Struttura e complessità di genomi procariotici ed eucariotici

Struttura e dimensioni di genomi cellulari

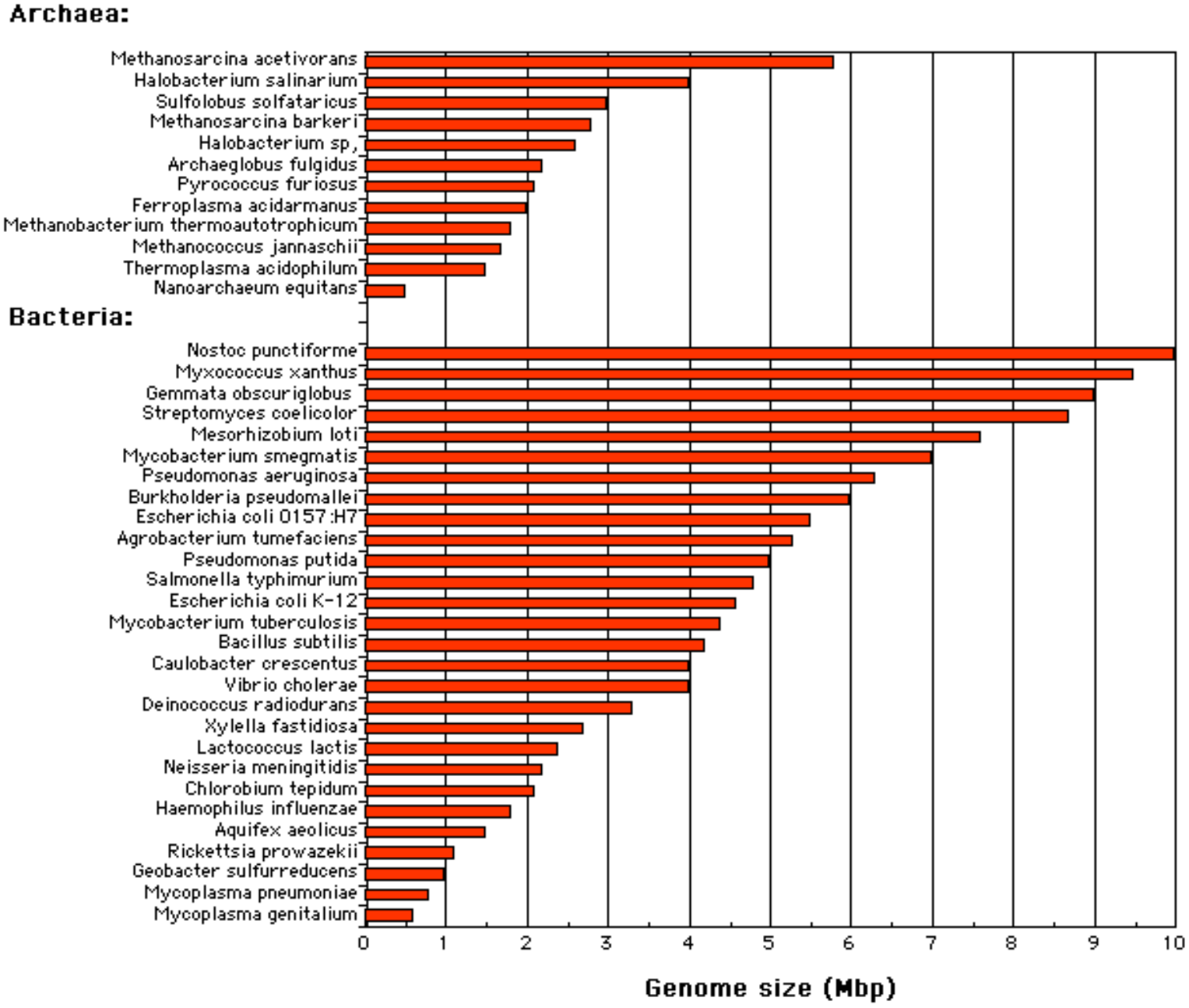
Procarioti (Bacteria e Archaea):

Tipicamente c'è **un solo cromosoma** nel citoplasma, contenente una molecola di **DNA circolare** lungo da **0,5 a 10 milioni** di paia di basi, circa. Spesso possono essere presenti anche uno o più **plasmidi**, piccoli DNA circolari accessori, costituiti da alcune migliaia di paia di basi. Sono note eccezioni: alcuni procarioti con due o più cromosomi circolari, o un cromosoma lineare.

Eucarioti (unicellulari, piante, funghi, animali):

Di regola c'è **più di un cromosoma nel nucleo**, mediamente poche decine, anche se sono noti casi di parecchie centinaia e, all'altro limite, di uno solo. Ciascun cromosoma contiene una molecola di DNA lineare. Nel nucleo può essere presente un solo corredo di cromosomi diversi (la cellula si dice **aploide**), due serie (**diploide**), e sono i casi più frequenti, o più (tri-, tetra-, esa-, …poli-ploide). Le **dimensioni di un corredo aploide** variano da **una decina di milioni** di paia di basi circa (ad esempio i lieviti o qualche alga unicellulare) a un **centinaio di miliardi** (alcune felci, i pesci polmonati e gli anfibi urodeli) o anche più in alcune amebe.
Al di fuori del nucleo è generalmente presente un certo numero di **mitocondri**, dotati di un proprio minigenoma, costituito da una molecola di **DNA circolare** di circa **15.000 – 100.000** paia di basi.
I vegetali contengono anche un certo numero di **cloroplasti**, pure dotati di un proprio minigenoma, costituito da una molecola di **DNA circolare** di circa **100.000** paia di basi.

Al di là di questo schema base si riscontrano **casi particolari.** Alcuni unicellulari sono privi di mitocondri (es.: *Giardia lamblia*), alcune alghe unicellulari (es. *Guillardia theta*) presentano due nuclei distinti ciascuno con più cromosomi (il minore è detto **nucleomorfo**). I Ciliati (es. *Tetrahymena*), accanto al nucleo normale, detto germinale o **micronucleo**, sviluppano secondariamente un ulteriore nucleo più grande, detto somatico o **macronucleo**, contenente molte migliaia di minicromosomi a DNA lineare, copie numerosissime di segmenti del DNA dei cromosomi del micronucleo. Le cellule salivari di *Drosophila* replicano il proprio DNA a cascata una decina di volte senza disgiungerlo dando origine a **cromosomi giganti** di oltre 1000 copie, detti **politenici**. Il batterio gigante *Epulopiscium* (batterio gigante simbionte del pesce chirurgo) contiene un numero molto elevato di copie del proprio genoma (poliploide).


Archaea:
Methanosarcina acetivorans
Halobacterium salinarium
Sulfolobus solfataricus
Methanosarcina barkeri
Halobacterium sp,
Archaeglobus fulgidus
Pyrococcus furiosus
Ferroplasma acidarmanus
Methanobacterium thermoautotrophicum
Methanococcus jannaschii
Thermoplasma acidophilum
Nanoarchaeum equitans
Bacteria:
Nostoc punctiforme
Myxococcus xanthus
Gemmata obscuriglobus
Streptomyces coelicolor
Mesorhizobium loti
Mycobacterium smegmatis
Pseudomonas aeruginosa
Burkholderia pseudomallei
Escherichia coli 0157:H7
Agrobacterium tumefaciens
Pseudomonas putida
Salmonella typhimurium
Escherichia coli K-12
Mycobacterium tuberculosis
Bacillus subtilis
Caulobacter crescentus
Vibrio cholerae
Deinococcus radiodurans
Xylella fastidiosa
Lactococcus lactis
Neisseria meningitidis
Chlorobium tepidum
Haemophilus influenzae
Aquifex aeolicus
Rickettsia prowazekii
Geobacter sulfurreducens
Mycoplasma pneumoniae
Mycoplasma genitalium
0
1
2
3
4
5
6
7
8
9
10
Genome size (Mbp)

Quanto DNA c'è in una cellula di *Escherichia coli*?

Massa molecolare delle basi:

-Adenina	**135 Da**
-Citosina	**111 Da**
-Guanina	**151 Da**
-Timina	**126 Da**

Massa molecolare media delle basi	**131 Da**
Massa molecolare media di un deossinucleoside	**247 Da**
Massa molecolare media di un deossinucleotide libero	**327 Da**
Massa molecolare media di un deossinucleotide in catena	**309 Da**
Massa molecolare media di un paio di deossinucleotidi in catena	**618 Da**

Il cromosoma di *E. coli* contiene circa 4.7×10^6 paia di basi, quindi ha una massa di circa 2.9×10^9 Da, corrispondenti a 4.8×10^{-15} g
(1Da=1.66×10^{-24} g)

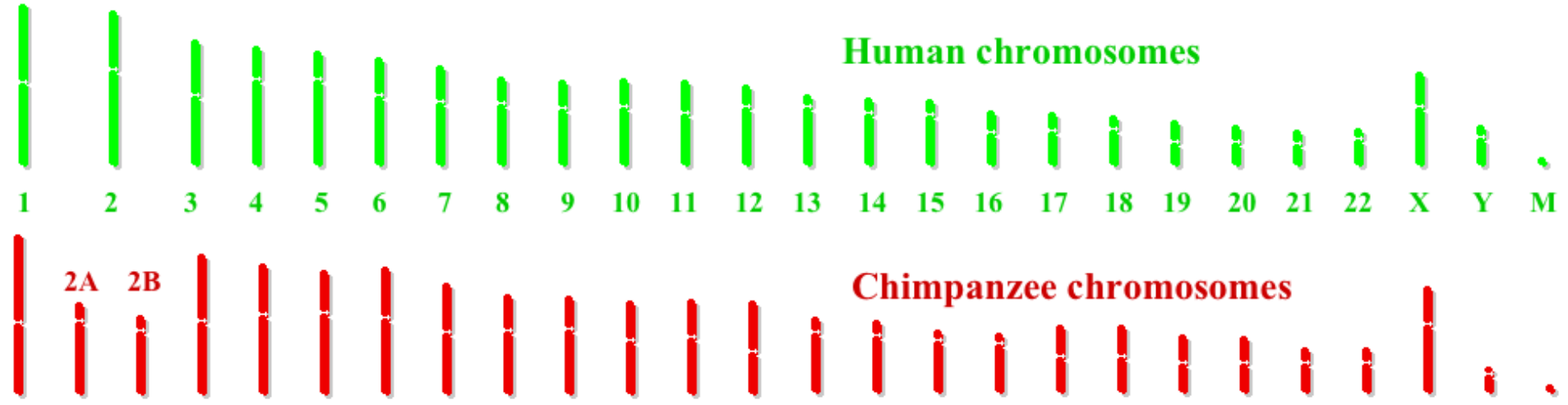
e in una cellula umana?

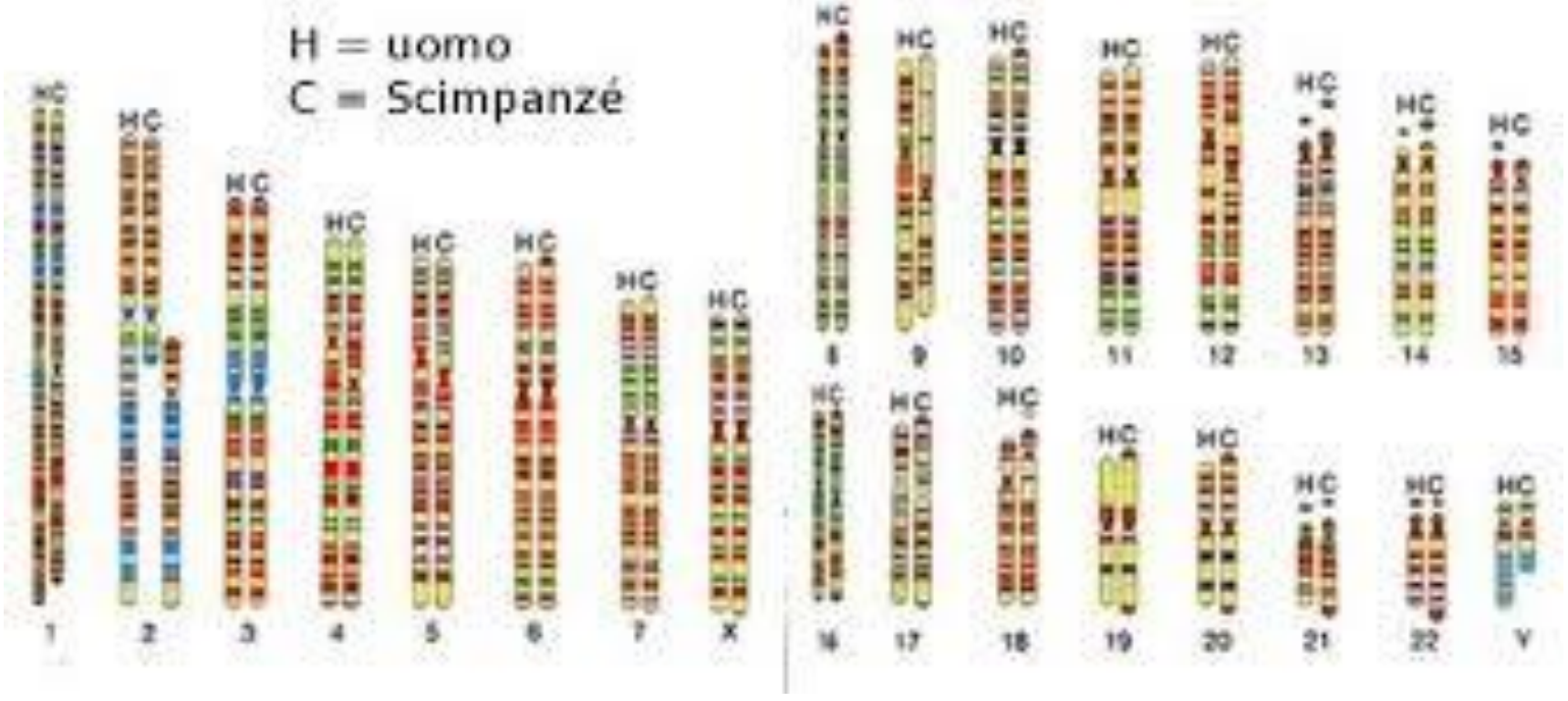
Un corredo aploide contiene circa 3.3×10^9 paia di basi, una cellula diploide il doppio, quindi una massa di circa 2.1×10^{12} Da, corrispondenti a 3.5×10^{-12} g, a cui andrebbe aggiunta una piccola quantità di DNA mitocondriale.
Essendo il numero stimato di cellule in un adulto dell'ordine di 10^{14}, per la gran parte diploidi, ne consegue un contenuto complessivo di DNA dell'ordine di qualche ettogrammo .

I 3 paradossi del genoma

Tra uomo e scimpanzè

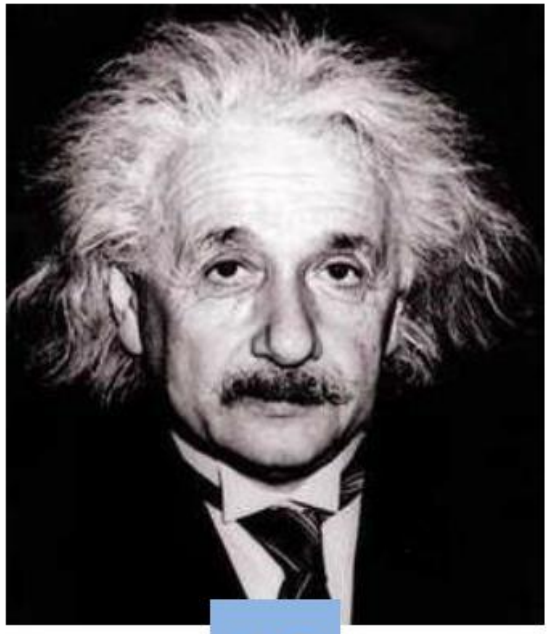
Differenza del 1-2% nel DNA


Human chromosomes
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y M
2A 2B
Chimpanzee chromosomes


H = uomo
C = Scimpanzé
1 2 3 4 5 6 7 X
8 9 10 11 12 13 14 15
16 17 18 19 20 21 22 Y

Paradosso del valore **K**: la complessità non correla con il **numero di cromosomi**.

Homo sapiens



46

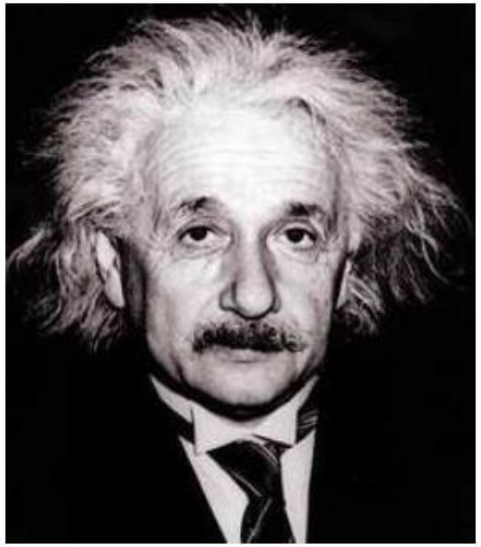
Lysandra atlantica

250

Ophioglossum reticulatum

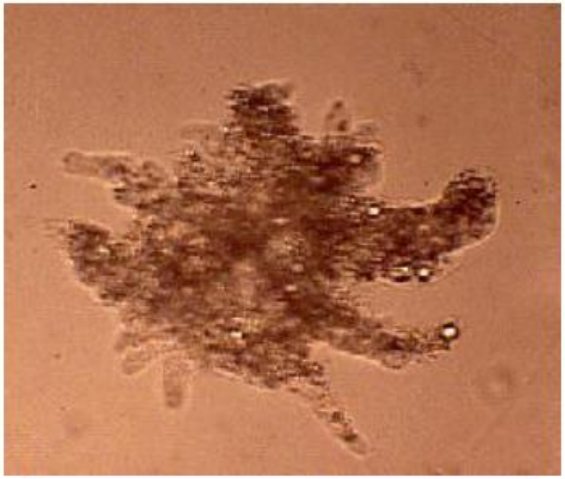
~1260

Paradosso del valore **C**: la complessità non correla con la **grandezza del genoma**.



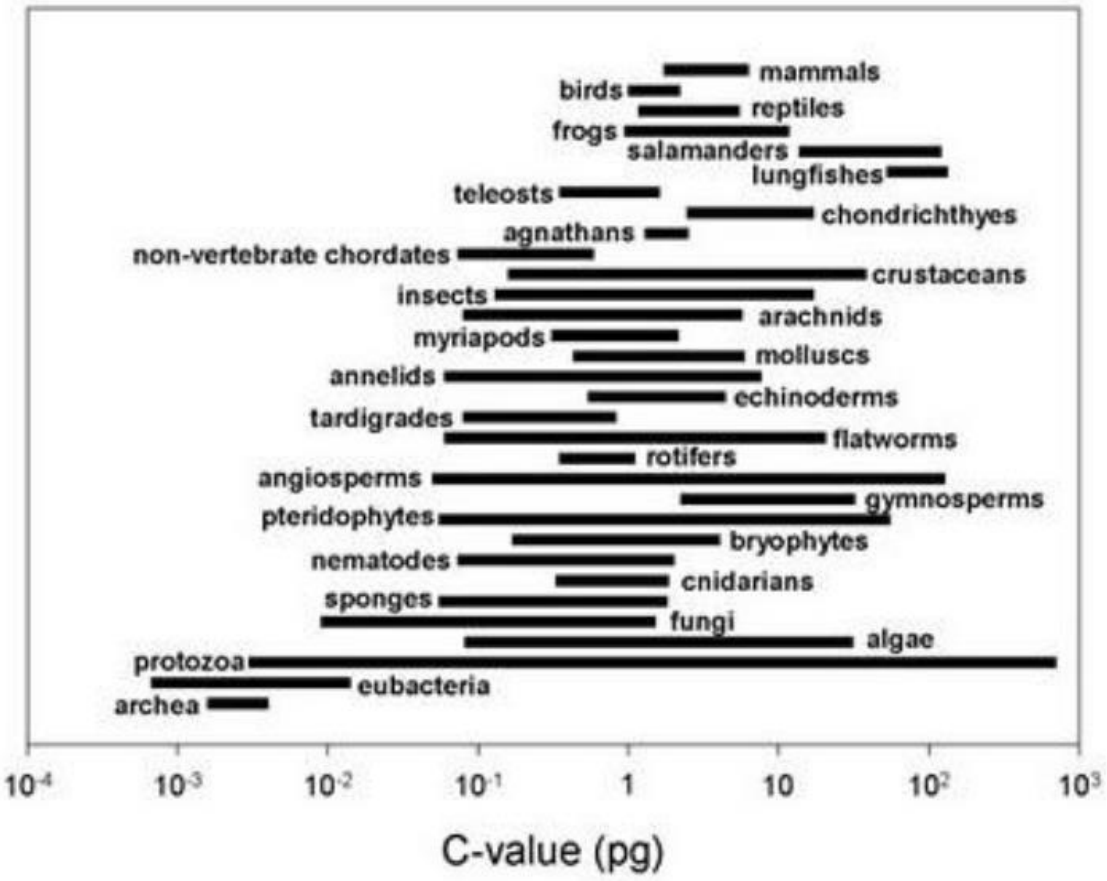
3.4×10^9 bp
Homo sapiens

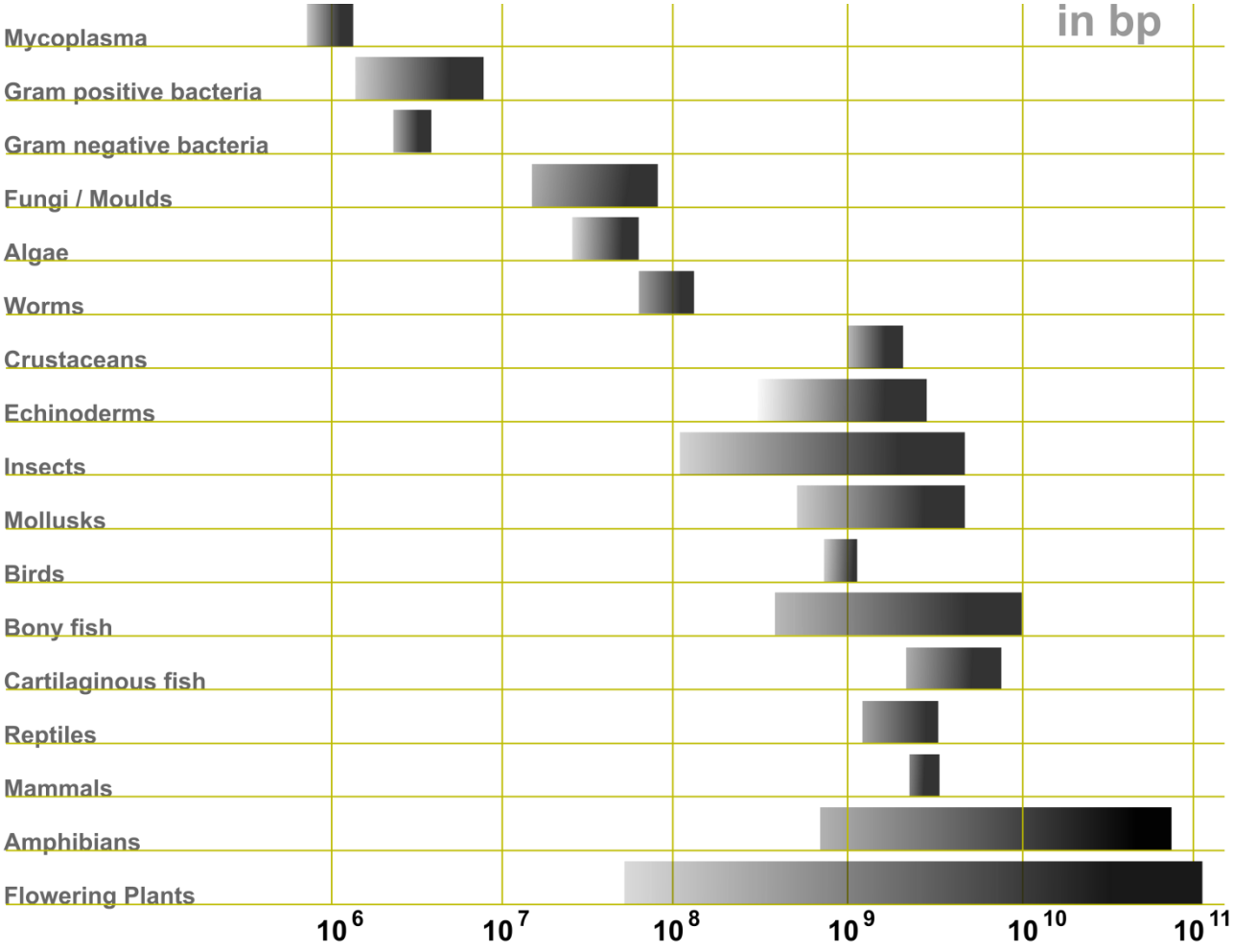
1.5×10^{10} bp
Allium cepa



6.8×10^{11} bp
Amoeba dubia

Il valore C (quantità di DNA contenuto nel nucleo di una cellula aploide di un organismo)




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

Il paradosso C dei genomi eucariotici

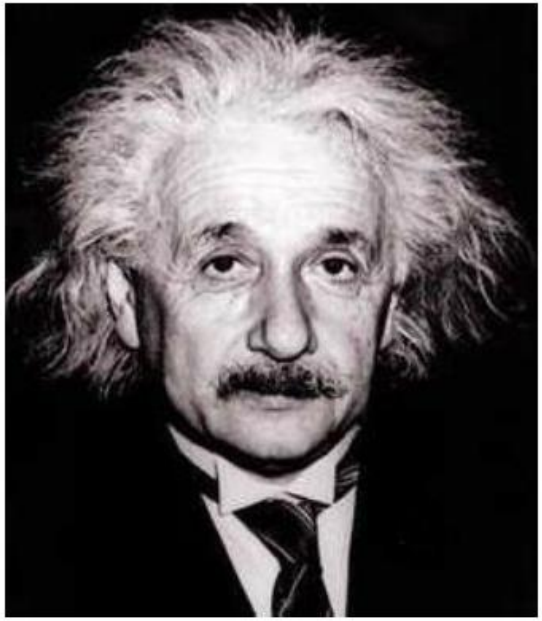
Nei **procarioti** numero di geni e dimensioni genomiche delle varie specie sono approssimativamente proporzionali, in ragione di circa **1000-1200 pb/gene**.
Appare ragionevole.

Le dimensioni genomiche (corredo aploide) nelle **specie eucariotiche** per contro **variano enormemente** (da circa **10^7** a più di **10^{11}**) **senza alcuna relazione con la complessità dell'organismo**. Ad esempio, negli unicellulari alcune amebe hanno le massime dimensioni genomiche mentre alcune alghe unicellulari e alcuni lieviti le minime. Nelle piante superiori si va da circa 10^8 a circa 10^{11}, e lo stesso succede con gli animali.
Questa osservazione è stata chiamata il **Paradosso C** (Complessità).

Il paradosso (in parte) rientra quando si scopre che la variazione nel **numero di geni** (che nella casistica attuale, non vasta ma certamente significativa, vanno da circa **5000** a circa **30000**) risulta molto più contenuta e approssimativamente connessa alla complessità dell'organismo.
Il grosso della differenza nelle dimensioni genomiche è dovuto alla **enorme variazione** nella quantità di **DNA non codificante** per proteine e in buona parte costituito da **sequenze ripetitive**. La ragione di tale variabilità non è ancora chiara.

Paradosso del valore **N**:

Il **numero di geni** e la complessità degli organismi non sono correlati



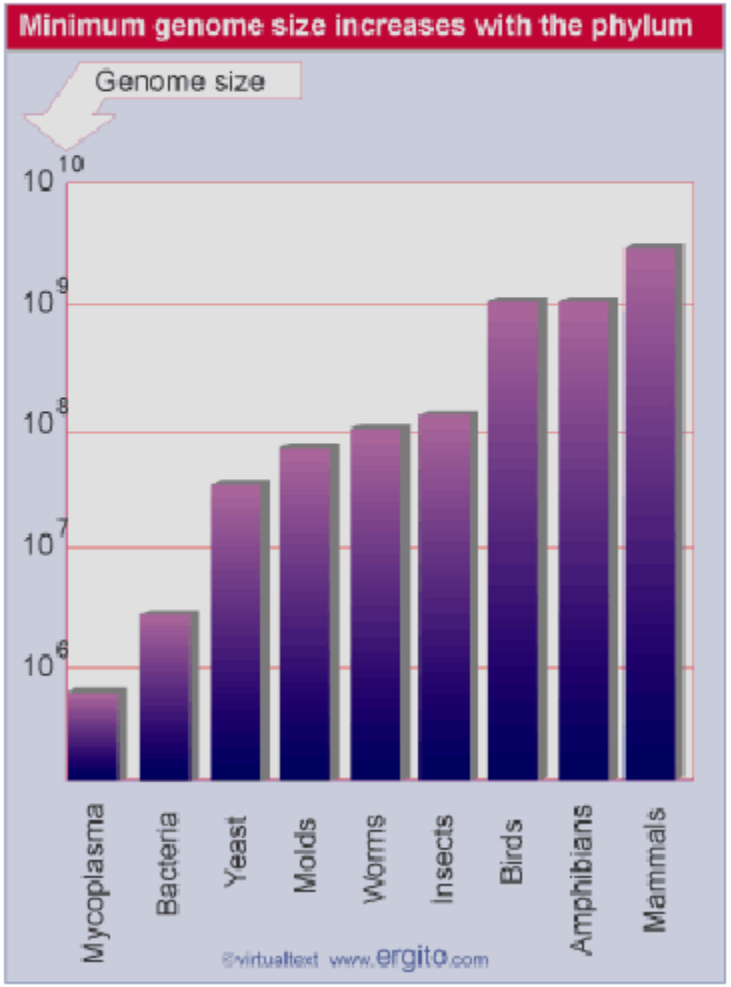
~21000 geni

~25000 geni

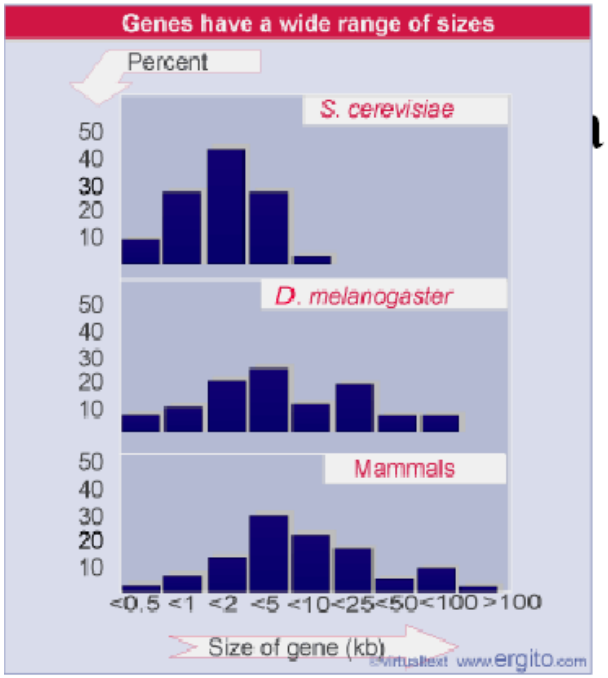
~60000 geni

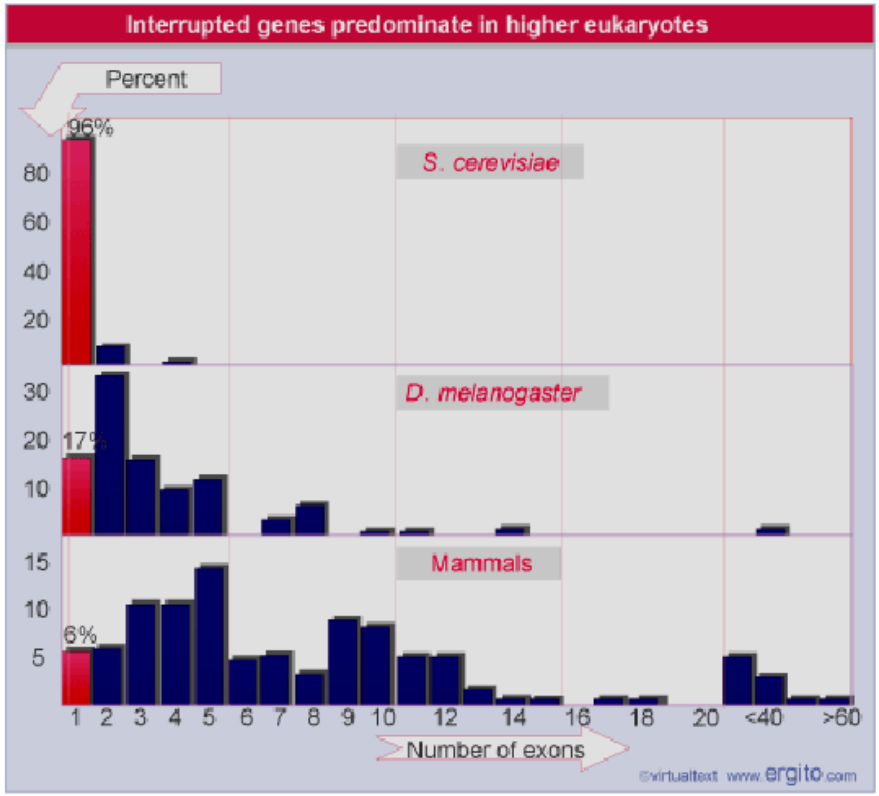
Quantità minima di DNA

- Il grafico indica che un aumento della grandezza del genoma è necessaria per la complessità dei procarioti ed eucarioti inferiori



La dimensione dei geni varia grandemente





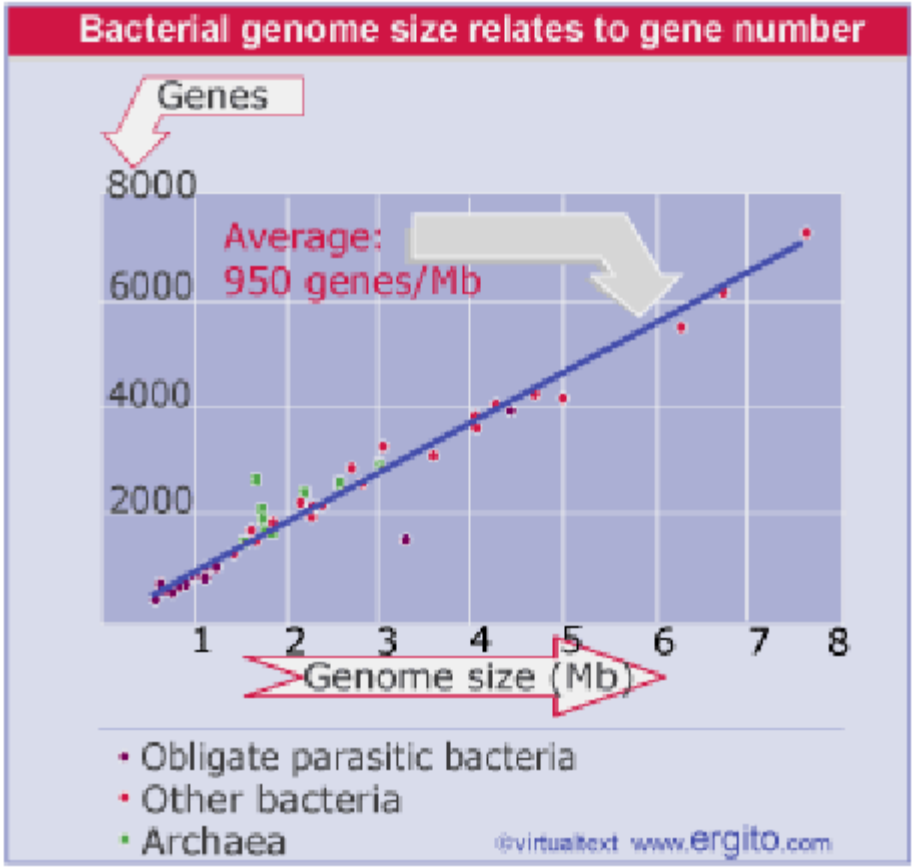
Dimensioni genomiche di alcuni organismi rappresentativi:

	N° cromosomi (corredo aploide)	M bp tot	N° geni approx	k bp gene
PROCARIOTI				
Nanoarchaeum equitans	**1**	**0.5**	**420**	**1,2**
Mycoplasma genitalium	**1**	**0.6**	**480**	**1,2**
Haemophilus influenzae	**1**	**1.8**	**1750**	**1,0**
Escherichia coli	**1**	**4.7**	**4300**	**1,1**
Bradyrhizobium japonicum	**1**	**9.1**	**8300**	**1,1**
EUCARIOTI				
Protisti				
Plasmodium falciparum	**14**	**23**	**5300**	**4,3**
Funghi				
Saccharomyces cerevisiae	**16**	**12**	**6000**	**2,0**
Aspergillus nidulans	**8**	**31**	**..**	
Piante				
Ophioglossum petiolatum	**510**	**160000**	**..**	
Arabidopsis thaliana	**5**	**120**	**25000 ca.**	**5**
Oryza sativa	**12**	**450**	**30000 ca.**	**15**
Triticum aestivum	**7/21**	**17000**	**..**	
Animali				
Caenorhabditis elegans	**6**	**100**	**19000**	**5**
Drosophila melanogaster	**4**	**180**	**13500**	**13**
Takifugu rubripes	**22**	**380**	**25000**	**13**
Danio rerio	**25**	**1900**	**25000**	**60**
Protopterus aethiopicus	**19**	**140000**	**..**	
Gallus gallus	**39**	**1200**	**25000**	**50**
Mus musculus	**20**	**3450**	**25000**	**140**
Homo sapiens	**23**	**3200**	**25000***	**130**

***** Secondo una stima di Craig Venter (nel 2007) i geni sarebbero 23.224, mentre secondo Jim Kent (2007) sarebbero 20.433 codificanti e 5.871 non codificanti.

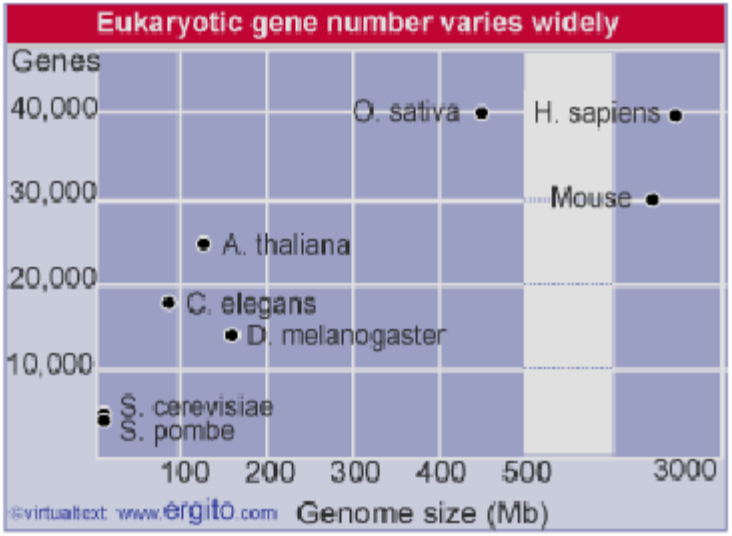
Geni batterici

- Nei batteri quasi tutto il DNA codifica proteine o RNA
- Il genoma varia entro un ordine di grandezza ed è proporzionale al numero di geni



Geni negli eucarioti

- Il genoma degli **eucarioti unicellulari** è simile a quello dei **procarioti**
- Gli eucarioti superiori hanno più geni, ma il loro numero non correla con la dimensione del genoma



Tre classi di geni

- I geni non sono sequenze casuali ma hanno caratteristiche ben precise.
- Buona parte dell'informazione contenuta in un gene viene "copiata" in una molecola di RNA; il resto del gene è coinvolto comunque nel processo di "copia" (trascrizione).
- Alcuni tipi di RNA vengono utilizzati per la sintesi delle proteine, altri svolgono svariati tipi di funzioni.
- Esistono tre classi di geni, che differiscono in base al tipo di RNA che viene prodotto con la loro espressione:
  - Geni della I classe
    - RNA ribosomiale (rRNA)
  - Geni della II classe
    - RNA messaggero (mRNA)
    - Piccoli RNA nucleari (snRNA)
    - Micro RNA (miRNA)
  - Geni della III classe
    - RNA transfer (tRNA)
    - Piccoli RNA nucleolari (snoRNA)
    - Piccoli RNA citoplasmatici (scRNA)
    - Micro RNA (miRNA)

Gli Pseudogeni

- Gli pseudogeni sono copie non funzionali di geni.
- Sono una sorta di relitti evolutivi.
- Gli pseudogeni convenzionali sono geni inattivati in seguito ad una o più mutazioni nella loro sequenza nucleotidica.
- Una volta che uno pseudogene è diventato completamente non funzionale si degraderà per accumulazione di ulteriori mutazioni e potrebbe addirittura non essere più riconosciuto come relitto genico.

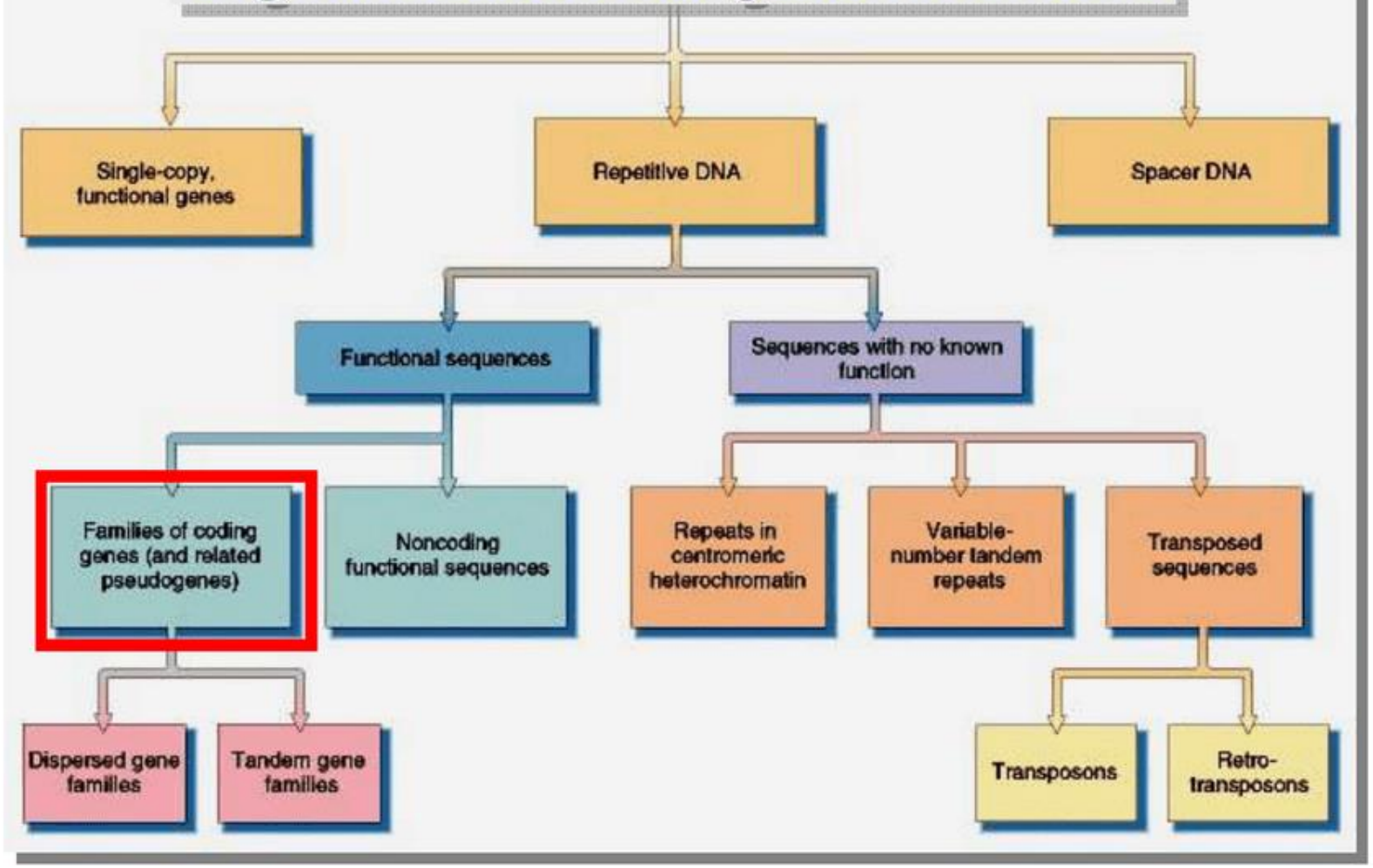
Pseudogeni

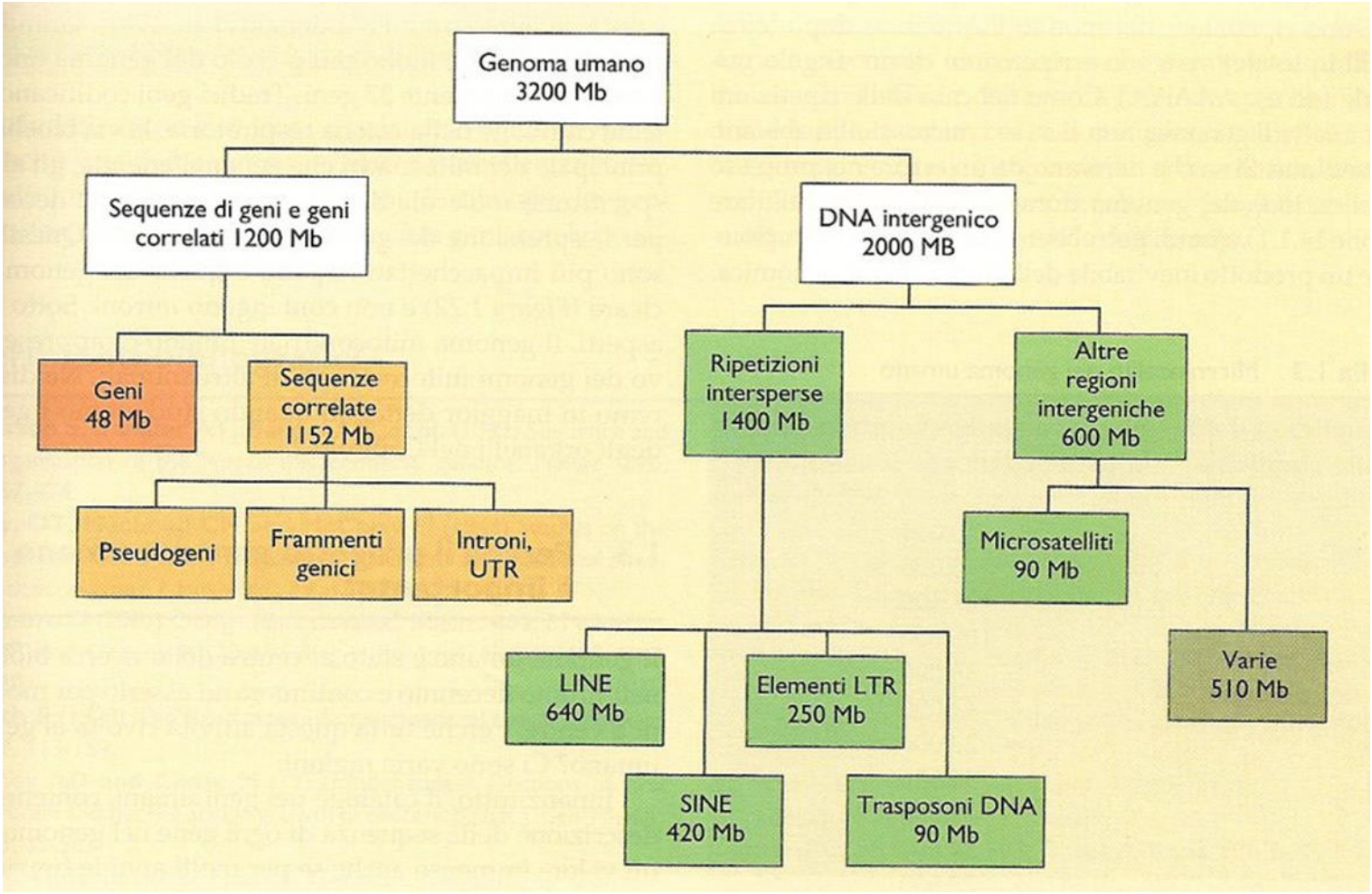
- Sono copie difettose di geni funzionali, prima ritenute un residuo evolutivo
- Contribuiscono alla formazione di qualche RNA attivo
- Si è notato che la loro compromissione determina la morte dell'organismo soggetto

Ripetizioni disperse e microsatelliti

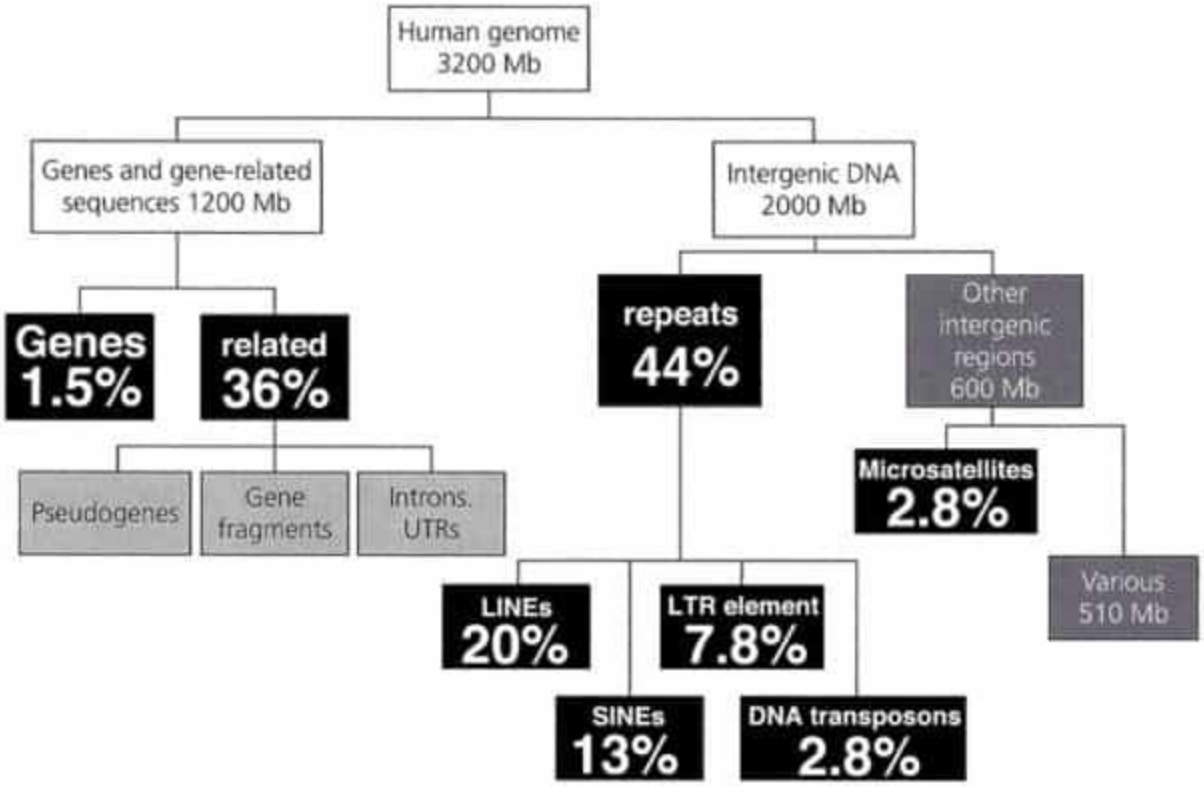
- La grande maggioranza del DNA intergenico è rappresentata da sequenze ripetute di vario tipo.
- Il DNA ripetitivo può essere diviso in due categorie:
  - Ripetizioni intersperse
  - DNA ripetuto in tandem

Organizzazione del genoma umano



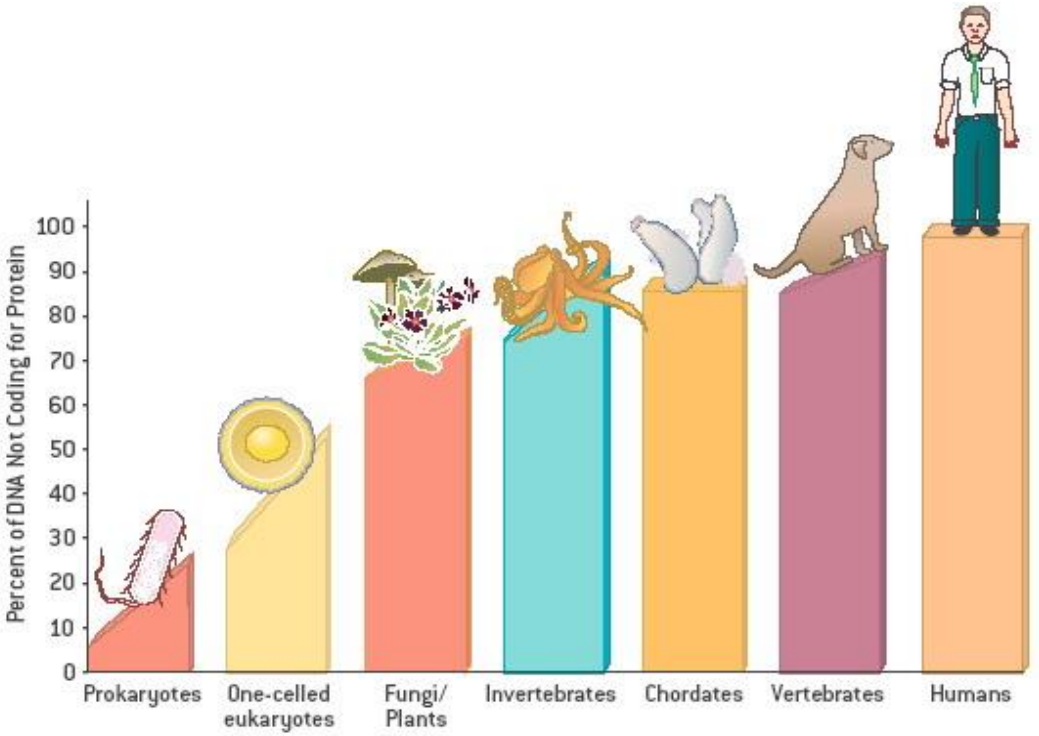

Genoma umano
3200 Mb
Sequenze di geni e geni correlati 1200 Mb
DNA intergenico
2000 MB
Geni
48 Mb
Sequenze correlate
1152 Mb
Ripetizioni intersperse
1400 Mb
Altre regioni intergeniche
600 Mb
Pseudogeni
Frammenti genici
Introni, UTR
Microsatelliti
90 Mb
LINE
640 Mb
Elementi LTR
250 Mb
Varie
510 Mb
SINE
420 Mb
Trasposoni DNA
90 Mb

Overview of human genome



Problema della complessità

Rapporto tra DNA
non codificante(ncDNA) e
DNA del genoma totale(tgDNA)
in alcuni genomi sequenziati.

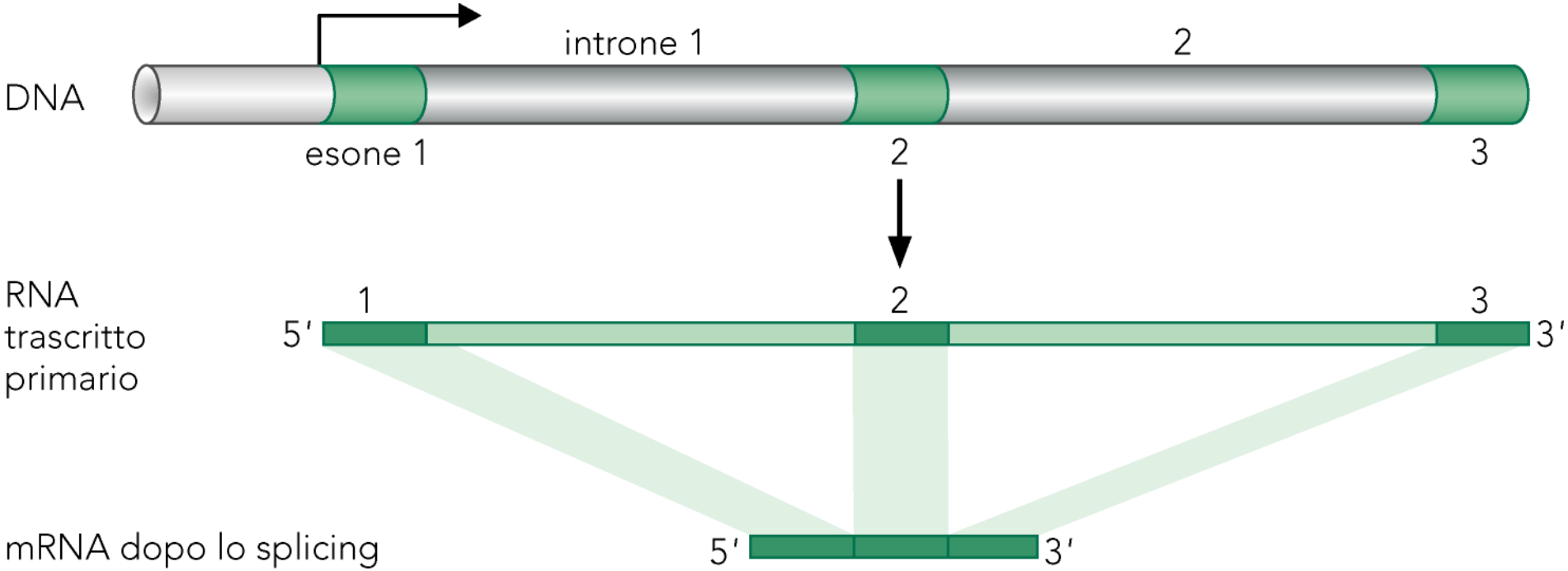


NONPROTEIN-CODING SEQUENCES make up only a small fraction of the DNA of prokaryotes. Among eukaryotes, as their complexity increases, generally so, too, does the proportion of their DNA that does not code for protein. The noncoding sequences have been considered junk, but perhaps it actually helps to explain organisms' complexity.

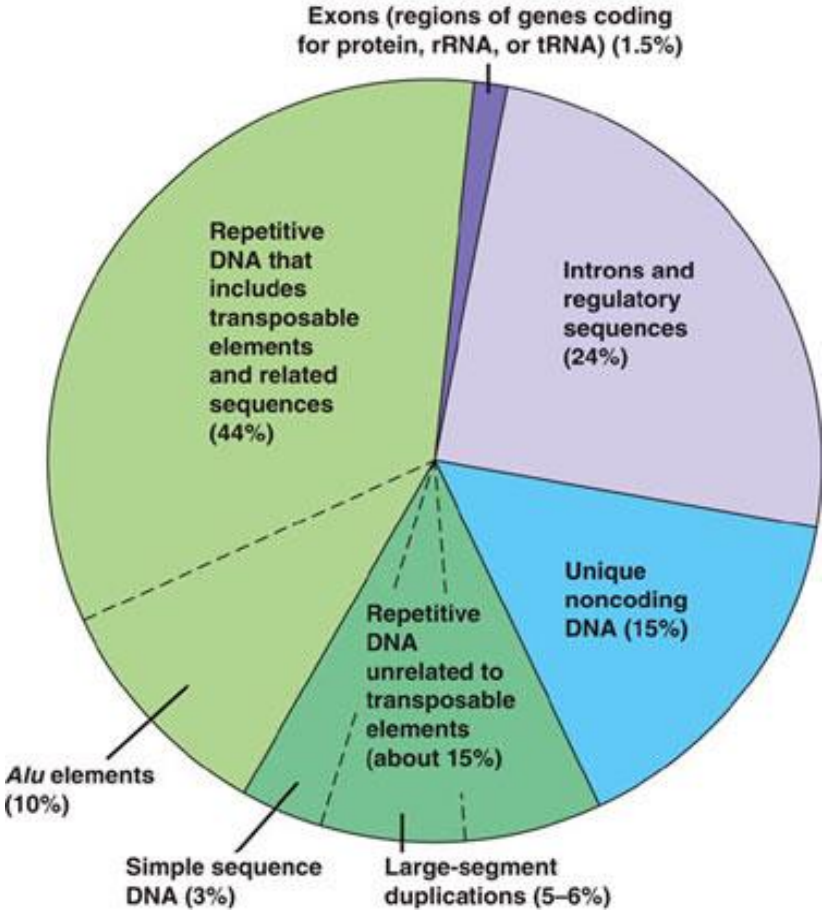
La complessità degli organismi è correlata alle dimensioni del genoma non codificante

Struttura dei Geni eucariotici codificanti

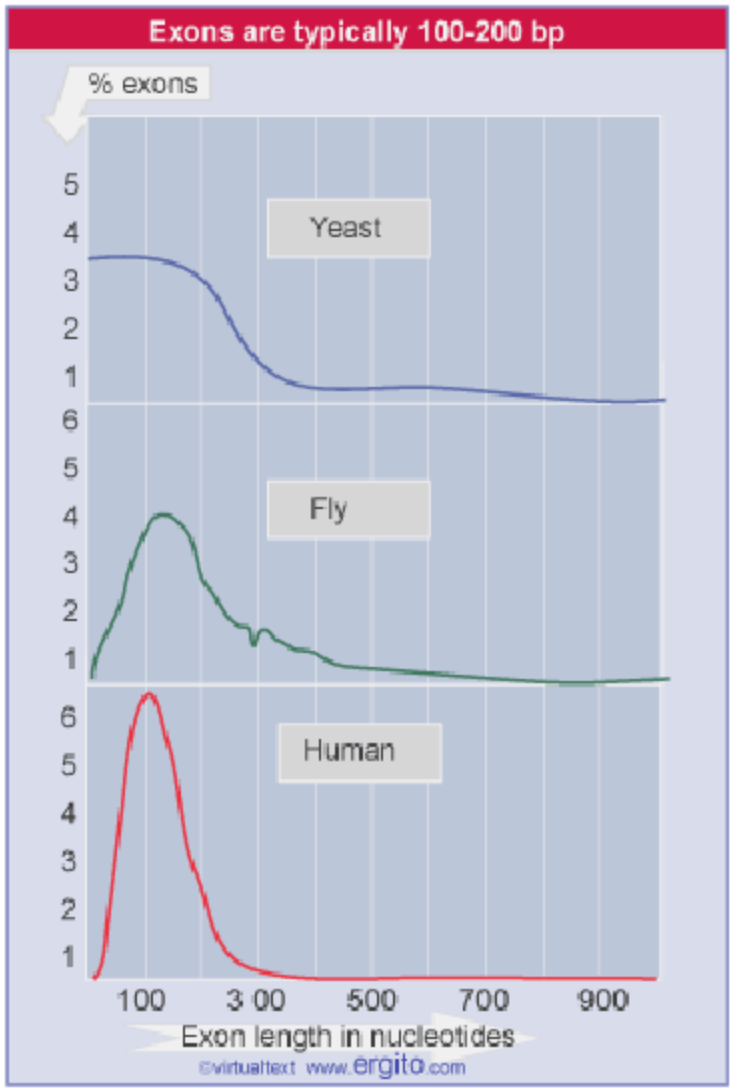
- I geni codificanti sono quelli che vengono trascritti in mRNA.
- Contengono una parte realmente codificante, che specifica la sequenza degli aminoacidi che costituiranno la proteina, ed una parte non codificante.
- A monte della sequenza che verrà trascritta in mRNA vi sono le sequenze regolatrici.
- La sequenza trascritta è costituita da due tipi di elementi, detti esoni ed introni.
- Solo gli esoni contengono informazioni per la sintesi della proteina.

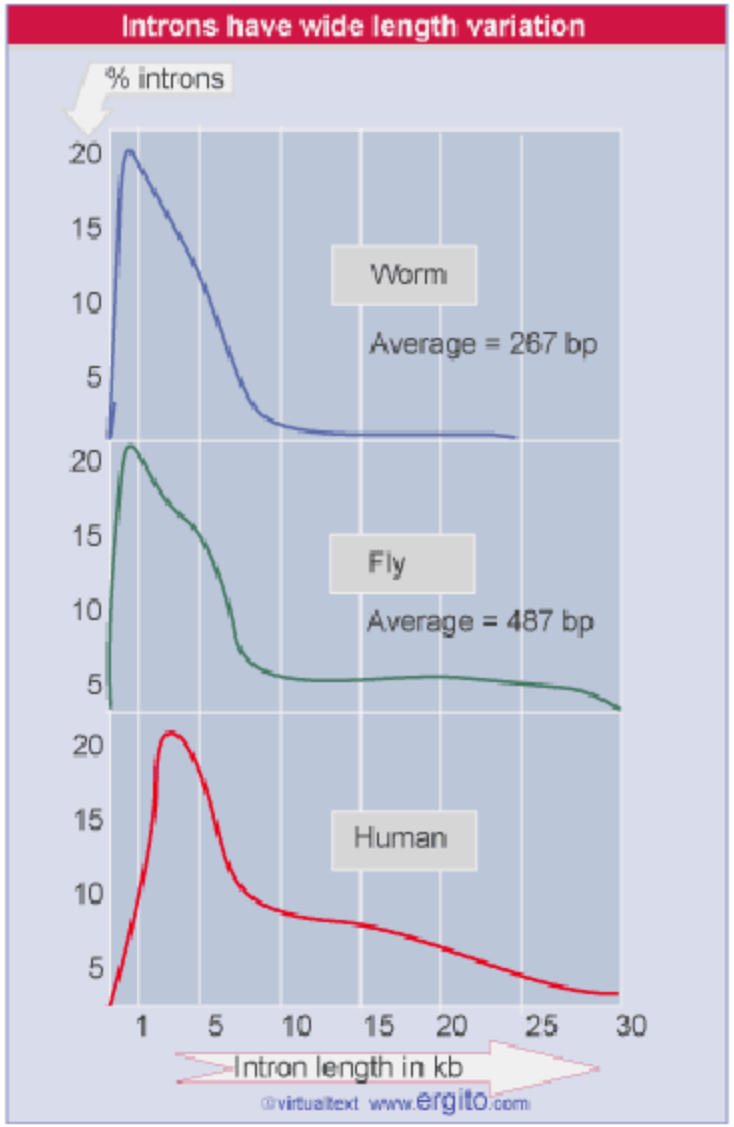


DNA codificante e non codificante nel genoma umano



Dimensione esoni e introni

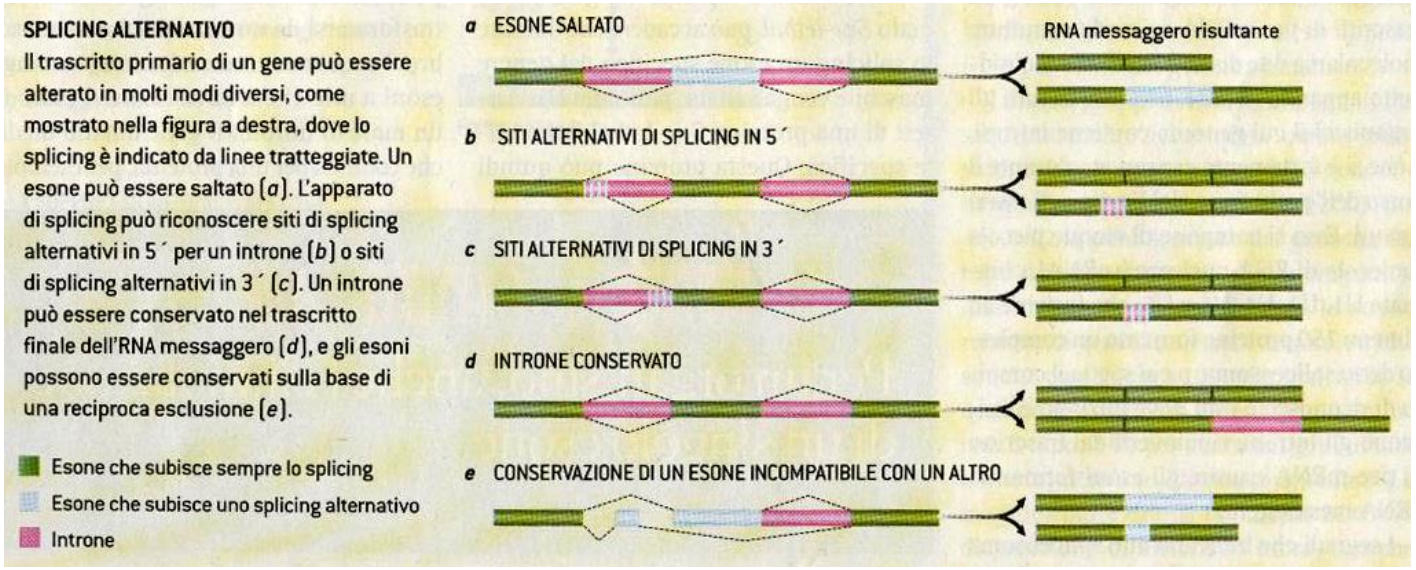




- La lunghezza media degli esoni non varia, quella degli introni aumenta con la filogenesi

Utilità degli introni

- Attraverso lo **splicing alternativo**, nel quale sequenze introniche sono conservate o eliminate, aumenta considerevolmente il numero di proteine codificabili per ogni gene



Utilità degli introni

- Le famiglie Alu, brevi sequenze trasposoniche di circa 300 basi (SINE) terminanti con la caratteristica coda poli-A, sembrano avere giocato un importante ruolo per l'evoluzione proteica.
- Sono circa il 10% del genoma.
- Inserendosi casualmente all'interno degli introni e subendo una mutazione (anche solo di una base), possono creare un sito di splicing alternativo, rendendo un pezzo di introne un nuovo esone.
- Sono presenti solo nei primati.
- Usano il macchinario delle LINE-1 per spostarsi.
- Si comportano come retrotrasposomi.
- Sono responsabili del 17% della variabilità interspecie.

Utilità degli introni

Una mutazione può portare vantaggi per l'organismo ospitante: la vecchia proteina è sempre conservata ma c'è la possibilità di crearne una nuova, che, se utile, può essere conservata.

Introni =Laboratori di sperimentazione per nuove proteine

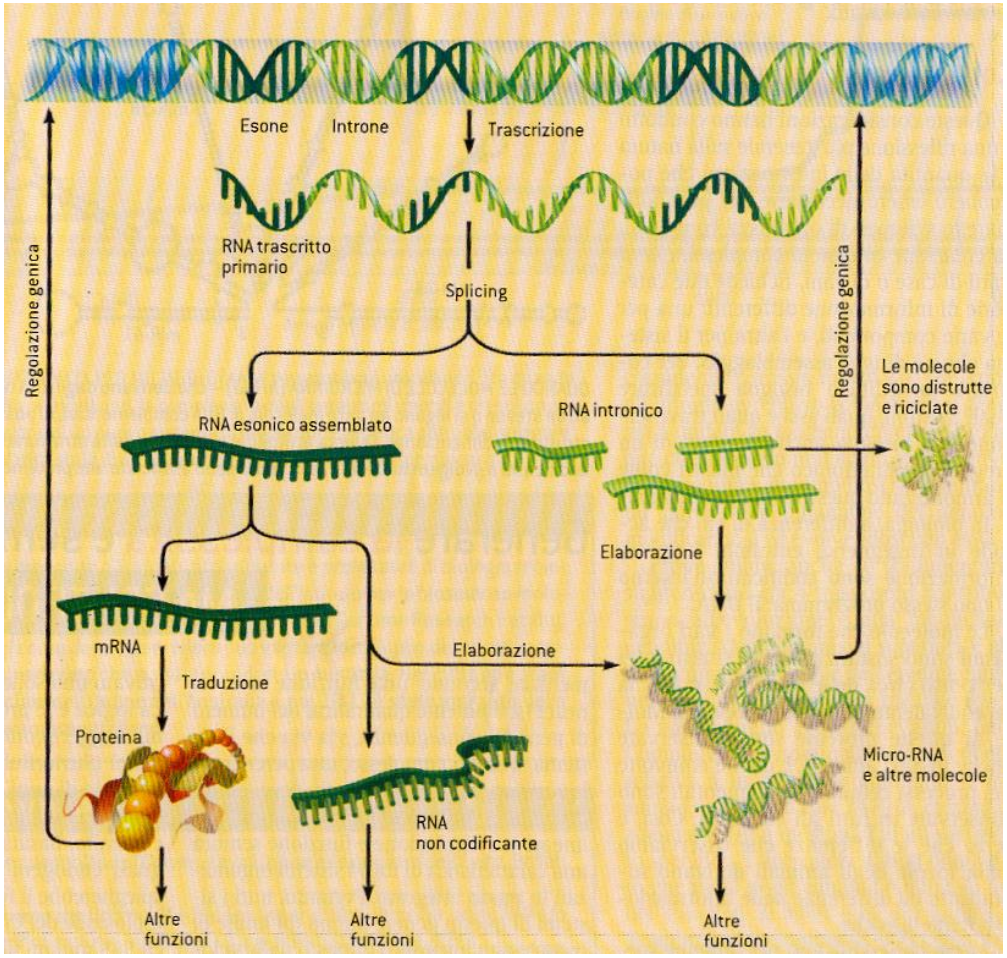
Utilità degli introni

- Recenti studi hanno mostrato che gli introni non vengono immediatamente riciclati dopo essere stati rimossi dall'RNA
- Importantissima funzione di regolazione tramite RNA attivi (miRNA e riboswitch)

micro-RNA

Riboswitch

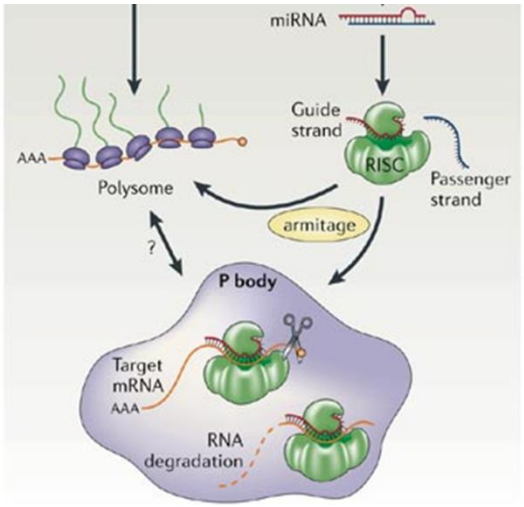
Recente teoria sull'attività genica negli eucarioti

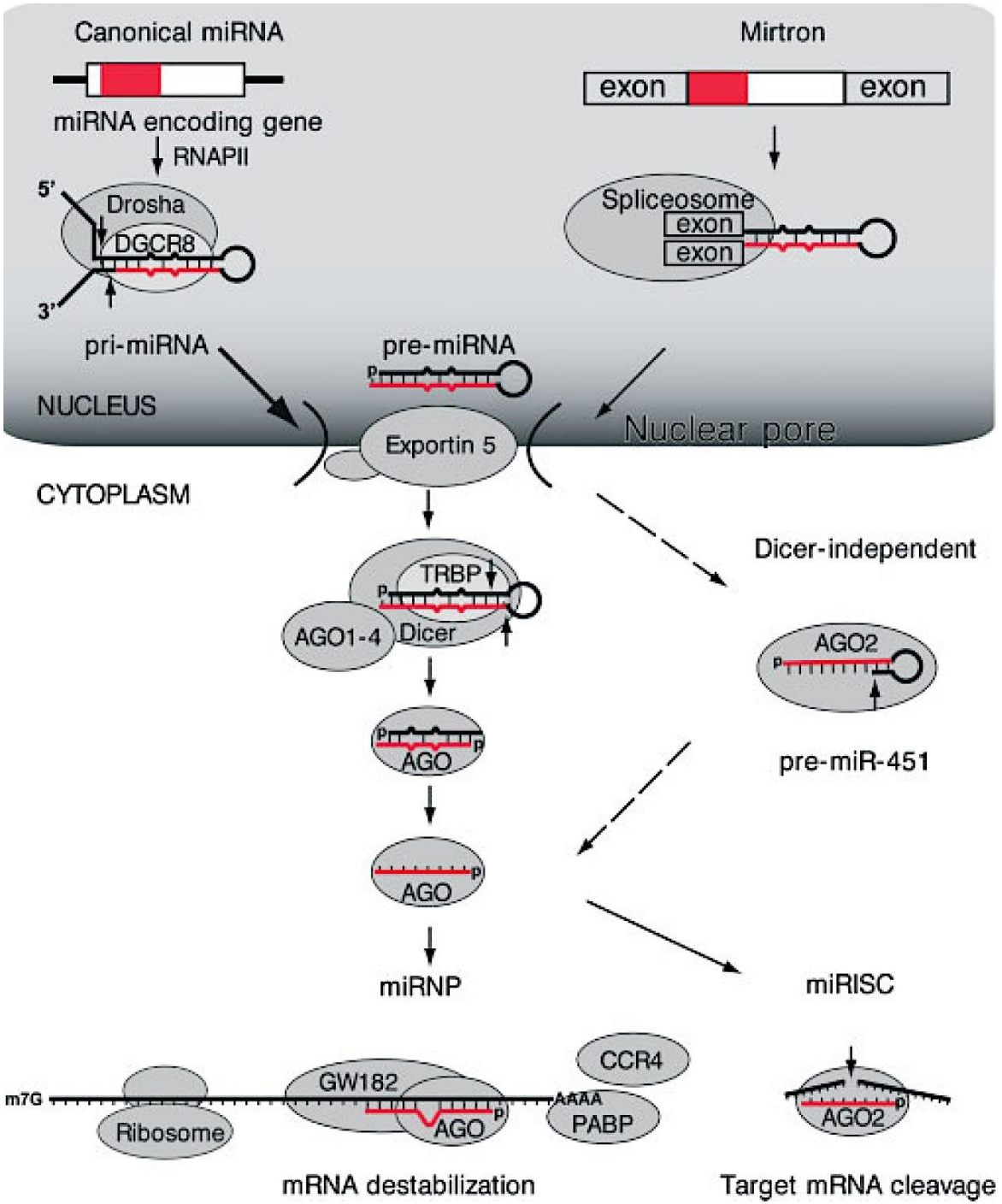


microRNA

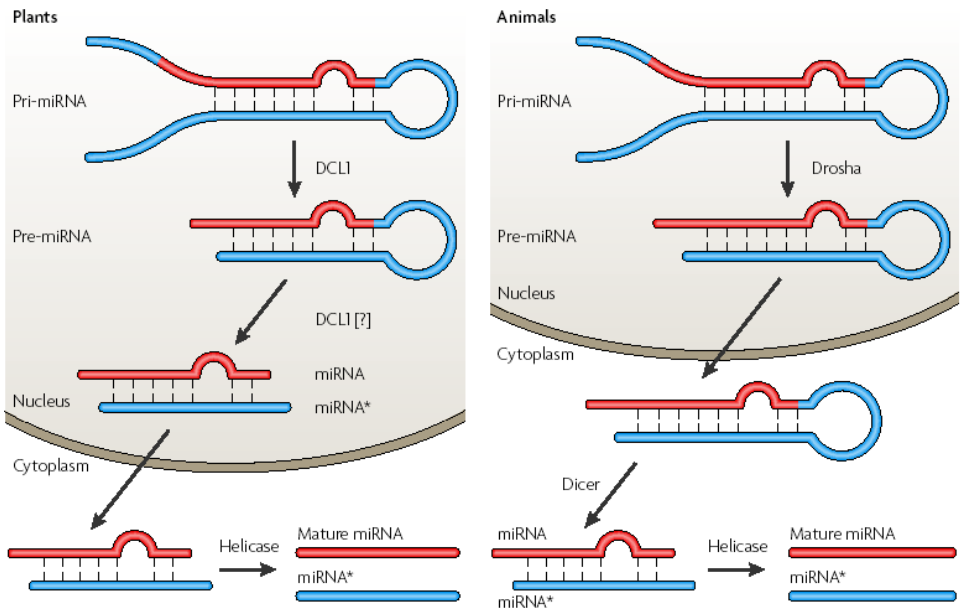
- RNA non codificante di 21-22 nt
- Il trascritto è un pre-miRNA di circa 70 nt che forma una struttura stem-loop
- Il pre-miRNA è processato in una molecola di 21-22 nt dall'enzima Dicer.
- La maggior parte dei miRNA regola l'espressione genica dei loro bersagli mRNA.
- La complementarietà perfetta con i loro bersagli porta a degradazione dell'mRNA.
- La complementarietà imperfetta con mRNA bersaglio porta all'inibizione della traduzione.
- **PRESENTI IN TUTTI GLI EUCARIOTI**

Biogenesi dei miRNA





Generation of miRNAs in Plants and Animals



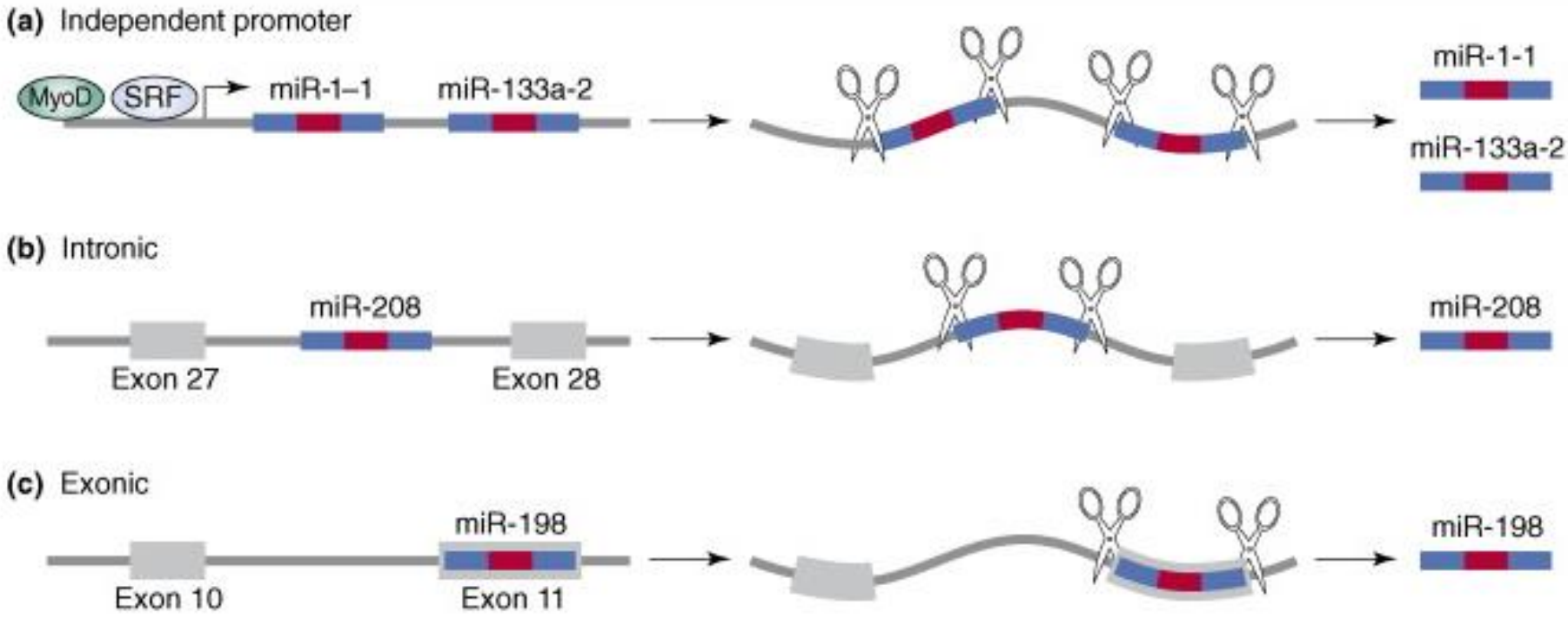
In plants, miRNA maturation occurs in the nucleus

In animals, pre-miRNA is formed in the nucleus and mature miRNA occurs in the cytoplasm

miRNAs regulate ~50% of the human transcriptome

from Chen and Rajewsky, *Nature Rev.Genet*. 8, 93 (2007)

Genomic Organization of miRNA Genes



•Intronic miRNAs often in antisense direction, made from own promoter

•Exonic miRNAs - non-coding (or in alternatively spliced exons)

Zhao Y, Srivastava D, TIBS 32:189,2007

Classification of MiRNAs

- MicroRNAs are classified in two group depending on their origin-

- **Intergenic or Exonic miRNAs:** located between the introns of genes & transcribed by RNA pol II or pol III as a stem loop structure called pri-miRNA

- **Interagenic or Intronic miRNAs:** miRNAs located within an intron of a protein coding gene & transcribed by RNA pol II as part of pre-mRNA

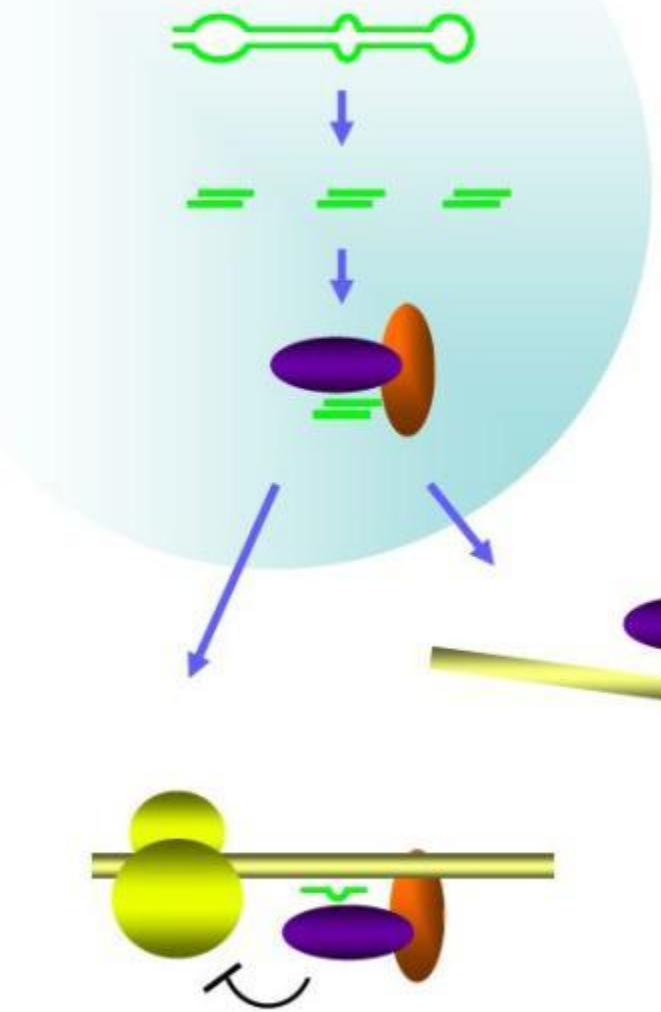
Human miRNAs (February 22, 2010)

- Total miRNA genes in 115 species 10,882
- Total number of miRNAs known 1,580
- Number human miRNAs identified 851
  (attualmente più di 1900)
- Number of human mRNA targets 34,788

- miRNAs can have multiple targets
- Target mRNAs can have multiple miRNA binding sites

miRBase @ http://www.mirbase.org/
MicroCosm @ http://www.ebi.ac.uk/enright-srv/microcosm/

Thousands of microRNAs act in multiple biological events



Developmental timing

Differentiation

Aging

Apoptosis

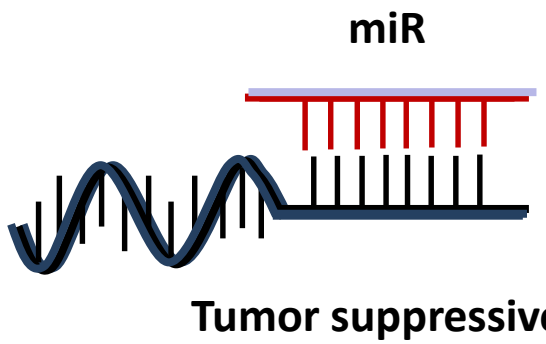
Metabolism

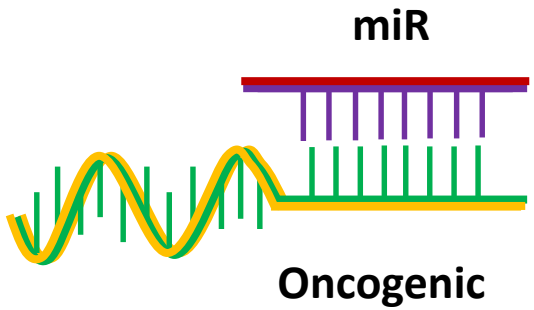
Cancer

…

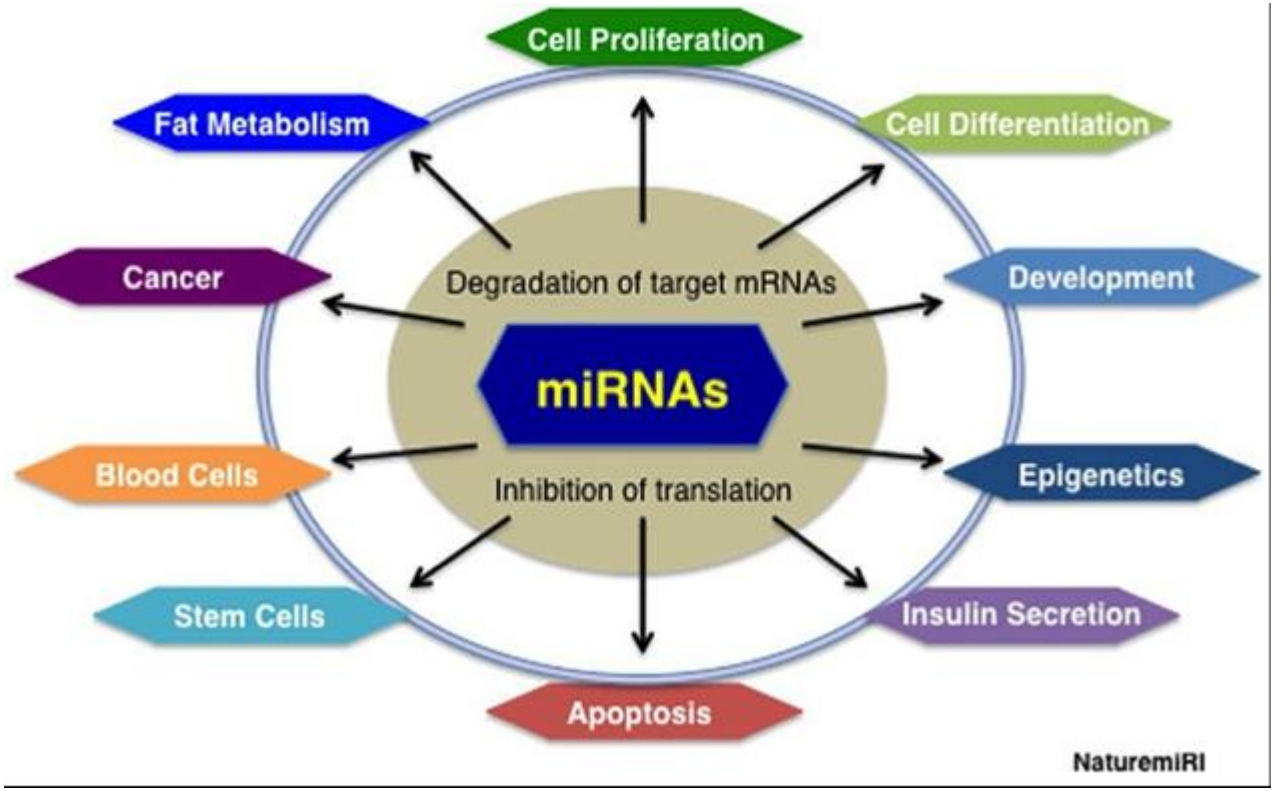
Cell fate/differentiation, Cell cycle…

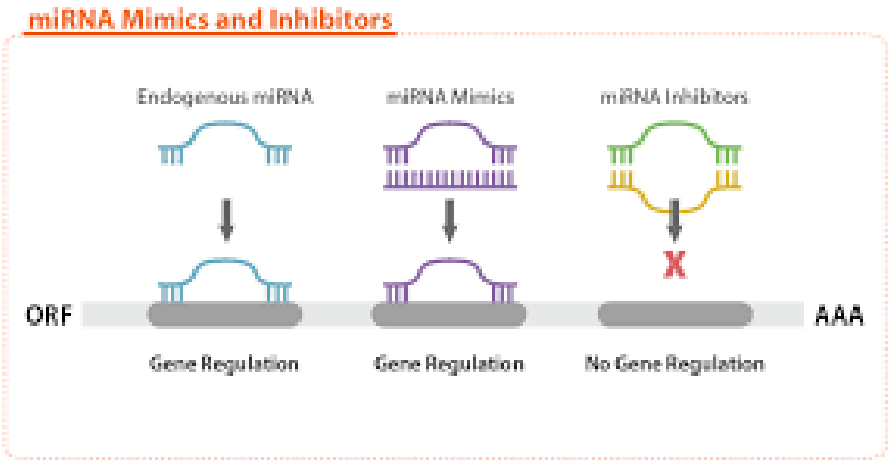
MicroRNA ACTIVITY IN CANCER: TUMOR SUPPRESSIVE OR ONCOGENIC



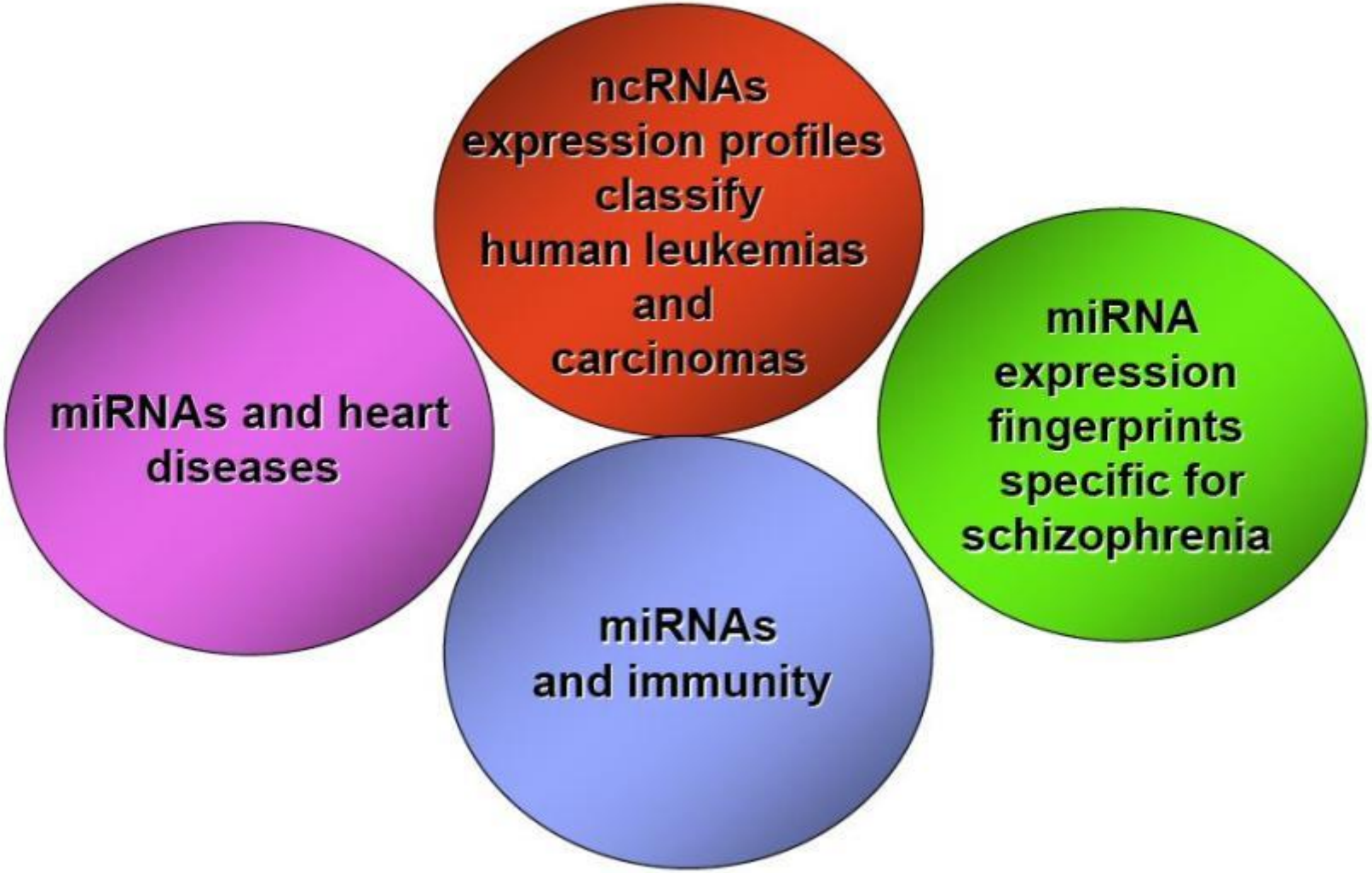


- Suppress expression of oncogenes, growth promoting, survival and angiogenic genes (low in tumors)
- Suppress expression of tumor suppressor, growth inhibitory, proapoptotic genes (high in tumors)


Cell Proliferation
Fat Metabolism
Cell Differentiation
Cancer
Degradation of target mRNAs
Development
miRNAs
Blood Cells
Inhibition of translation
Epigenetics
Stem Cells
Insulin Secretion
Apoptosis
NaturemiRI

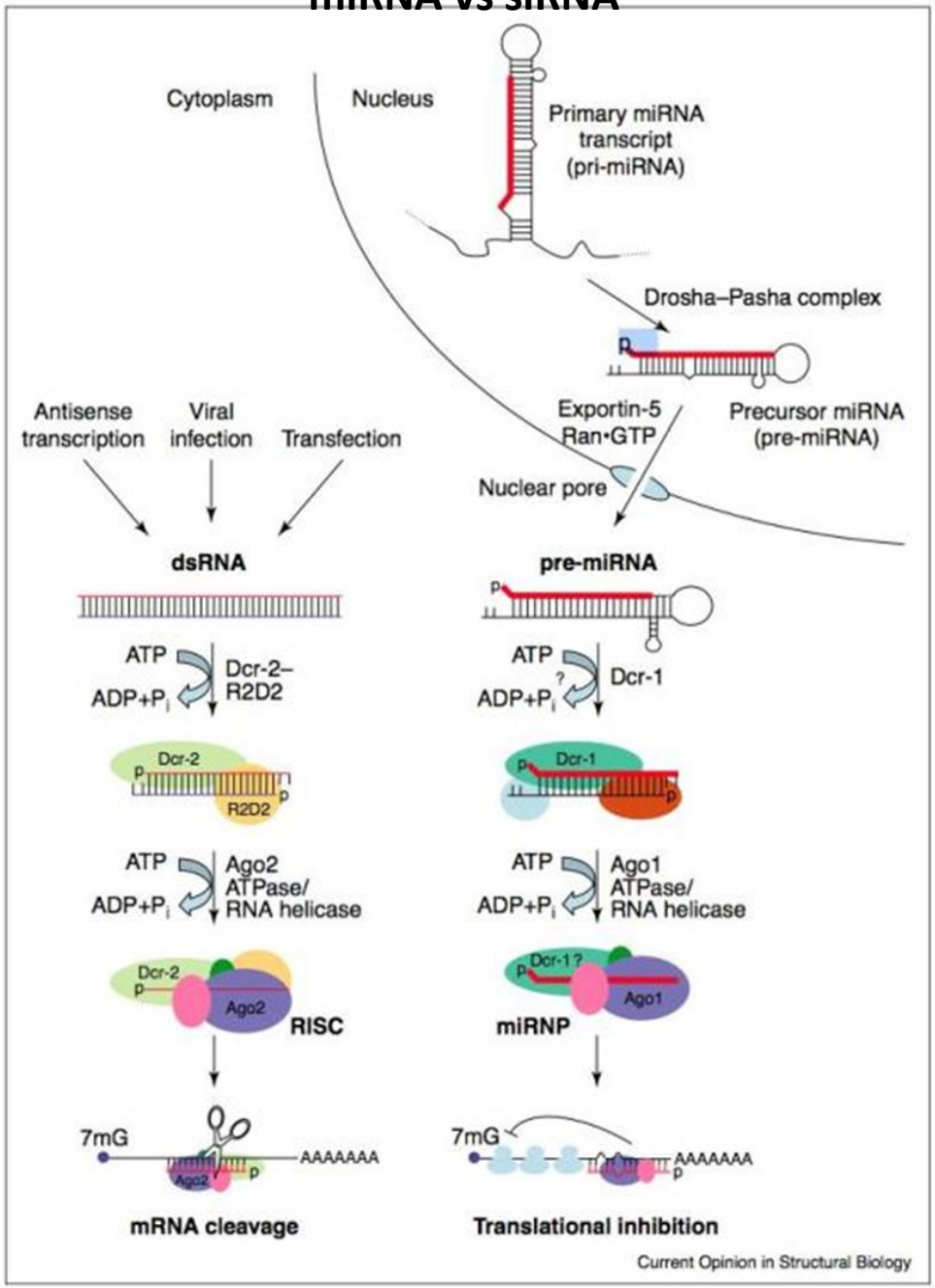

miRNA Mimics and Inhibitors
Endogenous miRNA
miRNA Mimics
miRNA Inhibitors
X
ORF
AAA
Gene Regulation
Gene Regulation
No Gene Regulation

ALTERATIONS OF NONCODING RNAS ARE FOUND IN EVERY TYPE OF HUMAN DISEASE

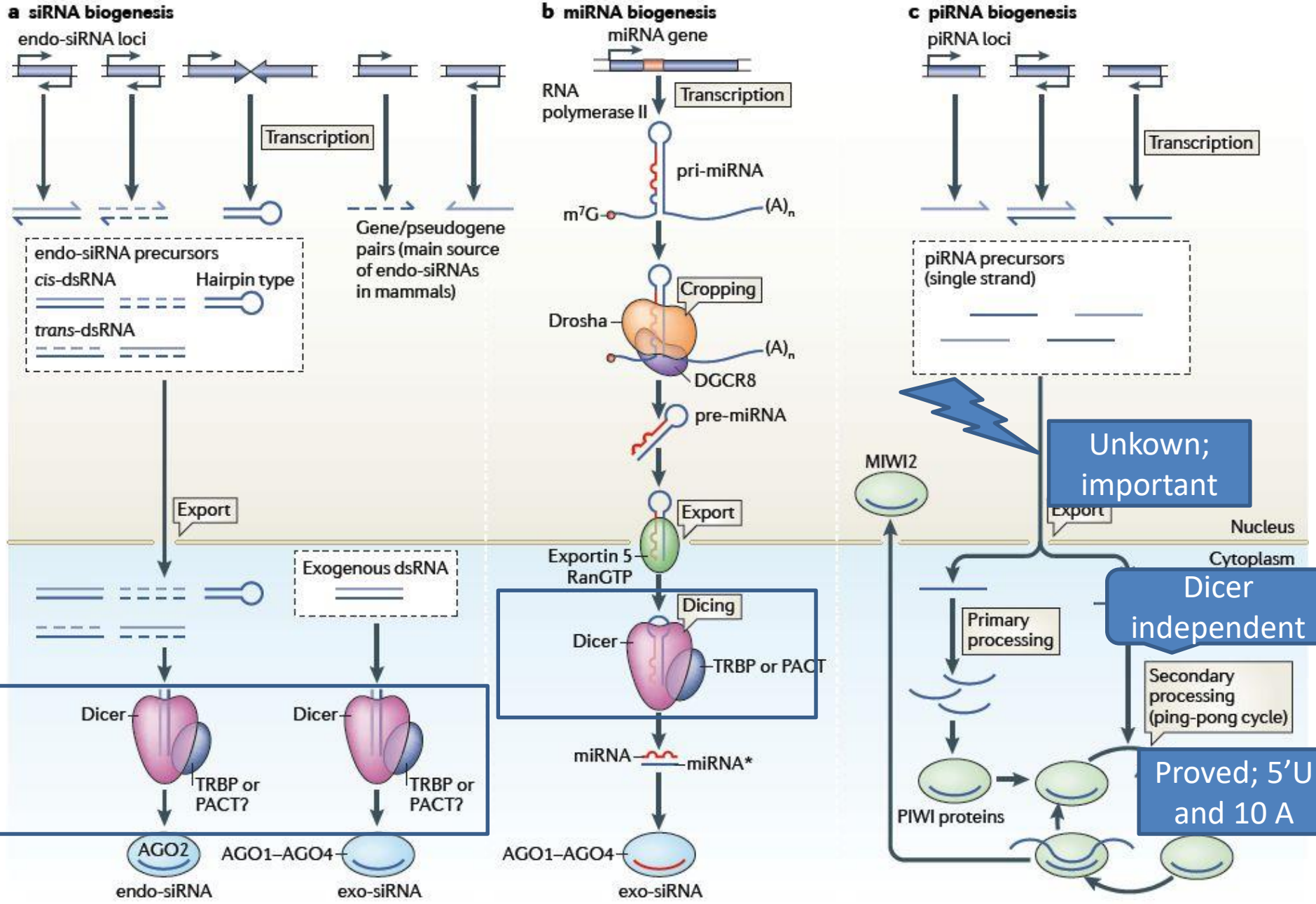


(Calin et al, PNAS 2002; Lu et al, Nature, 2005; Landgraf et al, Cell 2007; Perkins et al Genome Biol 2007; Hansen et al PLoS ONE, 2007; Beveridge et al, Hum Molec Genet 2008, Baltimore D, Nat Immunol 2008; van Rooij, Trends Genet, 2008)

miRNA vs siRNA

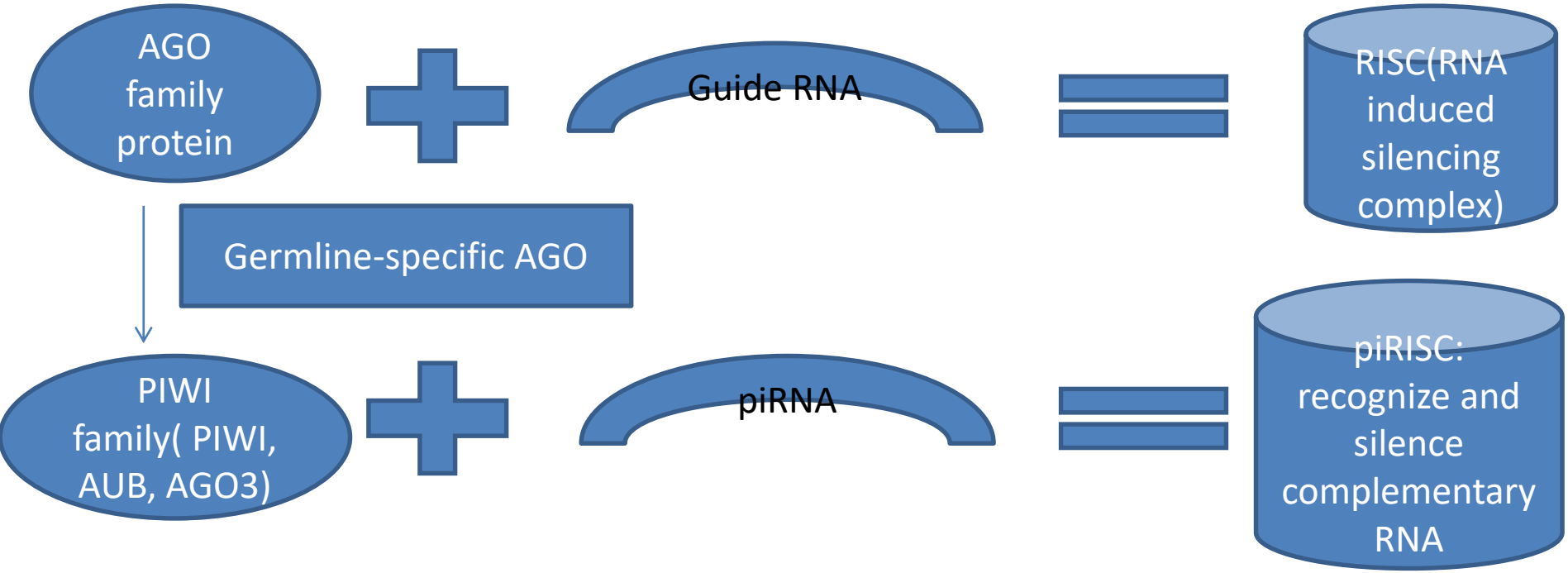


Biogenesis of small RNAs



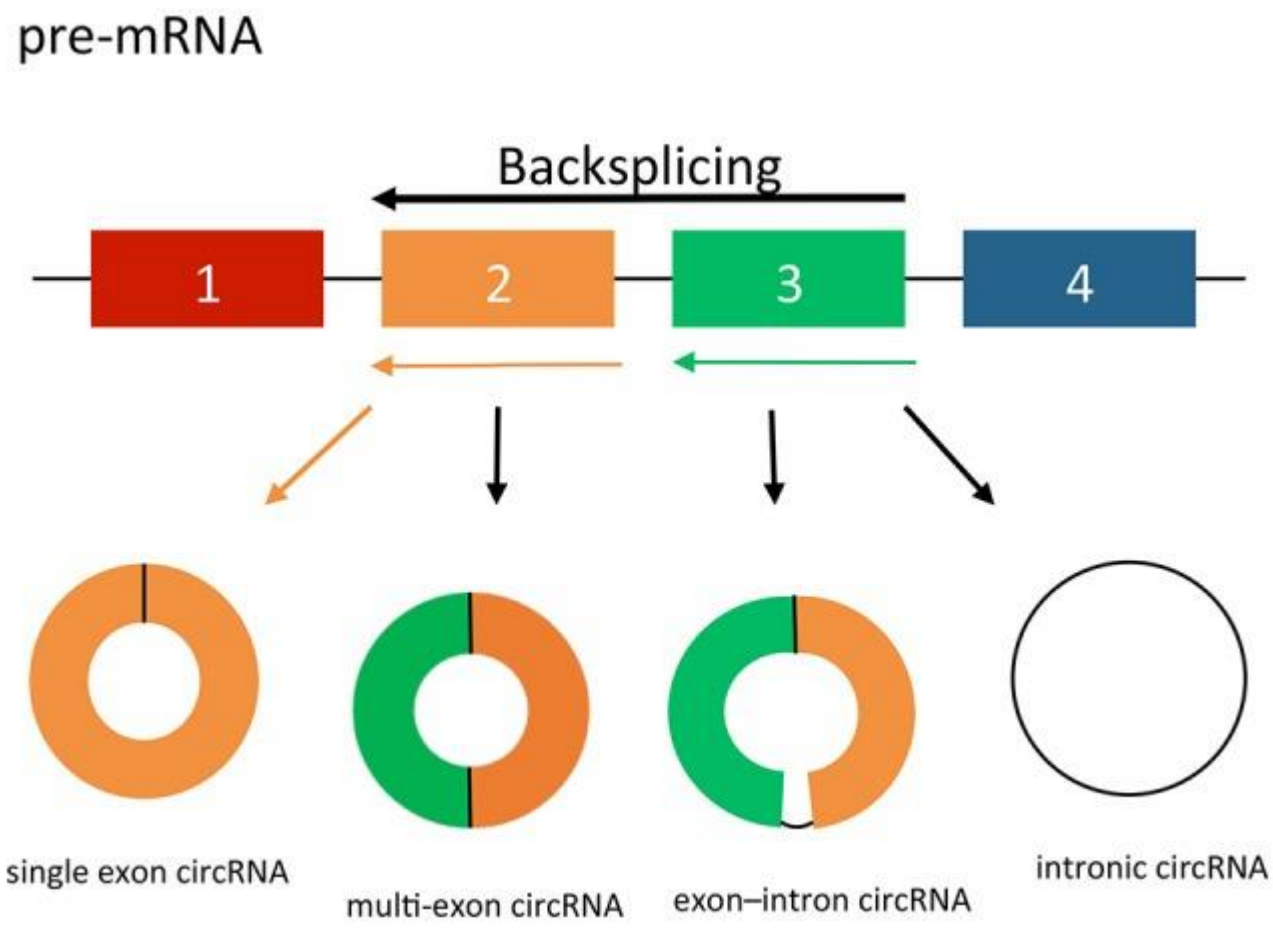
What is piRNA

- Small RNAs: piRNA, siRNA, miRNA.
- piRNA: derive from repetitive genomic element, interact with PIWI

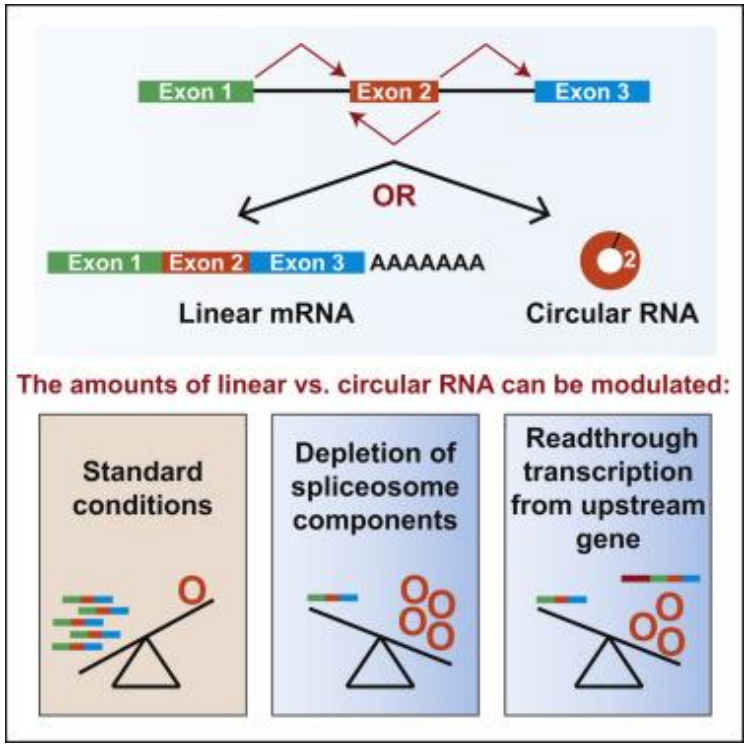


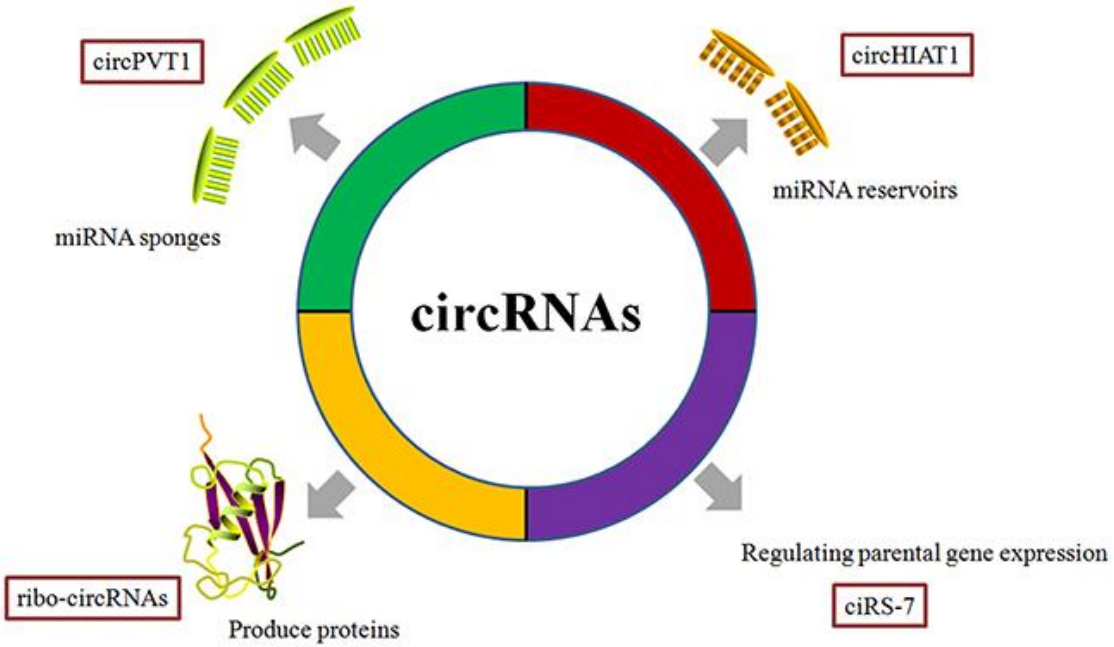
Current Evolutionary studies demonstrate

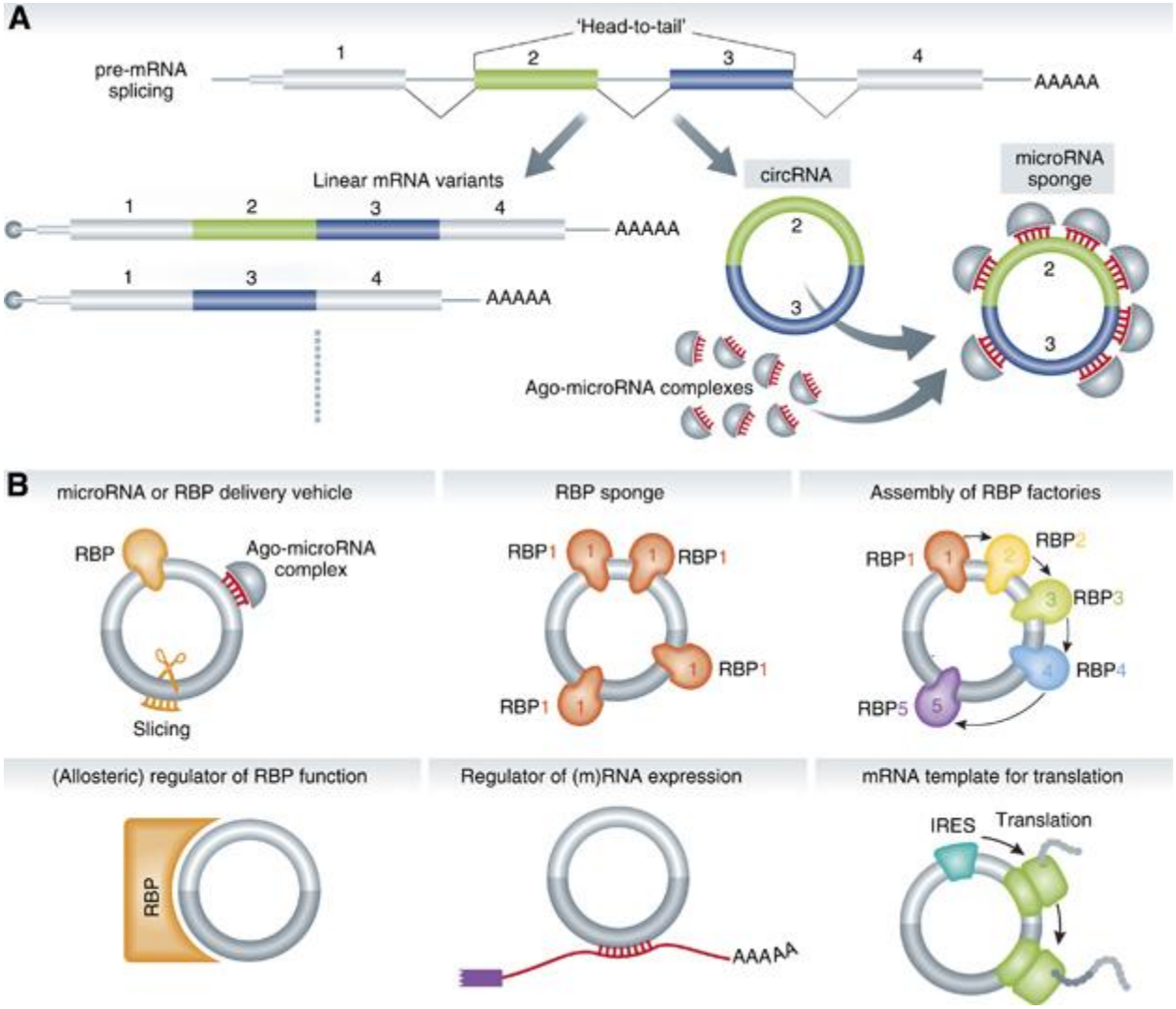
- piRNA strongly repress retrotransposons
- piRNA evolves very fast among species
- piRNA loci locates in low recombination region
- piRNA show a signature of selective constraint in African populations



Circular RNAs are a recently described class of stable RNAs that are naturally resistant to degradation by exonucleases and are generated from thousands of protein-coding genes


Exon 1
Exon 2
Exon 3
OR
Exon 1
Exon 2
Exon 3
AAAAAAA
2
Linear mRNA
Circular RNA
The amounts of linear vs. circular RNA can be modulated:
Standard conditions
Depletion of spliceosome components
Readthrough transcription from upstream gene


circPVT1
circHIAT1
miRNA sponges
miRNA reservoirs
circRNAs
Regulating parental gene expression
ciRS-7
ribo-circRNAs
Produce proteins

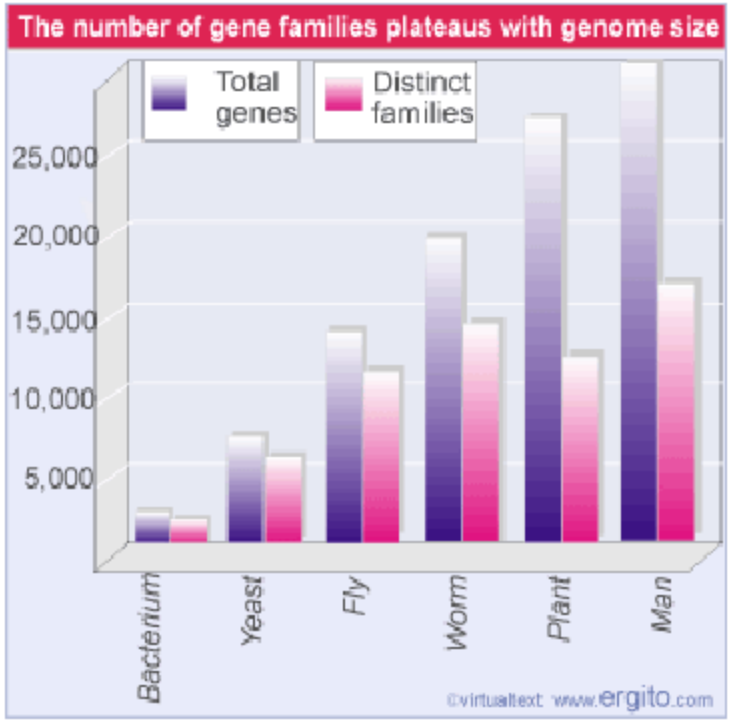

A
'Head-to-tail'
pre-mRNA splicing
1
2
3
4
AAAAA
Linear mRNA variants
1
2
3
4
AAAAA
1
3
4
AAAAA
circRNA
2
3
microRNA sponge
2
3
Ago-microRNA complexes
B
microRNA or RBP delivery vehicle
RBP
Ago-microRNA complex
Slicing
RBP sponge
RBP1
RBP1
RBP1
RBP1
Assembly of RBP factories
RBP1
RBP2
RBP3
RBP4
RBP5
(Allosteric) regulator of RBP function
RBP
Regulator of (m)RNA expression
AAAAA
mRNA template for translation
IRES
Translation

Regulatory RNAs

miRNA	microRNA	22-23 nt, highly conserved	post-transcriptional gene silencing
piRNA	piwi-interacting RNA	26-30 nt	transposon defense, germ cell development
siRNA	small interfering RNA	21-22 nt	gene regulation, transposon control and viral defence
PARs	promoter-associated RNAs	20-200 nt	a general term encompassing a suite of long and short RNAs that overlap promoters and TSSs. inhibit or enhance transcription of nearby genes.
eRNA	enhancer-associated RNAs	several hundreds to 1 kb	
circRNA	circular RNA	Head-to-tail splicing of exons	miRNA sponge, post-transcriptional regulators
ciRNA	circular intronic RNA	Introns of protein-coding genes	associated with Pol II to enhance transcription of host gene
lncRNA	**Long noncoding RNAs**	**>200nt, 5'cap, polyA, RNA pol II, poorly conserved, developmentally regulated, cell type-specific, low-level expression**	

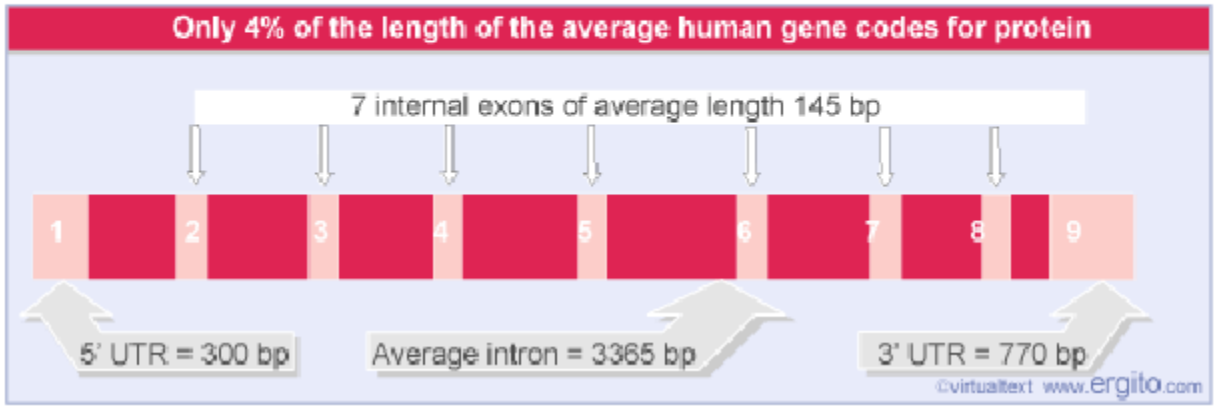
Famiglie di geni

- Alcuni geni sono unici, altri appartengono a famiglie
- Il n° di geni aumenta con la complessità, ma negli eucarioti superiori il n° di famiglie rimane costante



Il gene umano medio

- In media un gene umano ha 9 esoni e 7 introni ed è lungo 27 kb. Gli esoni terminali sono più lunghi. Solo 5% è codificante



I geni occupano il 25% del genoma umano ma solo 1% sono esoni!

Compattezza di alcuni genomi eucariotici

Proprietà del genoma	*S.cerevisiae*	*D.melanogaster*	*H. sapiens*
Densità genica (numero medio di geni per Mb)	479	79	9
Introni per gene (media)	0,04	3	7
% del genoma occupata dalle ripetizioni intersperse	3,4%	12%	44%

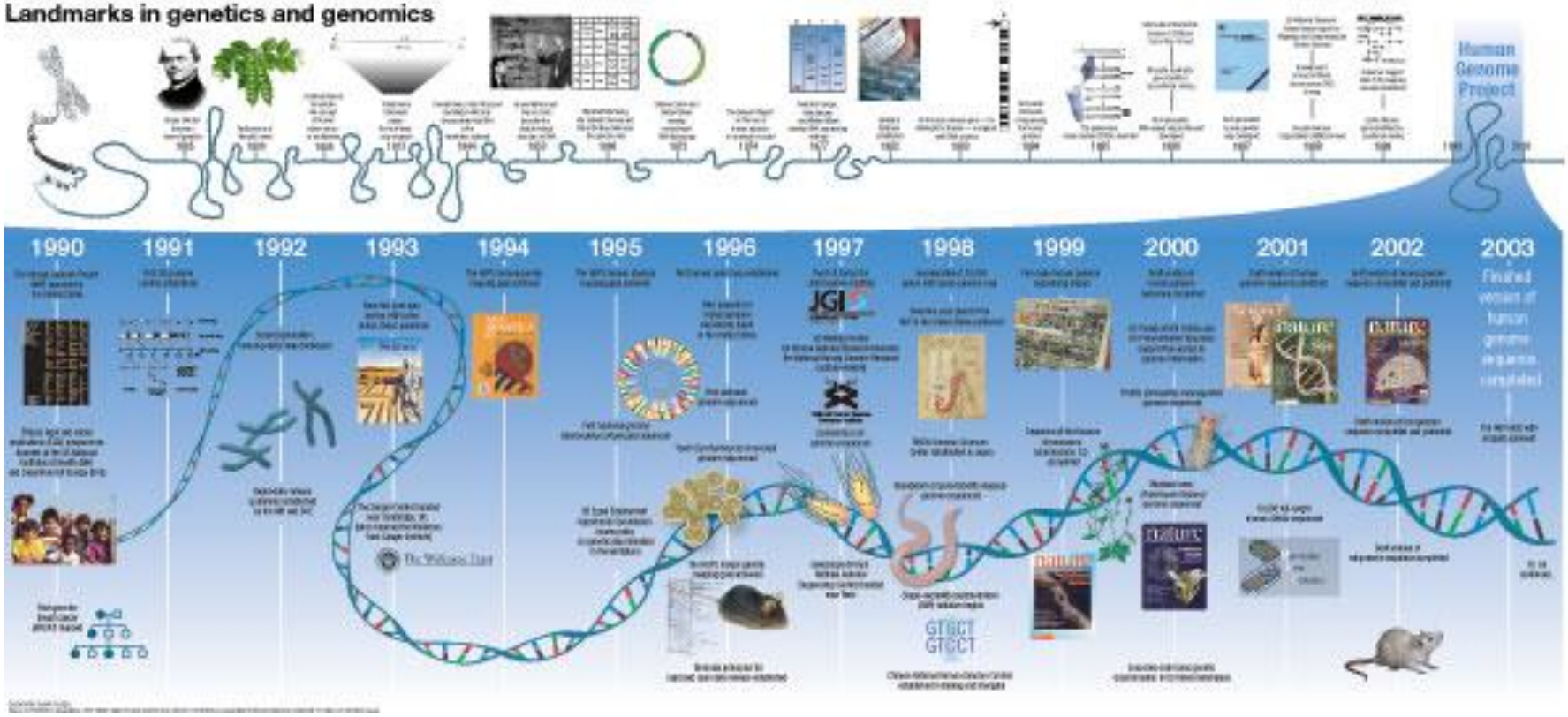
Densità genica:quantità di geni contenuti in una megabase di DNA genomico

Procarioti: 850-1000geni /Mb

Table 1–2 The Numbers of Gene Families, Classified by Function, That Are Common to All Three Domains of the Living World

GENE FAMILY FUNCTION	NUMBER OF "UNIVERSAL" FAMILIES
Information processing	
Translation	63
Transcription	7
Replication, recombination, and repair	13
Cellular processes and signaling	
Cell cycle control, mitosis, and meiosis	2
Defense mechanisms	3
Signal transduction mechanisms	1
Cell wall/membrane biogenesis	2
Intracellular trafficking and secretion	4
Post-translational modification, protein turnover, chaperones	8
Metabolism	
Energy production and conversion	19
Carbohydrate transport and metabolism	16
Amino acid transport and metabolism	43
Nucleotide transport and metabolism	15
Coenzyme transport and metabolism	22
Lipid transport and metabolism	9
Inorganic ion transport and metabolism	8
Secondary metabolite biosynthesis, transport, and catabolism	5
Poorly characterized	
General biochemical function predicted; specific biological role unknown	24

For the purpose of this analysis, gene families are defined as "universal" if they are represented in the genomes of at least two diverse archaea *(Archaeoglobus fulgidus* and *Aeropyrum pernix),* two evolutionarily distant bacteria *(Escherichia coli* and *Bacillus subtilis)* and one eucaryote (yeast, *Saccharomyces cerevisiae).* (Data from R.L. Tatusov, E.V. Koonin and D.J. Lipman, *Science* 278:631–637, 1997; R.L. Tatusov et al., *BMC Bioinformatics* 4:41, 2003; and the COGs database at the US National Library of Medicine.)


Landmarks in genetics and genomics
Human Genome Project
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
Finished version of human genome sequence completed
JGI
GTGCT
GTCCT