Computational Biology

ICS 491

Upcoming guest lectures

- November 14: Aekta Shah (Ex-Google) - Data Ethics

We may need to move around some discussion question presentations.

For next few classes, be prepared to present…

(if we have time)

1. Final update on the dataset you are using
2. Your data analysis plan
3. Your preliminary results/findings, if you have them

This will be participation credit

~1.5-2 minutes per student

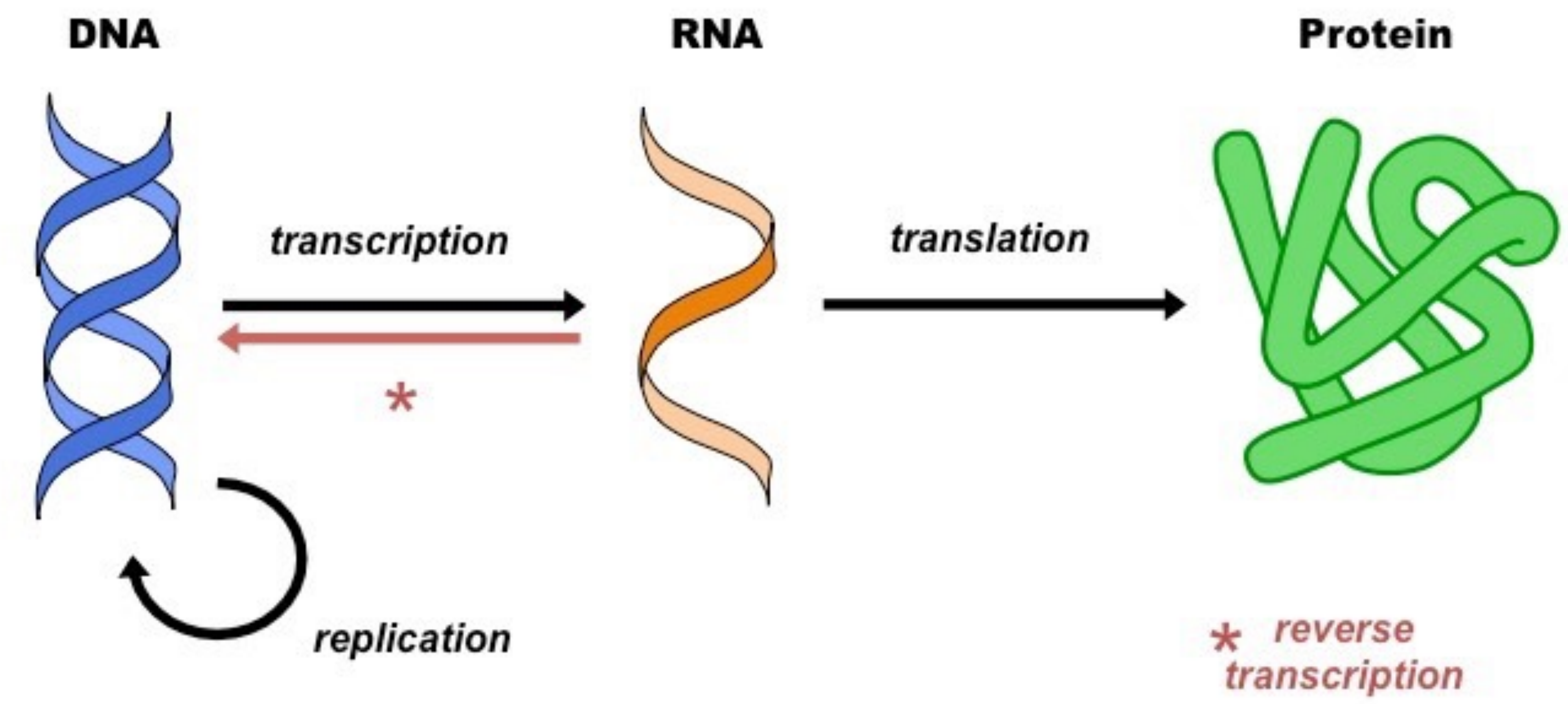
DNA

The language of DNA has 4 core characters:

- A
- T
- C
- G

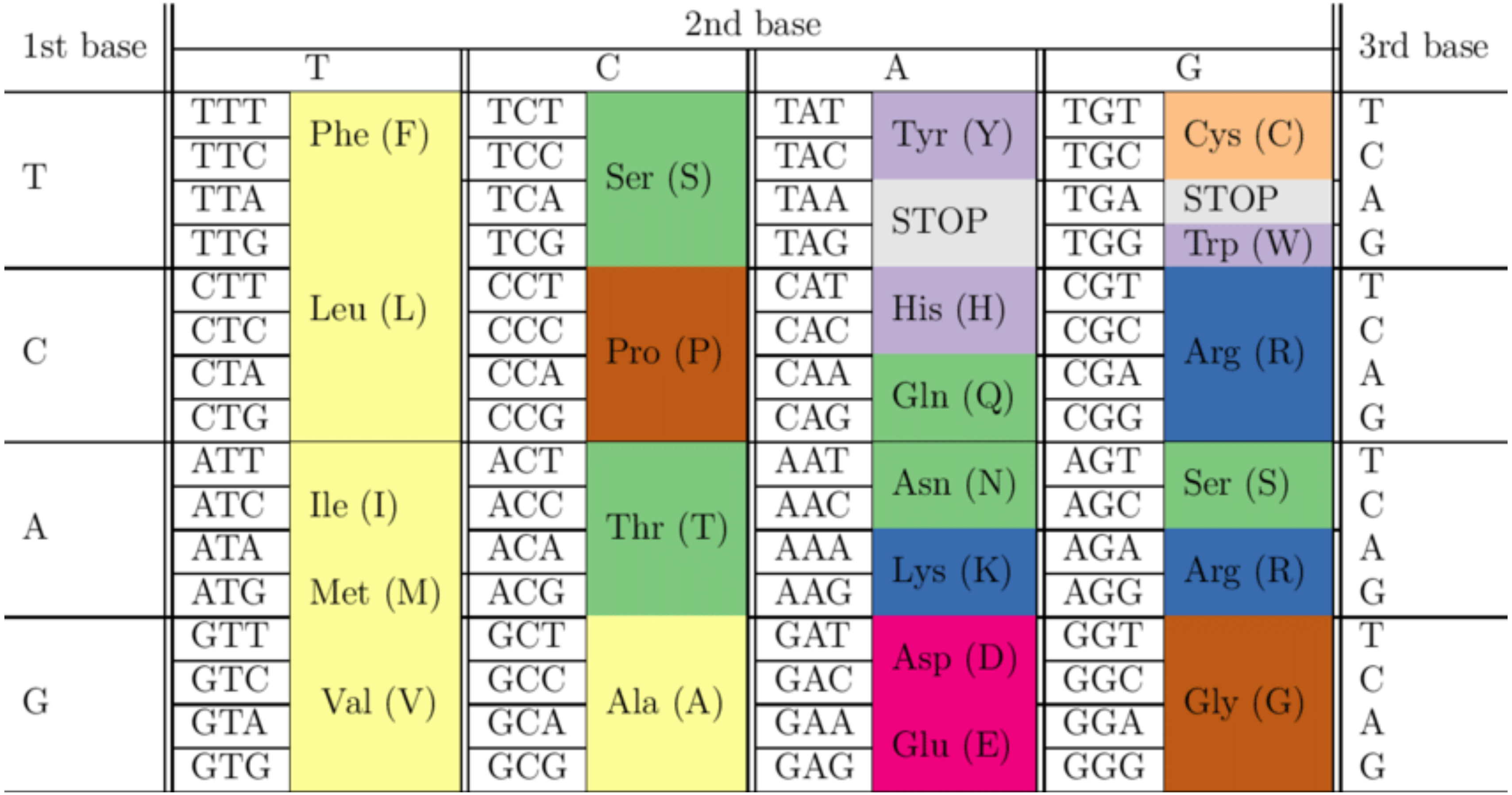
There are 3 billion base pairs in the human genome.

Central Dogma of Biology



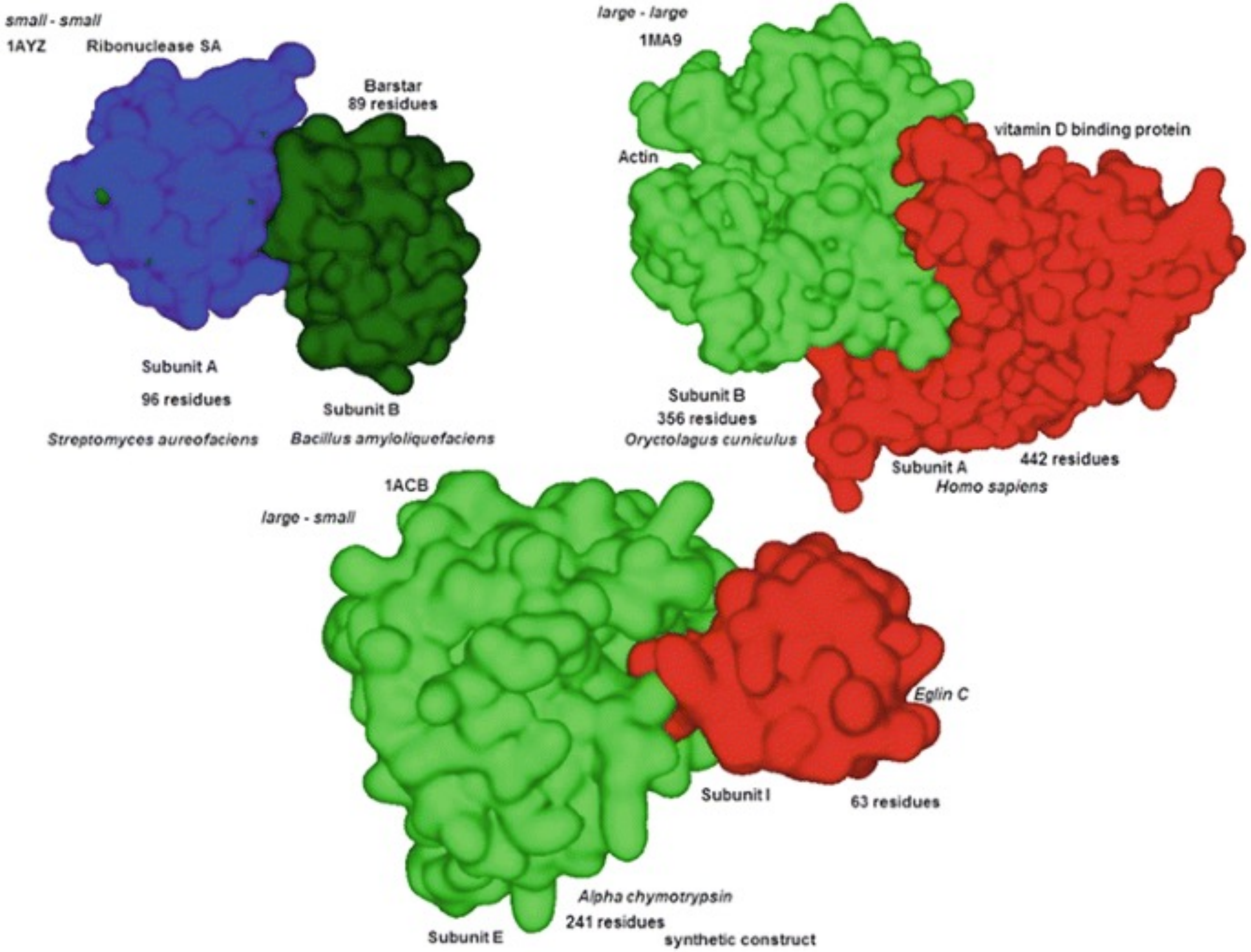
https://ib.bioninja.com.au/standard-level/topic-2-molecular-biology/27-dna-replication-transcri/central-dogma.html

Proteins are strings of amino acids

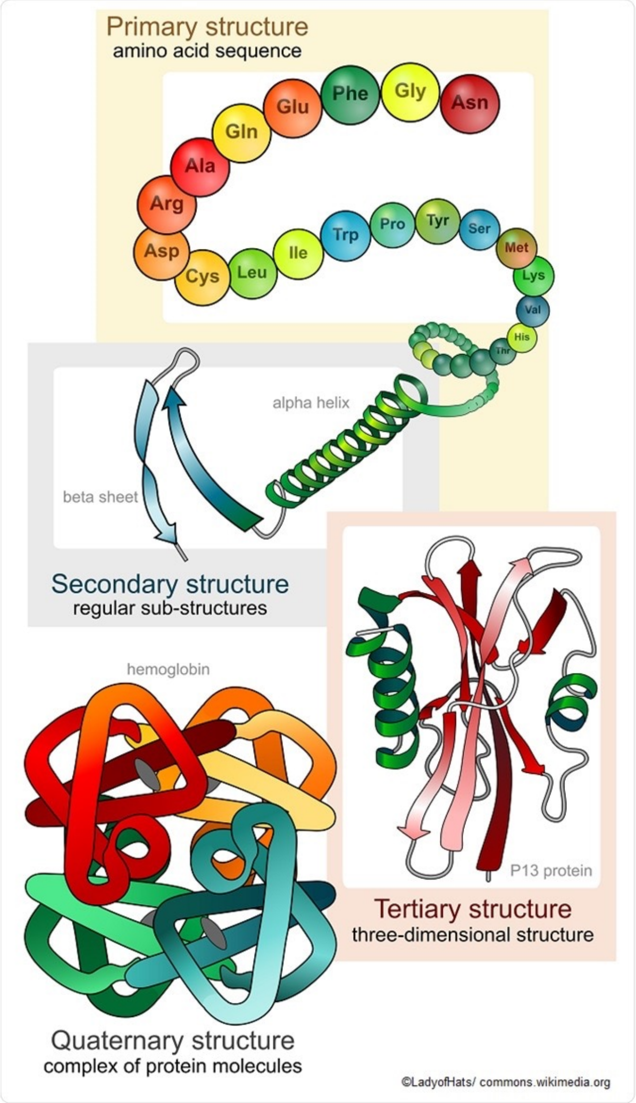


1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	Phe (F)	TCT	Ser (S)	TAT	Tyr (Y)	TGT	Cys (C)	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu (L)	TCA		TAA	STOP	TGA	STOP	A
	TTG		TCG		TAG		TGG	Trp (W)	G
C	CTT		CCT	Pro (P)	CAT	His (H)	CGT	Arg (R)	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln (Q)	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile (I)	ACT	Thr (T)	AAT	Asn (N)	AGT	Ser (S)	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys (K)	AGA	Arg (R)	A
	ATG	Met (M)	ACG		AAG		AGG		G
G	GTT	Val (V)	GCT	Ala (A)	GAT	Asp (D)	GGT	Gly (G)	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu (E)	GGA		A
	GTG		GCG		GAG		GGG		G

Proteins interact with each other

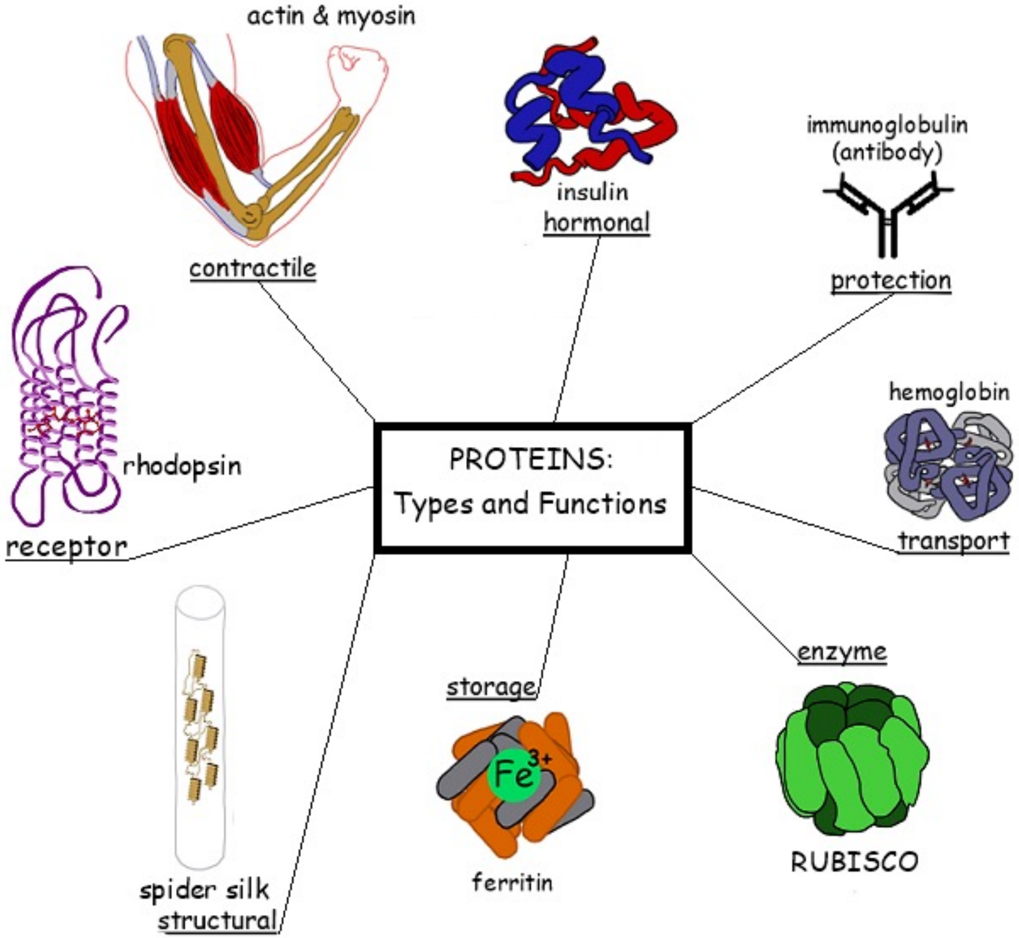


Protein structure affects function

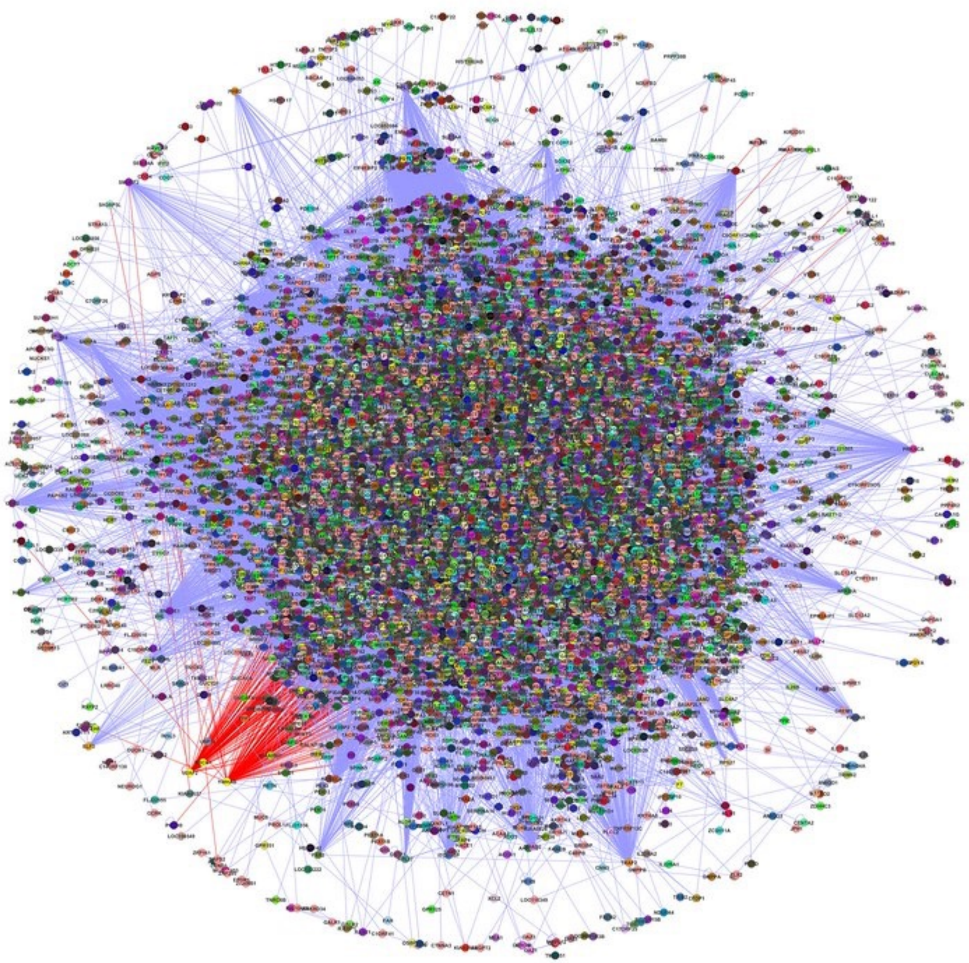


©LadyofHats/ commons.wikimedia.org

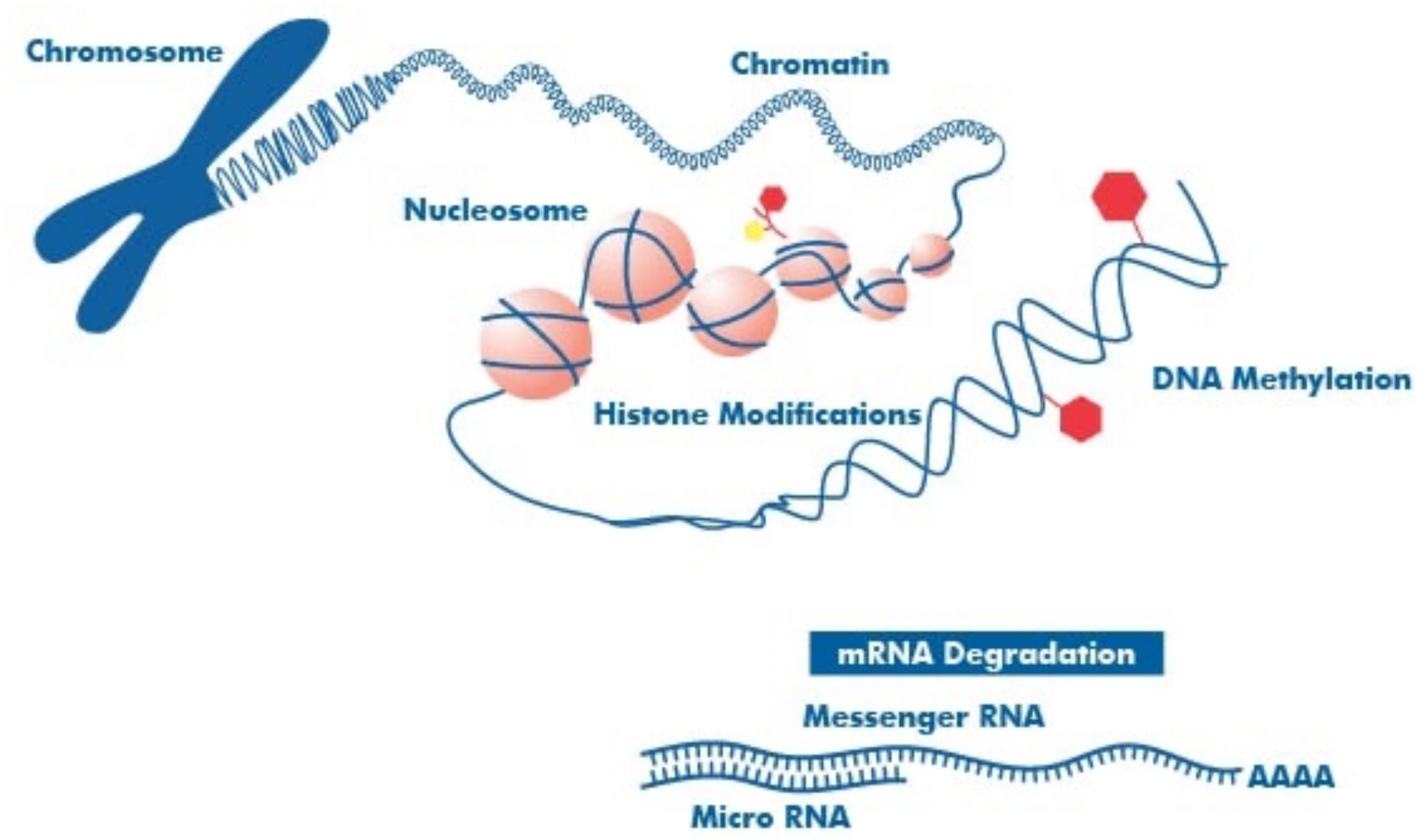
Protein structure affects function



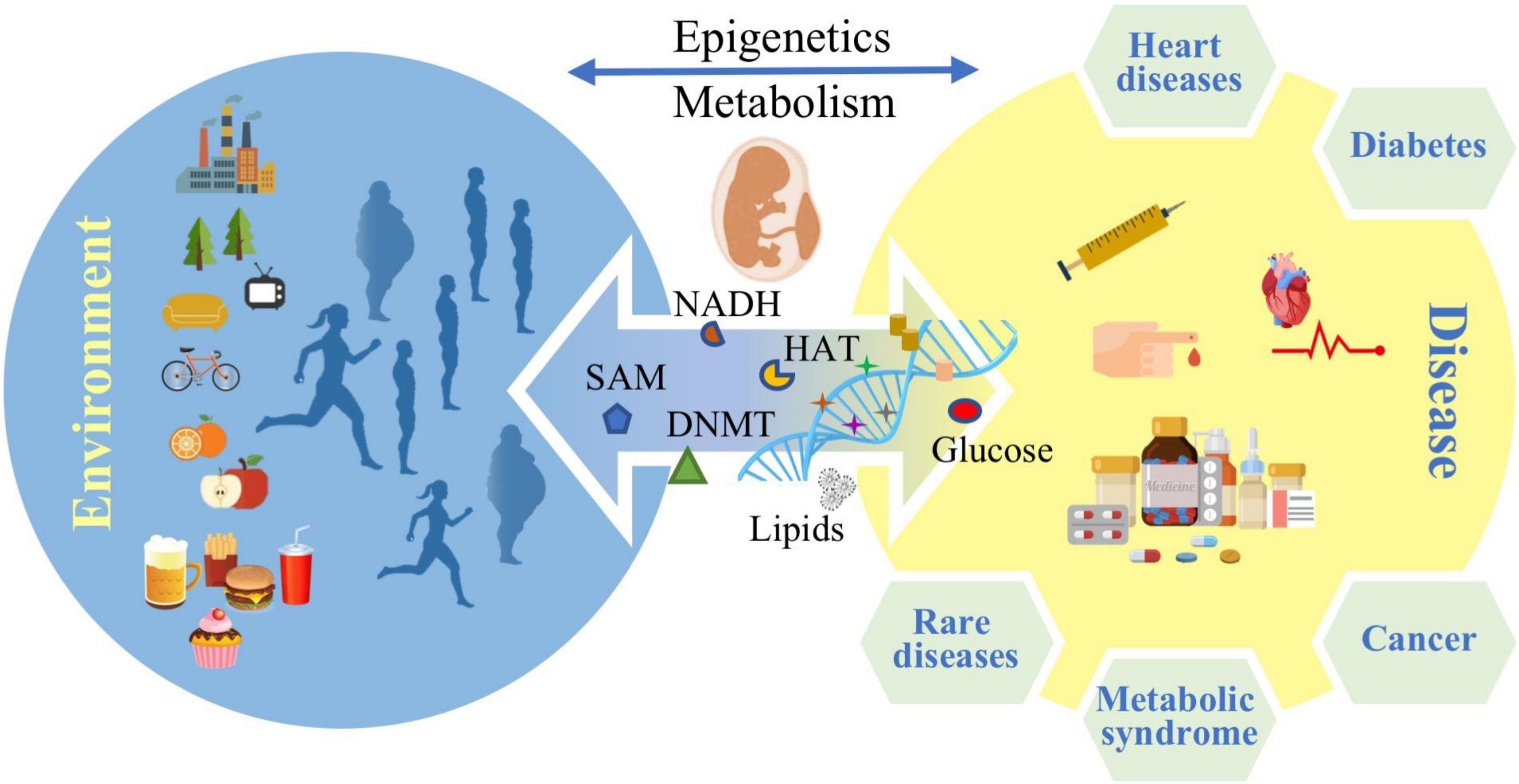
Protein-protein interaction networks



Epigenetics

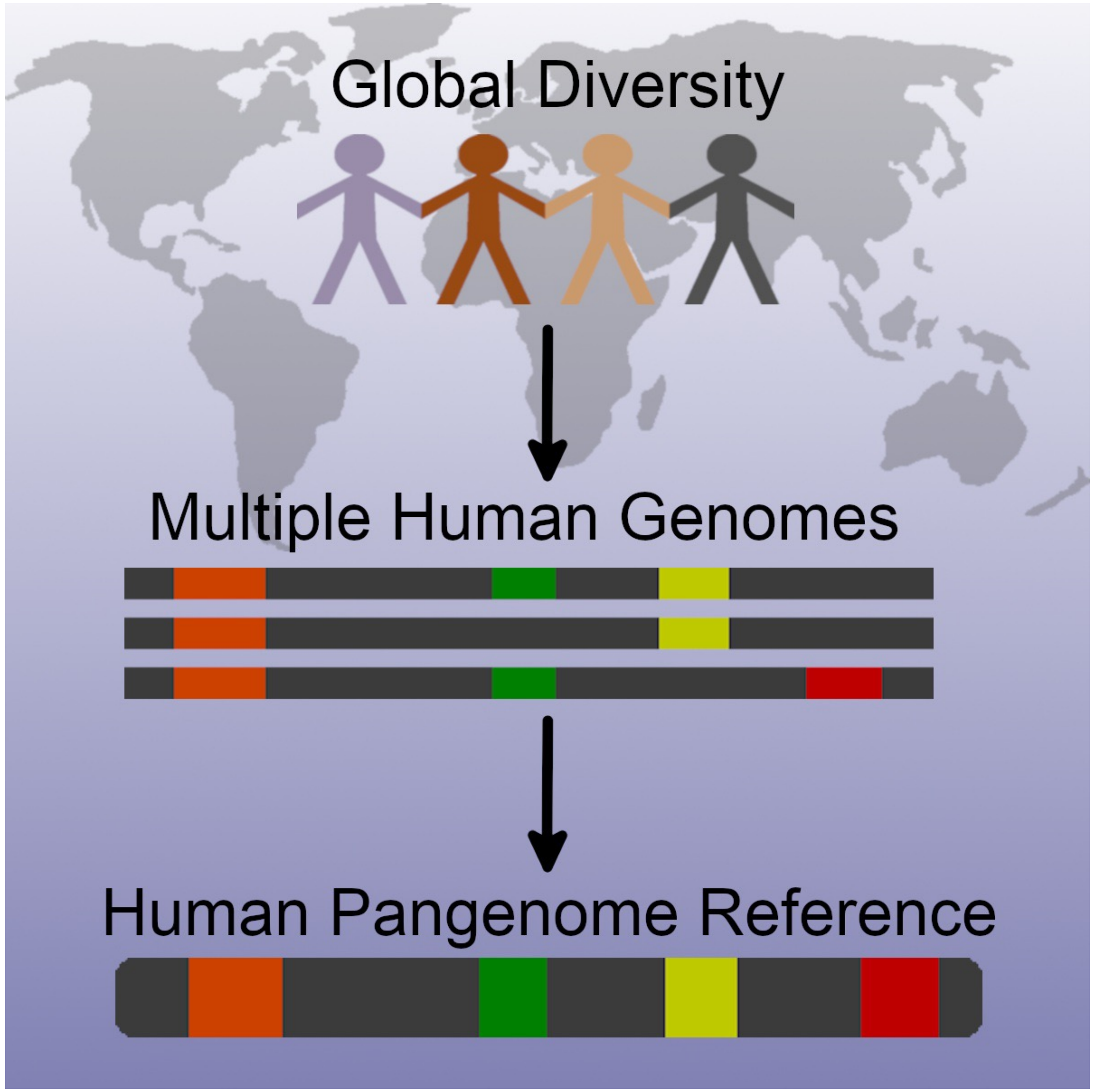


Epigenetics and Health

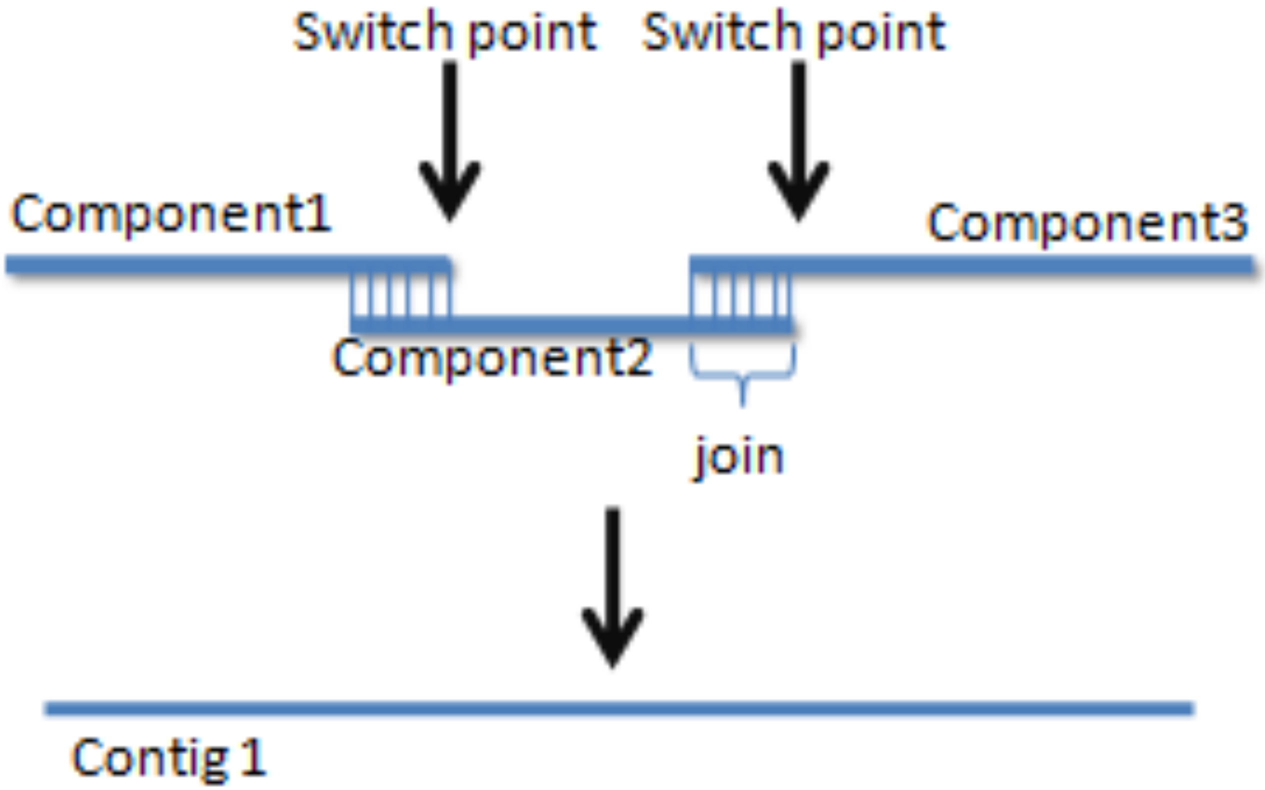


Bioinformatics

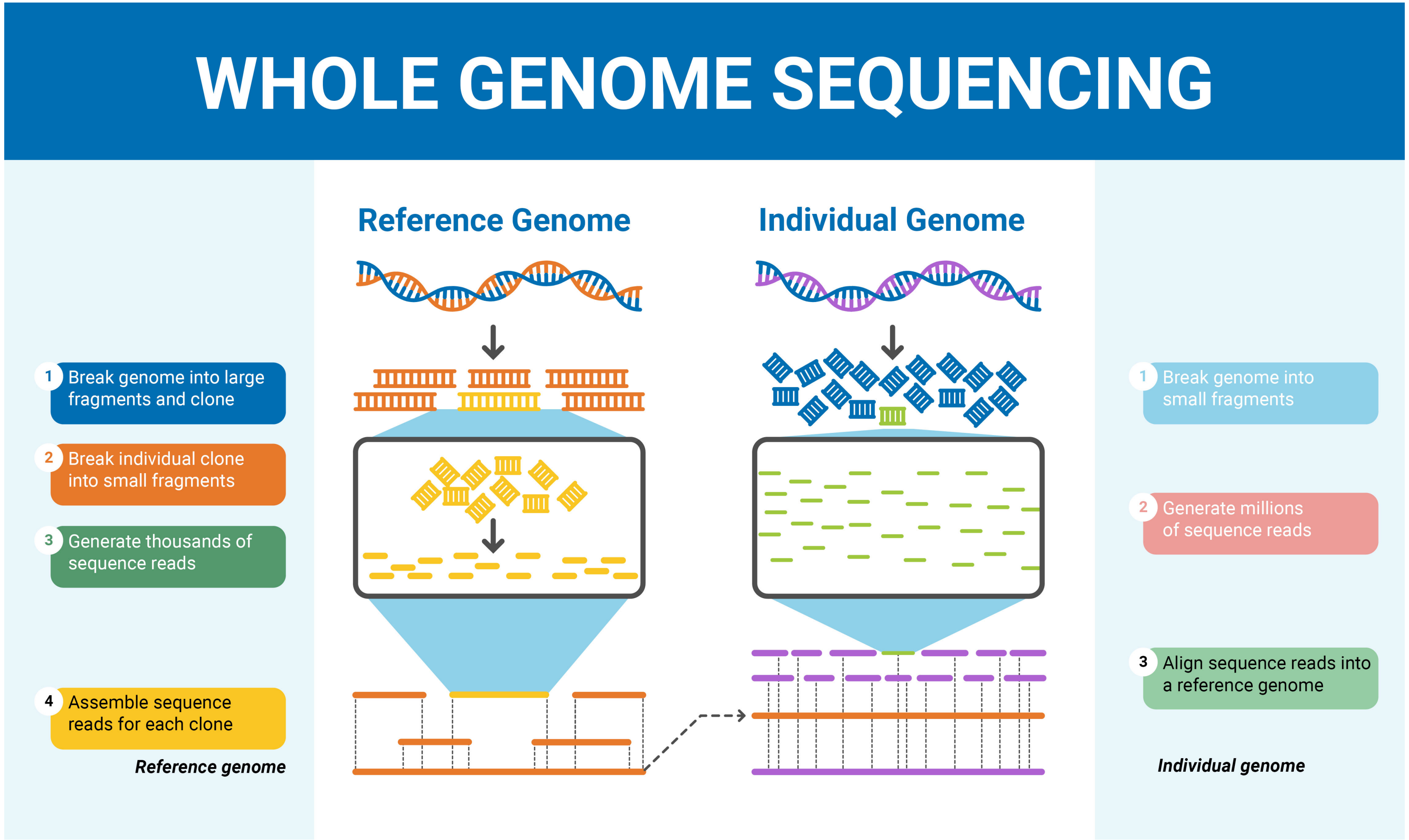
The Human Reference Genome



Reference Genome Construction



Whole Genome Sequencing



WHOLE GENOME SEQUENCING

Reference Genome

1. Break genome into large fragments and clone
2. Break individual clone into small fragments
3. Generate thousands of sequence reads
4. Assemble sequence reads for each clone

Reference genome

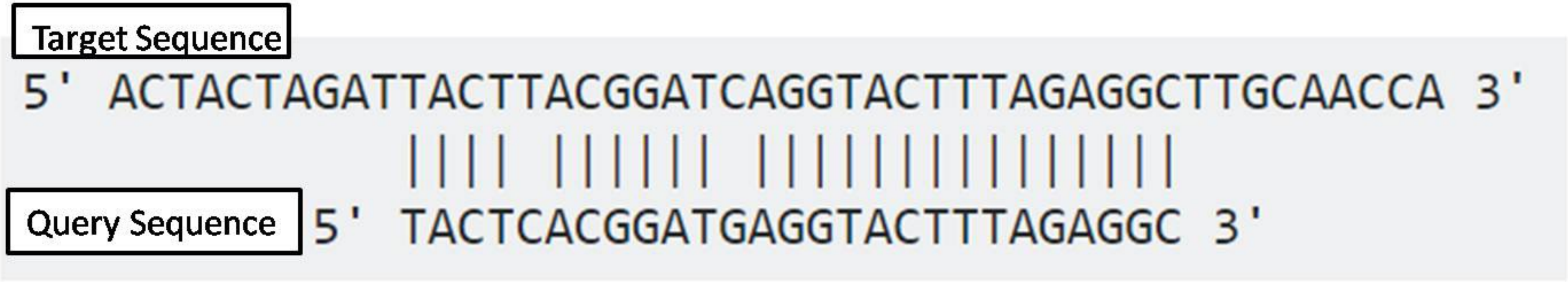
Individual Genome

1. Break genome into small fragments
2. Generate millions of sequence reads
3. Align sequence reads into a reference genome

Individual genome

Local vs Global Alignment

Local Alignment

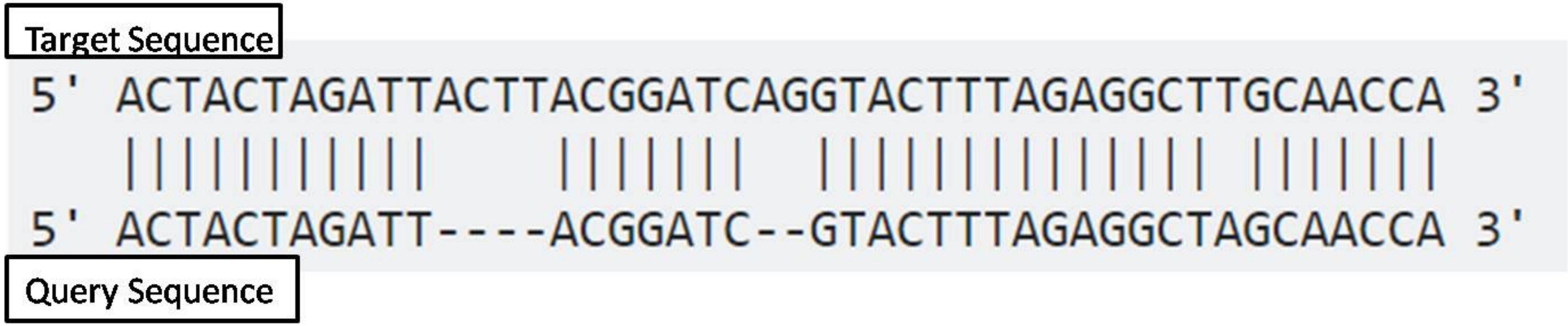


Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Query Sequence

Global Alignment

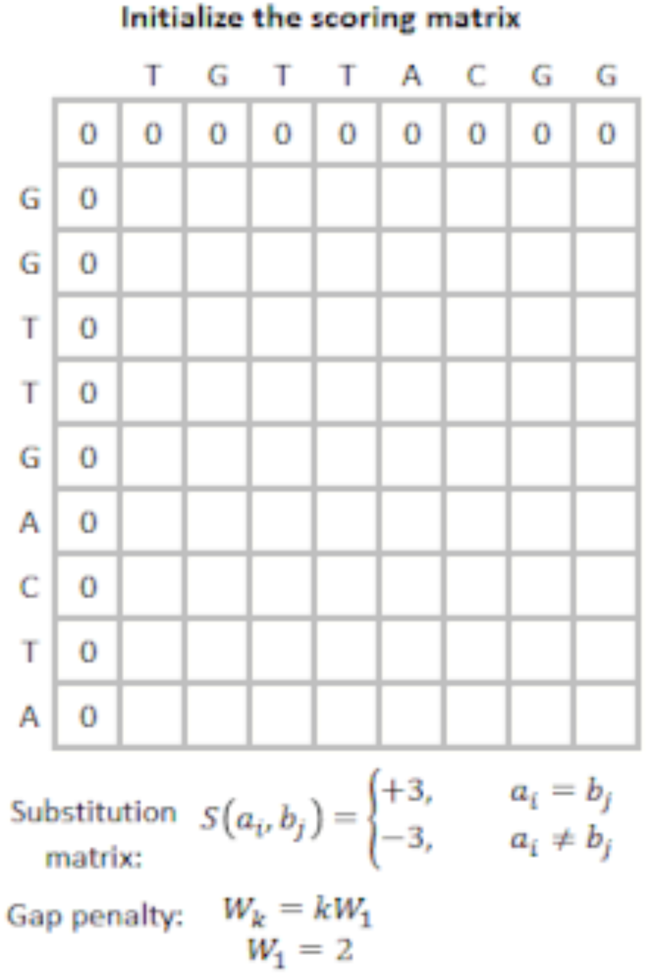


Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

Local Sequence Alignment: Smith-Waterman Algorithm



Initialize the scoring matrix

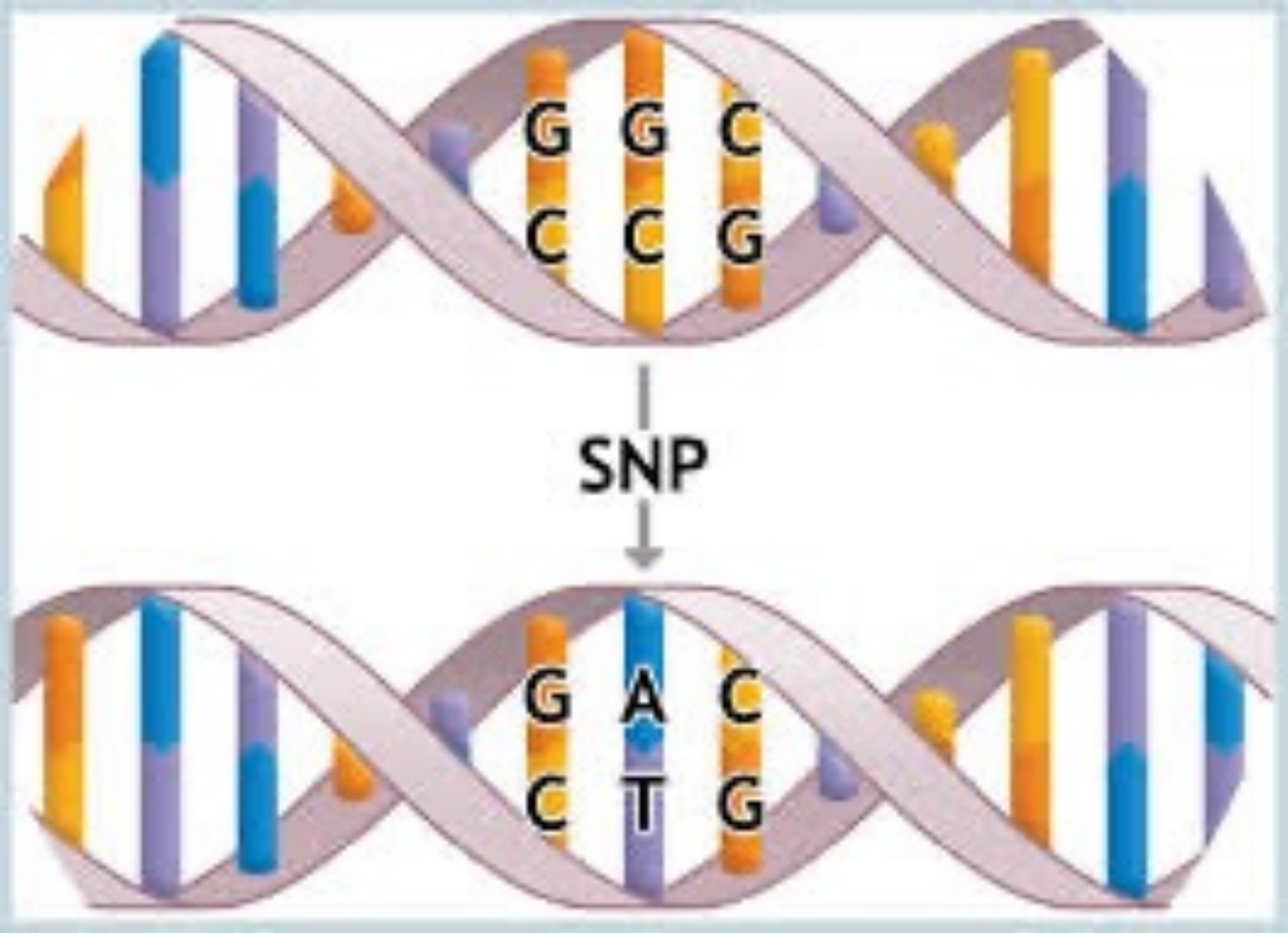
		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $S(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

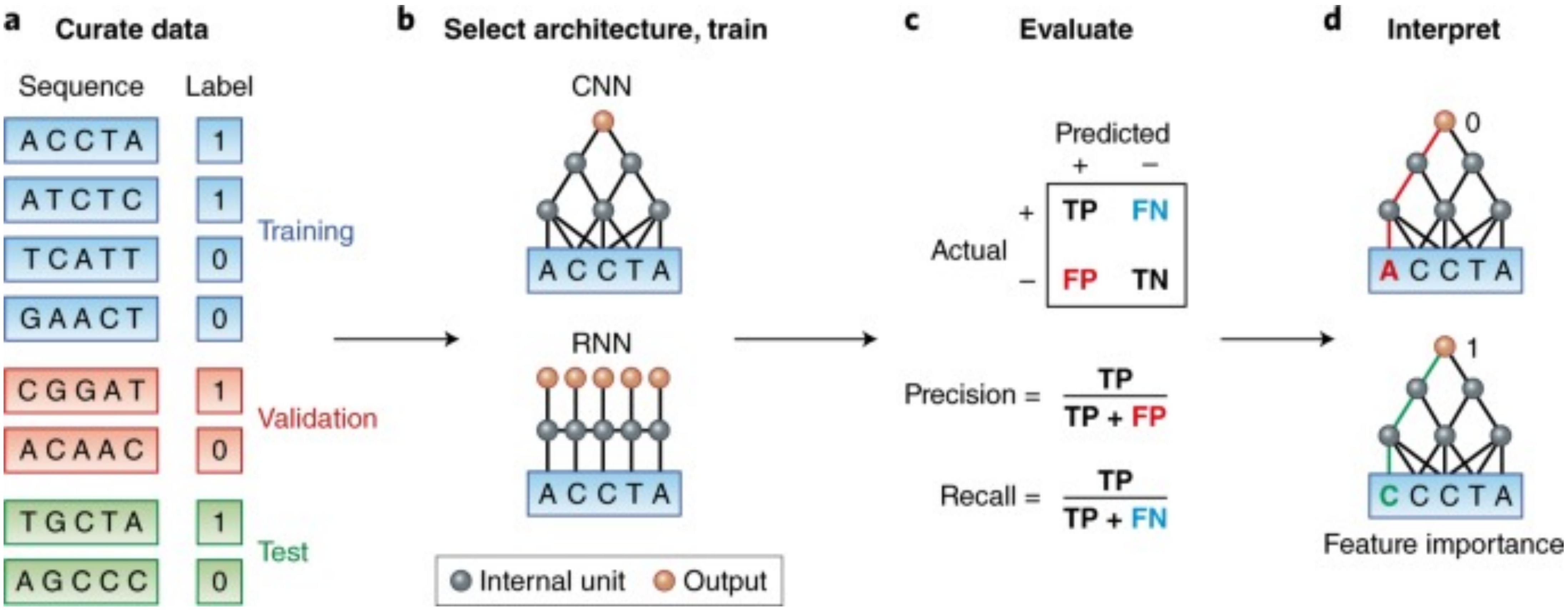
Gap penalty: $W_k = kW_1$

$W_1 = 2$

Single Nucleotide Polymorphism (SNPs)



AI in Genomics



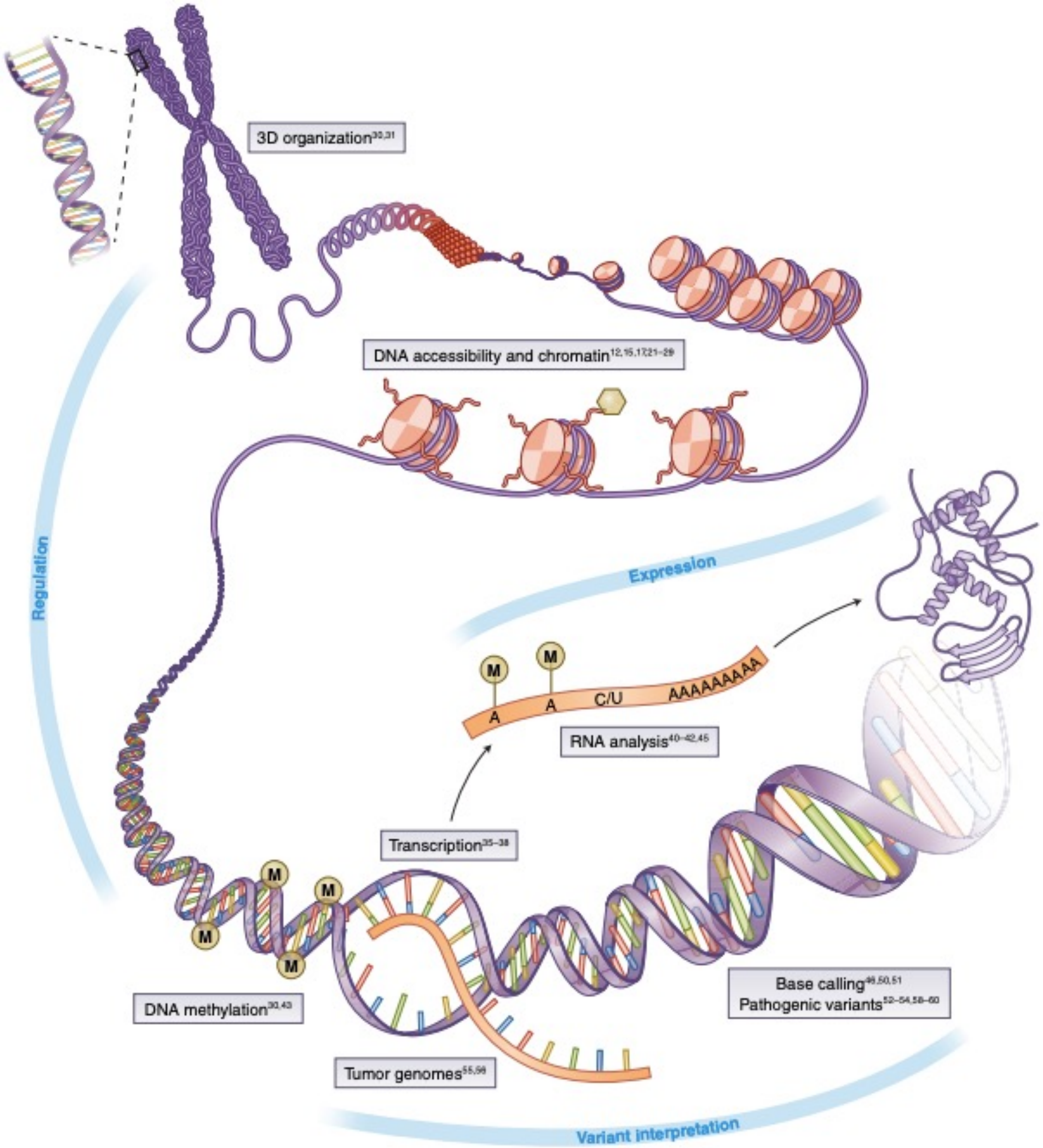
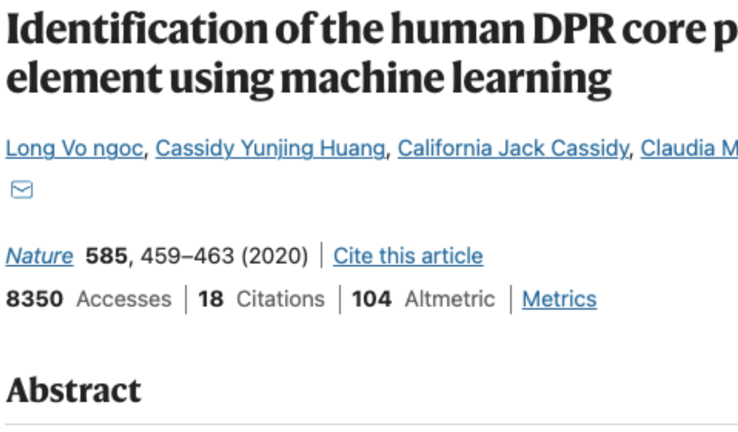


Fig. 2 | Applications of deep learning in genomics. The boxes highlight several application domains and references discussed in the text. Image adapted with permission from ref. [65], Springer Nature.

Understanding Human Diseases using AI



Identification of the human DPR core promoter element using machine learning

Long Vo ngoc, Cassidy Yunjing Huang, California Jack Cassidy, Claudia Medrano & James T. Kadonaga

Nature **585**, 459–463 (2020) | Cite this article

8350 Accesses | **18** Citations | **104** Altmetric | Metrics

Abstract

The RNA polymerase II (Pol II) core promoter is the strategic site of convergence of the signals

ription[1,2,3,4,5], but the downstream core promoter in

d[1,2,3]. Here we analyse the human Pol II core promoter

redictive models for the downstream core promoter

eloped a method termed HARPE (high-throughput

ents) to create hundreds of thousands of DPR (or TATA

iptional strength. We then analysed the HARPE data by

vide

and found that the SVR-based approach is m

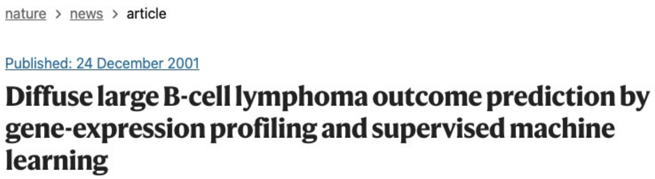
These re

hat is wi

DPR an

ese find

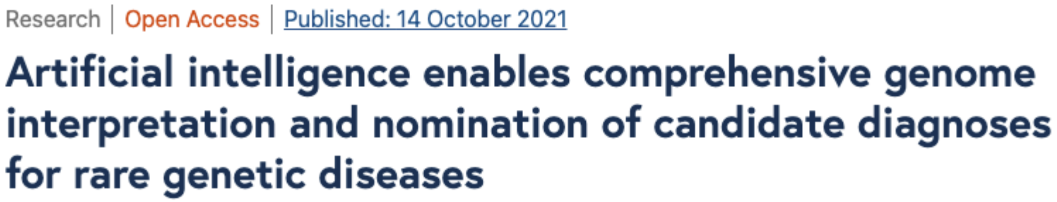
ysis of a



nature › news › article

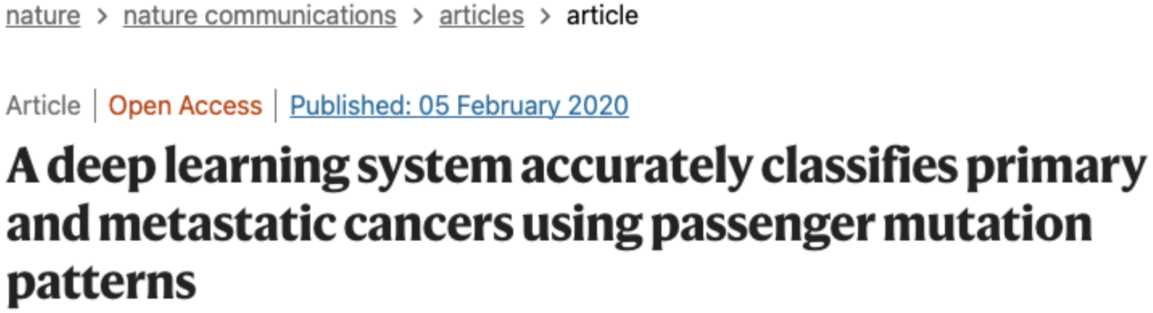
Published: 24 December 2001

Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning



Research | Open Access | Published: 14 October 2021

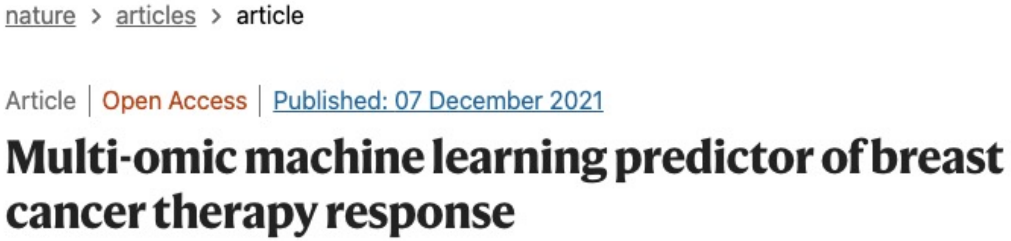
Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases



nature › nature communications › articles › article

Article | Open Access | Published: 05 February 2020

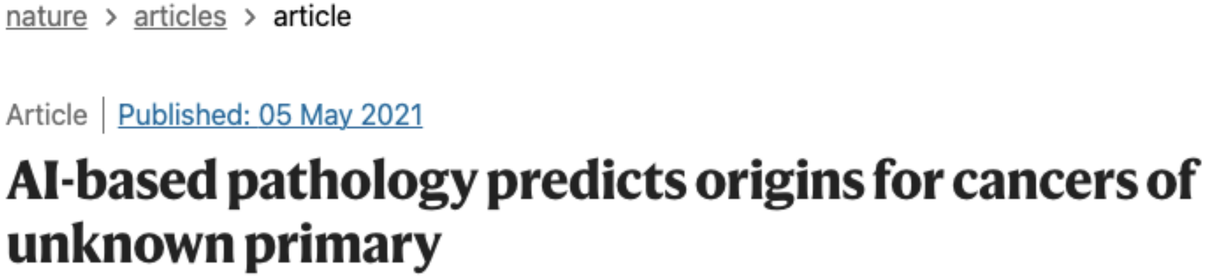
A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns



nature › articles › article

Article | Open Access | Published: 07 December 2021

Multi-omic machine learning predictor of breast cancer therapy response

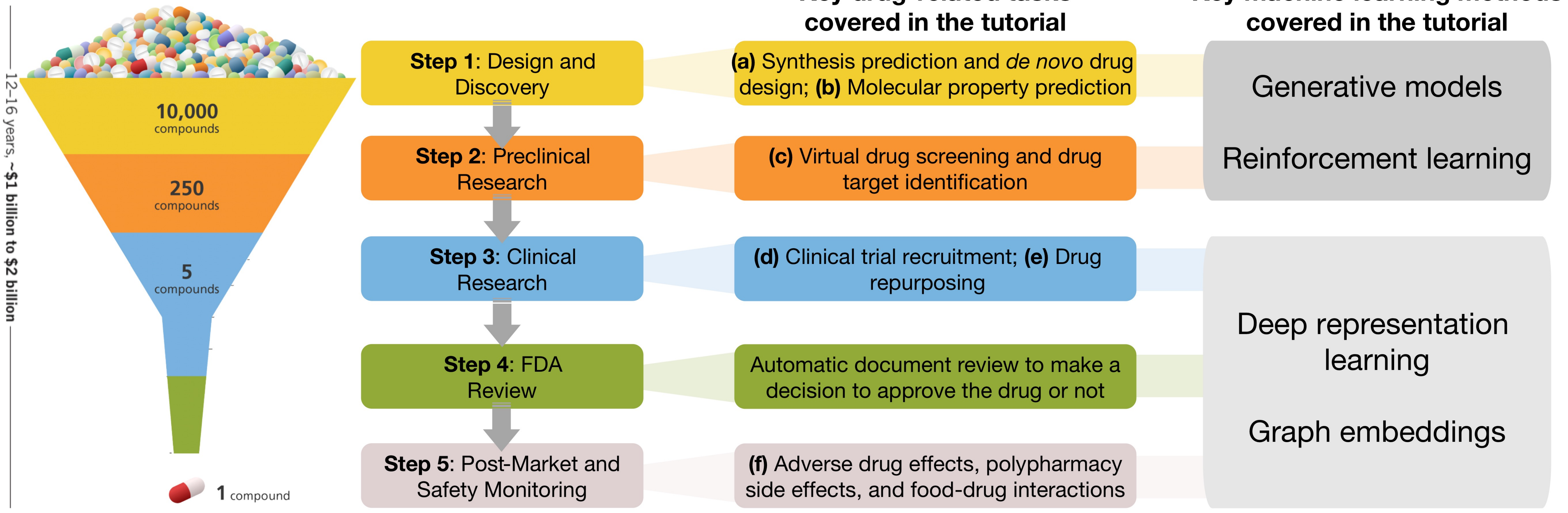


nature › articles › article

Article | Published: 05 May 2021

AI-based pathology predicts origins for cancers of unknown primary

Drug Discovery using AI



Key drug-related tasks covered in the tutorial

Key machine learning methods covered in the tutorial

Step 1: Design and Discovery

(a) Synthesis prediction and *de novo* drug design; **(b)** Molecular property prediction

Step 2: Preclinical Research

(c) Virtual drug screening and drug target identification

Step 3: Clinical Research

(d) Clinical trial recruitment; **(e)** Drug repurposing

Step 4: FDA Review

Automatic document review to make a decision to approve the drug or not

Step 5: Post-Market and Safety Monitoring

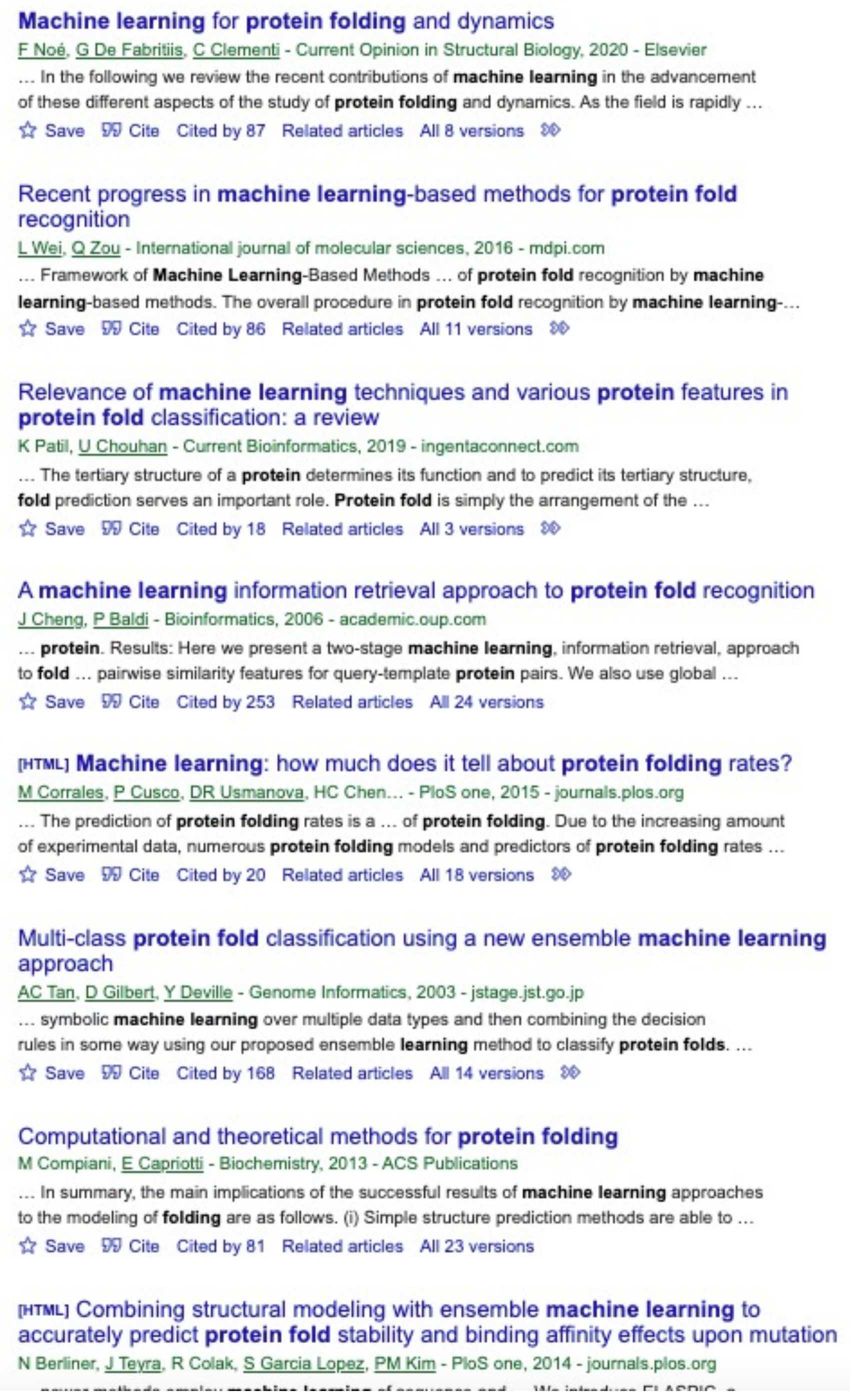
(f) Adverse drug effects, polypharmacy side effects, and food-drug interactions

Generative models

Reinforcement learning

Deep representation learning

Graph embeddings



Machine learning for **protein folding** and dynamics
F Noé, G De Fabritiis, C Clementi - Current Opinion in Structural Biology, 2020 - Elsevier
… In the following we review the recent contributions of **machine learning** in the advancement of these different aspects of the study of **protein folding** and dynamics. As the field is rapidly …
☆ Save ❞ Cite Cited by 87 Related articles All 8 versions »

Recent progress in **machine learning**-based methods for **protein fold** recognition
L Wei, Q Zou - International journal of molecular sciences, 2016 - mdpi.com
… Framework of **Machine Learning**-Based Methods … of **protein fold** recognition by **machine learning**-based methods. The overall procedure in **protein fold** recognition by **machine learning**-…
☆ Save ❞ Cite Cited by 86 Related articles All 11 versions »

Relevance of **machine learning** techniques and various **protein** features in **protein fold** classification: a review
K Patil, U Chouhan - Current Bioinformatics, 2019 - ingentaconnect.com
… The tertiary structure of a **protein** determines its function and to predict its tertiary structure, **fold** prediction serves an important role. **Protein fold** is simply the arrangement of the …
☆ Save ❞ Cite Cited by 18 Related articles All 3 versions »

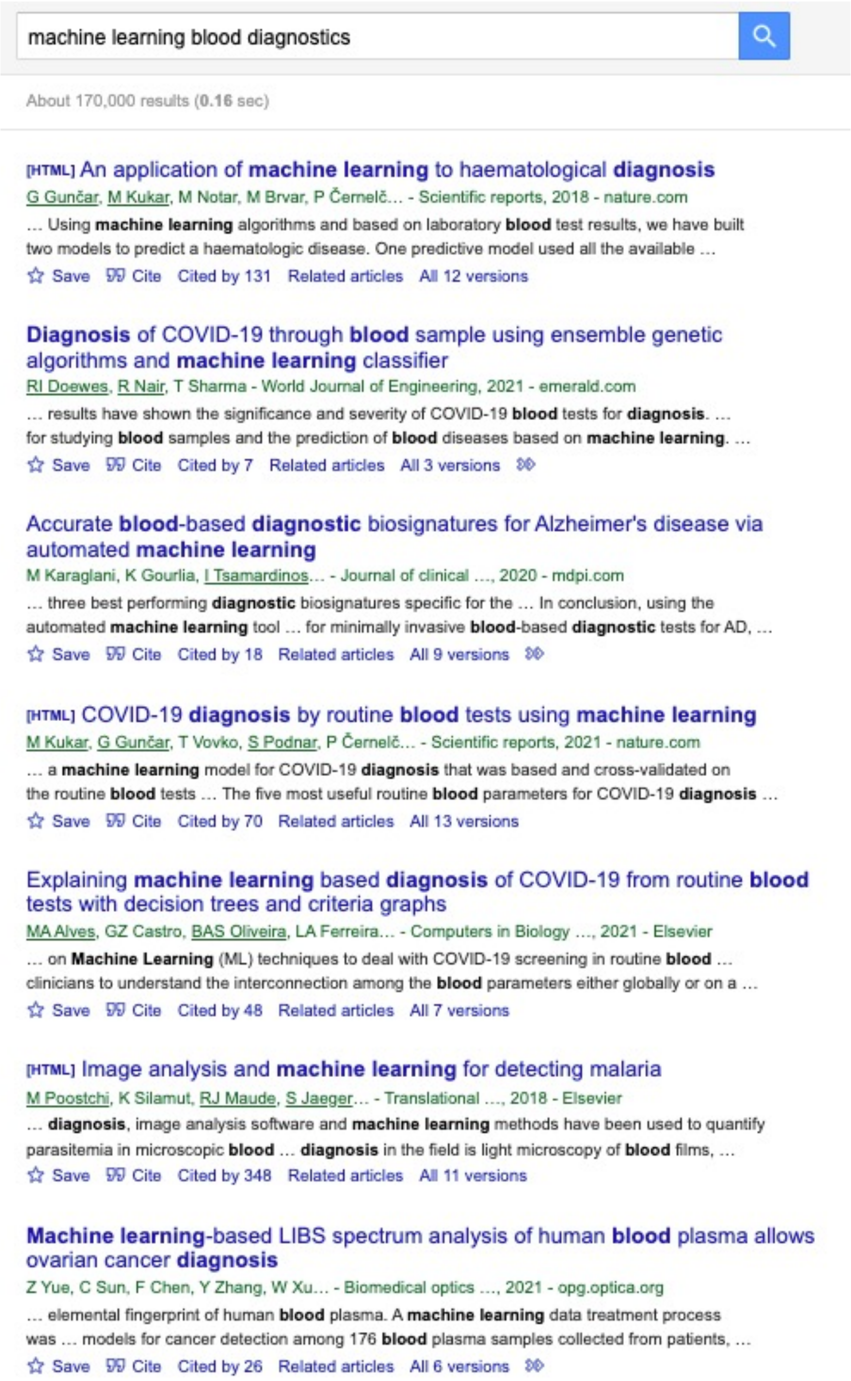
A **machine learning** information retrieval approach to **protein fold** recognition
J Cheng, P Baldi - Bioinformatics, 2006 - academic.oup.com
… **protein**. Results: Here we present a two-stage **machine learning**, information retrieval, approach to **fold** … pairwise similarity features for query-template **protein** pairs. We also use global …
☆ Save ❞ Cite Cited by 253 Related articles All 24 versions

[HTML] **Machine learning**: how much does it tell about **protein folding** rates?
M Corrales, P Cusco, DR Usmanova, HC Chen… - PloS one, 2015 - journals.plos.org
… The prediction of **protein folding** rates is a … of **protein folding**. Due to the increasing amount of experimental data, numerous **protein folding** models and predictors of **protein folding** rates …
☆ Save ❞ Cite Cited by 20 Related articles All 18 versions »

Multi-class **protein fold** classification using a new ensemble **machine learning** approach
AC Tan, D Gilbert, Y Deville - Genome Informatics, 2003 - jstage.jst.go.jp
… symbolic **machine learning** over multiple data types and then combining the decision rules in some way using our proposed ensemble **learning** method to classify **protein folds**. …
☆ Save ❞ Cite Cited by 168 Related articles All 14 versions »

Computational and theoretical methods for **protein folding**
M Compiani, E Capriotti - Biochemistry, 2013 - ACS Publications
… In summary, the main implications of the successful results of **machine learning** approaches to the modeling of **folding** are as follows. (i) Simple structure prediction methods are able to …
☆ Save ❞ Cite Cited by 81 Related articles All 23 versions

[HTML] Combining structural modeling with ensemble **machine learning** to accurately predict **protein fold** stability and binding affinity effects upon mutation
N Berliner, J Teyra, R Colak, S Garcia Lopez, PM Kim - PloS one, 2014 - journals.plos.org



machine learning blood diagnostics

About 170,000 results (**0.16** sec)

[HTML] An application of **machine learning** to haematological **diagnosis**
G Gunčar, M Kukar, M Notar, M Brvar, P Černelč… - Scientific reports, 2018 - nature.com
… Using **machine learning** algorithms and based on laboratory **blood** test results, we have built two models to predict a haematologic disease. One predictive model used all the available …
☆ Save 99 Cite Cited by 131 Related articles All 12 versions

Diagnosis of COVID-19 through **blood** sample using ensemble genetic algorithms and **machine learning** classifier
RI Doewes, R Nair, T Sharma - World Journal of Engineering, 2021 - emerald.com
… results have shown the significance and severity of COVID-19 **blood** tests for **diagnosis**. … for studying **blood** samples and the prediction of **blood** diseases based on **machine learning**. …
☆ Save 99 Cite Cited by 7 Related articles All 3 versions

Accurate **blood**-based **diagnostic** biosignatures for Alzheimer's disease via automated **machine learning**
M Karaglani, K Gourlia, I Tsamardinos… - Journal of clinical …, 2020 - mdpi.com
… three best performing **diagnostic** biosignatures specific for the … In conclusion, using the automated **machine learning** tool … for minimally invasive **blood**-based **diagnostic** tests for AD, …
☆ Save 99 Cite Cited by 18 Related articles All 9 versions

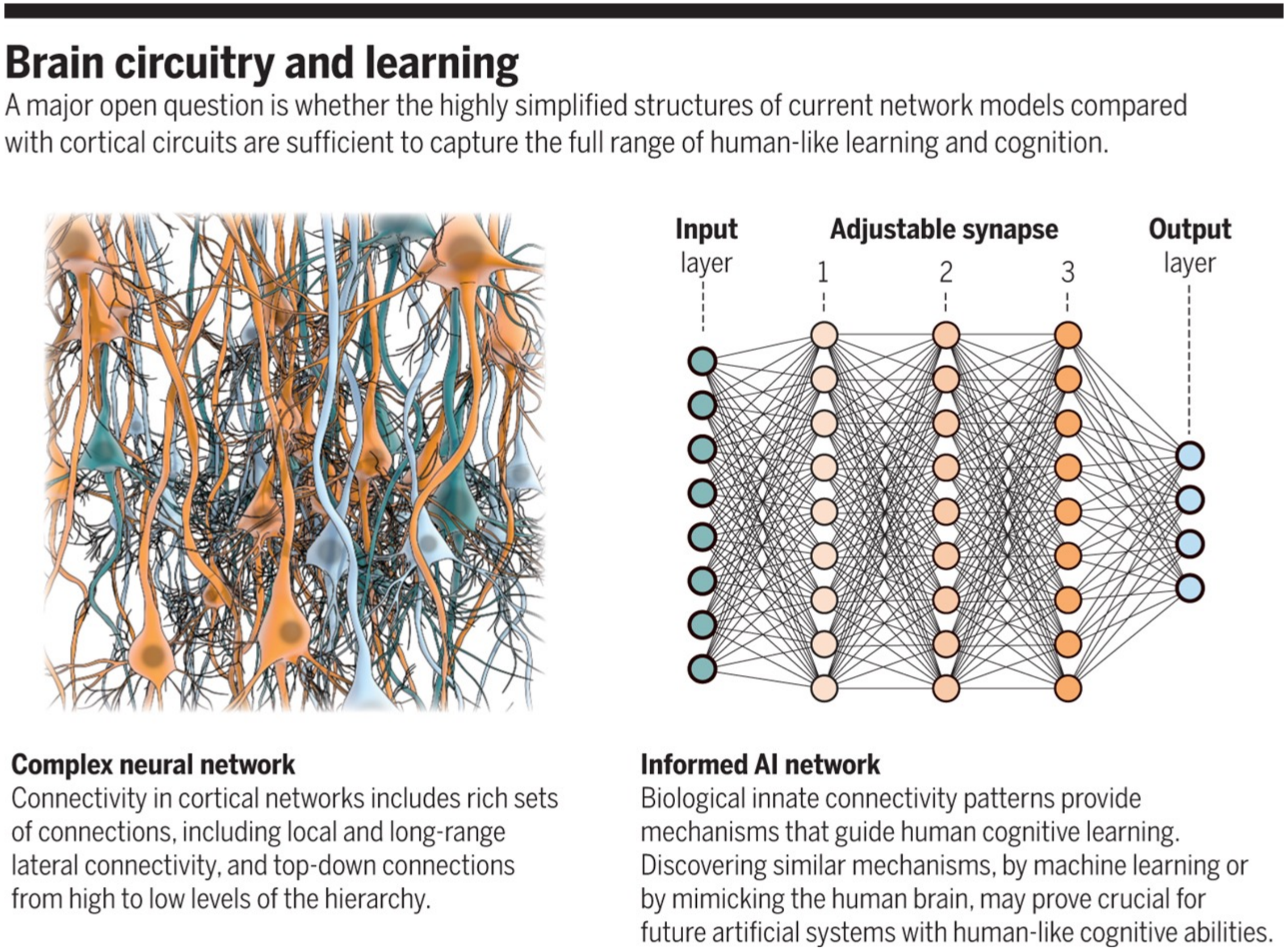
[HTML] COVID-19 **diagnosis** by routine **blood** tests using **machine learning**
M Kukar, G Gunčar, T Vovko, S Podnar, P Černelč… - Scientific reports, 2021 - nature.com
… a **machine learning** model for COVID-19 **diagnosis** that was based and cross-validated on the routine **blood** tests … The five most useful routine **blood** parameters for COVID-19 **diagnosis** …
☆ Save 99 Cite Cited by 70 Related articles All 13 versions

Explaining **machine learning** based **diagnosis** of COVID-19 from routine **blood** tests with decision trees and criteria graphs
MA Alves, GZ Castro, BAS Oliveira, LA Ferreira… - Computers in Biology …, 2021 - Elsevier
… on **Machine Learning** (ML) techniques to deal with COVID-19 screening in routine **blood** … clinicians to understand the interconnection among the **blood** parameters either globally or on a …
☆ Save 99 Cite Cited by 48 Related articles All 7 versions

[HTML] Image analysis and **machine learning** for detecting malaria
M Poostchi, K Silamut, RJ Maude, S Jaeger… - Translational …, 2018 - Elsevier
… **diagnosis**, image analysis software and **machine learning** methods have been used to quantify parasitemia in microscopic **blood** … **diagnosis** in the field is light microscopy of **blood** films, …
☆ Save 99 Cite Cited by 348 Related articles All 11 versions

Machine learning-based LIBS spectrum analysis of human **blood** plasma allows ovarian cancer **diagnosis**
Z Yue, C Sun, F Chen, Y Zhang, W Xu… - Biomedical optics …, 2021 - opg.optica.org
… elemental fingerprint of human **blood** plasma. A **machine learning** data treatment process was … models for cancer detection among 176 **blood** plasma samples collected from patients, …
☆ Save 99 Cite Cited by 26 Related articles All 6 versions

Computational Neuroscience



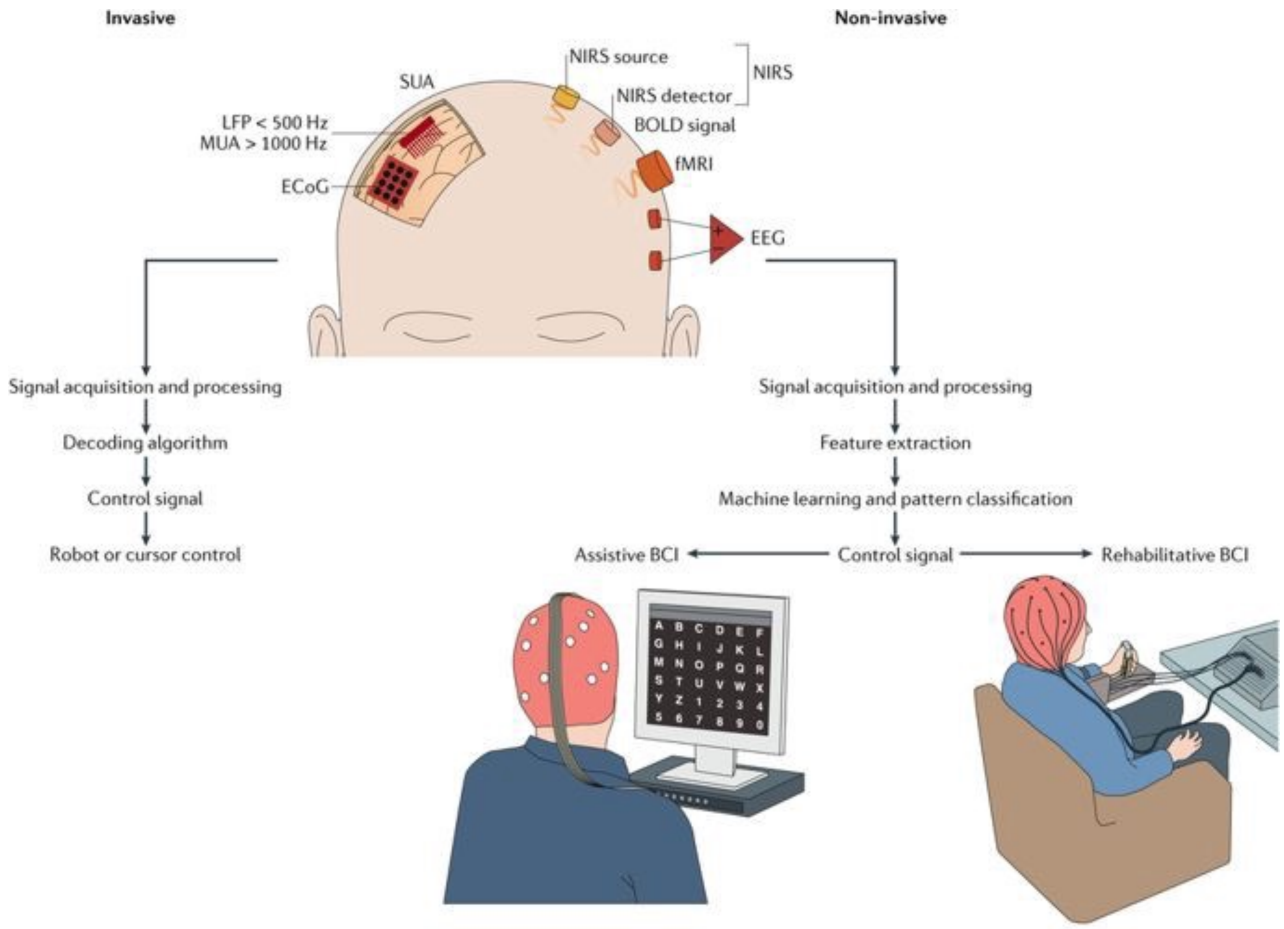
Brain circuitry and learning

A major open question is whether the highly simplified structures of current network models compared with cortical circuits are sufficient to capture the full range of human-like learning and cognition.

Complex neural network
Connectivity in cortical networks includes rich sets of connections, including local and long-range lateral connectivity, and top-down connections from high to low levels of the hierarchy.

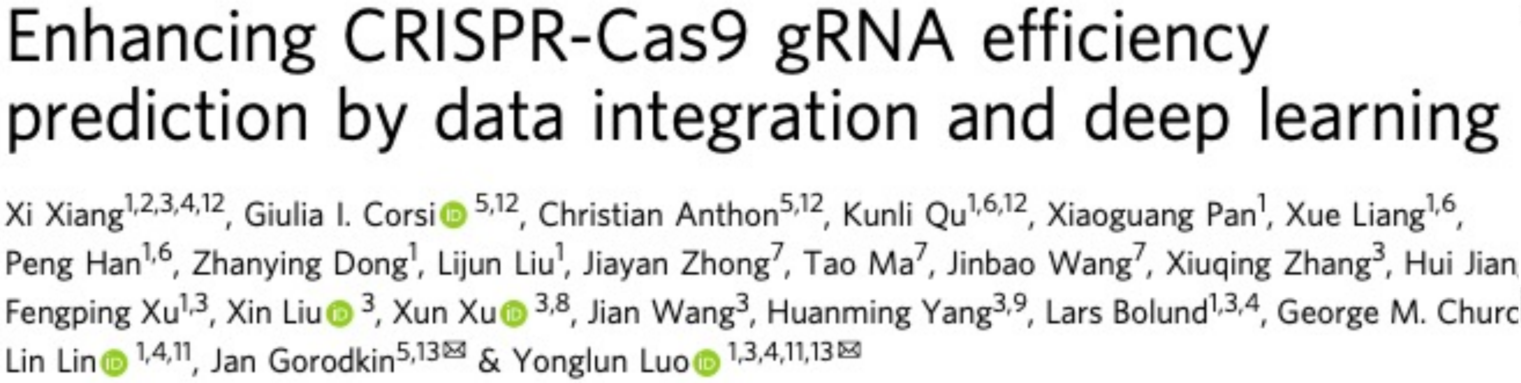
Informed AI network
Biological innate connectivity patterns provide mechanisms that guide human cognitive learning. Discovering similar mechanisms, by machine learning or by mimicking the human brain, may prove crucial for future artificial systems with human-like cognitive abilities.

Brain-Machine Interfaces



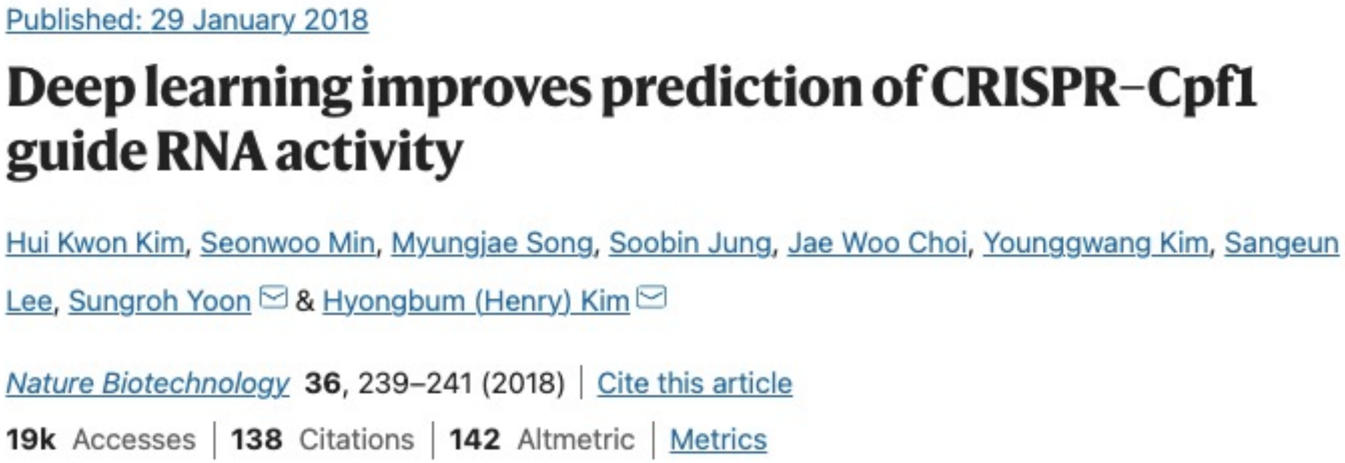
Nature Reviews | Neurology

Improving Gene Editing



Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning

Xi Xiang[1,2,3,4,12], Giulia I. Corsi[5,12], Christian Anthon[5,12], Kunli Qu[1,6,12], Xiaoguang Pan[1], Xue Liang[1,6], Peng Han[1,6], Zhanying Dong[1], Lijun Liu[1], Jiayan Zhong[7], Tao Ma[7], Jinbao Wang[7], Xiuqing Zhang[3], Hui Jian Fengping Xu[1,3], Xin Liu[3], Xun Xu[3,8], Jian Wang[3], Huanming Yang[3,9], Lars Bolund[1,3,4], George M. Churc Lin Lin[1,4,11], Jan Gorodkin[5,13] & Yonglun Luo[1,3,4,11,13]



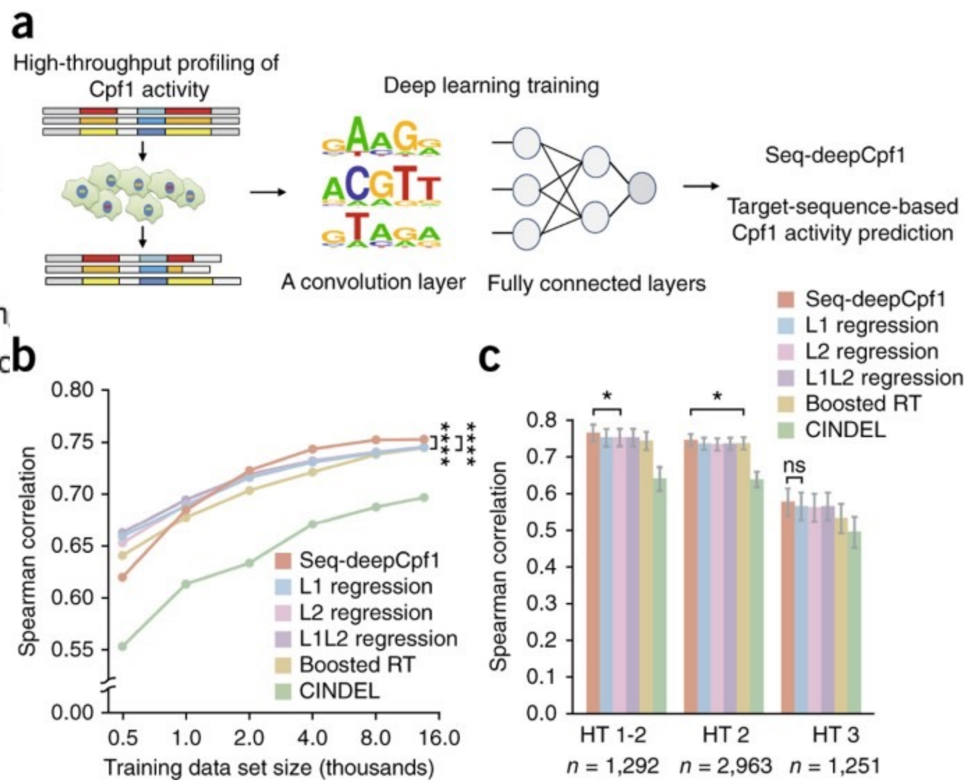
Published: 29 January 2018

Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity

Hui Kwon Kim, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon & Hyongbum (Henry) Kim

Nature Biotechnology **36**, 239–241 (2018) | Cite this article

19k Accesses | **138** Citations | **142** Altmetric | Metrics


a
High-throughput profiling of Cpf1 activity
Deep learning training
A convolution layer
Fully connected layers
Seq-deepCpf1
Target-sequence-based Cpf1 activity prediction
b
Spearman correlation
Training data set size (thousands)
****
****
Seq-deepCpf1
L1 regression
L2 regression
L1L2 regression
Boosted RT
CINDEL
c
Spearman correlation
*
*
ns
HT 1-2
n = 1,292
HT 2
n = 2,963
HT 3
n = 1,251