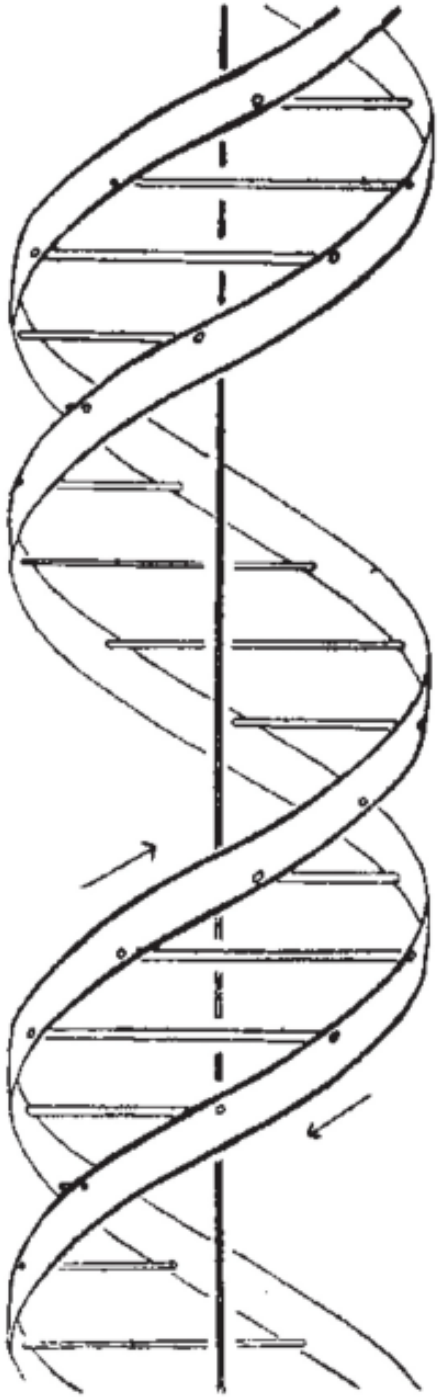
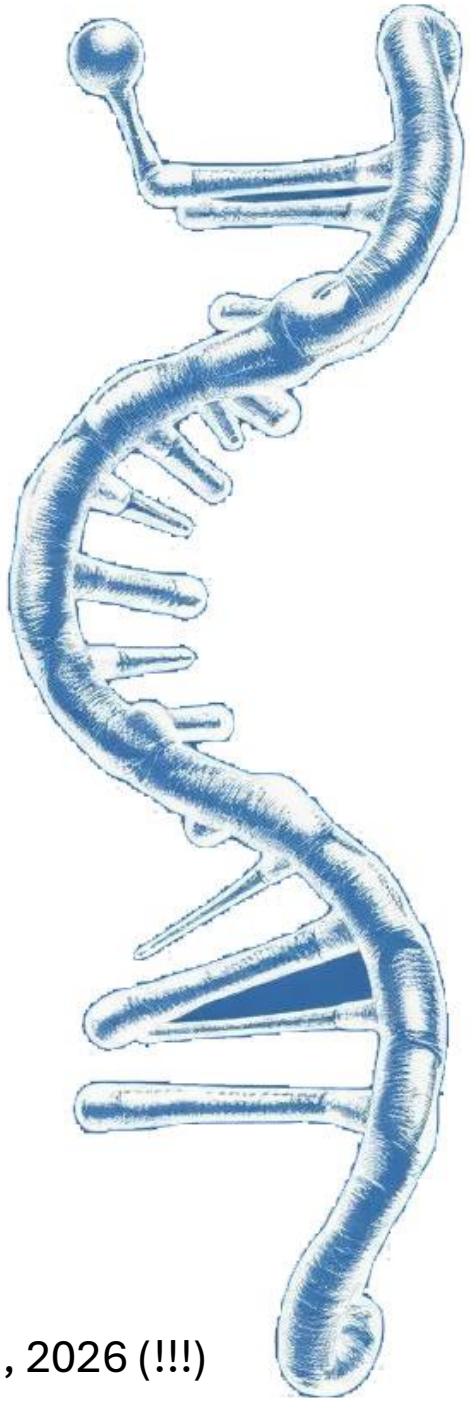


AI and Genetic Part 1

Jean Martin Beaulieu

January 29, 2026

Watson and Crick, Nature 1953



GPT 5.0 , 2026 (!!!)

Article

Advancing regulatory variant effect prediction with AlphaGenome

https://deepmind.google.com/science/alphagenome/

https://doi.org/10.1038/s41586-025-10014-0

Received: 16 May 2025

Accepted: 4 December 2025

Published online: 28 January 2026

Open access

Check for updates

Žiga Avsec[1,2]✉, Natasha Latysheva[1,2], Jun Cheng[1,2], Guido Novati[1,2], Kyle R. Taylor[1,2], Tom Ward[1,2], Clare Bycroft[1,2], Lauren Nicolaisen[1,2], Eirini Arvaniti[1,2], Joshua Pan[1,2], Raina Thomas[1], Vincent Dutordoir[1], Matteo Perino[1], Soham De[1], Alexander Karollus[1], Adam Gayoso[1], Toby Sargeant[1], Anne Mottram[1], Lai Hong Wong[1], Pavol Drotár[1], Adam Kosiorek[1], Andrew Senior[1], Richard Tanburn[1], Taylor Applebaum[1], Souradeep Basu[1], Demis Hassabis[1] & Pushmeet Kohli[1]✉

Deep learning models that predict functional genomic measurements from DNA sequences are powerful tools for deciphering the genetic regulatory code. Existing methods involve a trade-off between input sequence length and prediction resolution, thereby limiting their modality scope and performance[1–5]. We present AlphaGenome, a unified DNA sequence model, which takes as input 1 Mb of DNA sequence and predicts thousands of functional genomic tracks up to single-base-pair resolution across diverse modalities. The modalities include gene expression, transcription initiation, chromatin accessibility, histone modifications, transcription factor binding, chromatin contact maps, splice site usage and splice junction coordinates and strength. Trained on human and mouse genomes, AlphaGenome matches or exceeds the strongest available external models in 25 of 26 evaluations of variant effect prediction. The ability of AlphaGenome to simultaneously score variant effects across all modalities accurately recapitulates the mechanisms of clinically relevant variants near the *TAL1* oncogene[6]. To facilitate broader use, we provide tools for making genome track and variant effect predictions from sequence.

Data availability

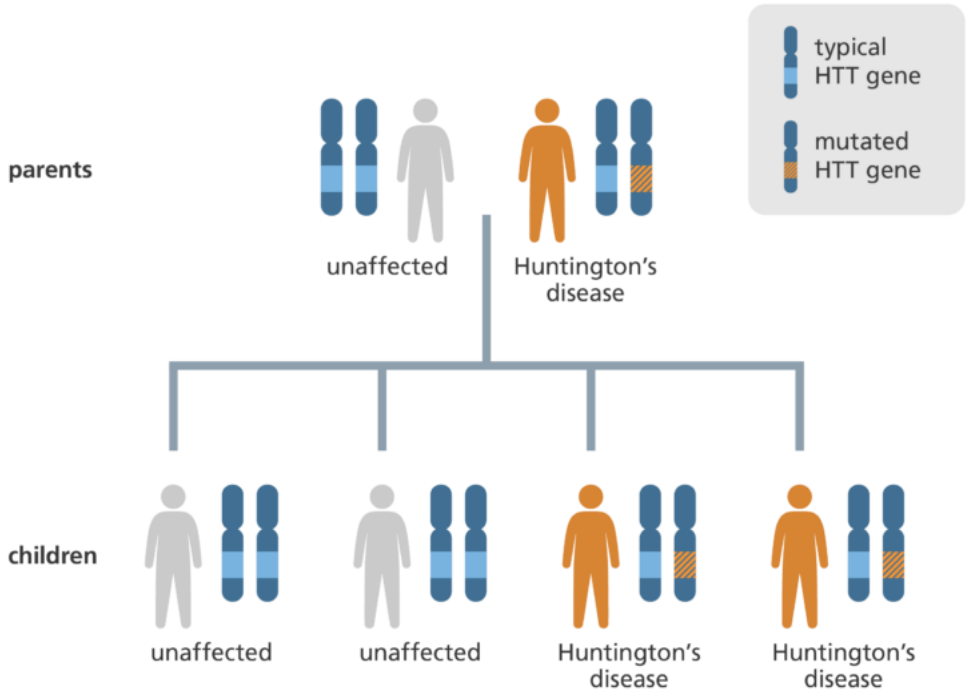
All primary experimental datasets used for the training and evaluation of AlphaGenome in this study were obtained from publicly accessible sources. A comprehensive manifest detailing these data sources, including specific repositories (such as ENCODE portal, GTEx portal, 4D Nucleome portal, ClinVar and gnomAD), individual accession numbers, relevant version information and direct URLs where applicable, is provided in Supplementary Table 2. This study did not generate new primary experimental data requiring deposition.

Code availability

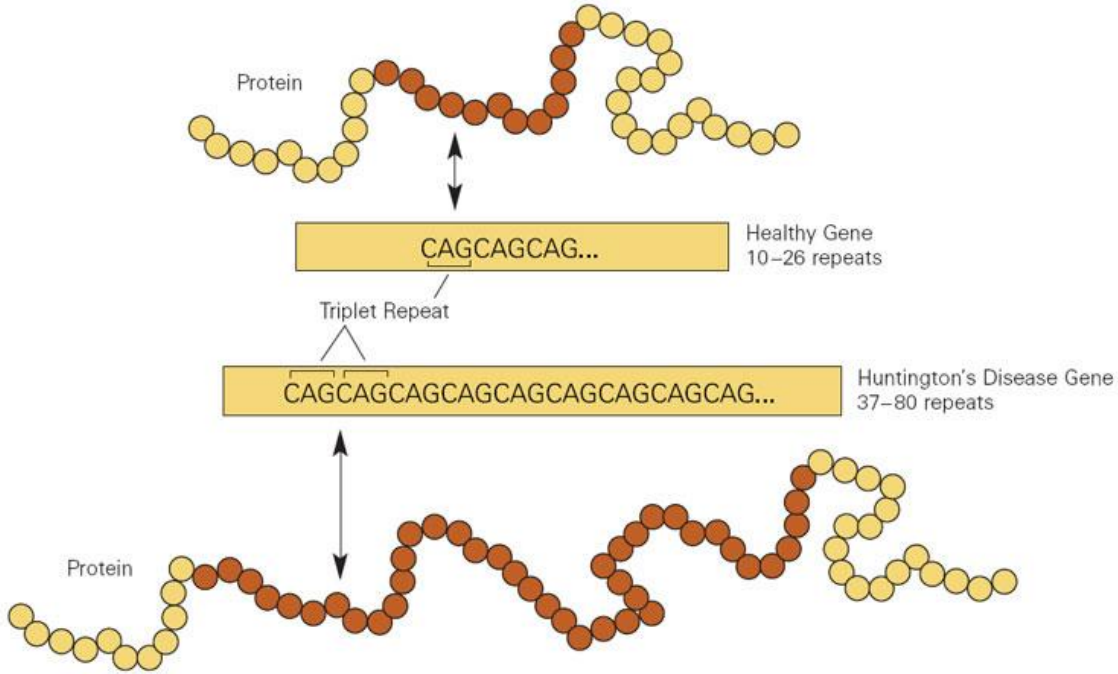
AlphaGenome is available for non-commercial use through an online API at http://deepmind.google.com/science/alphagenome, with an accompanying Python software development kit provided to interact with the model. We also provide a genome interpretation suite to facilitate the exploration and interpretation of AlphaGenome. This offers a range of functionalities, such as streamlined variant scoring with quantile calibration and identification of critical sequence regions through contribution scores from ISM-based experiments. The model source code, weights, variant scoring implementations and a selection of variant evaluation datasets and predictions are available at https://github.com/google-deepmind/alphagenome_research.


Part 1: Genomes and Genes

Genetic variants and disease causation



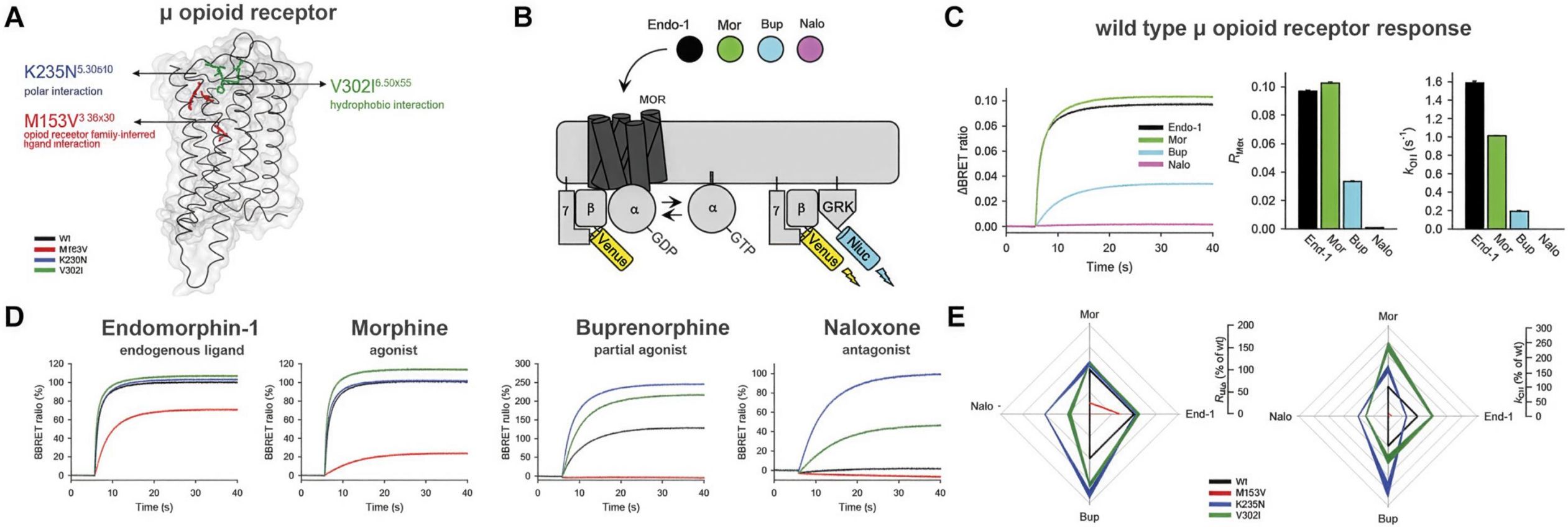
Family study



Gain of Function (GOF)
or
Loss of Function (LOF)

Impact of genetic variants on drug response (Pharmacogenetics)

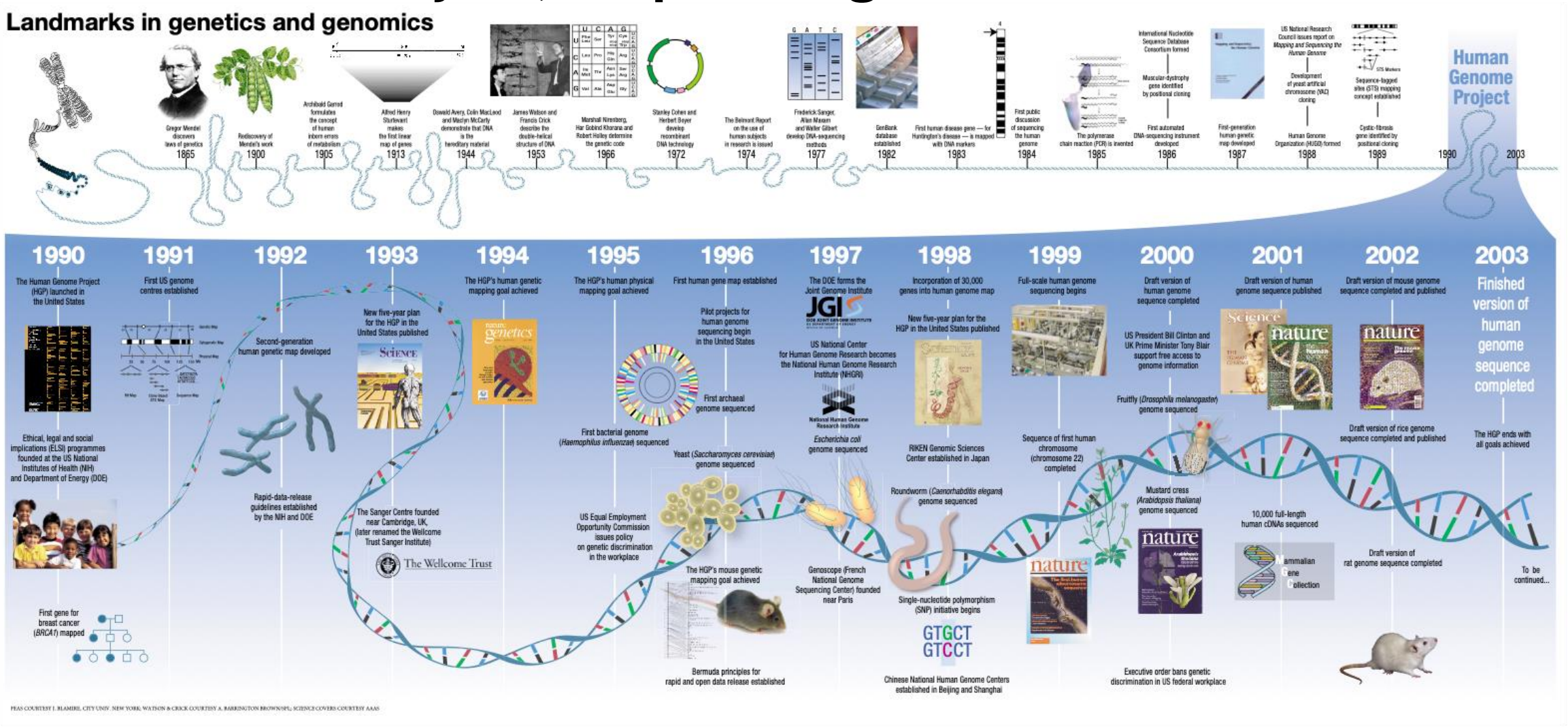
Impact of variants on functional response



Genetic variants can also affect compound metabolization, clearance etc…

Hauser et al., Cell 2017

Human Genome Project, Sequencing it all.



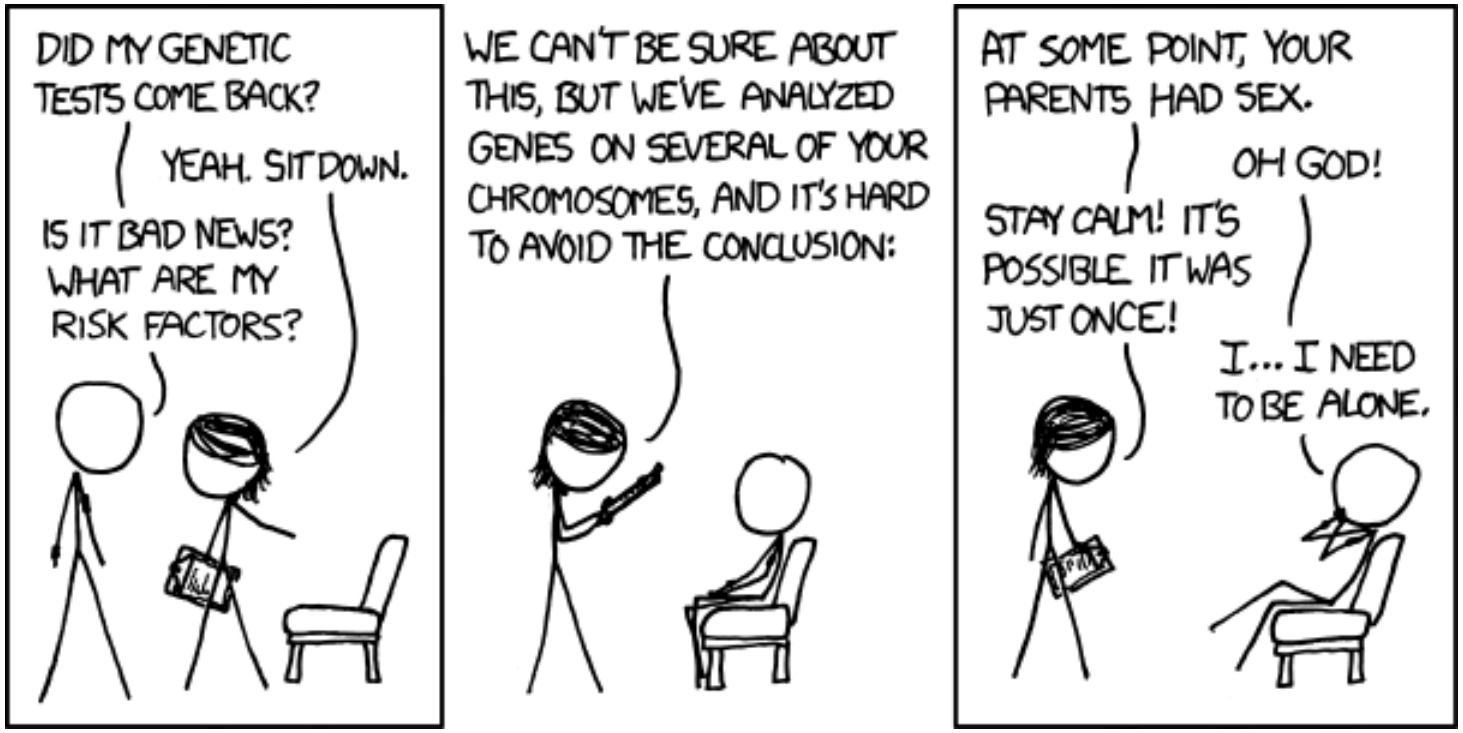
"It is humbling for me and awe inspiring to realise that we have caught the first glimpse of our own instruction book, previously known only to God." (Francis Collins, Project Director)

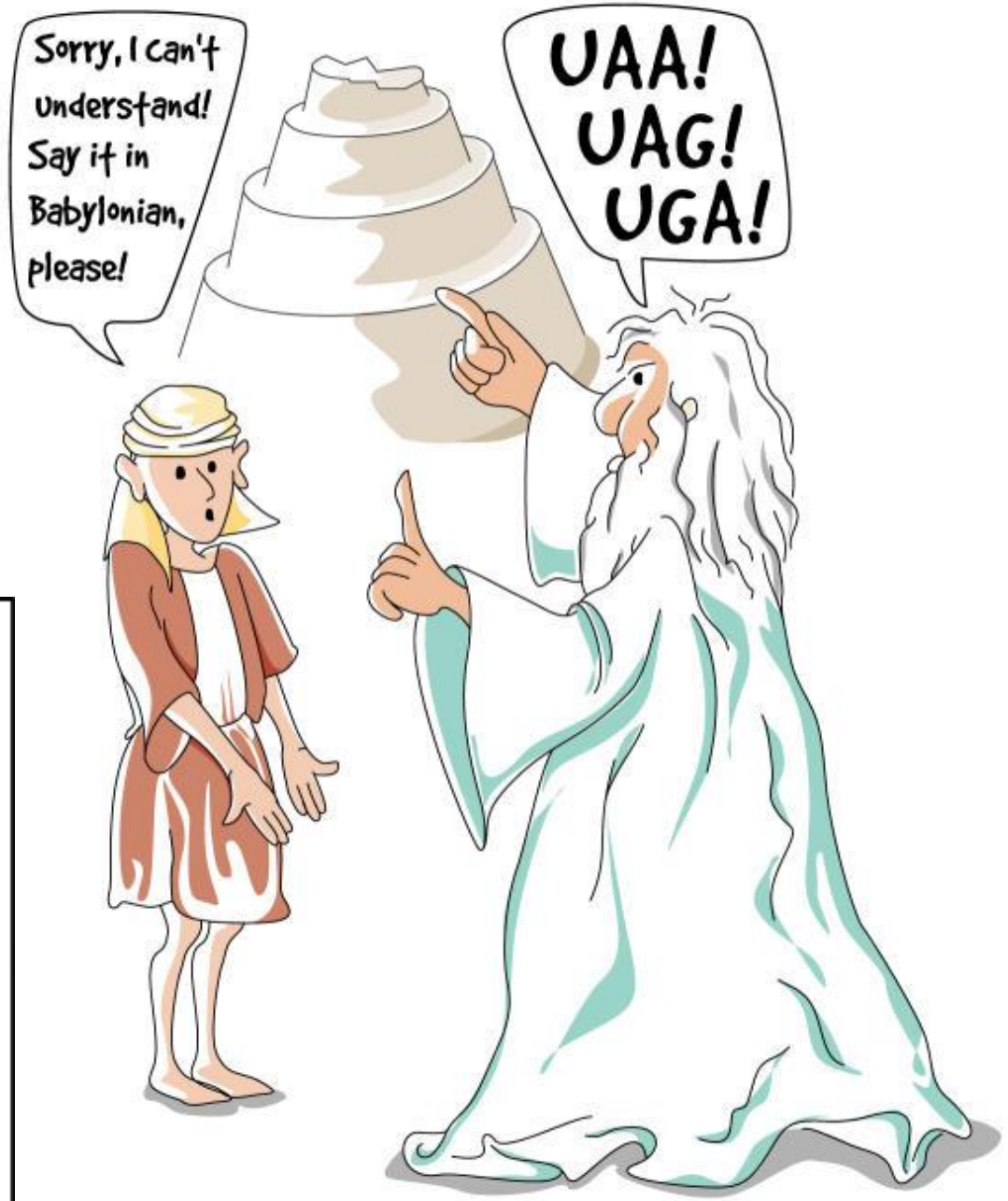
"We now have the possibility of achieving all we ever hoped for from medicine." (UK Science Minister Lord Sainsbury)

Or the limits of linear determinisms

- **"As was predicted… getting the sequence will be the easy part… The hard part will be finding out what it means." —** Sydney Brenner.

(Nobel Prize of Medicine 2002, for their discoveries concerning genetic regulation of organ development and programmed cell death)





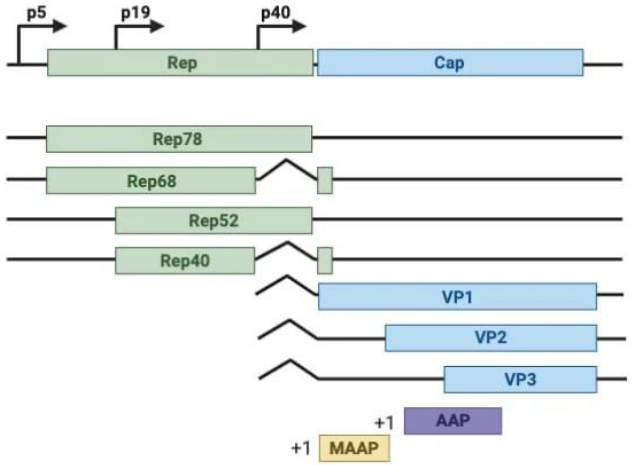
Weigmaan 2004, EMBO Reports

Genotype is one determinant of phenotype, it is not all … or is it?

What is a genome?

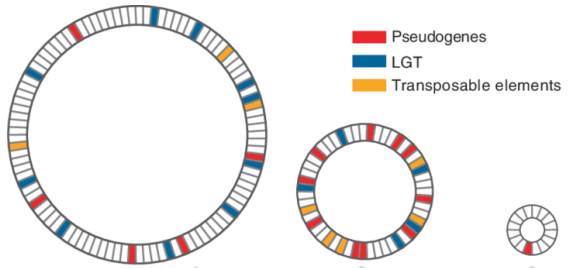
Viruses, Procaryotes and Eukaryotes

Virus
DNA or RNA
Single or double stranded
Data Compaction
-Small regulatory sequences
-Genes can sometimes overlap



Addgene

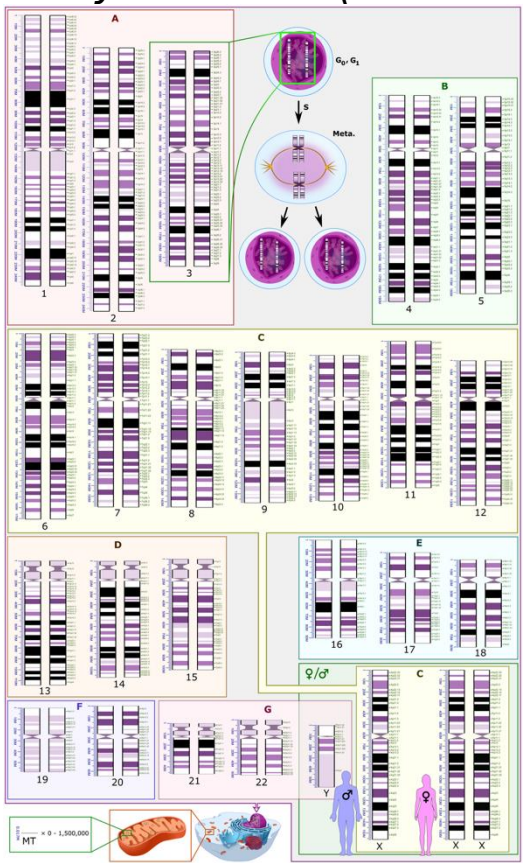
Procaryotes (no nucleus, bacteria)
Double stranded DNA
Data Compaction
Mostly Naked
Circular
Trans-specific information exchange (Plasmids)



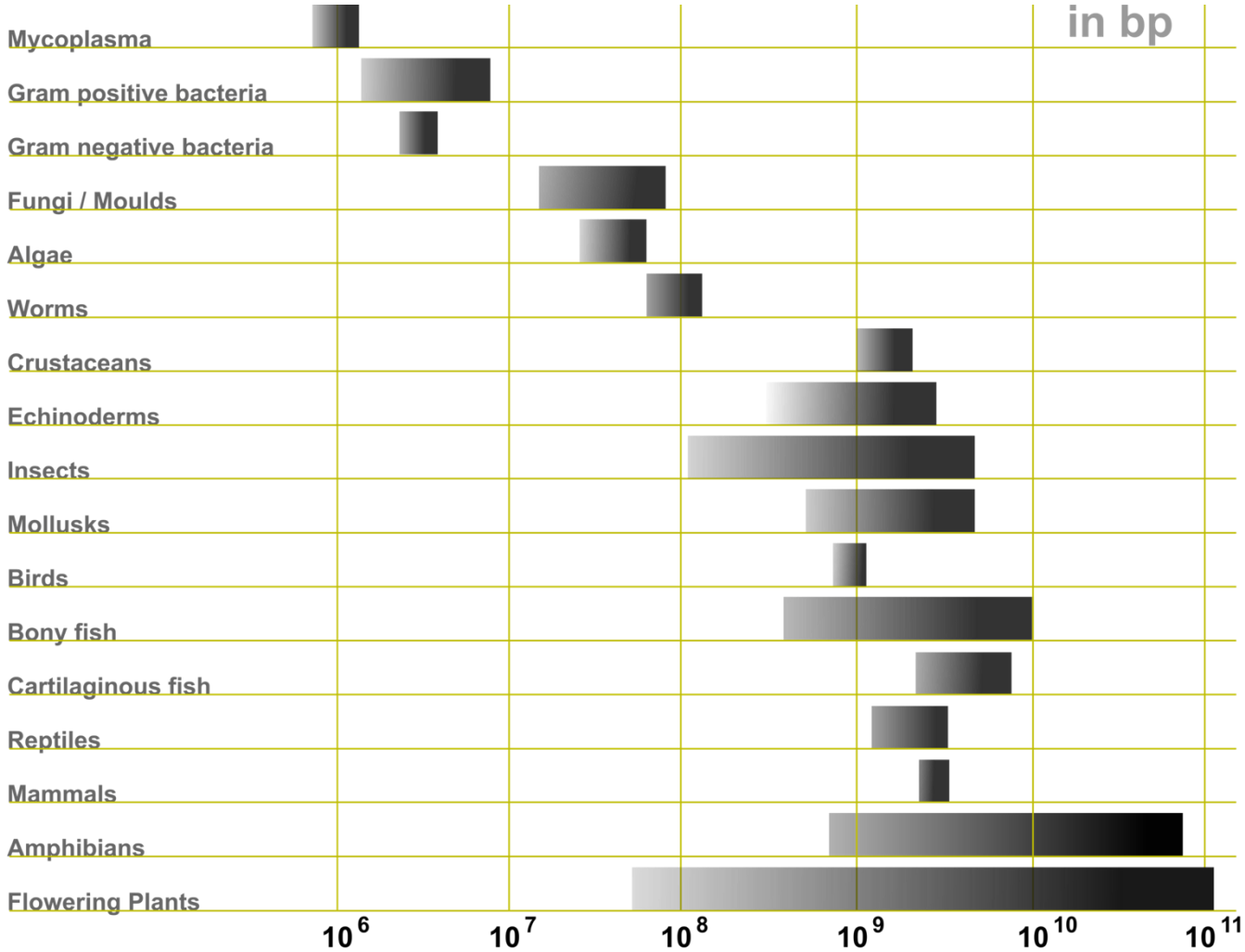
	Free-living	Recent or facultative pathogen	Obligate symbiont or pathogen
Genome size	Large (5-10 MB)	Intermediate (2-5 MB)	Small (0.5-1.5 MB)
Number of pseudogenes	Few	Many	Rare
Incidence of LGT	Frequent	Frequent to rare	Rare to none
Selfish genetic elements	Few	Common	Rare
Genome organization	Stable or unstable	Unstable	Stable
Effective population size	Large	Small	Small

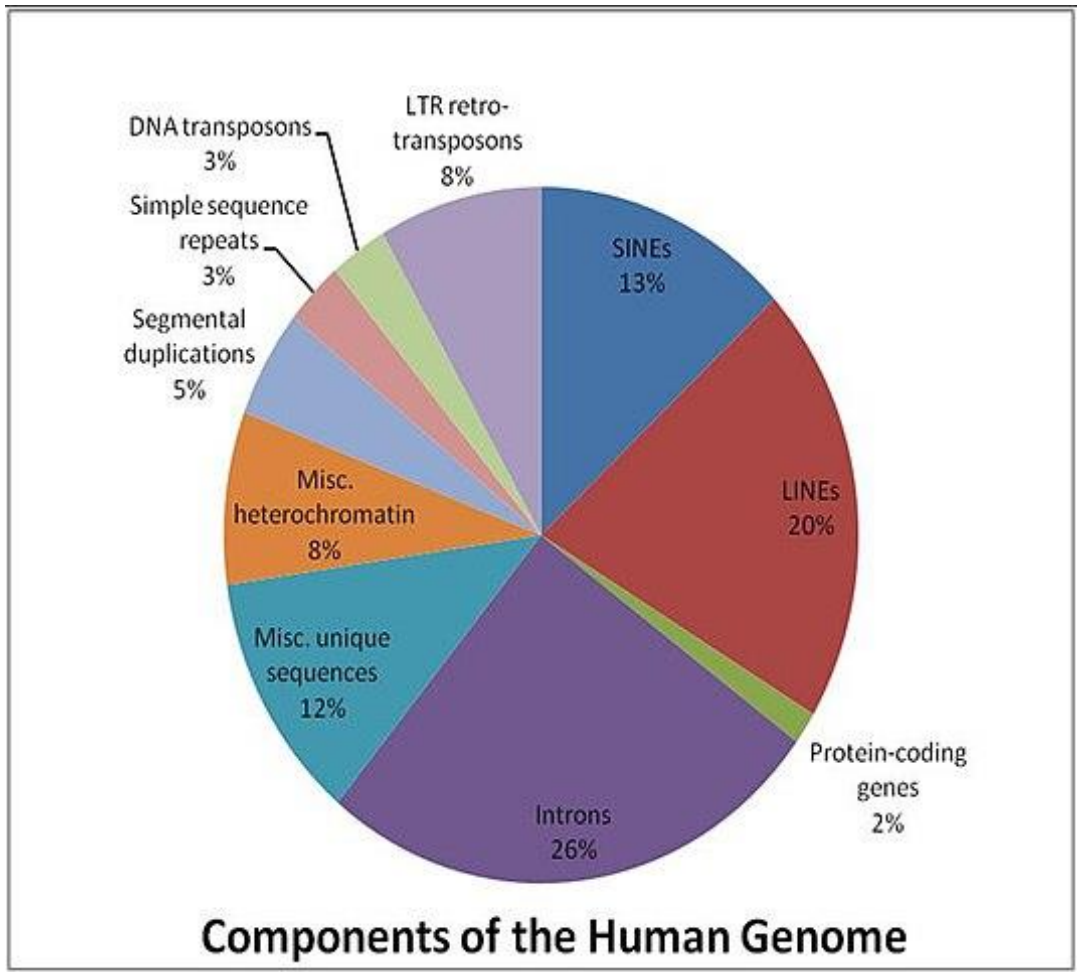
Ochman& Davalos, 2006 Science

Eukaryotes (cells with nucleus)
Double stranded DNA
Organized on chromozomes
Attached to proteins (chromatin)
Lots of non-coding sequences
Vestigial genes
Gene duplication
Heteroploidies
(more than one copies of each gene)
Secondary Genomes (mitochondria)

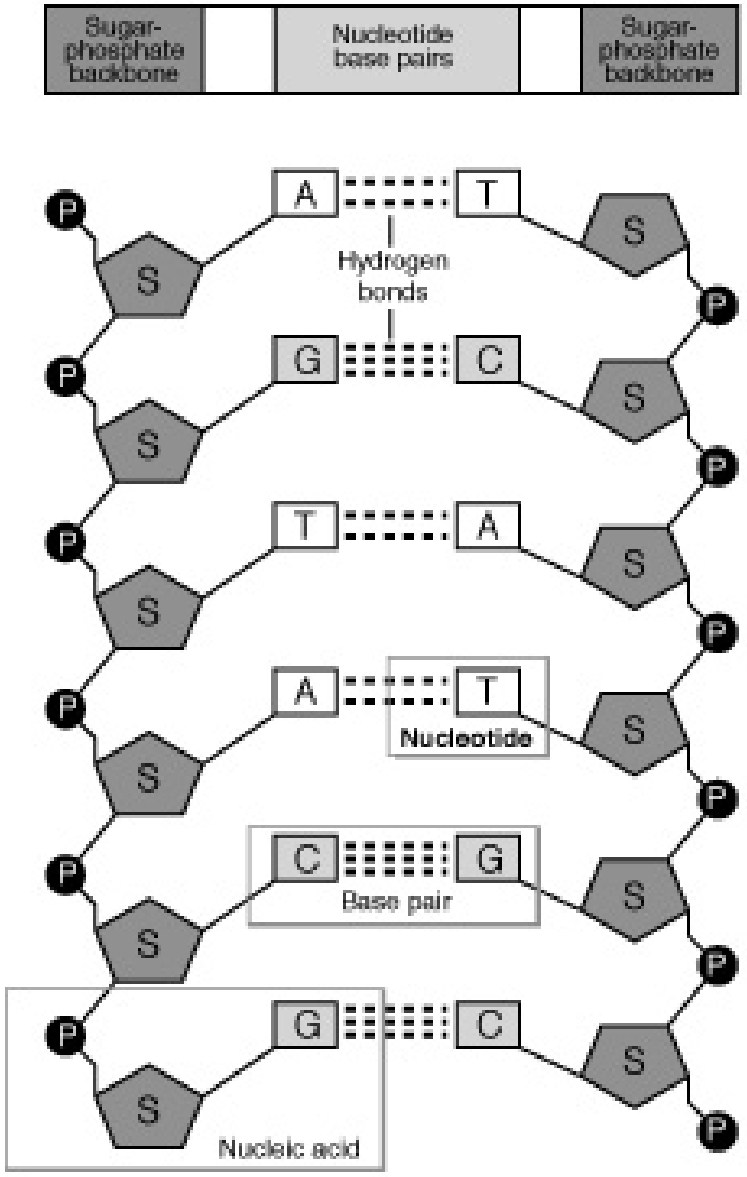


Genome size

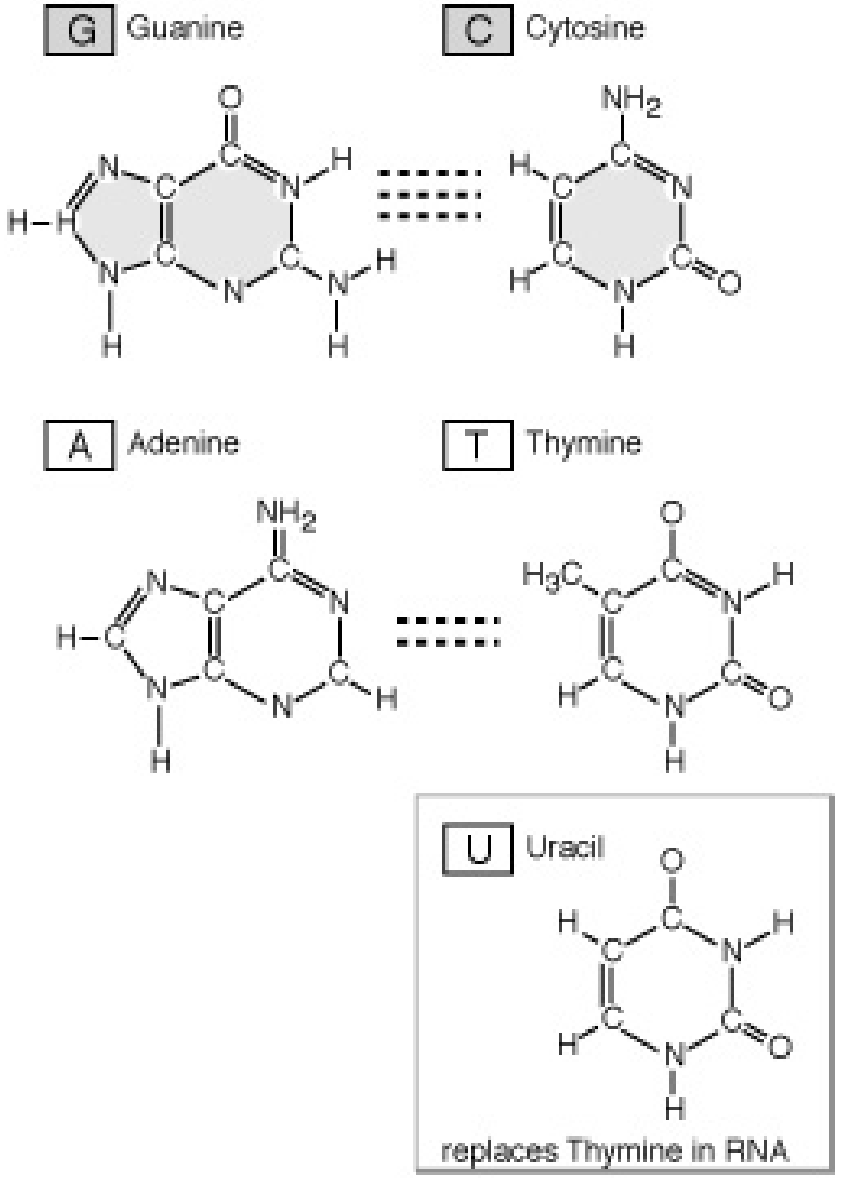

in bp
Mycoplasma
Gram positive bacteria
Gram negative bacteria
Fungi / Moulds
Algae
Worms
Crustaceans
Echinoderms
Insects
Mollusks
Birds
Bony fish
Cartilaginous fish
Reptiles
Mammals
Amphibians
Flowering Plants
10 6
10 7
10 8
10 9
10 10
10 11


DNA transposons 3%
LTR retro-transposons 8%
Simple sequence repeats 3%
Segmental duplications 5%
SINEs 13%
LINEs 20%
Misc. heterochromatin 8%
Misc. unique sequences 12%
Introns 26%
Protein-coding genes 2%
Components of the Human Genome

Deoxyribonucleic Acid (DNA)



Nucleotides



Why is nucleotide complementarity important ?

Biological Processes

Genome Replication

RNA synthesis

Protein Synthesis

Regulation of Gene expression

3D structure of the genome

RNA regulation

RNA transport

Methods

Gene Sequencing

PCR

In situ detection of nucleic acids (FISH)
- Genes
- Messengers

Modulation of Gene expression

Genome Engineering

Synthetic Biology.

Etc…

The complementarity of the bases is the mean to read and write information in nucleic acid.

DNA sequences encodes:

Genes

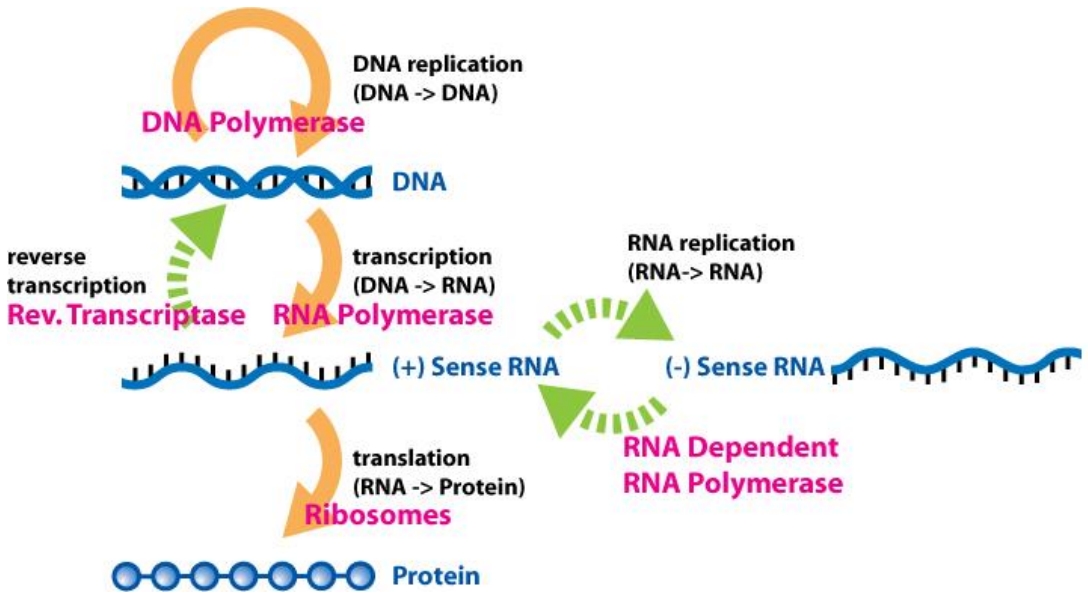
- The nature of the gene product : **Protein** or **RNA**
- The Expression level of genes: **Yes/No, How Much, When** and **Where**.

Genome

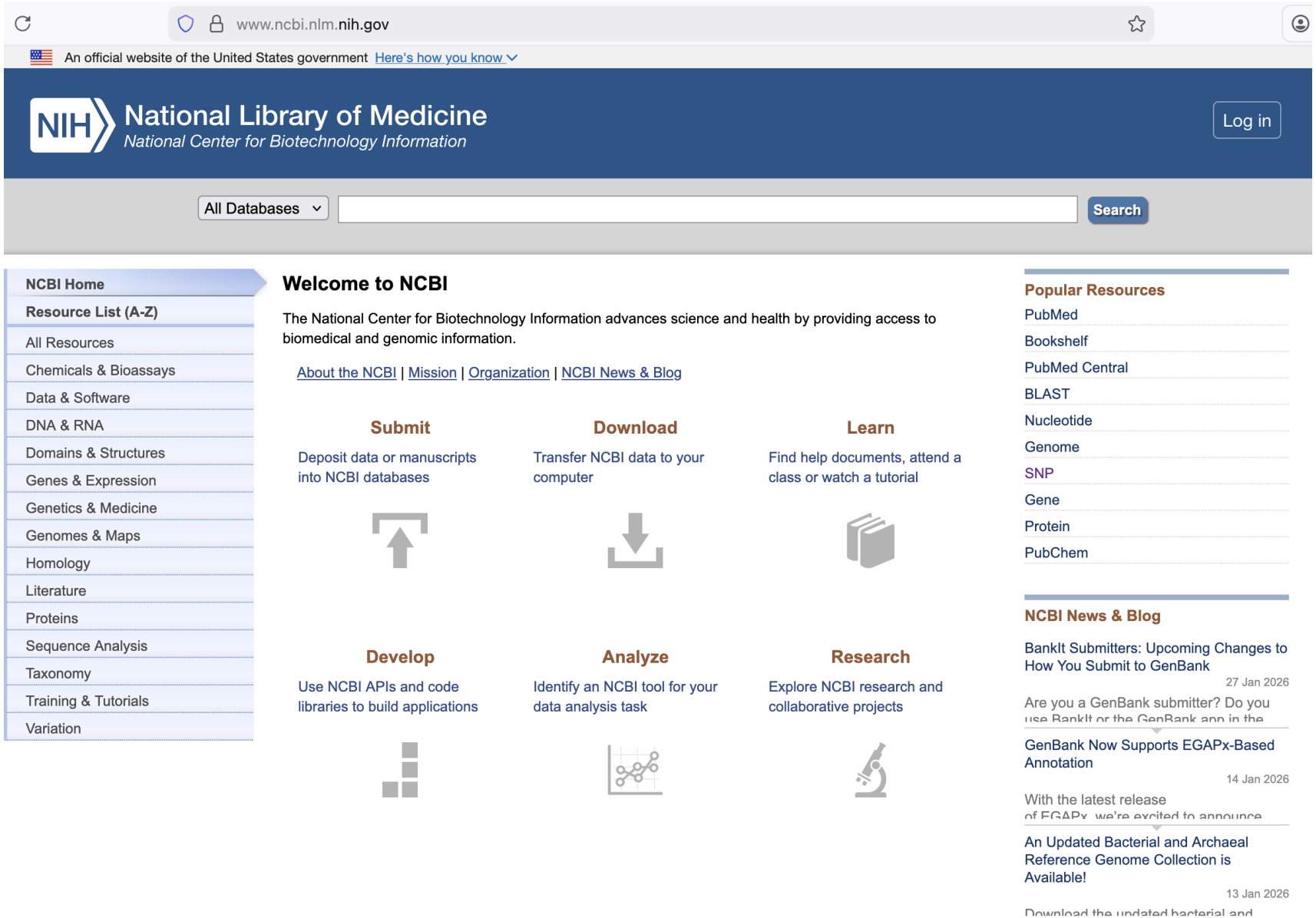
- Genome maintenance: Structure, Replication
- Junk, parasites or drivers of evolution ?

This information is encoded in multiple languages.
The genetic code use for protein synthesis is just one of these languages.

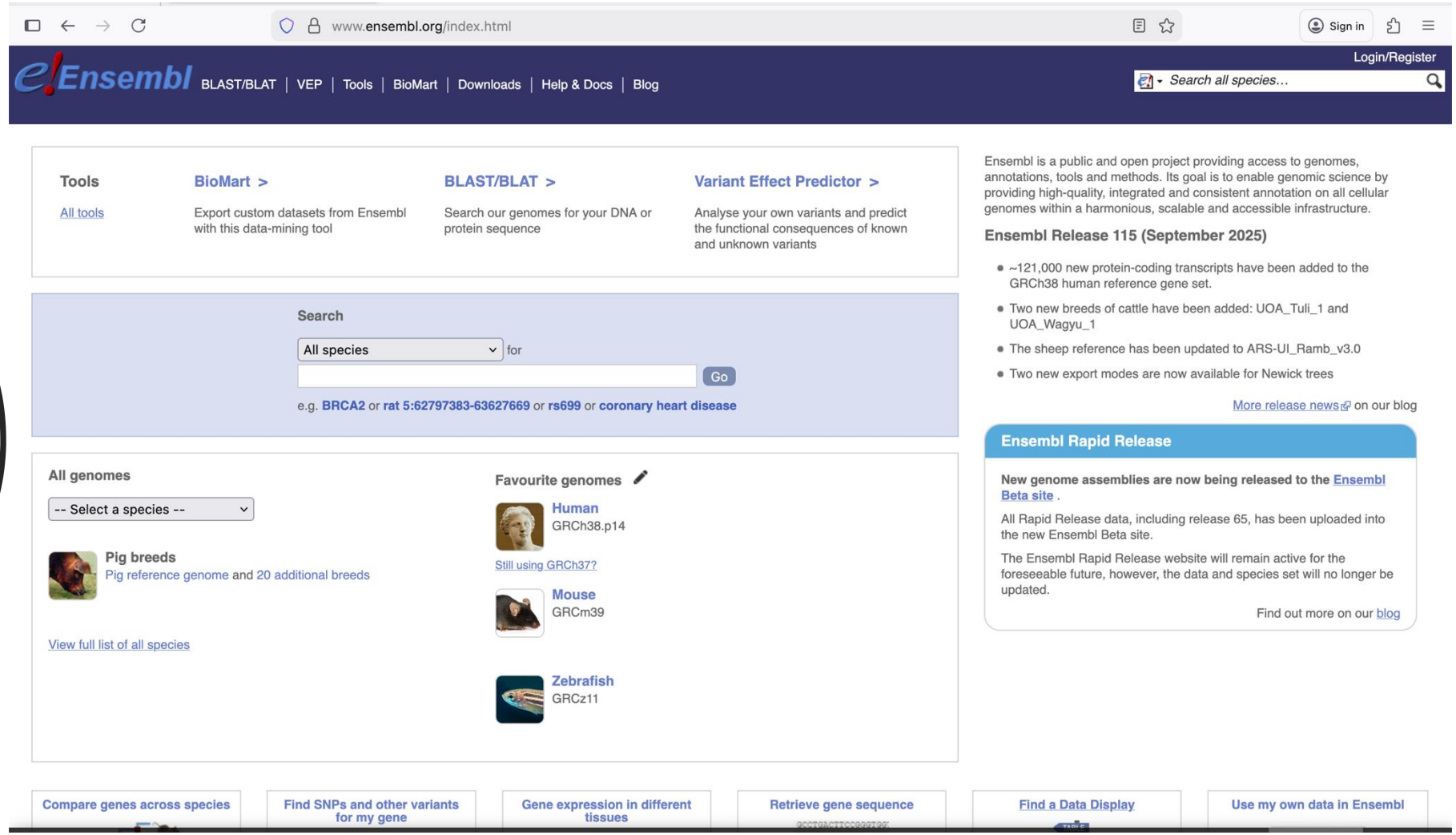
The Dogma:



General Genome Databases NCBI from NIH



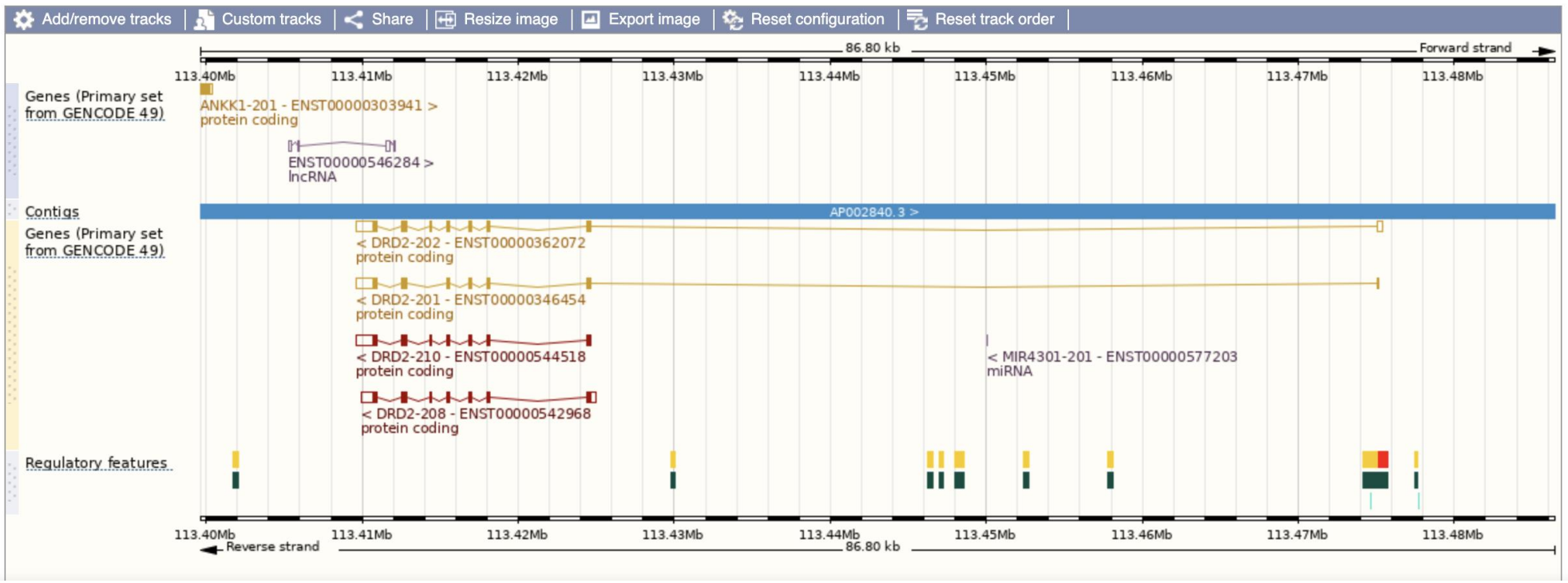
General Genome Databases Ensembl EMBL



Example of gene entry (human interface)

Name	DRD2 (HGNC Symbol)
MANE	This gene contains MANE Select ENST00000362072, ENSP00000354859
UniProtKB	This gene has proteins that correspond to the following UniProtKB identifiers: P14416
RefSeq	This Ensembl/Gencode gene contains transcript(s) for which we have selected identical RefSeq transcript(s). If there are other RefSeq transcripts available they will be in the External references table
CCDS	This gene is a member of the Human CCDS set: CCDS8361.1, CCDS8362.1
Ensembl version	ENSG00000149295.15
Other assemblies	This gene maps to 113,280,327-113,347,124 in GRCh37 coordinates. View this locus in the GRCh37 archive: ENSG00000149295
Gene type	Protein coding
Annotation method	Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article.

Go to Region in Detail for more tracks and navigation options (e.g. zooming)



There are several data bases one can use (an AI generated list).

Core nucleotide/genome sequence databases

- GenBank (NCBI) – primary public DNA sequence archive, part of the INSDC with daily exchange with ENA and DDBJ.
- European Nucleotide Archive (ENA, EMBL-EBI) – Europe's partner archive for nucleotide sequences.
- DNA Data Bank of Japan (DDBJ) – Japan's partner archive within INSDC.
- NCBI Genome resources – genome assemblies, annotations, variation viewers, BioProject/BioSample.

Human gene, variant, and GWAS resources

- GeneCards – integrated human gene database aggregating genomic, transcriptomic, proteomic, genetic and clinical info.
- dbSNP (NCBI) – catalog of short genetic variants.
- dbVar (NCBI) – structural variation database (CNVs, insertions, deletions, inversions, etc.).
- dbGaP (NCBI) – genotype–phenotype database for controlled-access human studies.

NHGRI–EBI GWAS Catalog – curated catalog of published human GWAS

General genome resource lists

- NHGRI "Online Research Resources" – curated list of major genome-related databases and tools.
- NCBI "All Resources" guide – consolidated entry point to NCBI genomic and related databases.
- U of T Genome Databases guide – field-specific links including model organism genome resources. (https://guides.library.utoronto.ca/c.php?g=251933&p=1675867)

Functional genomics and signatures

- GEO (Gene Expression Omnibus) – public repository for array- and sequence-based functional genomics data.
- GEO Datasets – curated GEO datasets with analysis tools.
- MSigDB (Molecular Signatures Database) – curated gene set collections for GSEA.

Population genomics reference resources

- 1000 Genomes / International Genome Sample Resource (IGSR) – deep catalog of common human genetic variation, ongoing maintenance and mapping to current assemblies.

Psychiatric Genomics Consortium (PGC) – Central hub for GWAS of schizophrenia, bipolar disorder, MDD, ADHD, ASD, PTSD, Tourette, OCD, etc.; provides extensive summary-statistics downloads per disorder and cross-disorder analyses.

Ageing and longevity genomics

- Human Ageing Genomic Resources (HAGR) – umbrella for ageing-related databases (GenAge, AnAge, GenDR, LongevityMap).
- LongevityMap – human variants associated with longevity.

Biobanks and population cohorts (with genomic data)

- "All of Us" Research Program (NIH) – large population biobank with genomic and environmental data.
- Biobank Graz, China Kadoorie Biobank, Estonian Biobank, EuroBioBank, UKBiobank etc. – large international biobanks with biospecimens and associated data.

Genome Geography, vocabulary

Locus: a location on the genome

Chromosome: A large nuclear structure that contains a long molecule of DNA associated to proteins (chromatin). Chromosomes encode subsets of the genome.

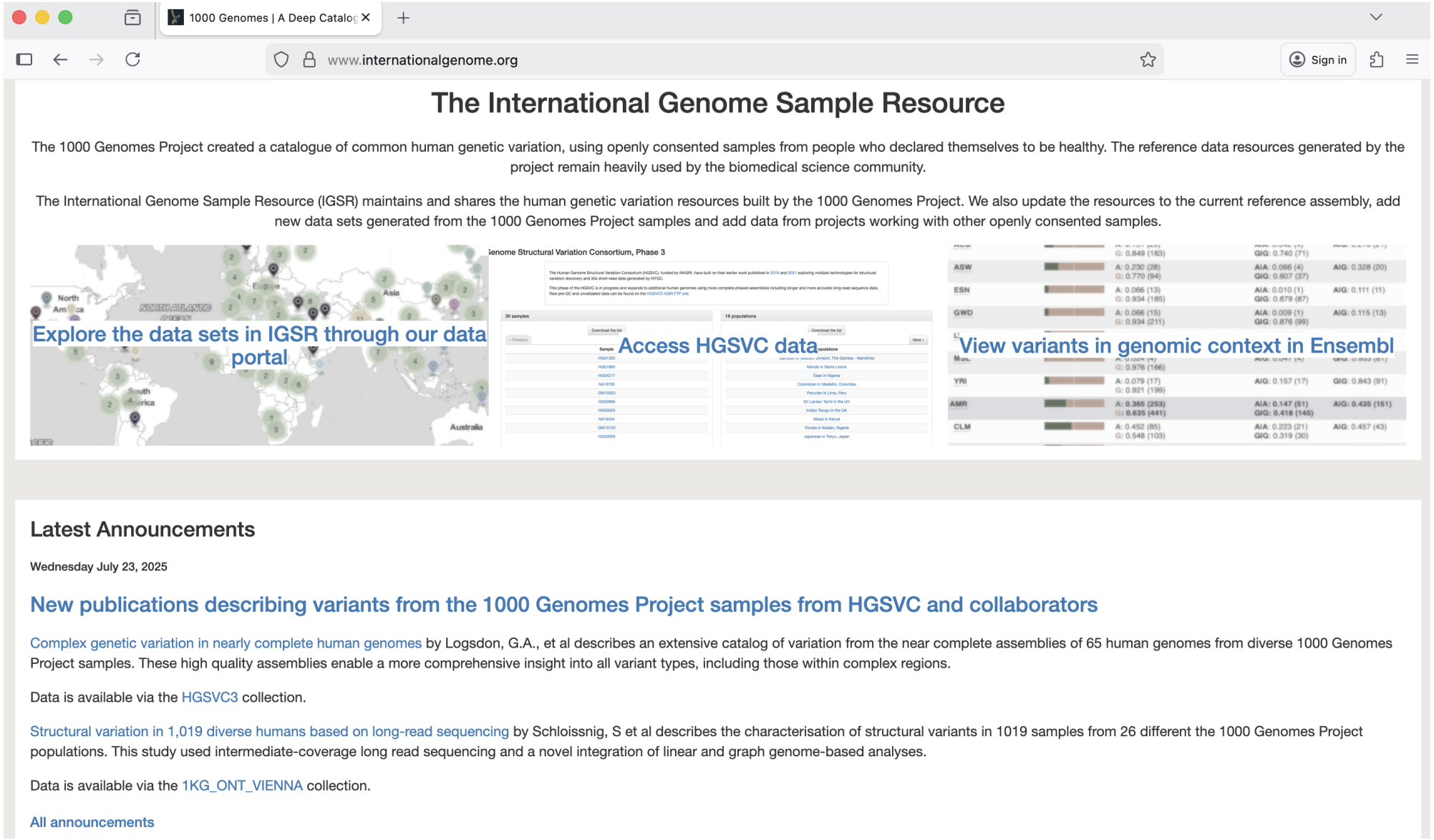
Gene: two meanings. The Mendelian gene is a basic unit of heredity. The molecular gene is a sequence in DNA that is transcribed to produce RNA.

Single Nucleotide Polymorphisms (SNPs): Variations between multiple genomes wittin a specie

Haplotypes: a set of closely linked DNA variations (like SNPs or alleles) on a single chromosome that are inherited together from a single parent

Haplogroup: haplotypes inherited from a single parent. (e.g.: y chromosome-father, mtDNA-mother) This is used to define ancestry.

HapMap/1000 Genomes. These are databases of human genetic variations based of haplotype frequency.



Genetic Vocabulary

Mutation: New variant (loaded term, use with care)

Penetrance: the proportion of individuals with a specific genotype (genetic variant) who express the corresponding, observable trait or phenotype

Expressivity: degree to which a specific genotype manifests as a particular phenotype (observable trait or disease) in an individual,

Types of genetic studies:

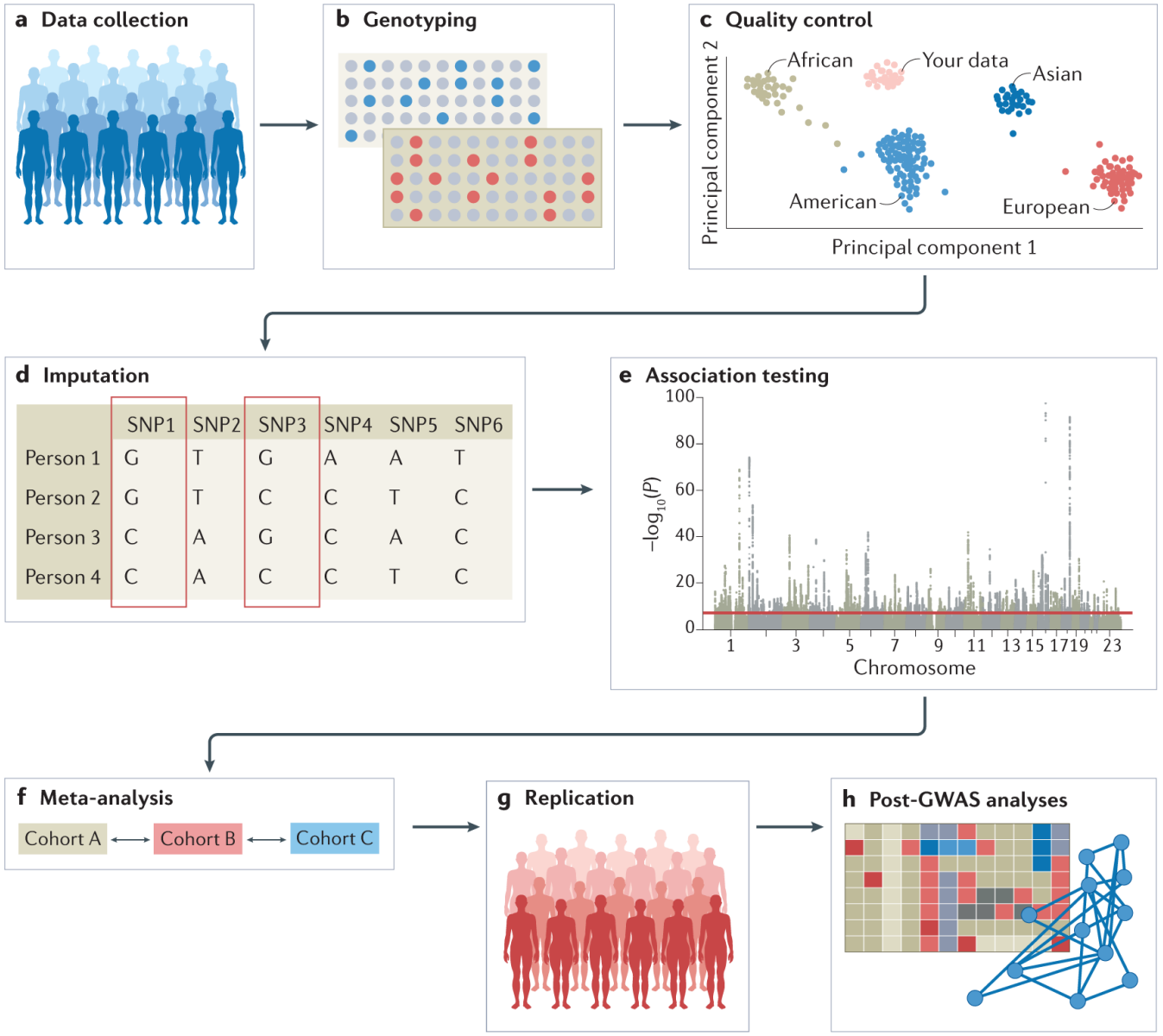
-***Family study:*** Study members of a family to identify variants that corelate with a phenotype.

-***Rare variants:*** Identify a variant (less than 1% of population (often new) that correlates with the apparition of a phenotype. Generally, these variants have a strong effect on the phenotype.

-***Genome Wide Association Study (GWAS):*** Study of the association of **multiple SNPs** covering the whole genome with a given continuous (e.g. BMI) or non-continuous (disease diagnostic) outcome .

-***Transcript Wide Association Study (TWAS):*** Study of the association of multiple **RNA Transcripts** covering the whole genome with a given continuous (e.g. BMI) or non- continuous (disease diagnostic) outcome .

Genome Wide Association Studies (GWAS)



Uffelmann et al., 2021 Nat, Rev methods

Overall ,a GWAS is a collection of regressions

In GWAS, one typically fit a separate regression model per variant. A common basic formula for a quantitative trait is a linear regression:

$$Y_i = \beta_0 + \beta_1 G_i + \beta_2 C_{i1} + \cdots + \beta_k C_{ik} + \varepsilon_i$$

- Y_i: phenotype (e.g., height) of individual i.
- G_i: genotype dosage for the tested SNP in individual i(coded 0, 1, 2 for copies of the effect allele).
- C_{ij}: covariates such as age, sex, principal components of ancestry, batch, etc.
- β_1: SNP effect size (per-allele change in phenotype).
- ε_i: residual error term.

A case–control GWAS compares genetic variants between people with a disease (cases) and people without it (controls) to find variants associated with disease risk. This is not a continous variable

Case–control GWAS, use logistic regression, for example:

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 G_i + \beta_2 C_{i1} + \cdots + \beta_k C_{ik}$$

Here $Y_i$is disease status (0 = control, 1 = case),

$\exp(\beta_1)$is the odds ratio per effect allele.

Can this be improved by ML in place of running ML on top of it?

For a good read on GWAS go see:
https://doi.org/10.1038/s43586-021-00056-9

Limitations of GWAS:

Sample composition: Population bias

Quality of the measures: What is the diagnostic

Pooling or not pooling: Heterogeneity inside groups

Most SNPs are not coding: No obvious functional outcome

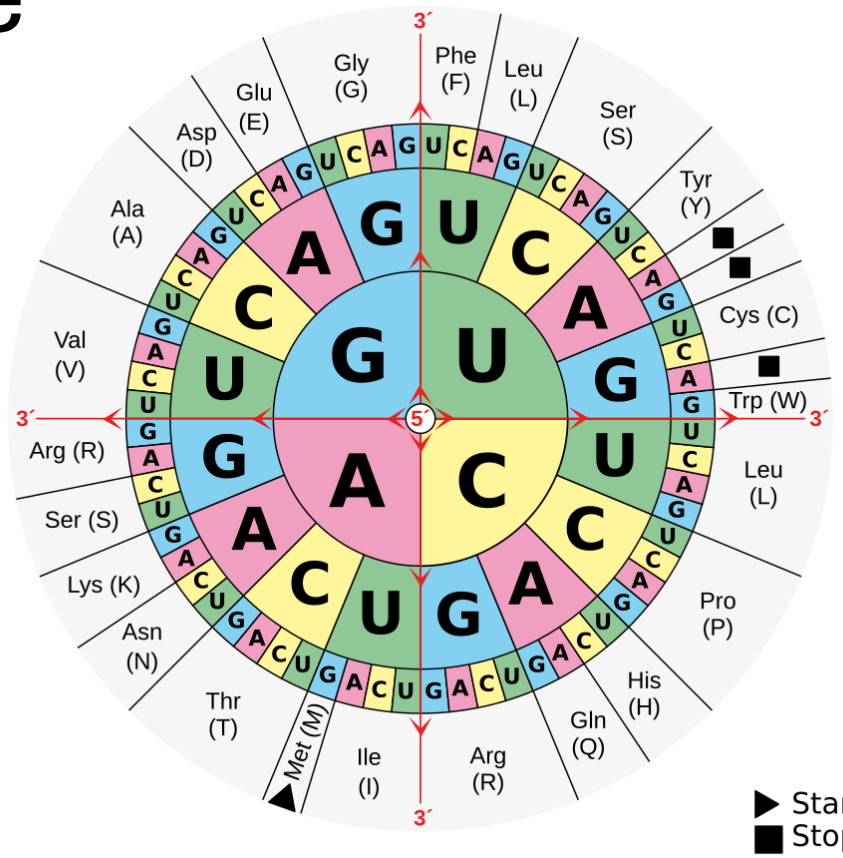
Identification of the affected gene is not always clear.

Polygenic Risk …

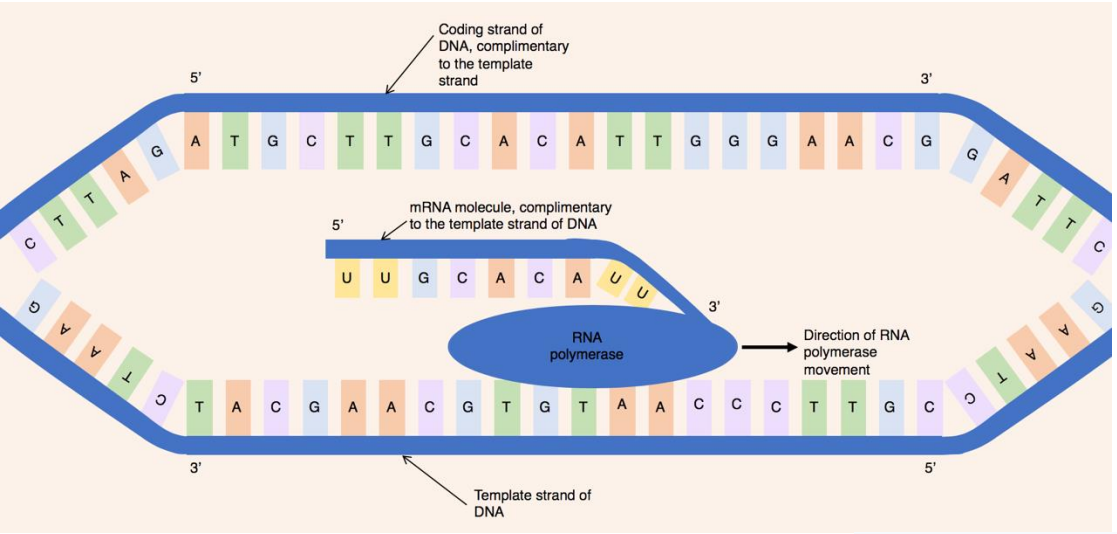
Part 2: Languages of the Genome

Some languages of the genome.

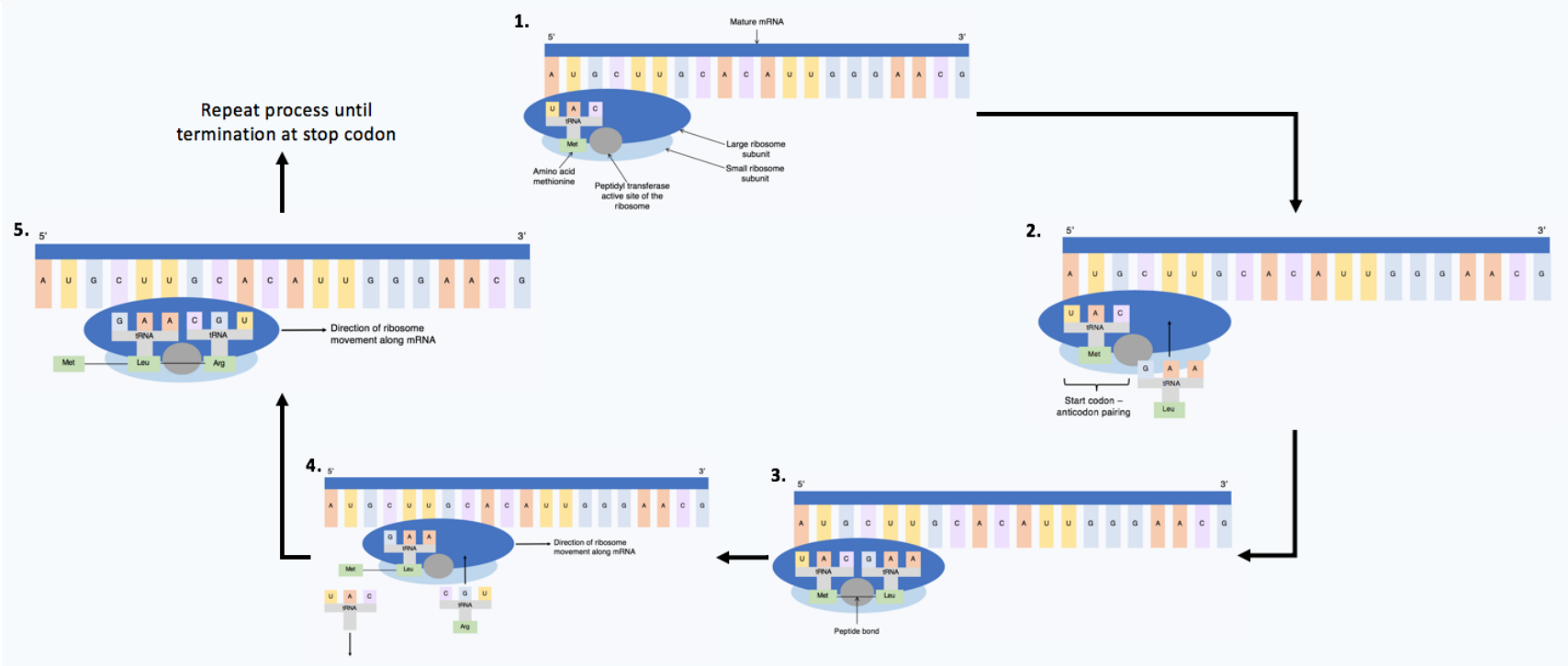
-Protein Coding
-mRNA splicing
-Regulation of Gene expression
-Regulation of Chromatin Structure
-Regulation mRNA translation



Transcription (nucleus)

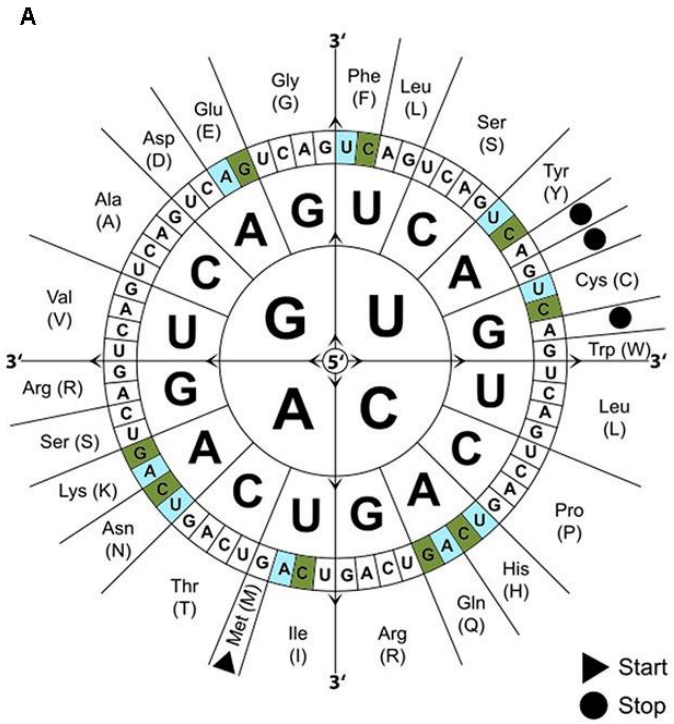


Translation (cytoplasm)



The codon language

Frequency
Can be regulatory

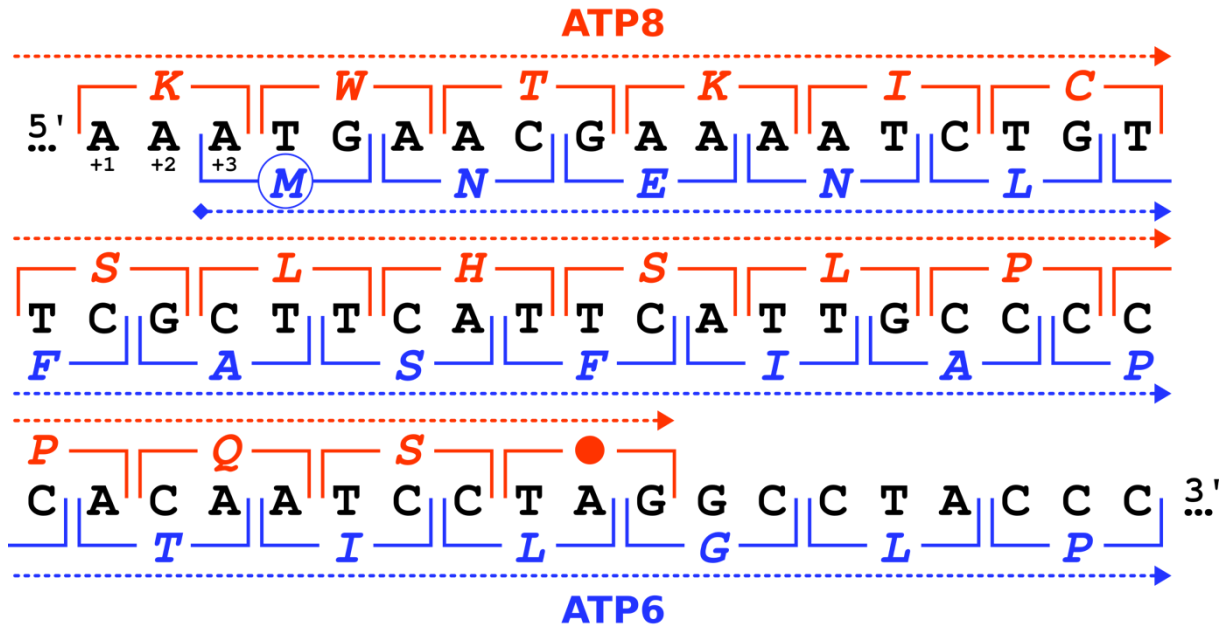


B

Amino acid	Codon	*P. patens*	*A. thaliana*	*S. cerevisiae*	*S. pombe*	*H. Sapiens*
Cys	UGC	+	+	-	+	+
	UGU	-	-	+	-	-
Glu	GAA	-	-	+	-	-
	GAG	+	+	-	+	+
Phe	UUC	+	+	+	+	+
	UUU	-	-	-	-	-
His	CAC	+	+	+	+	+
	CAU	-	-	-	-	-
Ile	AUA	-	-	/	-	-
	AUC	+	+	/	-	+
	AUU					
Lys	AAA	-	-	-	-	-
	AAG	+	+	+	+	+
Asn	AAC	+	+	+	+	+
	AAU	-	-	-	-	-
Gln	CAA	-	-	-	+	-
	CAG	+	+	+	-	+
Tyr	UAC	+	+	+	+	+
	UAU	-	-	-	-	-

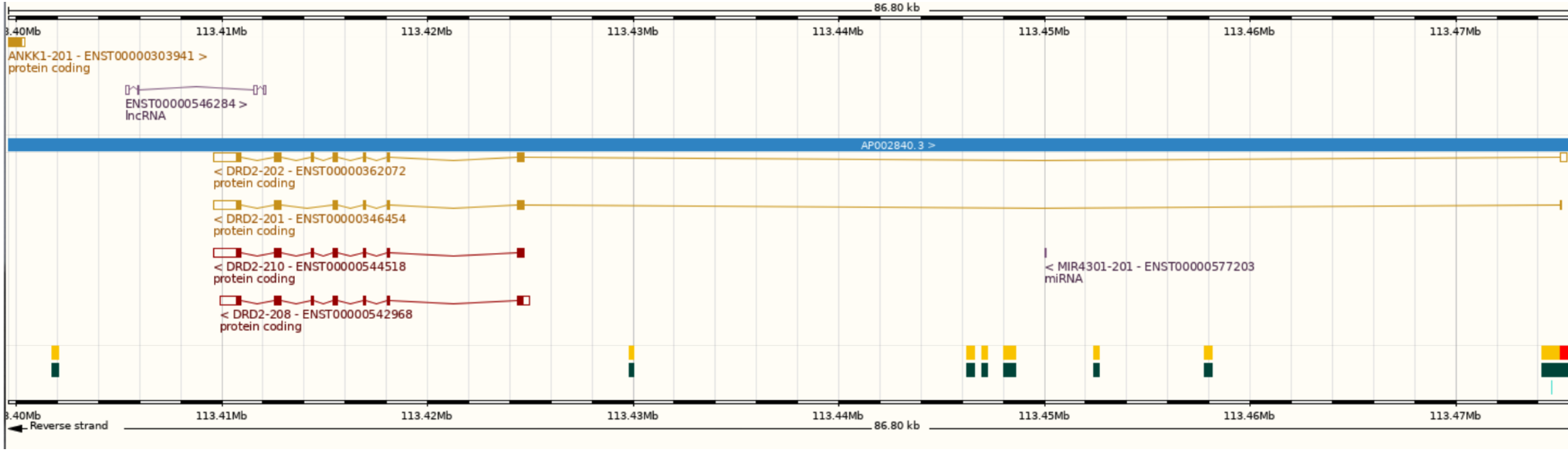
Reading Frame

e.g. Mitochondrial ATP synthase

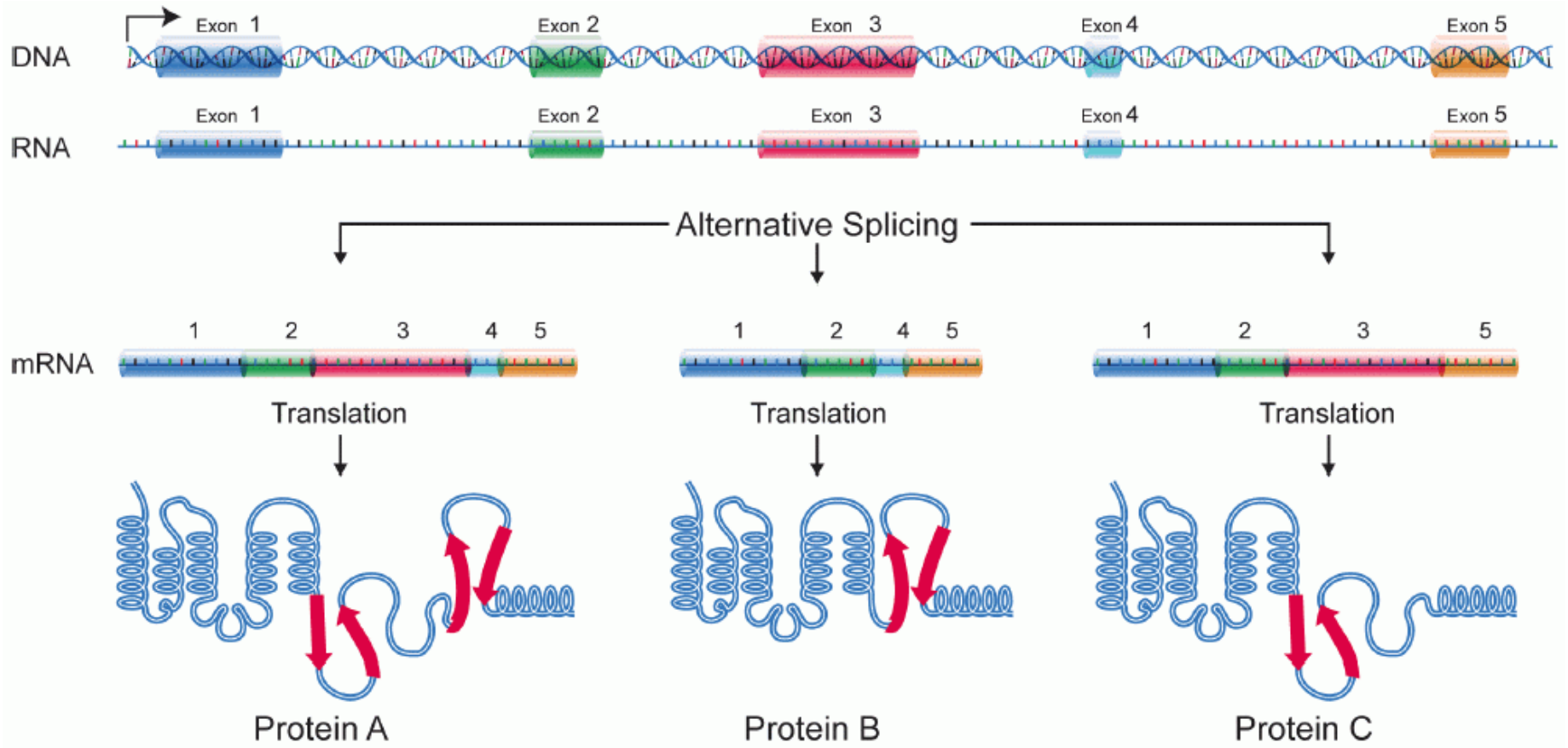


RNA Splicing

Exon: encodes protein
Intron: does not encode protein

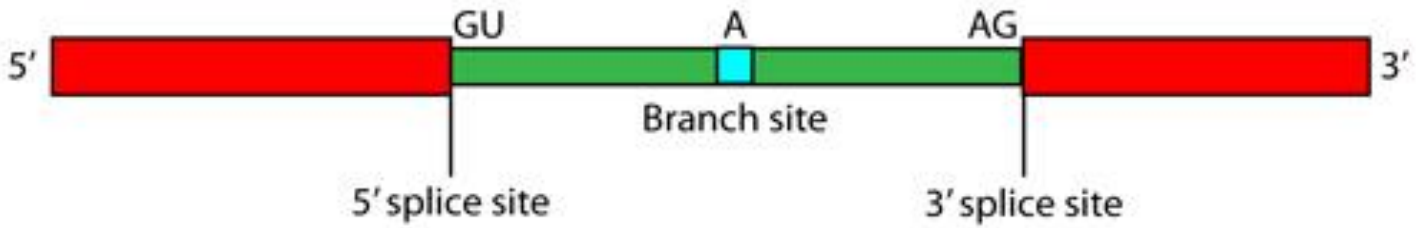


Alternative splicing

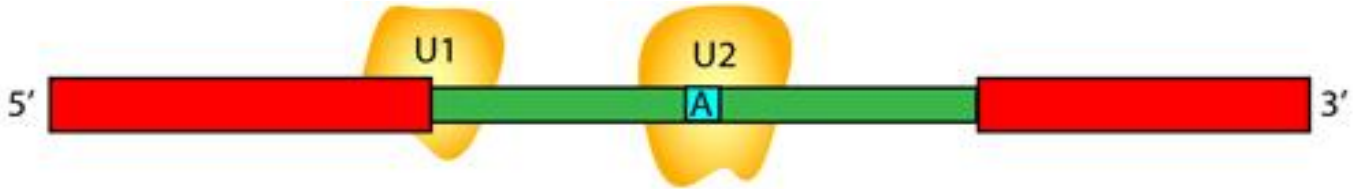


Splicing code

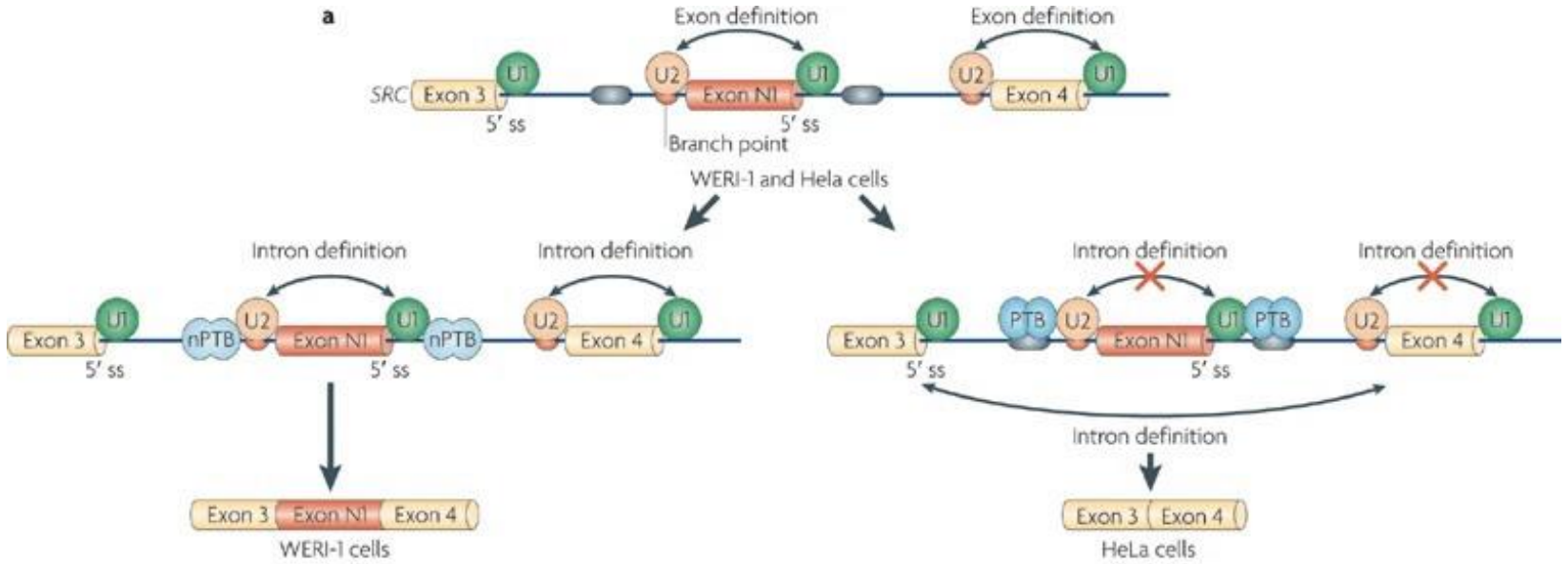
Located 18–40 nucleotides upstream of the 3' splice site. In mammals yUnAy. y= C/U n= any base



The Code



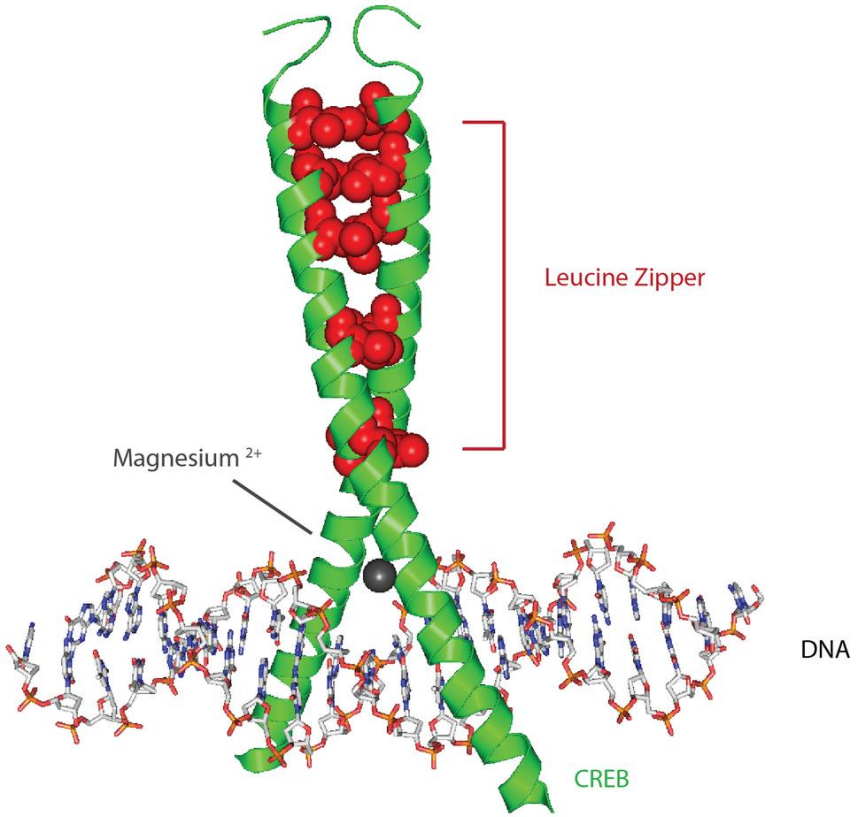
The readers (ribonucleoproteins)



Alternative regulation

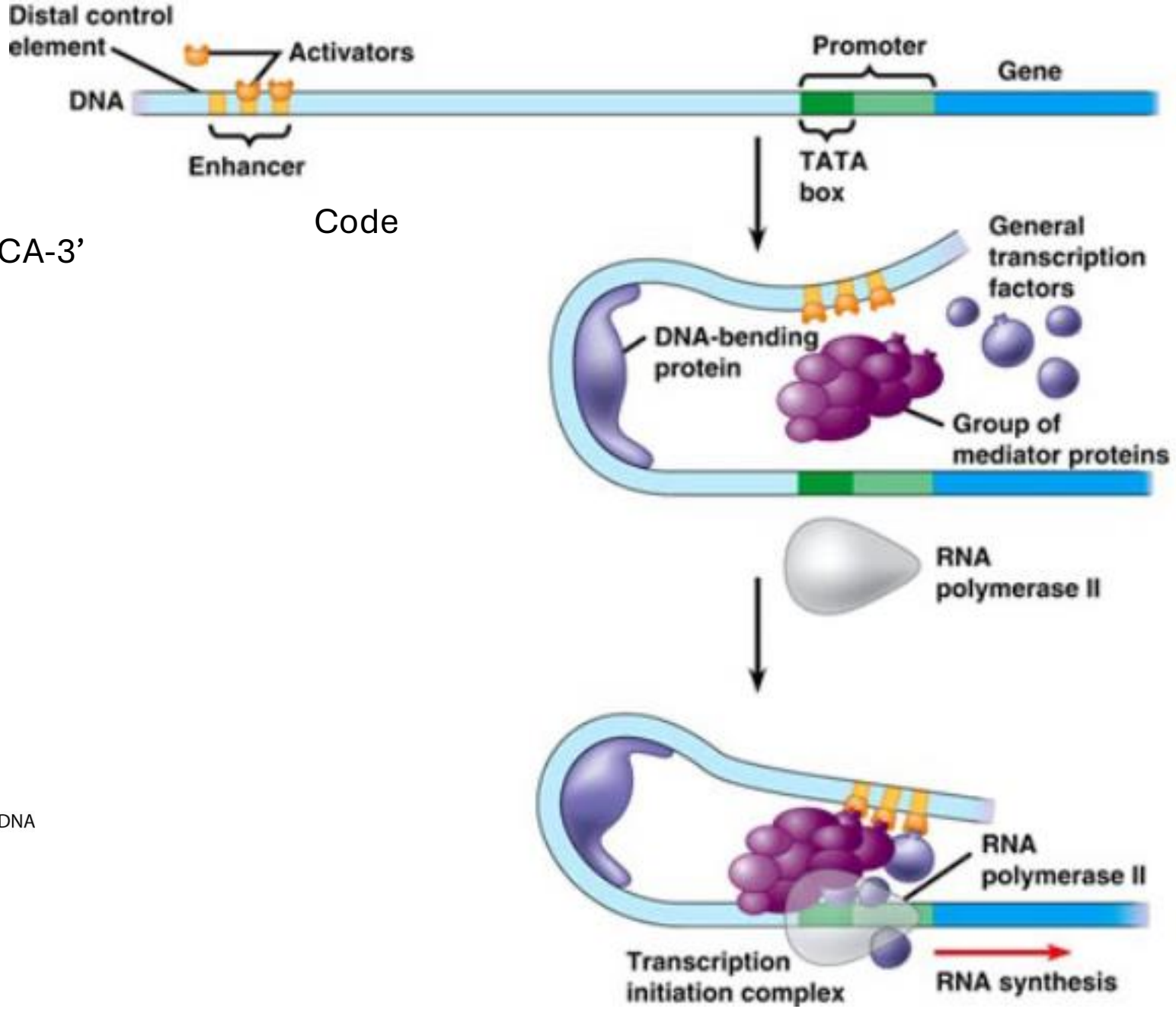
Regulating expression
Languages of transcription

e.g. cAMP Responsive Element: 5'-TGACGTCA-3’

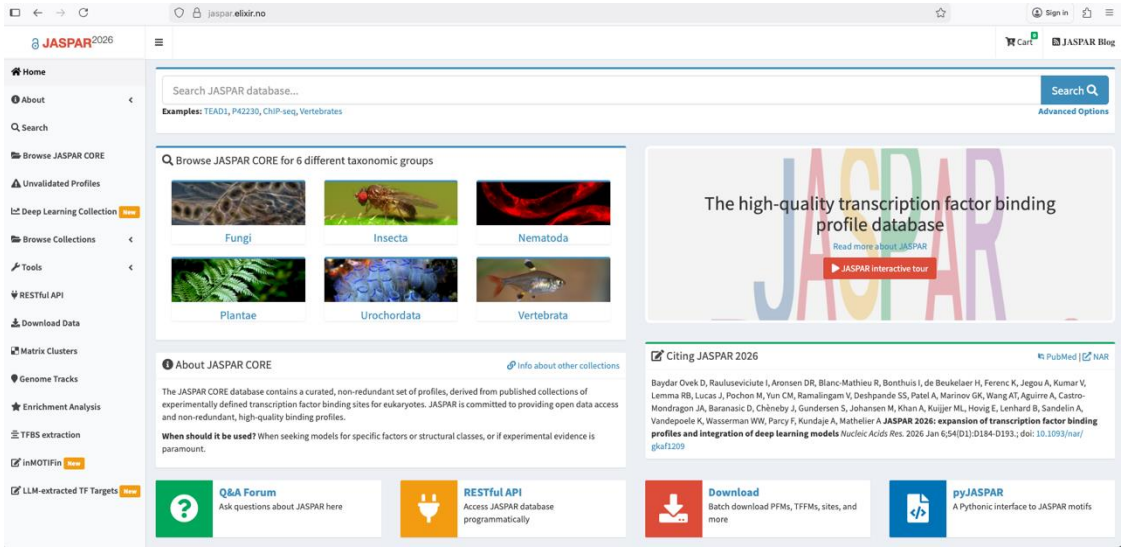


Reader

Code

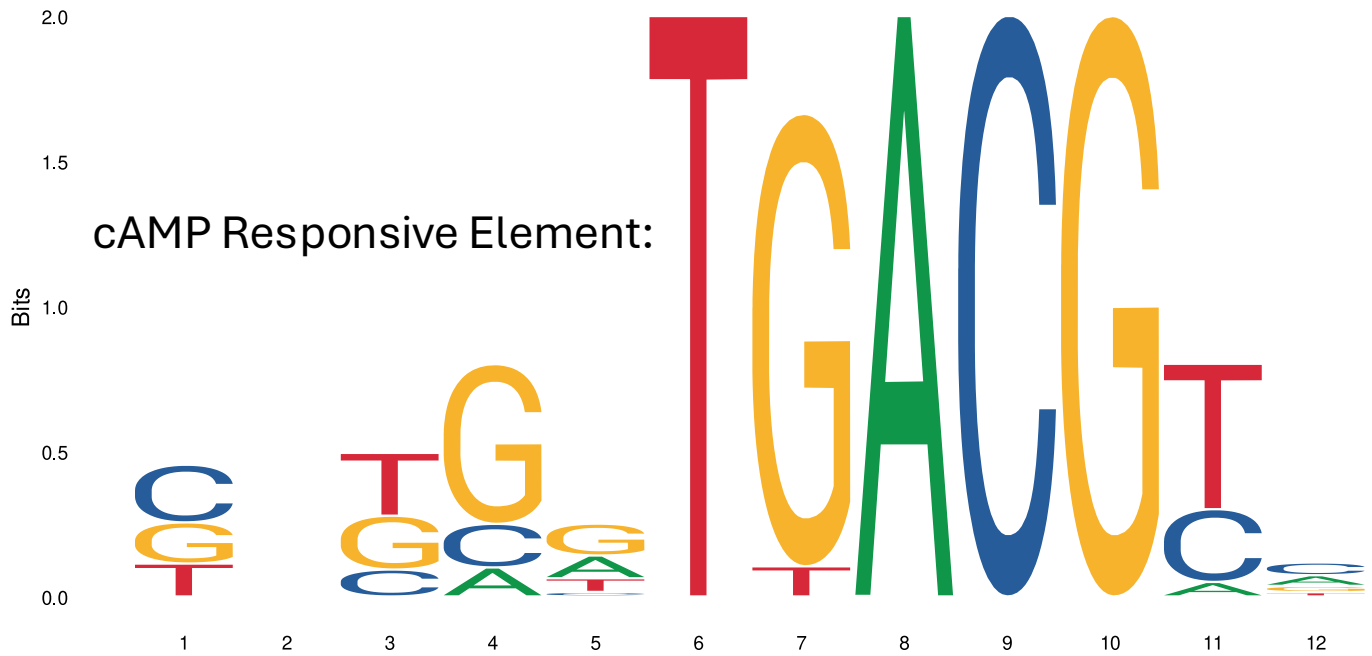


Database of binding sites

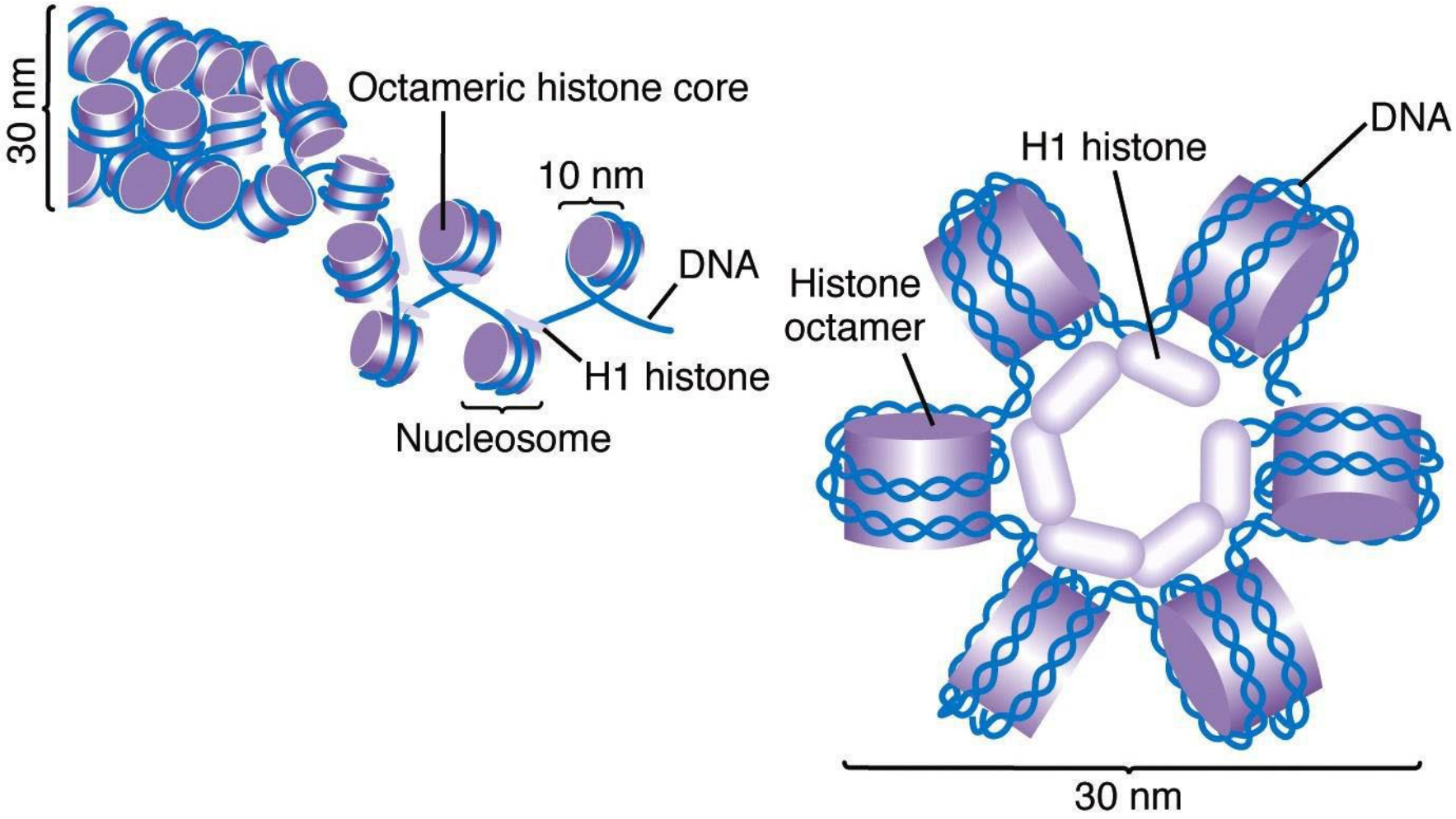


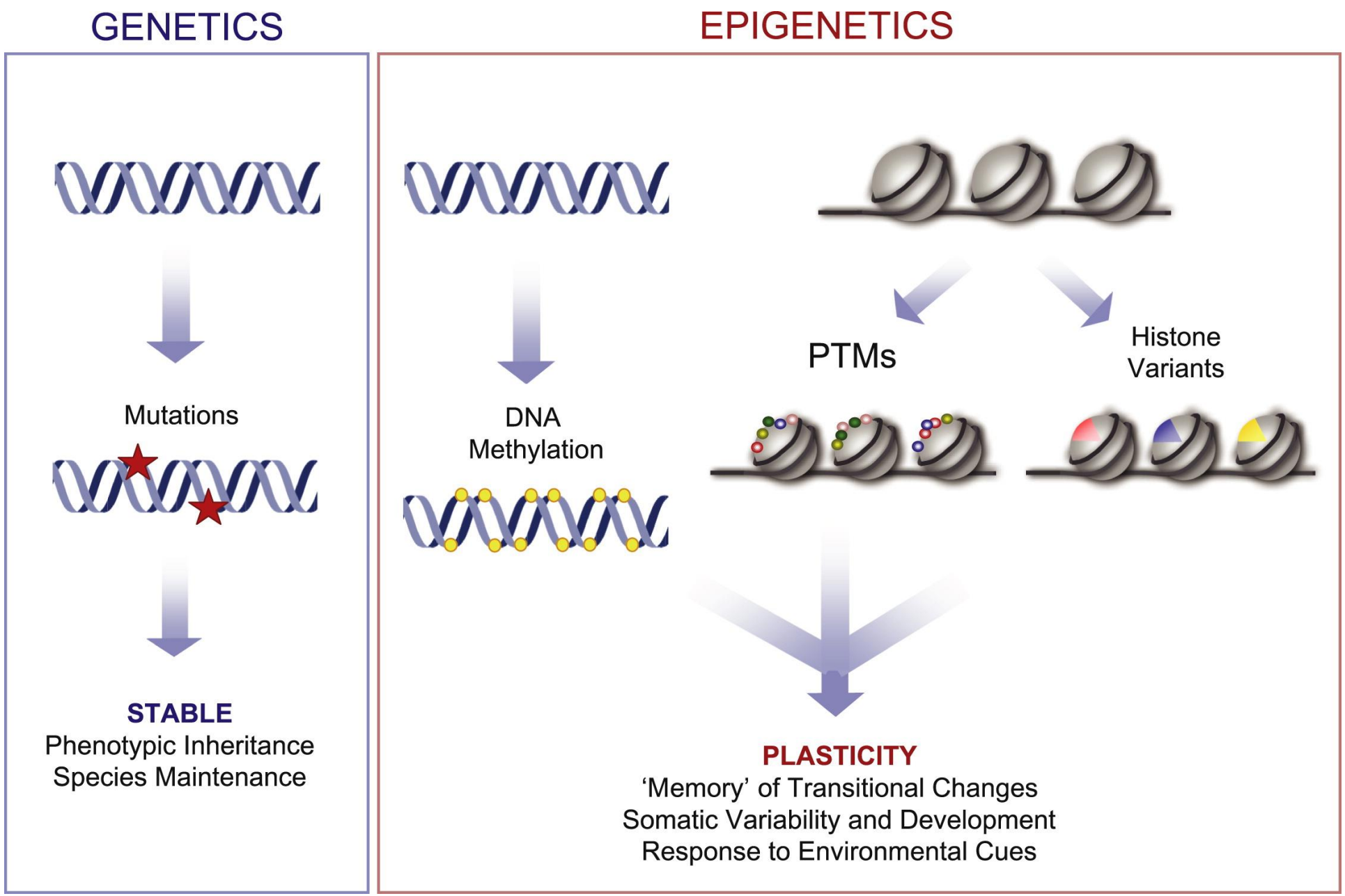
These binding site are often degenerated (receptor ligand analogy)

cAMP Responsive Element:



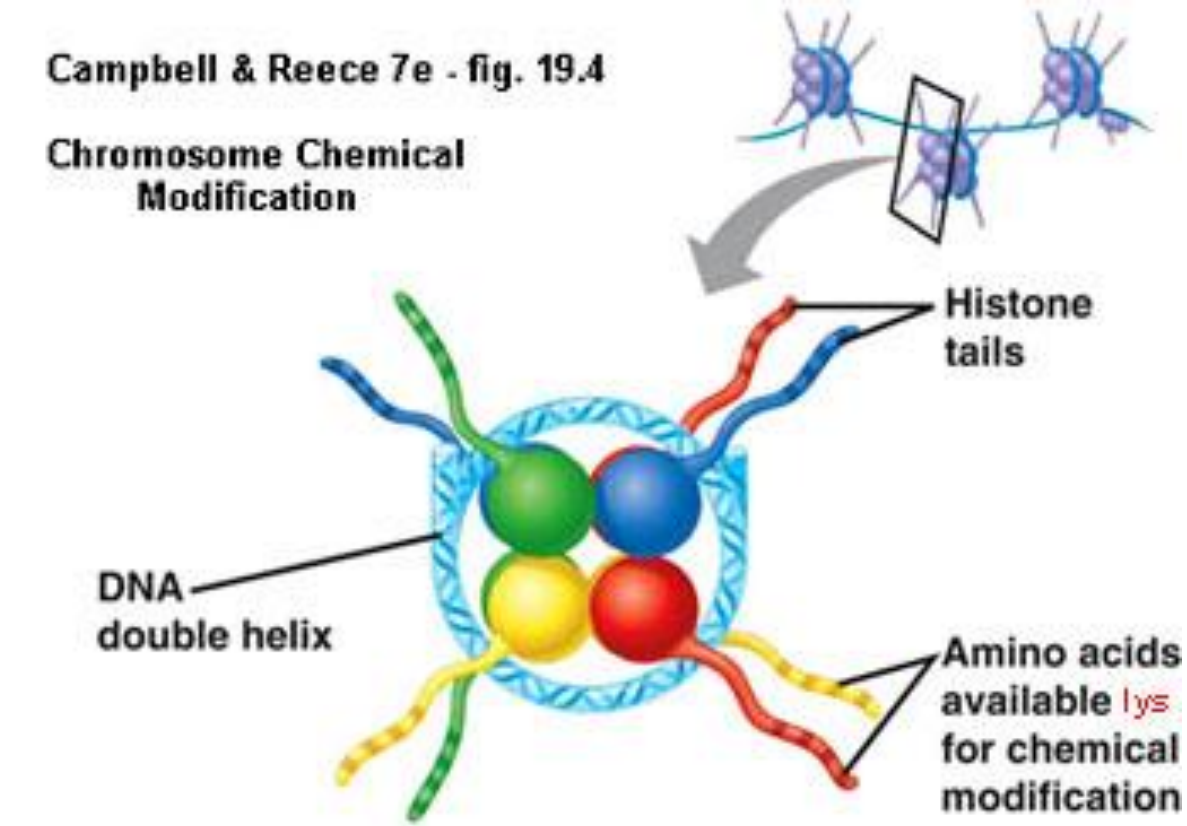
Chromosomes Eukaryotes



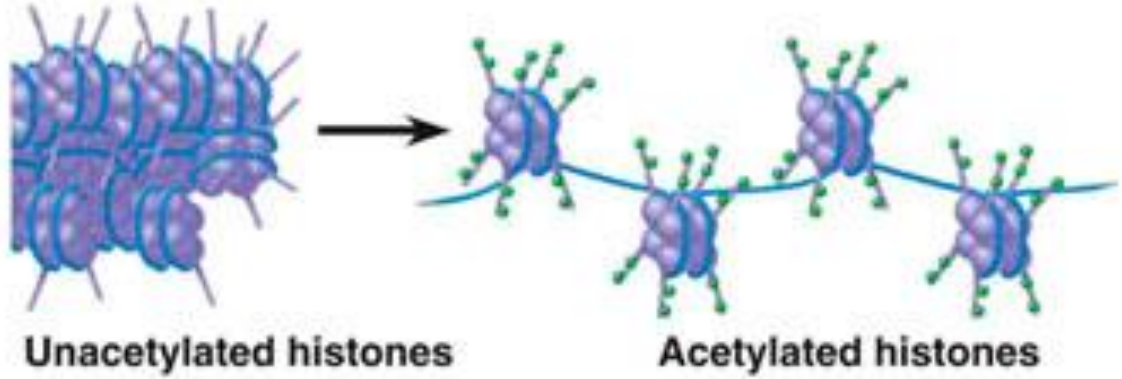

GENETICS
EPIGENETICS
Mutations
STABLE
Phenotypic Inheritance
Species Maintenance
DNA
Methylation
PTMs
Histone
Variants
PLASTICITY
'Memory' of Transitional Changes
Somatic Variability and Development
Response to Environmental Cues

Campbell & Reece 7e - fig. 19.4

Chromosome Chemical Modification



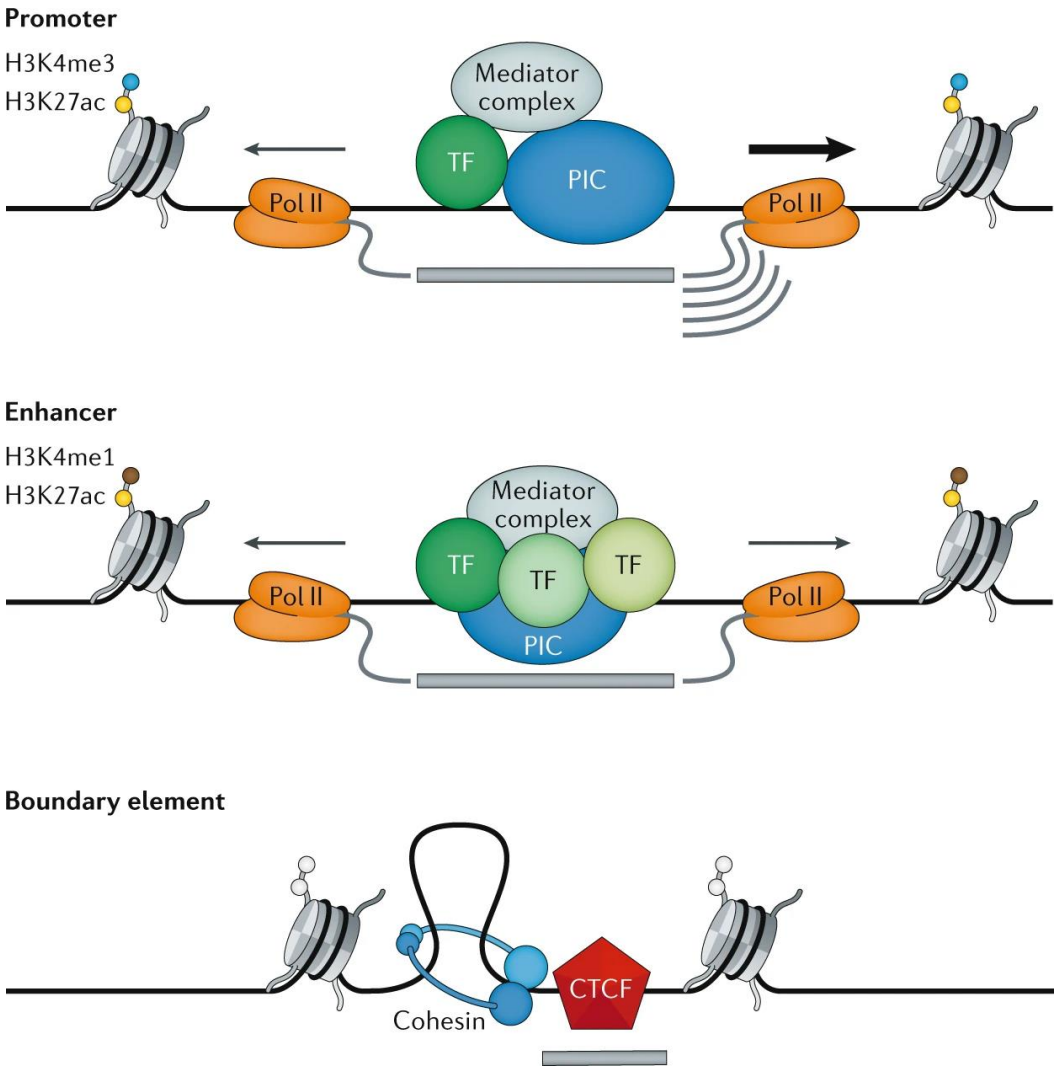
(a) Histone tails protrude outward from a nucleosome



(b) Acetylation of histone tails promotes loose chromatin structure that permits transcription

Boundary elements

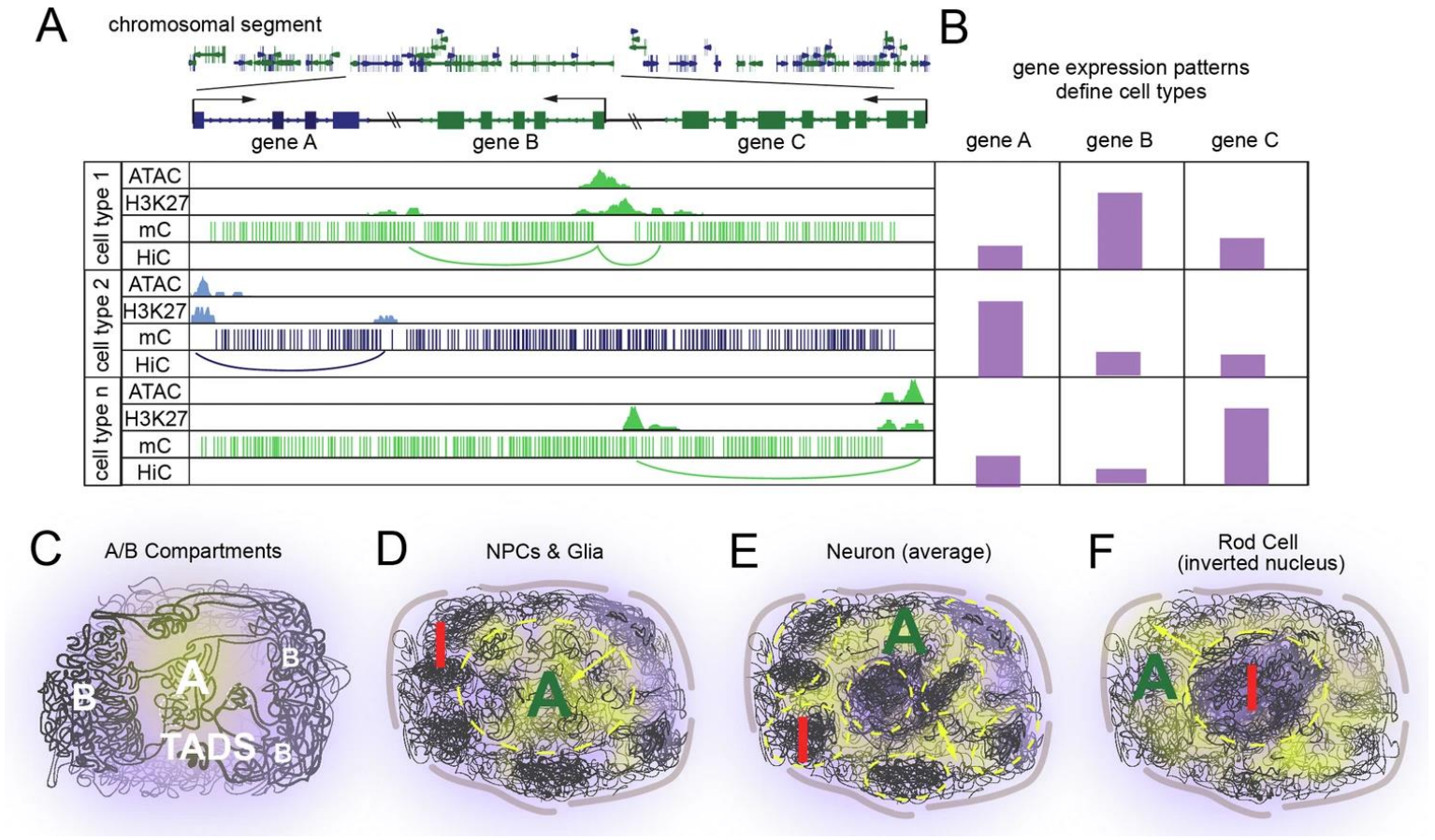
Sequences that isolate chromatin domains so that effects in one domain does not propagate



Overview of the three fundamental classes of regulatory elements in the genome: promoters, enhancers and boundary elements. The elements are indicated with a grey bar, and their bound factors and chromatin modifications when in an active state are shown. The transcription factors (TFs) in dark green represent ubiquitous TFs, whereas the TFs in light green represent tissue-specific and developmental stage-specific TFs. CTCF, CCCTC-binding factor*; H3K4me1, histone H3 monomethylated at Lys4; H3K4me3, histone H3 trimethylated at Lys4; H3K27ac, histone H3 acetylated at Lys27; PIC, pre-initiation complex; Pol II, RNA polymerase II. (Oudelaar & Higgs, Nature Rev Gen. 2020)

*CTCF is a DNA binding protein binding CCCTC repeats

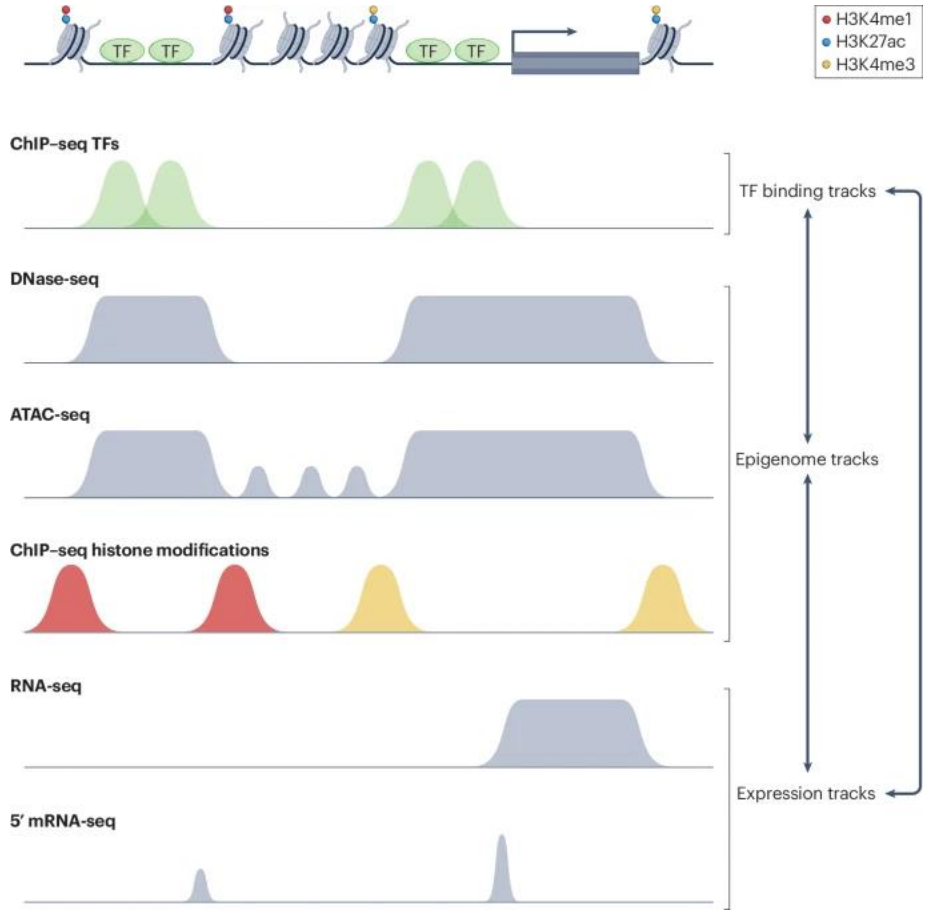
Differential regulation of gene expression defines cell type



A, B A combination of 4DN-type techniques is comparable in resolution with transcriptomics to define cell types. **A** Genome browser tracks for 3 cell types depicting ATAC, H3K27ac, DNA methylation, and HiC loops. **B** A cell type gene expression matrix for 3 cell types and 3 genes highlighting how cell types are defined by 4DN characteristics. **C–F** Neural precursors and glia (and most cells in the body) predominantly show inactive (I, red) chromatin localized at the nuclear exterior while active (A, green) chromatin is located centrally. Neurons do not seem to follow this rule as strictly, with the extreme example of nocturnal photoreceptors being complexly inverted. (Logeman et al.,2025 Mol Psy).

What data can be used to train deep learning models to predict gene transcription ?

Binding of transcription factors (TFs) influences gene expression and can be measured with chromatin immunoprecipitation followed by sequencing (ChIP–seq)[129]. Regulatory elements often reside in open chromatin, which can be measured with DNase I hypersensitive sites sequencing (DNase-seq) or assay for transposase-accessible chromatin with sequencing (ATAC-seq)[130]. Histone modifications that correlate with gene expression patterns can also be measured with ChIP–seq[131]. RNA sequencing (RNA-seq) or 5'mRNA sequencing (5'mRNA-seq) provides measurements of gene expression. 5'mRNA-seq techniques include cap analysis of gene expression[132], global run-on with cap capture[133] or transcription start site sequencing[134], which measure the 5'-capped ends of nascent RNAs (reviewed elsewhere[135]). Unlike RNA-seq, these techniques do not measure transcription termination, transcript stability or splicing events. All of these different data types can be used to train sequence-to-expression models There are extensive mutual causality links between these features (indicated by arrows): DNA sequence and chromatin accessibility dictate TF binding; some TFs can open up chromatin or recruit histone-modifying enzymes, and transcription is controlled by all of these features but conversely also affects chromatin states



https://doi.org/10.1038/s41576-025-00841-2 Nature Rev. Genetic

AlphaGenome Model

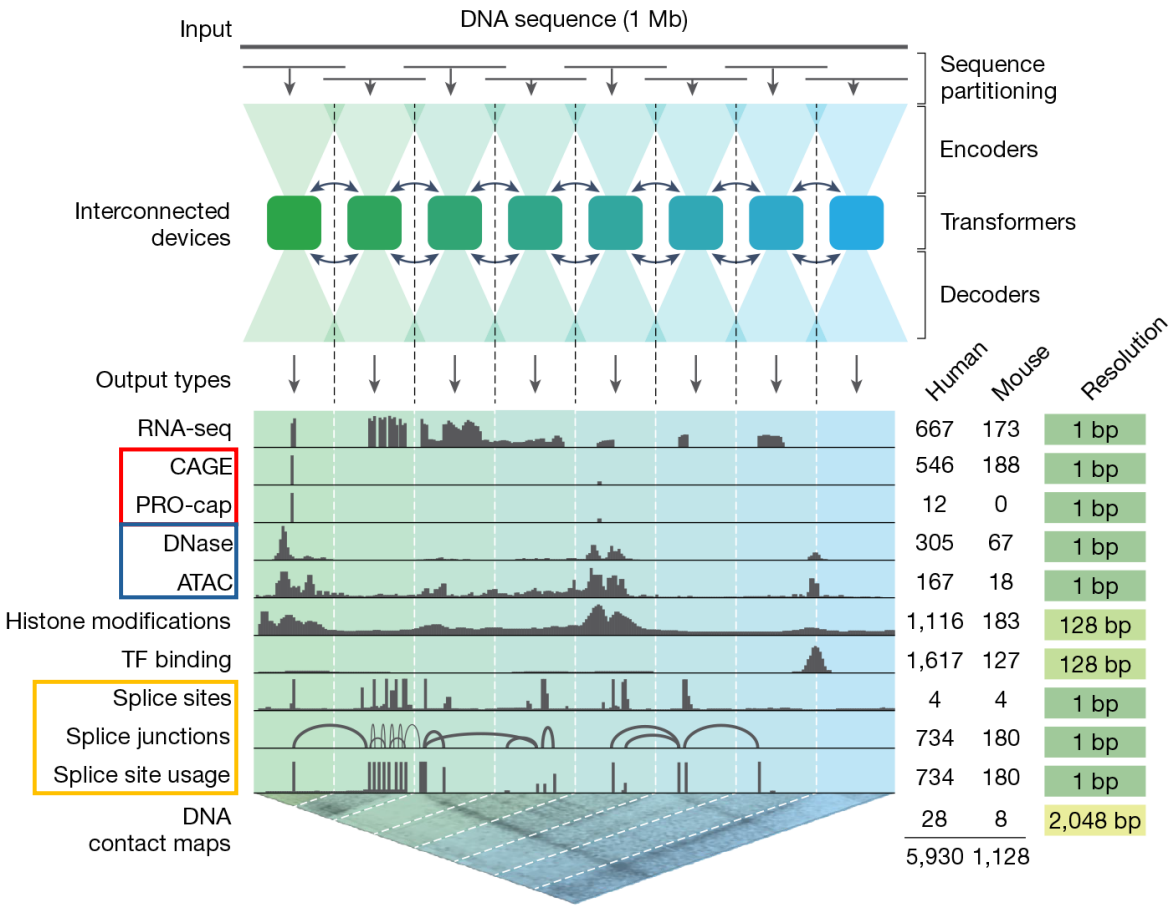


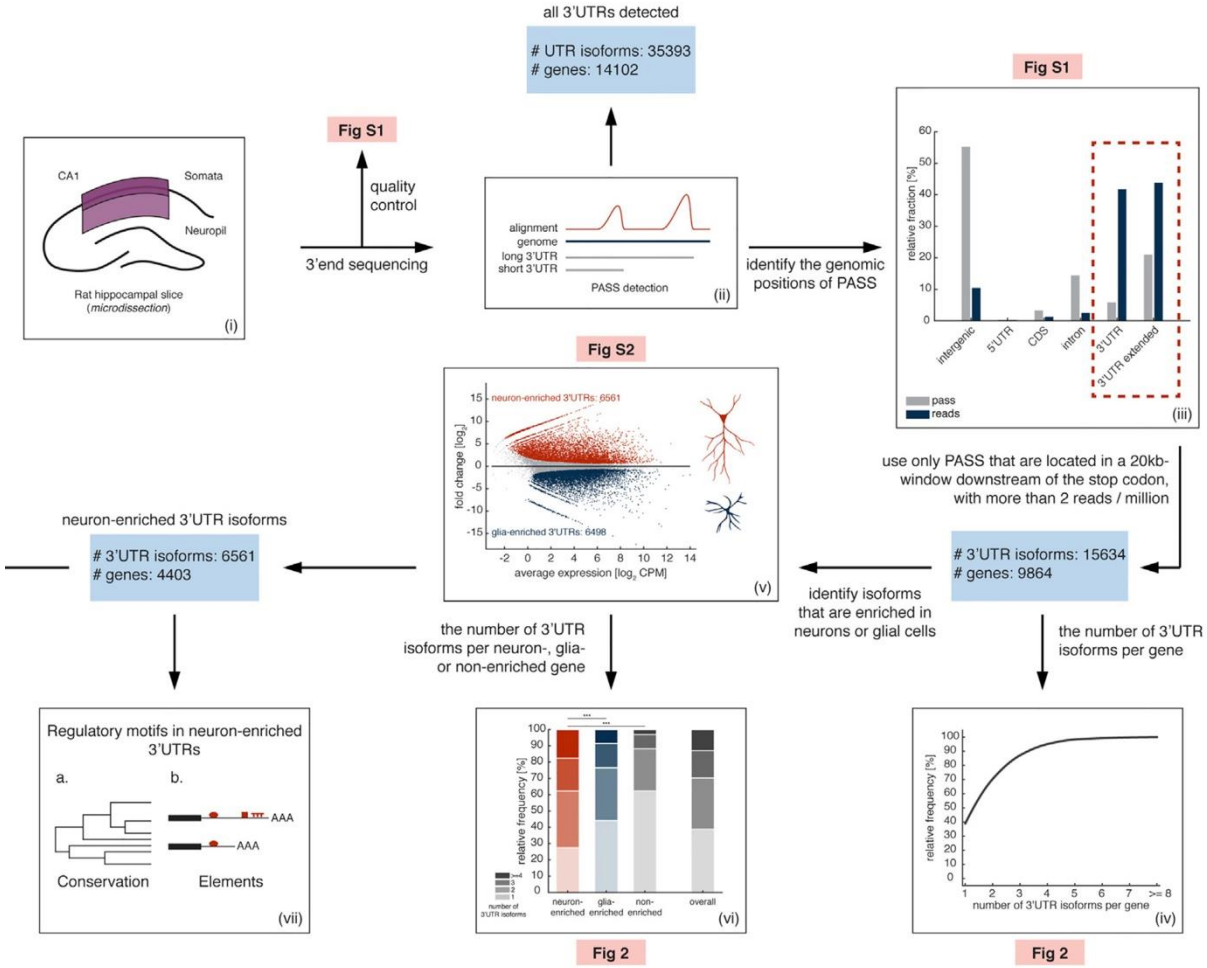
Fig. 1 | AlphaGenome model architecture, training regimes and comprehensive evaluation performance. a, Model overview. AlphaGenome processes 1 Mb of DNA sequences and species identity (human/mouse) to predict 5,930 human or 1,128 mouse genome tracks across diverse cell types and 11 output types at specific resolutions (far right). Computation leverages sequence parallelism, breaking the 1 Mb of DNA sequence into 131-kb chunks processed across devices. The core architecture features a U-Net-style design comprising an encoder (downsampling the sequence), transformers with inter-device communication and a decoder (upsampling), which feed into task-specific output heads at their respective resolutions (detailed in Extended Data Fig. 1).

Nature | Vol 649 | 29 January 2026

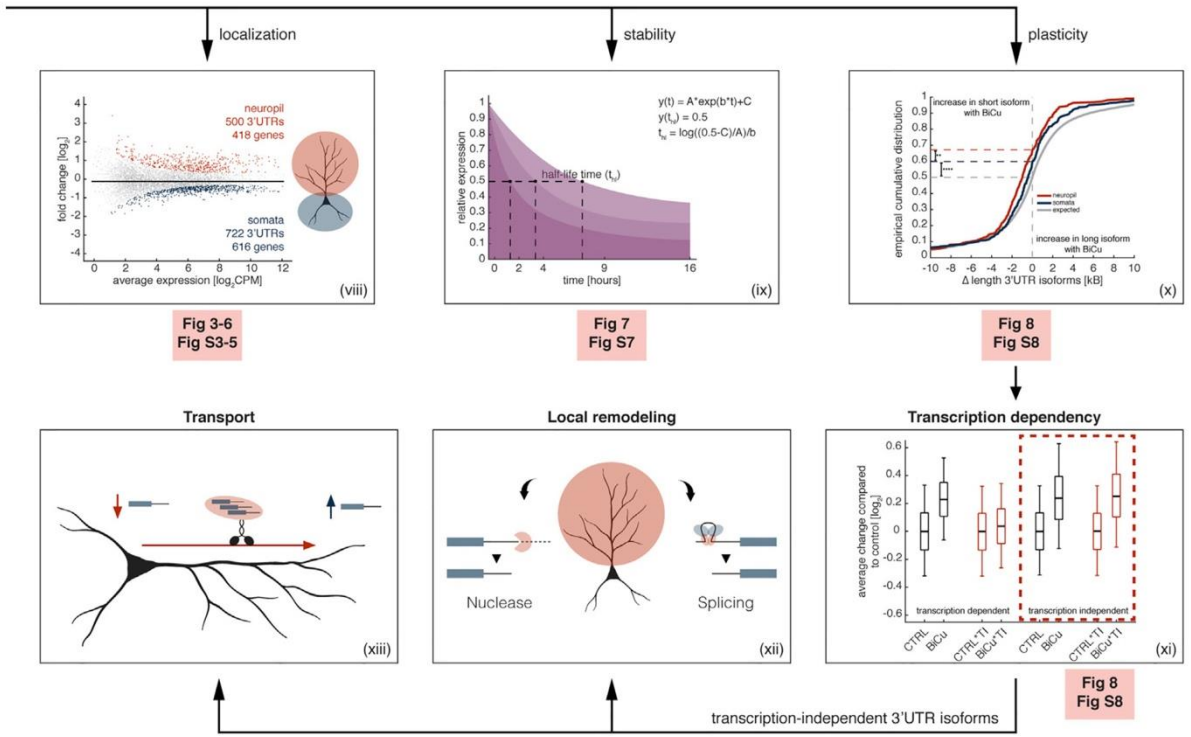
The other languages of RNA

5’ and 3’ Untranslated Regions (UTRs)

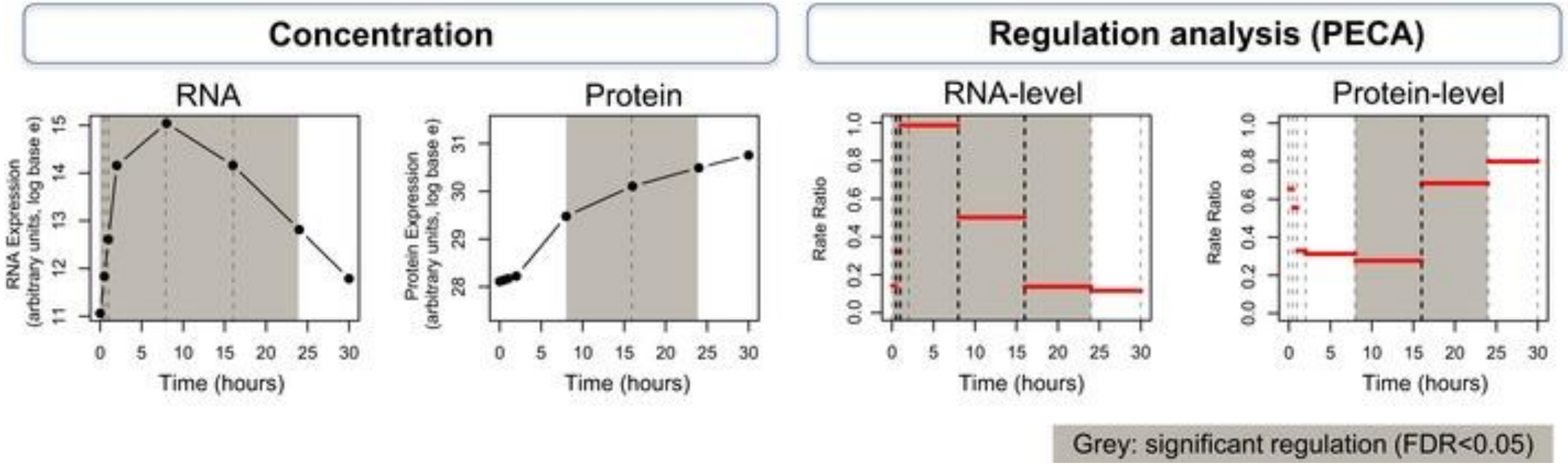
UTRs affect the localization
Translation and stability of their respective mRNAs.



Alternative 3'UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments



mRNA Levels are not directly related to protein levels



The example shows the chaperone GRP78, a key ER stress protein. mRNA and protein concentrations are shown on the left; PECA results are shown on the right for RNA and protein level, respectively. Intervals with significant regulation as determined by PECA are gray shaded (FDR < 0.05). The value of PECA is illustrated at the 16-h time point at which mRNA concentration decreases, while protein concentration still rises. PECA highlights that there is a significant RNA- and protein-level regulation around this time point—a signal that would otherwise likely have been overlooked.

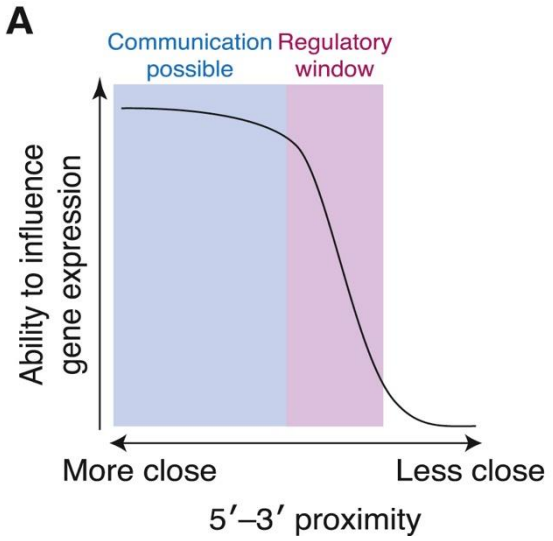
. 2016 Jan 20;12(1):855. doi: 10.15252/msb.20156423

For details on Protein Expression Control Analysis (PECA):
J Proteome Res. 2013 Nov 26;13(1):29–37. doi: 10.1021/pr400855q

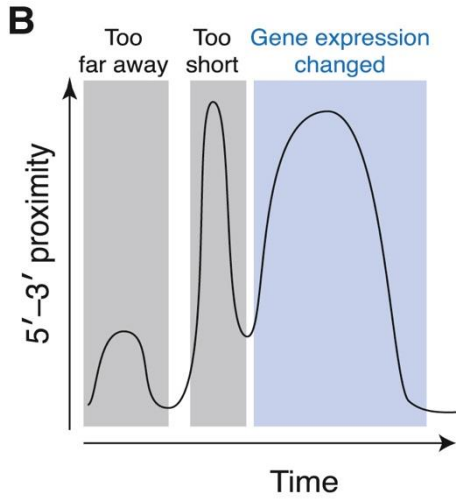
Not restricted to UTRs
mRNA form complex structures guided by Nucleotide interactions

The RNA sequence is not only a set of instruction to make proteins
It is also a mode of regulation

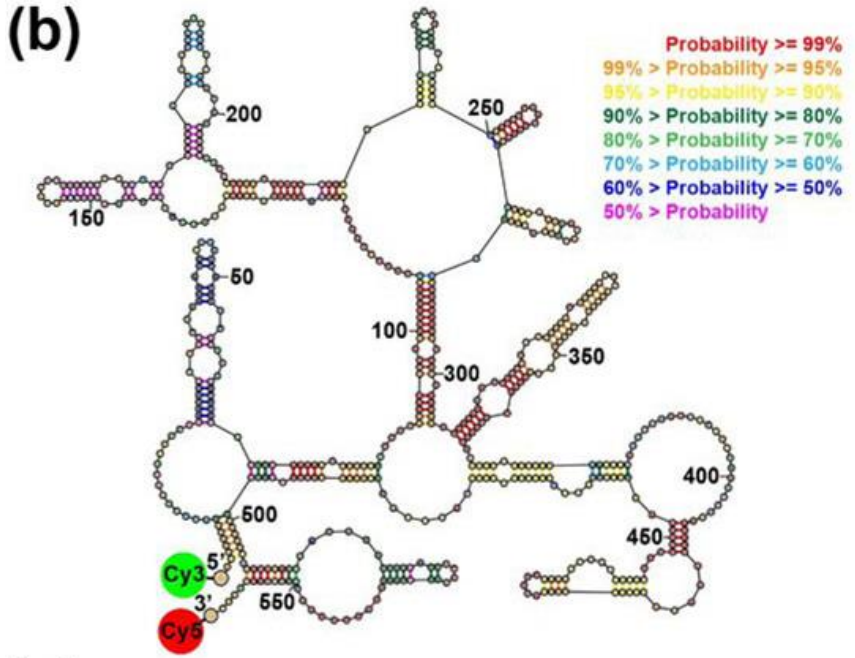
Proximity.

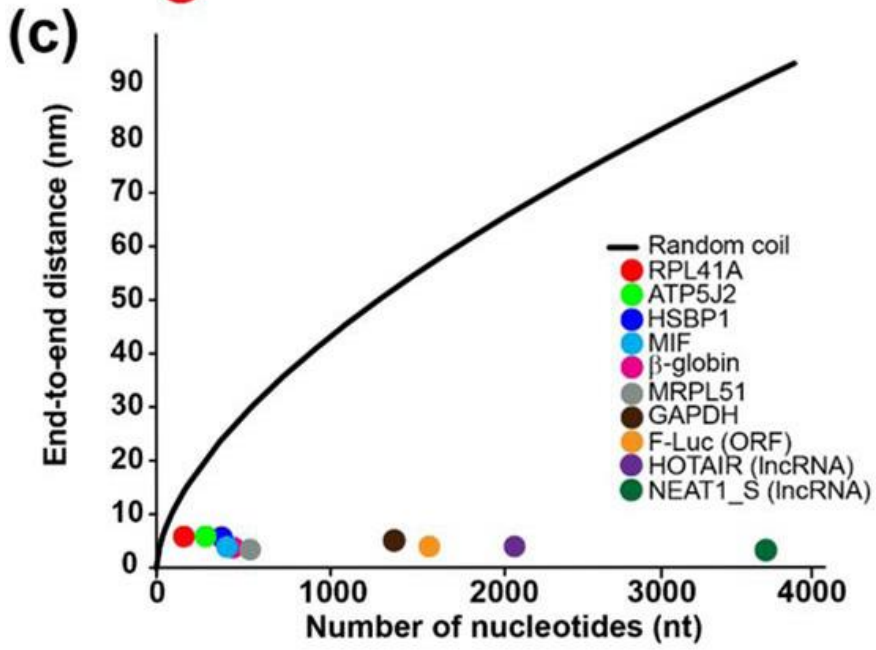


Duration of proximity



https://doi.org/10.1016/j.molcel.2018.10.047





https://www.biorxiv.org/content/10.1101/2020.04.29.069203v1.full

Three main mode of regulations

1- Non-specific regulators of protein synthesis.

2- Specific RNA binding proteins (RBPs); regulate stability, translation, transport.

3- Small regulating RNAs (e.g. miRNAs.)

Regulation of mRNAs by RBP and miRNA is critically determined by the formation of specific secondary structures in the mRNAs

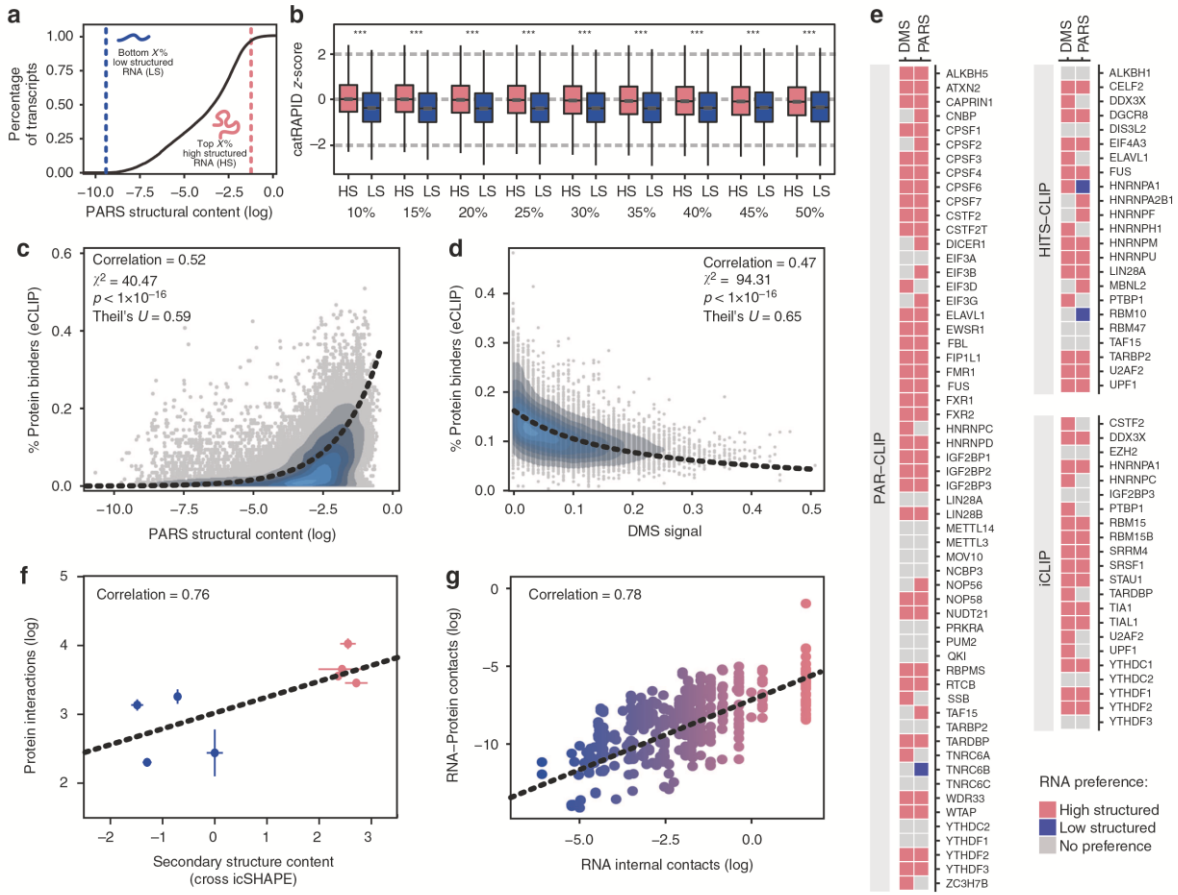
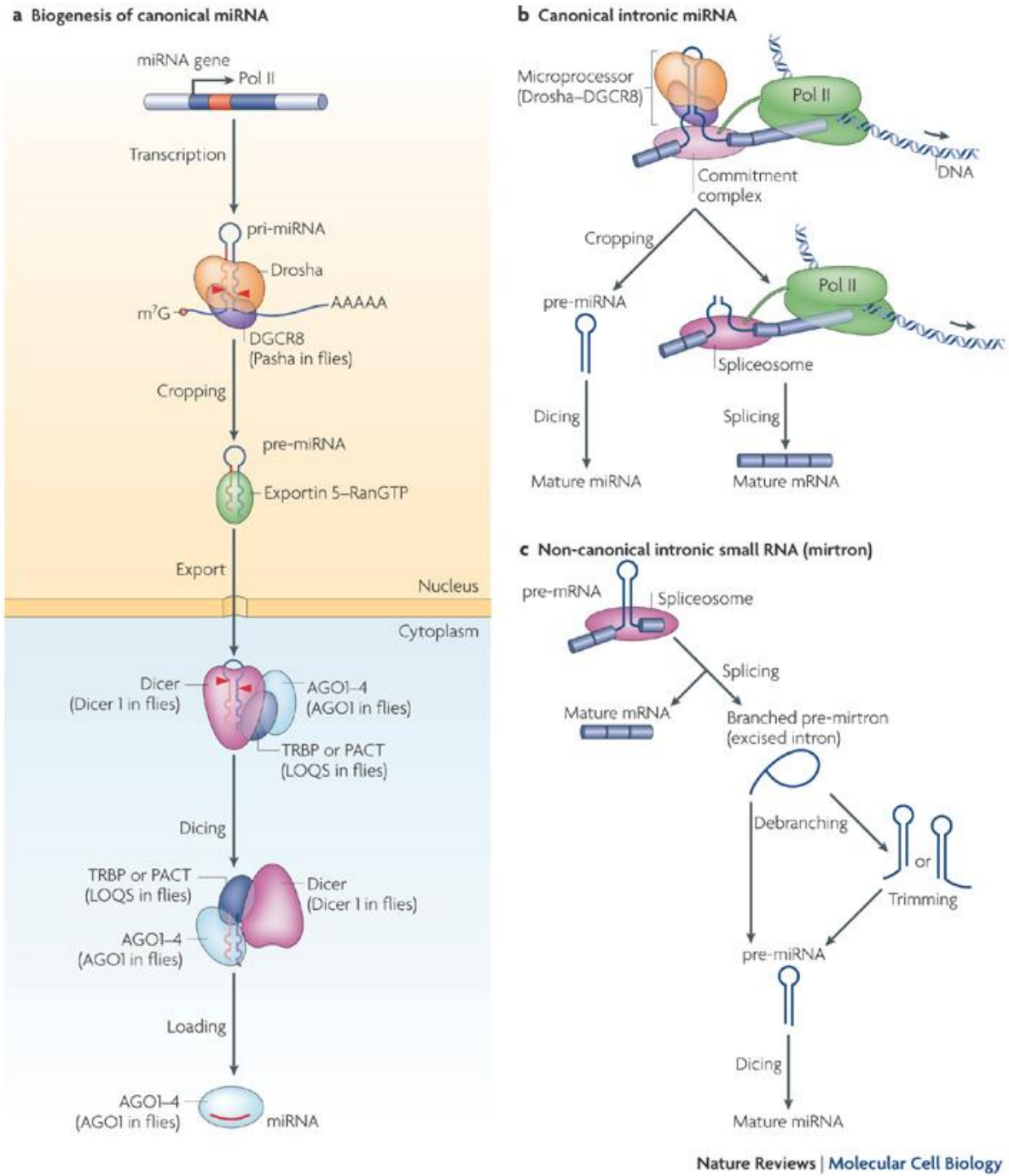


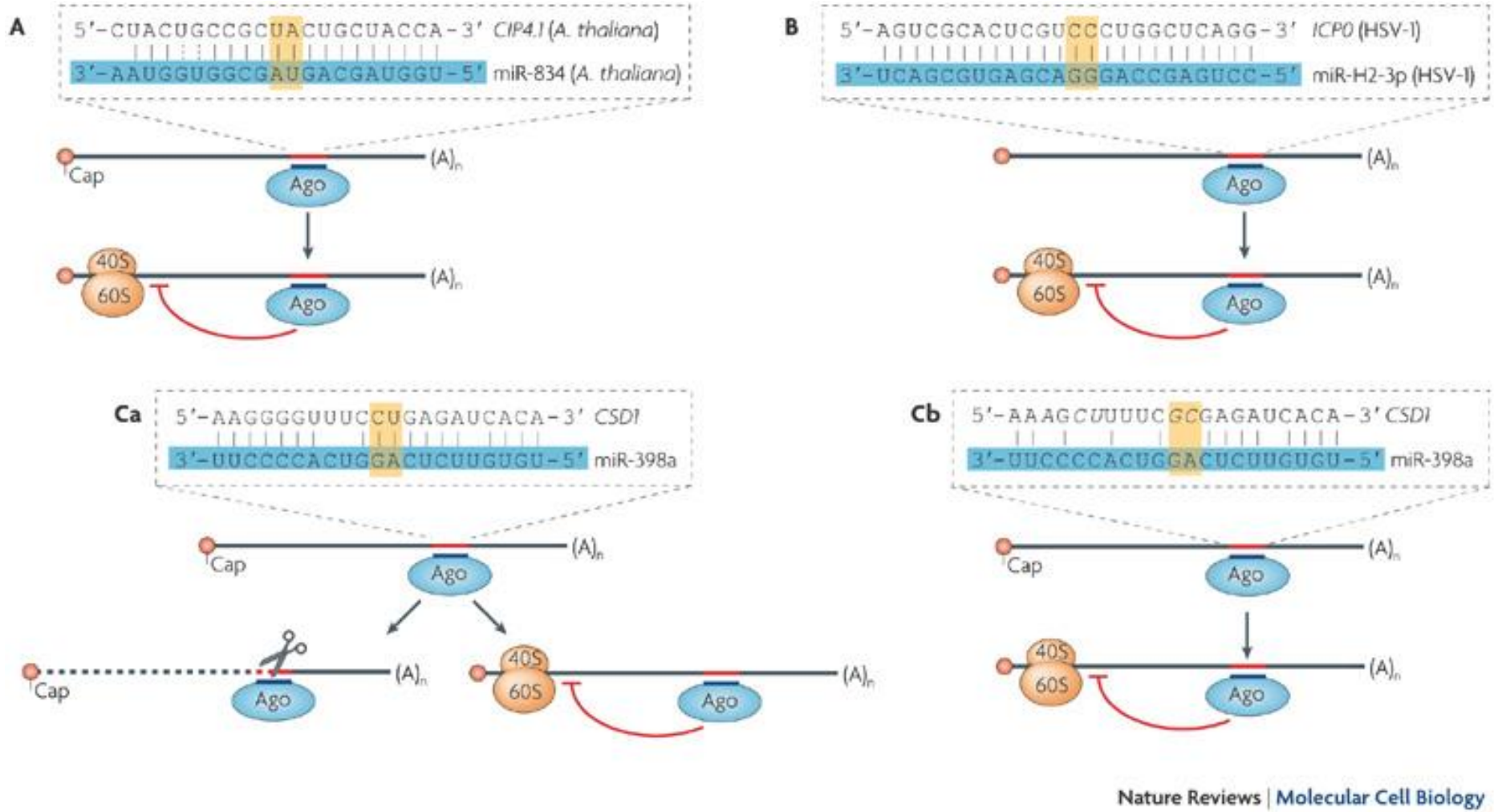
Fig. 1 The amount of protein structure correlates with the number of interactions. **a** Cumulative distribution function (CDF) for the secondary structure content of all human RNAs measured by parallel analysis of RNA structure (PARS)[8,69]. Vertical lines indicate a certain fraction (*X*%) of RNAs with the lowest secondary content (LS; blue) and the same fraction with the highest secondary content (HS; pink). **b** catRAPID predictions of protein interactions with human RNAs ranked by structural content measured by PARS (118 RNA-binding proteins (RBPs) for which enhanced crosslinking and immunoprecipitation (eCLIP) information is also available)[31]. The fractions 10%, 15%, ..., 50% refer to the comparison between equal-size HS and LS sets. The results indicate that catRAPID is able to distinguish HS and LS groups significantly and consistently through the different fractions (p value $<10^{-16}$; Kolmogorov–Smirnov (KS) test). The boxes show the interquartile range (IQR), the central line represents the median, the whiskers add 1.5 times the IQR to the 75 percentile (box upper limit) and subtract 1.5 times the IQR from the 25 percentile (box lower limit). s.d. is shown. **c** Relationship between number of protein interactions (eCLIP) and structural content measured by PARS[30]. The fitting line corresponds to the formula $y = \exp(\alpha + \beta x)$, where $\alpha = -0.75$; $\beta = 0.67$; p value estimated with KS test. **d** Relationship between number of protein interactions and structural content measured by dimethyl sulfate modification (DMS)[9]. The fitting line corresponds to the formula $y = 1/(\alpha + \beta x)$, where $\alpha = 2.60$; $\beta = 87.36$; p value estimated with KS test. **e** Structural preferences of RBPs measured with three different CLIP techniques (photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP), high-throughput sequencing-CLIP (HITS-CLIP) and individual nucleotide resolution CLIP (iCLIP)). The colour indicates the RNA-binding preference of each protein: pink, high structured; blue, low structured; grey, no preference. **f** Correlation between structural content (CROSS predictions of icSHAPE experiments) and protein interactions of eight transcripts revealed by protein microarrays (Pearson's correlation). s.d. is shown. **g** Analysis of Protein Data Bank (PDB) structures containing protein–RNA complexes reveals a trend between protein (inter) and RNA (intra) contacts (196 different pairs; Pearson's correlation)

Micro RNAs



Nature Reviews | Molecular Cell Biology

mIR Generaly stall the translation of mRNA or promote their degradation. However, there are evidence for other roles



Nature Reviews | Molecular Cell Biology

The regulation of mRNA translation and stability is an additional layer of regulation.

-It can allow for subcellular differences in gene expression.

-It can allow to prime a system by stocking RNA on stalled ribosomes for fast synthesis of proteins upon Changes of biological needs.

RNA levels are a function of:
-Gene transcription
-RNA stability
Presence of mRNA does not mean
the corresponding protein is made

Protein levels are a function of
-mRNA translation
-protein degradation
Presence of protein does not mean
It is active.

Neither mRNA level nor protein level is a sure predictor of functional response.
They are just indications that the system response status has changed.
Determination of a functional state requires a combination of datasets.
However, single datasets are a good shortcut sometimes.
It all depends on what the question is.

How do you think this can affect ML model predictions?
Are non-ML models possible?