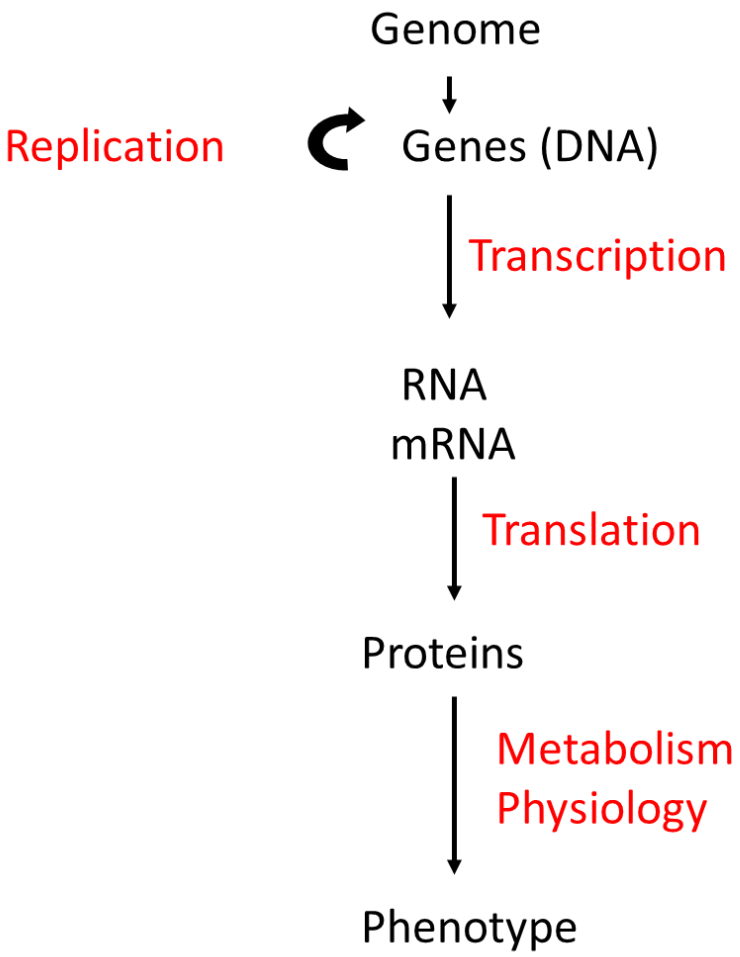
CENTER OF
PLANT SCIENCES

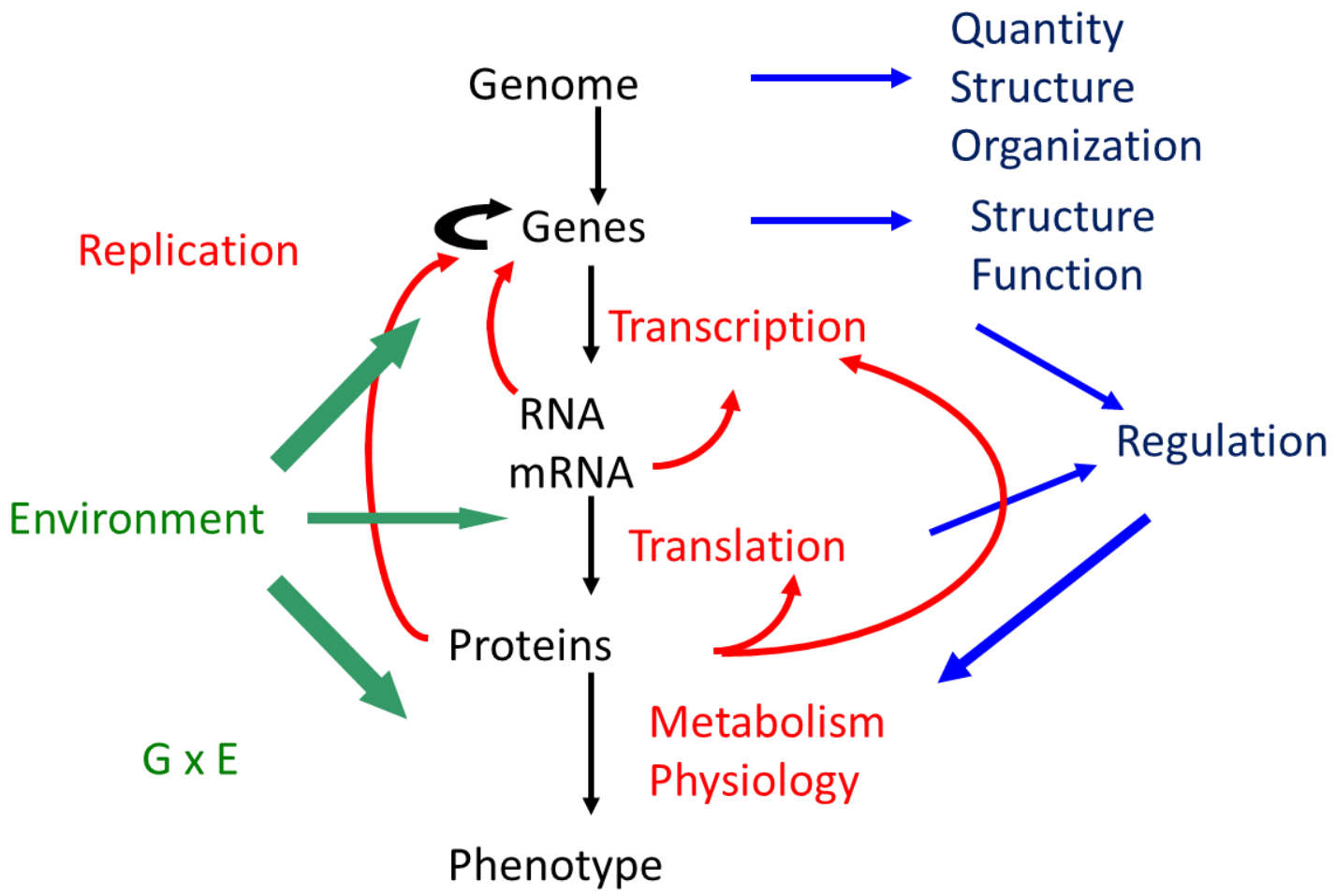
Sant'Anna
Scuola Universitaria Superiore Pisa

Advanced Genomics

The dogma of biology

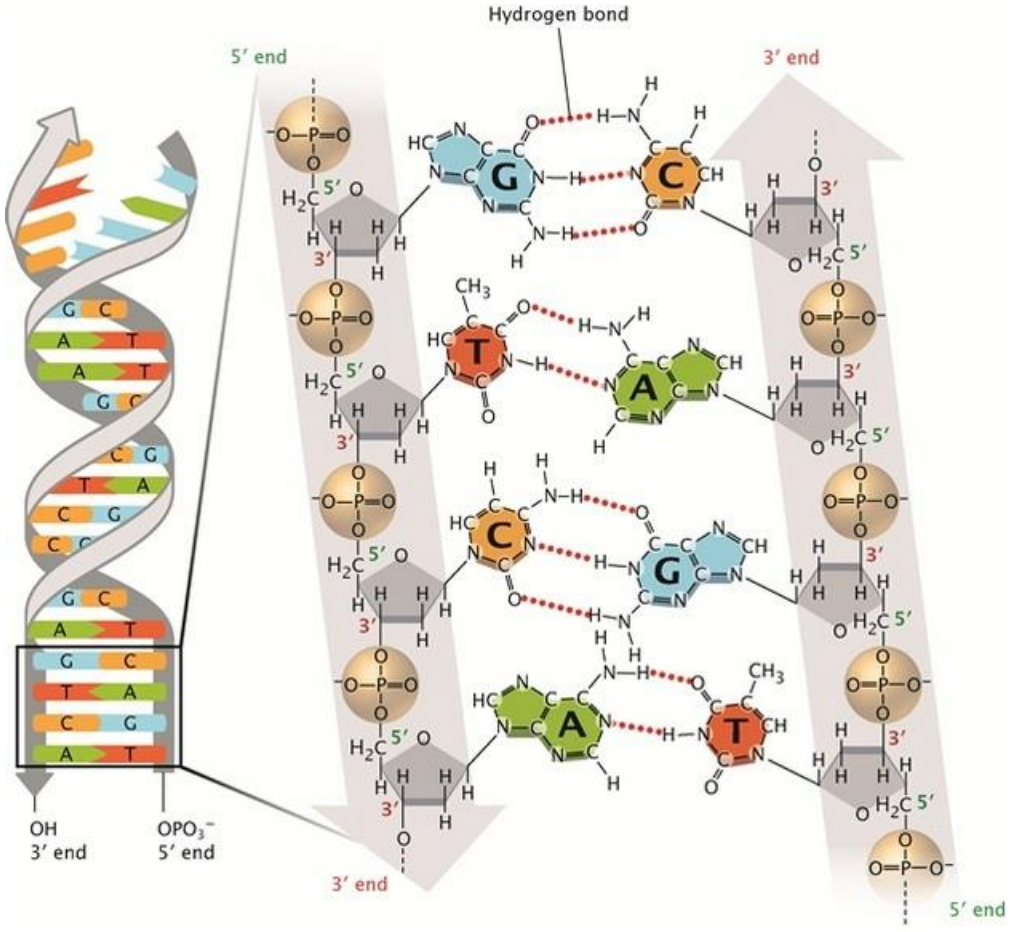
The central dogma of molecular biology states that "DNA makes RNA and RNA makes protein" Francis Crick (1957)

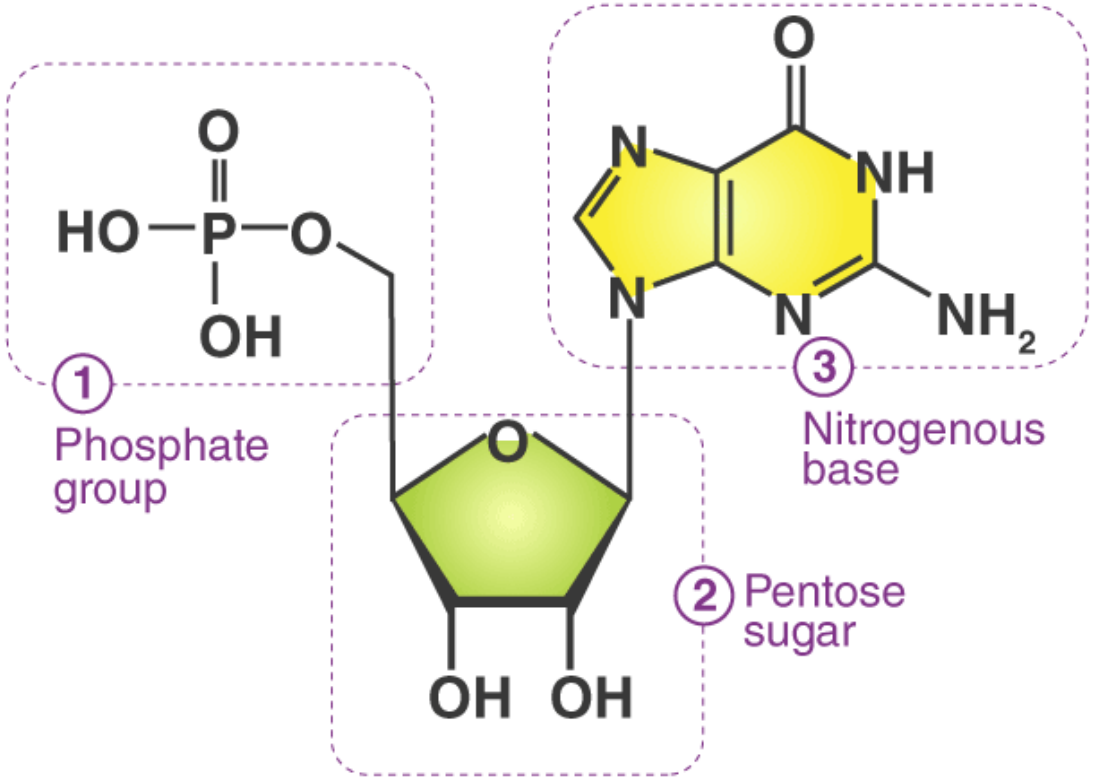




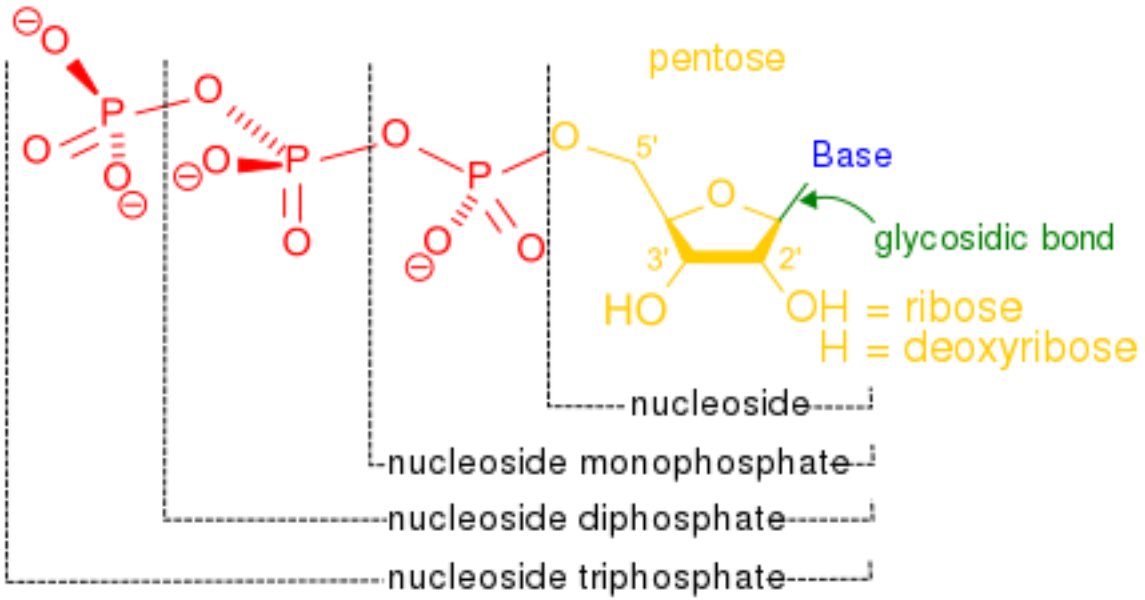
DNA

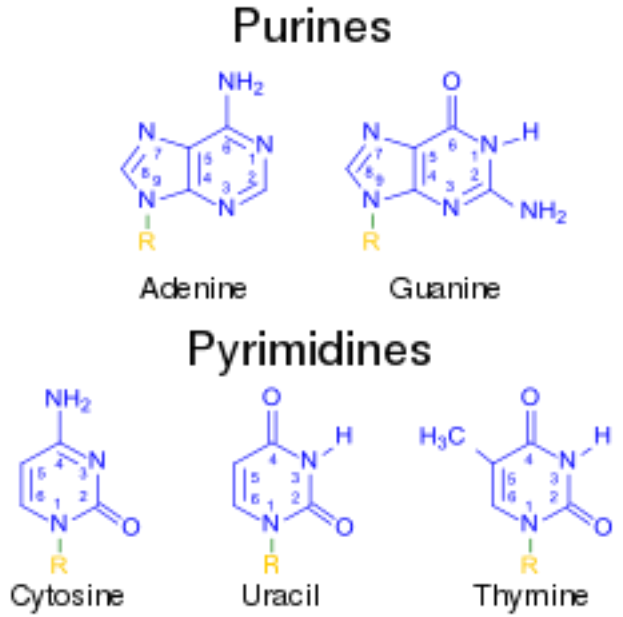
- Made of two linked strands (POLYMERS) paired in a double helix
- Backbone of sugar (deoxyribose) alternated with phospate
- Attached to each sugar is one of four bases: A, C, T(U), G
- Bases bond in pairs with hydrogen bonds: A & T(U), C & G
- The sequence of bases on a strand is what carries the information
- Bases of RNA are the same except of the fact that Thymine is replaced with Uracil (U)



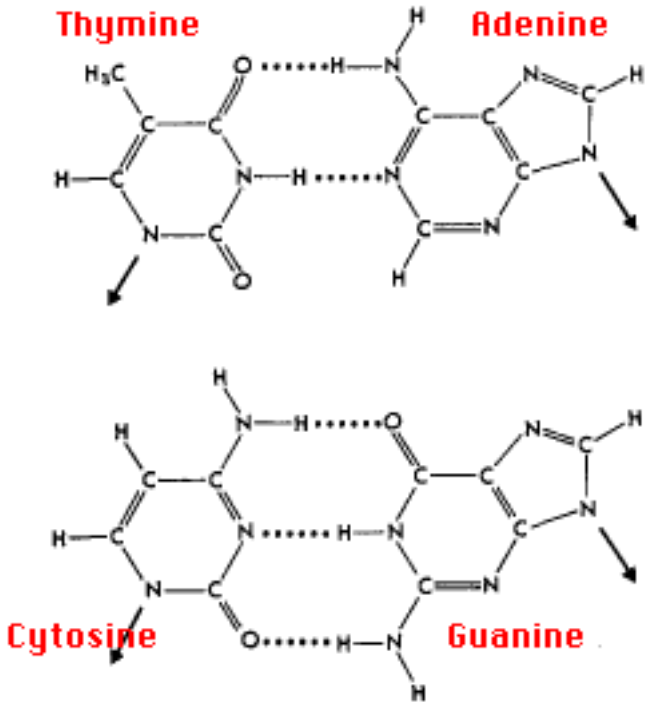


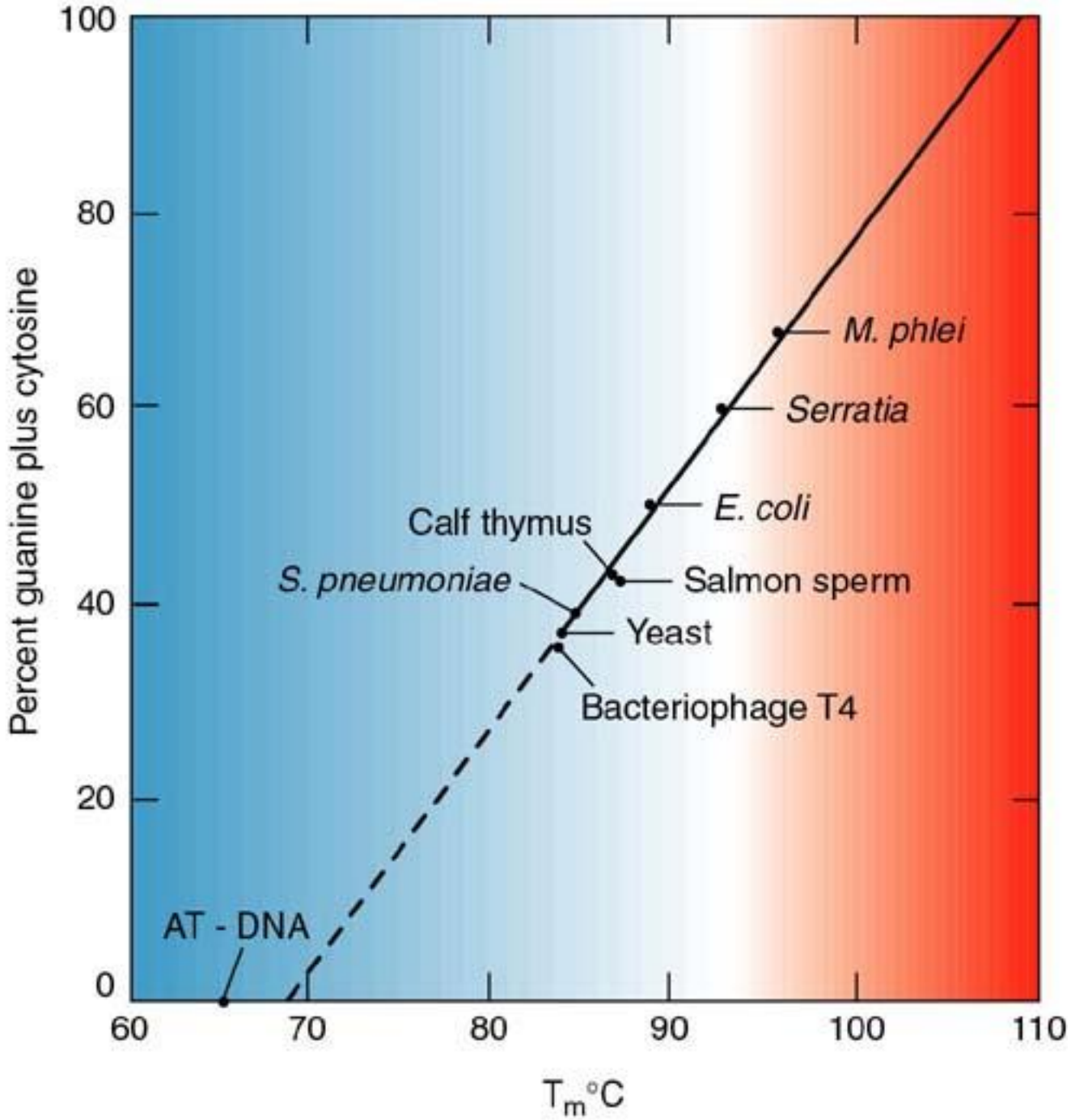
- The purines (A,G) have a two-ringed structure consisting of a nine-membered molecule with four nitrogen atoms
- The pyrimidines (C,T,U) only have one single ring, which has just six members and two nitrogen atoms





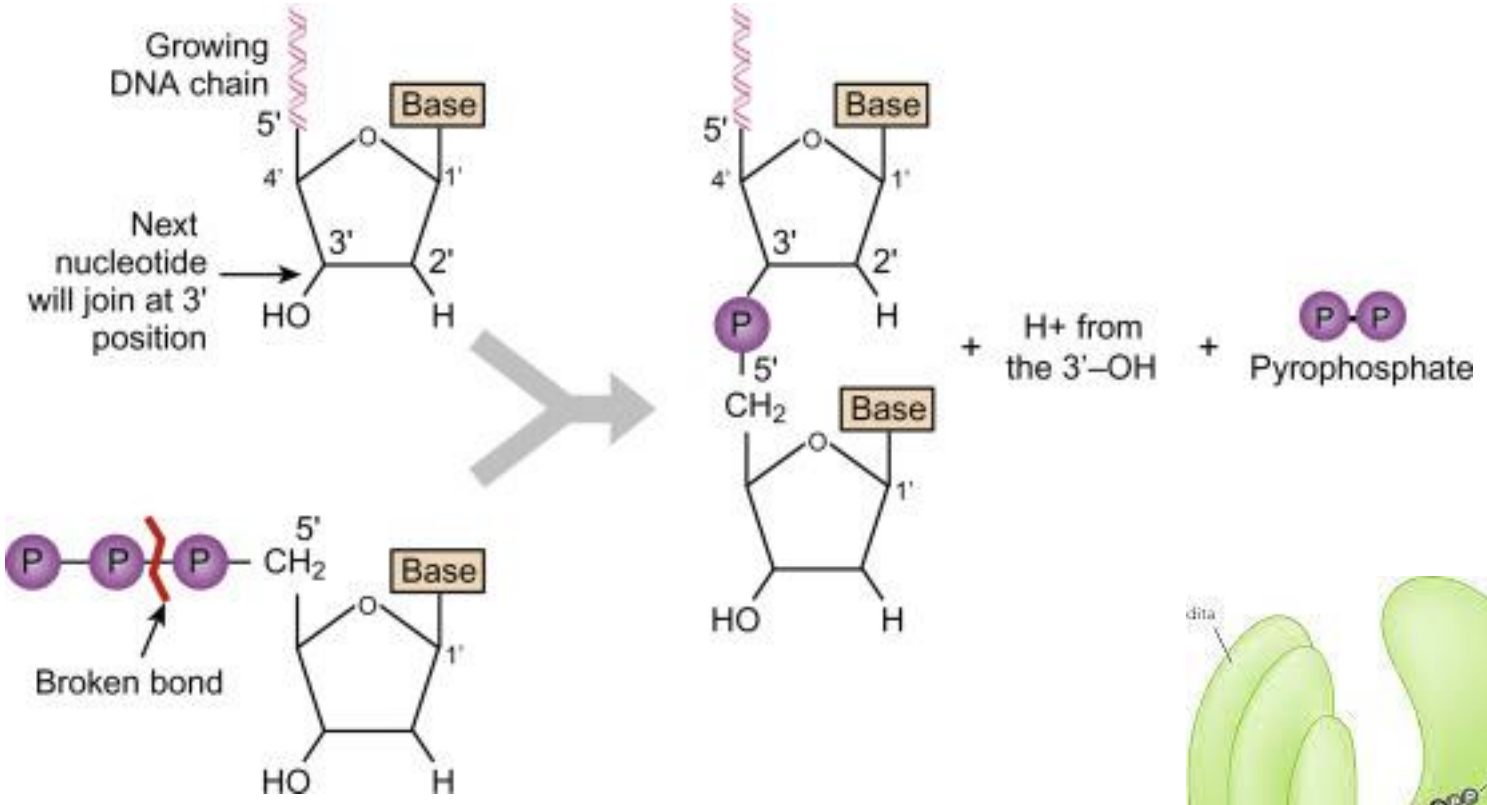
Sequence and energy content

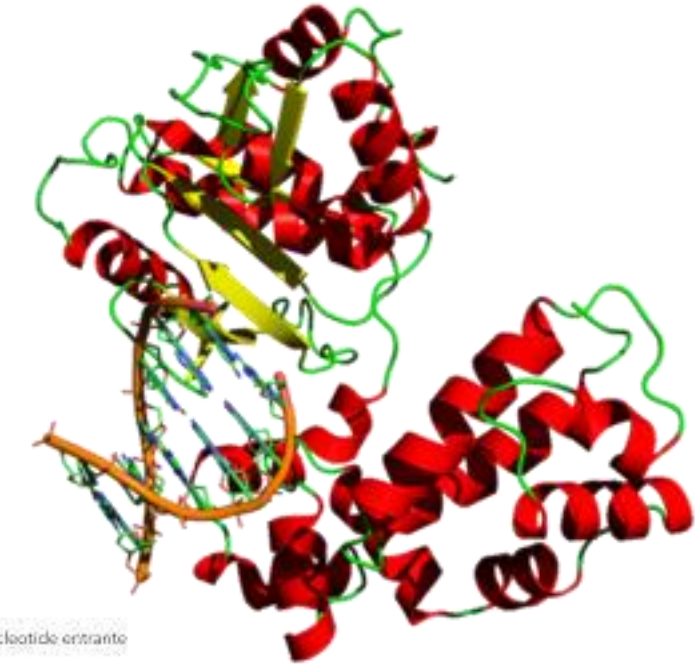


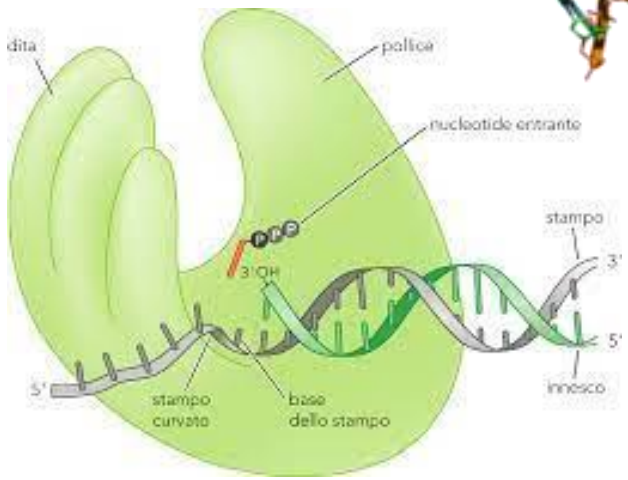


DNA synthesis works directionally adding a dNTP to the OH end of the polymer

There's a bunch of different polymerases doing it

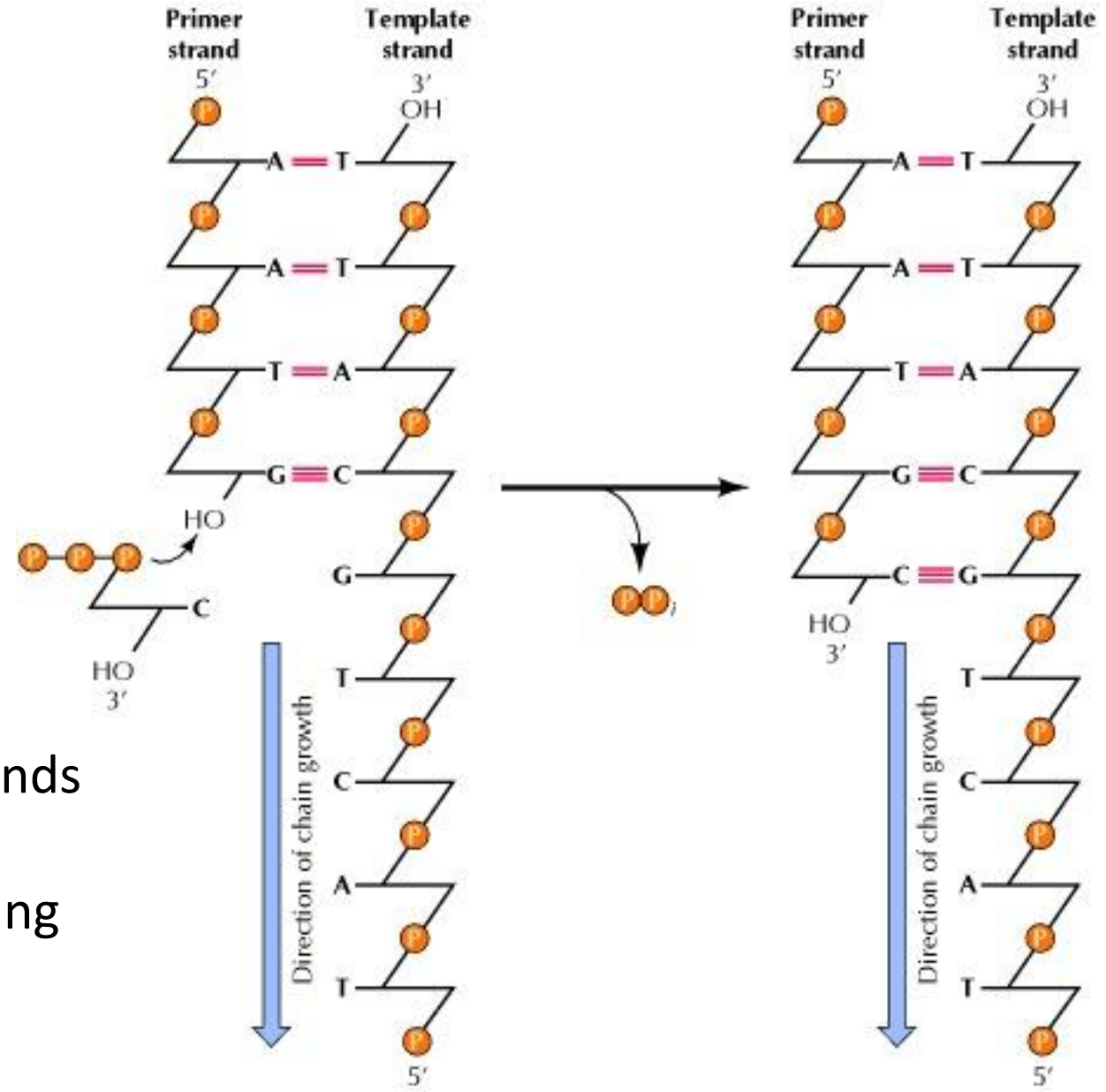






deoxynucleoside triphosphate + $DNA_n \rightleftharpoons$ pyrophosphate + DNA_{n+1}

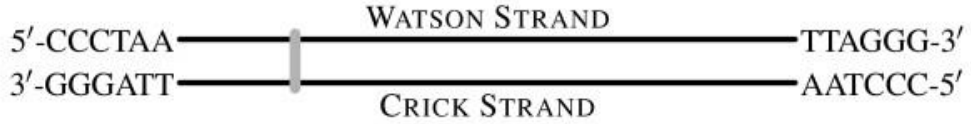
- You just need one sequence (+) to derive its complement
- DNA replication in vivo works **ALWAYS** from 5’ to 3’
  - The polymerases that assemble new strands generally rely on the energy produced by breaking nucleoside triphosphate bonds to attach new nucleoside monophosphates to the 3’-hydroxyl (−OH) group

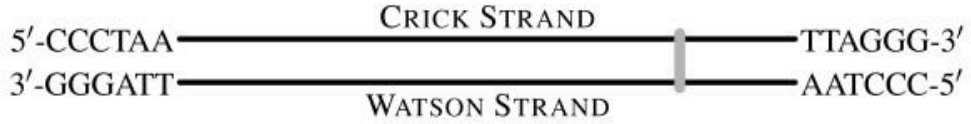


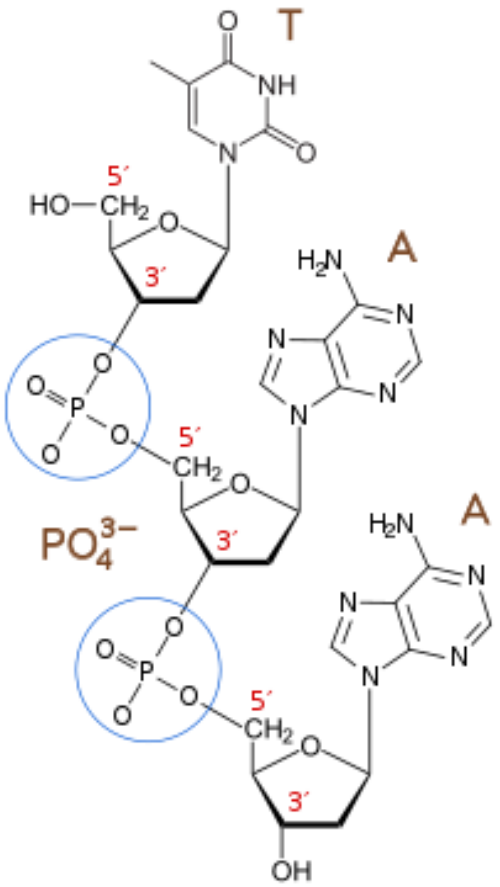
The relative positions of structures along strands of nucleic acid, including genes and various protein binding sites, are usually noted as being either *upstream* (towards the 5’-end) or *downstream* (towards the 3’-end)

Directionality of DNA/RNA molecules

- 5′ end which contains a phosphate group attached to the 5′ carbon of the ribose ring
- 3′ end typically is unmodified from the ribose -OH
- In a DNA double helix, the strands run in opposite directions to permit base pairing between them

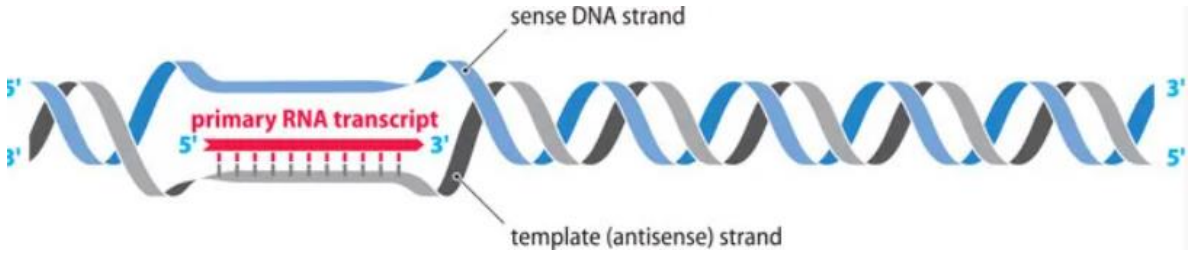






- When looking at genomes, the Watson strand is the strand of a chromosome that has its 5'-end at the short-arm telomere and its 3'-end at the long-arm telomere. The Watson strand is stored as the reference (+), the Crick strand is the other (-)

RNA

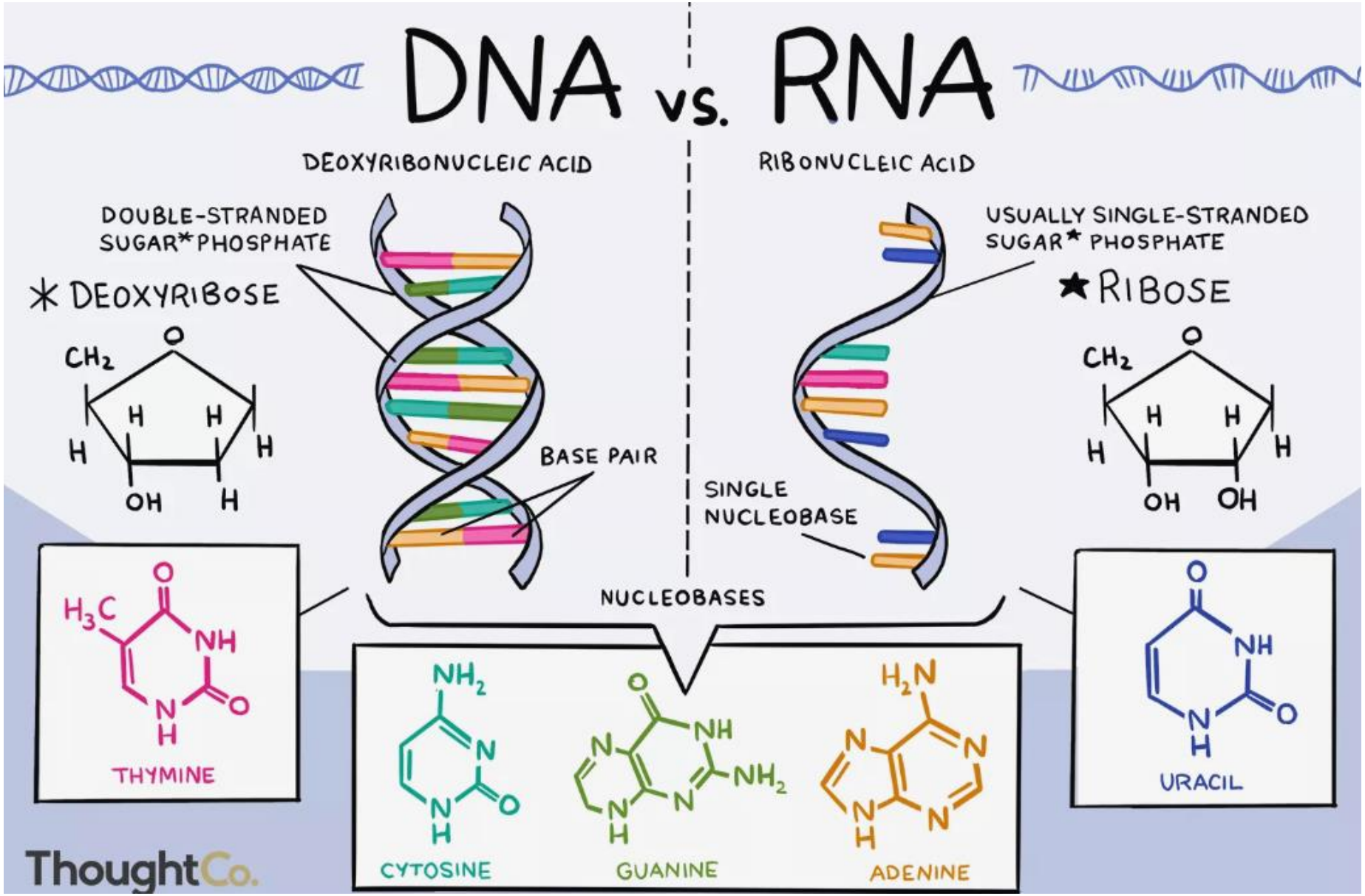


- One of the two DNA strands is used as a template for synthesizing an RNA copy whose sequence is complementary to the selected template DNA strand
- The RNA is then processed to either one of two broad classes:
  - Noncoding RNA (endpoint)
  - Coding RNA
- Coding RNAs are then moved ahead into proteins
- Proteins do their things and contribute to organismal function

The **transcriptome** is the initial product of genome expression

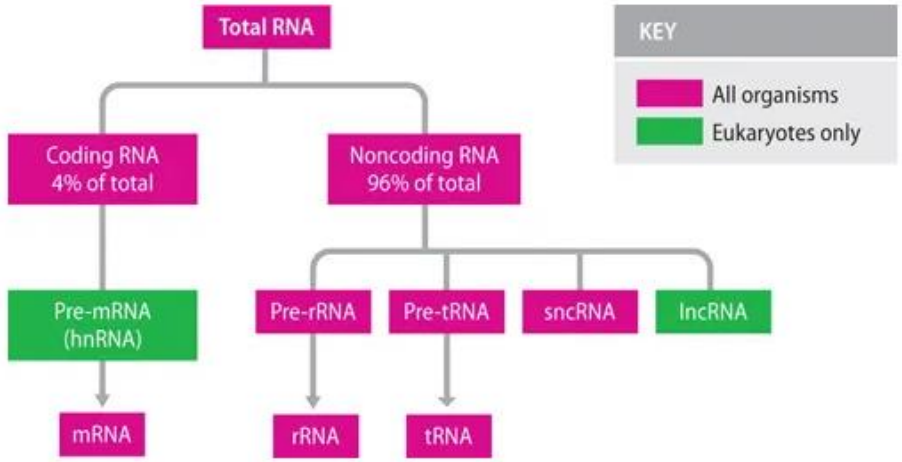
- RNA molecules are synthesized by transcription
- RNA is similar to DNA, but for two things:
  - The sugar in the RNA is a ribose
  - Uracil replaces thymine
- RNA also contains 3'-5' phosphodiester bonds, but these bonds are less stable than those in the DNA (due to the hydroxyl group at the 2' of the sugar) → RNAs are typically few thousands Nts in length
- Most often than not, RNA are single stranded

- A typical mammalian cell weight is 1% RNA (bacteria 6%)


DNA vs. RNA
DEOXYRIBONUCLEIC ACID
RIBONUCLEIC ACID
DOUBLE-STRANDED SUGAR* PHOSPHATE
* DEOXYRIBOSE
USUALLY SINGLE-STRANDED SUGAR* PHOSPHATE
★ RIBOSE
BASE PAIR
SINGLE NUCLEOBASE
NUCLEOBASES
THYMINE
CYTOSINE
GUANINE
ADENINE
URACIL
ThoughtCo.

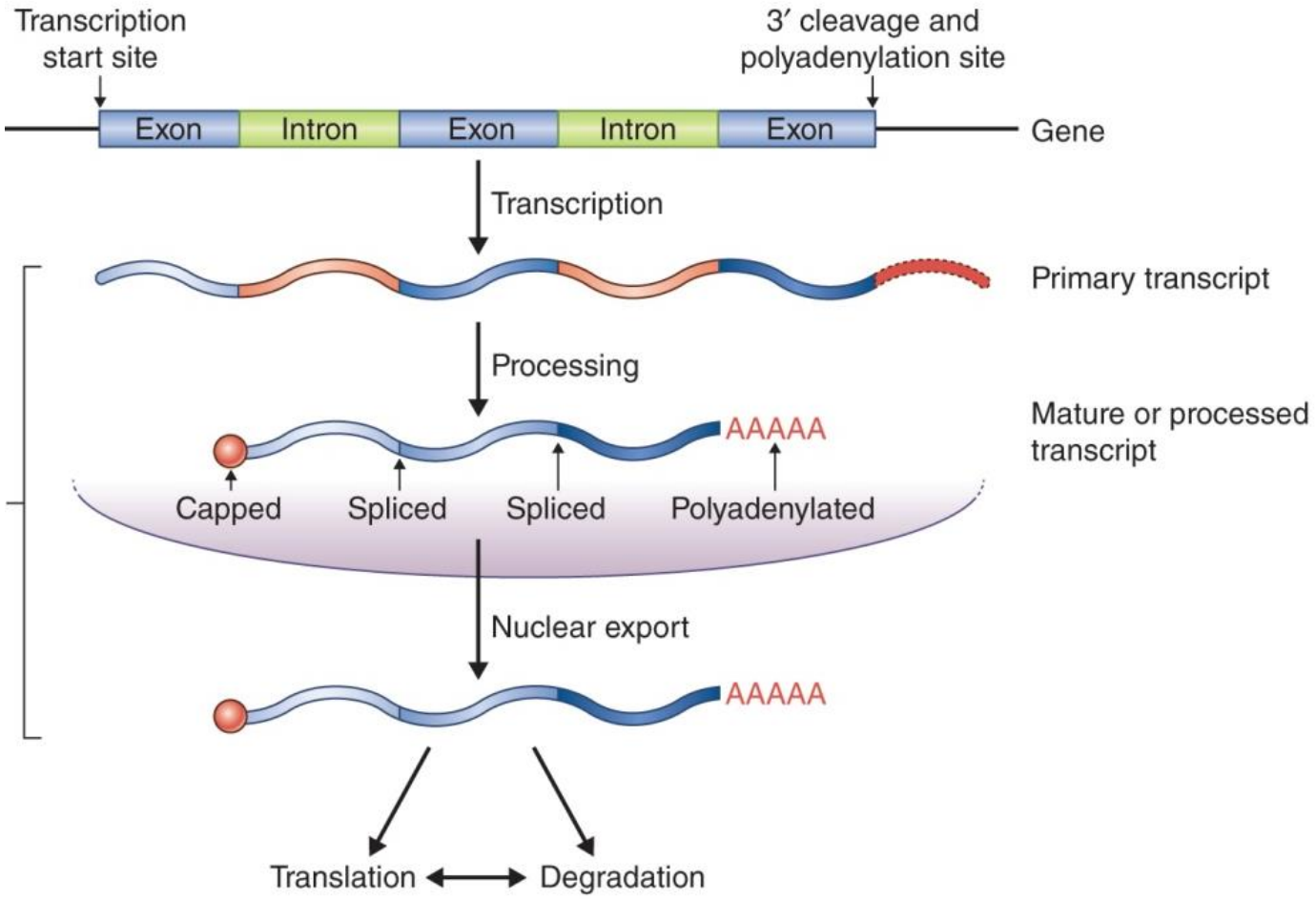
RNA comes in two flavours:

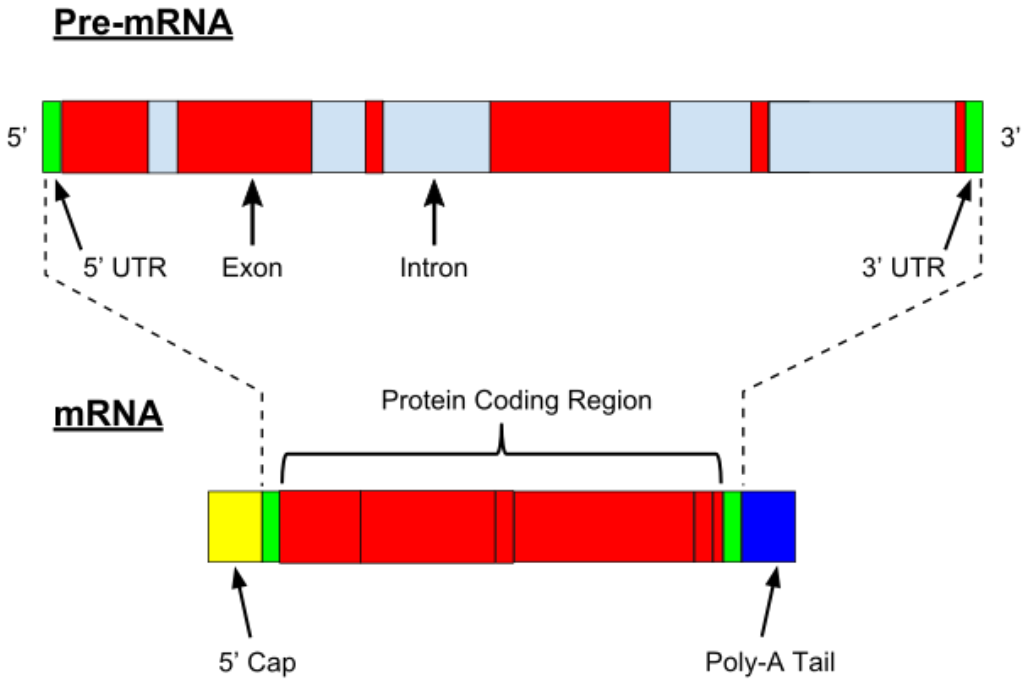
- Coding RNAs, or messenger RNA (mRNA)
  - About 6% of the total RNA in a given cell
  - Transcripts of protein coding genes
  - Degraded shortly after synthesis
- Non-coding RNAs
  - Ribosomal RNAs (rRNAs): about 80% of the total RNA in a cell, are a component of ribosomes
  - Transfer RNAs (tRNAs): needed to carry AA in protein synthesis
  - Short non coding RNAs (sncRNAs)
  - Long non coding RNAs (lncRNAs)

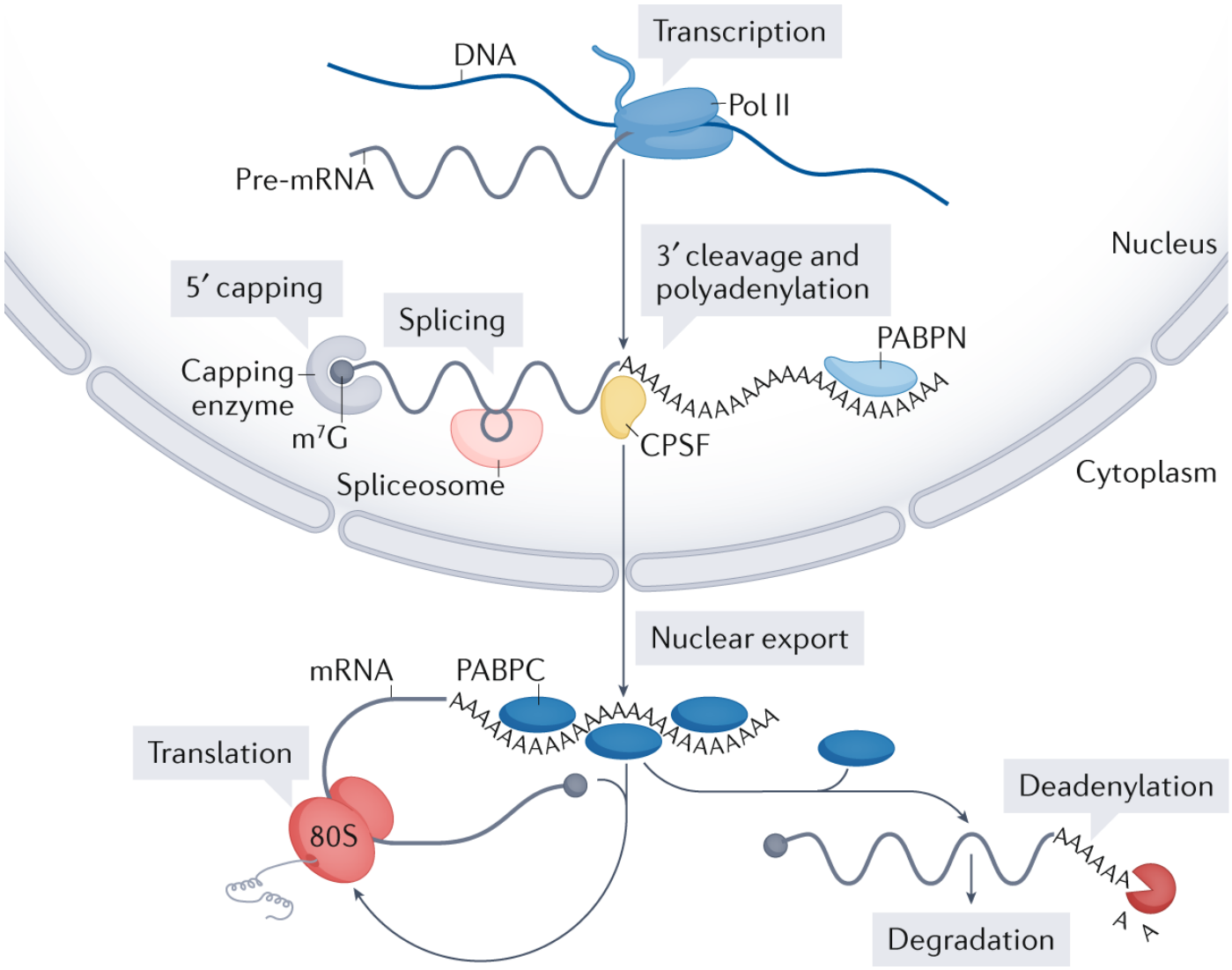


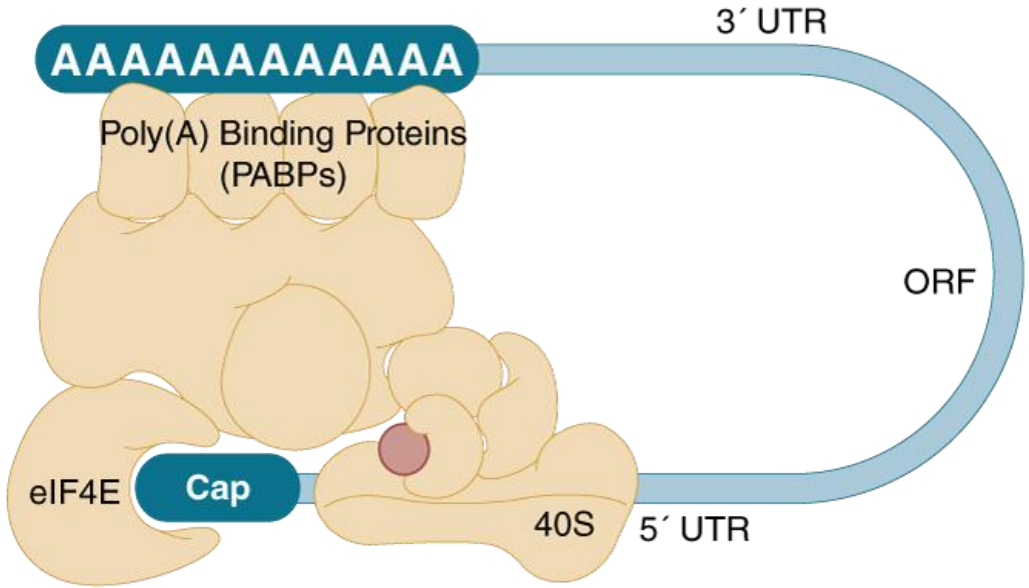
Coding RNA is the most well known RNA species

- It goes through modification steps (splicing, capping) to alter its function and/or improve its stability





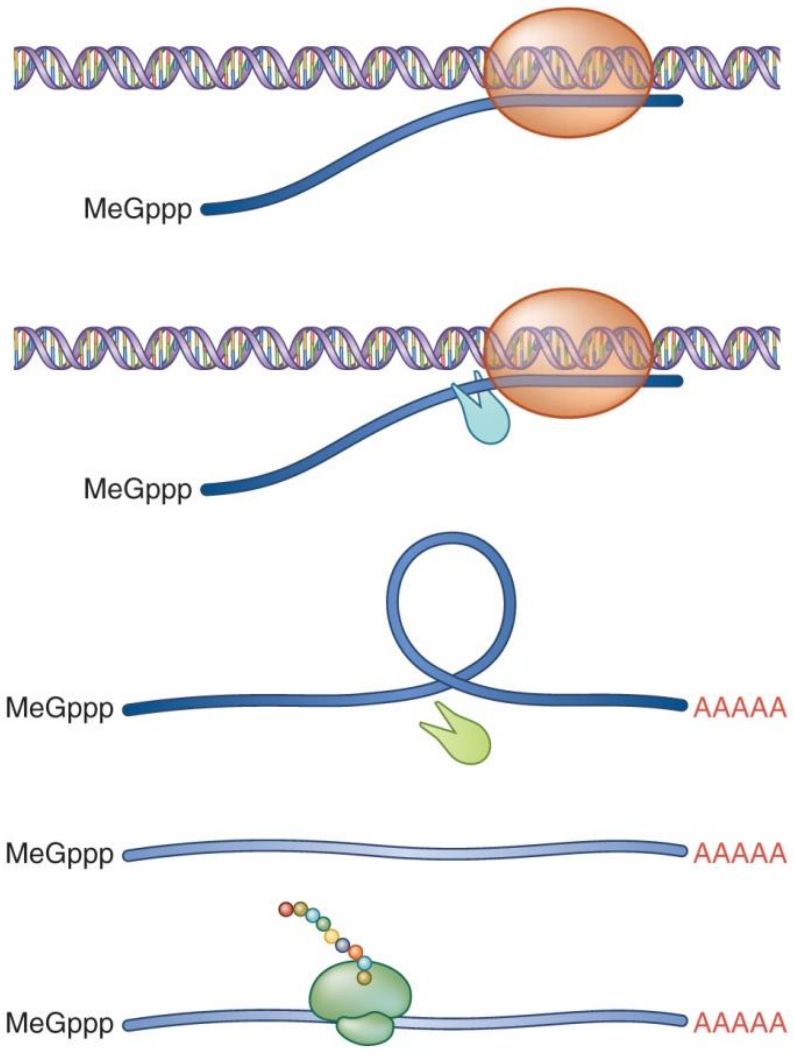




CAP and poly A are useful for stability as well as a signal for specific fates of the molecule

Turnaround of RNA production & modification can be measured in minutes -> fast response

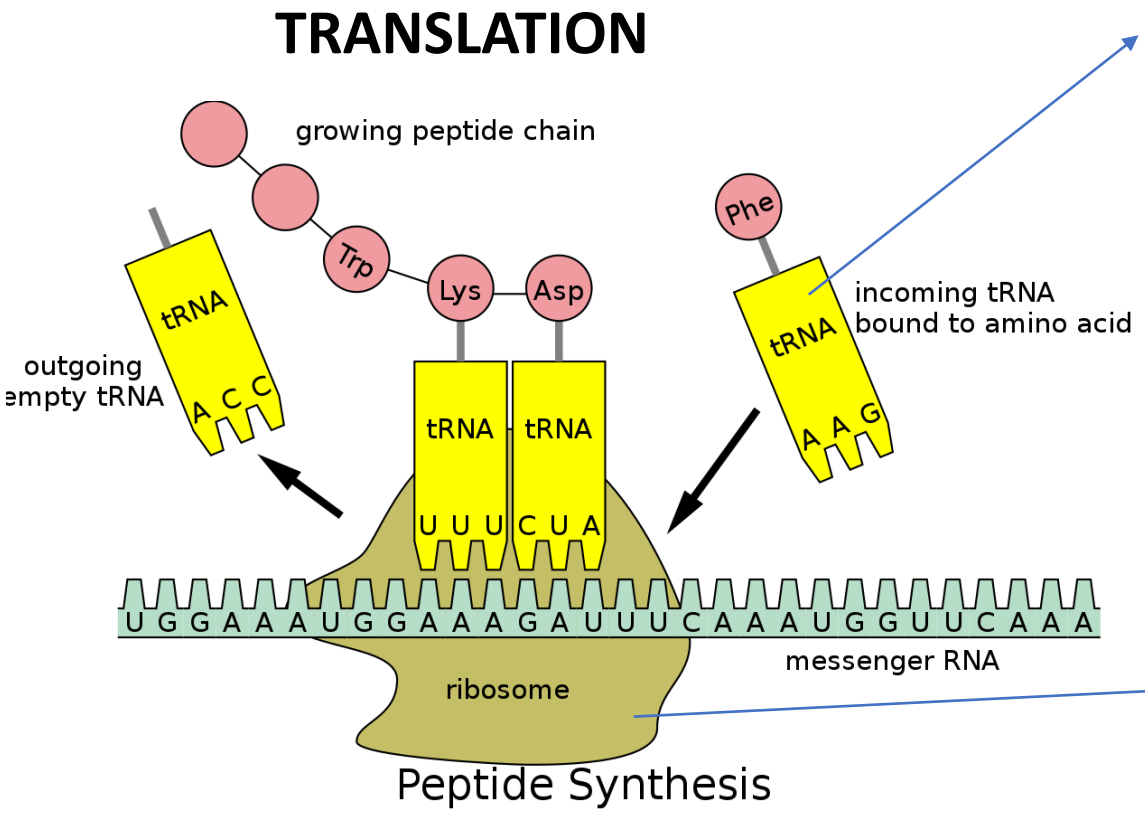
(non-coding RNAs also go through processes of maturation/modification)

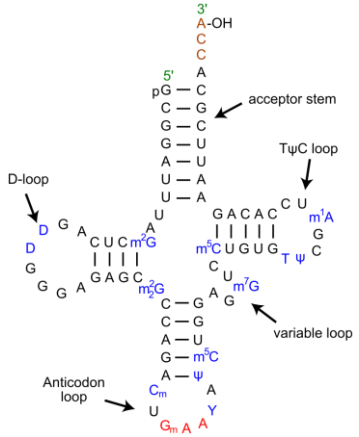


Processing step	Timeline
Transcription and 5′ end processing	~ 1 min
Transcription termination by 3′ end cleavage	~ 5–10 min
Polyadenylation and splicing	~ 10–20 min
Export to cytoplasm	~ 20–30 min
Translation	~ 30–240 min

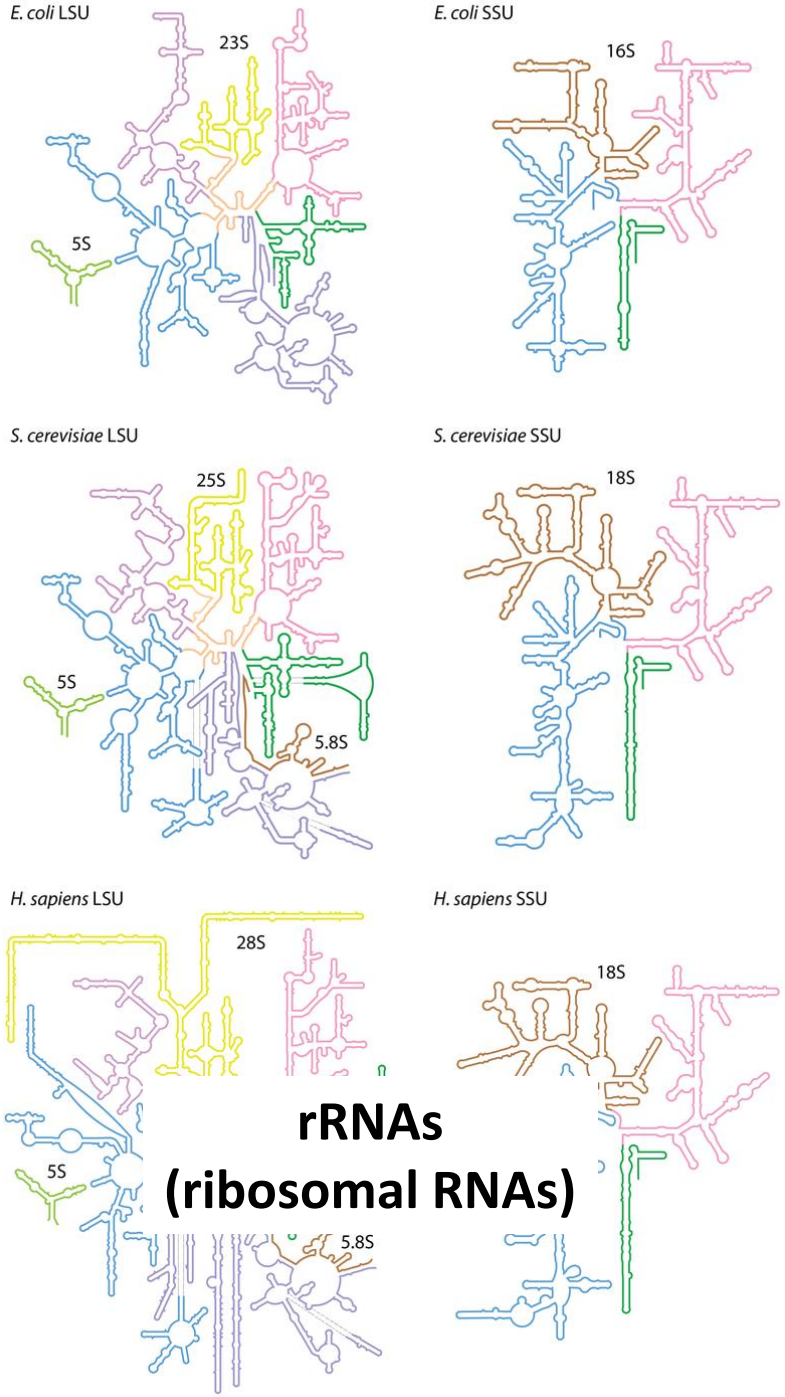
Figure 13.2 Timeline for events in mRNA processing. From Krebs *et al.*, 2010

non-coding RNAs make up the core machinery of translation

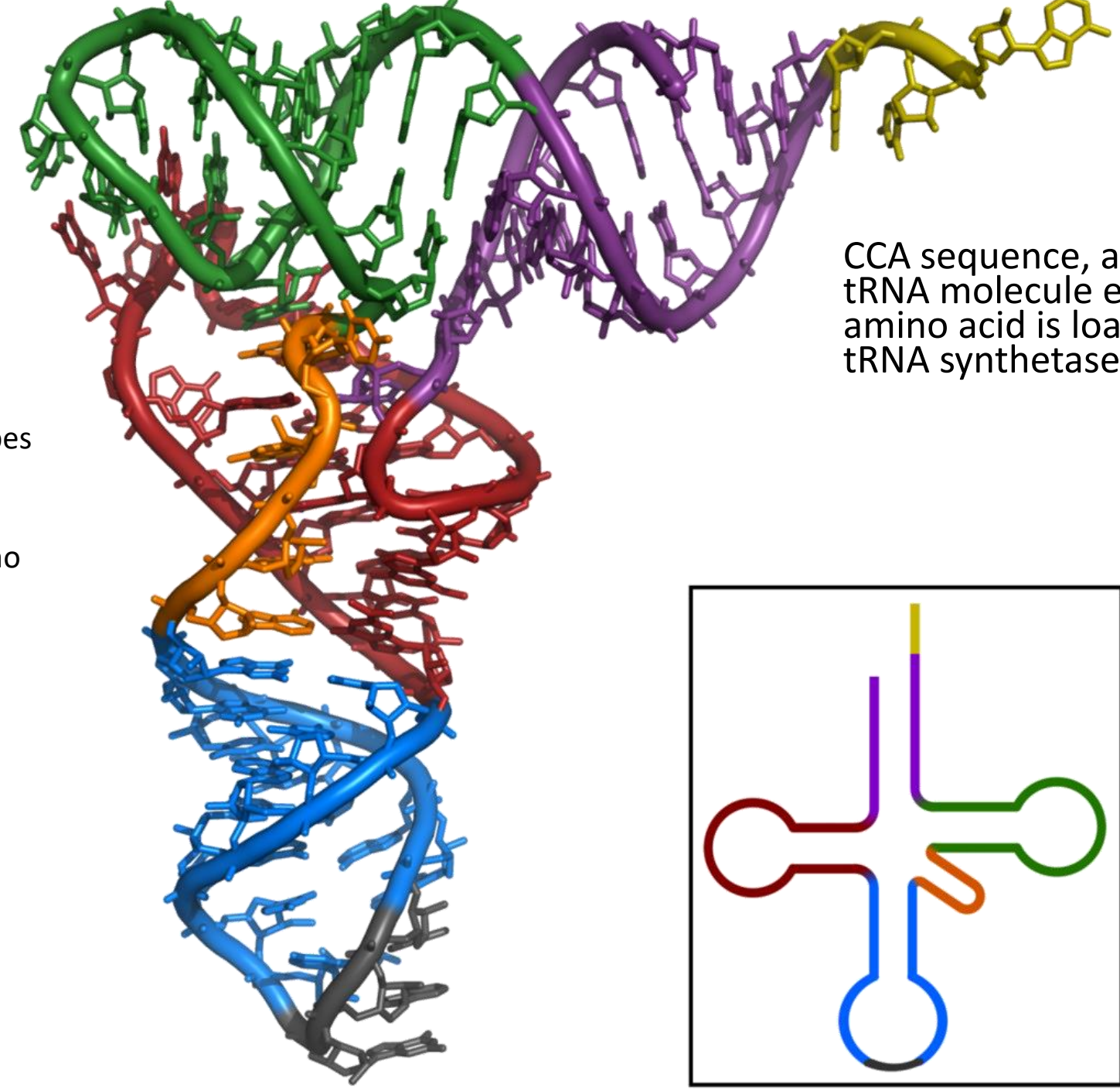




tRNAs (transfer RNAs)



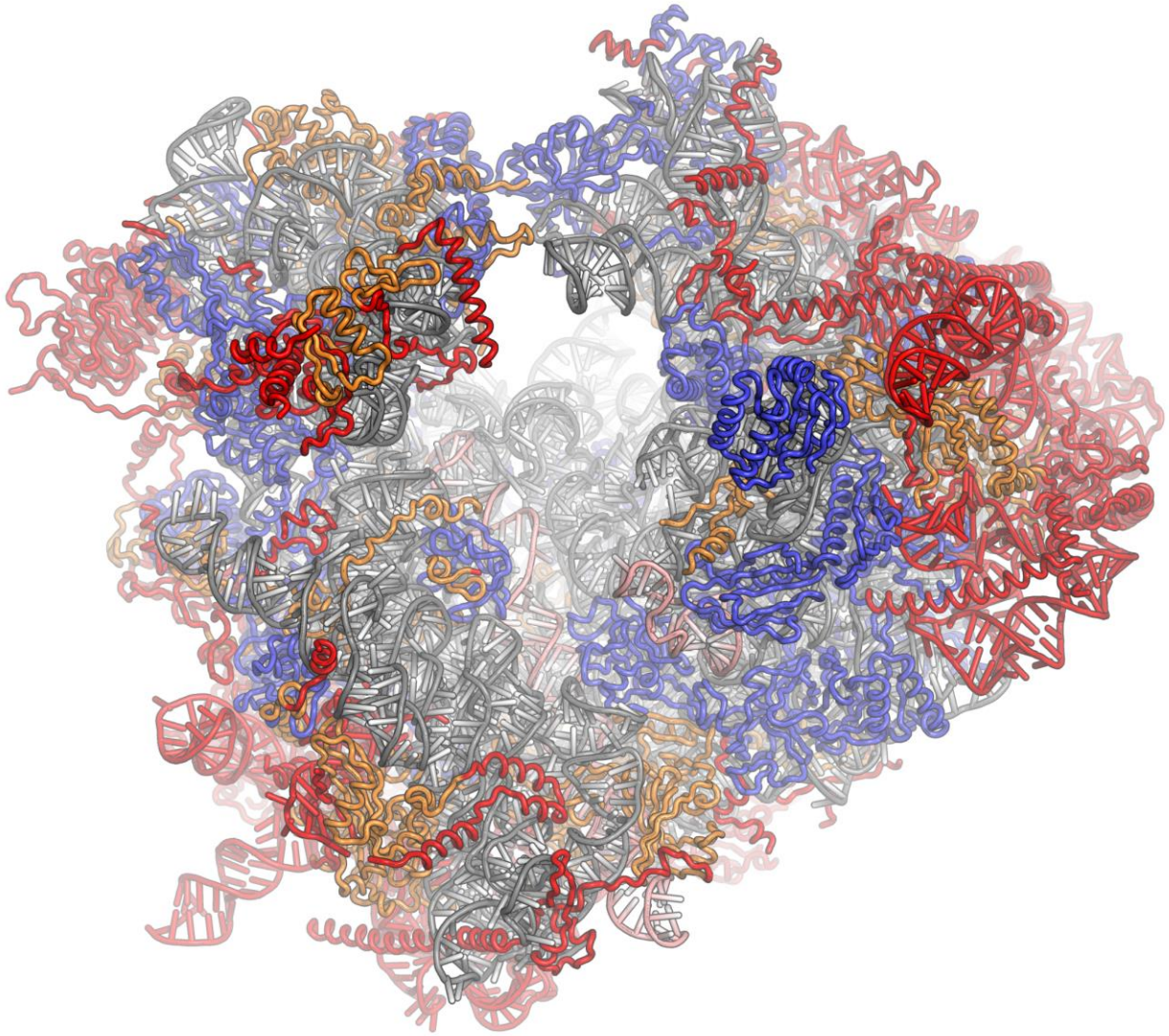
rRNAs (ribosomal RNAs)



CCA sequence, at the 3’ end of the tRNA molecule eand where the amino acid is loaded aminoacyl tRNA synthetase

In humans, the 20 different types of aa-tRNA are made by the 20 different aminoacyl-tRNA synthetases, one for each amino acid of the genetic code

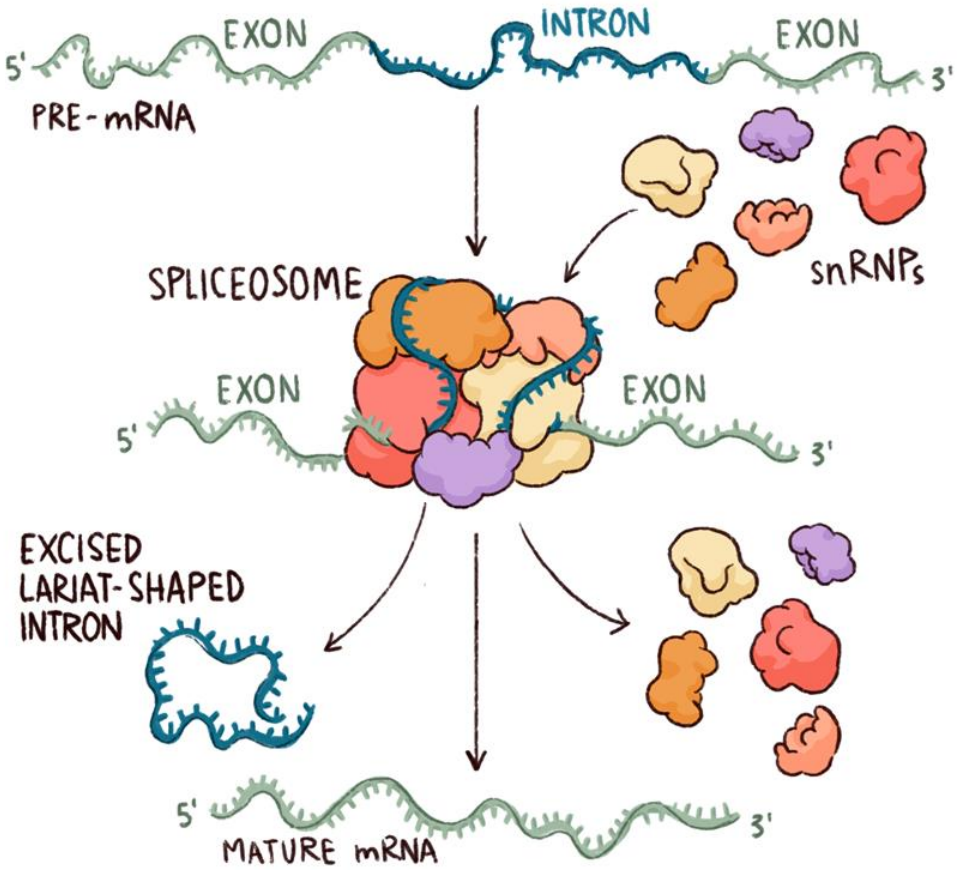
Anticodon, a unit of three nucleotides corresponding to the three bases of an mRNA codon



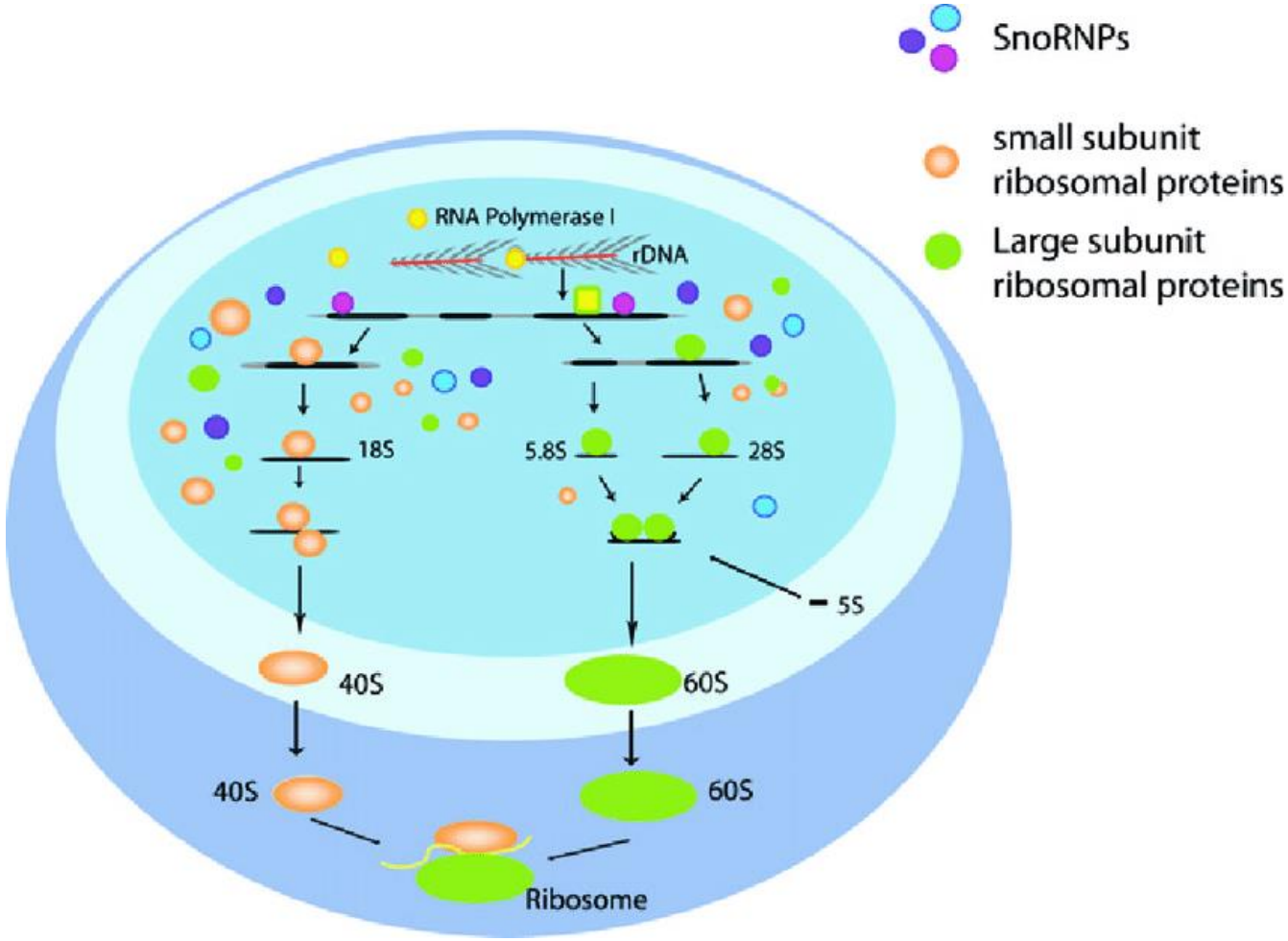
Eukaryotic ribosome

- 40S subunit is on the left
- 60S subunit on the right
- rRNA is the grey tube
- Expansion segments (extra block of RNA sequences) are red
- Universally conserved proteins are blue
- Proteins shared only between eukaryotes and archaea are orange
- Proteins specific to eukaryotes are red

		Eukaryotic[4]	Bacterial[4]
Ribosome	Sedimentation coefficient	80 S	70 S
	Molecular mass	~3.2×10^6 Da	~2.0×10^6 Da
	Diameter	~250–300 Å	~200 Å
Large subunit	Sedimentation coefficient	60 S	50 S
	Molecular mass	~2.0×10^6 Da	~1.3×10^6 Da
	Proteins	46	33
	rRNAs	• 25/28 S rRNA (3354 nucleotides) • 5 S rRNA (120 nucleotides) • 5.8 S rRNA (154 nucleotides)	• 23S rRNA (2839 nucleotides) • 5S rRNA (122 nucleotides)
Small subunit	Sedimentation coefficient	40 S	30 S
	Molecular mass	~1.2×10^6 Da	~0.7×10^6 Da
	Proteins	33	20
	rRNAs	• 18S rRNA (1753 nucleotides)	• 16S rRNA (1504 nucleotides)

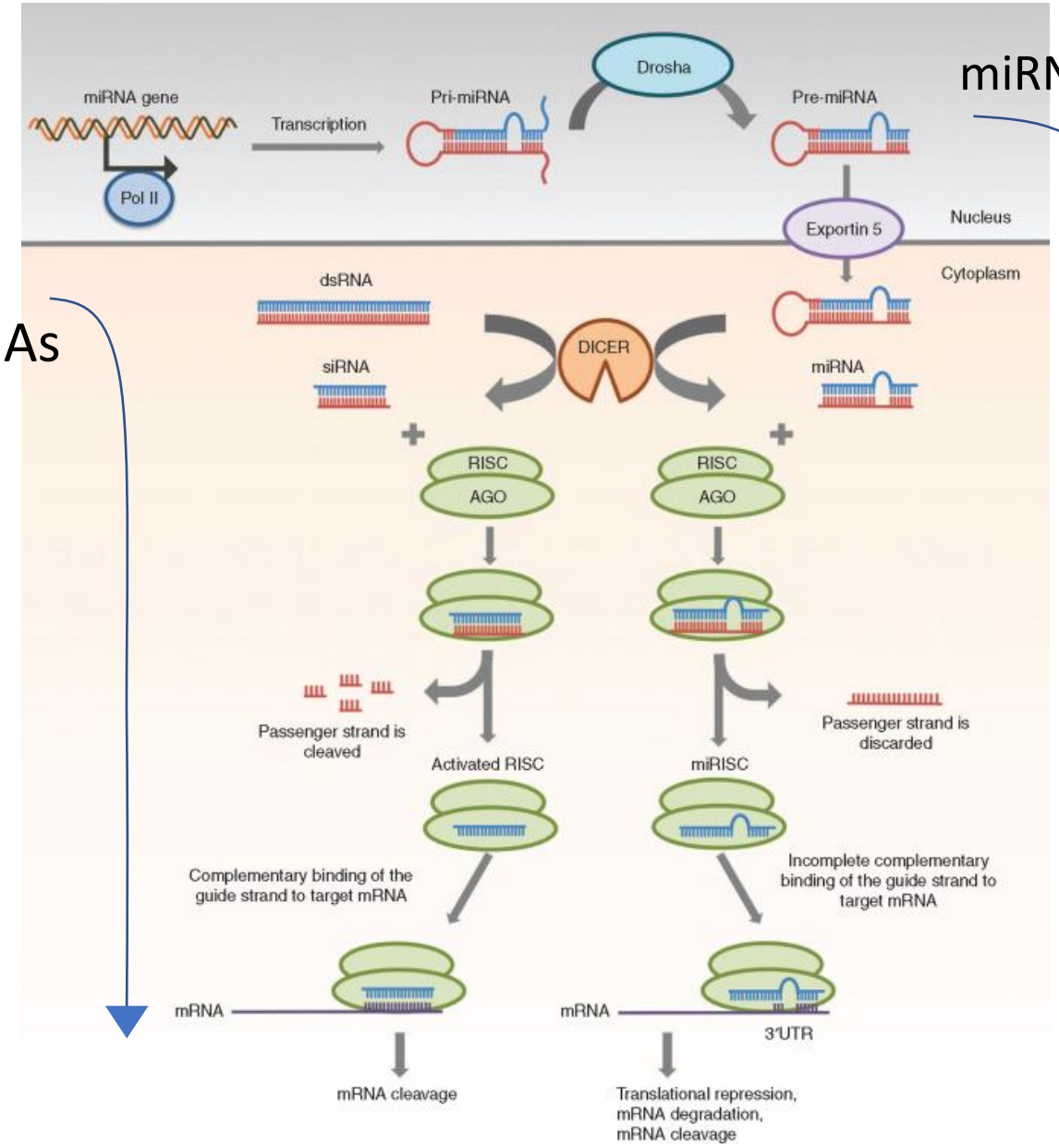


Among the main functions of small nuclear RNA (snRNA) is splicing (part of the spliceosome)



small nucleolar RNAs (snoRNA) stabilize the rRNA structure during ribosome biogenesis via
2'-O-methylation and pseudouridylation

miRNAs and siRNAs



miRNAs

siRNAs

RNA Interference

Micro RNAs and small interfering RNAs suppress gene expression by:

- Degradation of mRNA
- Inhibition of translation
- Heterochromatin formation (epigenetics)

- The initial gene transcript is called primary miRNA (pri-miRNA)
- In the cell nucleus, these hairpin-loop molecules are cut to form double-stranded precursor miRNA (pre-miRNA)
- The pre-miRNA is transported to the cytoplasm. There, it is further cut to form a functional mature miRNA (mature miRNA molecules are about 22 nucleotides long)
- The mature miRNA first binds with a molecule called the RNA interference silencing complex, or RISC
- Then the miRNA binds with its target messenger RNA (mRNA), thereby blocking its translation or prompting its degradation

Lam et al 2015

nature plants

ARTICLES

https://doi.org/10.1038/s41477-021-01005-w

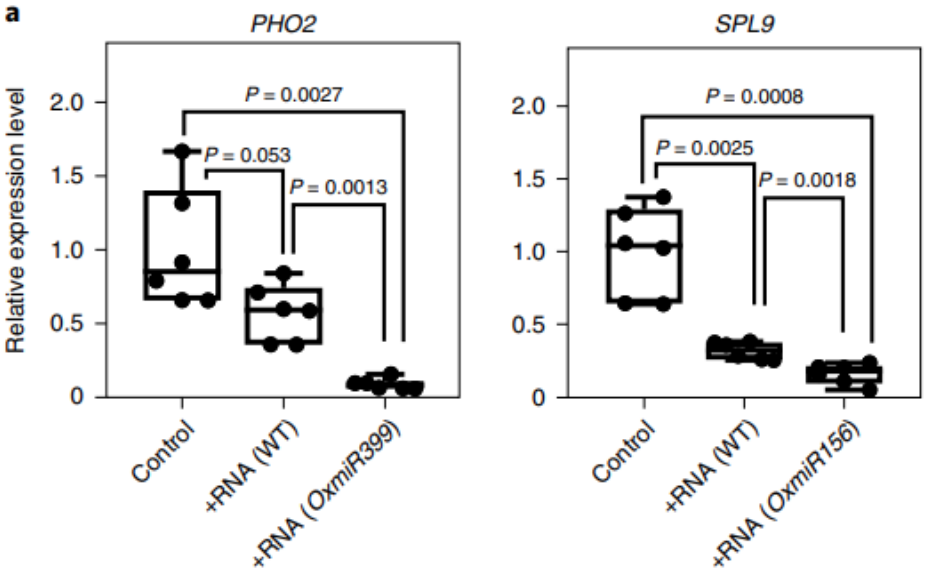
Check for updates

OPEN

Exogenous miRNAs induce post-transcriptional gene silencing in plants

Federico Betti[1,6], Maria Jose Ladera-Carmona[1,6], Daan A. Weits[1], Gianmarco Ferri[2], Sergio Iacopino[3], Giacomo Novi[1], Benedetta Svezia[1], Alicja B. Kunkowska[1], Antonietta Santaniello[4], Alberto Piaggesi[4], Elena Loreti[5]✉ and Pierdomenico Perata[1]✉

Plants seem to take up exogenous RNA that was artificially designed to target specific genes, followed by activation of the RNA interference (RNAi) machinery. It is, however, not known whether plants use RNAs themselves as signalling molecules in plant-to-plant communication, other than evidence that an exchange of small RNAs occurs between parasitic plants and their hosts. Exogenous RNAs from the environment, if taken up by some living organisms, can indeed induce RNAi. This phenomenon has been observed in nematodes and insects, and host *Arabidopsis* cells secrete exosome-like extracellular vesicles to deliver plant small RNAs into *Botrytis cinerea*. Here we show that micro-RNAs (miRNAs) produced by plants act as signalling molecules affecting gene expression in other, nearby plants. Exogenous miRNAs, such as *miR156* and *miR399*, trigger RNAi via a mechanism requiring both AGO1 and RDR6. This emphasizes that the production of secondary small interfering RNAs is required. This evidence highlights the existence of a mechanism in which miRNAs represent signalling molecules that enable communication between plants.



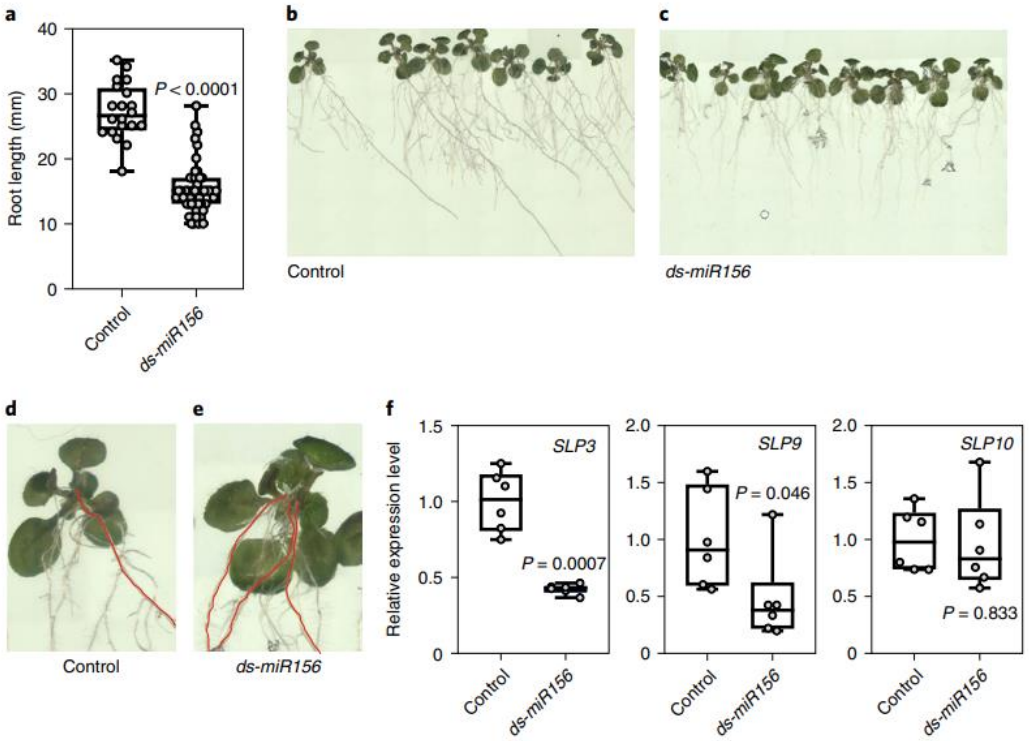


Fig. 2 | Exogenous *ds-miR156* influences the root phenotype in *Arabidopsis* seedlings. a, Primary root length in 10-day-old seedlings germinated on control medium ($n = 20$) or on a medium supplemented with 0.2 µM synthetic *ds-miR156* ($n = 48$). **b**,**c**, Photographs of representative control (**b**) and *ds-miR156*-treated (**c**) vertical plates (15 days after sowing). **d**,**e**, Magnified seedlings from **b** (**d**) and **c** (**e**) showing (in red) the presence of adventitious roots. **f**, Transcript level of *SPL3*, *SPL9* and *SPL10* extracted from roots at the stage depicted in **b** and **c**. For all boxplots, the bottom and top of each box denote the first and third quartile, respectively ($n = 6$ biological replicates). In the boxplots, dots represent single data points, whiskers denote minimum/maximum values, the box defines the interquartile range, the centre represents the median and box bounds represent the lower and upper quartiles. Welch's *t*-test (two-sided) values are shown.

Persistence of RNA in environment

Release and degradation of environmental DNA and RNA in a marine system

Susanna A. Wood [a,*], Laura Biessy [a], Janie L. Latchford [a], Anastasija Zaiko [a,b], Ulla von Ammon [a], François Audrezet [a], Melania E. Cristescu [c], Xavier Pochon [a,b]

[a] Coastal and Freshwater Group, Cawthron Institute, Nelson, New Zealand
[b] Institute of Marine Science, University of Auckland, Auckland, New Zealand
[c] Department of Biology, McGill University, Montreal, QC, Canada

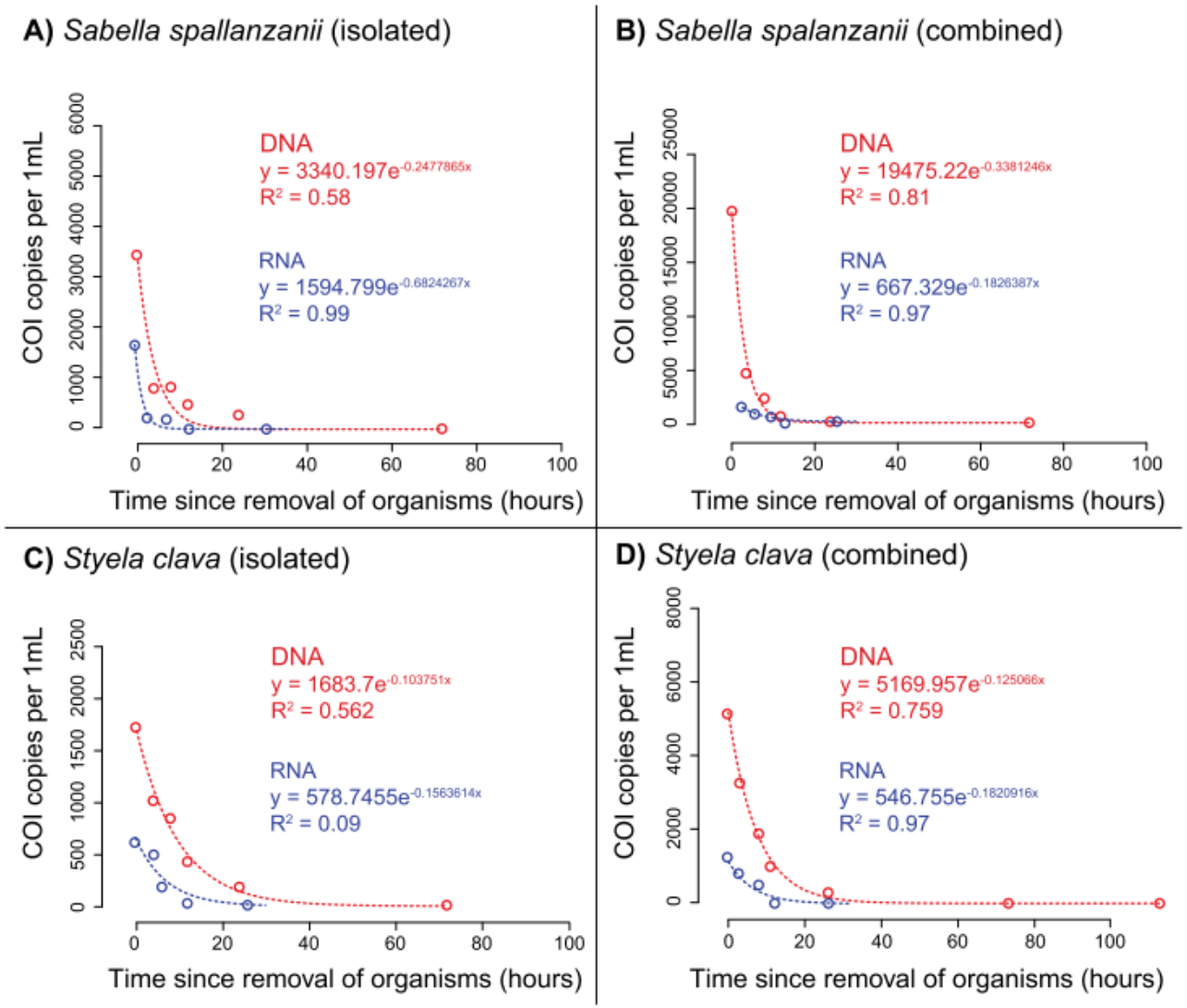
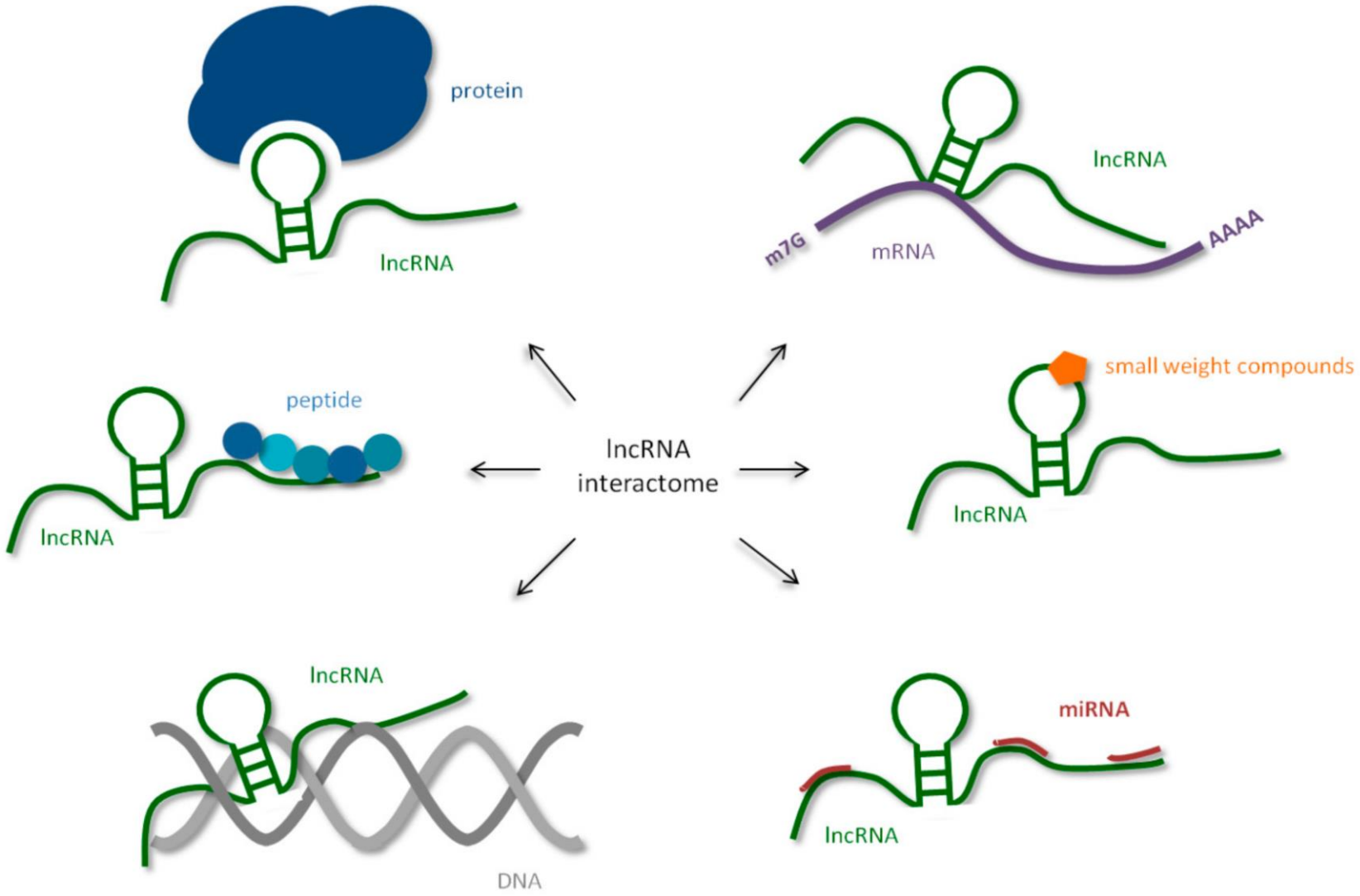


Fig. 3. Time-dependent changes in average environmental DNA and RNA concentration (based on detection of the *Cytochrome c Oxidase* subunit 1 [COI] gene) for: (A) *Sabella spallanzanii* in isolation, (B) *S. spallanzanii* when combined with *Styela clava*, (C) *S. clava* in isolation, and (D) *S. clava* when combined with *S. spallanzanii*. Equations show the rate of exponential decay after applying the decay model $N(t) = N_0 e^{-\lambda t}$ to all raw data. R^2 values indicate the closeness of fit of raw data to the fitted decay model. Individual curves for DNA and RNA are shown in Fig. S1.

Long, non-coding RNAs



nature reviews molecular cell biology https://doi.org/10.1038/s41580-022-00566-8

Consensus statement

Check for updates

Long non-coding RNAs: definitions, functions, challenges and recommendations

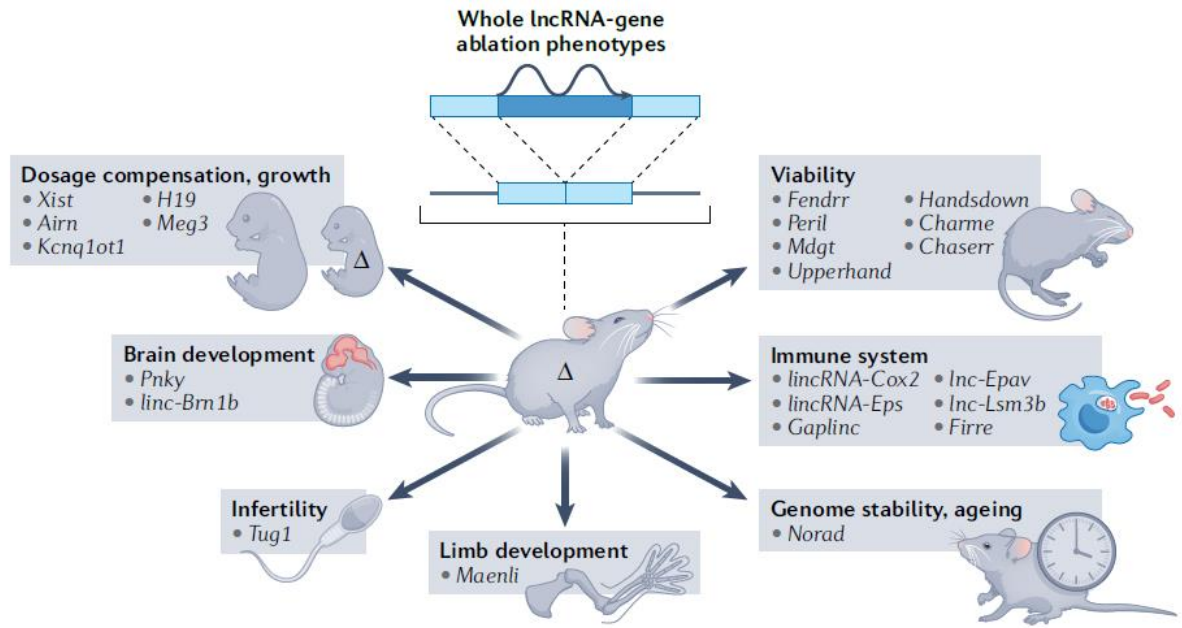


Fig. 1 | Visible phenotypes of mutations in long non-coding RNA genes in mice[163]. The following long non-coding RNAs (lncRNAs) are listed in the figure underneath their associated phenotypes: *Airn*, antisense of IGF2R non-protein-coding RNA[147,435]; *Charme*, chromatin architect of muscle expression[436]; *Chaserr*, *CHD2* adjacent, suppressive regulatory RNA[437]; *Fendrr*, *FOXF1* adjacent non-coding developmental regulatory RNA[165,438]; *Firre*, functional intergenic repeating RNA element[316]; *Gaplinc*, gastric adenocarcinoma predictive long intergenic non-coding RNA[200]; *H19*, clone pH19 (ref. 439); *Handsdown*, downstream of the protein-coding gene *Hand2* (ref. 440) *Kcnq1ot1*, *Kcnq1* overlapping antisense transcript 1 (ref. 441); *linc-Brn1b*, long intergenic non-coding RNA (lincRNA) downstream of the *Brn1* protein-coding gene[165]; *linc-Epav*, endogenous retrovirus-derived lncRNA positively regulates antiviral responses[442]; *lincRNA-Cox2*, lincRNA downstream of the inflammation response gene *Cox2* (ref. 443); *lincRNA-Eps*, lincRNA involved in erythroid prosurvival[201]; *lnc-Lsm3b*, interferon-inducible non-coding splice variant of the U6 small nuclear RNA-associated Sm-like protein *lsm3* gene[444]; *Maenli*, master activator of *engrailed1* in the limb[165]; *Mdgt*, midget[165]; *Meg3*, maternally expressed gene 3 (also known as *Gtl2*)[445,446]; *Norad*, non-coding RNA activated by DNA damage[447]; *Peril*, perinatal lethal long non-coding RNA[165]; *Pnky*, pinky (also known as *lnc-Pou3f2*)[448]; *Tug1*, taurine upregulated gene 1 (refs. 165,166,449) *Upperhand*, lncRNA upstream of the *Hand2* cardiomyocyte transcription factor locus[318]; *Xist*, X-inactive-specific transcript[450]. Figure courtesy of Daniel Andergassen and John Rinn.

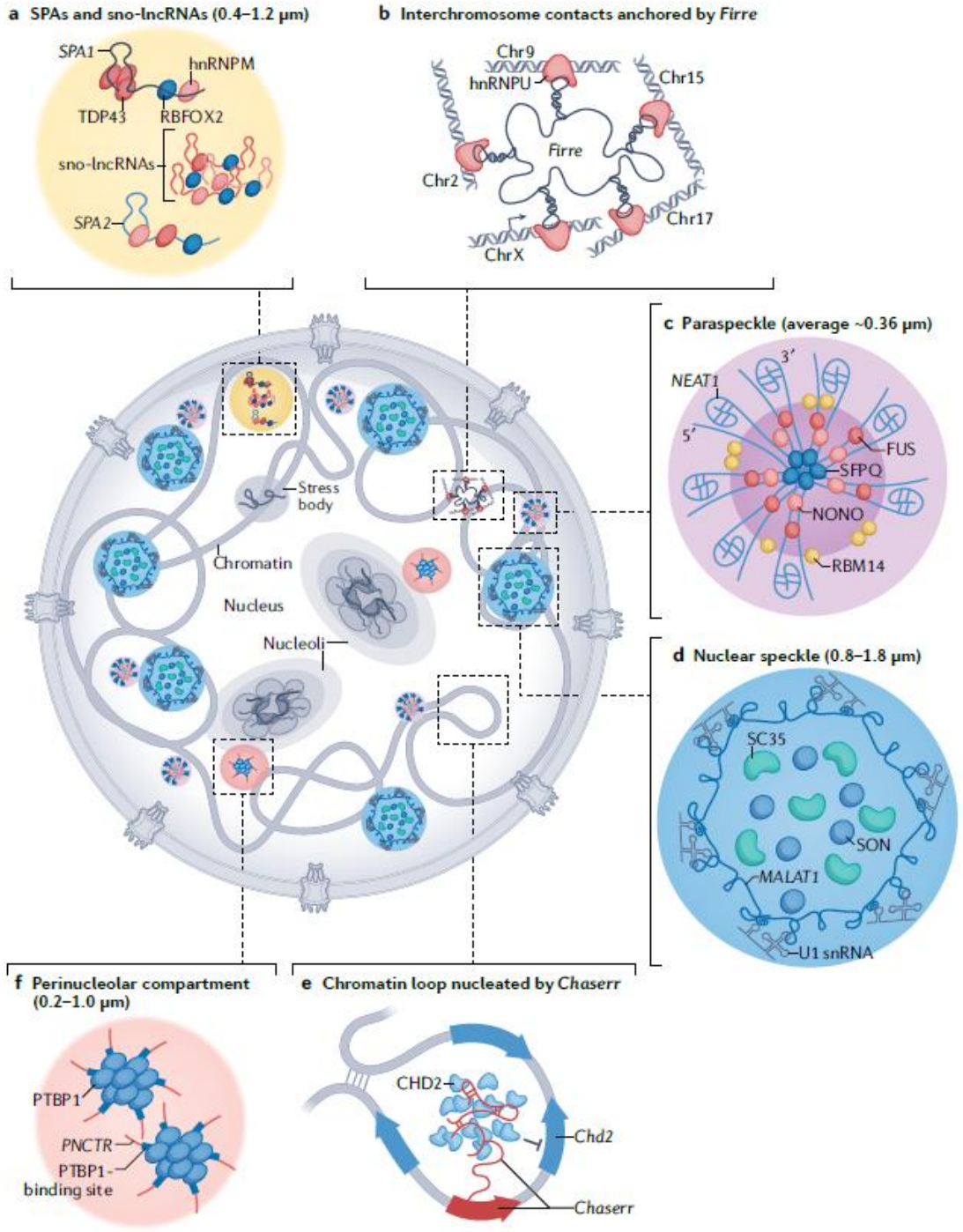
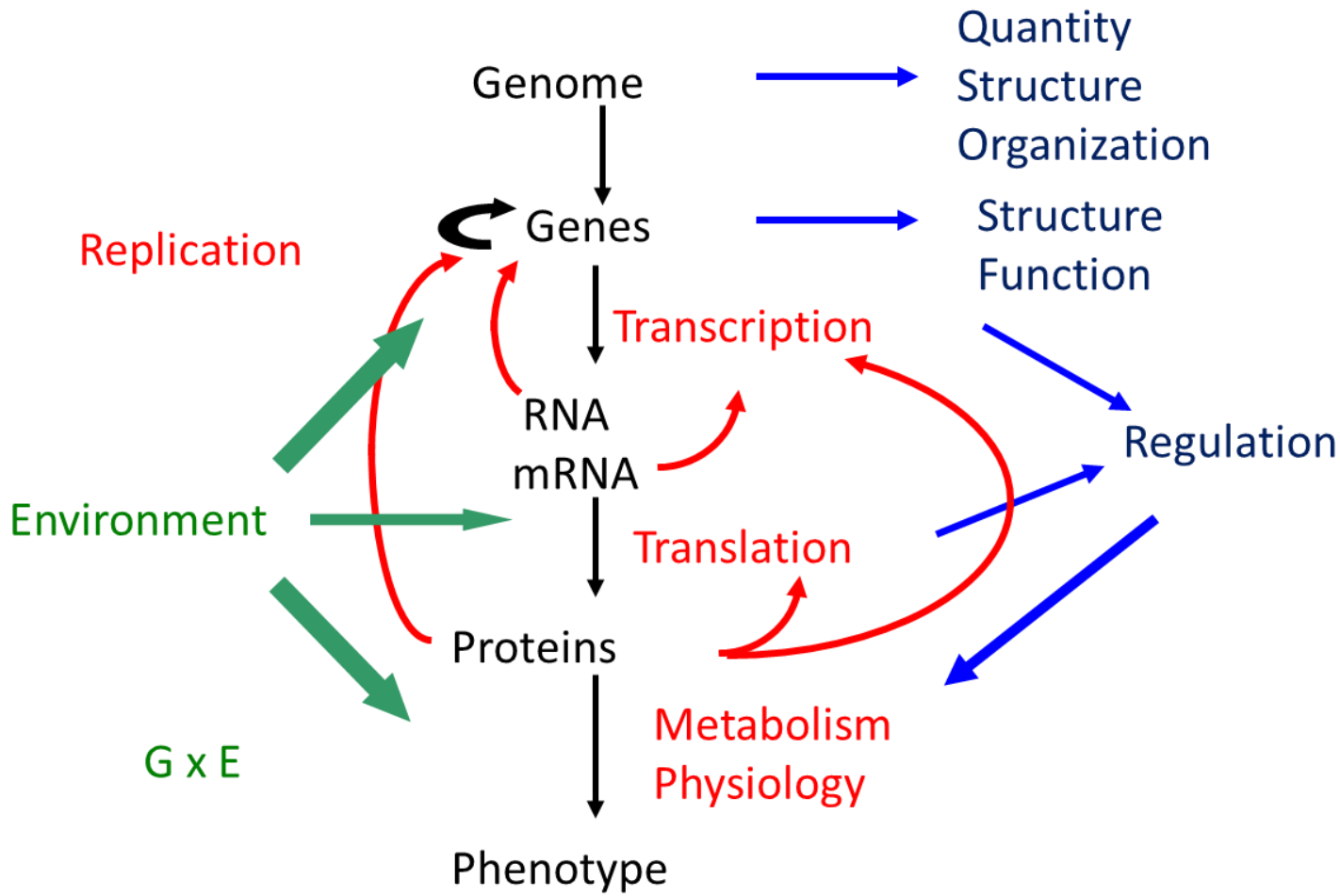


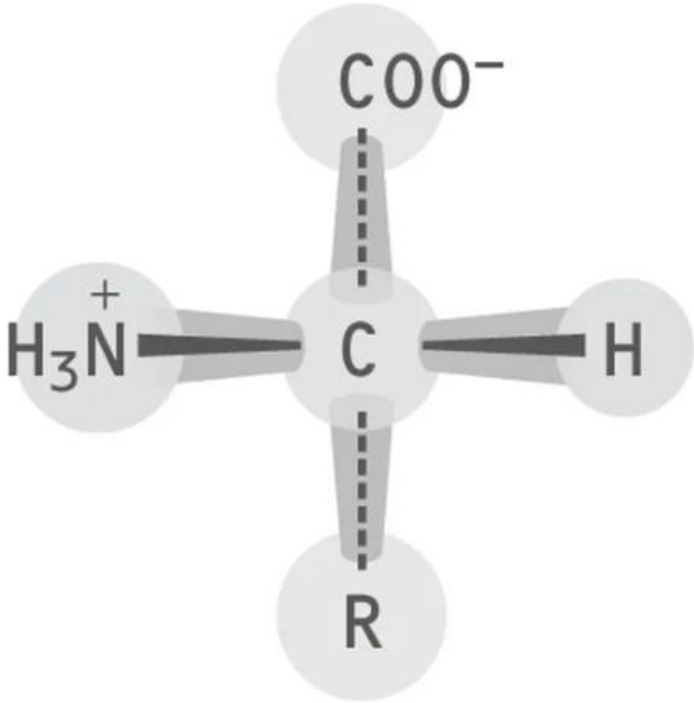
Fig. 2 | Roles of long non-coding RNAs in nuclear organization. a, 5′ small nucleolar RNA-capped and 3′-polyadenylated long non-coding (lncRNAs) (SPAs)[42] and small nucleolar RNA-related lncRNAs (sno-lncRNAs)[41] accumulate at their sites of transcription and interact with several splicing factors such as RNA-binding protein FOX-1 homologue 2 (RBFOX2), TAR DNA-binding protein 43 (TDP43) and heterogeneous nuclear ribonucleoprotein M (hnRNPM) to form a microscopically visible nuclear body that is involved in the regulation of alternative splicing[42]. **b**, The lncRNA functional intergenic repeating RNA element (*Firre*) is transcribed from the mouse X chromosome and interacts with the nuclear matrix factor hnRNPU to tether chromosome X (chrX), chr2, chr9, chr15 and chr17 into a nuclear domain[451,452]. **c**, The lncRNA nuclear paraspeckle assembly transcript 1 (*NEAT1*) is essential for the formation of paraspeckles[178]. *NEAT1* sequesters numerous paraspeckle proteins to form a highly organized core–shell (dark and light purple, respectively) spheroidal nuclear body[453]. The middle region of *NEAT1* is localized in the centre of paraspeckles, and the 3′-end and 5′-end regions are localized in the periphery[453]. Different paraspeckle proteins are embedded by *NEAT1* into the spheroidal structure in the core region (non-POU domain-containing octamer-binding protein (NONO), fused in sarcoma (FUS) and splicing factor, proline- and glutamine-rich (SFPQ)) or in the shell region (RNA-binding motif protein 14 (RBM14))[453]. **d**, The lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) is localized at the periphery of nuclear speckles[172,454] and is involved in the regulation of pre-mRNA splicing[339,455]. *MALAT1* interacts with the U1 small nuclear RNA (U1 snRNA)[428], whereas proteins such as SON DNA- and RNA-binding protein and splicing component 35 kDa (SC35) are localized at the centre of nuclear speckles[456]. **e**, The lncRNA *CHD2* adjacent, suppressive regulatory RNA (*Chaserr*) forms a compartment within a region of the mouse chromosome corresponding to a topologically associating domain that includes its own gene as well as the *Chd2* gene (encoding chromodomain DNA helicase protein 2 (CHD2))[437]. *Chaserr* limits in *cis* the expression of *Chd2*, which is important for proper regulation of many genes (not shown). **f**, The perinucleolar compartment contains the lncRNA pyrimidine-rich non-coding transcript (*PNCTR*), which sequesters pyrimidine tract-binding protein 1 (PTBP1) and thus suppresses PTPBP1-mediated pre-mRNA splicing elsewhere in the nucleoplasm[369]. The size of nuclear bodies is indicated where relevant[457]. Figure adapted from ref. 80, Springer Nature. Part **e** courtesy of Inna-Marie Strazhnik and Mitch Guttman.

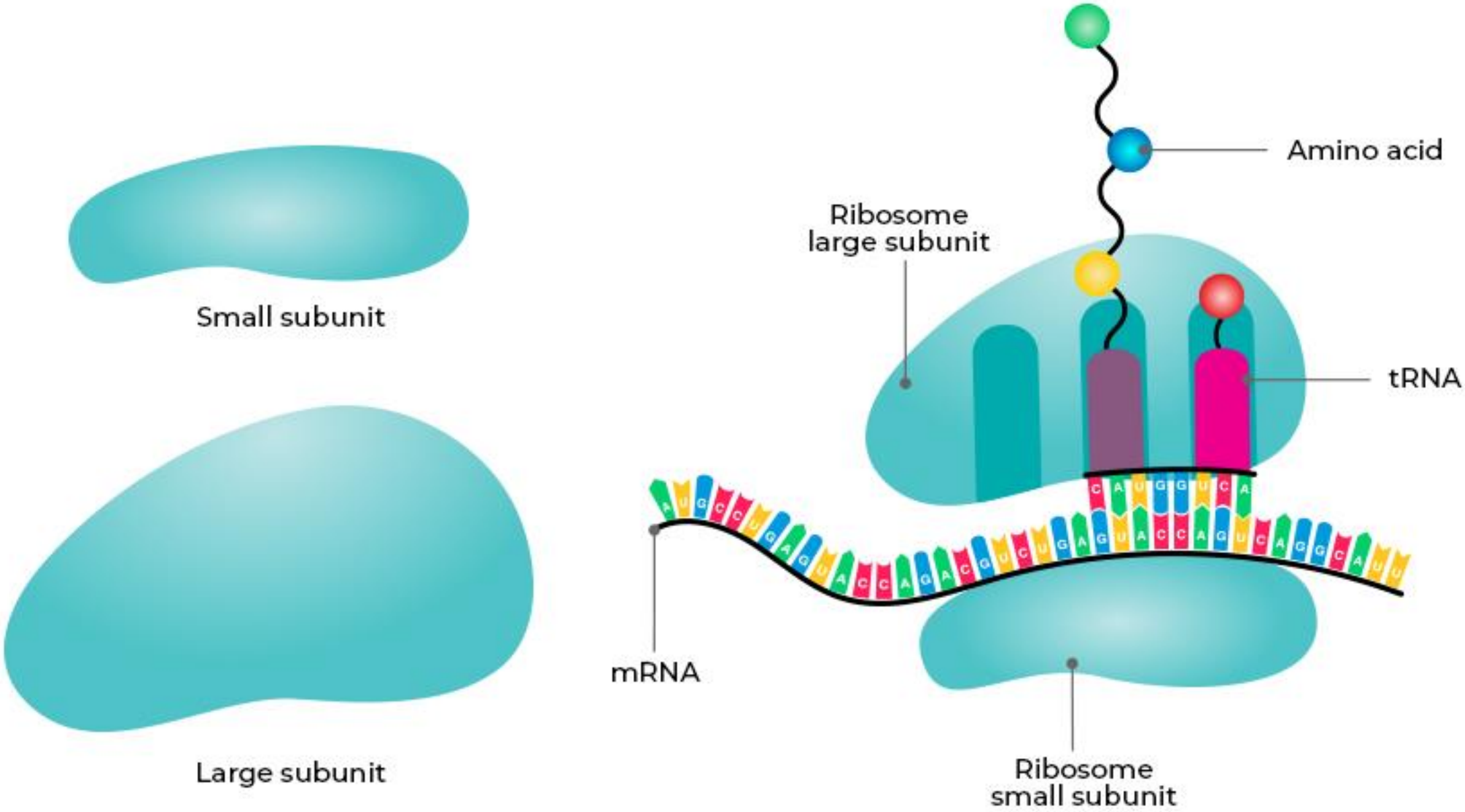

Genome
Quantity
Structure
Organization
Genes
Structure
Function
Replication
Transcription
RNA
mRNA
Regulation
Environment
Translation
Proteins
Metabolism
Physiology
G x E
Phenotype

Proteins

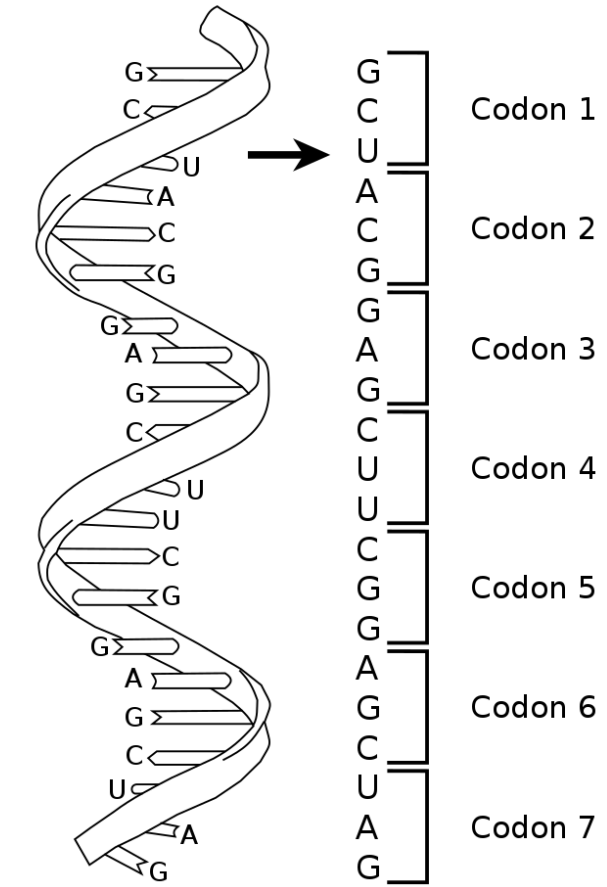
The **proteome** comes after the transcriptome

- This is the cell's repertoire of proteins which specifies the nature of the biochemical reactions that the cell is able to carry out
- Derives from translation of the RNA; proteins are also linear unbranched polymers
- The monomeric subunits of proteins are amino acids; together, the buid polypeptides tipically shorter than 2K AA
- There's about 500 existing amino acids out there; 22 α-amino acids are the only ones appearing in the genetic code




Small subunit
Large subunit
Amino acid
Ribosome
large subunit
tRNA
mRNA
Ribosome
small subunit

The genetic code is degenerate; Multiple triplets encode the same AA

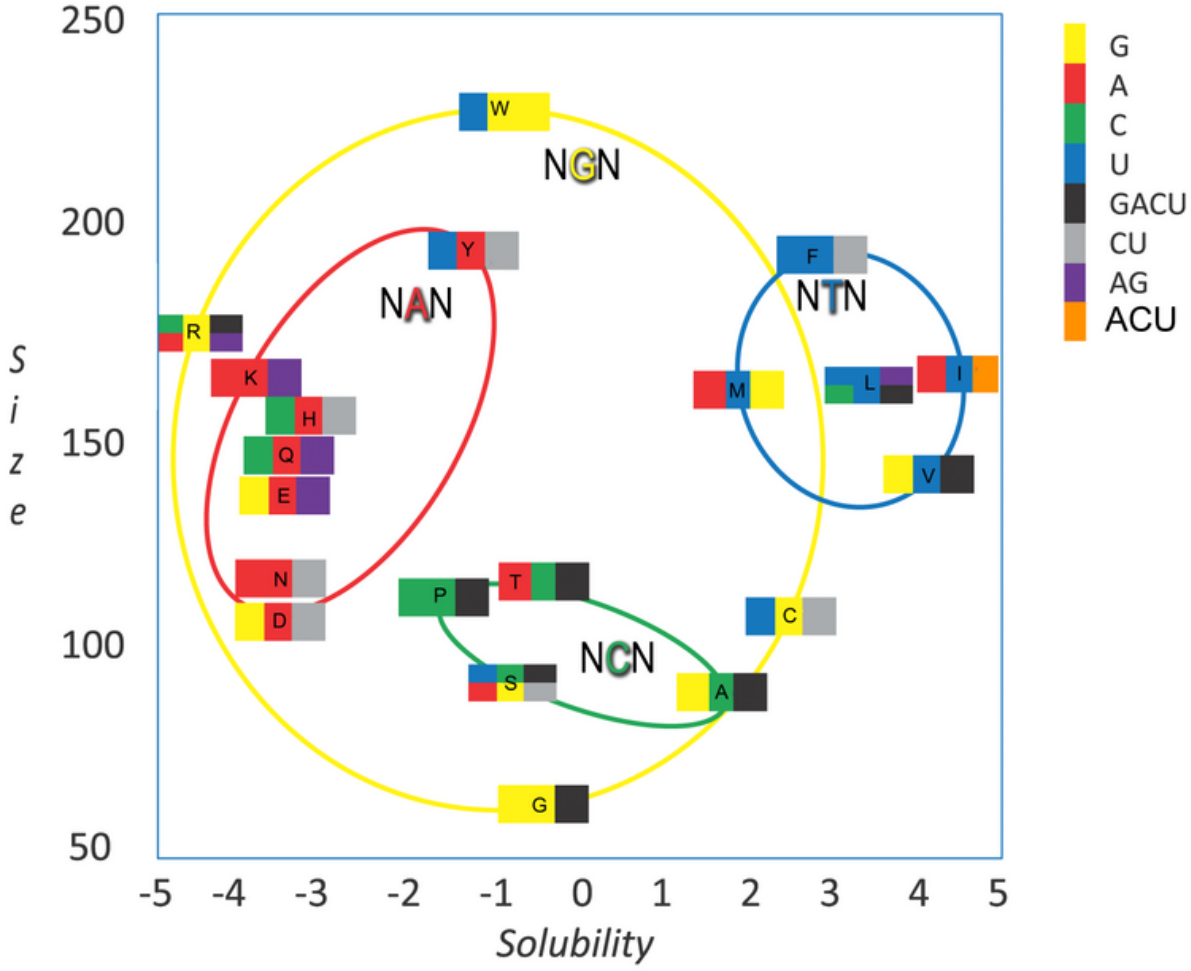


RNA

Ribonucleic acid

First Letter	Second Letter: T	Second Letter: C	Second Letter: A	Second Letter: G	Third Letter
T	TTT Phe TTC Phe TTA Leu TTG Leu	TCT Ser TCC Ser TCA Ser TCG Ser	TAT Tyr TAC Tyr TAA Stop TAG Stop	TGT Cys TGC Cys TGA Stop TGG Trp	T C A G
C	CTT Leu CTC Leu CTA Leu CTG Leu	CCT Pro CCC Pro CCA Pro CCG Pro	CAT His CAC His CAA Gln CAG Gln	CGT Arg CGC Arg CGA Arg CGG Arg	T C A G
A	ATT Ile ATC Ile ATA Ile ATG Met	ACT Thr ACC Thr ACA Thr ACG Thr	AAT Asn AAC Asn AAA Lys AAG Lys	AGT Ser AGC Ser AGA Arg AGG Arg	T C A G
G	GTT Val GTC Val GTA Val GTG Val	GCT Ala GCC Ala GCA Ala GCG Ala	GAT Asp GAC Asp GAA Glu GAG Glu	GGT Gly GGC Gly GGA Gly GGG Gly	T C A G

- Redundancy!
- Evolutionary processes; chemical features of degerenate triplets are somehow similar
- Taxon-specificity: in the nuclear genetic code, UGA is a stop codon; in the mitochondrial genetic code it codes for tryptophan; In the nuclear genetic code, AUA codes for isoleucine; for methionine in the mitochondrial genetic code



Examples of notable **Mutations**

Selection of notable mutations, ordered in a standard table of the genetic code of amino acids.

Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic, polar or nonpolar, while nonsense mutations result in a stop codon.

1st base \ 3rd base in each row	2nd base: U	2nd base: C	2nd base: A	2nd base: G
U	UUU (Phe/F) Phenylalanine	UCU (Ser/S) Serine	UAU (Tyr/Y) Tyrosine	UGU (Cys/C) Cysteine
	UUC (Phe/F) Phenylalanine	UCC (Ser/S) Serine	UAC (Tyr/Y) Tyrosine	UGC (Cys/C) Cysteine
	UUA (Leu/L) Leucine	UCA (Ser/S) Serine	UAA Ochre (Stop)	UGA Opal (Stop)
	UUG (Leu/L) Leucine	UCG (Ser/S) Serine	UAG Amber (Stop)	UGG (Trp/W) Tryptophan
C	CUU (Leu/L) Leucine	CCU (Pro/P) Proline	CAU (His/H) Histidine	CGU (Arg/R) Arginine
	CUC (Leu/L) Leucine	CCC (Pro/P) Proline	CAC (His/H) Histidine	CGC (Arg/R) Arginine
	CUA (Leu/L) Leucine	CCA (Pro/P) Proline	CAA (Gln/Q) Glutamine	CGA (Arg/R) Arginine
	CUG (Leu/L) Leucine	CCG (Pro/P) Proline	CAG (Gln/Q) Glutamine	CGG (Arg/R) Arginine
A	AUU (Ile/I) Isoleucine	ACU (Thr/T) Threonine	AAU (Asn/N) Asparagine	AGU (Ser/S) Serine
	AUC (Ile/I) Isoleucine	ACC (Thr/T) Threonine	AAC (Asn/N) Asparagine	AGC (Ser/S) Serine
	AUA (Ile/I) Isoleucine	ACA (Thr/T) Threonine	AAA (Lys/K) Lysine	AGA (Arg/R) Arginine
	AUG (Met/M) Methionine	ACG (Thr/T) Threonine	AAG (Lys/K) Lysine	AGG (Arg/R) Arginine
G	GUU (Val/V) Valine	GCU (Ala/A) Alanine	GAU (Asp/D) Aspartic acid	GGU (Gly/G) Glycine
	GUC (Val/V) Valine	GCC (Ala/A) Alanine	GAC (Asp/D) Aspartic acid	GGC (Gly/G) Glycine
	GUA (Val/V) Valine	GCA (Ala/A) Alanine	GAA (Glu/E) Glutamic acid	GGA (Gly/G) Glycine
	GUG (Val/V) Valine	GCG (Ala/A) Alanine	GAG (Glu/E) Glutamic acid	GGG (Gly/G) Glycine

ΔF508 deletion in cystic fibrosis

- Myotonic dystrophy
- SCA 8

β-Thalassemia

McArdle's disease

β-Thalassemia

Fragile X Syndrome

Prostate cancer

Colorectal cancer

Sickle-cell disease

Friedreich's ataxia

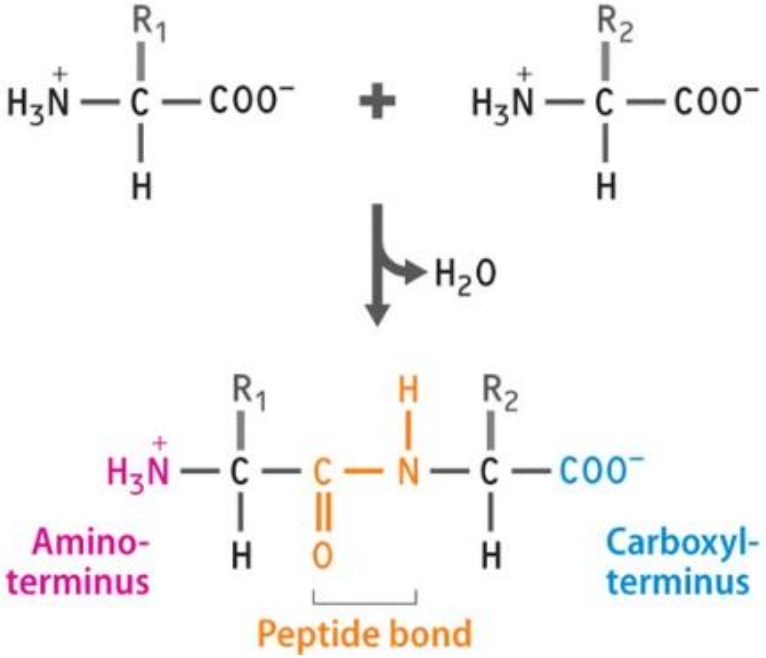
Amino acids
- Basic
- Acidic
- Polar
- Nonpolar (hydrophobic)

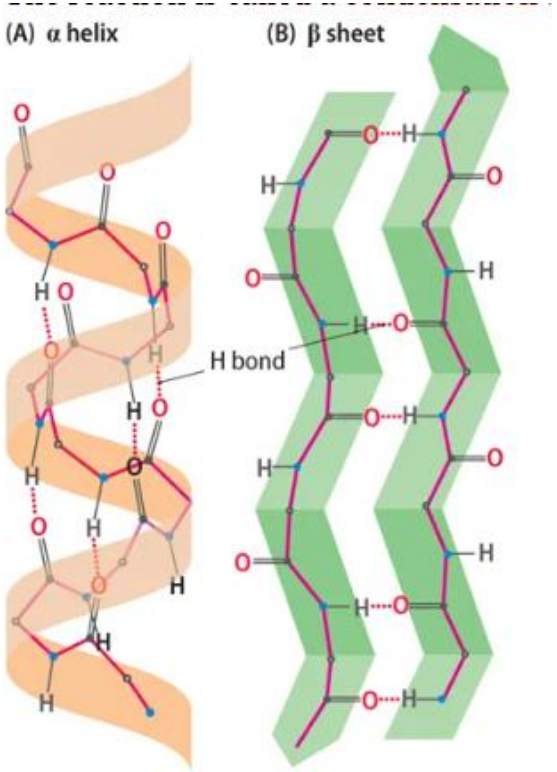
Polyglutamine (PolyQ) Diseases
- Huntington's disease
- Spinocerebellar ataxia (SCA) (most types)
- Spinobulbar muscular atrophy (Kennedy disease)
- Dentatorubral-pallidoluysian atrophy

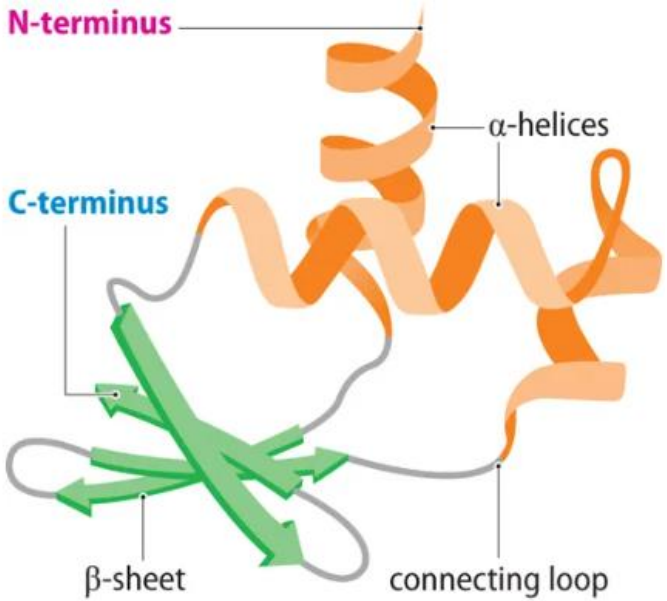
Mutation type
- = Trinucleotide repeat
- = Deletion
- = Missense
- = Nonsense

Proteins have four distinct levels of structure:

- **Primary**: the linear sequence deriving from joining amino acids in the polypeptide (also polypeptides have two distinct ends: N terminus, C terminus
- **Secondary**: the different conformation that the polypeptide can assume: «alpha helix» or «beta sheet». These depend on H bonds
- **Tertiary**: three dimensional configuration resulting from combination of secondary structures. Stabilized by various different chemical bonds, hydrophobic effects, disulfide bonds
- **Quaternary**: resulting from the assocation of 2+ polypeptides. These are multisubunit proteins held together by revertible bonds







https://alphafold.ebi.ac.uk/