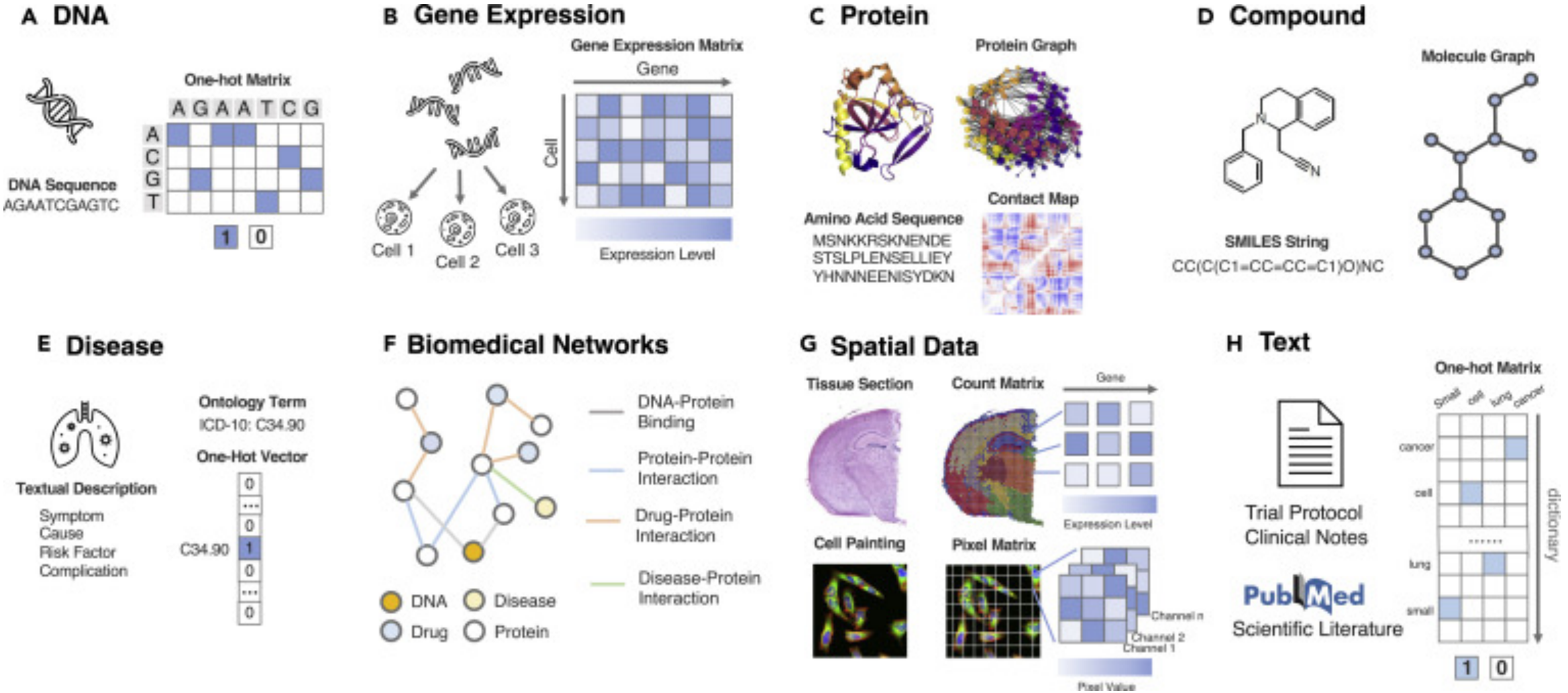
Bioinformatics for Beginners

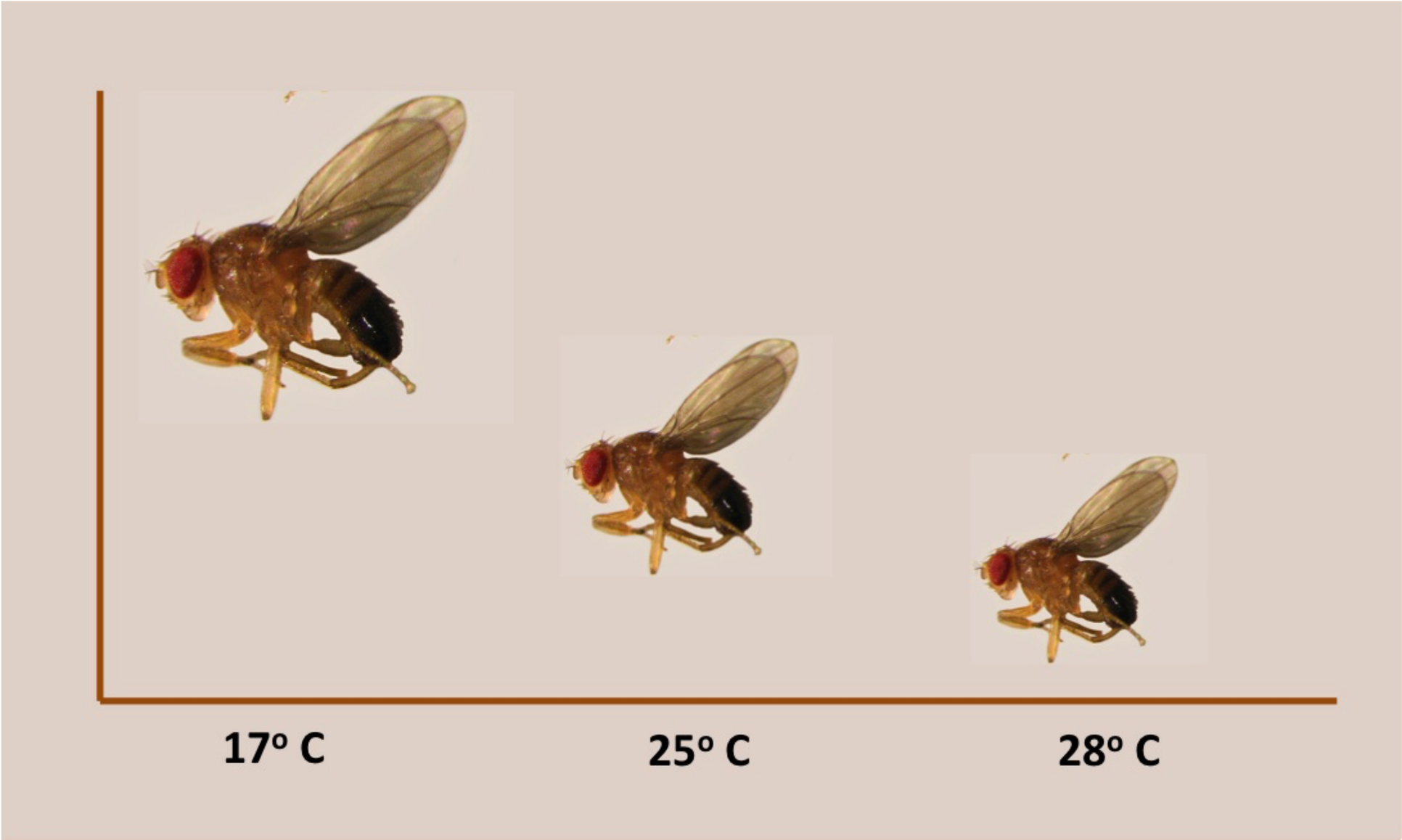
Introduction to Bioinformatics: An Overview

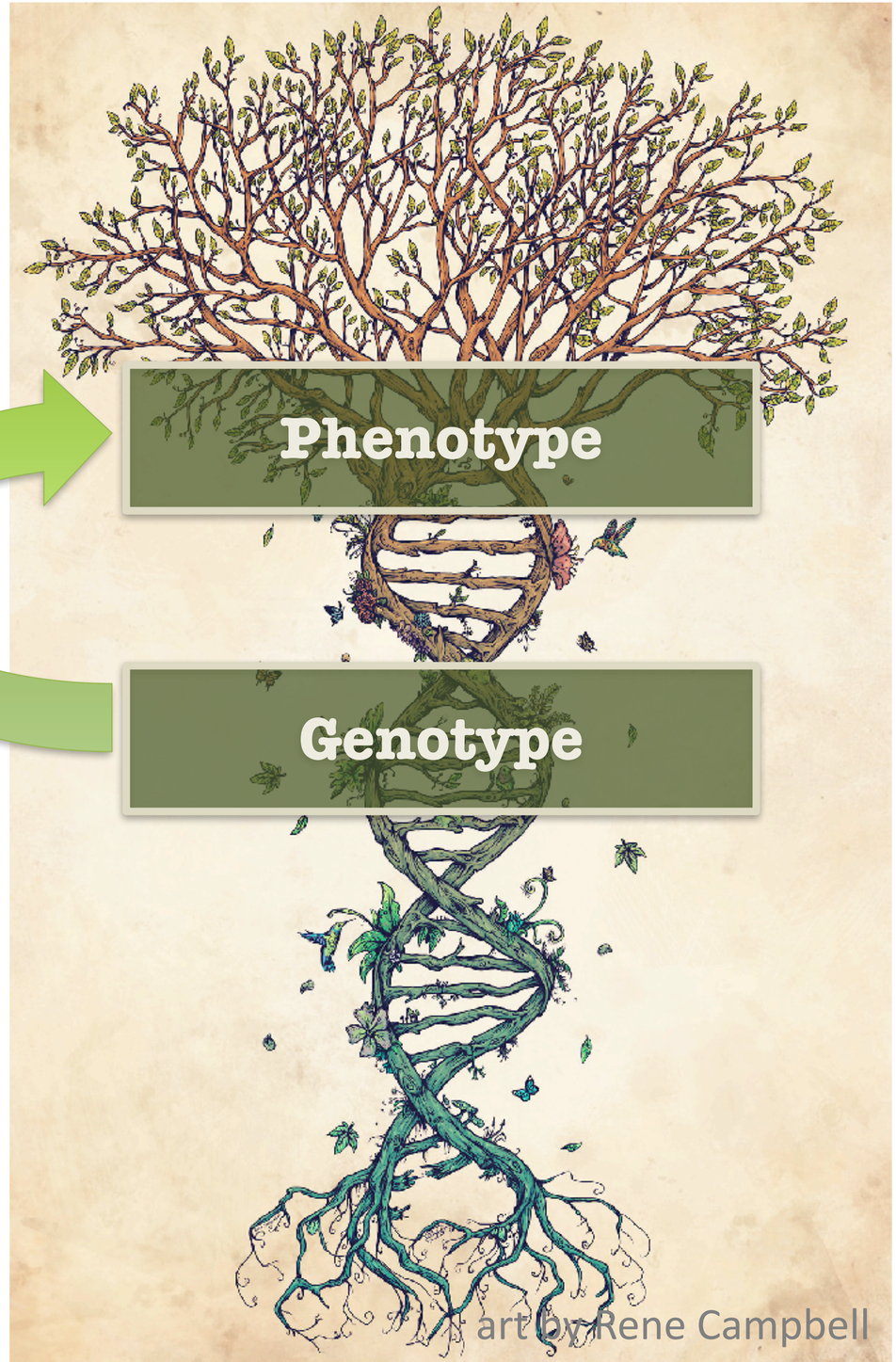
Tugce Bilgin Sonay, ZHAW, May 2023

Bioinformatics

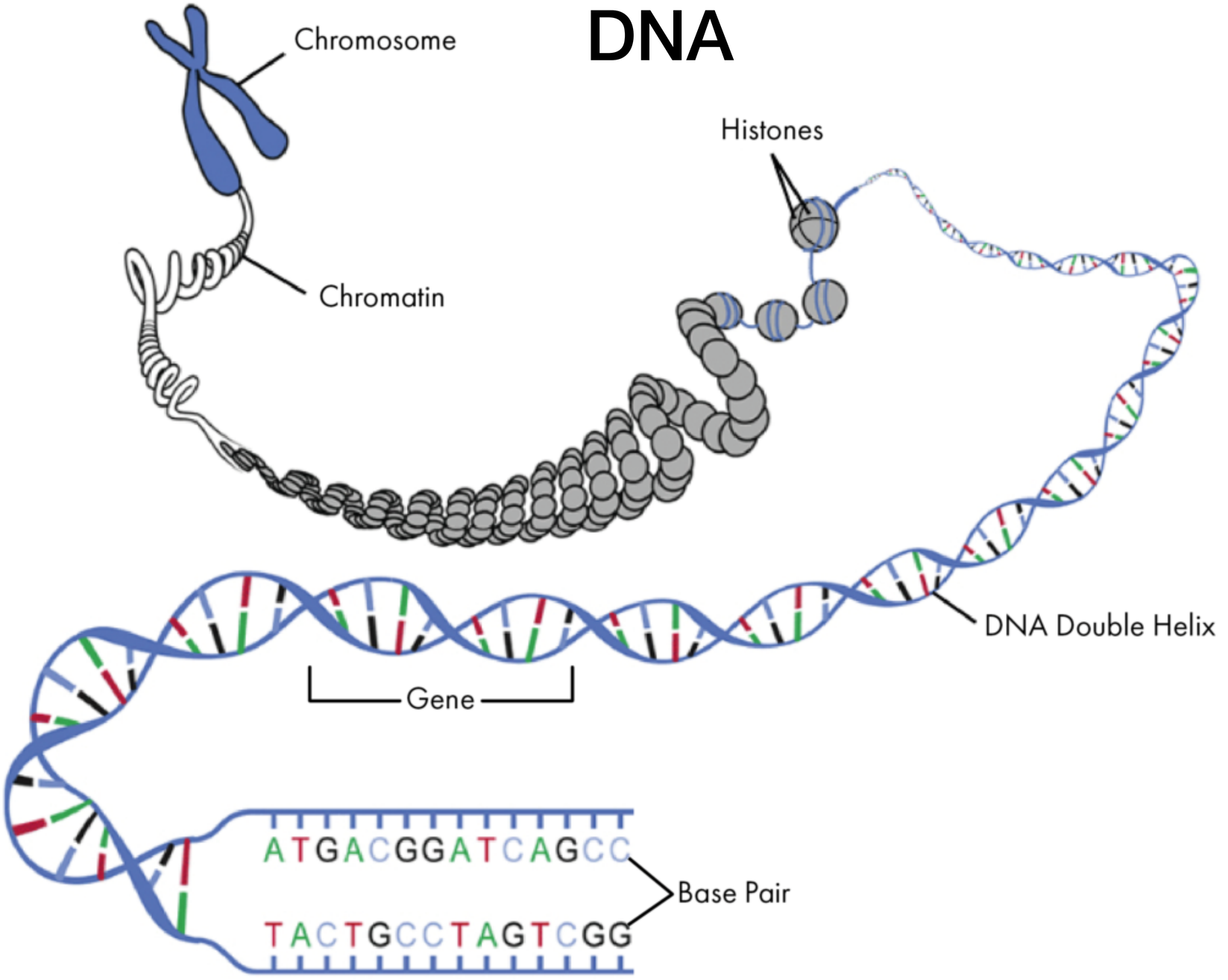


Phenotypic Elasticity

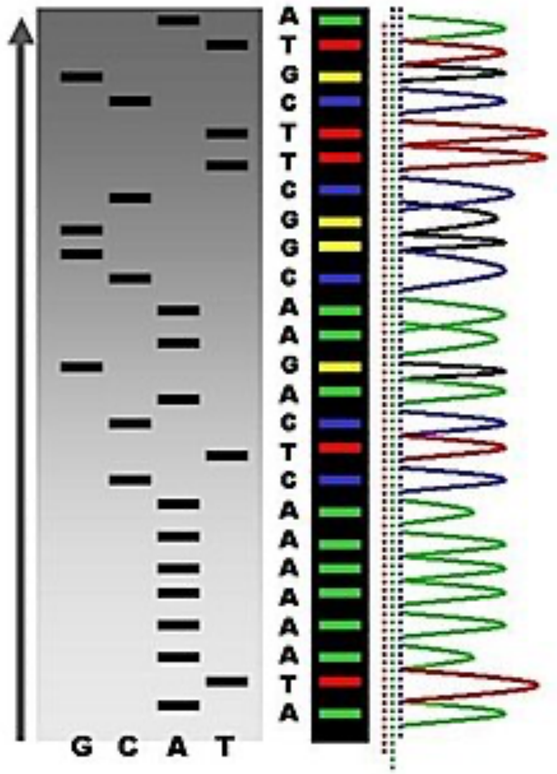



Phenotype
Genotype
art by Rene Campbell

DNA


Chromosome
Chromatin
Histones
DNA Double Helix
Gene
ATGACGGATCAGCC
TACTGCCTAGTCGG
Base Pair

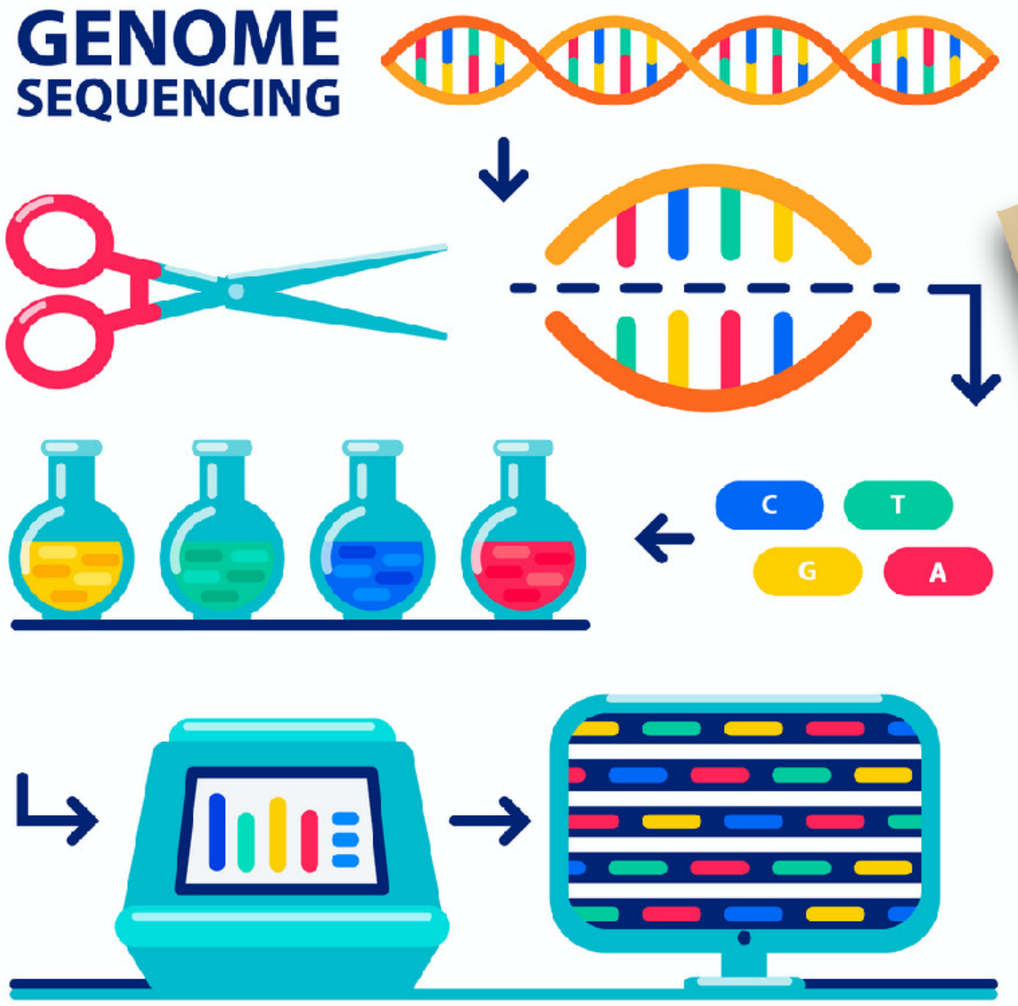
Sanger Sequencing



1977: First bacterial genome
1986: Genes for color blindness
1987: Sequence of the CRISPR
1989: Gene involved in cystic fibrosis
1994: Genes involved cancer
1995: Genome of H.influenza
1996: First eukaryotic genome
1998: First multicellular genome

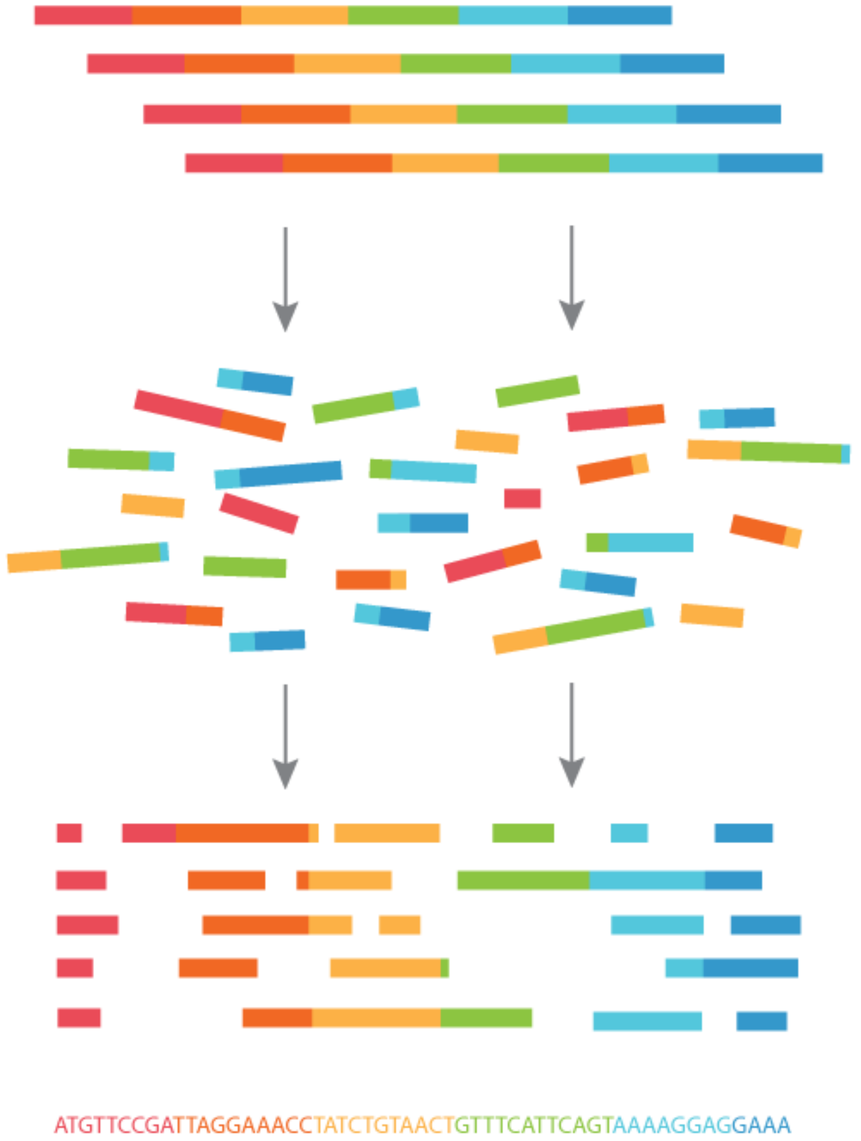
Human Genome Project
1990–2003

13 years
$3 billion
20 countries



VectorStock®
VectorStock.com/22459411

Next Generation Sequencing



A few important findings of the HGP

- ☑ Ethical breakthrough: instant data sharing
- ☑ data analysis and storage technologies
- ☑ We have ~20,000 genes
- ☑ Our chromosomes have different lengths
- ~ Most of the human genome is not coding proteins
- ~ sequence is NOT equal to function

DNA Sequencing Techniques

Next Generation Sequencing

$3,000,000,000	**2003** Human Genome Project	
$20,000,000	**2006** 1st individual genome	
$2,000,000	**2007** 1st NGS Genome	
$200,000	**2008** 1st 30x genome	
$10,000	**2010** 1st sub-10K genome	
$1,000	**2014** 1st $1,000 genome	

DNA Sequencing Techniques

Next Generation Sequencing

$3,000,000,000	20
$20,000,000	2006
$2,000,000	2007
$200,000	
$10,000	
$1,000	2

If you printed all your DNA
(6.4 billion letters), it would fill

4,200 books*

Veritas
sequences your
whole genome

(equivalent to
4,200 books)

* Assuming that a book, like Darwin's *Origin of Species,* has 500 pages.

SEQUENCED GENOME SIZES

35
30
25
20
15
10
5
0

INFLUENZA VIRUS
13,500
BASE PAIRS OF DNA

WALNUT TREE
650 MILLION
BASE PAIRS OF DNA

HUMAN
3 BILLION
BASE PAIRS OF DNA

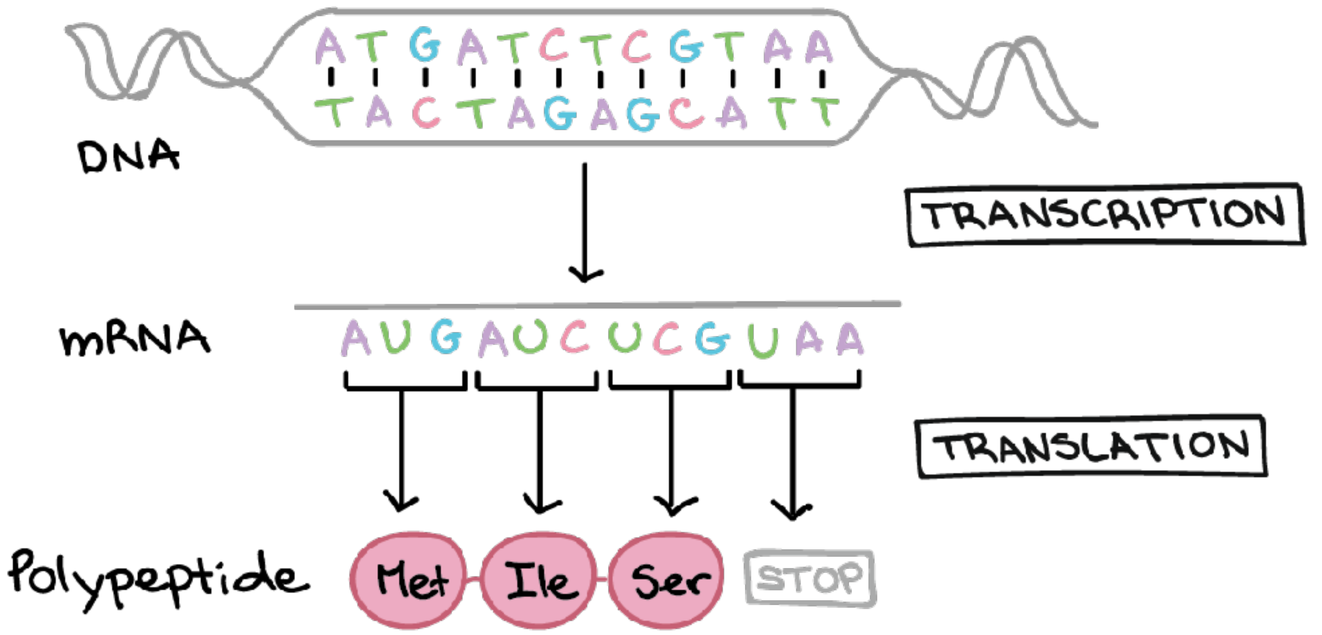
GIANT SEQUOIA
8 BILLION
BASE PAIRS OF DNA

COAST REDWOOD
27 BILLION
BASE PAIRS OF DNA

AXOLOTL SALAMANDER
28.4 BILLION
BASE PAIRS OF DNA

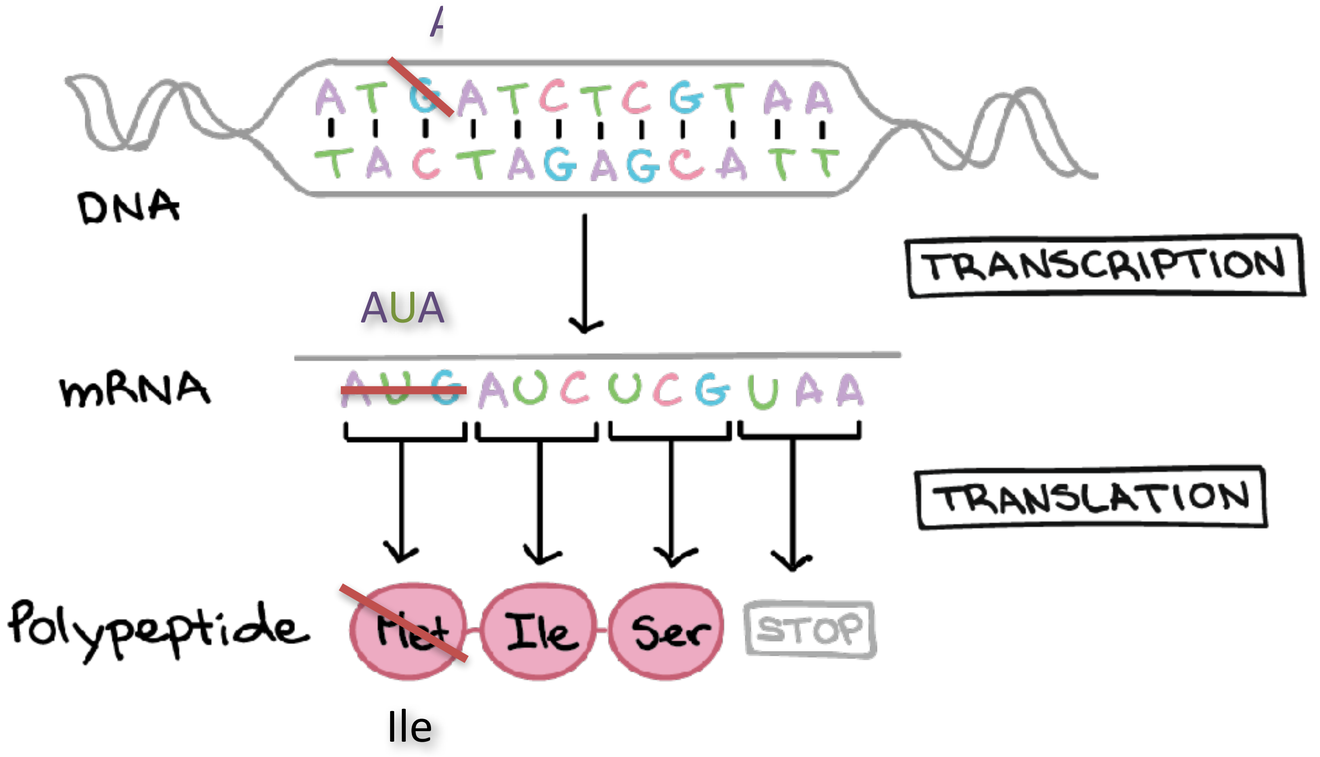
The coast redwood and giant sequoia genomes were sequenced by Save the Redwoods League, University of California, Davis, and Johns Hopkins University. Learn more at SaveTheRedwoods.org/RedwoodGenome.

THE CENTRAL DOGMA



GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

THE CENTRAL DOGMA



GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU	UAA UAG UGA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	stop
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

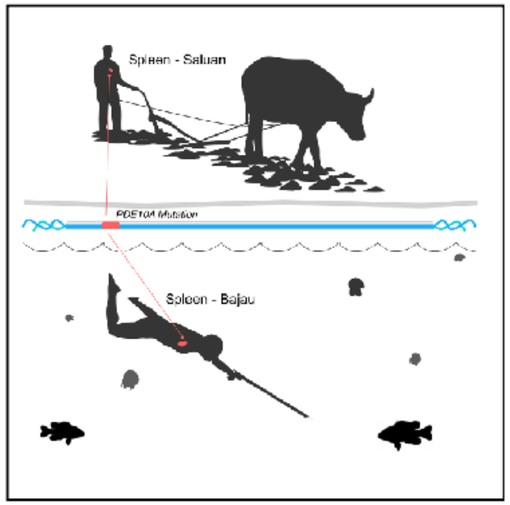

11.
AYKUT KENCE
EVRİM
KONFERANSI
18-19 ŞUBAT 2017
WWW.ODTUAKEK.ORG
Dr. Tuğçe Bilgin Sonay
Sunum Konusu: DNA'nin en oynak elemanı: ardışık kısa tekrar dizileri ve fenotipik evrim
Yeditepe Üniversitesi Moleküler Biyoloji ve Genetik bölümü mezunun olan Dr. Tuğçe Bilgin Sonay, Zürich Üniversitesi'nde Evrimsel Biyoloji üzerine doktorasını bitirdi, yine aynı yerde araştırma görevlisi olarak çalışıyor. Evrim dışında psikoloji ve ikisinin kesişim noktaları olan bilincin ve empatinin ortaya çıkışı gibi konulara ilgi duyuyor.
ODTÜ KKM - Kemal Kurdaş Konferans Salonu

Article

Cell

Physiological and Genetic Adaptations to Diving in Sea Nomads

Graphical Abstract



Authors

Melissa A. Ilardo, Ida Moltke, Thorfinn S. Korneliussen, ..., Suhartini Salingkat, Rasmus Nielsen, Eske Willerslev

Correspondence

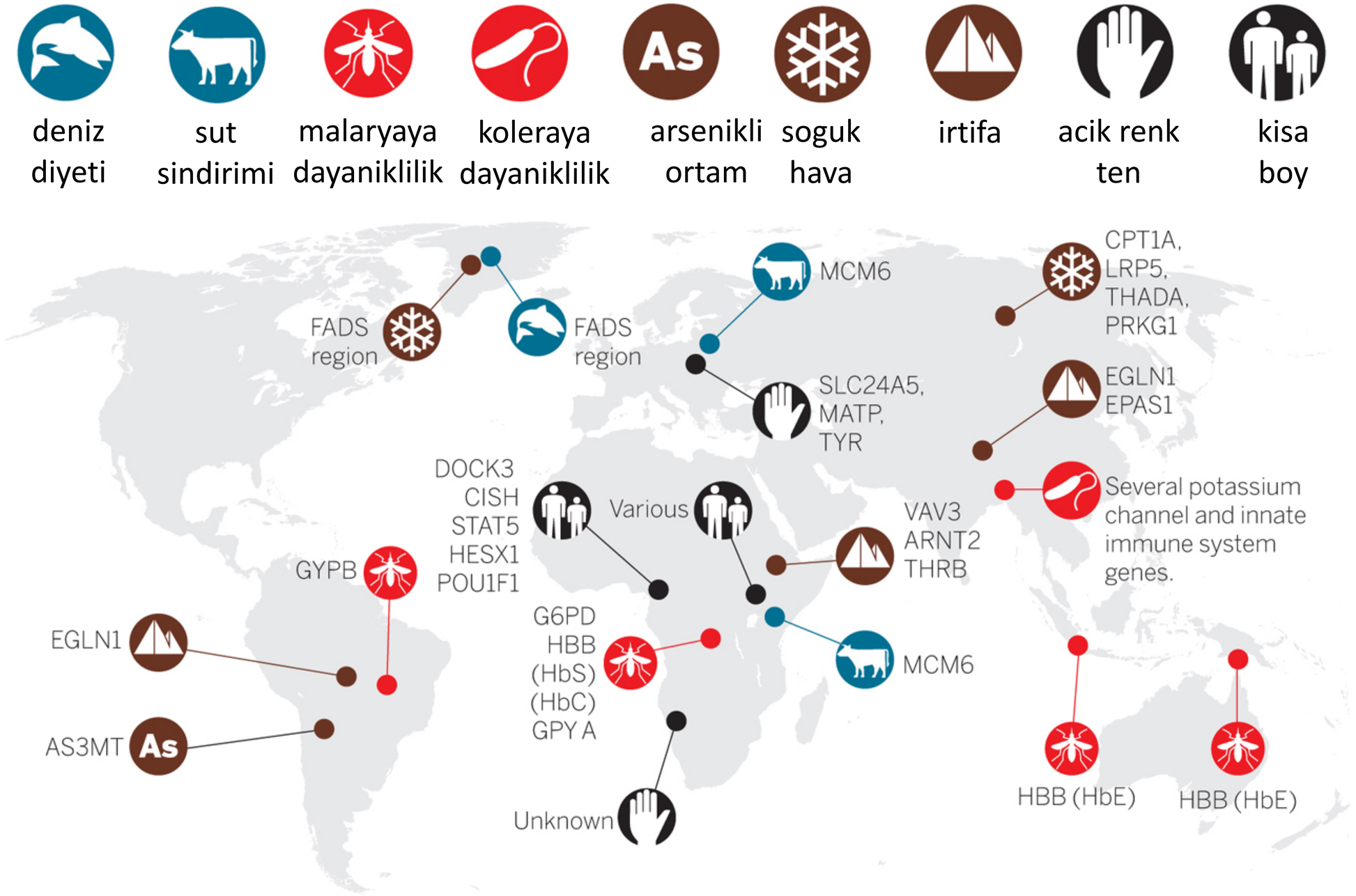
rasmus_nielsen@berkeley.edu (R.N.), ewillerslev@snm.ku.dk (E.W.)

In Brief

Genetic and physiological adaptations enable the remarkable breath-holding ability of marine nomads.

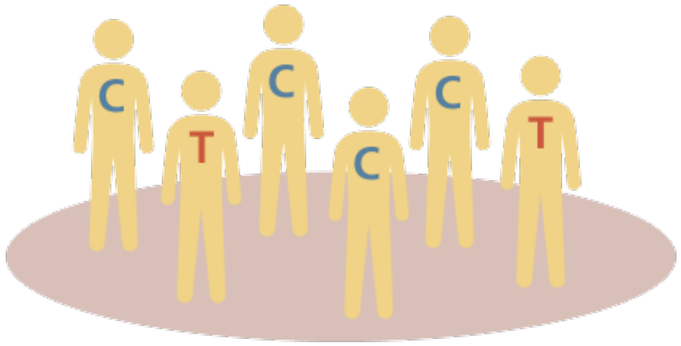
Highlights

- The Bajau, or "Sea Nomads," have engaged in breath-hold diving for thousands of years
- Selection has increased Bajau spleen size, providing an oxygen reservoir for diving

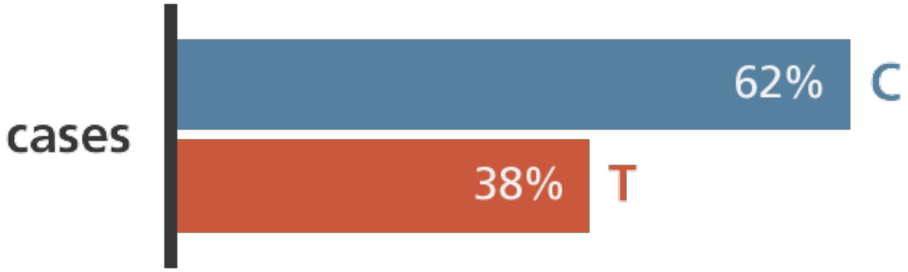


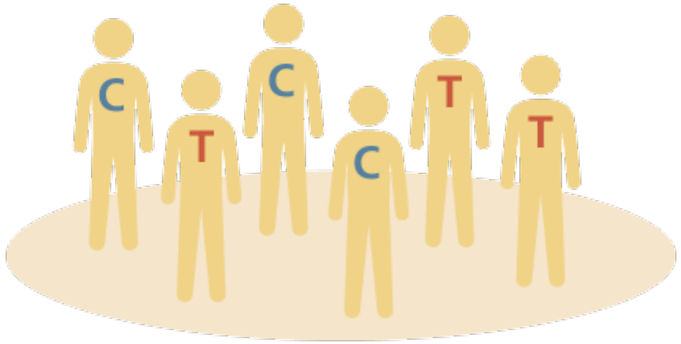
Tishkoff, Science, 2015

Genome Wide Association Studies



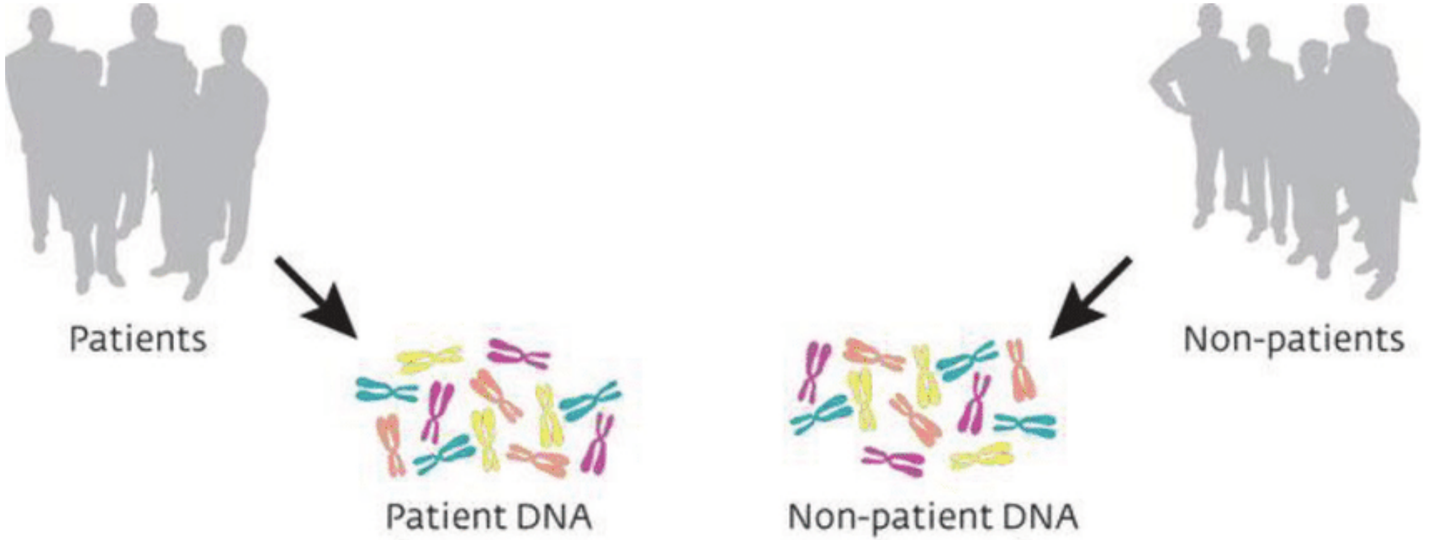
cases (n=1,000)
people with heart disease

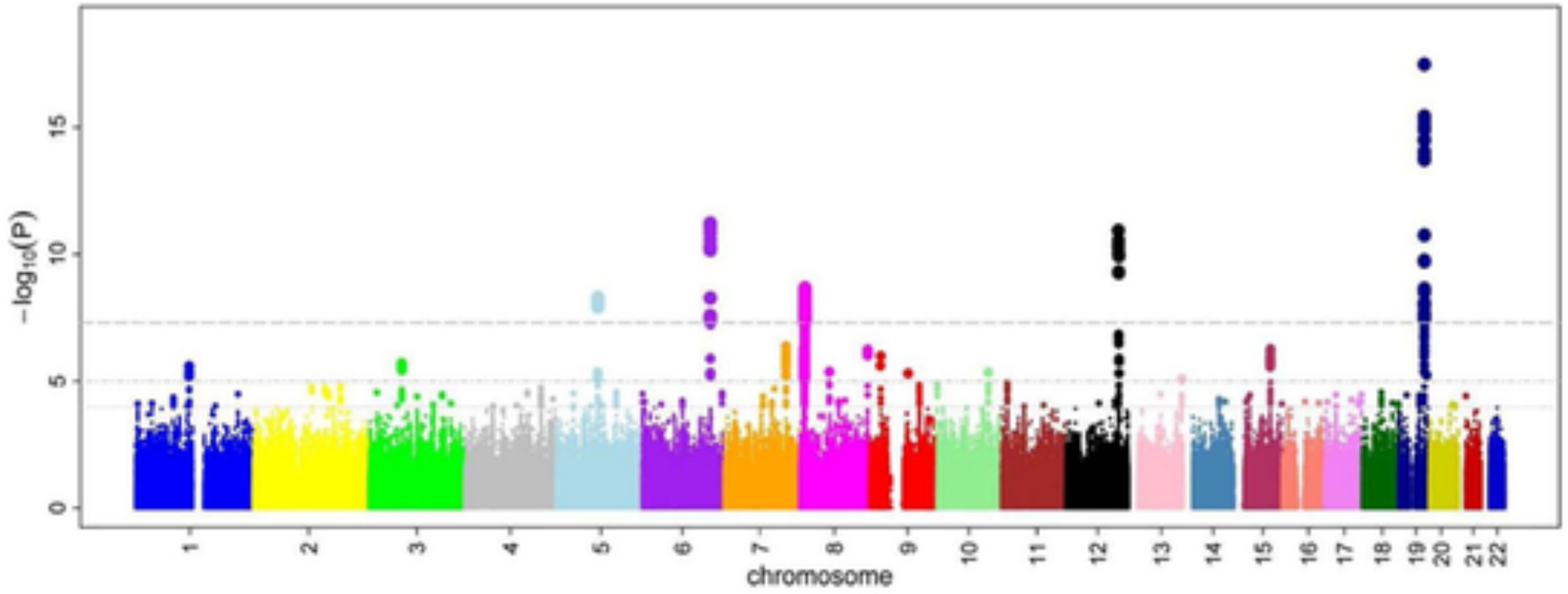




controls (n=1,000)
people without heart disease

Genome Wide Association Studies



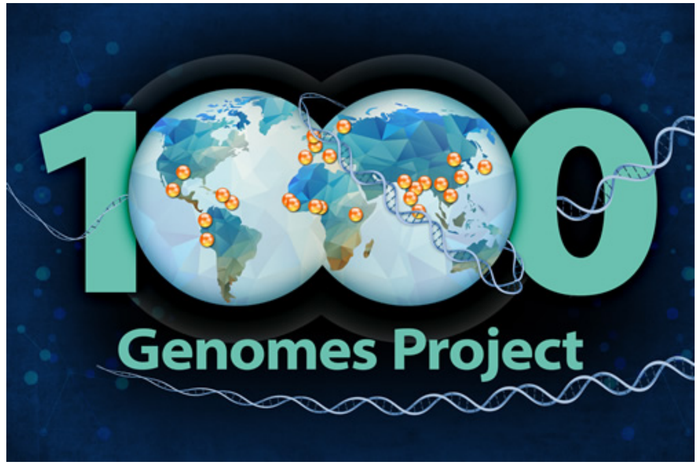


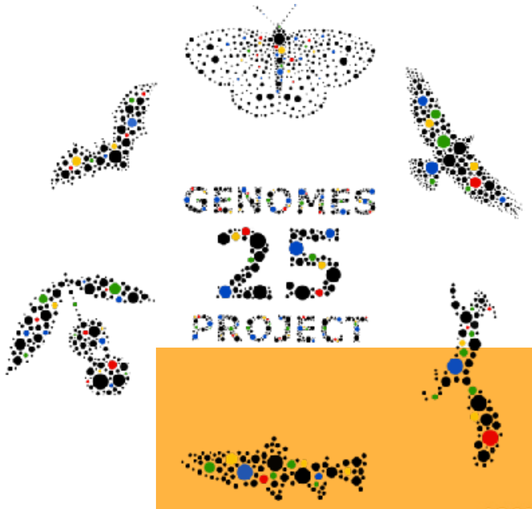
Genome Wide Association Studies

colaus
psycolaus

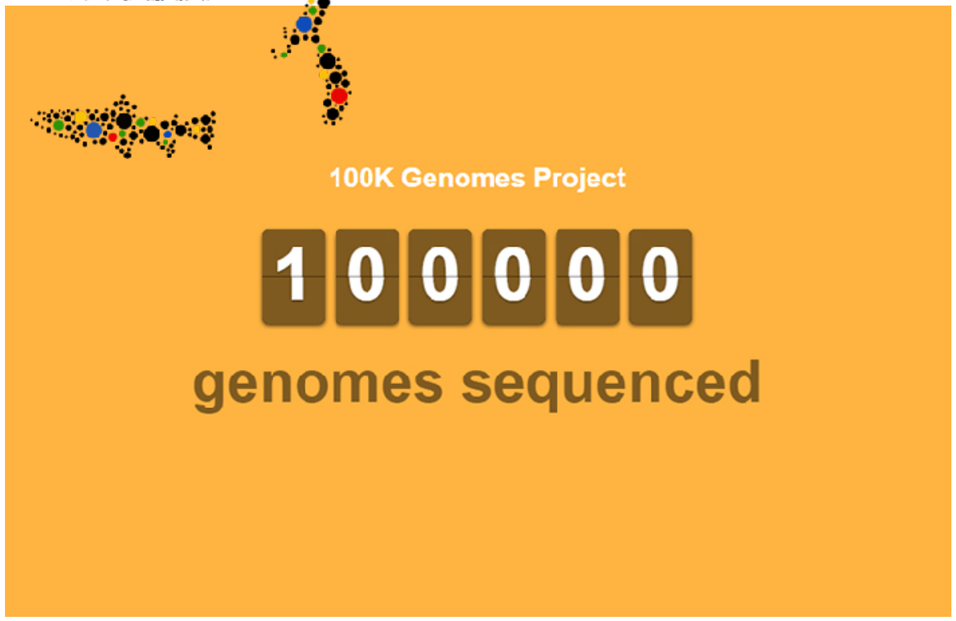
From 6734 people:
DNA sequence
Diseases, BMI, habits, ethnicity

Genome Projects





NIH HUMAN MICROBIOME PROJECT



Genomics

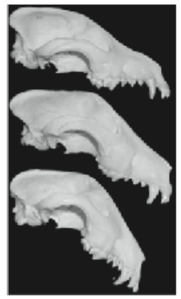
Biological Diversity and Conservation

GCMP
Global Coral Microbiome Project

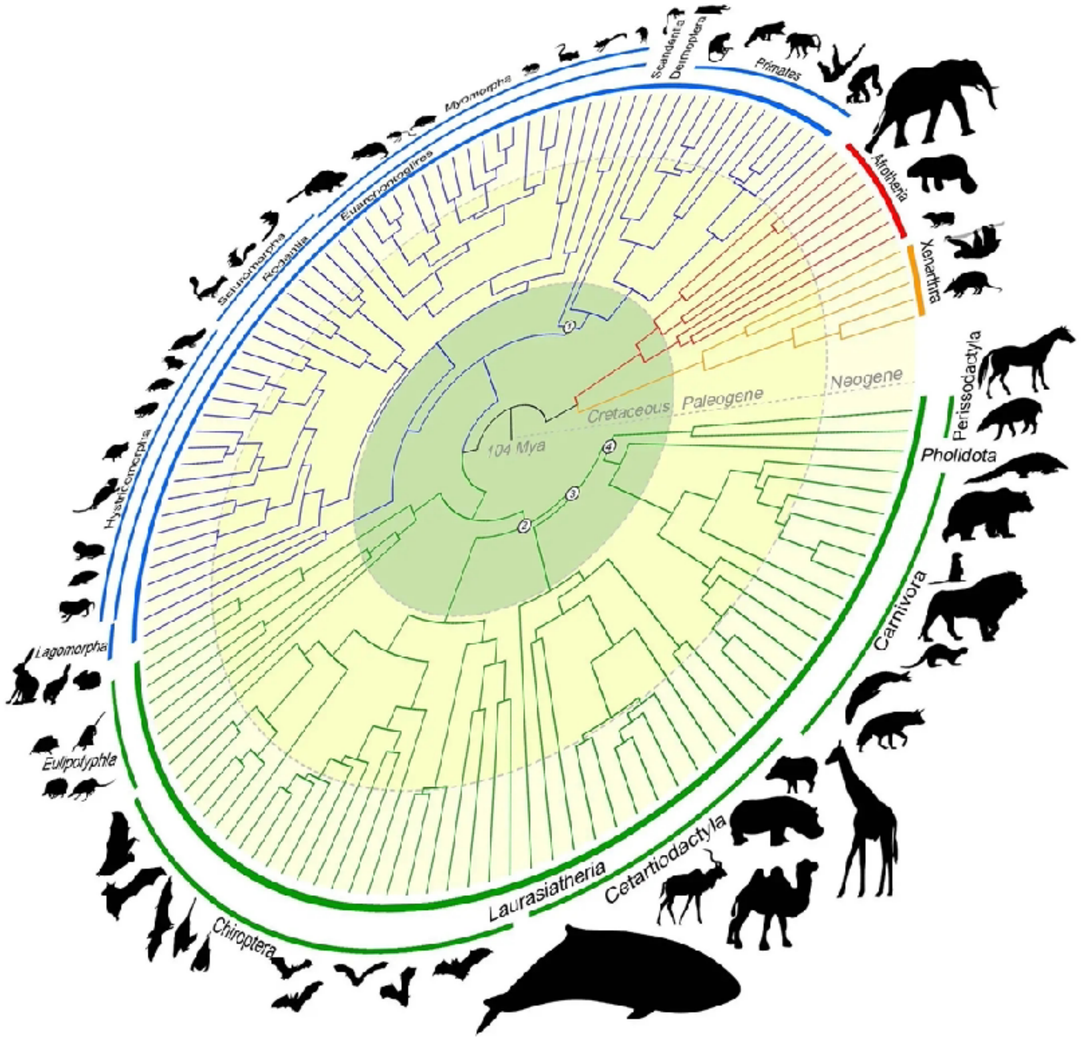
GENOMES
2.5
PROJECT

VGP
VERTEBRATE
GENOMES
PROJECT
DIGITAL NOAH'S ARK GENOME LIBRARY

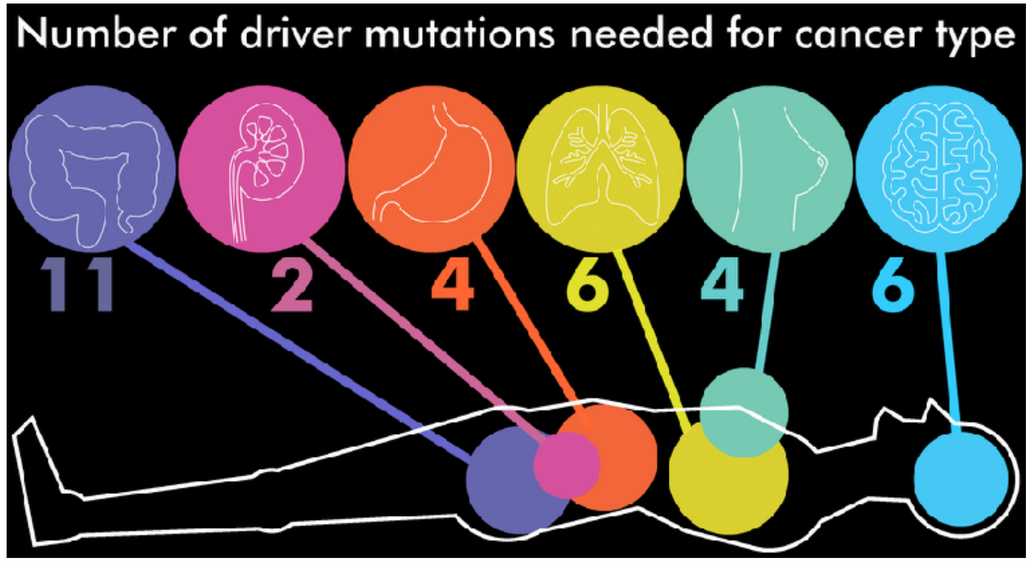



Phylogenetic Trees

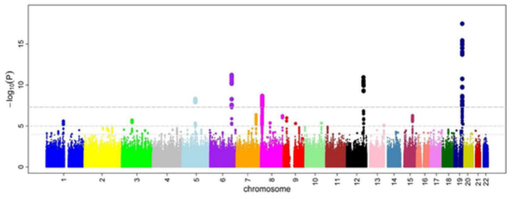


Most mutations are neutral



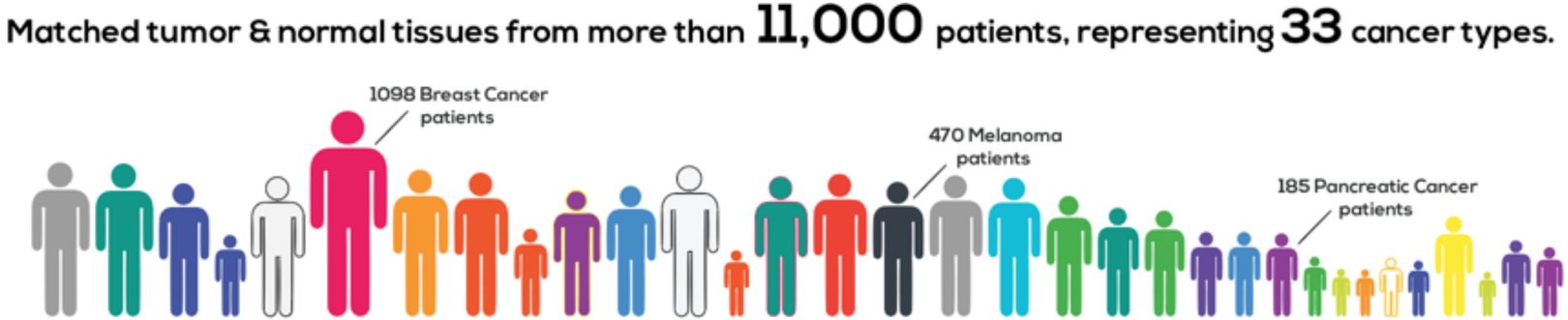
Which mutations?

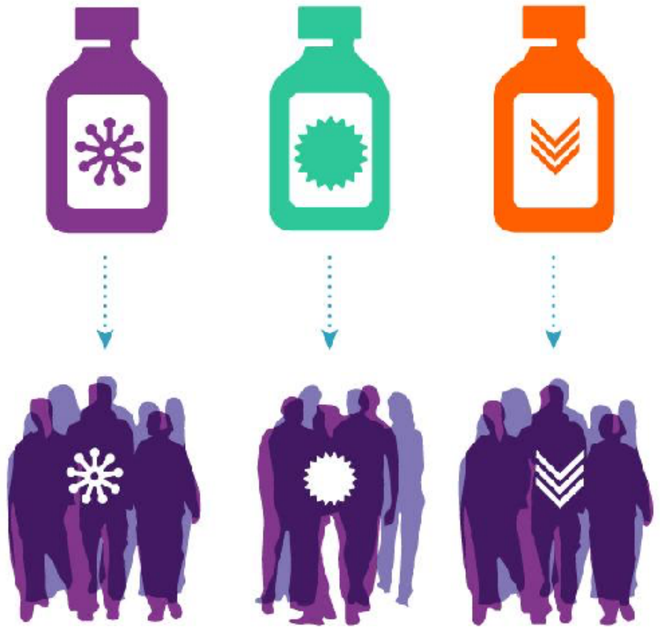
- In silico candidates - statistical prediction
- In vitro mutagenesis - experimentally mutate

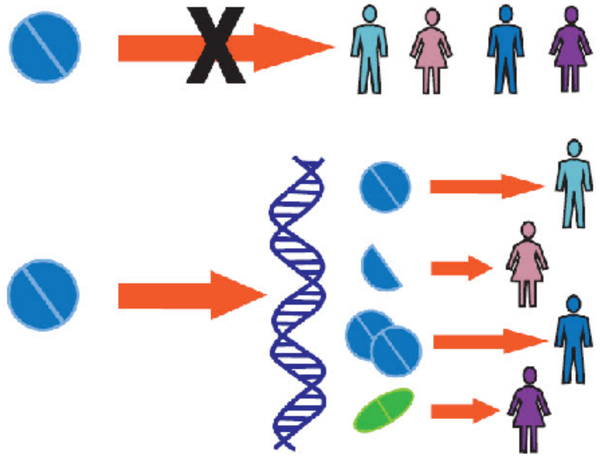


Genomics

Precision Medicine

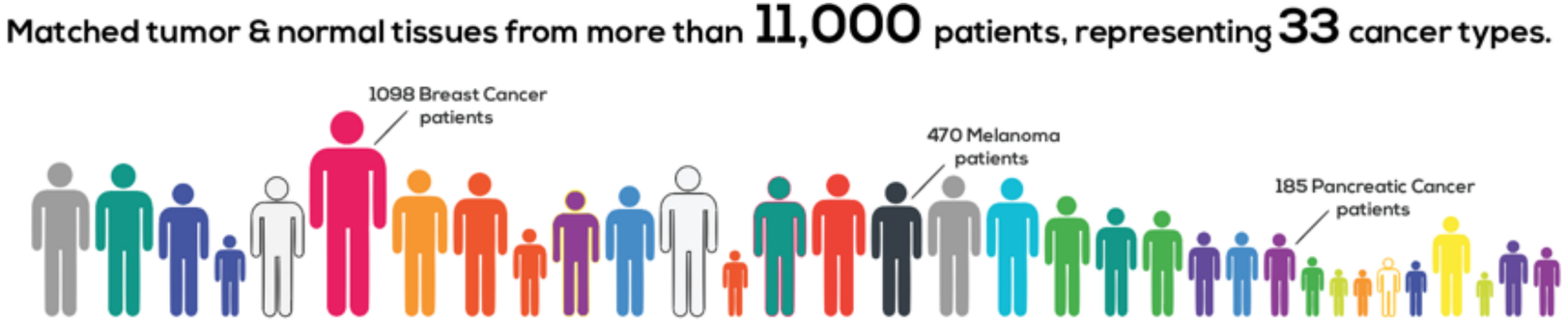


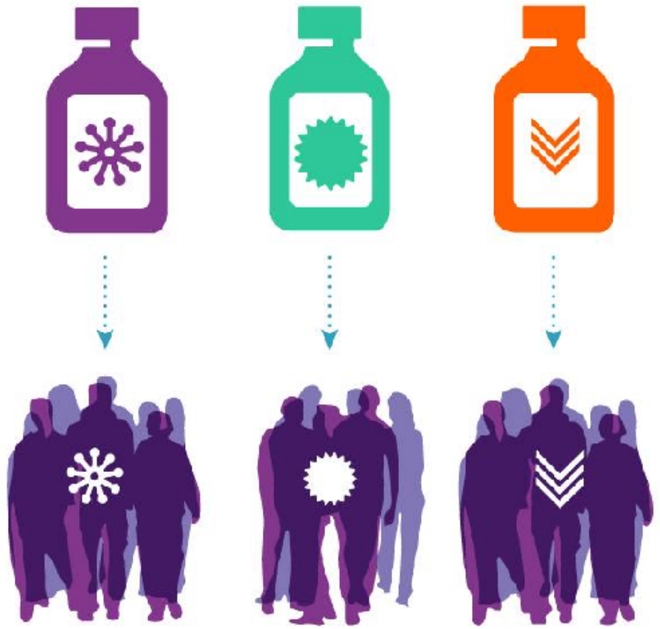


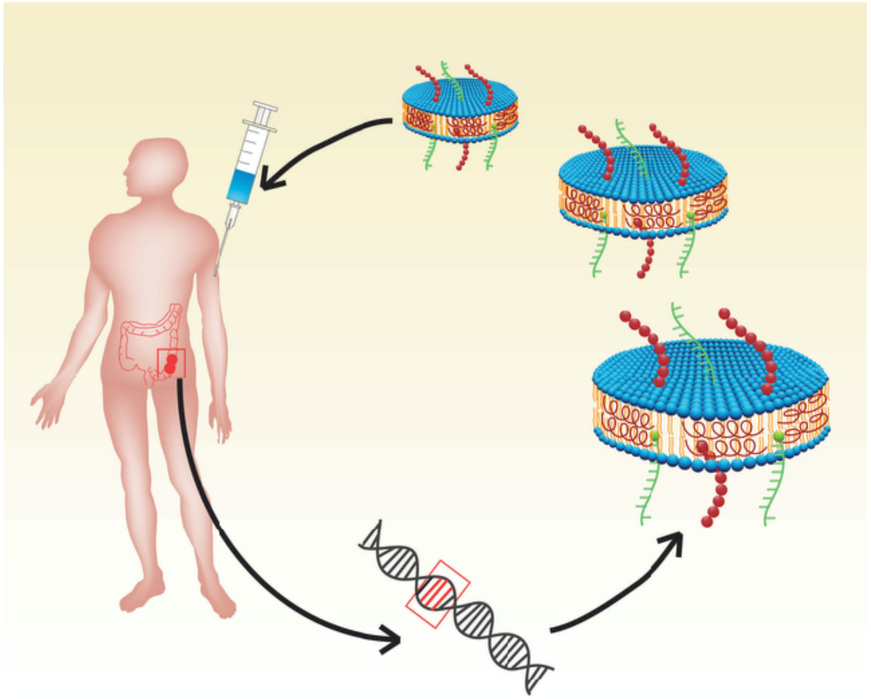


Genomics

Precision Medicine



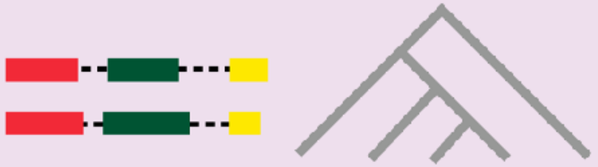




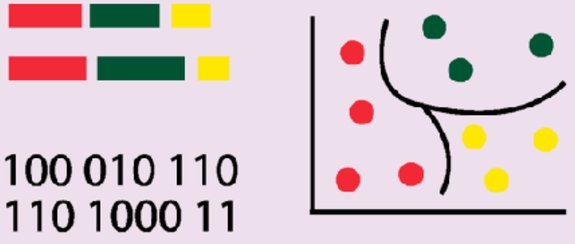
Machine and deep learning integration with bioinformatics

Molecular evolution

Phylogenetic inference



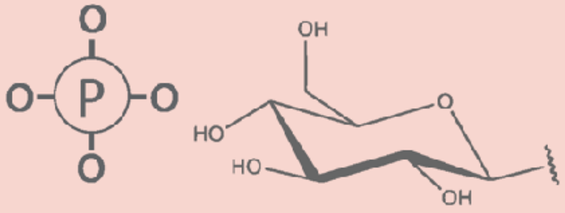
Alignment-free sequence classification



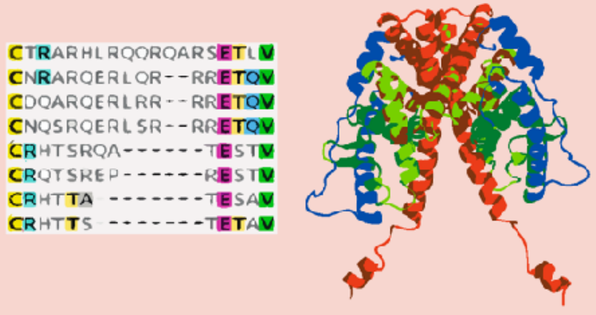
- *Inference of tree topology*
- *Sequence classification*
- *Viral sequence identification*
- *functional annotation*

Protein structure Analysis

Post translational modification



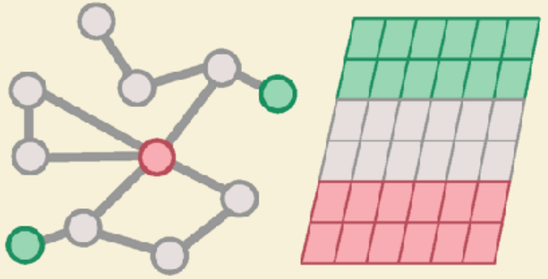
Folding and structure



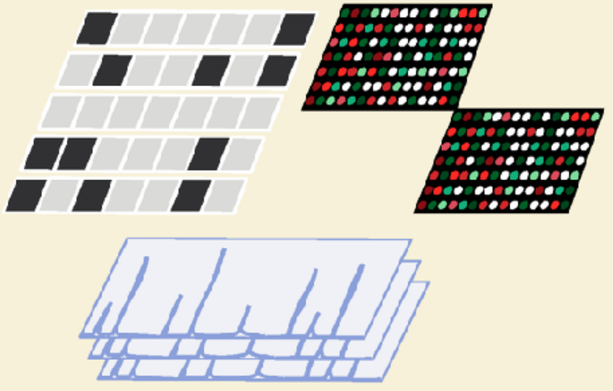
- *Phosphorylation site prediction*
- *Protein glycosylation prediction*
- *Protein contact maps prediction*
- *Structural homology prediction*

Systems biology

Biological Networks



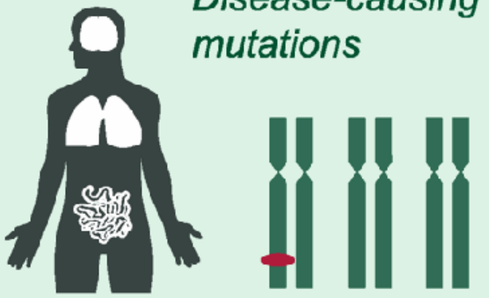
Multi-Omics integration



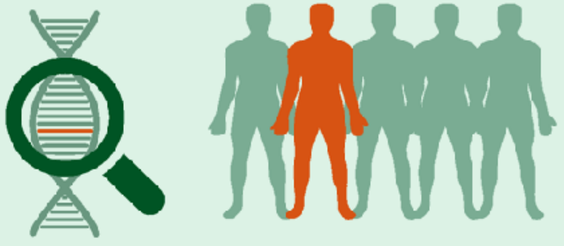
- *Biological networks construction*
- *Biological interactions prediction*
- *Pathway dynamics prediction*
- *Platform integration frameworks*

Genomics for Disease Research

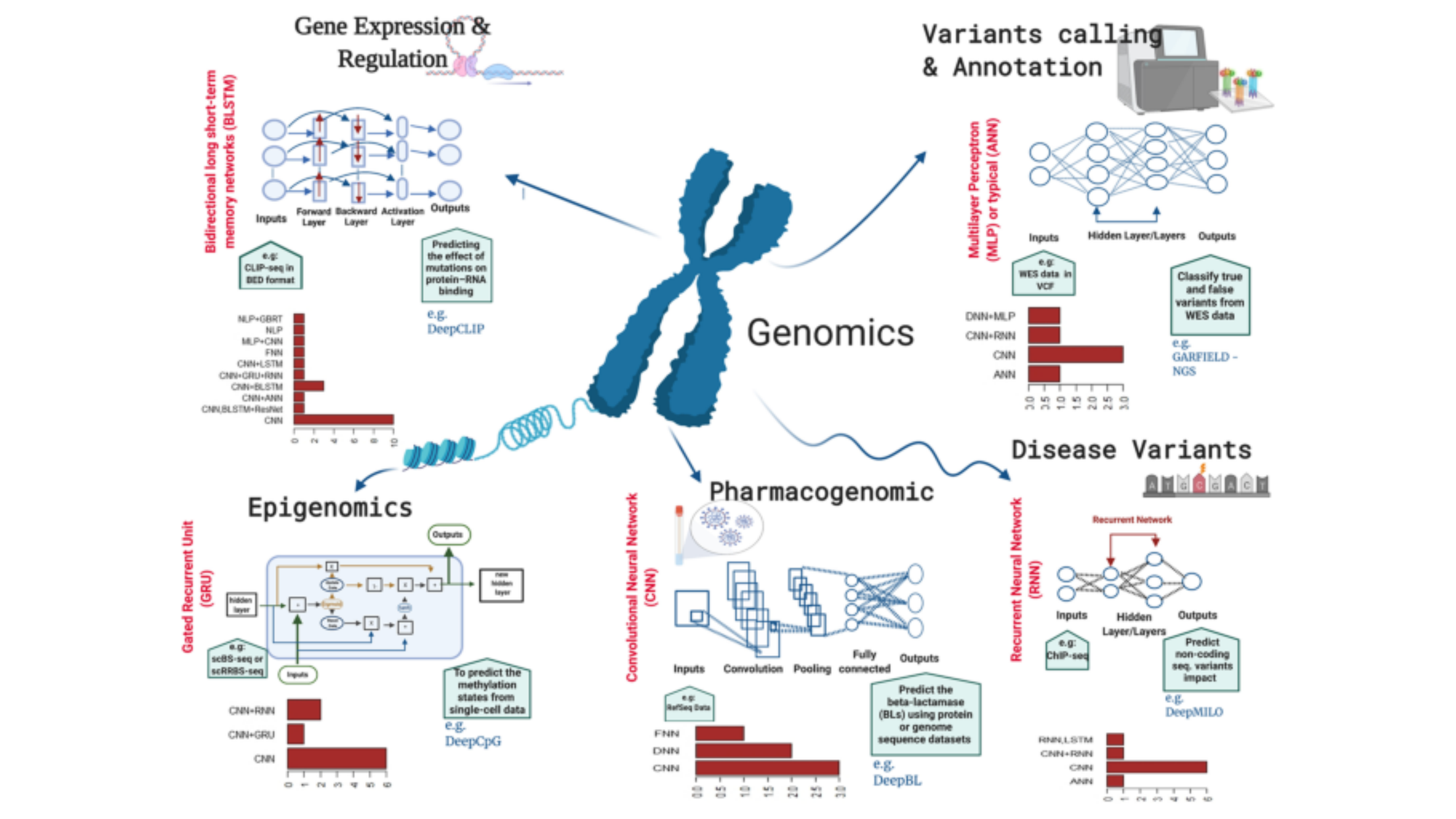
Disease-causing mutations



Biomarkers discovery



- *Diesease associated genes and mutations*
- *Biomarkers*
- *Precision medicine applications*


Gene Expression & Regulation
Bidirectional long short-term memory networks (BLSTM)
Inputs
Forward Layer
Backward Layer
Activation Layer
Outputs
e.g: CLIP-seq in BED format
Predicting the effect of mutations on protein–RNA binding
e.g. DeepCLIP
NLP+GBRT
NLP
MLP+CNN
FNN
CNN+LSTM
CNN+GRU+RNN
CNN+BLSTM
CNN+ANN
CNN,BLSTM+ResNet
CNN
Variants calling & Annotation
Multilayer Perceptron (MLP) or typical (ANN)
Inputs
Hidden Layer/Layers
Outputs
e.g: WES data in VCF
Classify true and false variants from WES data
e.g. GARFIELD - NGS
DNN+MLP
CNN+RNN
CNN
ANN
Genomics
Epigenomics
Gated Recurrent Unit (GRU)
Outputs
hidden layer
new hidden layer
e.g: scBS-seq or scRRBS-seq
Inputs
To predict the methylation states from single-cell data
e.g. DeepCpG
CNN+RNN
CNN+GRU
CNN
Pharmacogenomic
Convolutional Neural Network (CNN)
Inputs
Convolution
Pooling
Fully connected
Outputs
e.g: RefSeq Data
Predict the beta-lactamase (BLs) using protein or genome sequence datasets
e.g. DeepBL
FNN
DNN
CNN
Disease Variants
Recurrent Neural Network (RNN)
Recurrent Network
Inputs
Hidden Layer/Layers
Outputs
e.g: ChIP-seq
Predict non-coding seq. variants impact
e.g. DeepMILO
RNN,LSTM
CNN+RNN
CNN
ANN

Questions?

Introduction to Bioinformatics: An Overview

Tugce Bilgin Sonay, ZHAW, May 2023

Bioinformatics for Beginners

Mutations

Tugce Bilgin Sonay, ZHAW, May 2023

Mutations:

- **Single Nucleotide Polymorphisms – SNPs**

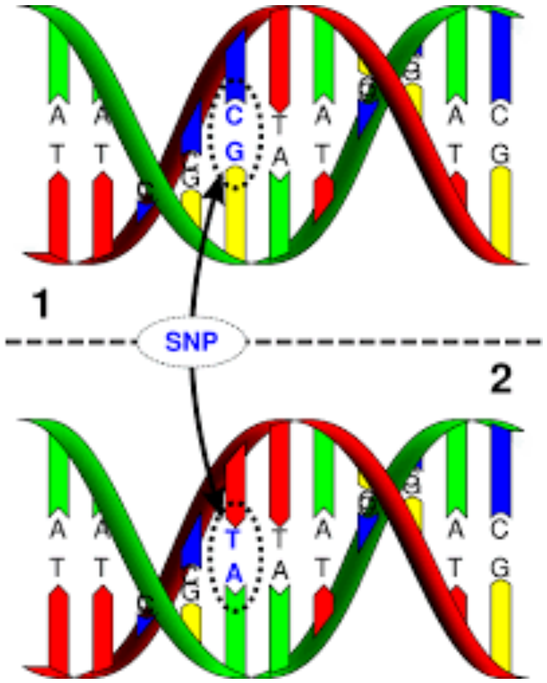
Mutation in 10-100 million bases per generation per cell

- **Indels**

Insertions and deletions of sequences, less frequent

- **Copy Number Mutations**

Long or short sequences



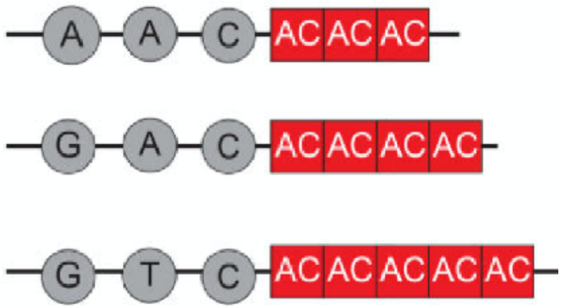
Short Tandem Repeats:

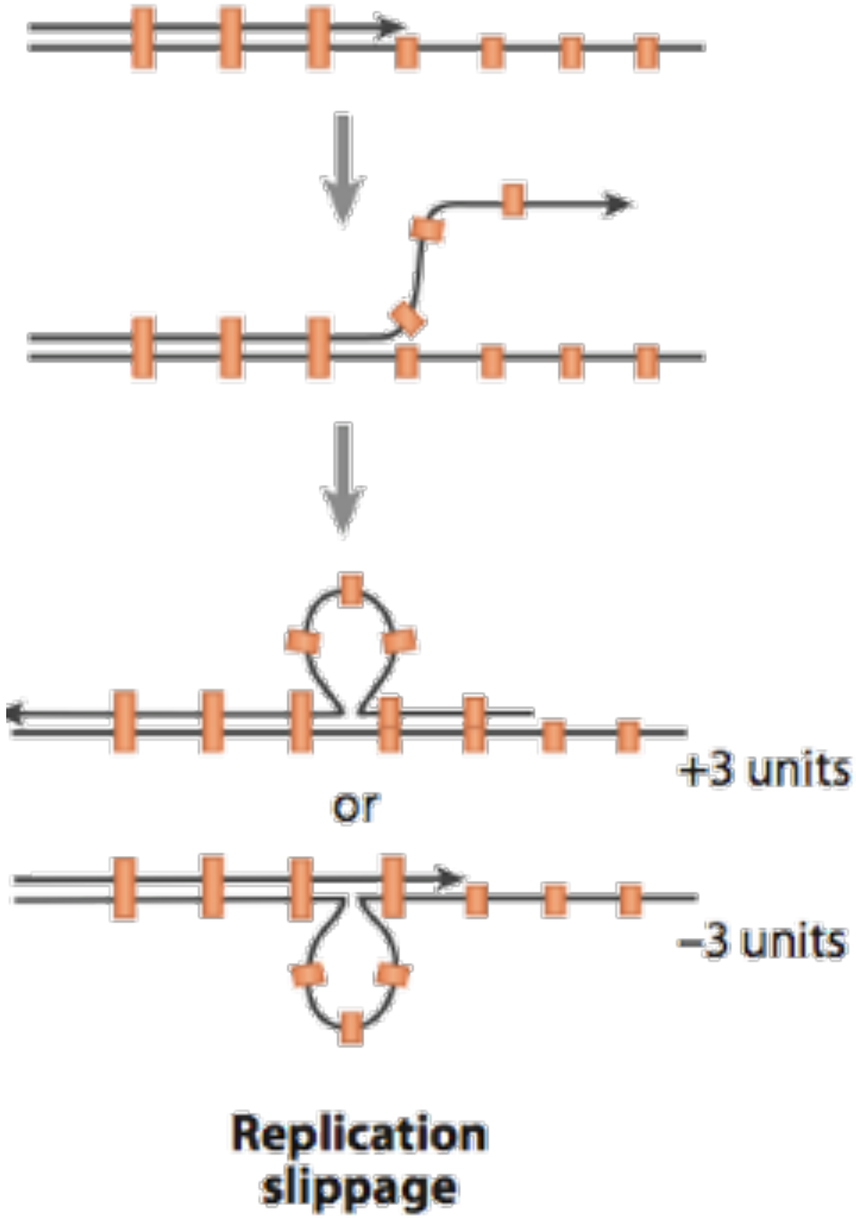
-Tandem repeat instability
20 % of functional variants

…ATTTAATTTAATTTAATTTA…

extremely high mutation rates: 10^{-2}-10^{-5} per cell division

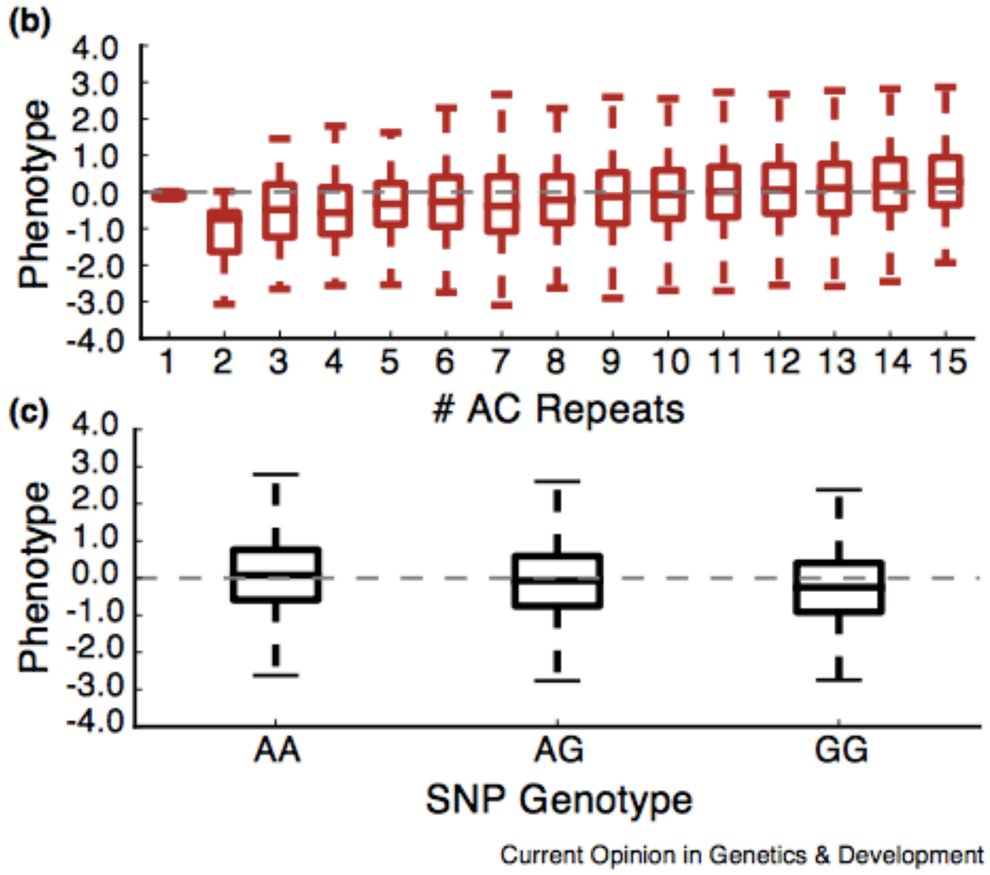
Microsatellites: Short Tandem Repeats in DNA



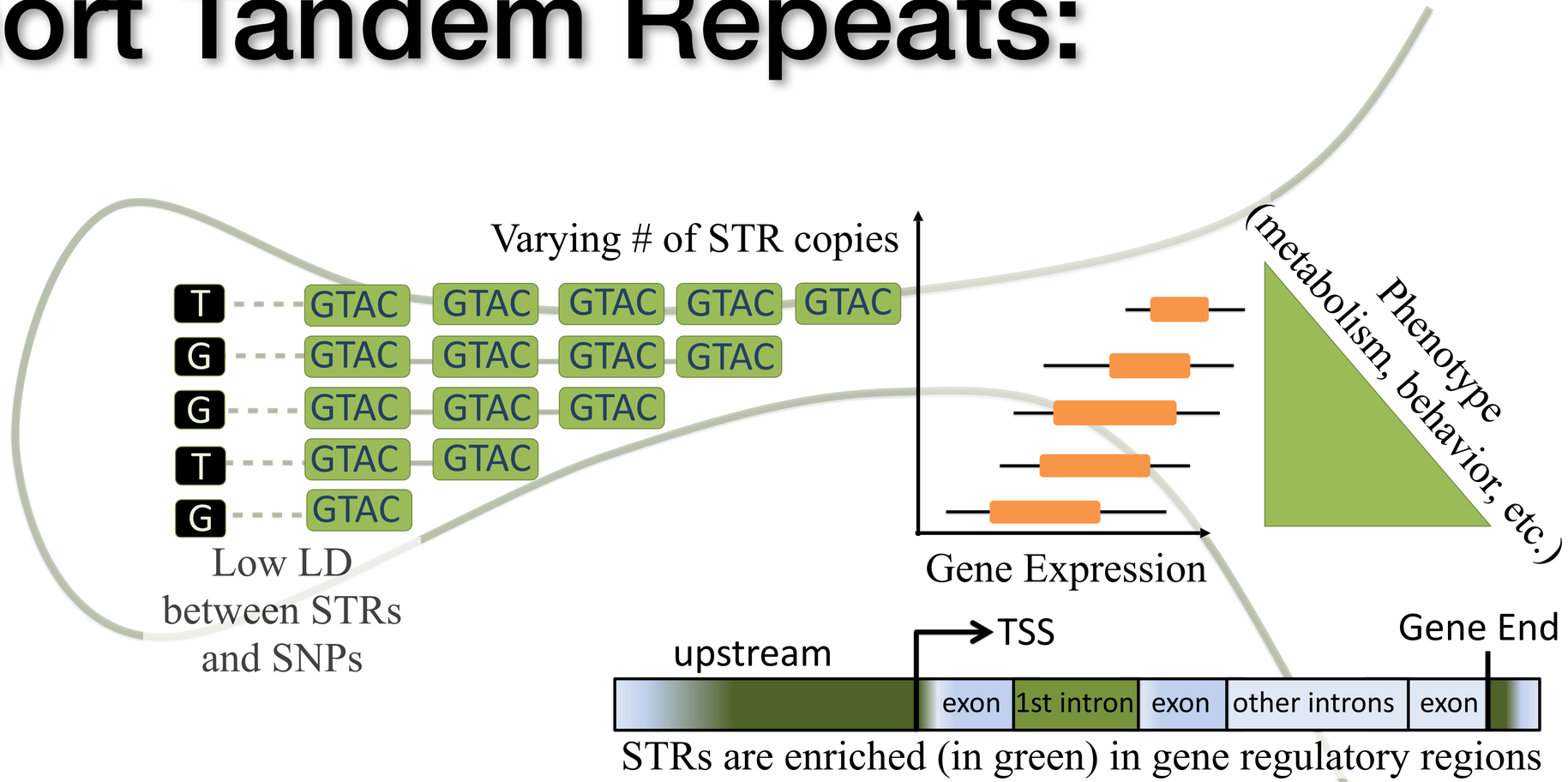


Gamayel *et al., Annu. Rev. Genet.*, 2010

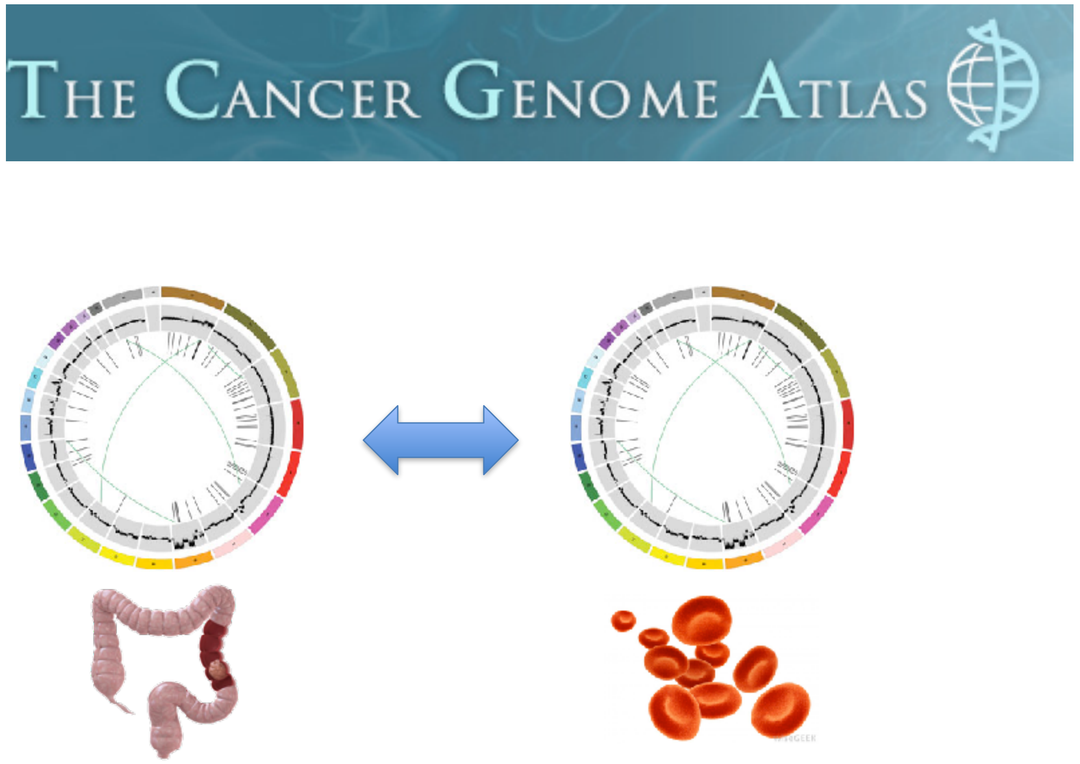
Short Tandem Repeats offer much more genotypic variation than SNPs



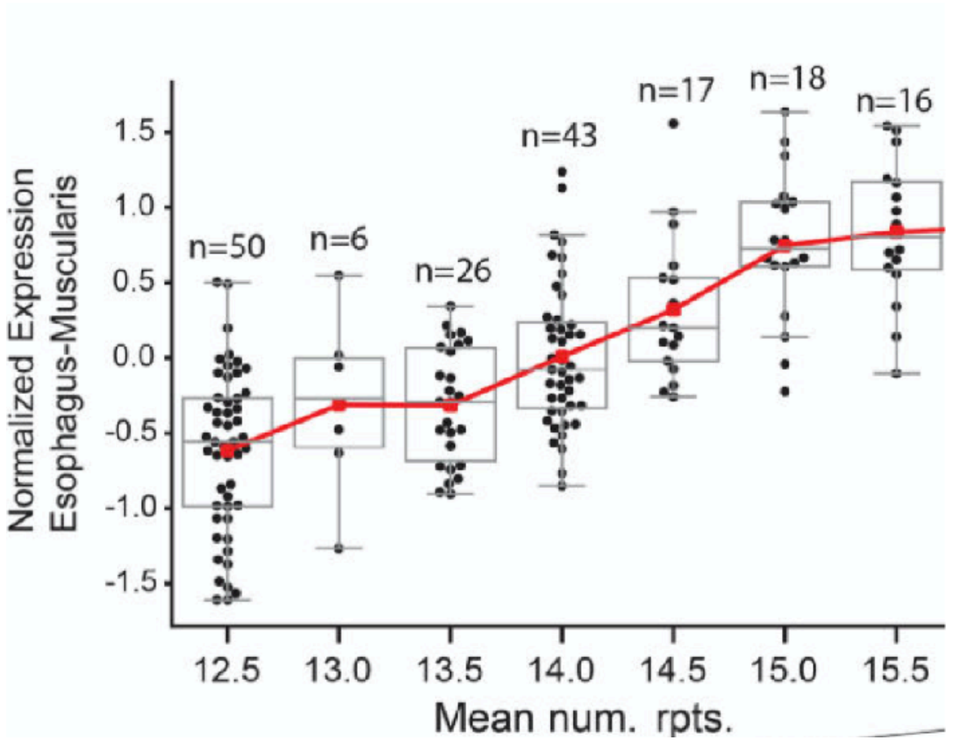
Short Tandem Repeats:



Short Tandem Repeats and Cancer:



Bilgin Sonay et al., BMC Genomics, 2015



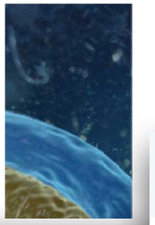
Fosting et al., Nat Genet, 2020

Short Tandem Repeats and Cancer:

NIH NATIONAL CANCER INSTITUTE

Developing the First Precision Immunotherapy

In May 2017, the Food and Drug Administration (FDA) approved the first drug to treat tumors based on their genetic characteristics, regardless of where in the body the cancer originated. Until now, drugs have been approved based on their cell or tissue of origin, such as the breast or the lung. But pembrolizumab (Keytruda®) doesn't target the genetic abnormalities of cancer cells specifically—it targets the immune system.



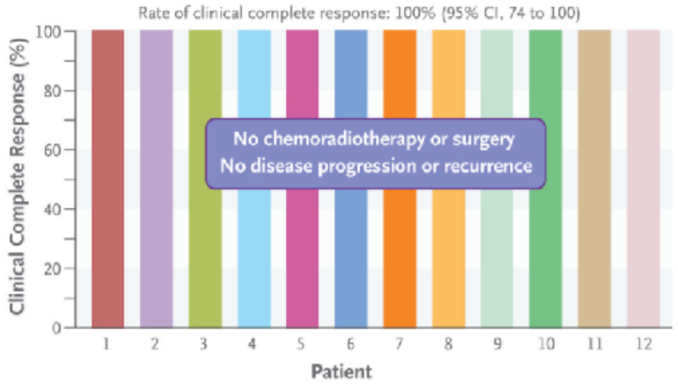
The NEW ENGLAND JOURNAL of MEDICINE

RESEARCH SUMMARY

PD-1 Blockade in Mismatch Repair–Deficient, Locally Advanced Rectal Cancer

Cercek A et al. DOI: 10.1056/NEJMoa2201445

Overall Response to Dostarlimab in 12 Patients



nature COMMUNICATIONS

Article | OPEN | Published: 06 June 2017

A molecular portrait of microsatellite instability across multiple cancers

nature biotechnology

Article | Published: 11 September 2017

Analysis of somatic microsatellite indels identifies driver events in human tumors

Cell

The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes

nature medicine

Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer

Science

Cite as: D. T. Le *et al.*, *Science* 10.1126/science.aan6733 (2017).

Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade

Short Tandem Repeats and Cancer:

Maria Anisimova

Intrinsic Disorder as a potential link between tandem repeats and cancer

Verbiest, Delucchi, Bilgin Sonay and Anisimova, 2021

Mutation and selection processes regulating short tandem repeats

Verbiest, …, Anisimova, Gymrek, Bilgin Sonay, invited review in JEB, 2022

Questions?

Mutations

Tugce Bilgin Sonay, ZHAW, May 2023

Bioinformatics for Beginners

Bioinformatics Resources

Tugce Bilgin Sonay, ZHAW, May 2023

Protein Databases

- Protein-protein interaction: https://string-db.org/
- Protein databases: https://www.uniprot.org/ https://www.rcsb.org/

Bioinformatics Forums

- Best friend of a bioinformatician: https://www.biostars.org
- Another best friend: https://stackoverflow.com

Notebooks

- Jupyter Notebook, bioinformatician's diary: https://jupyter.org

Downstream Genomic

- Bedtools : https://bedtools.readthedocs.io/en/latest/

Omics Databases

- USCS Table Browser: https://genome.ucsc.edu/cgi-bin/hgTables
- Biomart Ensembl: https://useast.ensembl.org/index.html
- ncbi entrez: https://www.ncbi.nlm.nih.gov/

Cancer Databases

- Pan Cancer, TCGA: https://cancergenome.nih.gov

Gene Databases

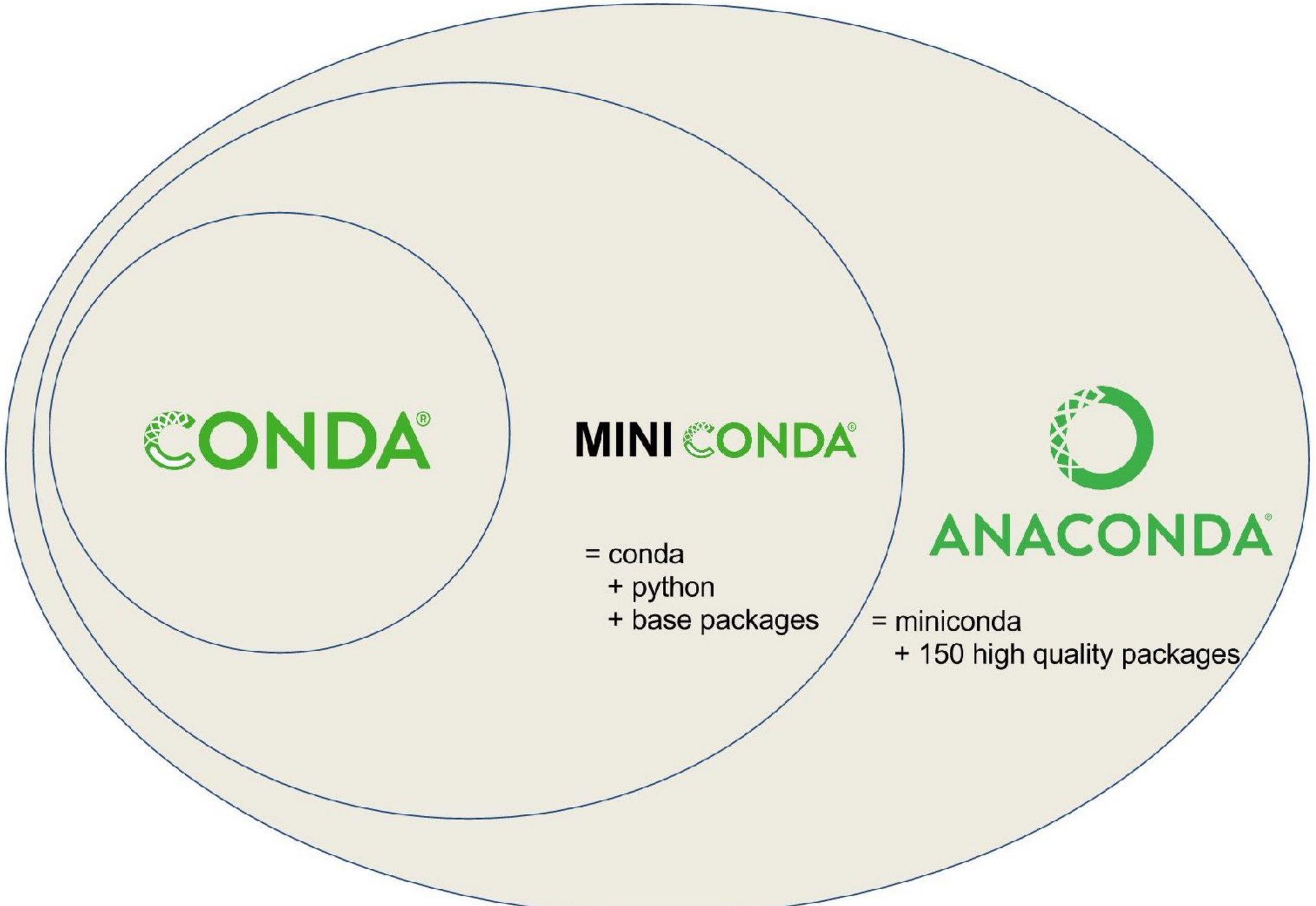
- Gene Cards: https://www.genecards.org
- DAVID for Gene Ontology terms: https://david.ncifcrf.gov/summary.jsp

Genome Databases

1000 genomes new name : https://www.internationalgenome.org

Coding

- R Circos - https://cran.r-project.org/web/packages/BioCircos/vignettes/BioCi
- R packages: https://www.bioconductor.org
- python packages: Biopython


CONDA
MINICONDA
= conda
+ python
+ base packages
ANACONDA
= miniconda
+ 150 high quality packages

Questions?

Bioinformatics Resources

Tugce Bilgin Sonay, ZHAW, May 2023