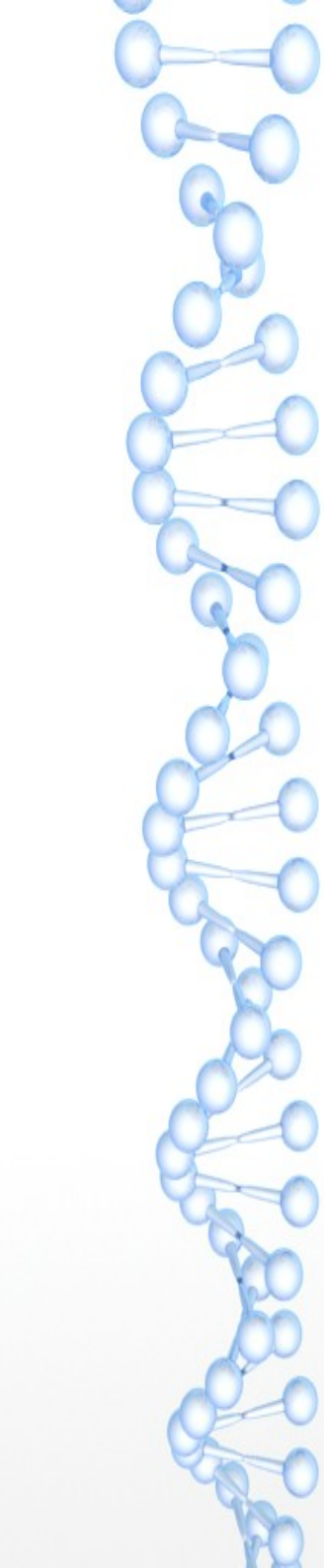


II. Gene and genetic code

Definition of term gene

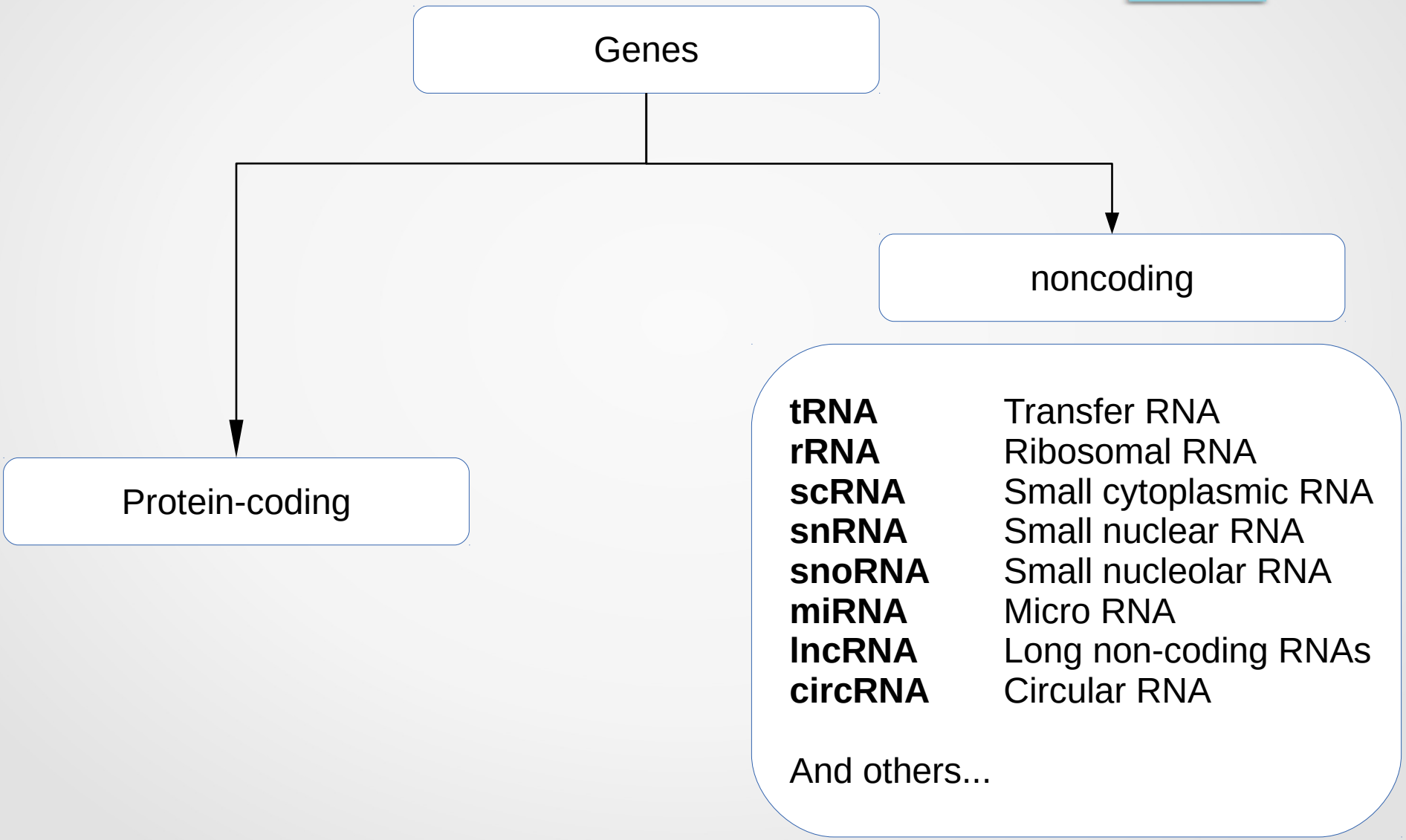
Gene is

- basic unit of inheritance
- region of DNA encoding function
- unit of hereditary information that occupies a fixed position (locus)
- nucleotide sequence that stores the information which specifies the order of the monomers in a final functional polypeptide or RNA molecule, or set of closely related isoforms

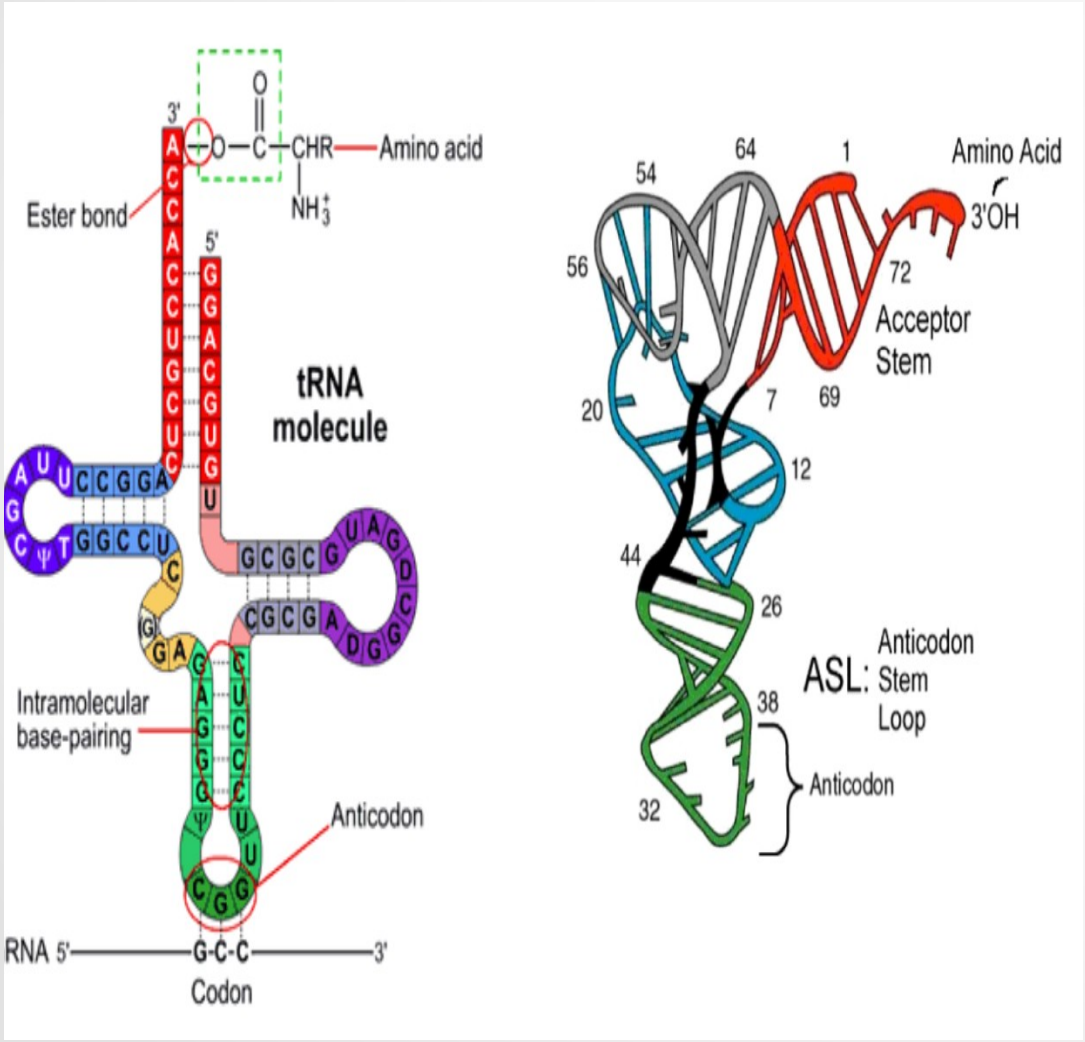
Main problem of gene definition:

How to unify phenotype effect, molecular basis and enhiritance

Classification of genes



tRNA



Features of tRNAs:

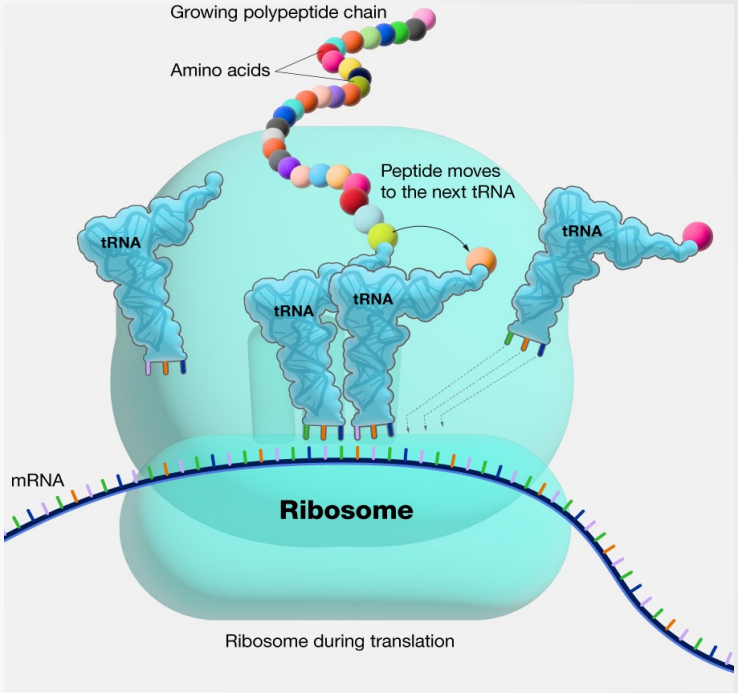
- Conservative secondary and tertiary structures
- Modified bases
- Necessity of intron presence for mature (mainly for modification of nucleotides)
- pre-tRNAs are transcribed by RNApol III

Ribosomal RNA (rRNAs)

Taxon	rRNA organization
Bacteria	Operon(16S/23S/5S)
Eukaryota	Operon(18S/28S/5.8S) + 5S
Mitochondria	12S + 16S

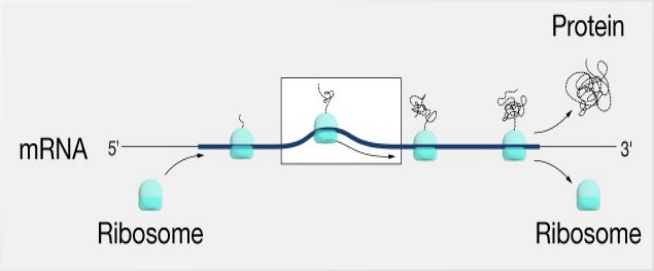
Features

- Most abundant RNA in cell;
- Part of ribosome
- Responsible for protein synthesis.
- Many copies
- High GC content – very difficult to sequence
- only fragments are present in assembly
- Is transcribed by RNA pol I

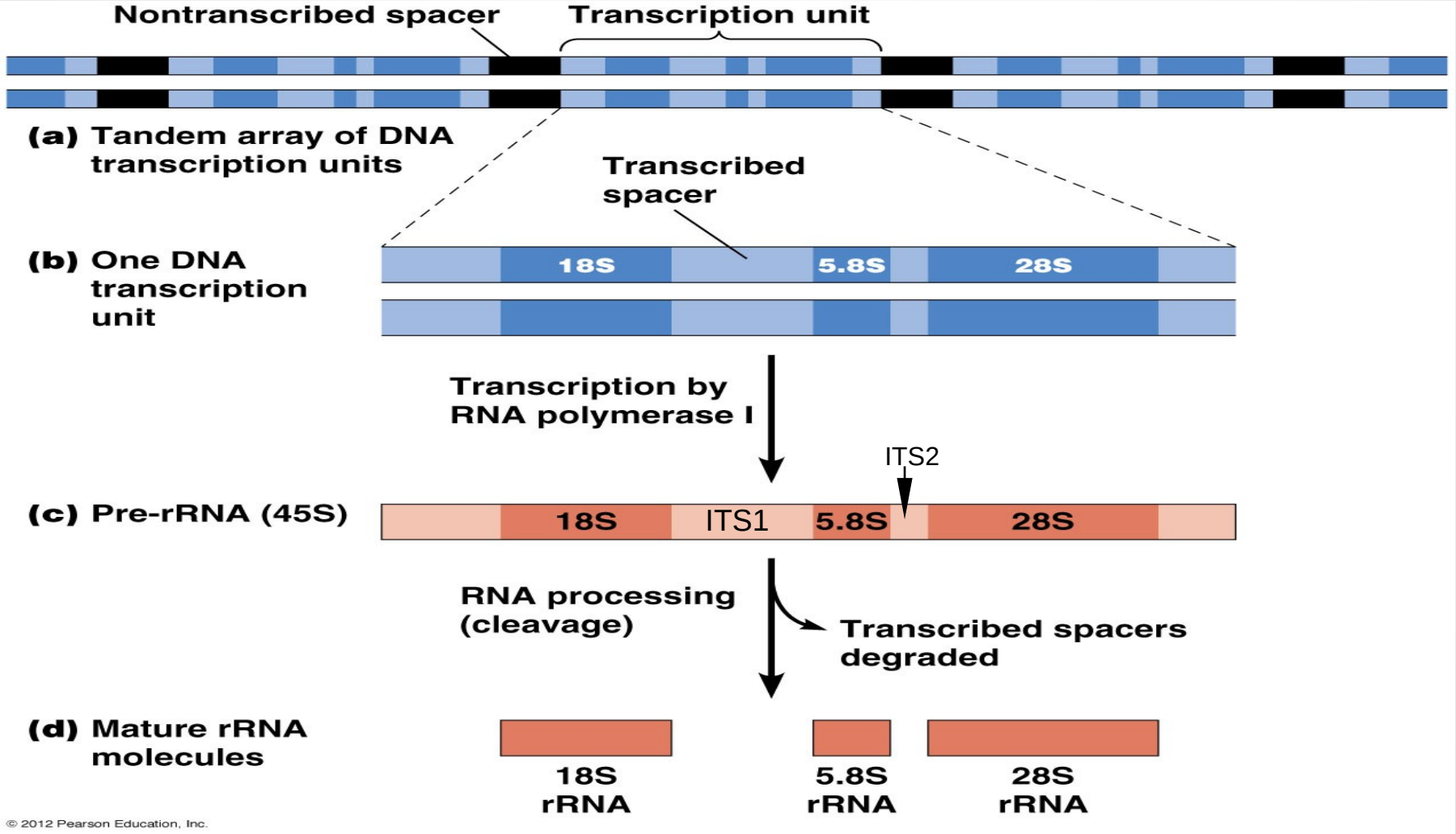


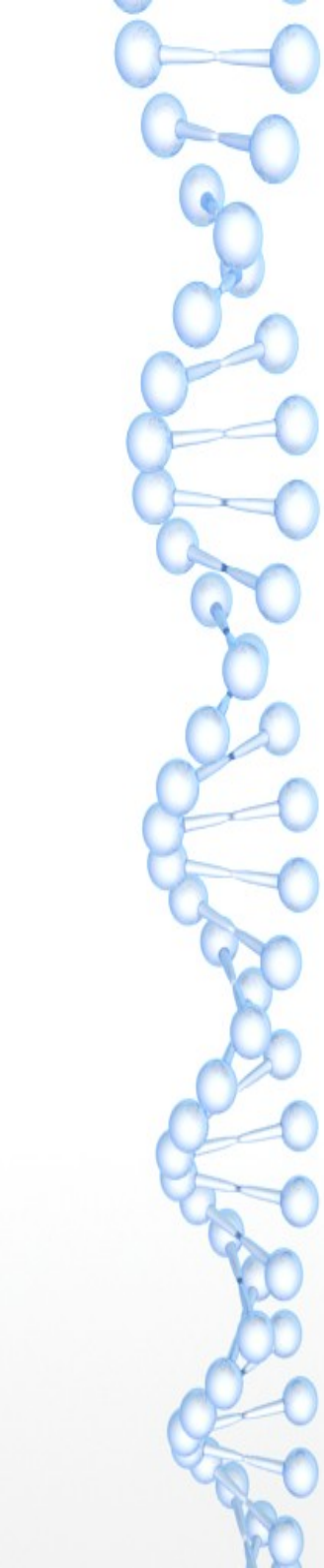
rRNAs and ribosome

Type	Size	Large subunit	Small subunit
Bacterial	70S	50S (5S, 23S)	30S (16S)
Eukaryotic	80S	60S (5S, 5.8S, 28S)	40S (18S)



Processing of ribosomal operon

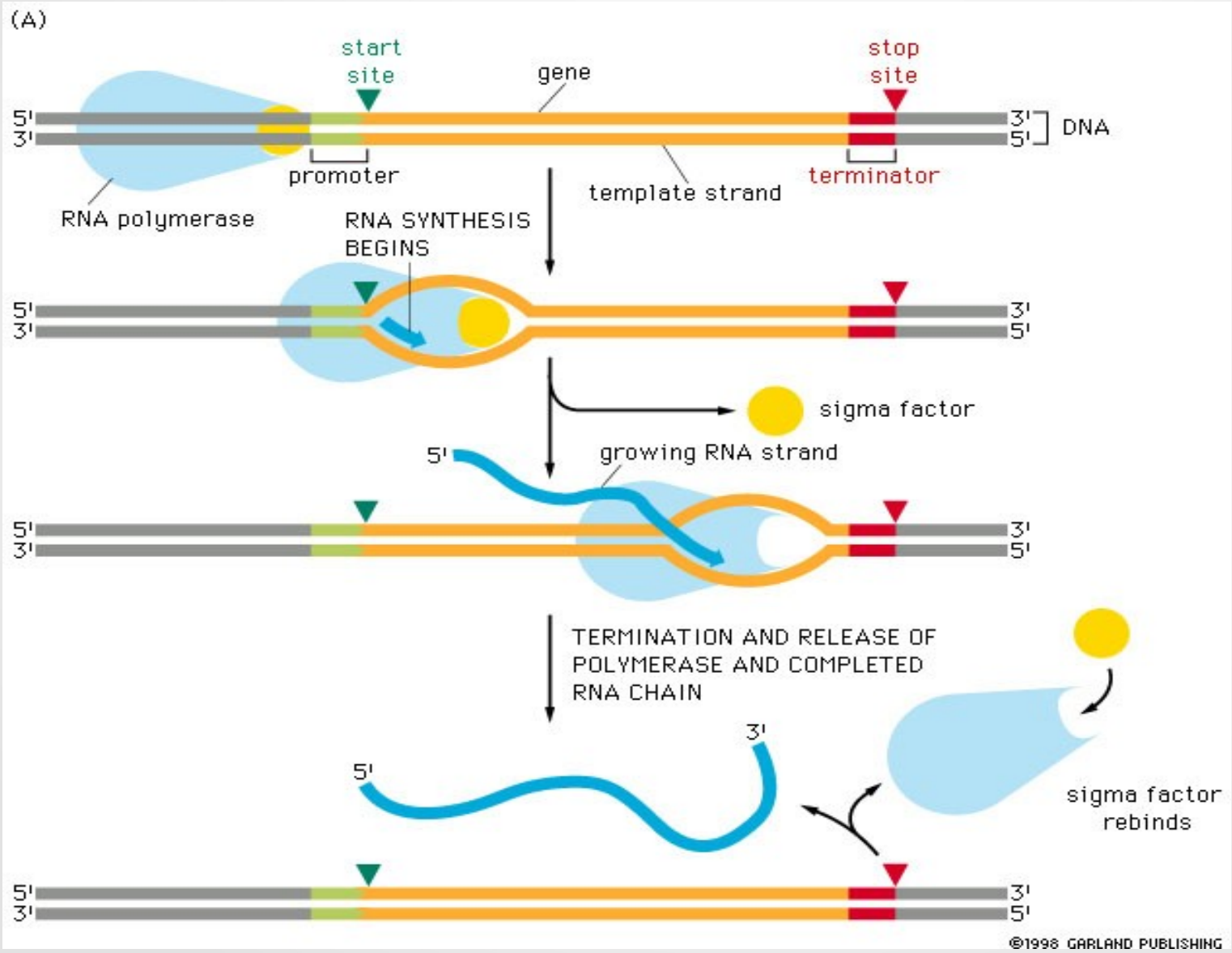




II. Gene and genetic code

Structure of protein coding genes

Transcription



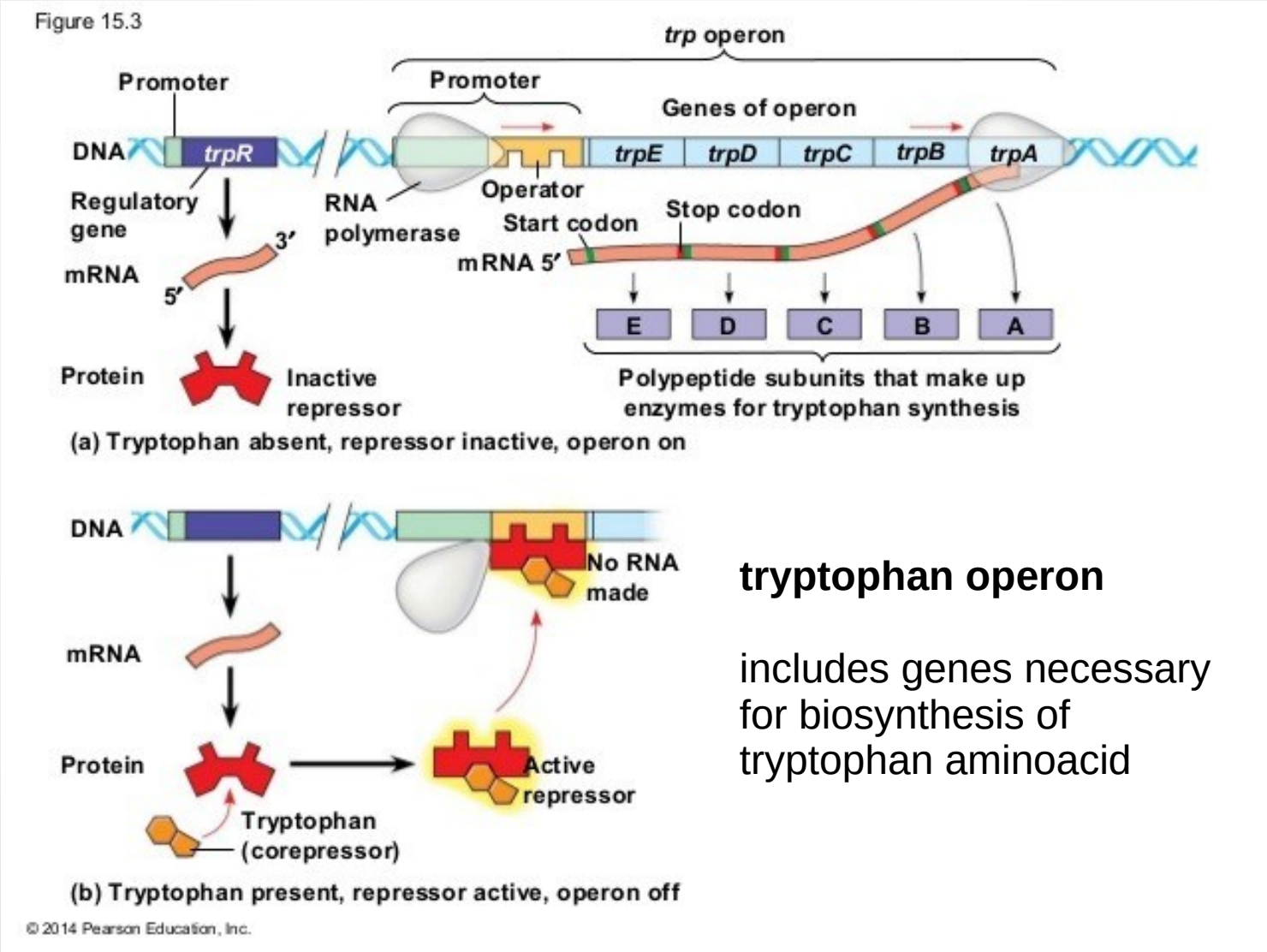
3 phases:

I. Initiation

II. Elongation

III. Termination

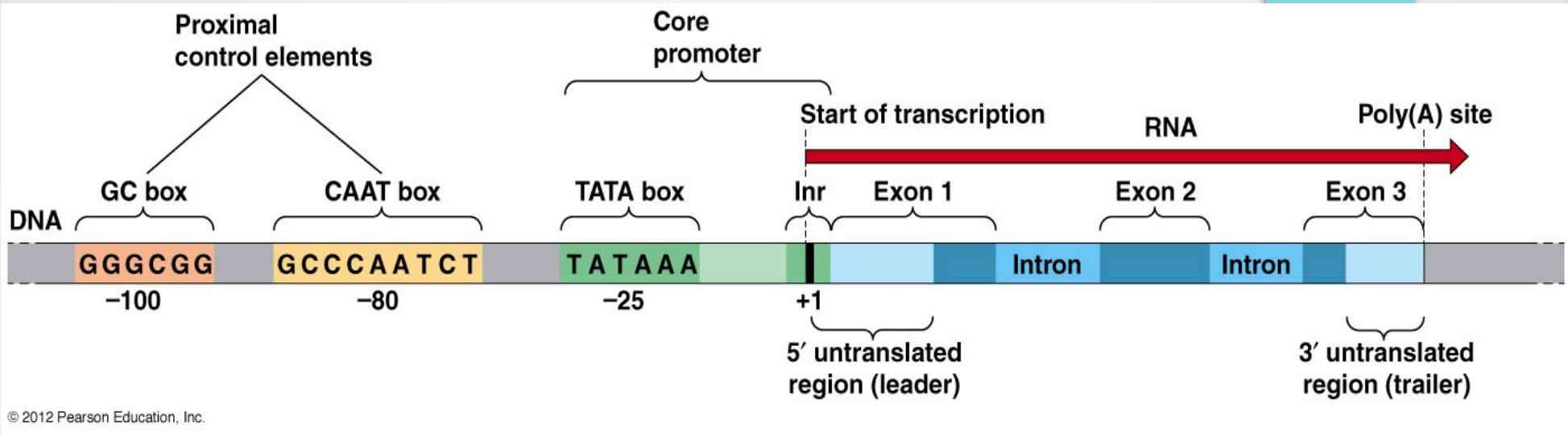
Protein-coding genes in Prokaryotes



tryptophan operon

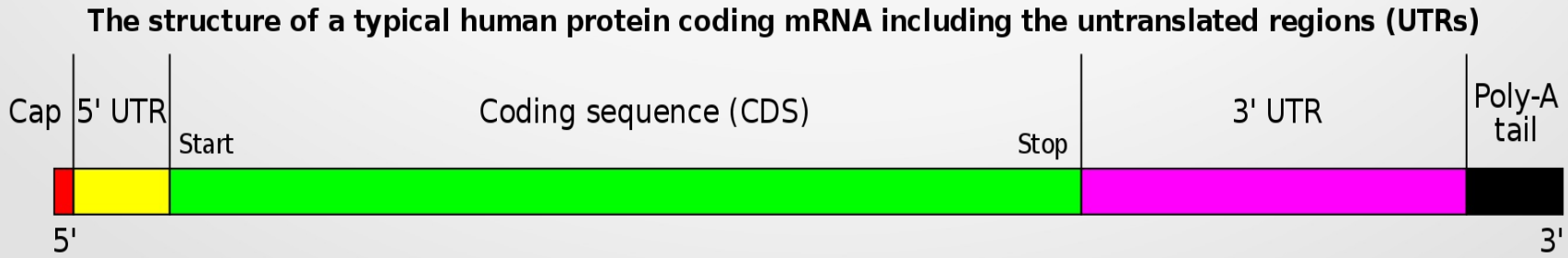
includes genes necessary for biosynthesis of tryptophan aminoacid

Protein-coding genes in Eukaryota

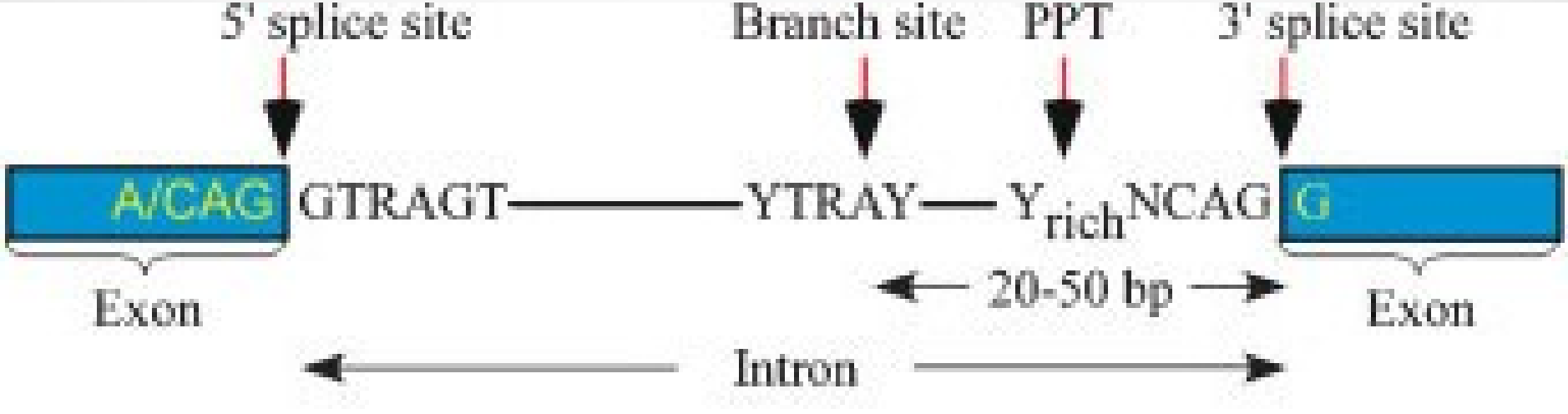


Major “post”-transcriptional modification of mRNAs:

- 5’ capping: 5-methyl-guanine is added to 5’ end of transcript
- splicing: excision of introns
- 3’ polyadenilation: addition multiple adenine stretch to 3’ end of transcript

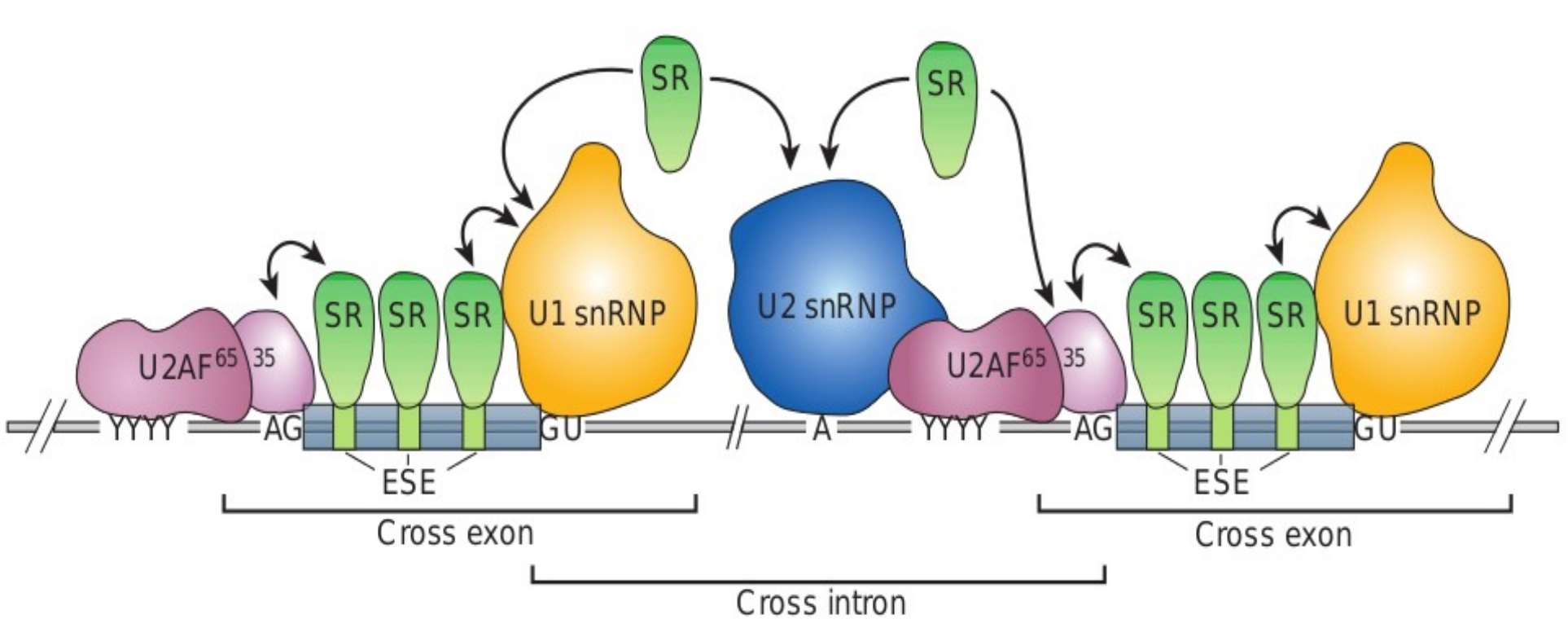


Splicing signals



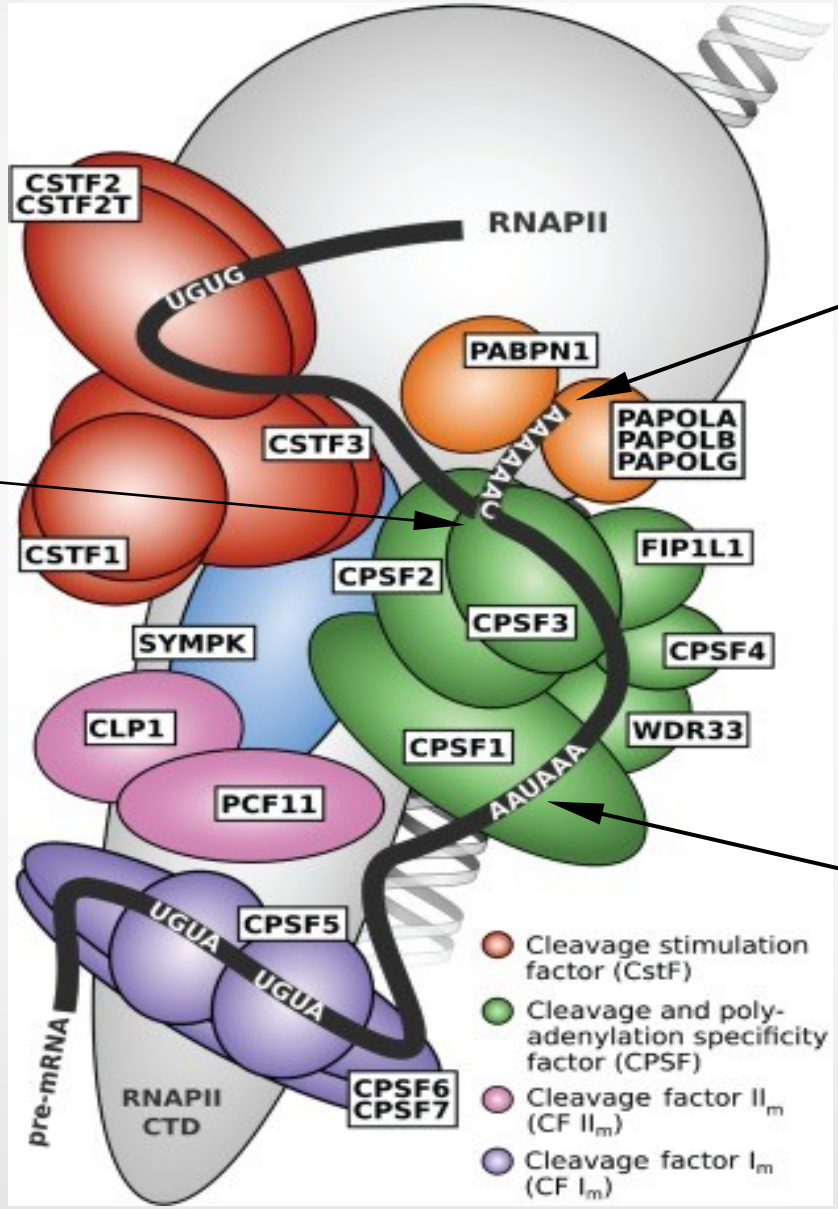
Kim et al, 2008

Recognition of splicing signals



Maniatis and Tasic, 2002

Polyadenilation

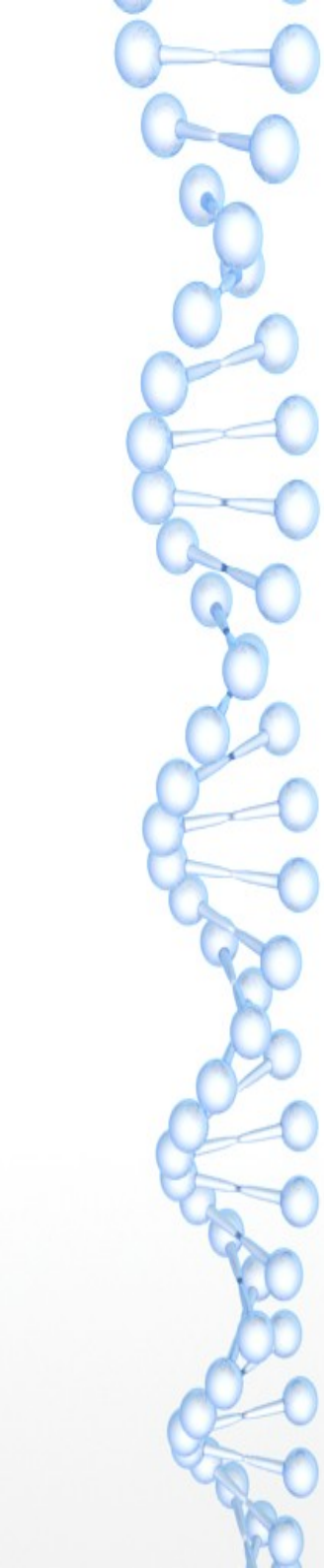


growing polyA tail

cleavage site

polyadenilation site (PAS)

Gruber et al, 2014



II. Gene and genetic code

”Alternative events”

G-value paradox

	Number of protein coding genes
Saccharomyces cerevisiae (yeast)	~6000
Caenorhabditis elegans (flat worm)	~20500
Drosophila melanogater	~14000
Homo sapiens (human)	~20000
Gallus gallus (chicken)	~20000
Arabidopsis thaliana (plant)	~25000

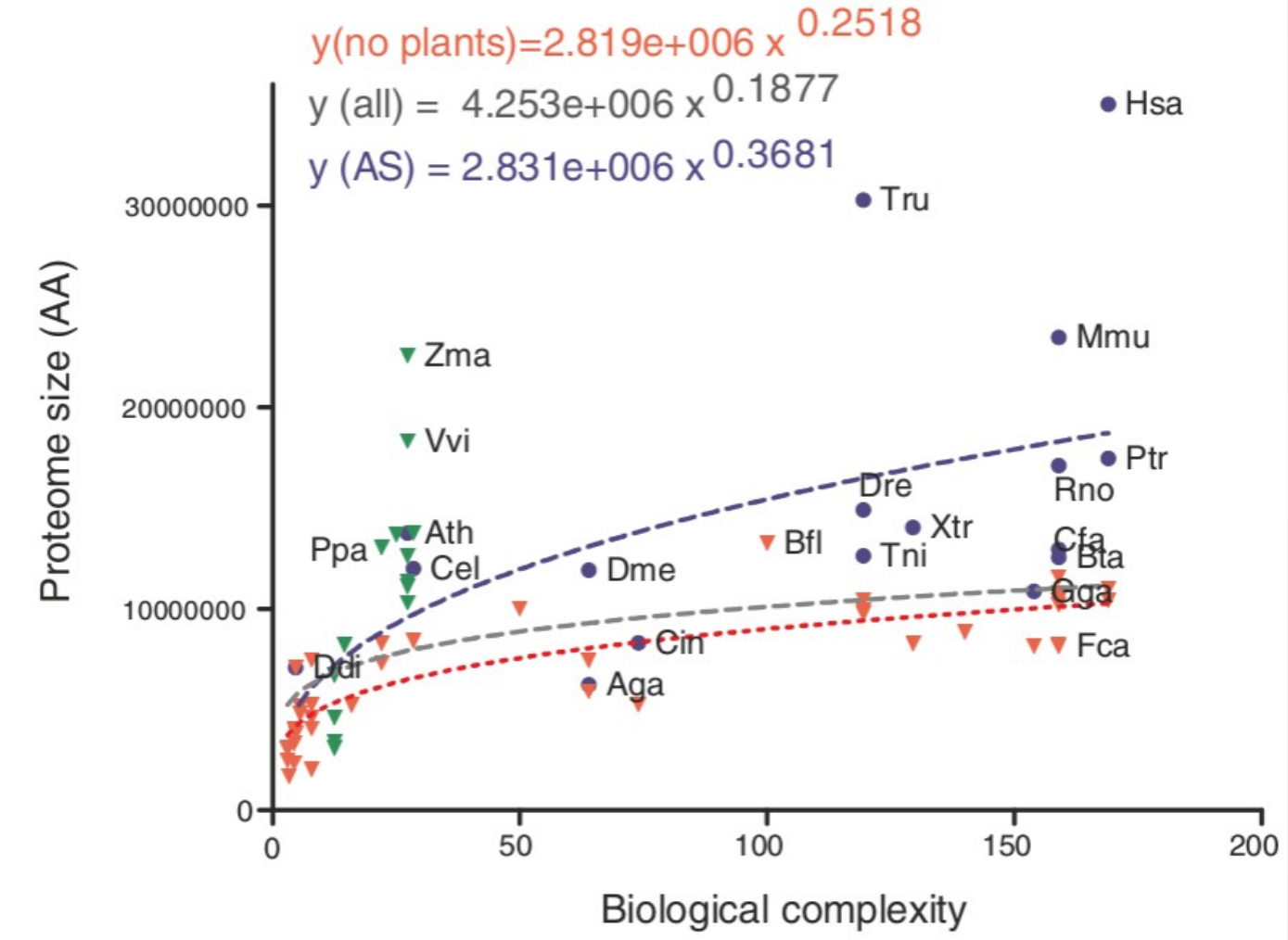
G-value paradox - absence of correlation between biological complexity and number of genes

Possible solutions for G-value paradox

Without increasing number of genes, complexity of organism might be increased by:

- complication of gene expression regulation networks by increasing the number of transcription factors and non-coding RNAs (Chen and Rajewsky, 2007, Levine and Tjian, 2003)
- acquisition of additional functions by genes
- significant increase in transcriptome and proteome size (Schad et al, 2011; Kim et al, 2008)

Solution of G-value paradox for animals



Shad et al, 2011

Proteome size (here) - total number of aminoacids in all proteins of organism

Biological complexity (here) - number of cell lines

With plants

$R^2 = 0.1333$

p-value = 0.0072

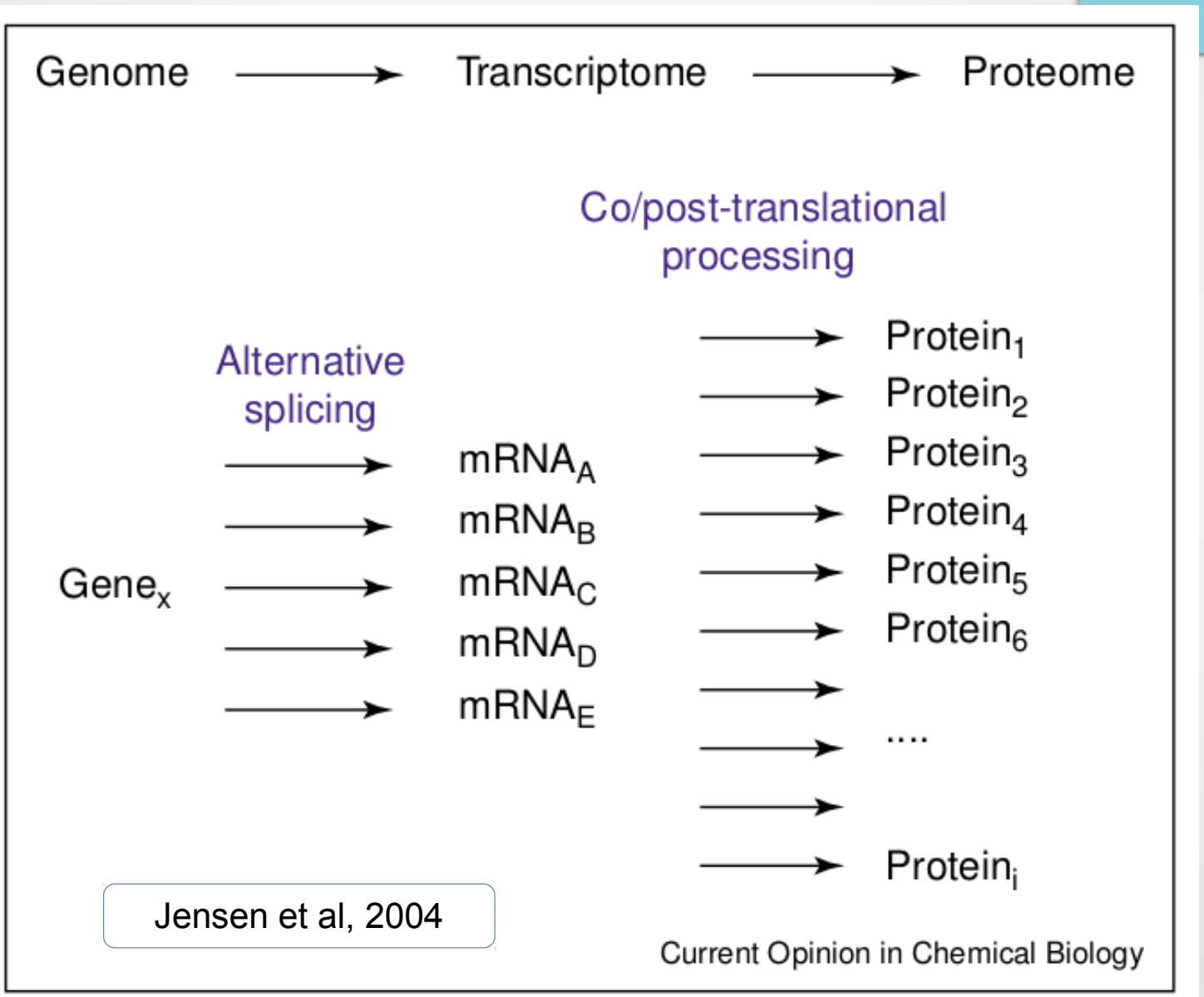
Without plants

$R^2 = 0.6326$

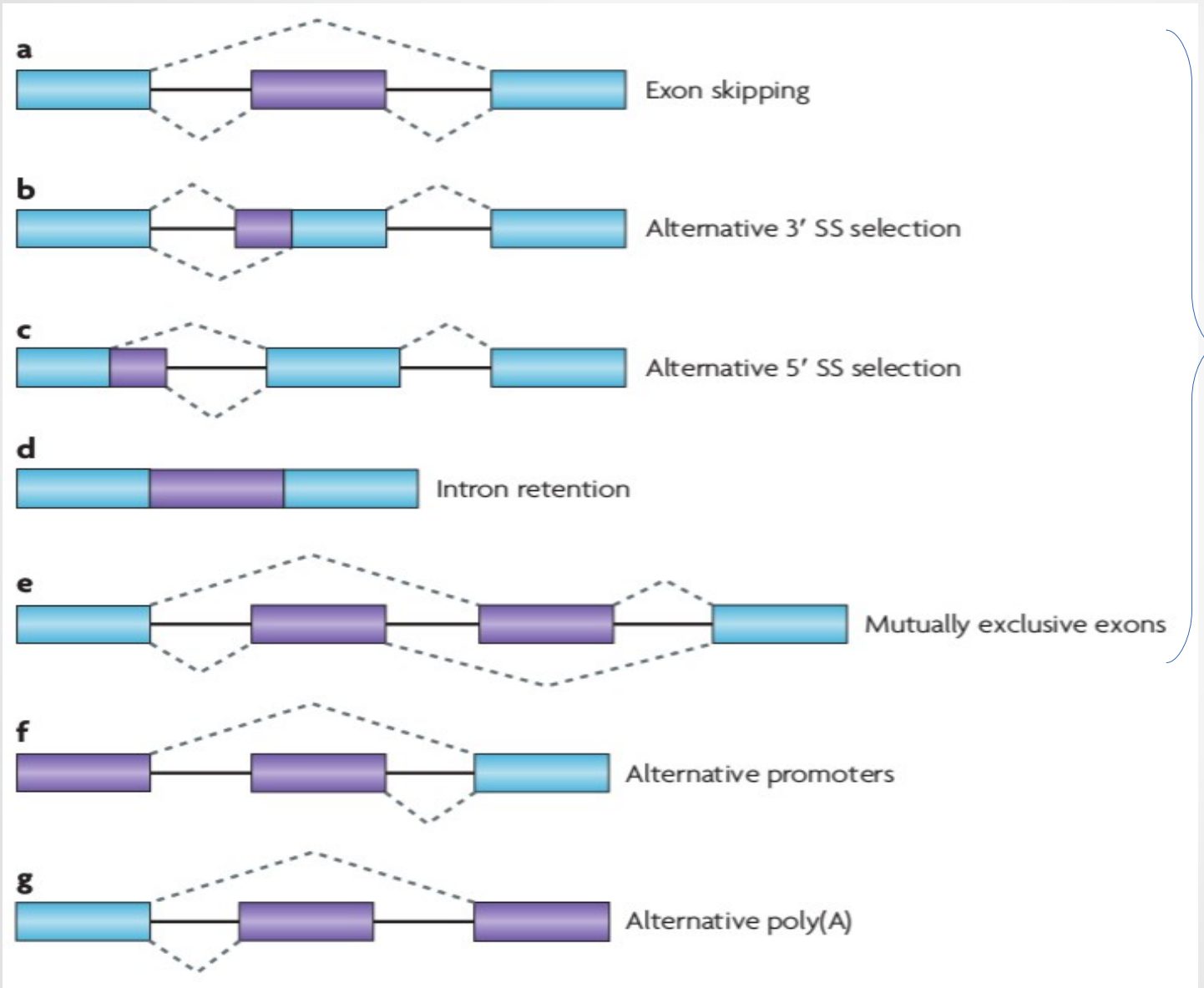
p-value < 0.0001

AS - curve for data including proteins generated by alternative splicing

From genome to proteome



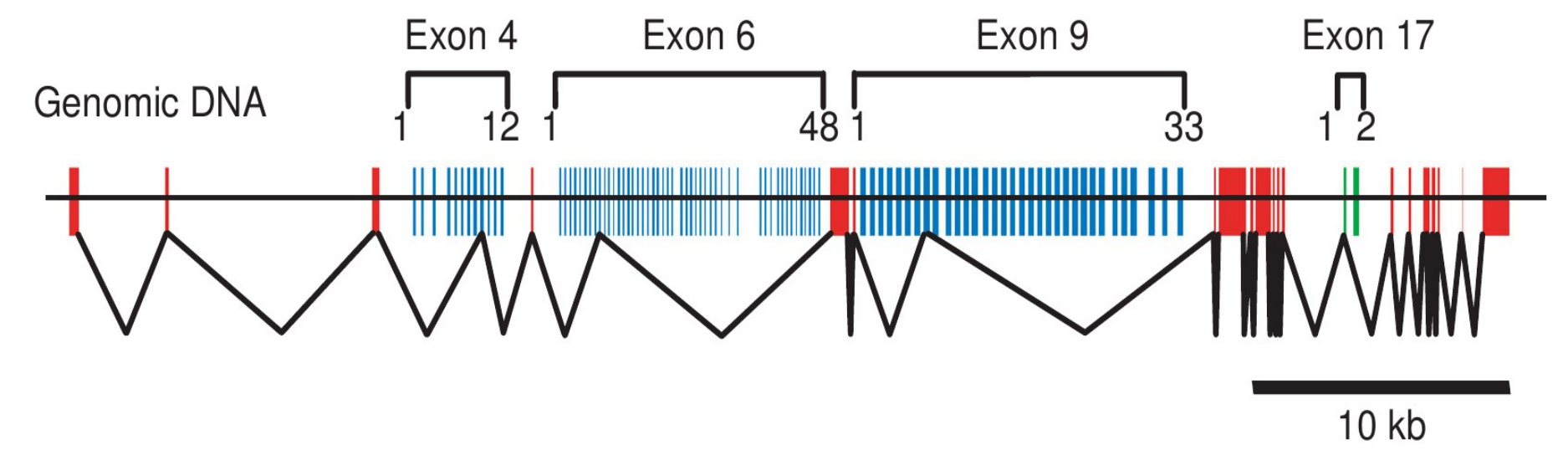
Types of “alternative” events



alternative splicing

Keren et al, 2010

Case of drosophila *DSCAM* gene

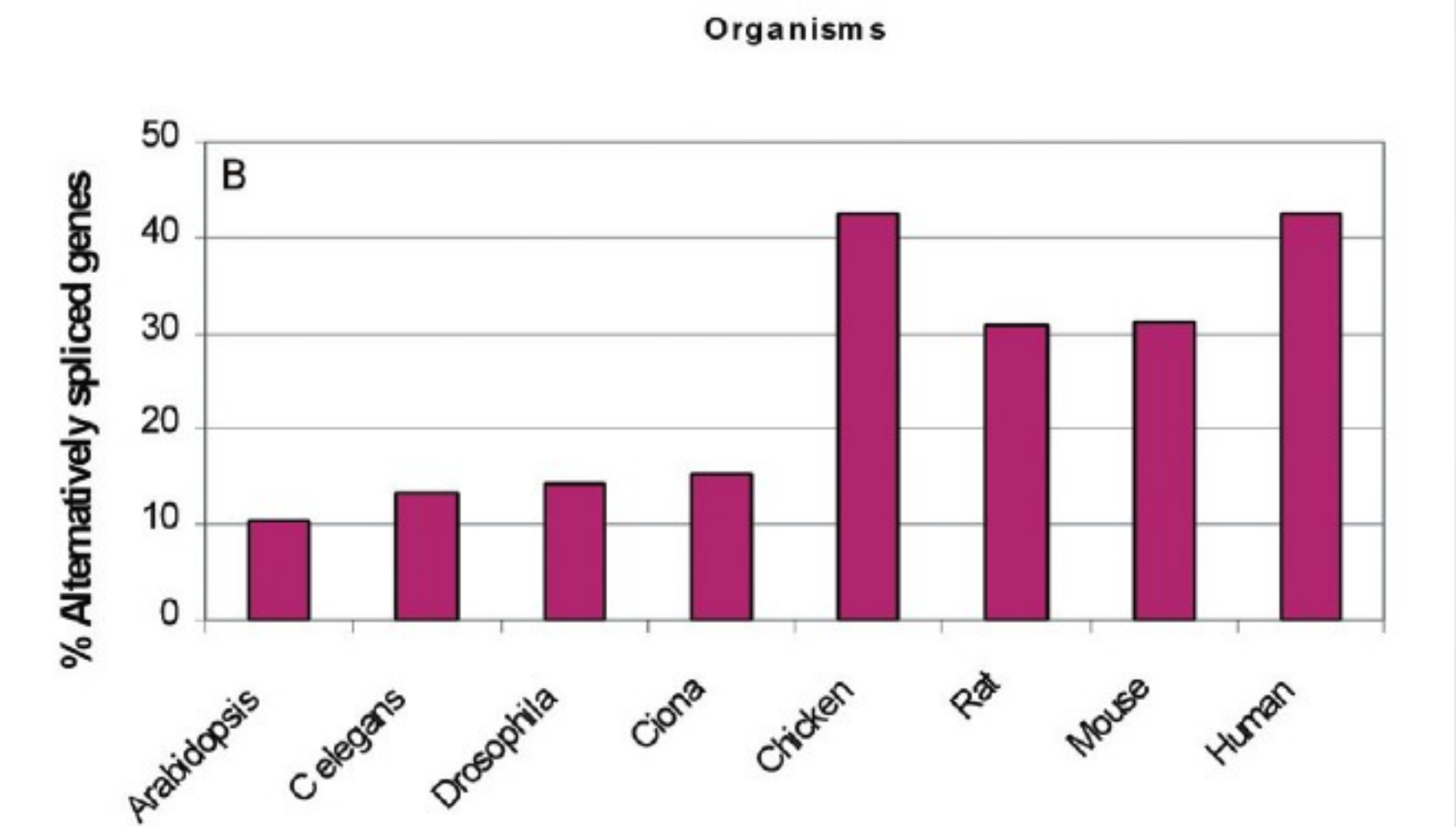


4 exon cassettes

12*48*33*2 = 38016 potential transcripts

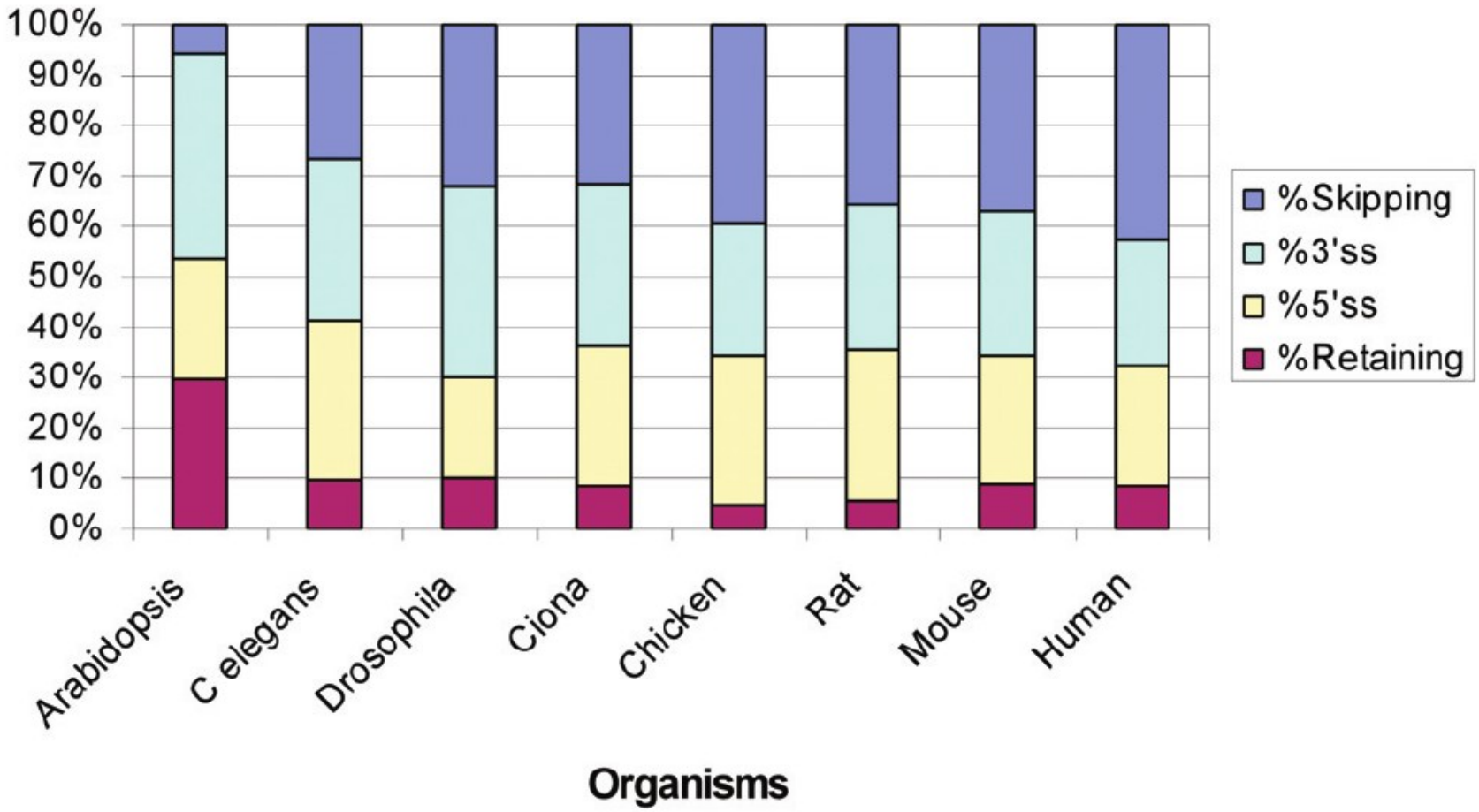
Neves et al, 2004

“Alternative” splicing is widespread



Kim et al, 2007

Different types of “alternative” splicing are preferred in different organisms

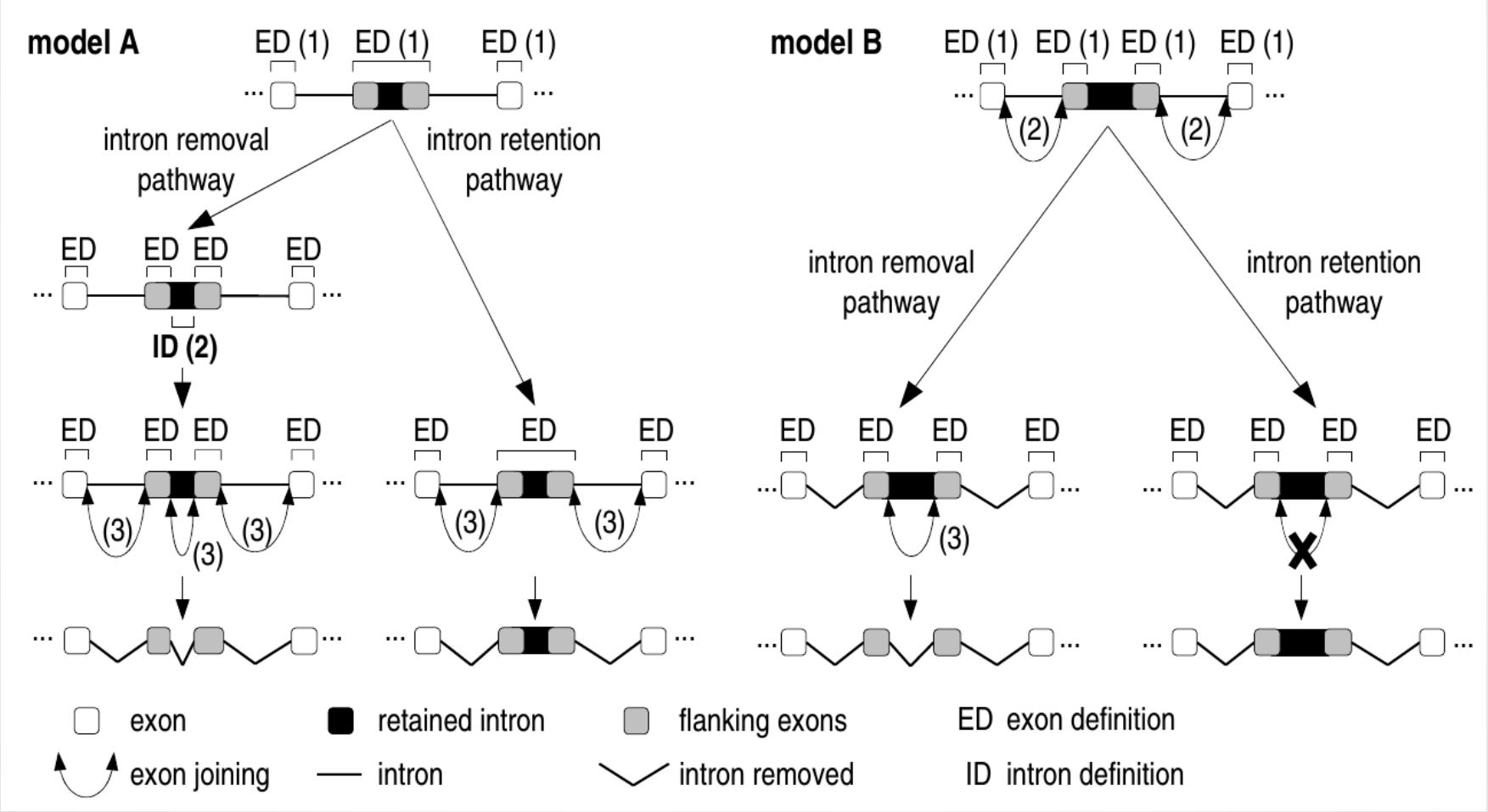


Kim et al, 2007

Intron retention models

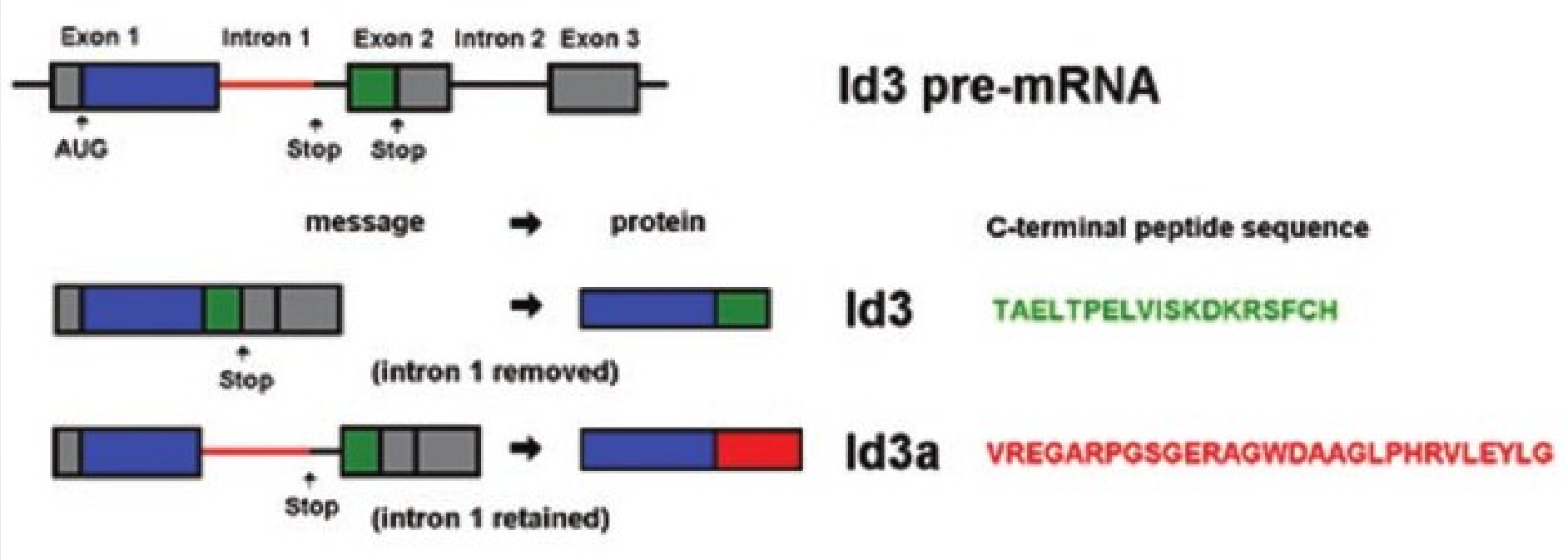
intron + flanking exons < 400 bp

intron + flanking exons > 400 bp



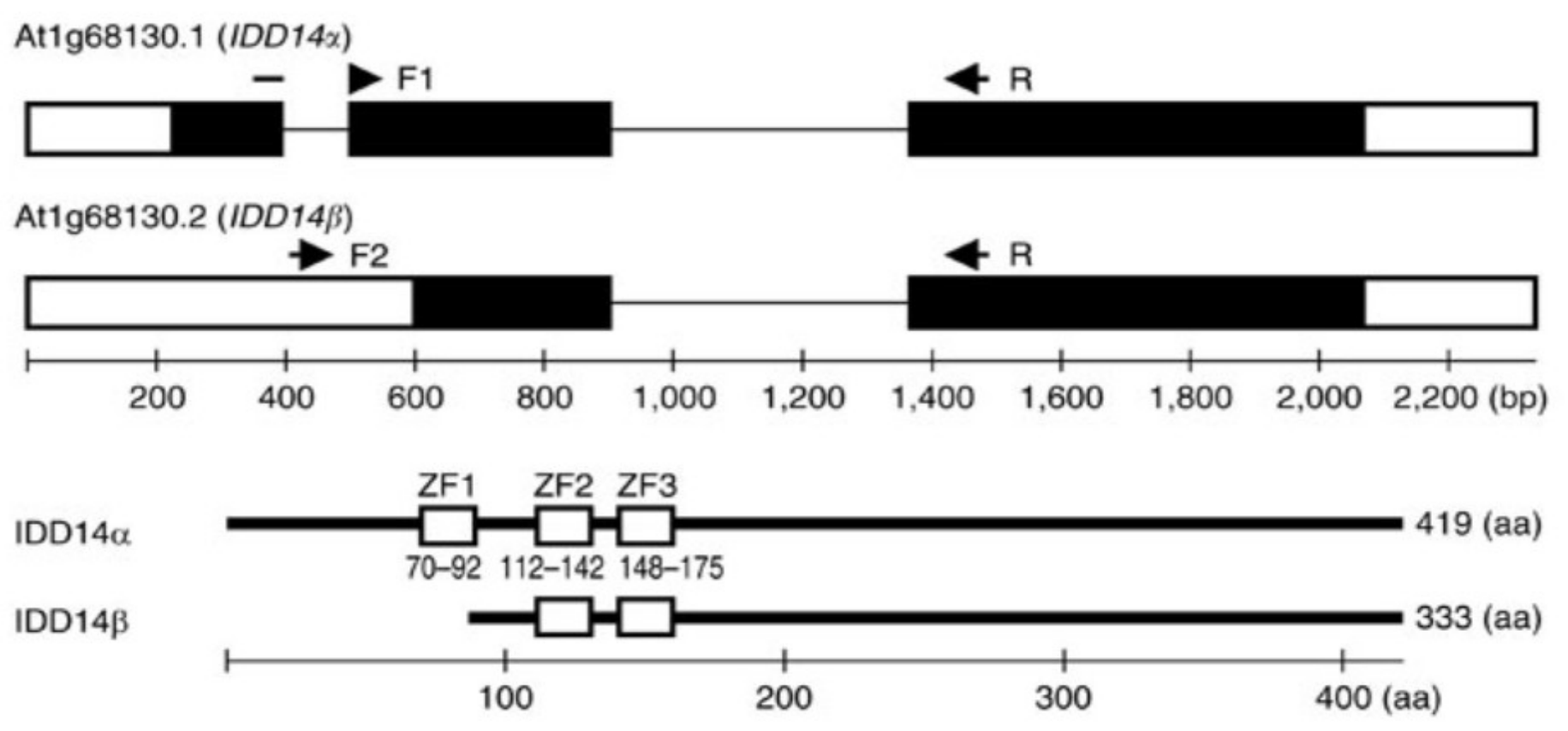
Sakabe et al, 2007

Intron retention in Id3 gene of mouse



Sakabe et al, 2007

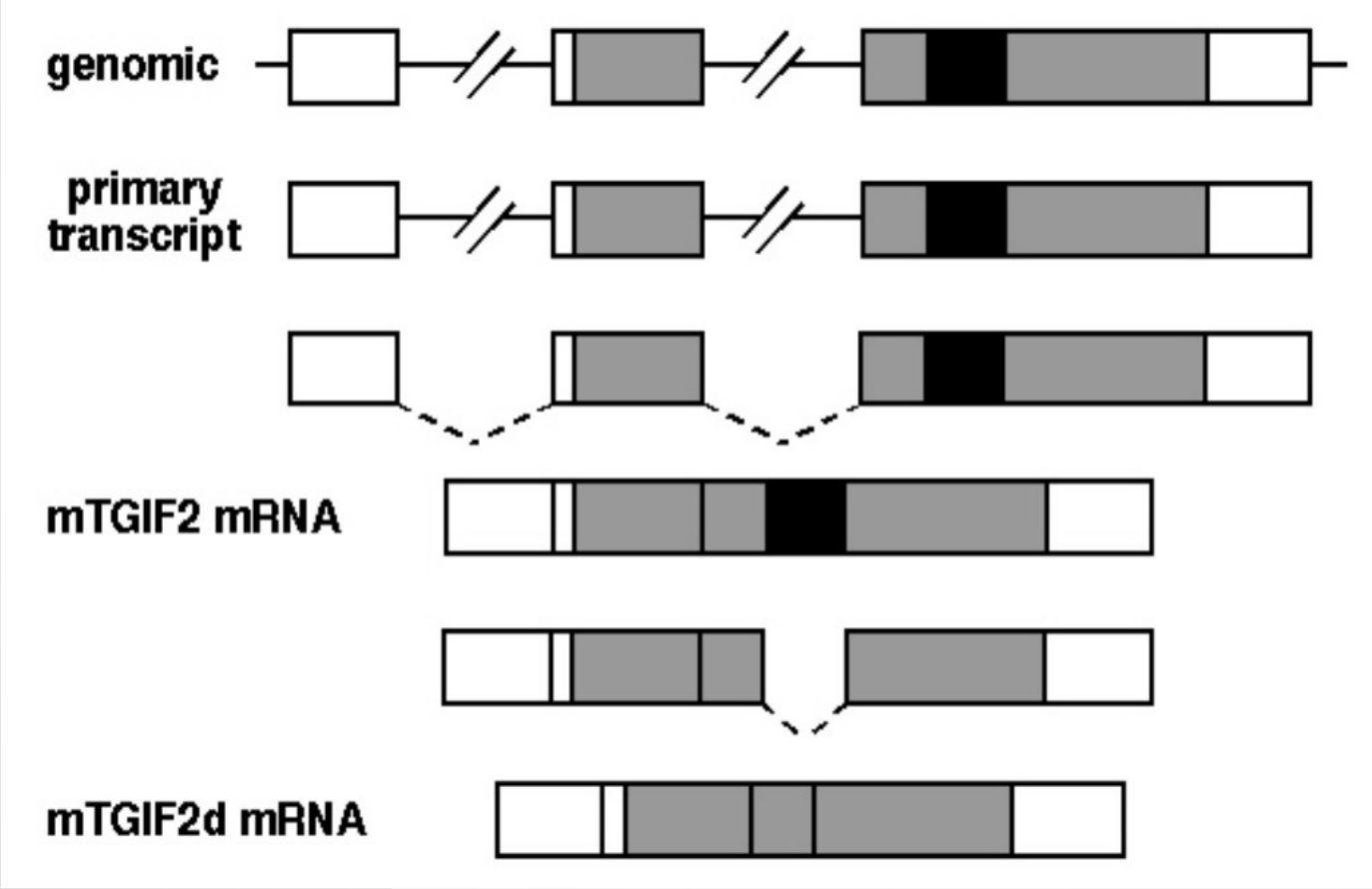
Intron retention in IDD14 gene *Arabidobsis thaliana*



Unique case: intron retention results in shortening of protein from N-terminus!

Seo et al, 2011

Intron retention in TGF2 gene



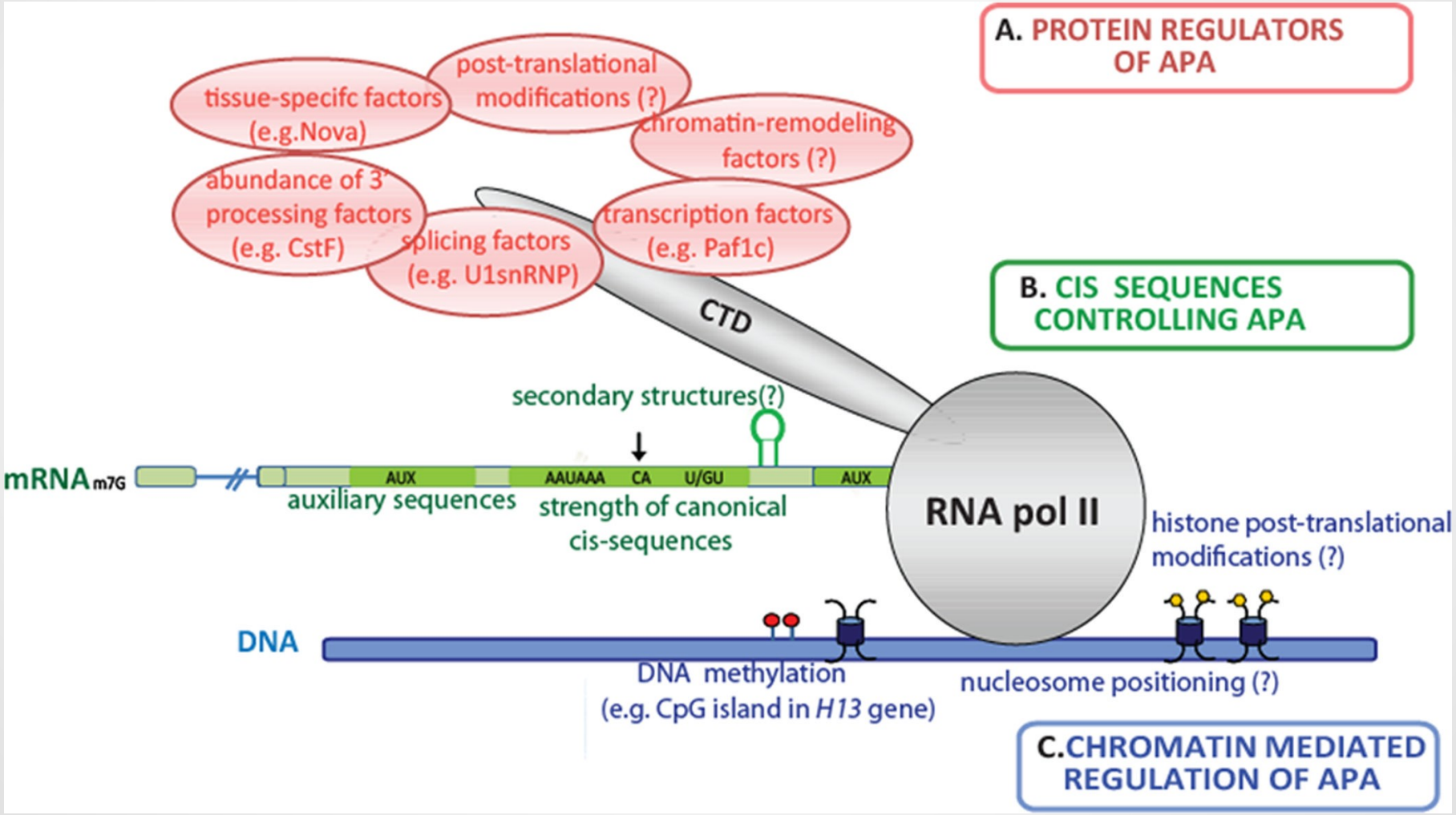
in mouse:
2 transcripts:
with intron
without intron

in human:
1 transcript:
with intron!

Is this intron in mouse a novel invention?

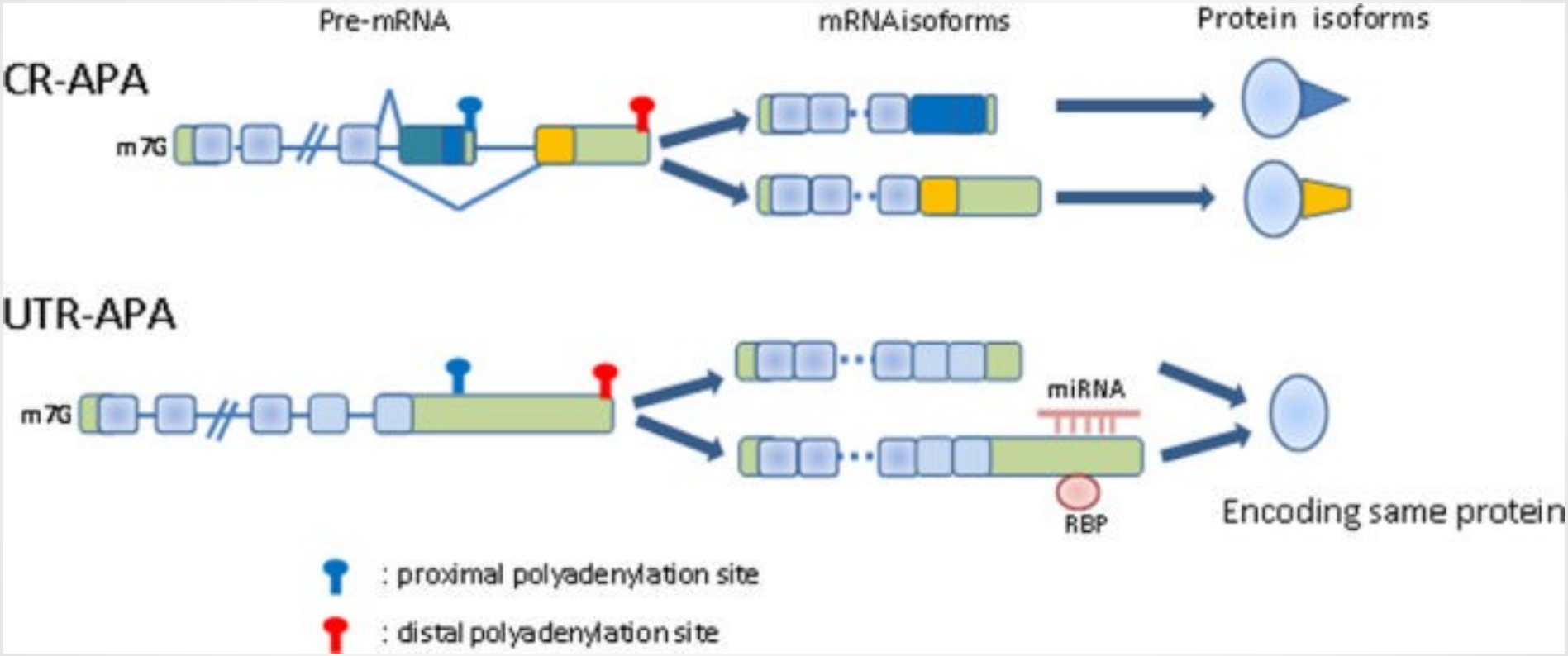
Melhuish and Wotton, 2006

Factors controlling alternative polyadenilation



Giammartino et al, 2011

Types of alternative polyadenilation (APA)



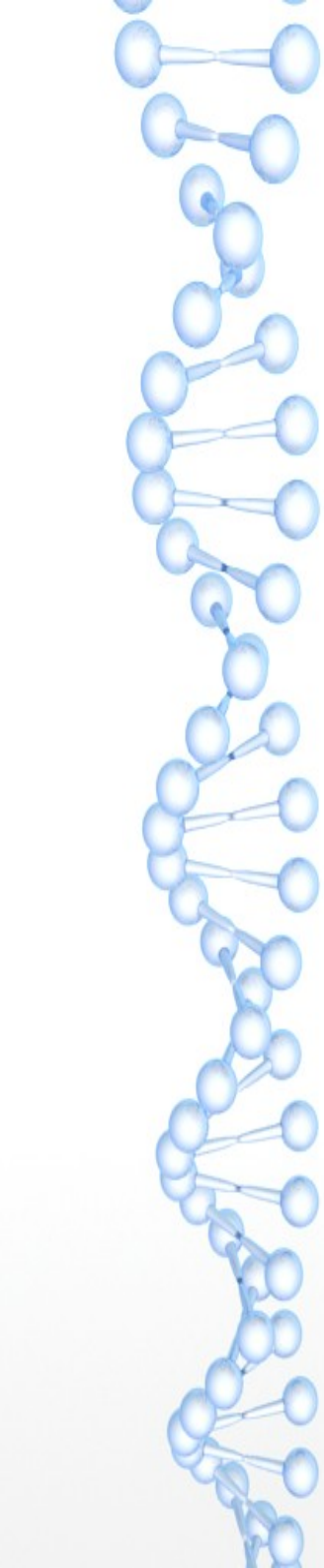
CR-APA - APA in coding region

UTR-APA - APA in untranslated region

Giammartino et al, 2011

Summary

- Alternative events (AR) during transcription are not alternative.
- AE are common and widespread.
- AE are responsible from transition of ~20k genes to 100k+ transcripts
- Sometimes difference between exon and intron is very small



II. Gene and genetic code

Genetic code
Features

Features of genetic code

Classical view

Genetic code is

1. Degenerated (redundant)
2. Triplet
3. Continuous
4. Unambiguous
5. Non-overlapping
6. Unidirectional
7. Universal

Modern view

Genetic code is

1. Degenerated (redundant)
2. Triplet
3. (Quasi)continuous
4. Quasiunambiguous
5. Quasinon-overlapping
6. (Quasi)unidirectional
7. Quasiuniversal

Genetic code is quasiuniversal

Standard

```
  AAs  = FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = ---M---------------M---------------M----------------------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

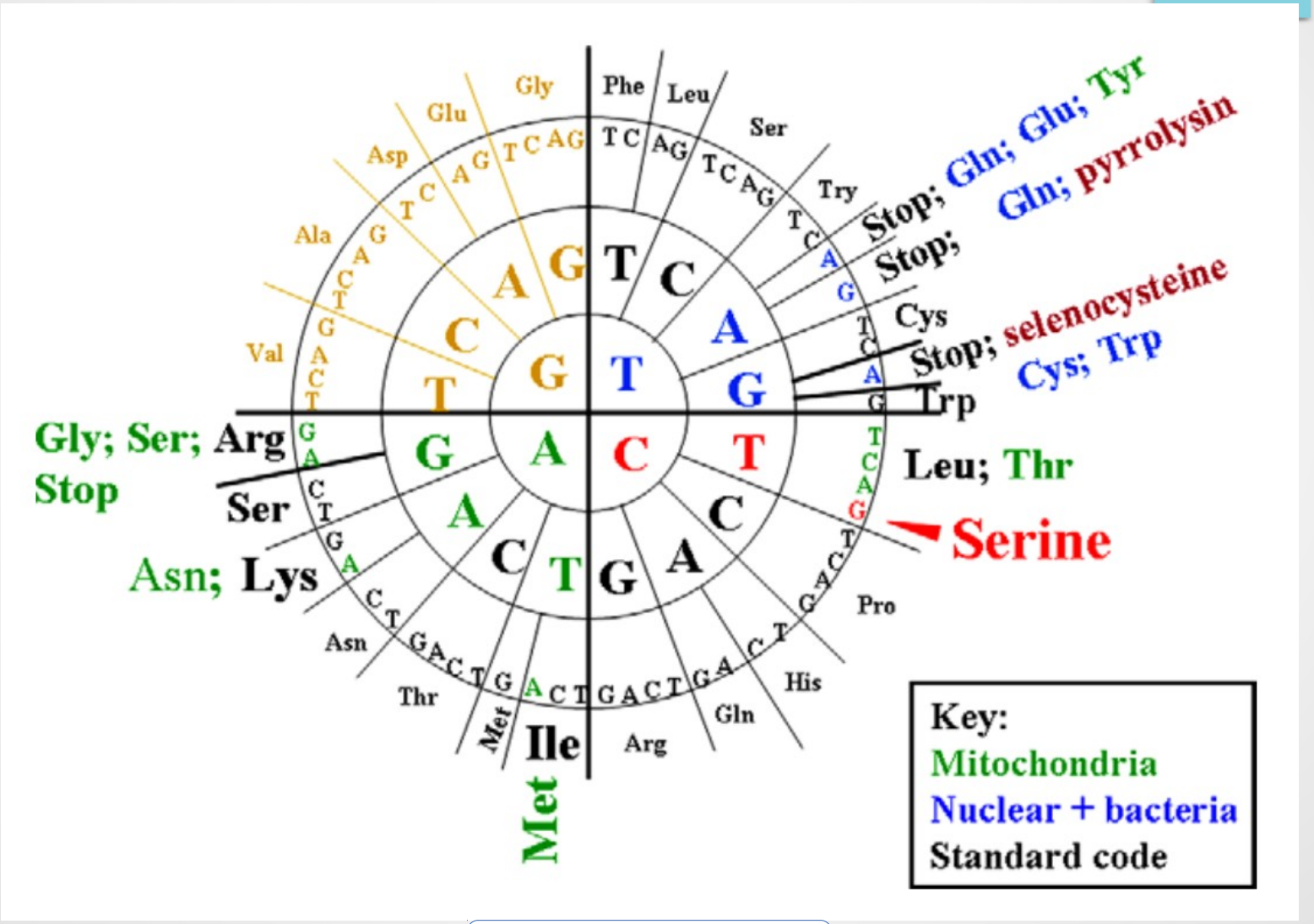
Mitochondrial *Vertebrate*

```
  AAs  = FFLLSSSSYY**CCWWLLLLPPPPHHQQRRRRIIMMTTTTNNKKSS**VVVVAAAADDEEGGGG
Starts = --------------------------------MMMM---------------M------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Nuclear *Ciliate*

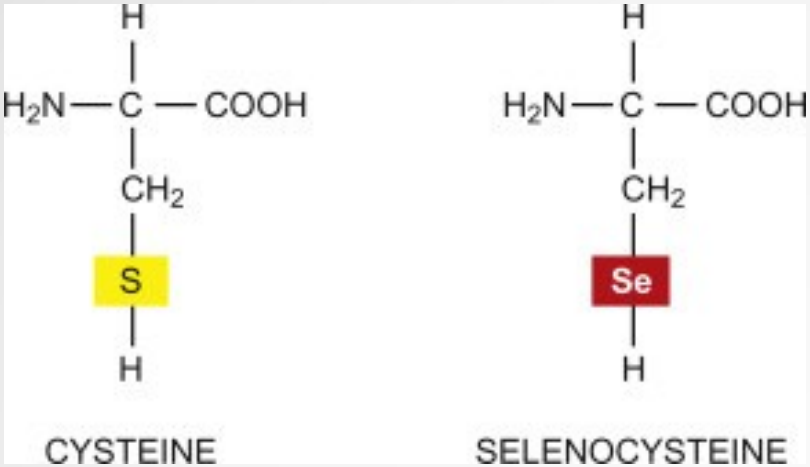
```
  AAs  = FFLLSSSSYYQQCC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
Starts = --------------------------------M----------------------------
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG
```

Different variants of genetic code

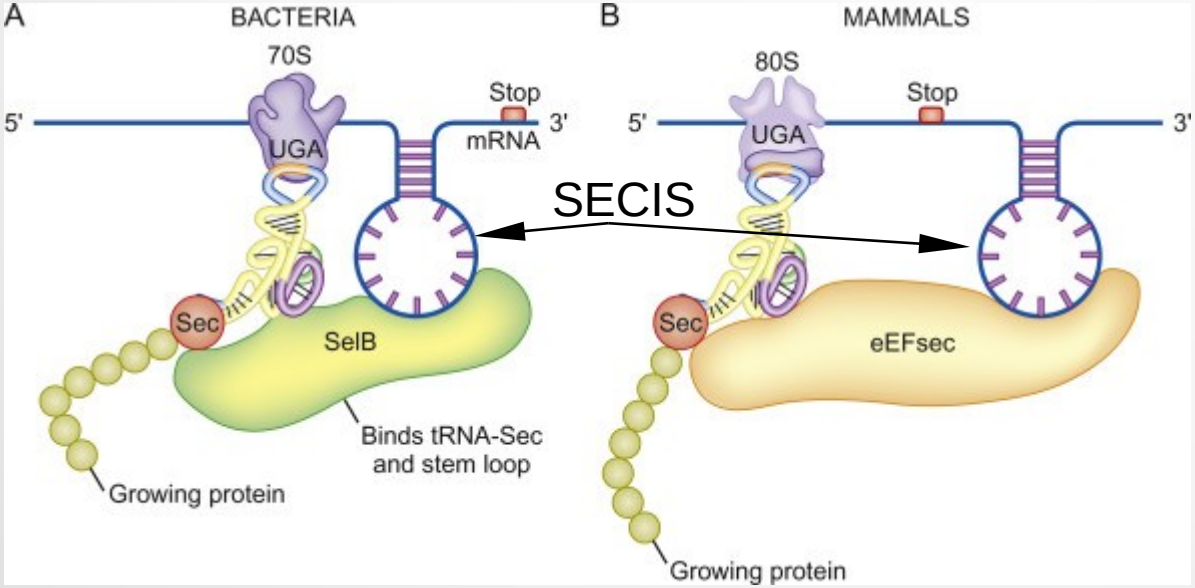


Moura et al, 2010

Noncanonical aminoacids: selenocystein (1)



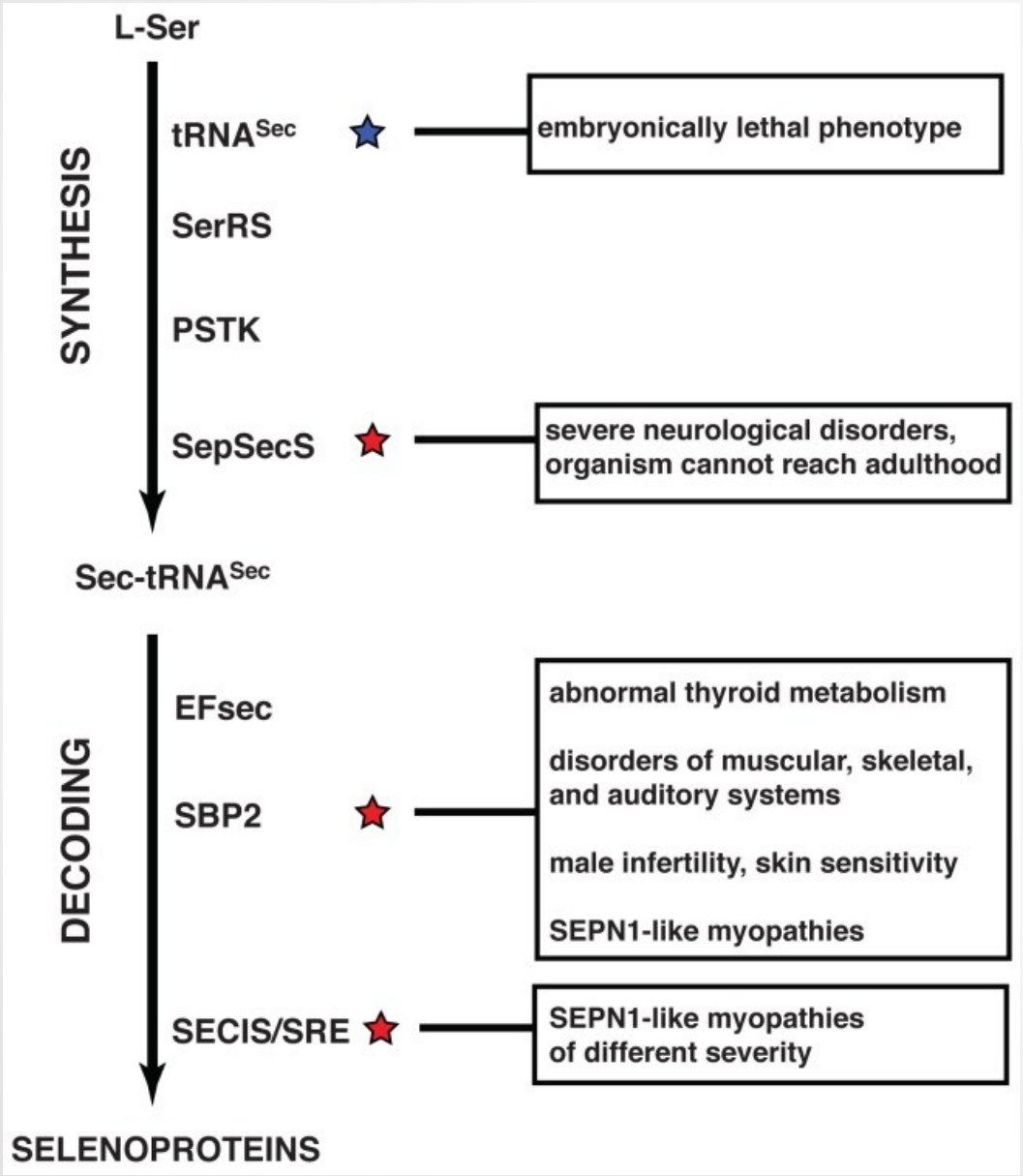
- Encoded by UGA codon, which usually it is serves as stop codon.
- SECIS (selenocysteine insertion sequence) is required to recognize UGA as selenocystein
- selenoproteins are present in multiple taxa, but absent in plants



zebrafish selenoprotein P contains 15 selenocysteins

Clark et al, 2019

Noncanonical aminoacids: selenocystein (2)



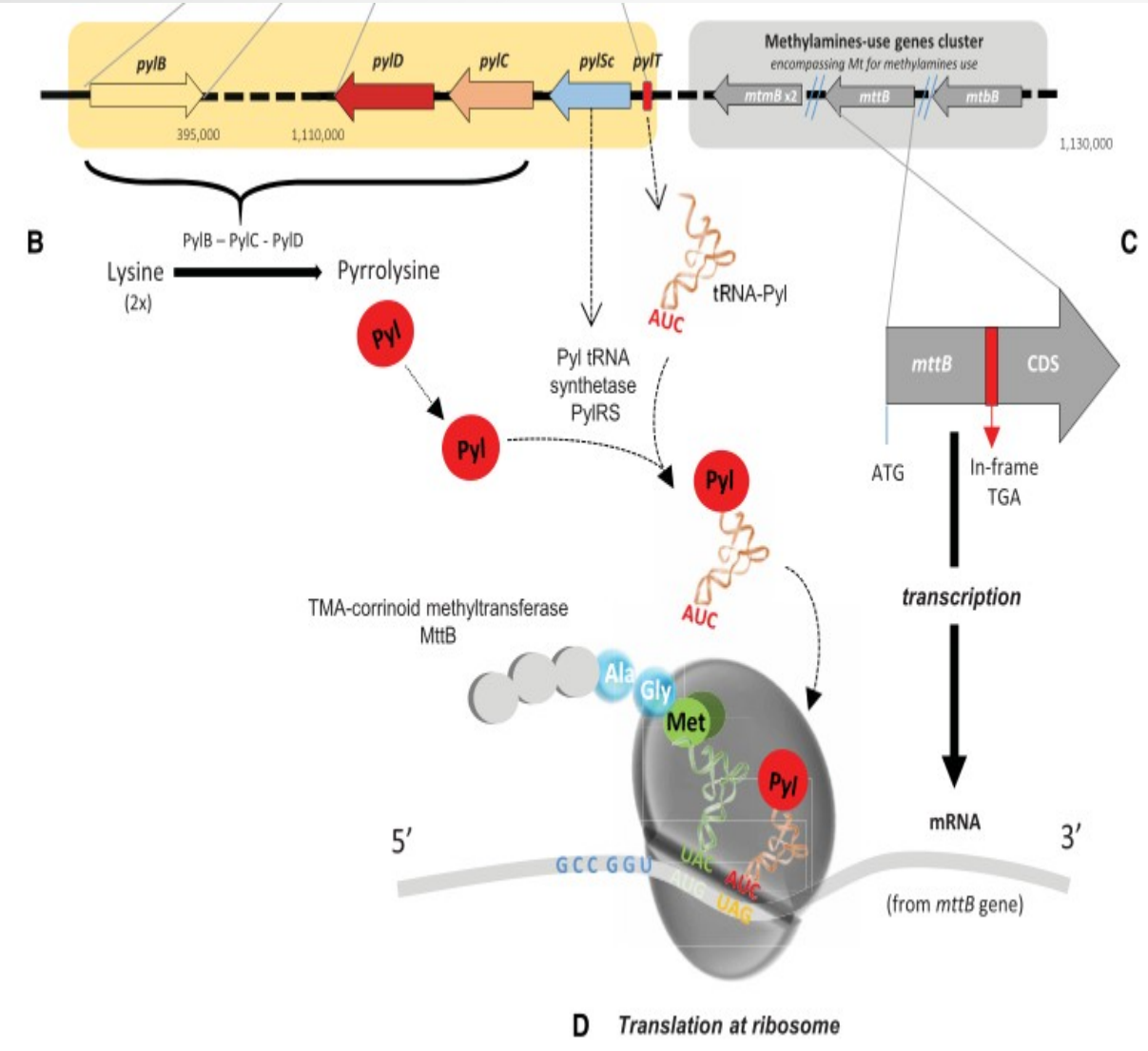
Human proteom includes >20 selenoproteins

Why selenocystein sometimes is required instead cystein is unclear

Mutation in genes related to synthesis of selenocystein and its recognition in mRNA often result in diseases

Schmidt and Simonović, 2012

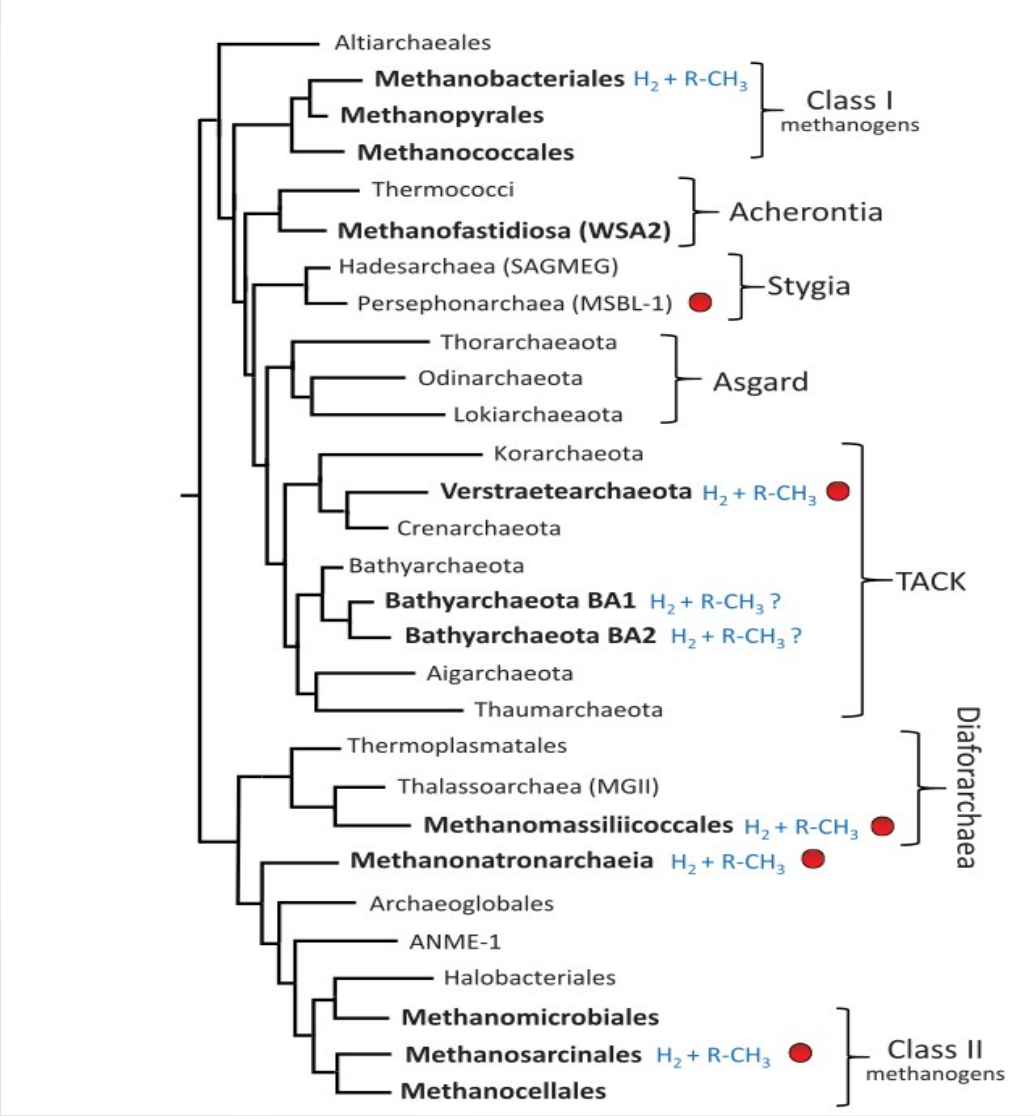
Noncanonical aminoacids: pyrrolysin (1)



- Encoded by UAG codon, which usually it is serves as stop codon.
- pyrroproteins are present in some Archea and in some Bacteria
- is linked to anaerobic methylamine metabolism

Schmidt and Simonović, 2012

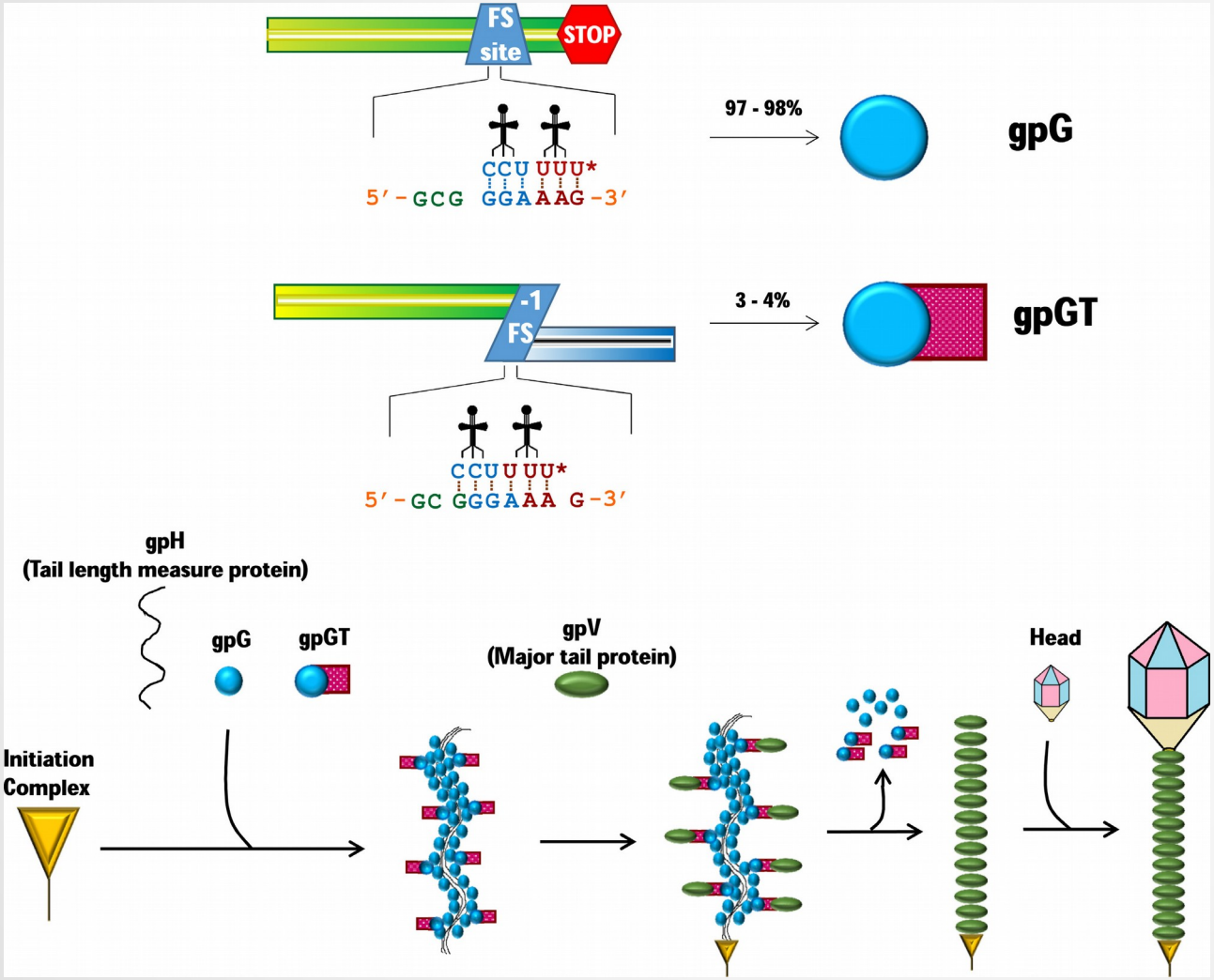
Noncanonical aminoacids: pyrrolysin (2)



- red dot - lineages with pyrroproteins
- bold - lineages possibly with pyrroproteins (predicted by protein homology)
- pyrrolysin is very ancient invention
- probably, it appeared **once** post-LUCA (last universal caommon ancestor), but very soon
- another hypothesis suggest that might appeared earlier in non-LUCA lineage (now extinct) and was transferred by HGT to descendants of LUCA

Brugere et al, 2018

Ribosome sliding



In some cases ribosome could slide 1 bp (+1) forward or 1 bp backward (-1) and continue synthesis.

For some genes it is obligatory.

Most common in viruses

Atkins et al, 2016

Ribosome sliding (FSDB database)

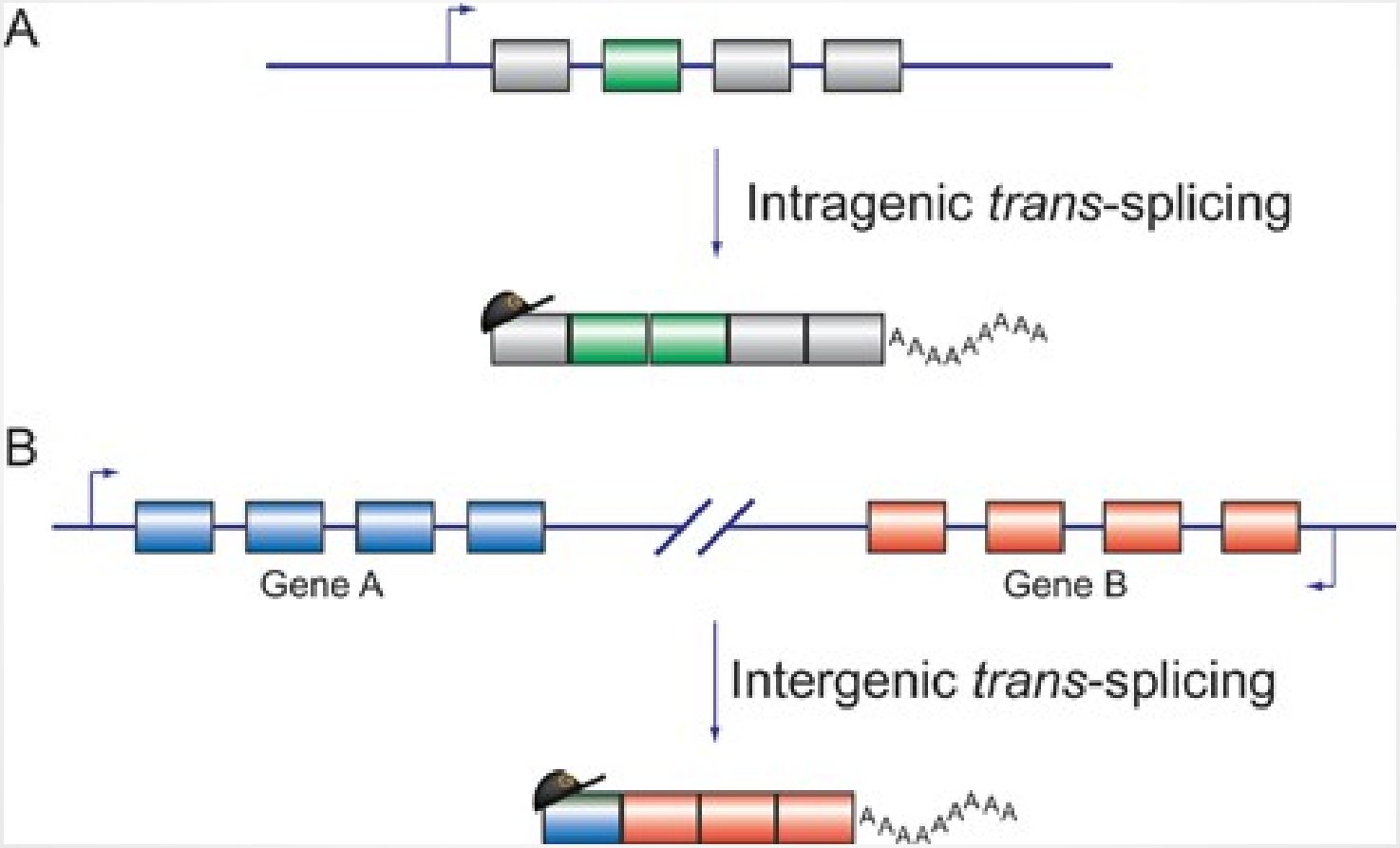
Current statistics (Blue and red numbers are clickable)

Type	Viruses		Prokaryota		Eukaryota		Total
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
-1 frameshifting	38	75	7	6	3	13	142
+1 frameshifting	1	0	2	83	12	13	111
Total	114		98		41		253
Experimental data: 63			Predicted data: 190				

ornithine decarboxylase antizyme

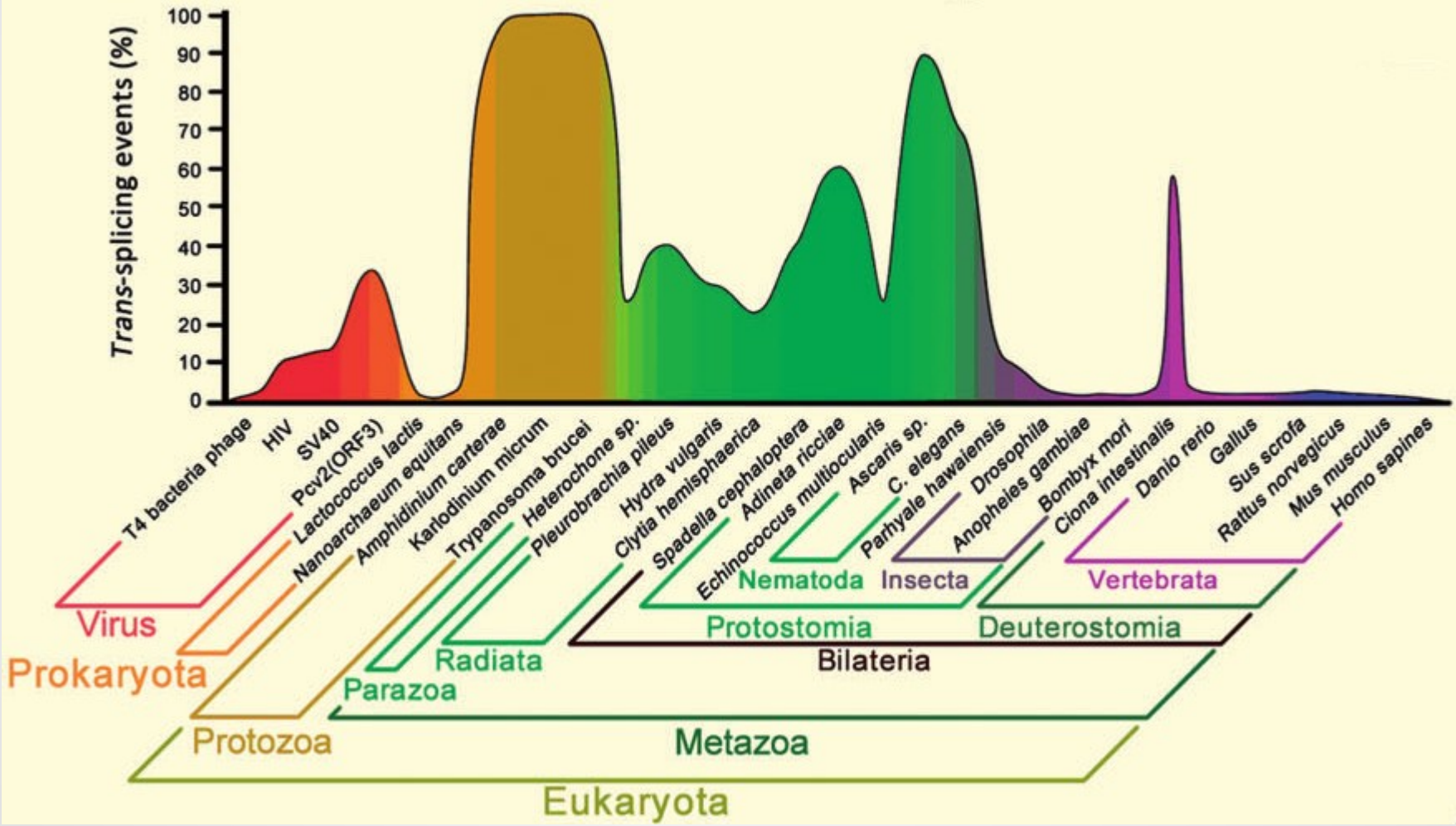
http://wilab.inha.ac.kr/fsdb/

Trans-splicing



Horiuchi and Aigaki, 2006

How common is trans-splicing?

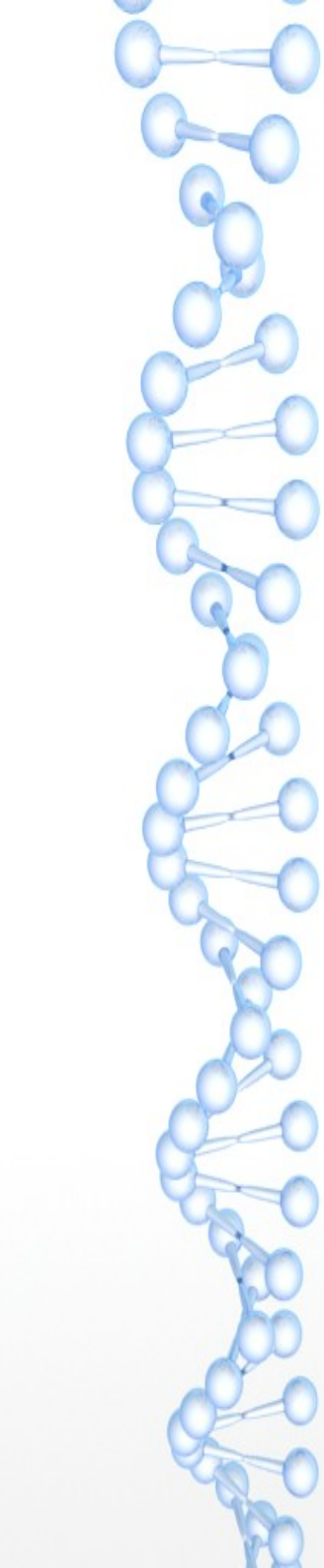


Lei et al, 2016

Transsplicing events (%) % of total gene number in species

Summary

Classical view		Modern view
Genetic code is		**Genetic code is**
1. Degenerated (redundant)		1. Degenerated (redundant)
2. Triplet		2.Triplet
3. Continuous	trans-splicing →	3. (Quasi)continuous
4. Unambiguous	Sec, Pyr →	4. Quasiunambiguous
5. Non-overlapping	ribosomal frameshift →	5. Quasinon-overlapping
6. Unidirectional	ribosomal frameshift →	6. Quasiunidirectional
7. Universal	difference between taxa →	7. Quasiuniversal



II. Gene and genetic code

Genetic code
Evolution

Are features of genetic code random?

1. There are > 10^{84} variant of triplet genetic code encoding 20 aminoacids and stop codons
2. Standard genetic code is resistant to errors but there many more resistant variants

3. Codons with U in second position encoded hydrophob aminoacids

4. There is a negative correlation between molecular weight of aminoacid and number of corresponding codons

5. There is a positive correlaion between number of codons encoding particular aminoacid and frequency of it in proteins

Theories describing origin and evolution of genetic code

0. Theory of frozen random origin
- Code originated once randomly and is frozen

1. Stereochemical theory
- Structure of genetic code is dependent on affinity between aminoacids and corresponding codons or anticodons

2. Adaptive theory
- Structure of genetic code is result of natural selection, which minimized deleterious effects of point mutations and transcriptional errors on tertiary structure and function of proteins

3. Coevolutionary theory
- Structure of genetic code is result of co-evolution with aminoacid biosynthesis pathways

Stereochemical theory

Major postulate

-Structure of genetic code is dependent on affinity between aminoacids and corresponding codons or anticodon**s**

- amino acids predominantly bind to short RNAs enriched with the corresponding triplets
- only a small fraction of possible random codes show a better correlation with these data. For codons, the correlation with the standard genetic code is stronger than for 90.3% of random codes, for anticodons than for 99.8%

- affinity of amino acids to their corresponding codons and anticodons, although statistically significant, is rather weak (compared to the others).
- for different amino acids, there is affinity either to codons or anticodons, or to both of them.

Adaptive theory

Major postulate

Structure of genetic code is result of natural selection, which minimized deleterious effects of point mutations and transcriptional errors on tertiary structure and function of proteins

Minimization of point mutations effect	Minimization of transcription/translation errors
• degeneracy of genetic code	• degeneracy of genetic code • similarity of physical and chemical properties of aminoacids encoded by similar codons

Weak points

- Assessment of the physicochemical similarity of amino acid properties is problematic
- There are many more optimal variants of genetic code
- When modeling evolution by selection, the standard genetic code proves to be unstable

Co-evolutionary theory

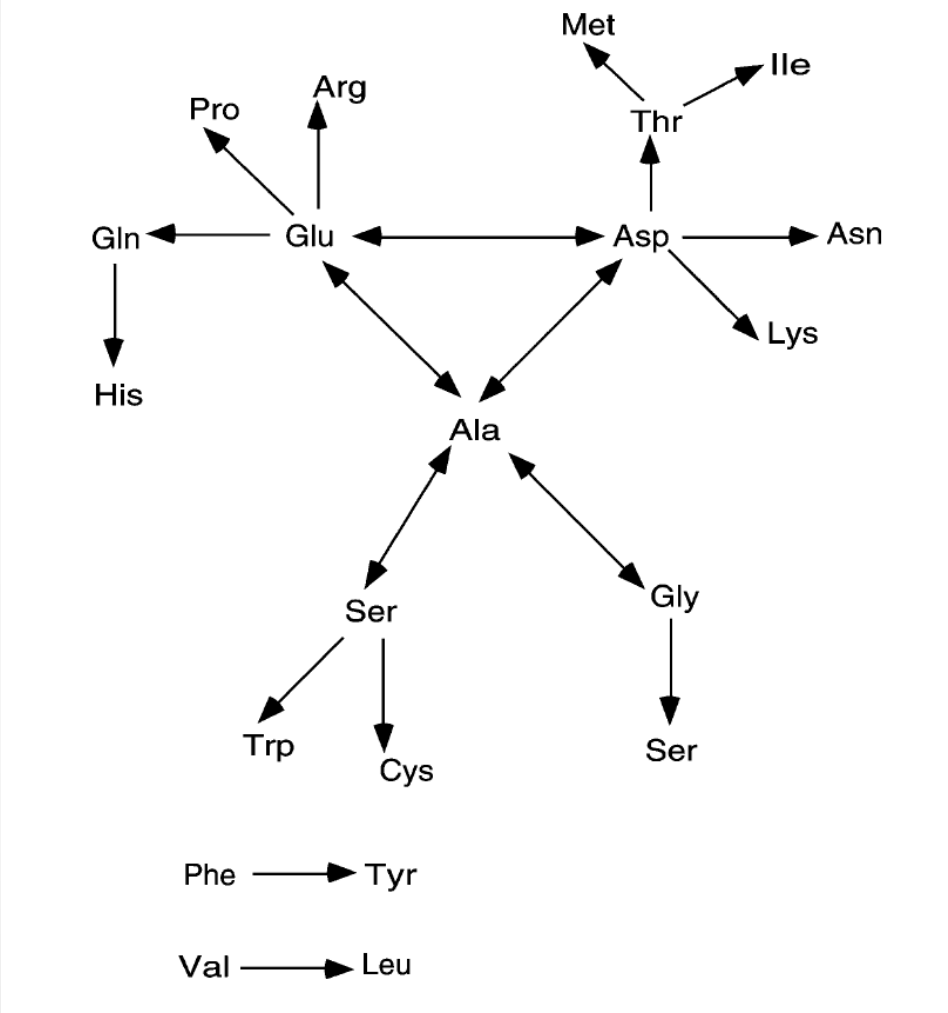
Major postulate

- *Prebiotic synthesis could not ensure the synthesis of all proteinogenic amino acids*
- *Biosynthetic pathways had to develop for some amino acids before they could be incorporated into the genetic code*
- *The evolution of the genetic code and amino acid biosynthesis pathways went in parallel*

- Amino acid biosynthesis pathways
- In some organisms, some amino acids are synthesized from their tRNA-bound precursors-the "fossil" pathways of biosynthesis
- Block structure of genetic code

- The evolutionary scenario is highly sensitive to the choice of amino acids presumably synthesized at the prebiotic stage
- When considering the NNY codons (tRNAs do not distinguish them) as one, the evolutionary scenario loses statistical support

Interconversion of aminoacids



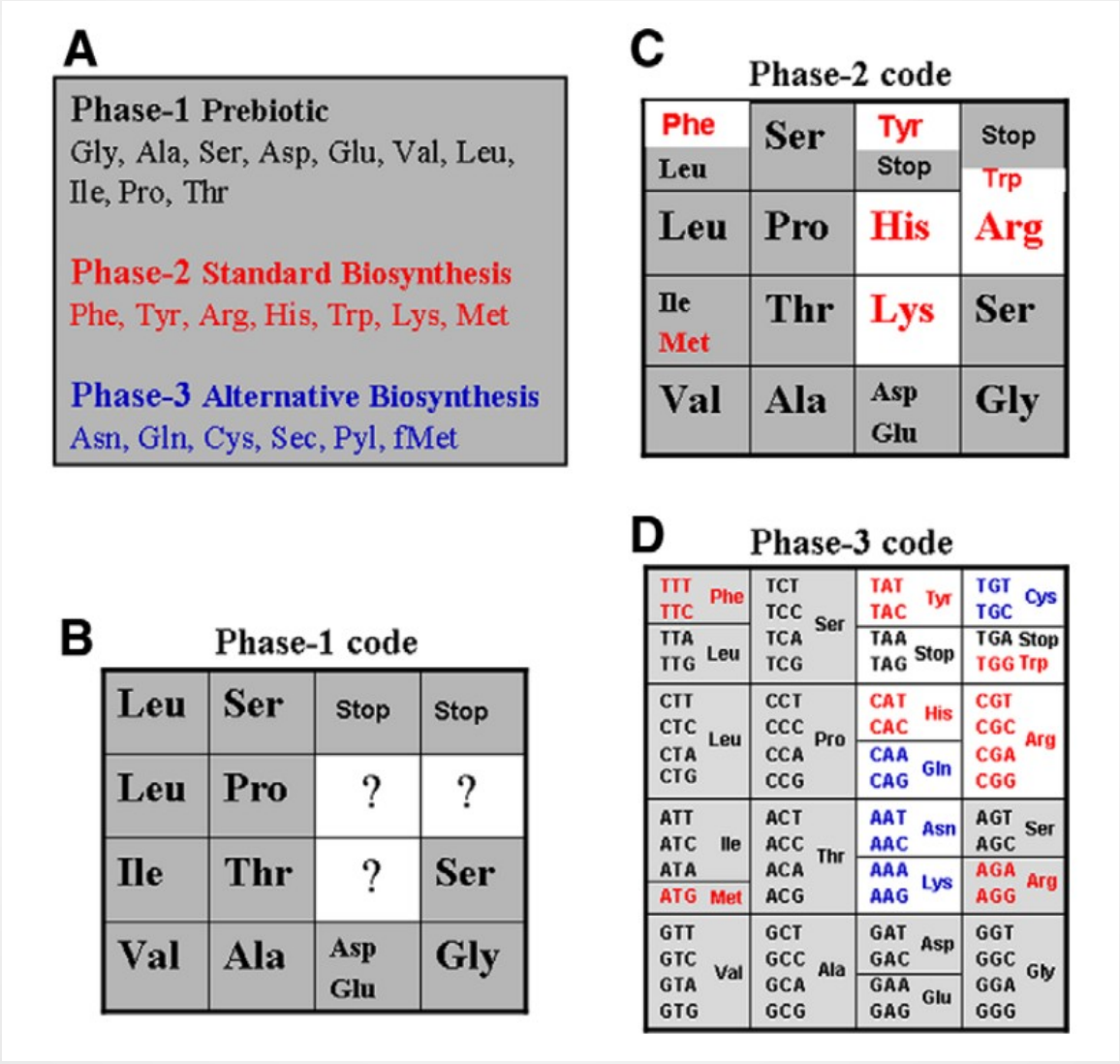
Giulio et al, 2004

Aminoacids synthesized in connection with tRNAs

Pathways	Phylogenetic distribution
$Glu\text{-}tRNA^{Gln} \rightarrow Gln\text{-}tRNA^{Gln}$	Bacteria and Archaea
$Asp\text{-}tRNA^{Asn} \rightarrow Asn\text{-}tRNA^{Asn}$	Bacteria (present in minority) and Archaea
$Ser\text{-}tRNA^{Sec} \rightarrow Sec\text{-}tRNA^{Sec}$	Bacteria, Archaea, and Eucarya
$Met\text{-}tRNA^{fMet} \rightarrow fMet\text{-}tRNA^{fMet}$	Bacteria, organelles
$Lys\text{-}tRNA^{Pyl} \rightarrow Pyl\text{-}tRNA^{Pyl}$	Some Archaea and Bacteria

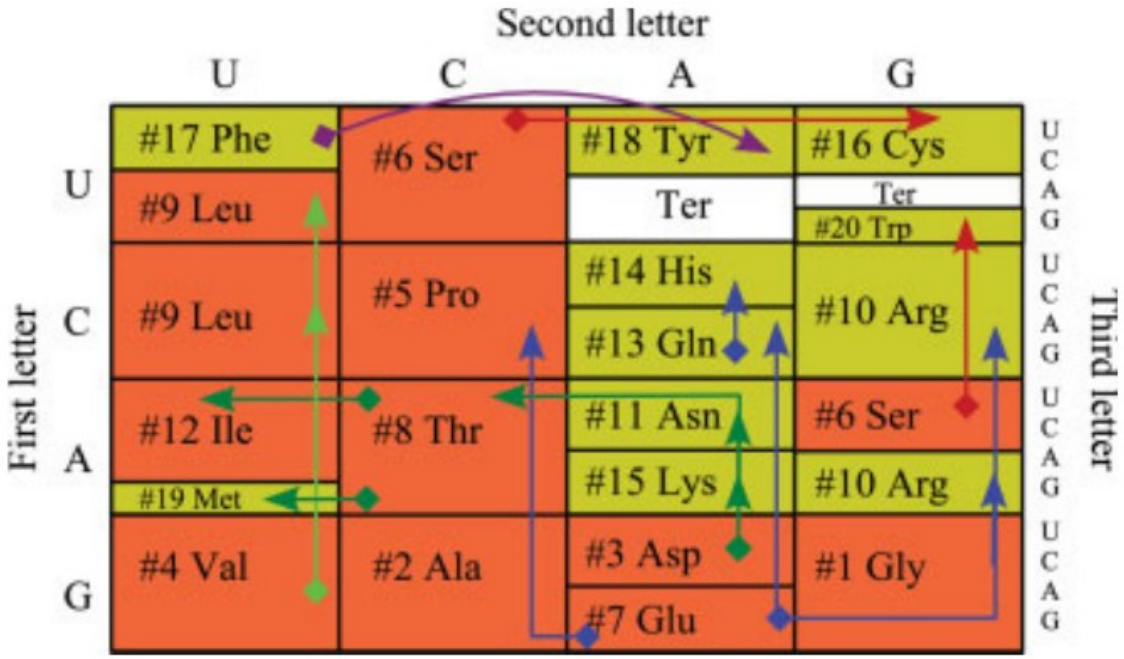
Giulio et al, 2004

Evolution of genetic code according to co-evolutionary theory

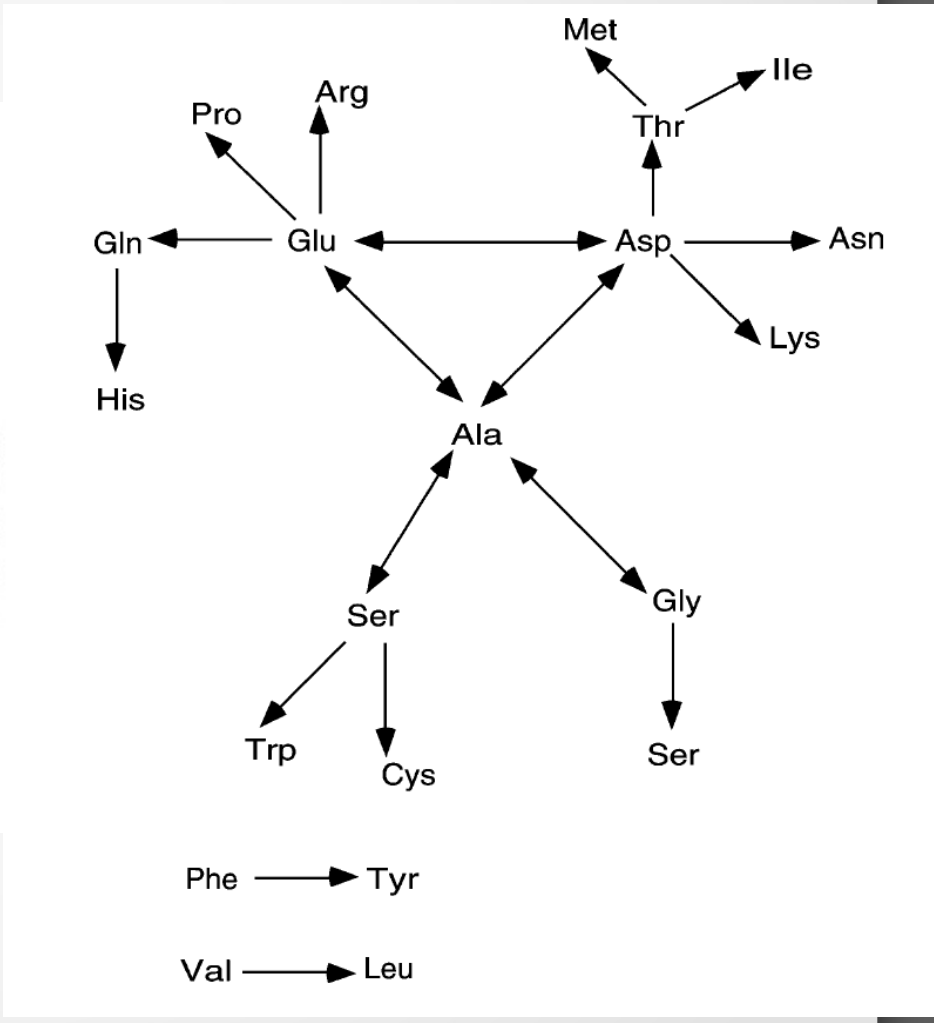


Moura et al, 2010

Evolution of genetic code according to co-evolutionary theory

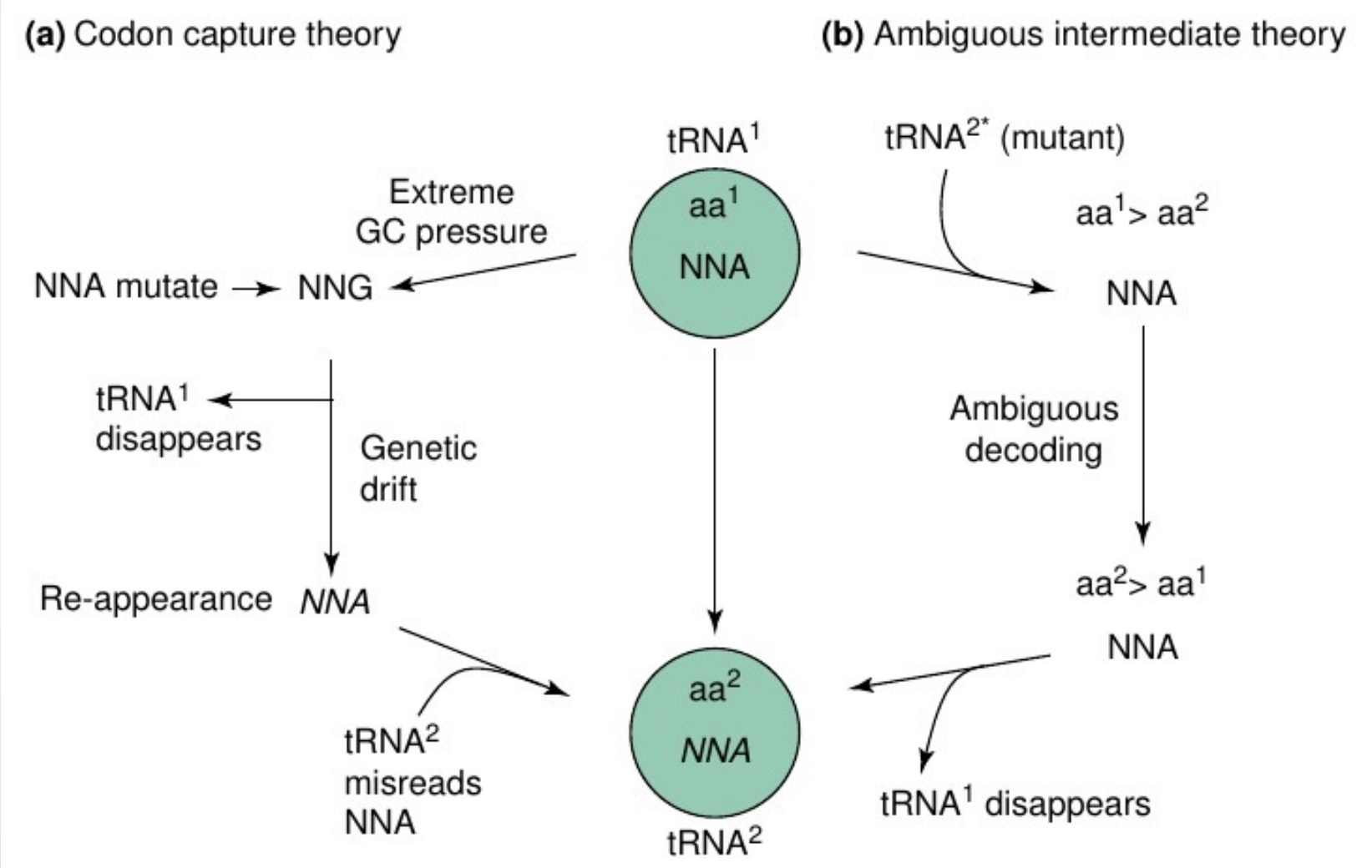


Moura et al, 2010



Moura et al, 2010

Possible mechanisms for modification of genetic code



Santos et al, 2004

Examples of deviations from the standard genetic code

Case	Codon	Standard AA	Case AA
Some *Candida* species (fungi)	CUG	Leu	Ser
Mitochondria (*Saccharomyces cerevisiae*)	CU(U, C, A, G)	Leu	Ser
Mitochondria of higher plants	CGG	Arg	Trp
Mitochondria (all species)	UGA	Stop	Trp
Prokaryota	GUG	Val	Start
Eukaria (seldom)	CUG	Leu	Start
Eukaria (seldom)	GUG	Val	Start
Prokaryota (seldom)	UUG	Leu	Start
Eukaria (seldom)	ACG	Thr	Start
Mammalian mitochondria	AGC, AGU	Ser	Stop

Summary

- Genetic code is not universal. It implies that it is not stable
- There are three compatible theories for origin and evolution of genetic code. Each has weak points.
- There are at least two hypotetical mechanisms explaining change of genetic code