.21           N
ATOM   4102  CA  ASN A 703     199.699 208.320 187.153  1.00 84.21           C
ATOM   4103  C   ASN A 703     200.907 207.559 186.626  1.00 84.21           C
ATOM   4104  O   ASN A 703     201.988 207.573 187.218  1.00 84.21           O
ATOM   4105  CB  ASN A 703     199.550 209.661 186.421  1.00 84.21           C
ATOM   4106  CG  ASN A 703     200.860 210.409 186.290  1.00 84.21           C
ATOM   4107  OD1 ASN A 703     201.717 210.349 187.168  1.00 84.21           O
ATOM   4108  ND2 ASN A 703     201.020 211.126 185.185  1.00 84.21           N
ATOM   4109  N   SER A 704     200.707 206.882 185.501  1.00 80.18           N
ATOM   4110  CA  SER A 704     201.755 206.092 184.866  1.00 80.18           C
ATOM   4111  C   SER A 704     201.885 206.559 183.427  1.00 80.18           C
ATOM   4112  O   SER A 704     201.023 206.258 182.596  1.00 80.18           O
ATOM   4113  CB  SER A 704     201.438 204.597 184.928  1.00 80.18           C
ATOM   4114  OG  SER A 704     202.408 203.840 184.227  1.00 80.18           O
ATOM   4115  N   VAL A 705     202.953 207.304 183.135  1.00 73.36           N
ATOM   4116  CA  VAL A 705     203.196 207.730 181.766  1.00 73.36           C
ATOM   4117  C   VAL A 705     203.334 206.500 180.885  1.00 73.36           C
```



6VSB

Prefusion 2019-nCoV spike glycoprotein with a single receptor-binding domain up

DOI: 10.2210/pdb6VSB/pdb **EMDataResource:** EMD-21375

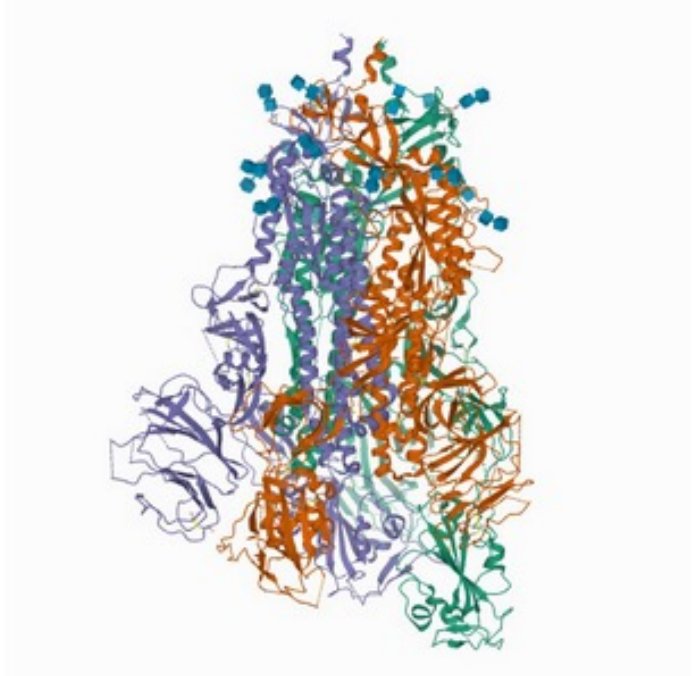
Classification: VIRAL PROTEIN
Organism(s): Severe acute respiratory syndrome coronavirus 2
Expression System: Homo sapiens
Mutation(s): Yes

Deposited: 2020-02-10 **Released:** 2020-02-26

RCSB PDB PROTEIN DATA BANK

174507 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

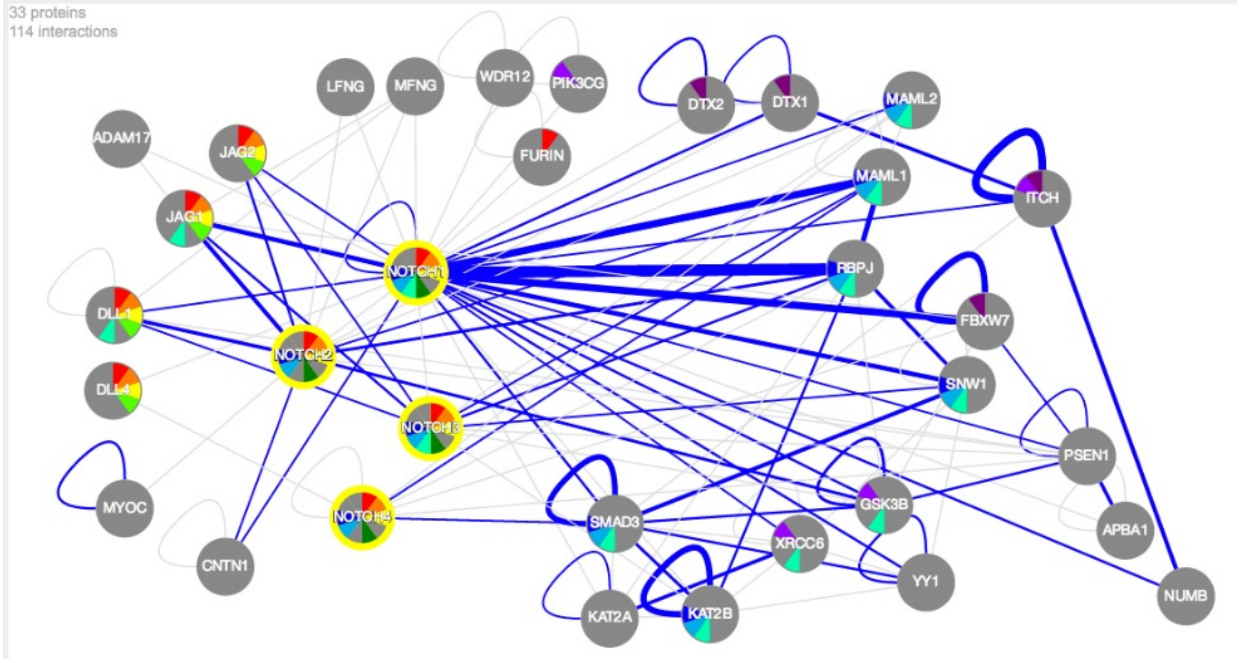
Fasta file



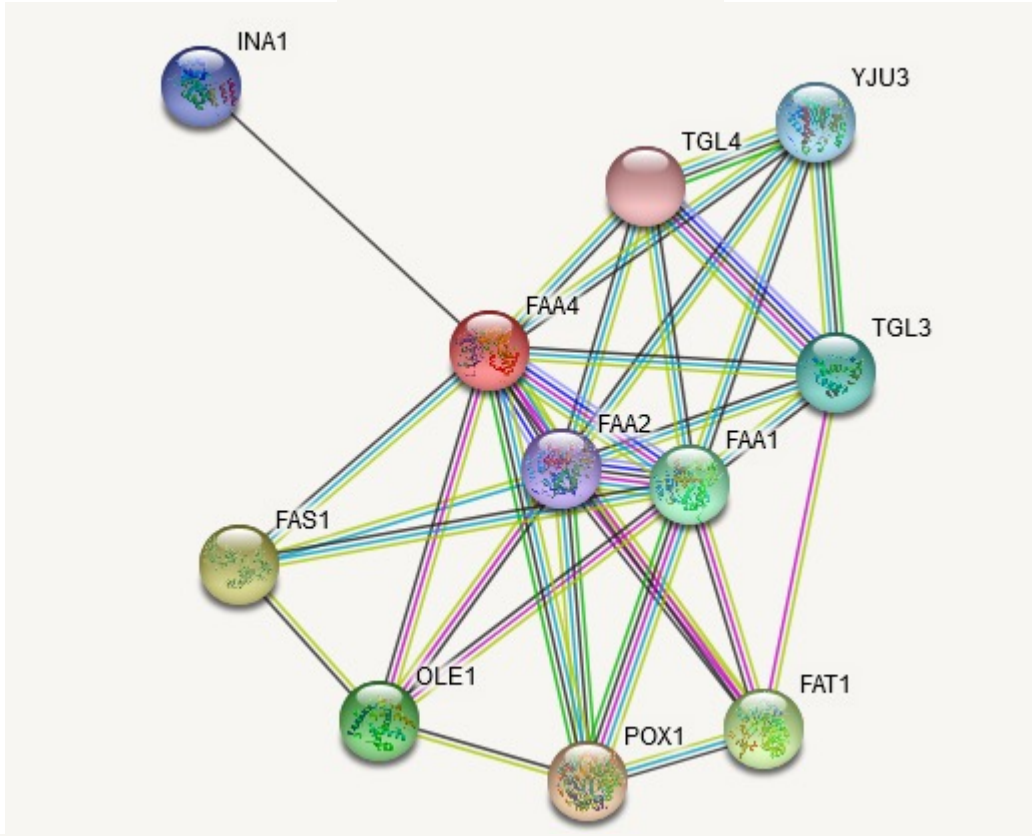
>6VSB_1|Chains A,B,C|Spike glycoprotein|Severe acute respiratory syndrome coronavirus 2 (2697049)
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLL
IVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPF
LMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINI
TRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCAL
DPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAW
NRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAP
GQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDIST
EIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPK
KSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDI
TPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQ
TRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPGSASSVASQSIIAY (…)

Protein-protein interaction

Agile Protein Interactomes DataServer



STRING



Node Color

colored nodes:
query proteins and first shell of interactors

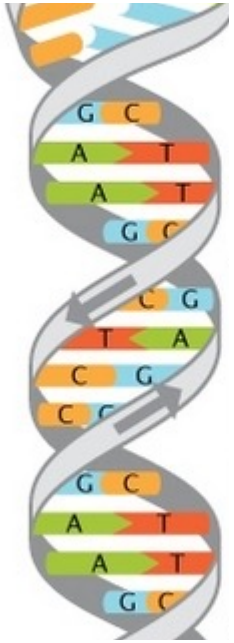
white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

Whole Genome Sequence

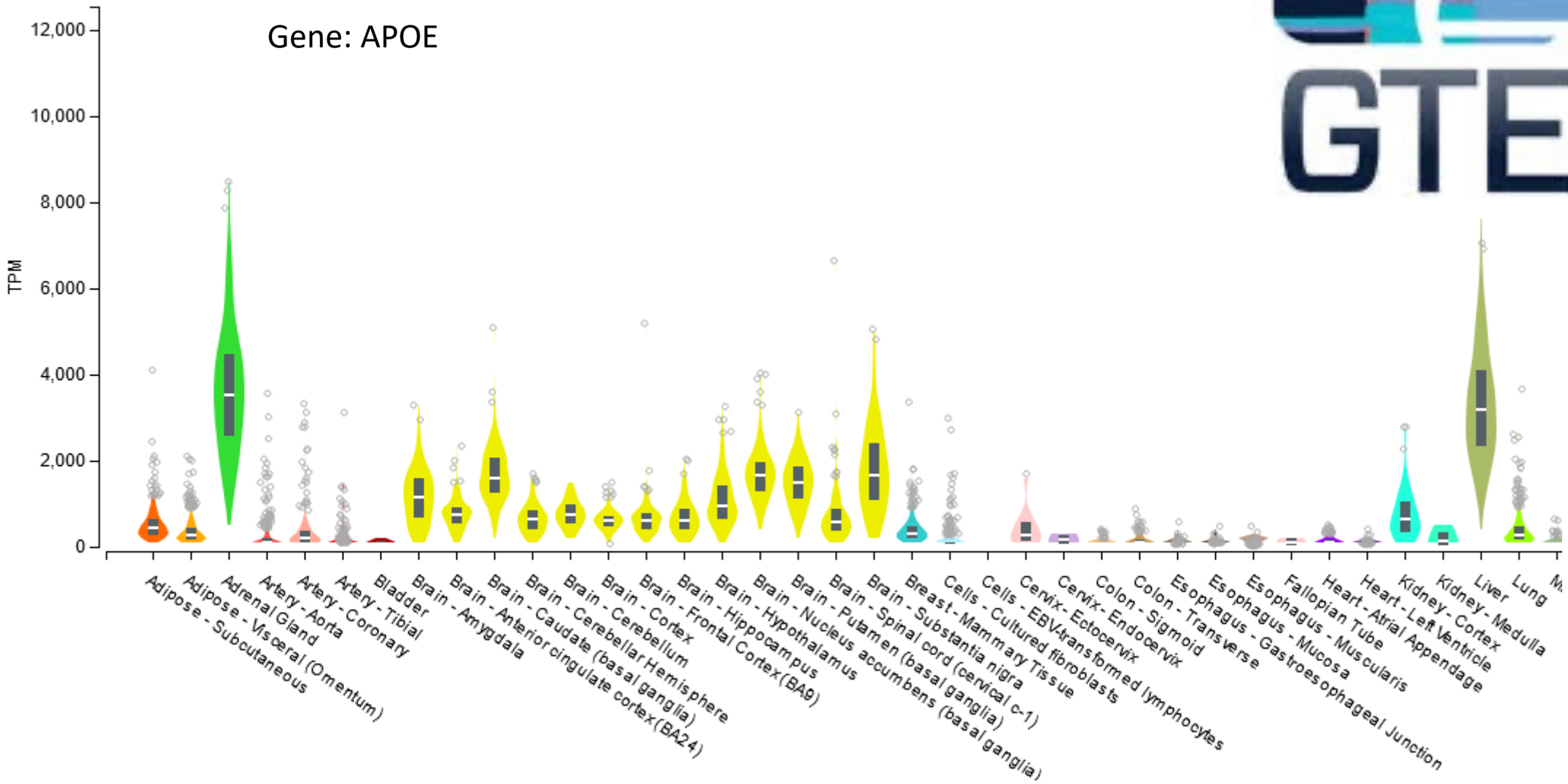


```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      N
A00001        NA00002        NA00003
20     14370   rs6054257 G       A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:
HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T       A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:
HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A       G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:
HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T       .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:
HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT    G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP
0/1:35:4       0/2:17:2       1/1:40:3
```

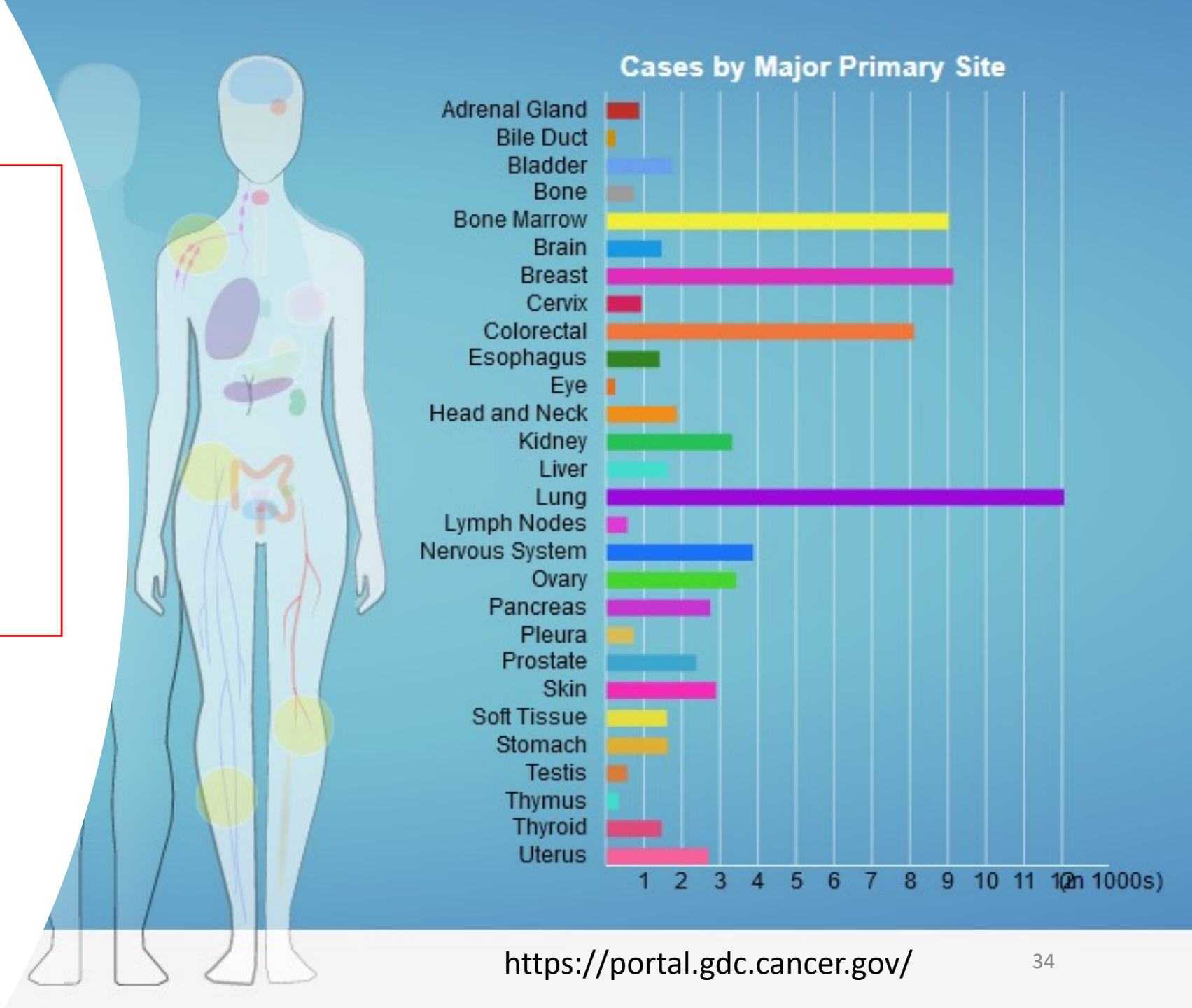
VCF file

https://www.internationalgenome.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40/

Gene expression - RNASeq



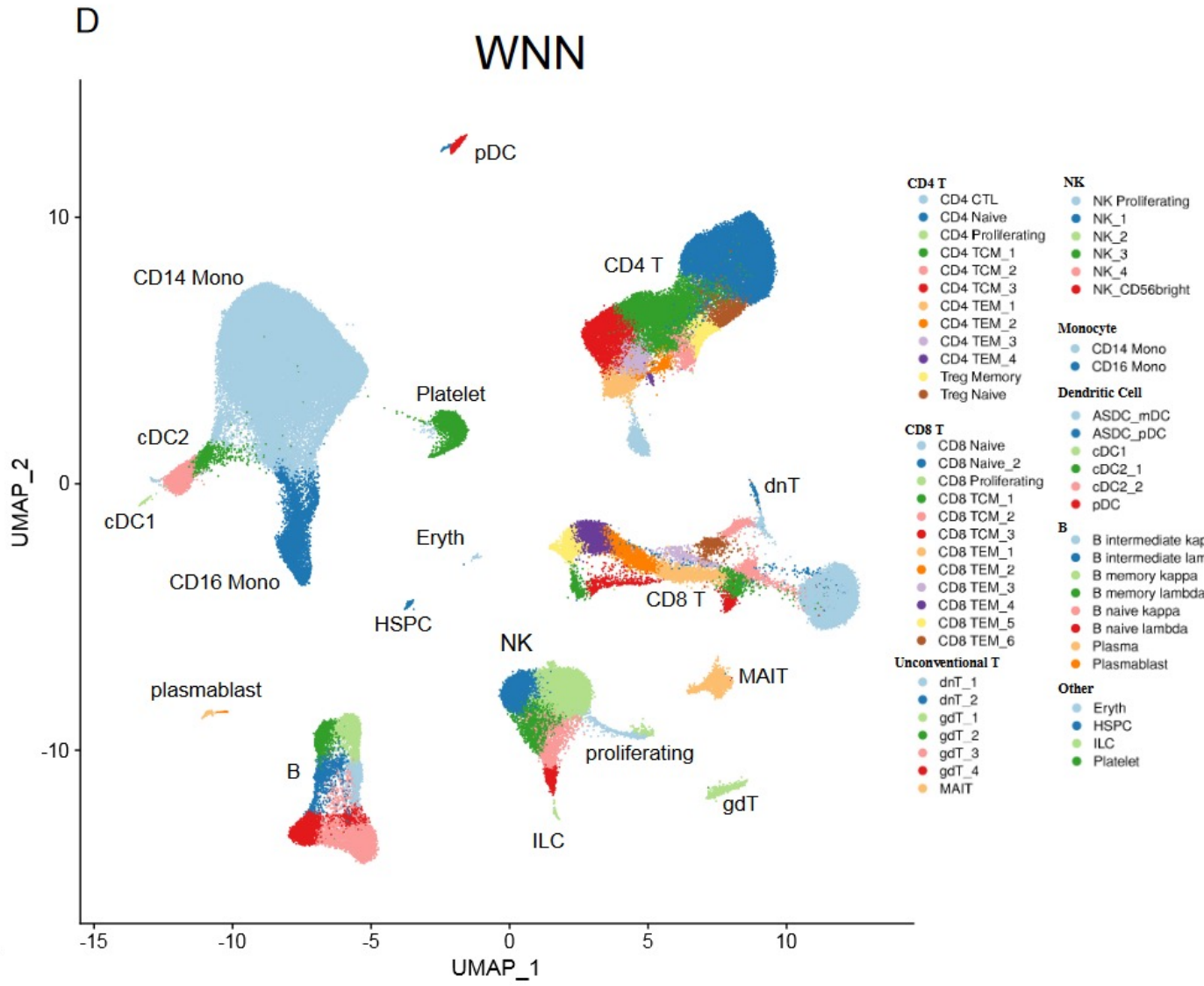
The Cancer Genome Atlas TCGA



https://portal.gdc.cancer.gov/

Single-cell analysis

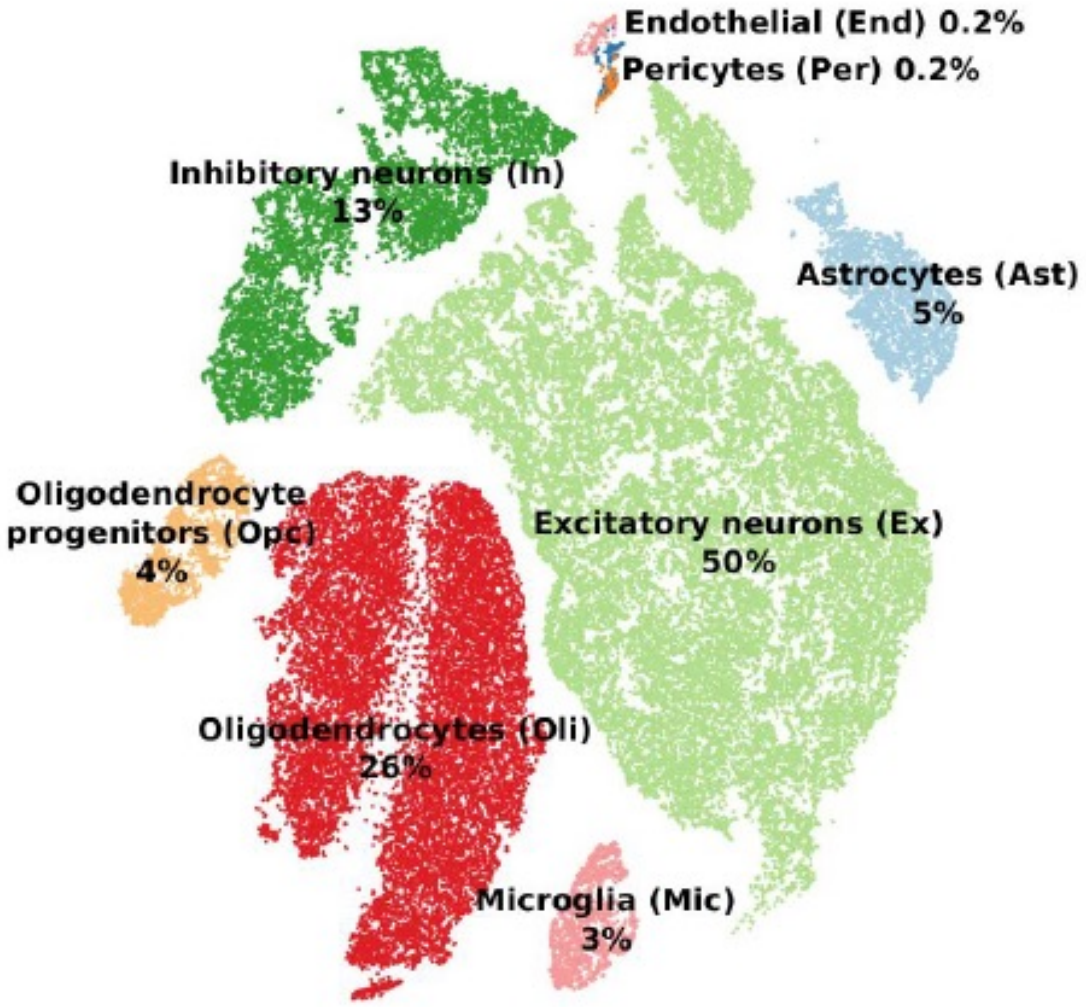
What are the cell types in a drop of blood?



Hao et al, Biorxiv 2020.

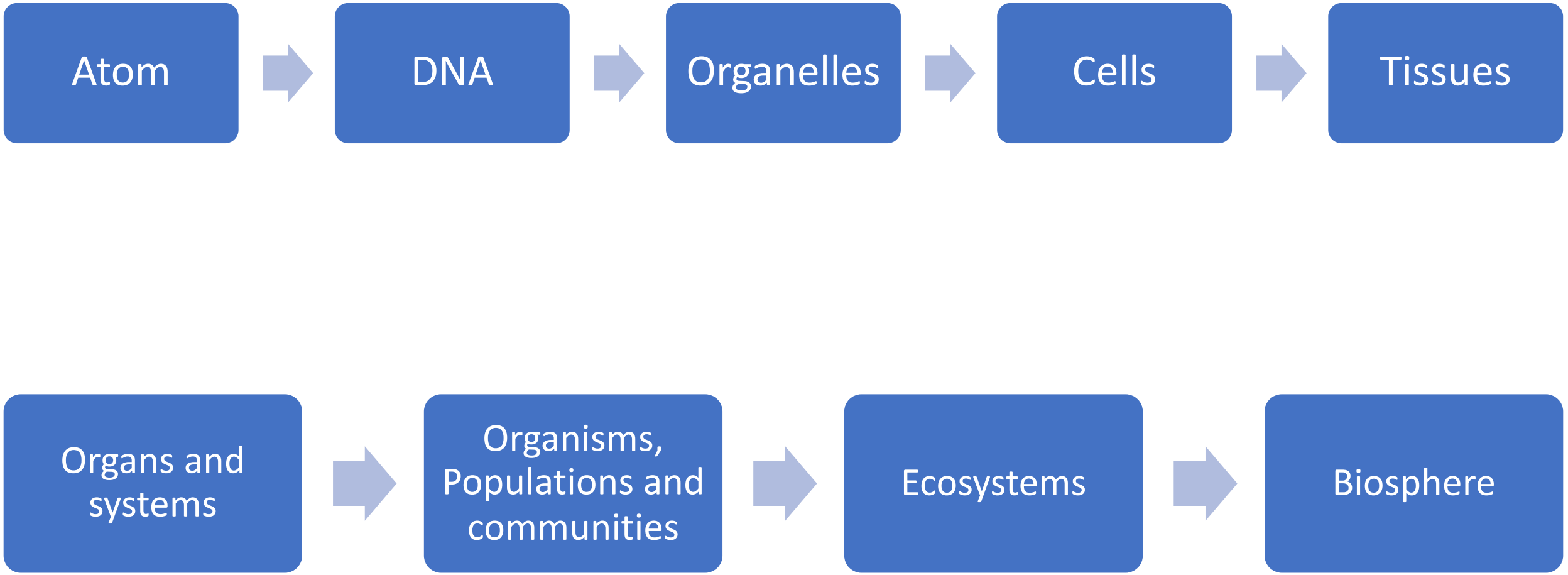
Single-cell analysis

What are the cell types in the human brain?

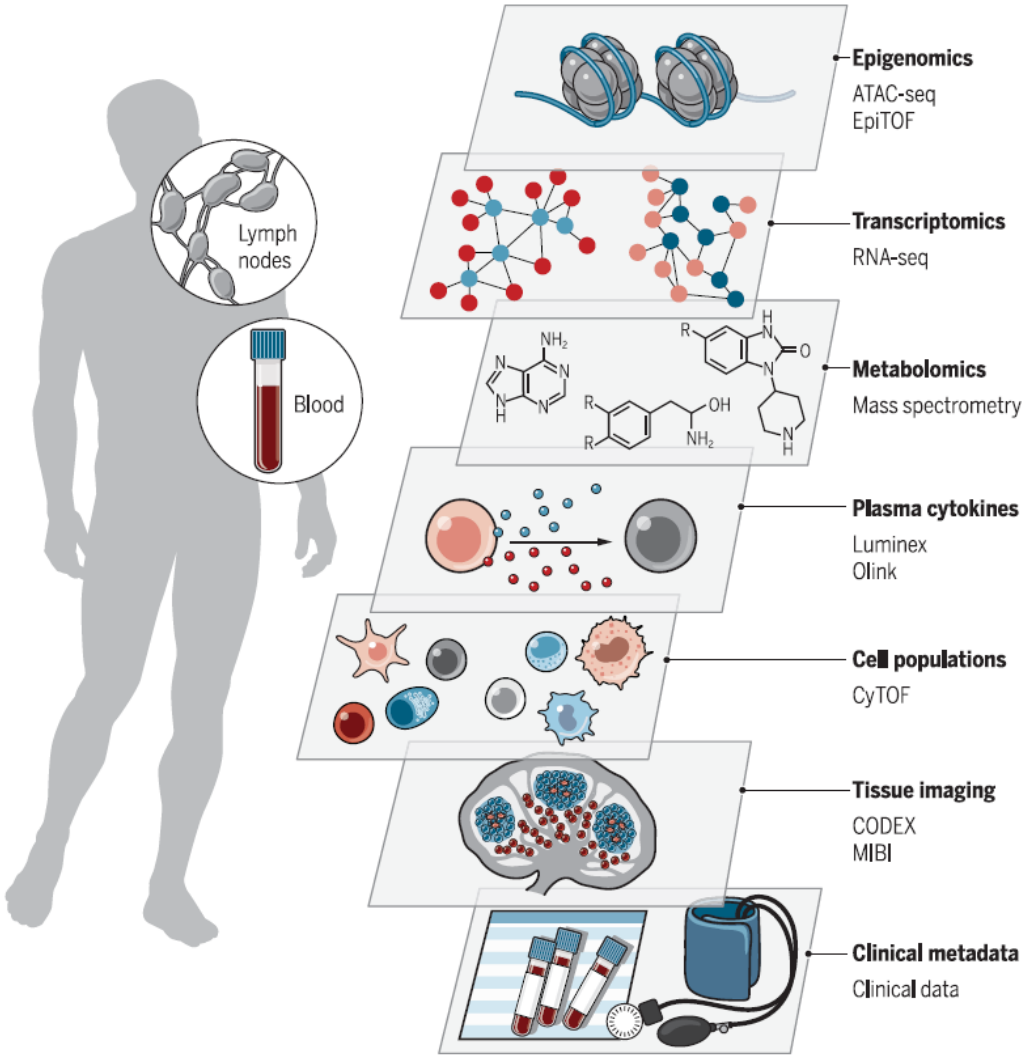


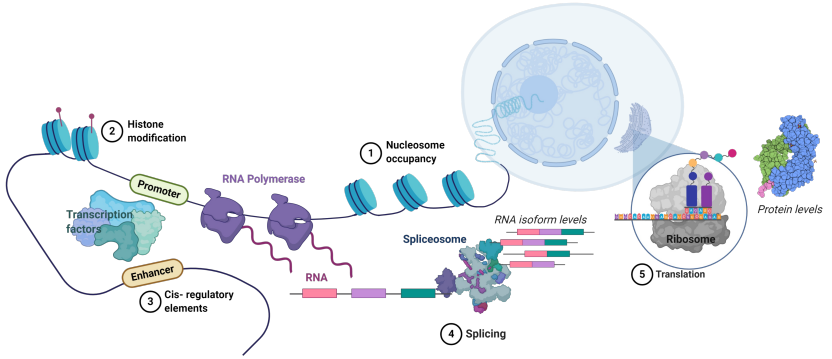
Mathys et al. Nature, 2019.

Levels of organization



Distinct methods to study distinct systems





Data integration

Bali Pulendran and Mark M. Davis. Science, 2020.

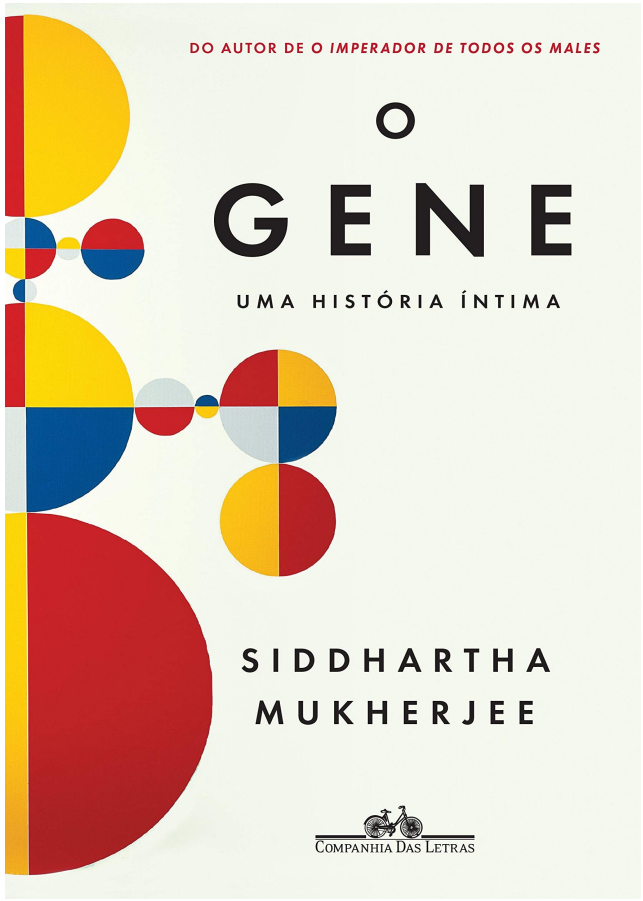
Tasks

- Q1: How does the availability of vast and diverse biological data sources impact the advancement of systems biology research?

- Q2: Where the *SPP1* gene is expressed? https://www.gtexportal.org/home/

- Q3: List the information we can get from a single-cell UMAP representation.

Reading recommendations



REVIEW SUMMARY

IMMUNOLOGY

The science and medicine of human immunology

Bali Pulendran* and Mark M. Davis

https://science.sciencemag.org/content/369/6511/eaay4014

Thank you!

katiaplopes@gmail.com
@lopeskp