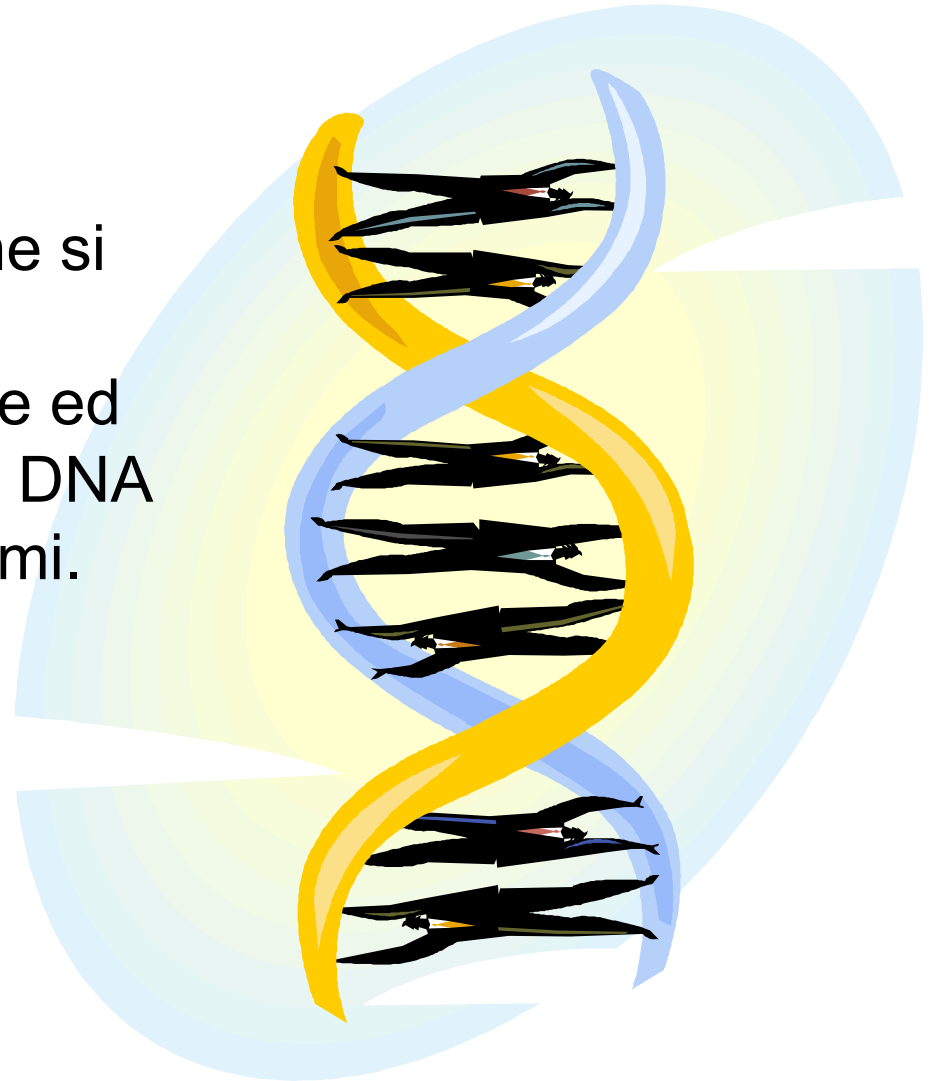


# Quali sono e come sono organizzate le sequenze nel genoma?

**La genomica:** disciplina che si occupa dello studio dell'organizzazione, funzione ed evoluzione delle sequenze di DNA contenute nei diversi genomi.



# Il Genoma Umano

$10^{13}$  cellule/individuo.

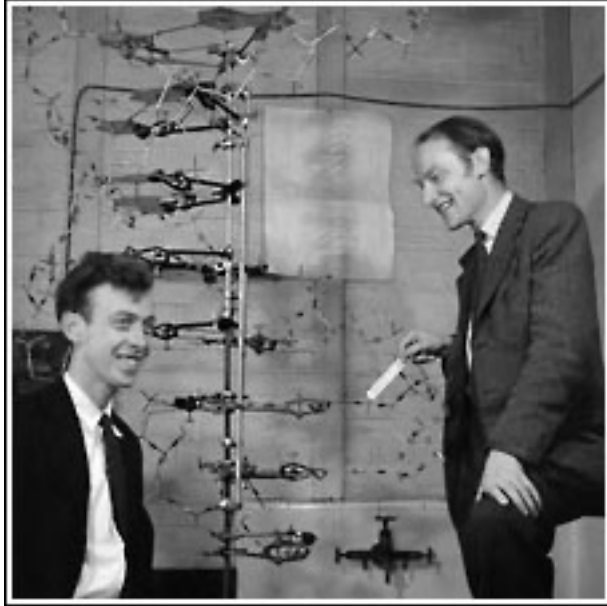
Ciascuna con un genoma nucleare  
e molti genomi mitocondriali.

## Il DNA nucleare

contiene ca  $3,1 \times 10^9$  coppie di basi,  
suddivise in 23 molecole lineari: i  
cromosomi. Il più piccolo  
contiene ca. 50.000.000 di nucleotidi; il  
più grande ca. 250.000.000.



# La genetica moderna



**1953** James Watson e Francis Crick determinano la struttura del DNA

**1977** Gli scienziati americani Allan Maxam and Walter Gilbert e l'inglese Frederick Sanger mettono a punto 2 diversi metodi per sequenziare il DNA, cioè per "leggere" la successione di basi che lo compongono. Il metodo di Sanger, in forma automatizzata, è quello tuttora utilizzato.

**1985** Lo scienziato americano Kary Mullis inventa la PCR, una tecnica che permette di moltiplicare artificialmente il DNA, anche se presente in quantità minima.

**1986** Il premio Nobel Renato Dulbecco e Leroy Hood lanciano l'idea di sequenziare l'intero genoma Umano.



Negli Stati Uniti nasce ufficialmente lo **Human Genome Project** (HGP), sotto la guida di James Watson. Negli anni successivi Regno Unito, Giappone, Francia, Germania, Cina si uniscono al progetto formando un consorzio pubblico internazionale. In Italia il progetto genoma si interrompe nel 1995.

**Il U.S. Human Genome Project inizia formalmente nel 1990 sotto il coordinamento del U.S. Department of Energy e del National Institutes of Health (NIH)..**

**Gli obiettivi principali del progetto erano:**

**Costruire una mappa completa del genoma umano**

**Immagazzinare i dati ottenuti in banche dati**

**Migliorare gli strumenti per l'analisi delle sequenze**

**La sequenza completa del genoma umano e di genomi di organismi modello**

**Trasferire i dati e le tecnologie sviluppate al settore privato**

**Dare una regolamentazione legislativa ed etica**

**La previsione era di terminare il progetto in 15 anni ma lo sviluppo tecnologico ha, di fatto, accelerato le tappe**

# Il progetto genoma umano comprendeva il sequenziamento completo di:

- un genoma umano (3100 Mb);
- un batterio (*Escherichia coli*, 4,6 Mb);
- un lievito (*Saccharomyces cerevisiae*, 12,1 Mb);
- un nematode (*Caenorhabditis elegans*, 97 Mb);
- un insetto (*Drosophila melanogaster*, 180 MB);
- un mammifero (*Mus musculus*, 2600 Mb).



Successivamente a questi cinque organismi è stata aggiunta  
-una pianta *Arabidopsis thaliana* (120 Mb).



# Mappe fisiche

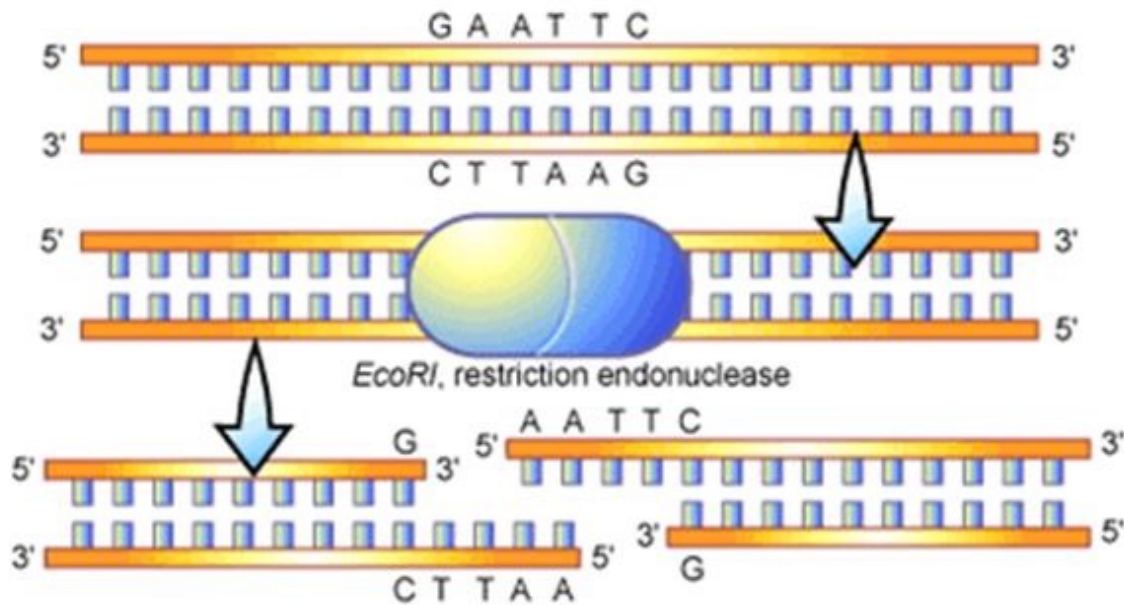
Si allestiscono mediante analisi diretta del DNA e permettono di posizionare le diverse sequenze in relazione a distanze misurate in numero di basi (pb, kb, Mb).

Le mappe fisiche più utilizzate sono quelle allestite a partire da cloni di librerie genomiche (mappe di sequenze).

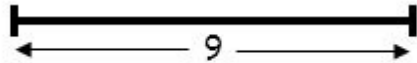
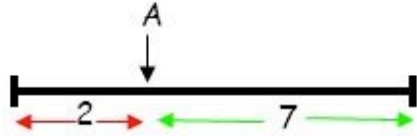
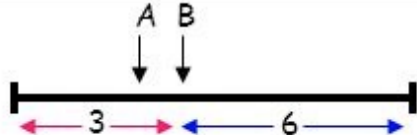
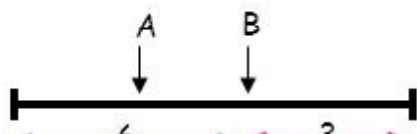
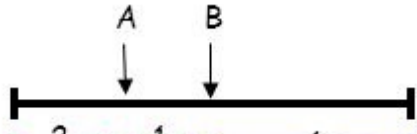

I diversi tipi di mappa variano fondamentalmente per il loro grado di risoluzione, ovvero per la capacità di misurare la separazione di elementi vicini gli uni agli altri. Maggiore è la risoluzione, migliore è la mappa.

# enzimi di restrizione

Sono enzimi presenti nei procarioti che difendono il batterio dal DNA estraneo, tagliandolo in punti specifici.

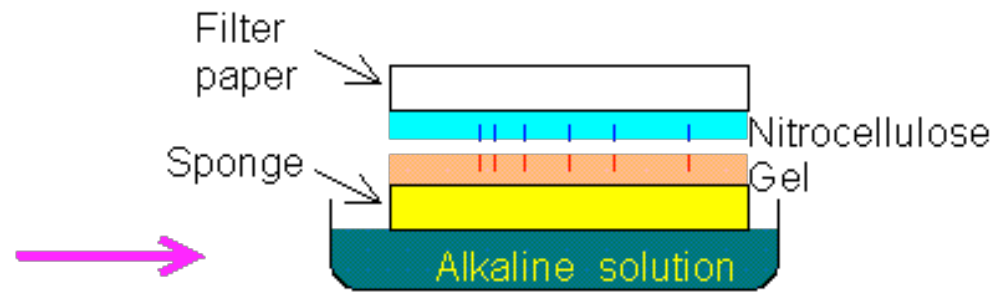


# ...MAPPATURA di RESTRIZIONE

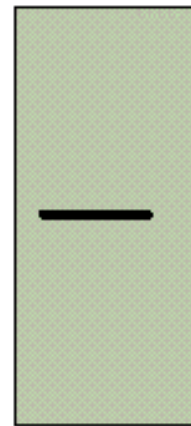
TRATTAMENTO	DIMENSIONE dei FRAMMENTI (Kb)	INTERPRETAZIONE
Nessuna digestione	9	
Enzima A	2 + 7	
Enzima B	3 + 6	
		
Enzima A + B	1 + 2 + 6	
	2 + 3 + 4	



# Gel electrophoresis



Hybridize with probe

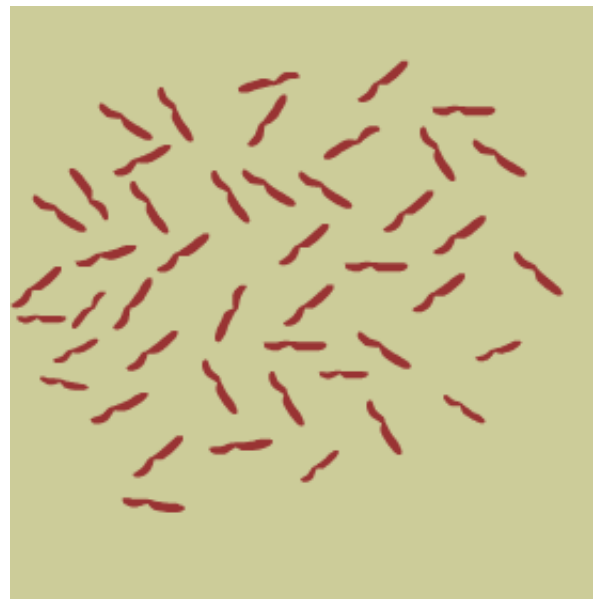


Autoradiogram

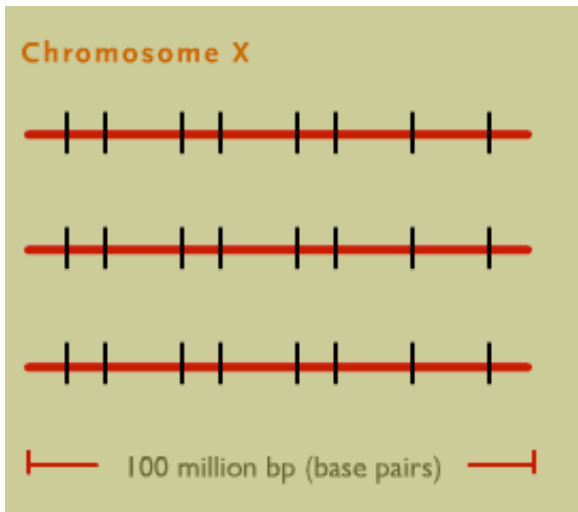
Nitrocellulose filter

# Creazione di librerie genomiche

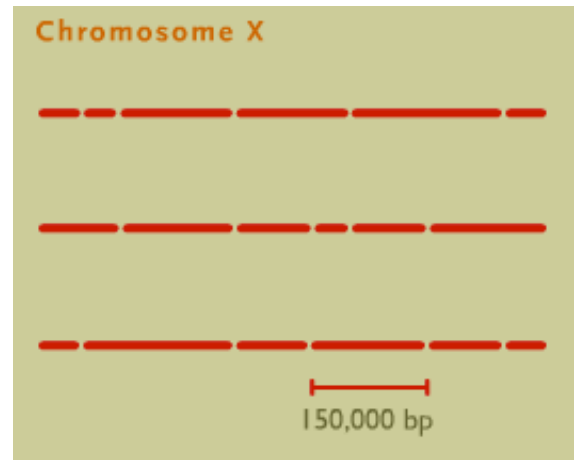
## Il genoma umano

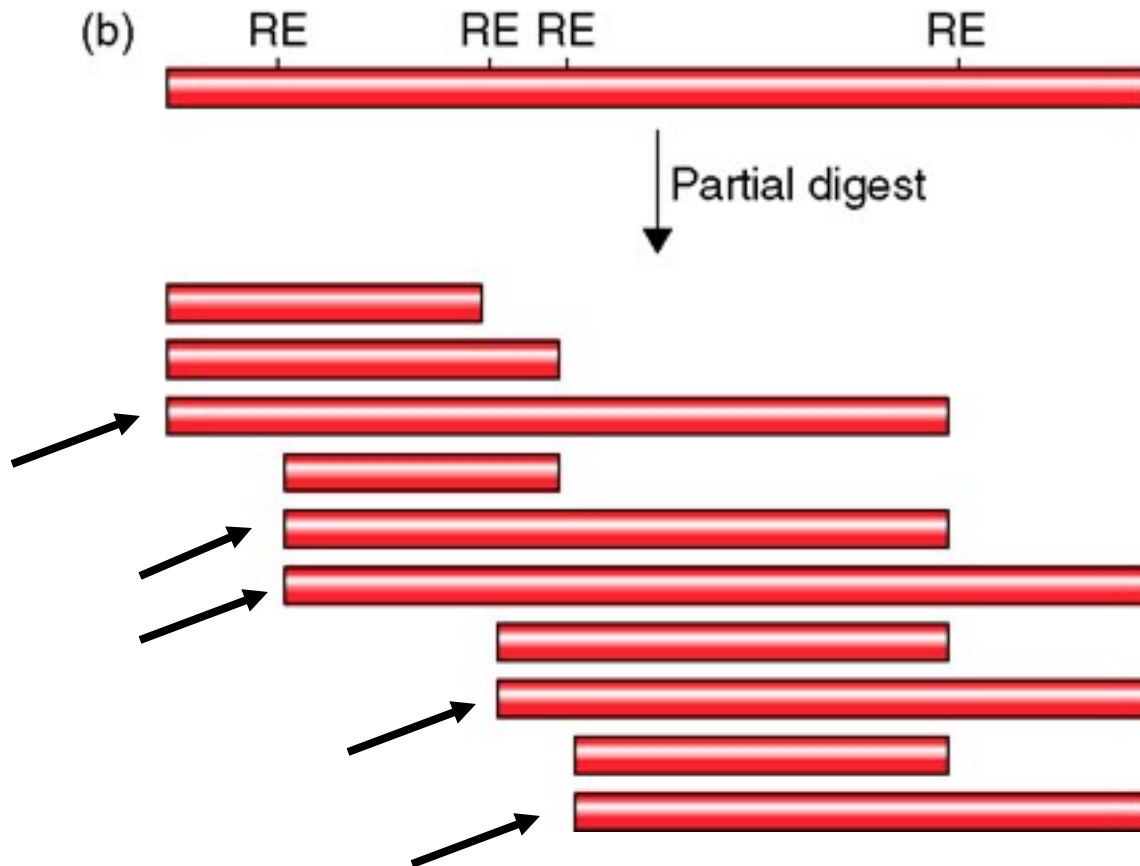
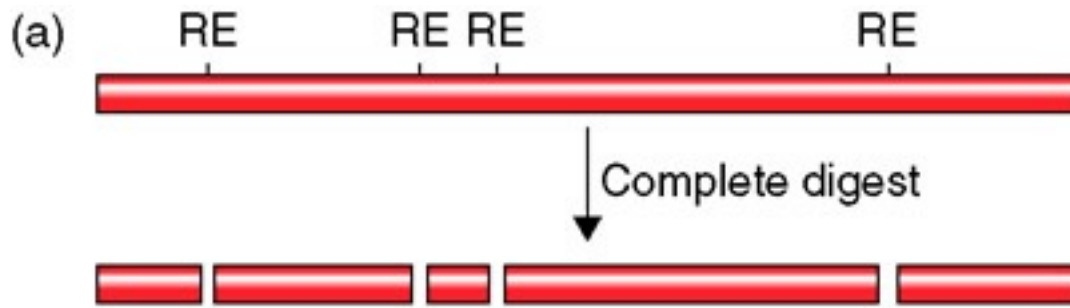


## Siti di restrizione potenziali per un dato enzima



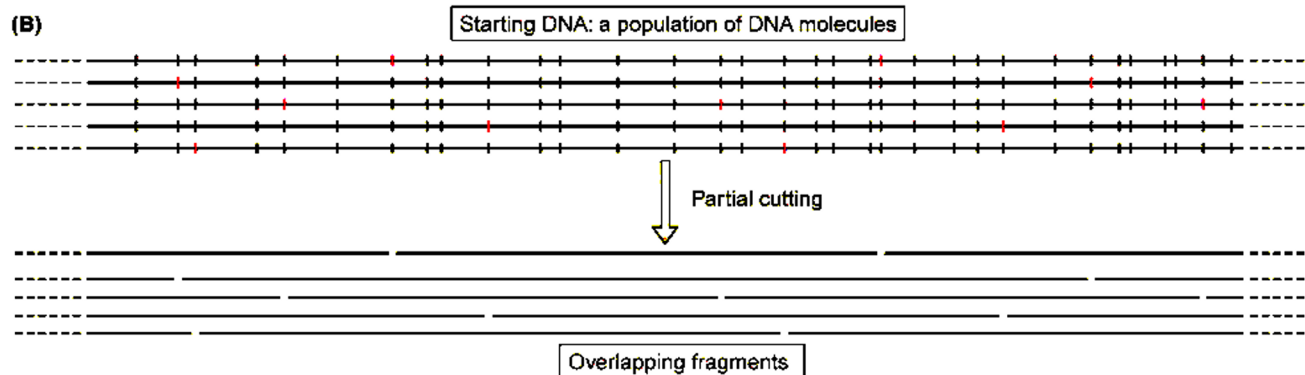
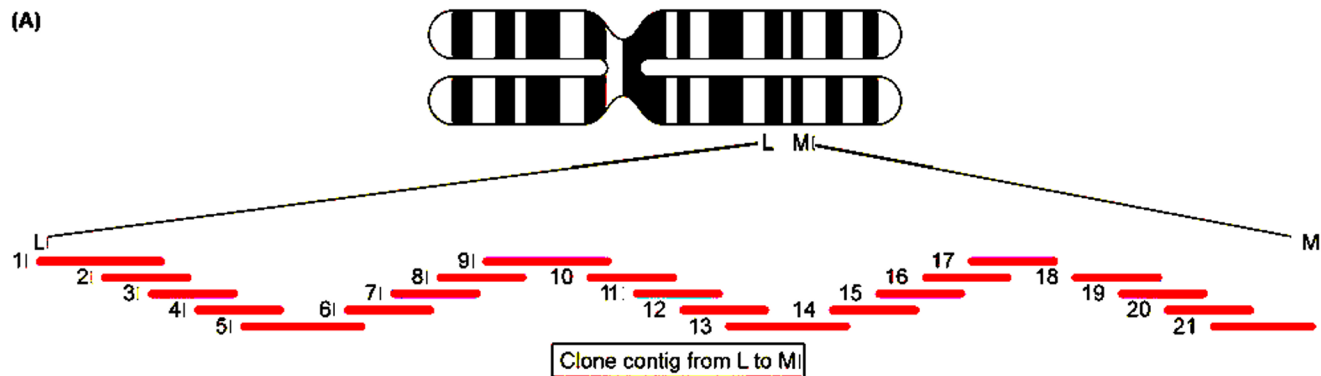
## Digestione parziale



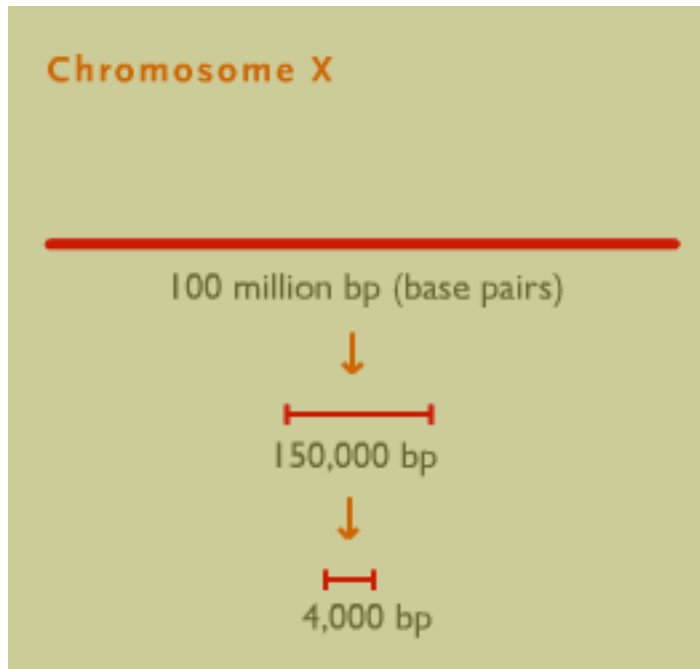


# Costruzione della mappa fisica di una regione

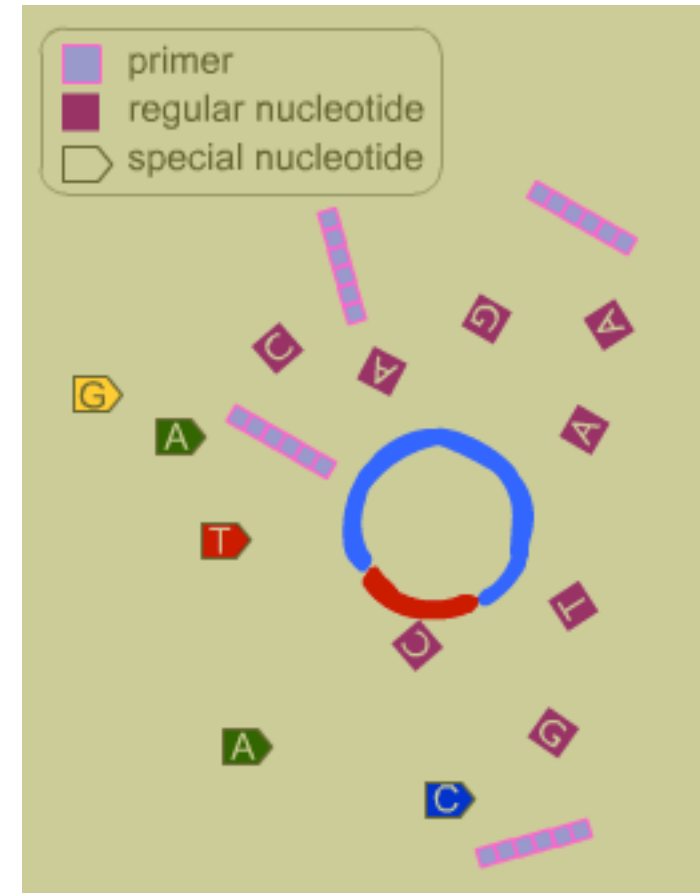
Obiettivo: ordinare lungo il cromosoma cloni di DNA con inserti che si sovrappongono - **CONTIGUI** -



Ogni clone (150-200 kb) mappato, viene  
ulteriormente frazionato e  
sub-clonato in vettore plasmidico



## Sequenziamento

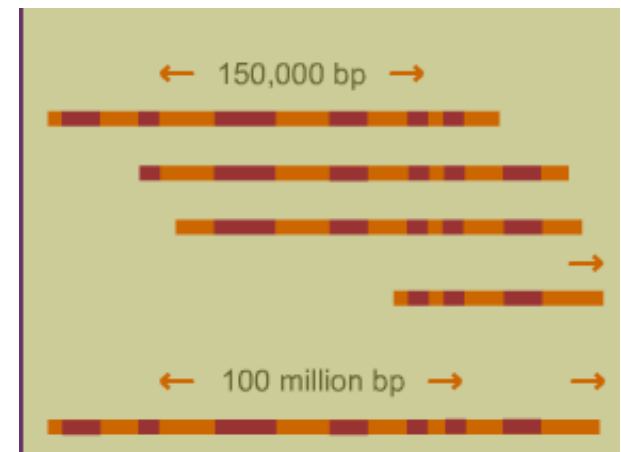
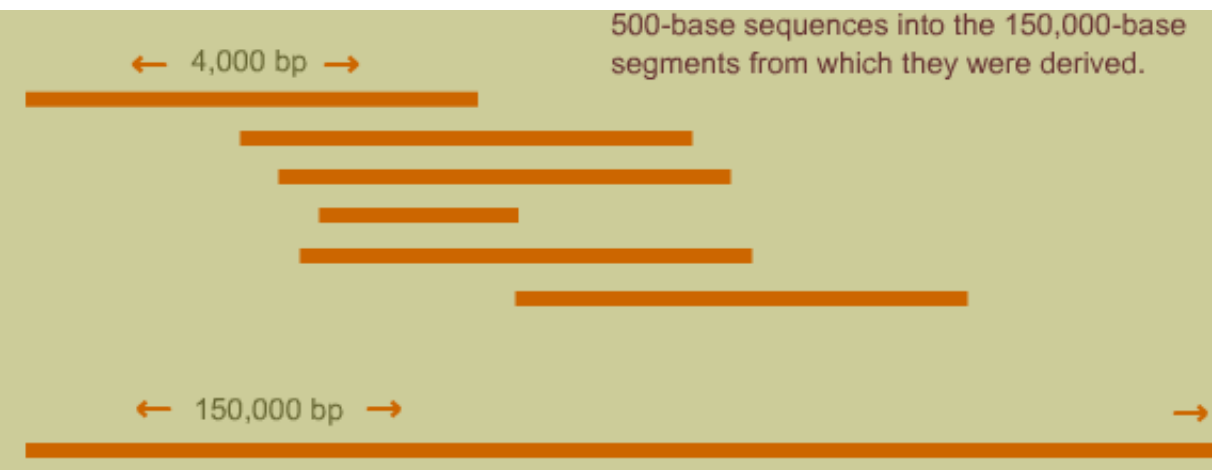


ACAAGTAAC

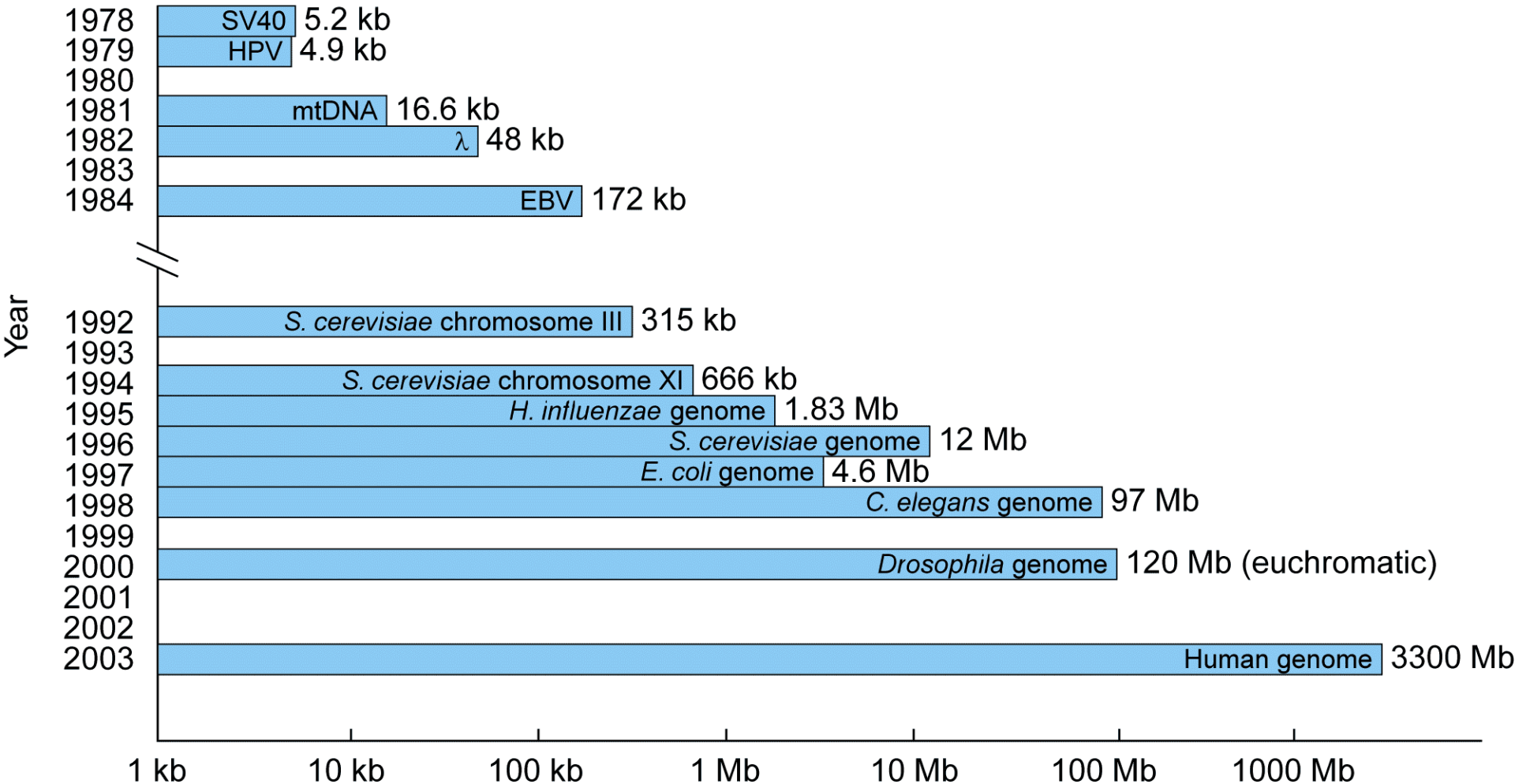
le sequenze dei sub-cloni vengono allineate.....

```
GGGCAACAAGT
  ACAAGTAAC
    AGTAACTTAT
      AAGTAACTTATCGT
        CTTATCGTTTCAGCC
```

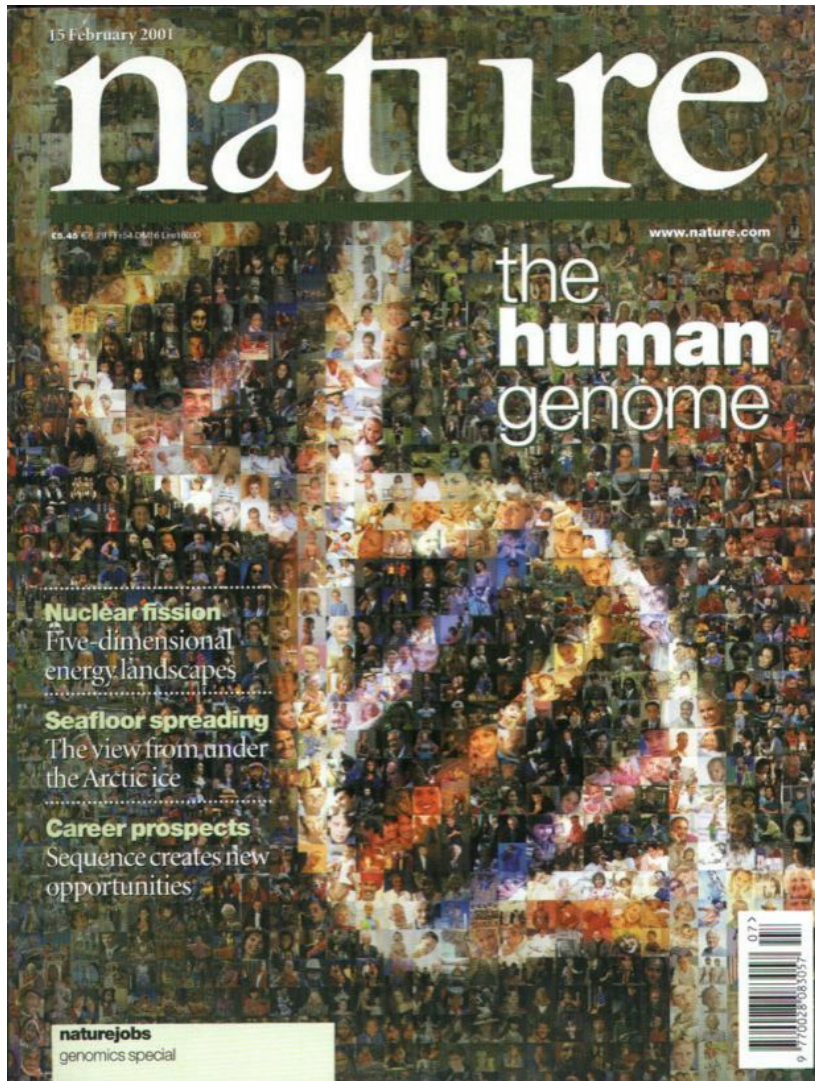
...fino ad ottenere la sequenza completa del clone genomico da cui derivavano



# Sequenziamento dei genomi nel tempo



Febbraio 2001



Finito nel 2003



## RESEARCH ARTICLE

## HUMAN GENOMICS

## The complete sequence of a human genome

2023

Sergey Nurk<sup>1†</sup>, Sergey Koren<sup>1†</sup>, Arang Rhie<sup>1†</sup>, Mikko Rautiainen<sup>1†</sup>, Andrey V. Bzikadze<sup>2</sup>, Alla Mikheenko<sup>3</sup>, Mitchell R. Vollger<sup>4</sup>, Nicolas Altemose<sup>5</sup>, Lev Uralsky<sup>6,7</sup>, Ariel Gershman<sup>8</sup>, Sergey Aganezov<sup>9†</sup>, Savannah J. Hoyt<sup>10</sup>, Mark Diekhans<sup>11</sup>, Glennis A. Logsdon<sup>4</sup>, Michael Alonge<sup>9</sup>, Stylianos E. Antonarakis<sup>12</sup>, Matthew Borchers<sup>13</sup>, Gerard G. Bouffard<sup>14</sup>, Shelise Y. Brooks<sup>14</sup>, Gina V. Caldas<sup>15</sup>, Nae-Chyun Chen<sup>9</sup>, Haoyu Cheng<sup>16,17</sup>, Chen-Shan Chin<sup>18</sup>, William Chow<sup>19</sup>, Leonardo G. de Lima<sup>13</sup>, Philip C. Dishuck<sup>4</sup>, Richard Durbin<sup>19,20</sup>, Tatiana Dvorkina<sup>3</sup>, Ian T. Fiddes<sup>21</sup>, Giulio Formenti<sup>22,23</sup>, Robert S. Fulton<sup>24</sup>, Arkarachai Fungtammasan<sup>18</sup>, Erik Garrison<sup>11,25</sup>, Patrick G. S. Grady<sup>10</sup>, Tina A. Graves-Lindsay<sup>26</sup>, Ira M. Hall<sup>27</sup>, Nancy F. Hansen<sup>28</sup>, Gabrielle A. Hartley<sup>10</sup>, Marina Haukness<sup>11</sup>, Kerstin Howe<sup>19</sup>, Michael W. Hunkapiller<sup>29</sup>, Chirag Jain<sup>1,30</sup>, Miten Jain<sup>11</sup>, Erich D. Jarvis<sup>22,23</sup>, Peter Kerpedjiev<sup>31</sup>, Melanie Kirsche<sup>9</sup>, Mikhail Kolmogorov<sup>32</sup>, Jonas Korlach<sup>29</sup>, Milinn Kremitzki<sup>26</sup>, Heng Li<sup>16,17</sup>, Valerie V. Maduro<sup>33</sup>, Tobias Marschall<sup>34</sup>, Ann M. McCartney<sup>1</sup>, Jennifer McDaniel<sup>35</sup>, Danny E. Miller<sup>4,36</sup>, James C. Mullikin<sup>14,28</sup>, Eugene W. Myers<sup>37</sup>, Nathan D. Olson<sup>35</sup>, Benedict Paten<sup>11</sup>, Paul Peluso<sup>29</sup>, Pavel A. Pevzner<sup>32</sup>, David Porubsky<sup>4</sup>, Tamara Potapova<sup>13</sup>, Evgeny I. Rogaev<sup>6,7,38,39</sup>, Jeffrey A. Rosenfeld<sup>40</sup>, Steven L. Salzberg<sup>9,41</sup>, Valerie A. Schneider<sup>42</sup>, Fritz J. Sedlazeck<sup>43</sup>, Kishwar Shafin<sup>11</sup>, Colin J. Shew<sup>44</sup>, Alaina Shumate<sup>41</sup>, Ying Sims<sup>19</sup>, Arian F. A. Smit<sup>45</sup>, Daniela C. Soto<sup>44</sup>, Ivan Sovic<sup>29,46</sup>, Jessica M. Storer<sup>45</sup>, Aaron Streets<sup>5,47</sup>, Beth A. Sullivan<sup>48</sup>, Françoise Thibaud-Nissen<sup>42</sup>, James Torrance<sup>19</sup>, Justin Wagner<sup>35</sup>, Brian P. Walenz<sup>1</sup>, Aaron Wenger<sup>29</sup>, Jonathan M. D. Wood<sup>19</sup>, Chunlin Xiao<sup>42</sup>, Stephanie M. Yan<sup>49</sup>, Alice C. Young<sup>14</sup>, Samantha Zarate<sup>9</sup>, Urvashi Surti<sup>50</sup>, Rajiv C. McCoy<sup>49</sup>, Megan Y. Dennis<sup>44</sup>, Ivan A. Alexandrov<sup>3,7,51</sup>, Jennifer L. Gerton<sup>13,52</sup>, Rachel J. O'Neill<sup>10</sup>, Winston Timp<sup>8,41</sup>, Justin M. Zook<sup>35</sup>, Michael C. Schatz<sup>9,49</sup>, Evan E. Eichler<sup>4,53\*</sup>, Karen H. Miga<sup>11,54\*</sup>, Adam M. Phillippy<sup>1\*</sup>

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.



# LA COMPARAZIONE DEI GENOMI

I genomi degli organismi eucariotici sono di dimensioni superiori rispetto a quelli procariotici.

In generale, le specie pluricellulari possiedono più DNA rispetto ai più semplici eucarioti unicellulari.

I genomi eucariotici contengono un maggior numero di geni rispetto ai procarioti;

gli eucarioti pluricellulari ne possiedono di più rispetto agli eucarioti unicellulari.

TUTTAVIA:

tra gli organismi pluricellulari non esiste una stretta relazione tra l'ampiezza del genoma, il numero dei geni e la complessità.

# Distribuzione genomica dei geni negli eucarioti

Genomi semplici hanno generalmente:

- una densità genica alta.
  - 1/0,5-1-2 Kb in mitocondri umani, E. coli, S. cerevisiae.
- geni parzialmente sovrapposti.

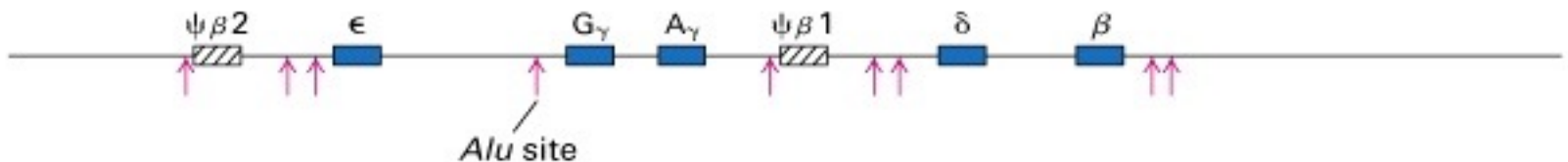
Genomi complessi hanno generalmente:

- Densità genica bassa.
- Geni molto grandi e meno clusterizzati

(a) *S. cerevisiae* (chromosome III)



(b) Human  $\beta$ -globin gene cluster (chromosome 11)



## I geni umani non sono distribuiti uniformemente nei cromosomi

- le regioni di eterocromatina costitutiva sono prive di geni
- le regioni eucromatiche presentano una diversa distribuzione genica  
urban centers vs. desert areas
- la concentrazione dei geni correla con il contenuto in GC
- cromosomi diversi presentano una differente concentrazione di geni  
crom 19 (alta) vs. crom 5 (bassa)

# La distribuzione dei geni nei cromosomi non è omogenea

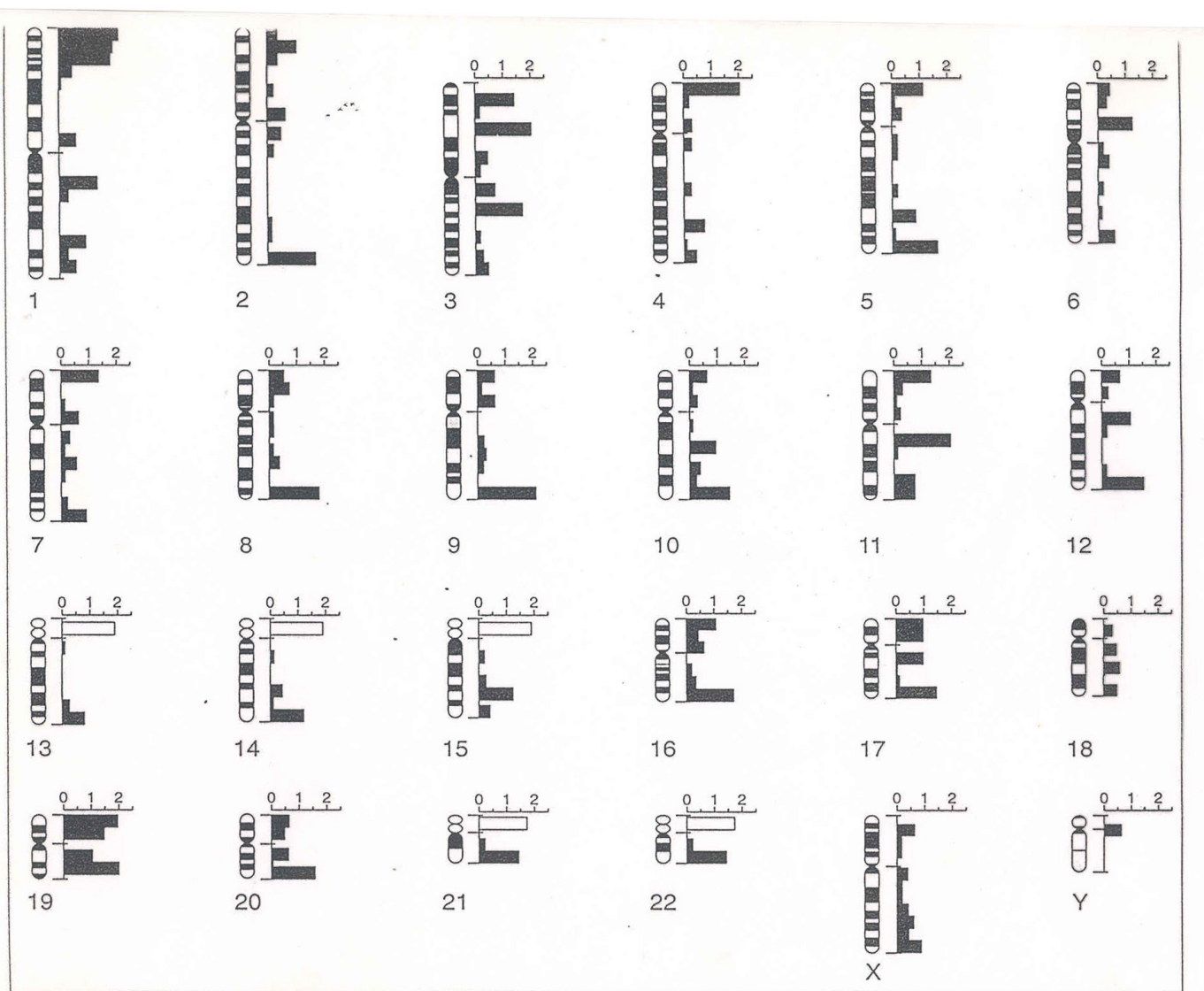


FIGURE 1. Histograms showing the distribution of H3 isochores on human G-banded chromosomes. Open bars, rRNA-encoding DNA; solid bars, non-rRNA. Scales are percentage of total number of signals. (Reprinted, with permission, from Ref. 10.)

Nelle regioni genomiche con alta densità genica si possono trovare geni overlappanti.

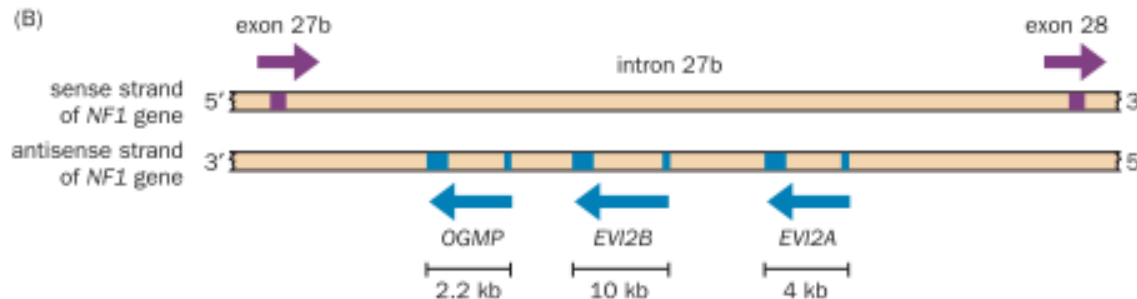
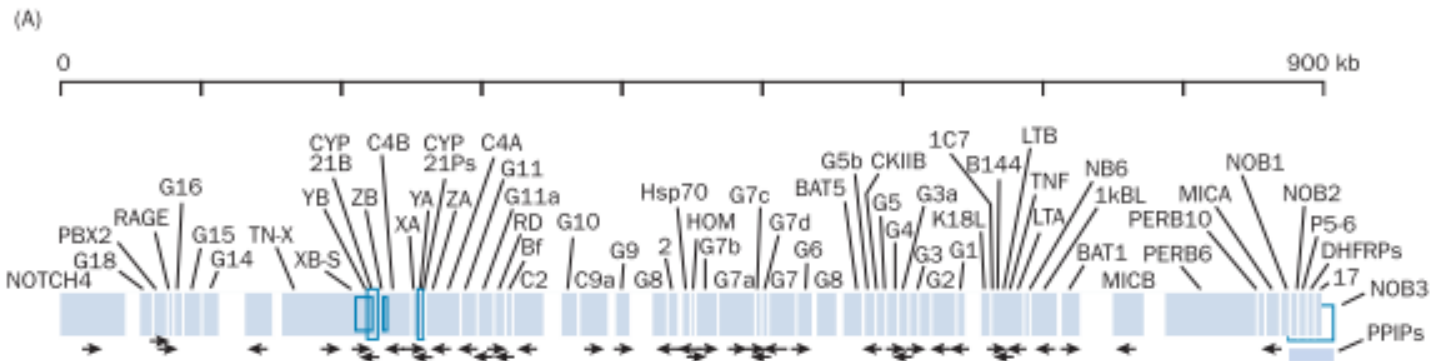
Es. la regione del complesso HLA di classe III in 6p21.3 (A).

ca. il 9% dei protein-coding genes del genoma umano sono in effetti in overlapping

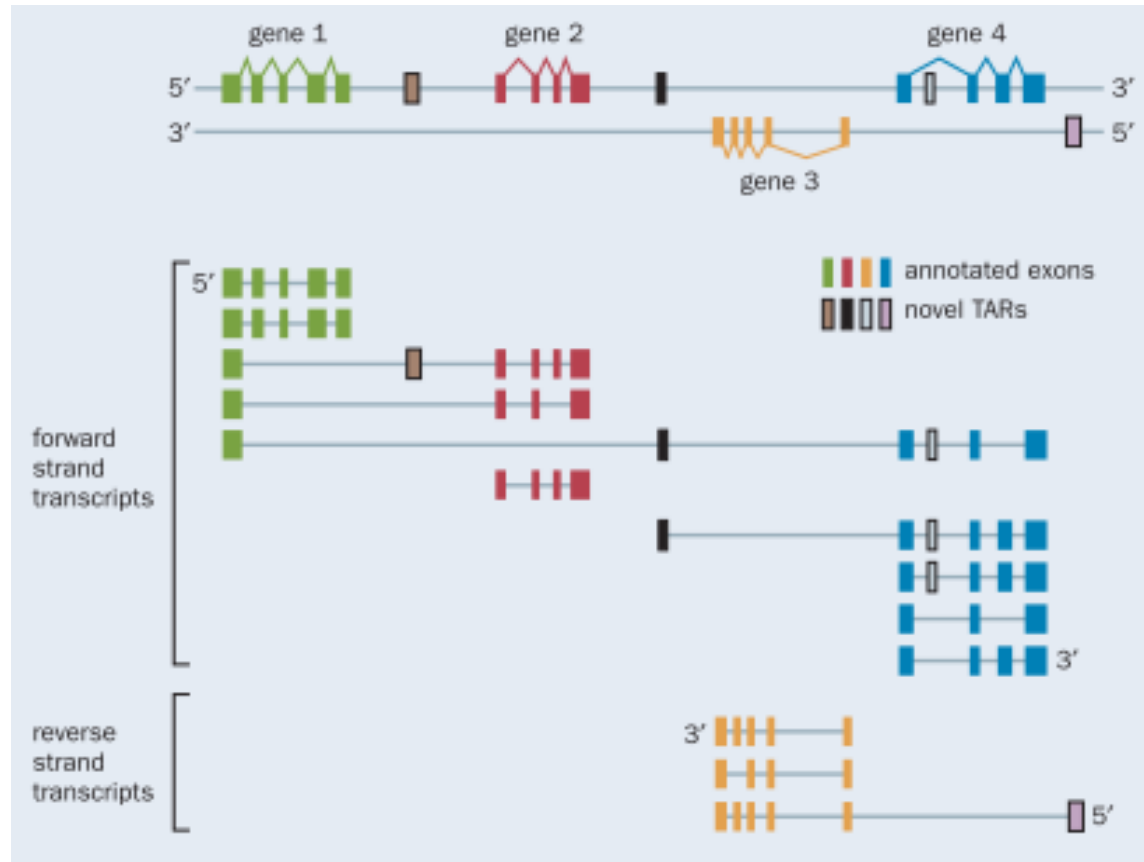
Per il 90% sono in overlapping su strand opposti;

A volte l'overlapping è parziale ma in alcuni casi è all'interno (introni) di geni più grandi (B);

RNA genes sono spesso in overlapping con protein-coding genes.



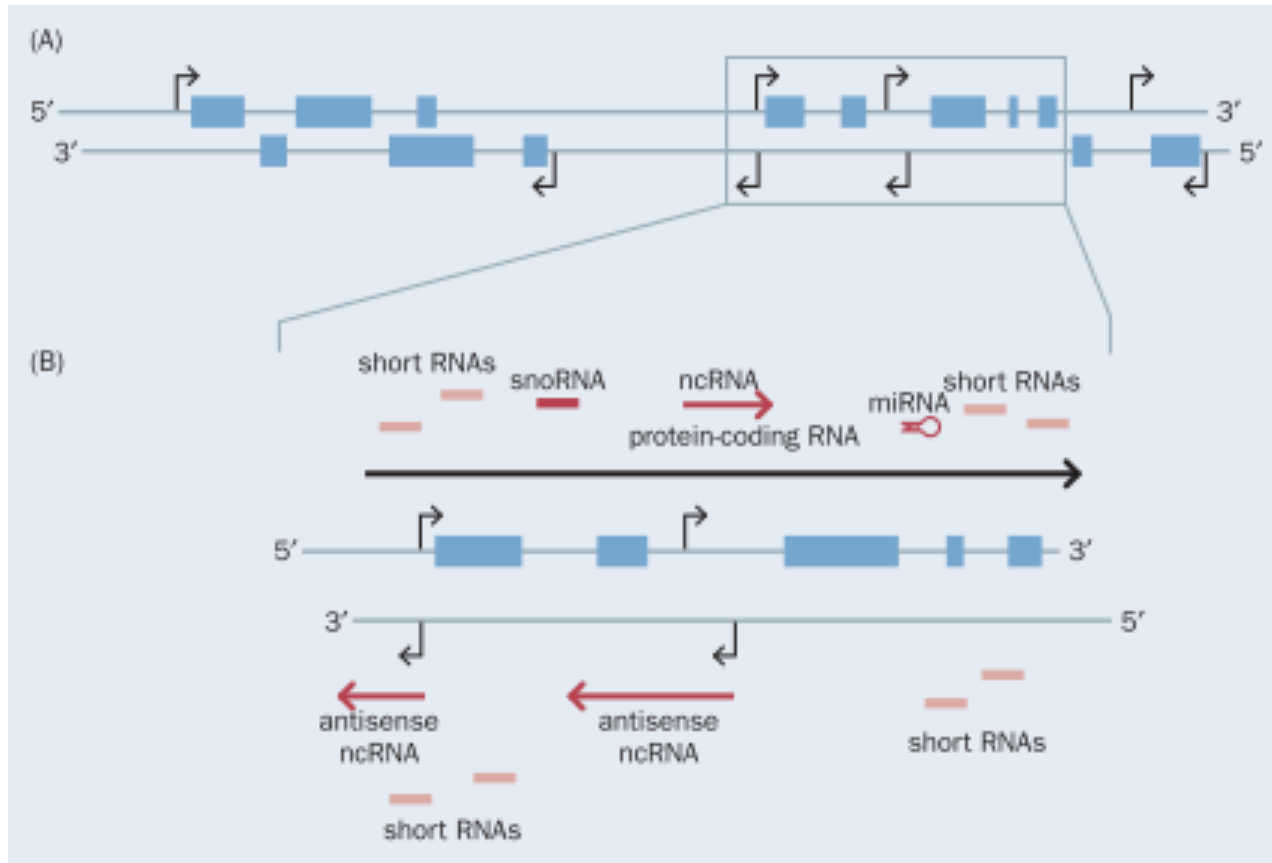
# The concept of a gene in the post-genome era



A variety of transcripts often links exons in neighboring genes. The transcripts frequently include sequences from previously unsuspected transcriptionally active regions (TARs).

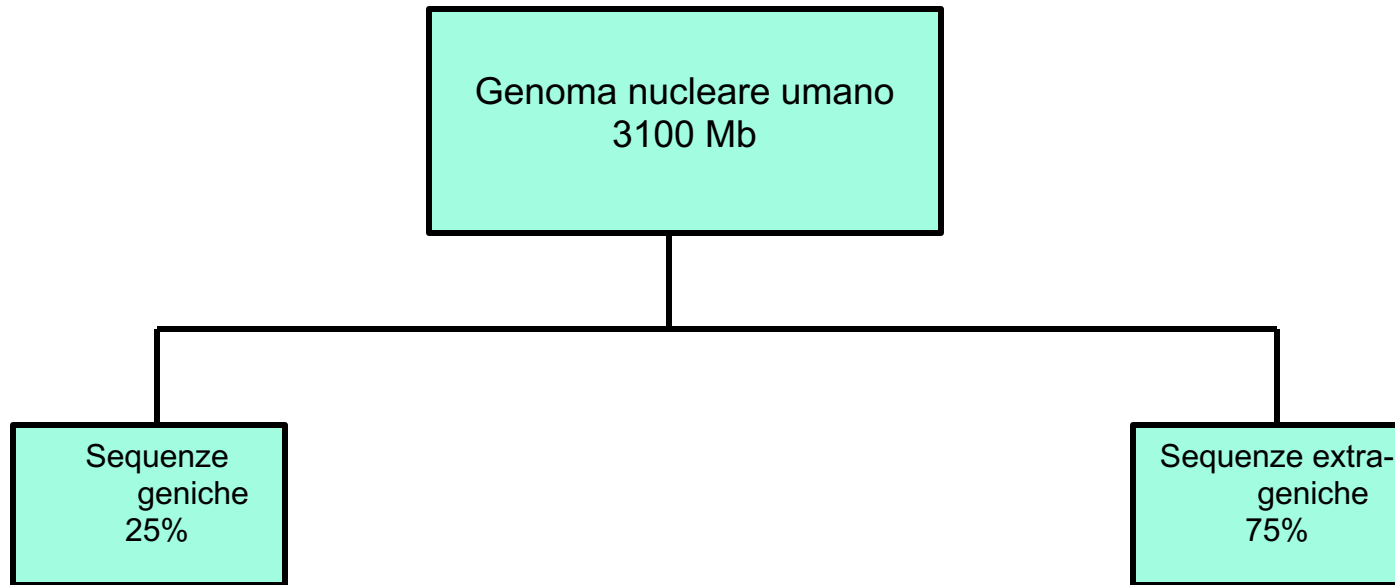


Human genes are frequently transcribed on both strands

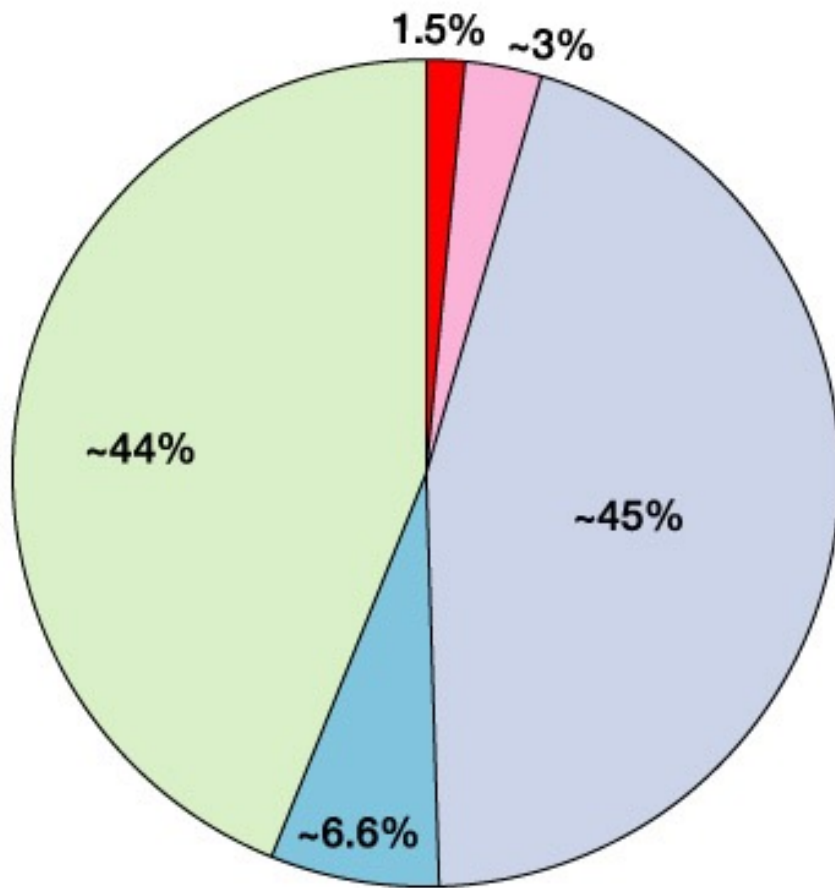


A single gene can have multiple transcriptional start sites as well as many interleaved coding and non coding transcripts

# Quali sono e come sono organizzate le sequenze nel genoma?



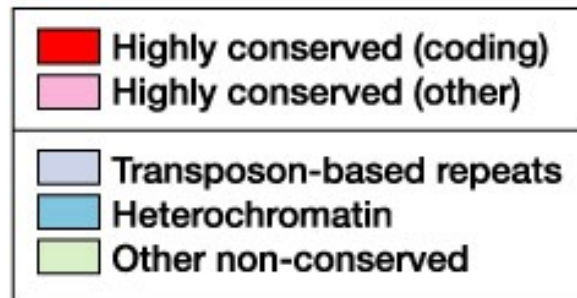
JUNK DNA?



## Il genoma umano

## I geni dell'uomo

Caratteristica	Valore medio
Numero degli esoni	8.8
Dimensione di un esone	145 pb
Dimensioni di un introne	3365 pb
Dimensioni di una regione 5' non tradotta	300 pb
Dimensioni di una regione 3' non tradotta	770 pb
Lunghezza totale di un gene	27000 pb

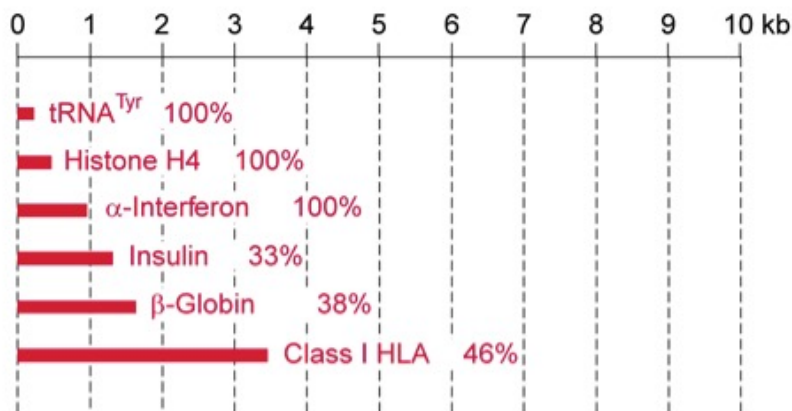


# Organizzazione dei geni

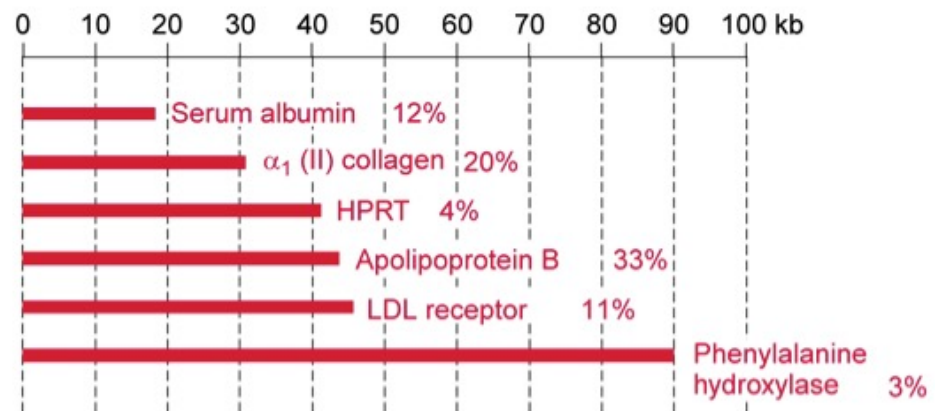
Geni che codificano per prodotti proteici  
(protein-coding genes)

# Human protein-coding genes show enormous variation in size and internal organization

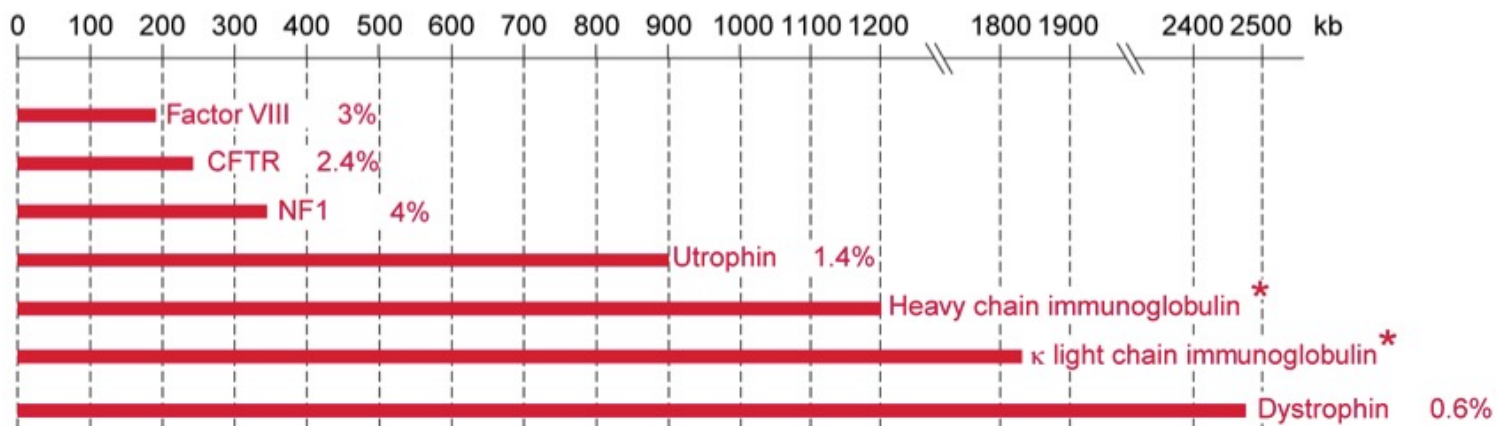
(A) Less than 10 kb



(B) Less than 100 kb



(C) More than 100 kb



Generalmente vi è una relazione diretta tra grandezza del gene e grandezza del prodotto.

Tuttavia:

Una piccola proporzione di protein-coding genes manca di introni e quindi sono generalmente piccoli;

Per quelli che li posseggono vi è una correlazione inversa tra grandezza del gene e frazione di DNA codificante

Questo non è dovuto a una minor estensione degli esoni, bensì ad una maggior estensione degli introni.

Geni molto espressi hanno generalmente introni corti o mancano completamente di introni

La trascrizione degli introni costa tempo ed energia! La trascrizione delle 2.4 Mb del gene della distrofina avviene in circa 16 ore.

**TABLE 9.4 STRUCTURAL VARIATION IN SIZE AND ORGANIZATION OF HUMAN PROTEIN-CODING GENES**

Human protein	Size of protein (no. of amino acids)	Size of gene (kb)	No. of exons	Coding DNA (%)	Average size of exon (bp)	Average size of intron (bp)
SRY	204	0.9	1	94	850	–
β-Globin	146	1.6	3	38	150	490
p16	156	7.4	3	17	406	3064
Serum albumin	609	18	14	12	137	1100
Type VII collagen	2928	31	118	29	77	190
p53	393	39	10	6.0	236	3076
Complement C3	1641	41	29	8.6	122	900
Apolipoprotein B	4563	45	29	31	487	1103
Phenylalanine hydroxylase	452	90	26	3	96	3500
Factor VIII	2351	186	26	3	375	7100
Huntingtin	3144	189	67	8.0	201	2361
RB1 retinoblastoma protein	928	198	27	2.4	179	6668
CFTR (cystic fibrosis transmembrane receptor)	1480	250	27	2.4	227	9100
Titin	34,350	283	363	40	315	466
Utrophin	3433	567	74	2.2	168	7464
Dystrophin	3685	2400	79	0.6	180	30,770

# I protein-coding genes nell'uomo

<b>Caratteristica</b>	
<b>Numero medio di esoni per gene</b>	<b>9.8</b>
<b>Dimensione di un esone</b> Gli esoni alle estremità 3' tendono ad essere più grandi	<b>288 pb</b>
<b>Esone più piccolo</b>	<b>4 pb</b>
<b>Esone più grande</b>	<b>18200 pb</b>
<b>Dimensioni introne più piccolo</b>	<b>&lt;30 pb</b>
<b>Dimensioni introne più grande</b>	<b>1.1 Mb</b>
<b>Dimensioni di una regione 5' non tradotta</b>	<b>300 pb</b>
<b>Dimensioni di una regione 3' non tradotta</b>	<b>770 pb</b>
<b>Lunghezza media di un gene</b>	<b>53600 pb</b>



Una certa % dei geni umani è costituita da membri di famiglie di sequenze di DNA.

Ovvero geni che condividono tra loro un buon grado di omologia

A seconda del grado di omologia tra i diversi geni di una stessa famiglia, si distinguono:

**FAM DI GENI RIPETUTI**

ORGANIZZAZIONE IN TANDEM,  
IN UNO O PIU' CLUSTERS

(i componenti sono fondamentalmente identici tra loro. es. rRNA e Istoni)

**FAM GENICHE CLASSICHE**

(i componenti mostrano una certa divergenza tra loro)

ORGANIZZAZIONE IN CLUSTERS  
(es. GLOBINE E HLA)

ORGANIZZAZIONE INTERSPERSA  
(es. actina)

**SUPERFAMIGLIE GENICHE**

(geni che codificano prodotti funzionalmente correlati, ma con minima omologia di sequenza)

ORGANIZZAZIONE INTERSPERSA

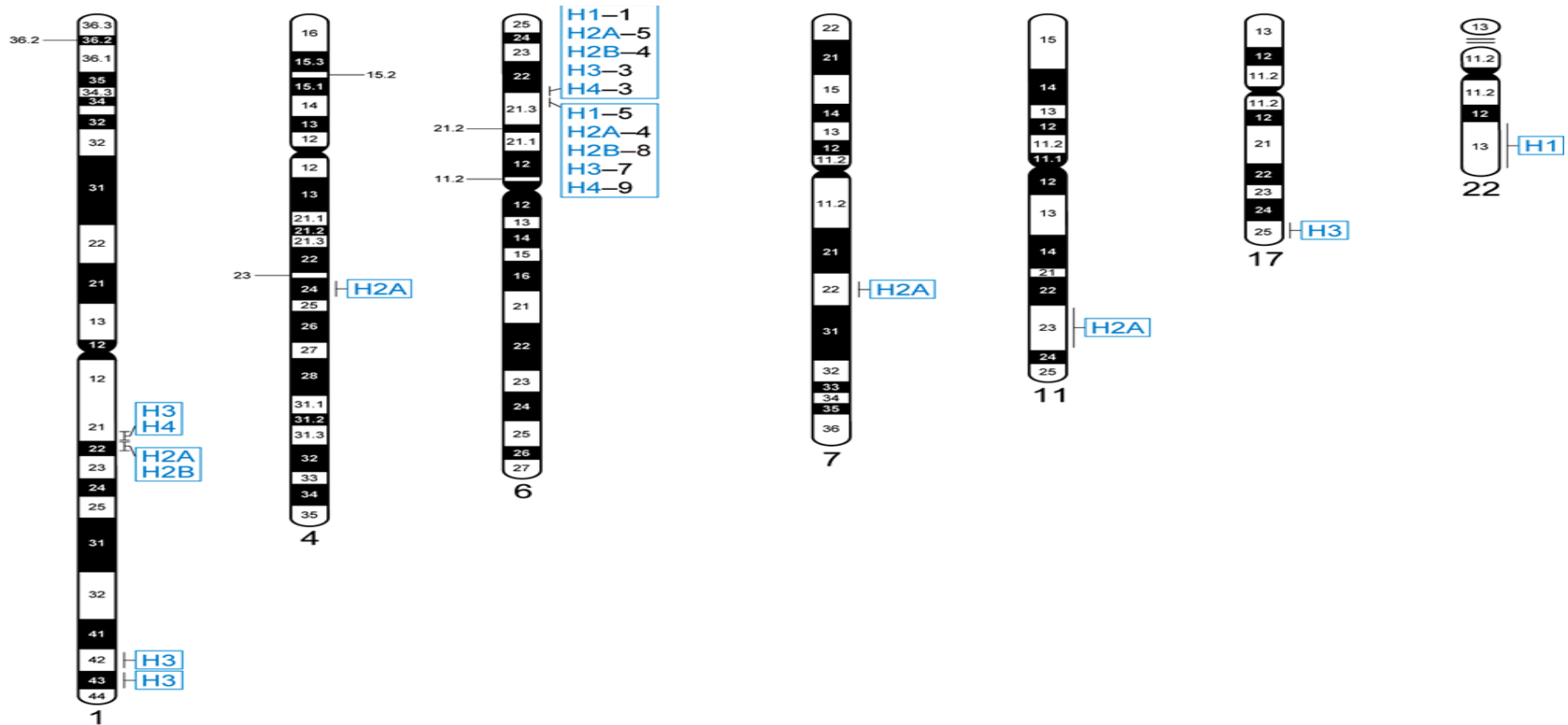
# Perché esistono le famiglie geniche?

- 1-Richiesta di grandi quantità di prodotto.
- 2-Funzioni specializzate.
- 3-Espressione differenziale.

**TABLE 9-2 Protein Families in Vertebrates and Invertebrates**

Family	Number of Proteins in Family
<b>COMMON PROTEINS</b>	
Actins	5-30
70 K Heat-shock proteins	3
Keratins	>20
Myosin, heavy chain	5-10
Protein kinases	10-100s
Transcription factors	10-100s (?)
Tubulins, $\alpha$ and $\beta$	3-15
<b>INSECT PROTEINS</b>	
Eggshell proteins (silk moth and fruit fly)	50
<b>VERTEBRATE PROTEINS</b>	
Globins (many species)	
$\alpha$ -globin	1-3
$\beta$ -like globins	5
Immunoglobins, variable regions (many species)	500
Ovalbumin (chicken)	3
Transplantation antigens (mouse and human)	50-100
Visual pigment protein (human)	4
Vitellogenin (frog, chicken)	5

# Richiesta di grandi quantità di prodotto



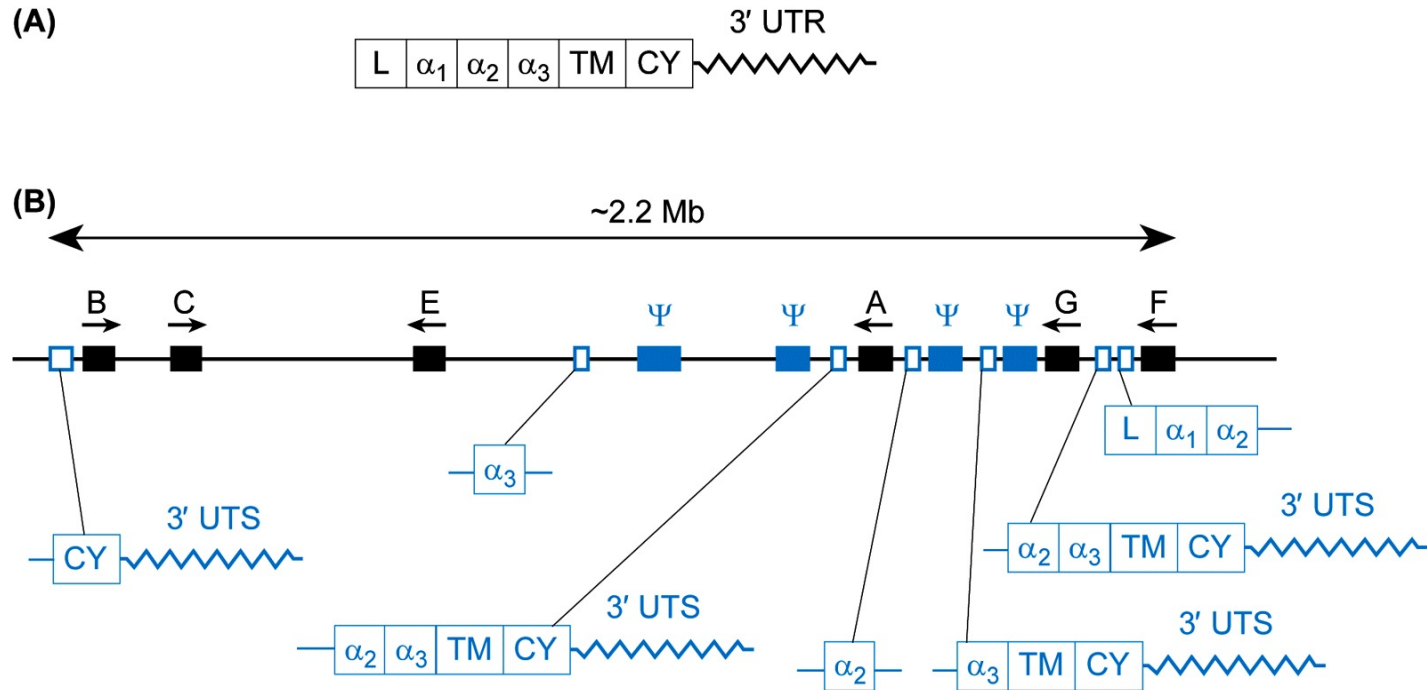
## La famiglia genica degli istoni

La maggior parte degli 11 clusters dei geni per gli istoni (in totale circa 60 geni) sono raggruppati in 6p

# Richiesta di funzioni specializzate

Famiglie geniche classiche:

i geni per la catena pesante HLA (MHC dell'uomo) di classe I

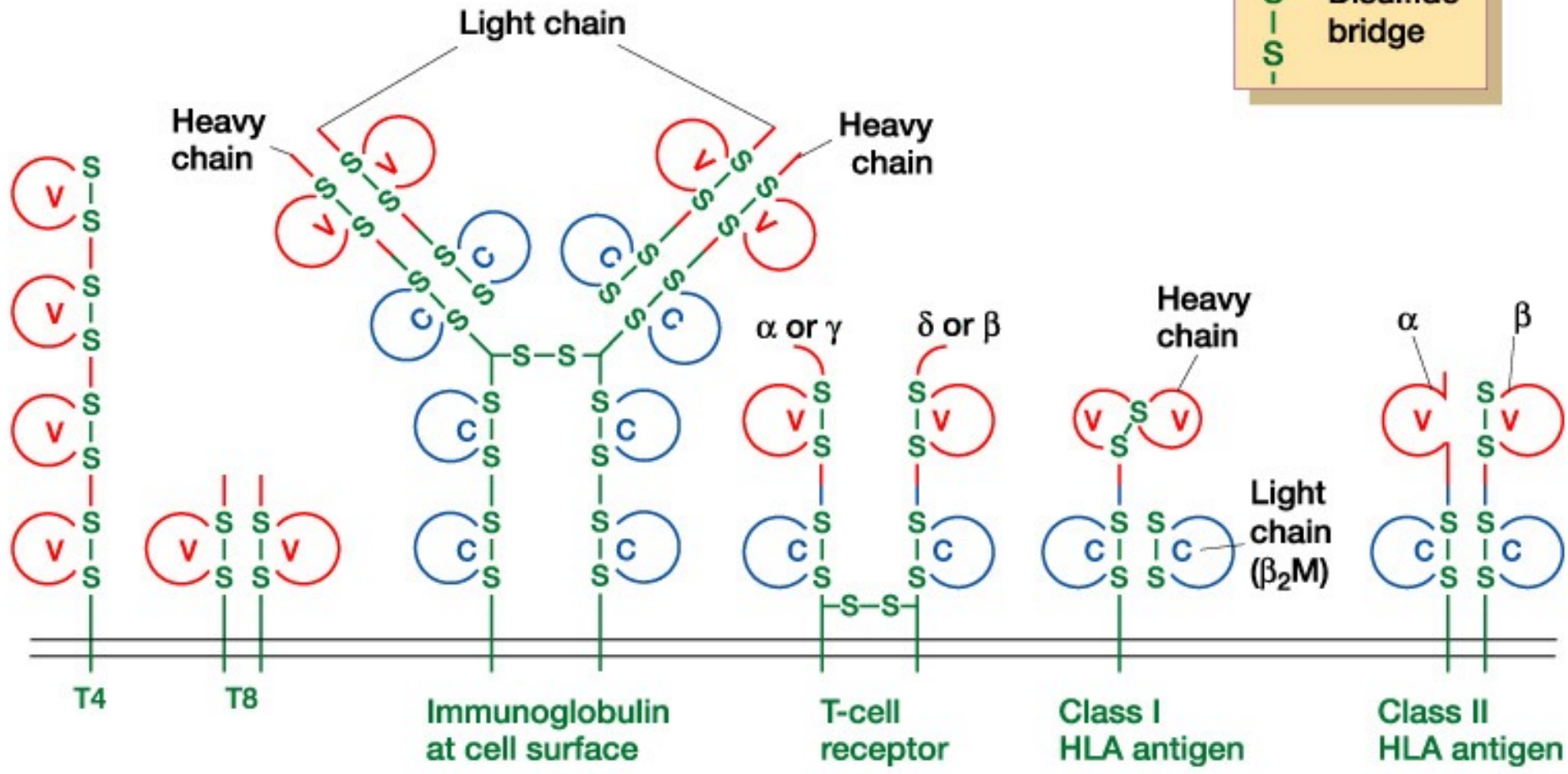


I singoli membri presentano un elevato grado di omologia di sequenza per tutta la lunghezza del gene o almeno per la parte codificante.

Le famiglie geniche contengono spesso pseudogeni non processati o frammenti genici.

**Superfamiglia geniche:** i geni codificano prodotti funzionalmente correlati ma che non mostrano elevati gradi di omologia di sequenza né motivi amminoacidici particolarmente conservati

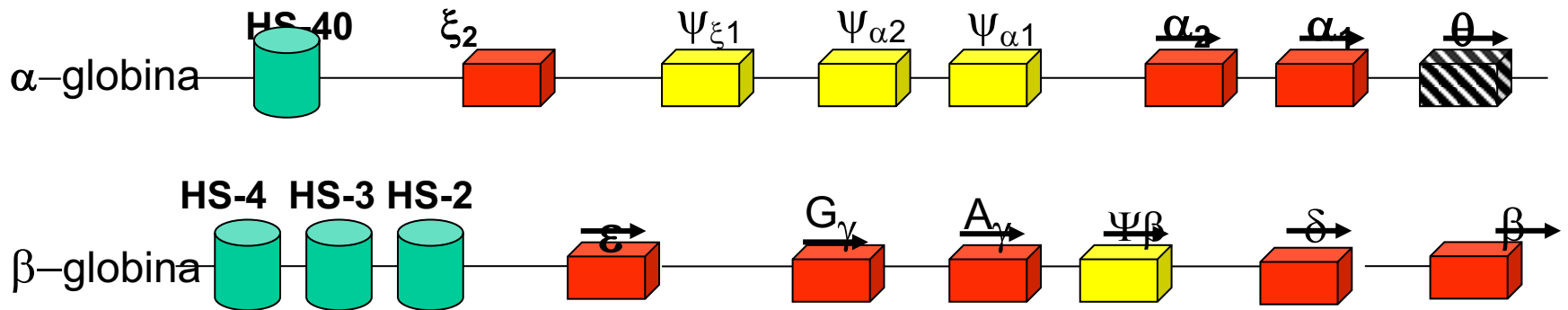
**Key:**  
 Disulfide bridge



I membri della superfamiglia delle Ig sono proteine di superficie con strutture e domini simili tra loro

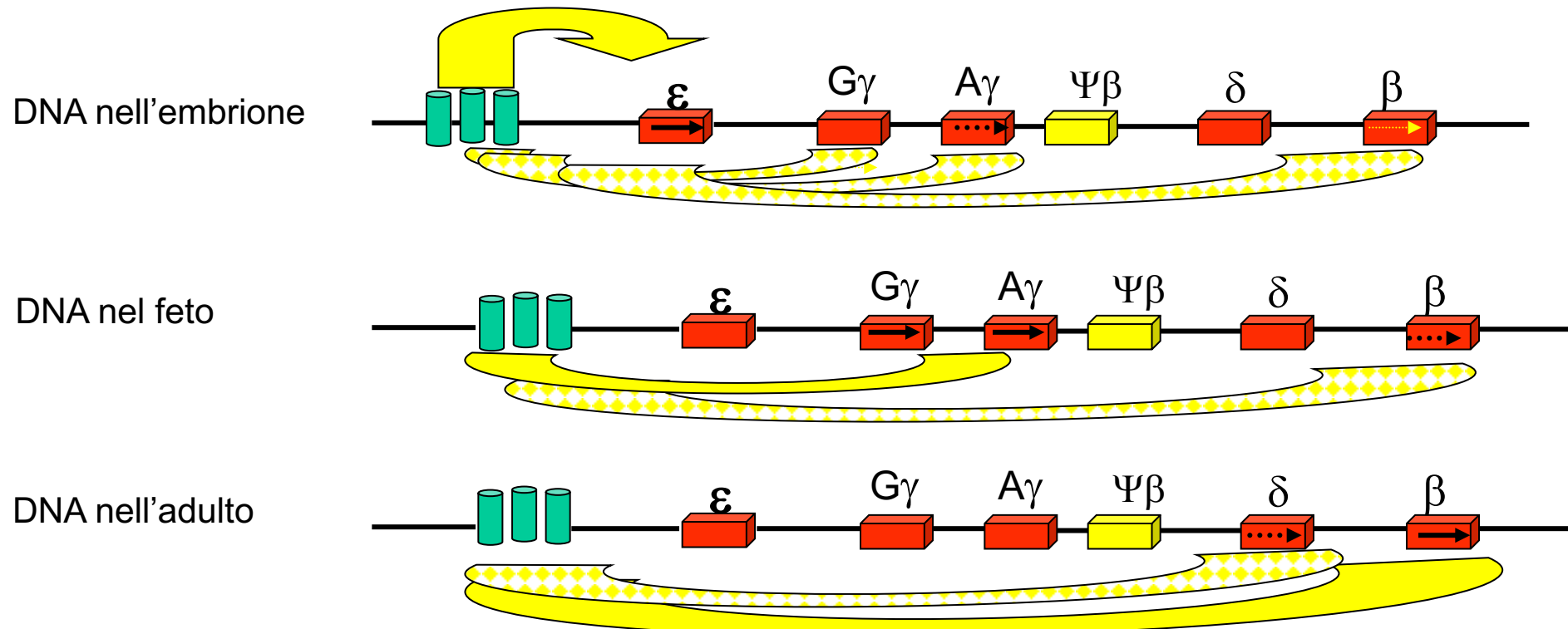
# Espressione differenziale

Famiglie geniche classiche:  
raggruppamento  $\alpha$  e  $\beta$  globinico



## ORGANIZZAZIONE IN CLUSTER -> CONTROLLO FUNZIONALE:

i singoli geni sono fisicamente distanziati tra loro, ma strettamente raggruppati; questo permette di essere soggetti a un meccanismo di regolazione comune. Le LCR sono brevi sequenze *enhancer* che agiscono *in cis*, che vengono riconosciute da **fattori di trascrizione eritroide-specifici**. Si ritiene che l'alternanza delle emoglobine sia legata oltre che a fenomeni di competizione dei geni per interagire con le LCR, anche all'intervento di silenziatori gene-specifici, modulati durante lo sviluppo.



# Famiglie geniche classiche con organizzazione interspersa:

Non vi è una relazione fisica tra i membri di una famiglia. Possono derivare da eventi di duplicazione genica oppure da eventi di trasposizione.

famiglia	n° copie	caratteristiche
aldolasi	5	3 geni funzionali e 2 pseudogeni, su 5 cromosomi diversi
NF1	>12	1 gene funzionale (17q), copie difettose, non processate
gliceraldeide 3-fosfato-deidrogenasi	>18	1 gene funzionale
actina	>20	4 geni funzionali
PAX	9	tutti geni funzionali

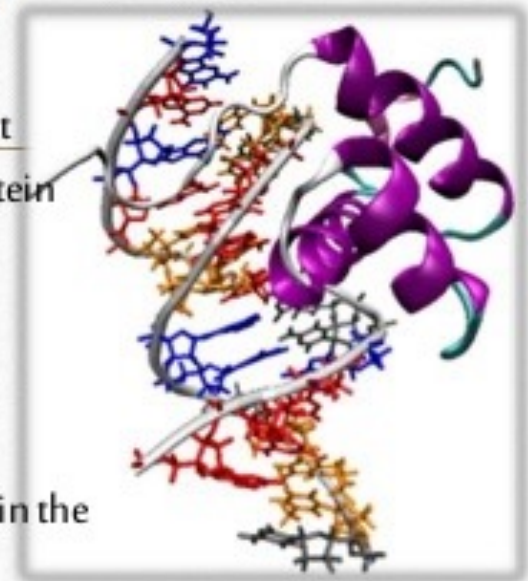


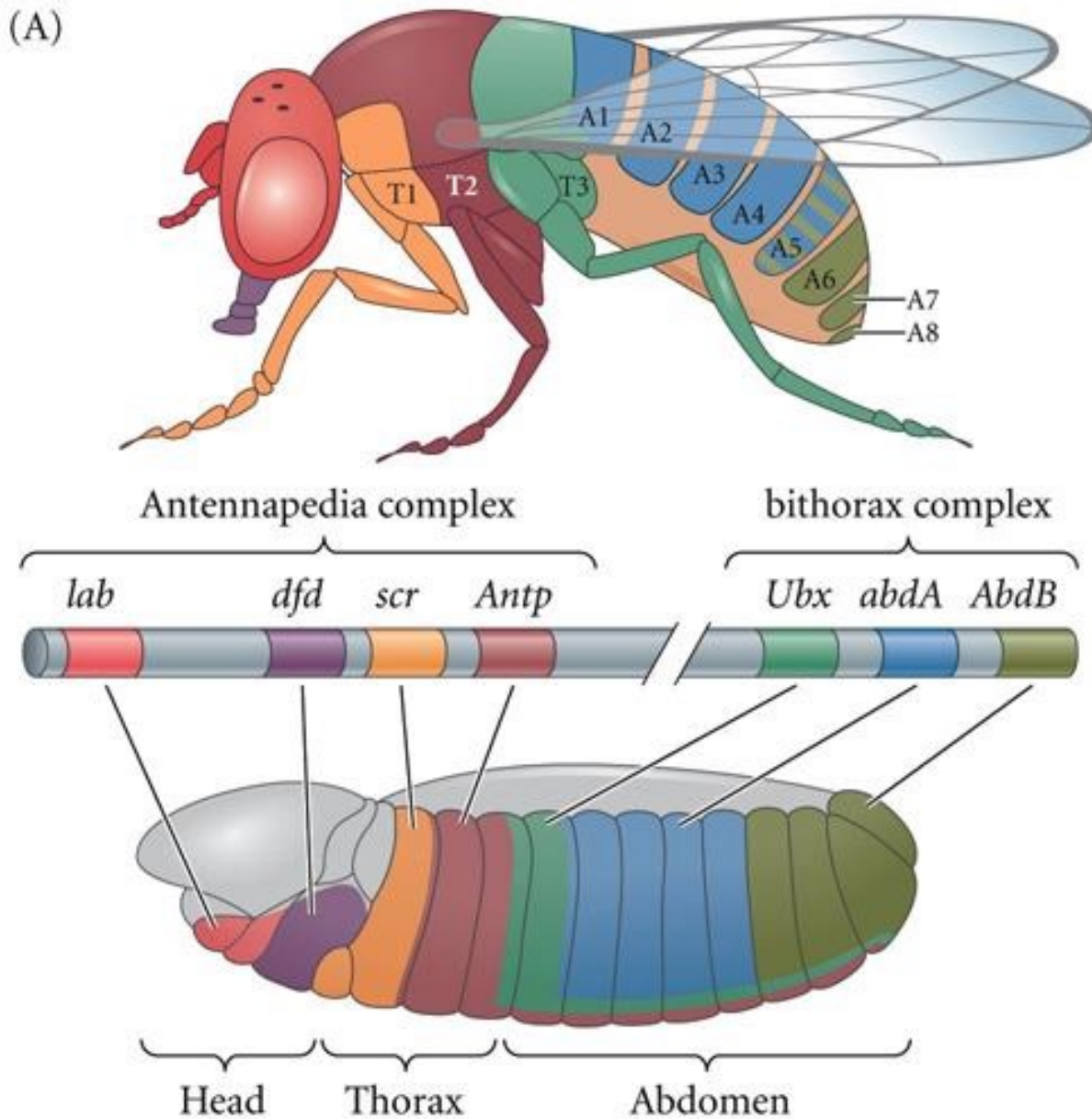
## Homeodomain

• Homeobox genes contain a 180 base pairs DNA sequence that provides instructions for making a string of 60 amino acids protein building blocks known as the *homeodomain*.

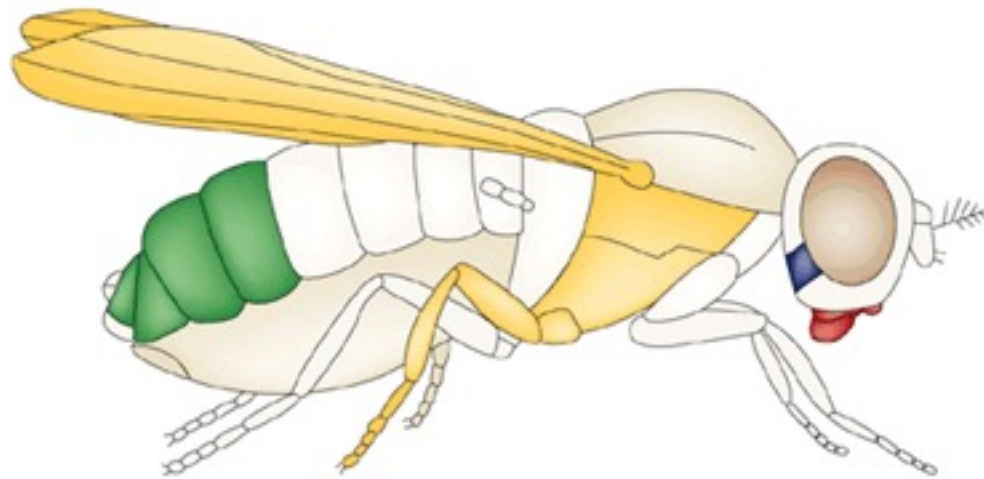
• Homeodomain act as transcription factors,

- Bind to DNA and controls transcription of other genes in the cell
- Initiate patterns of gene expression

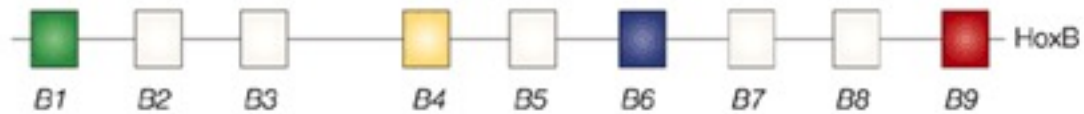




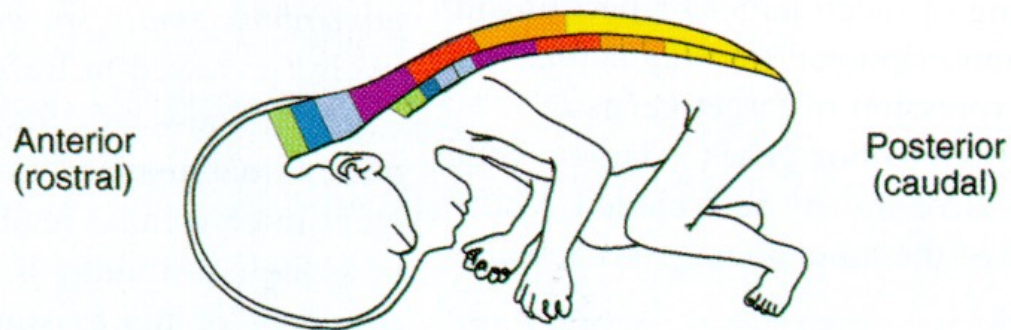
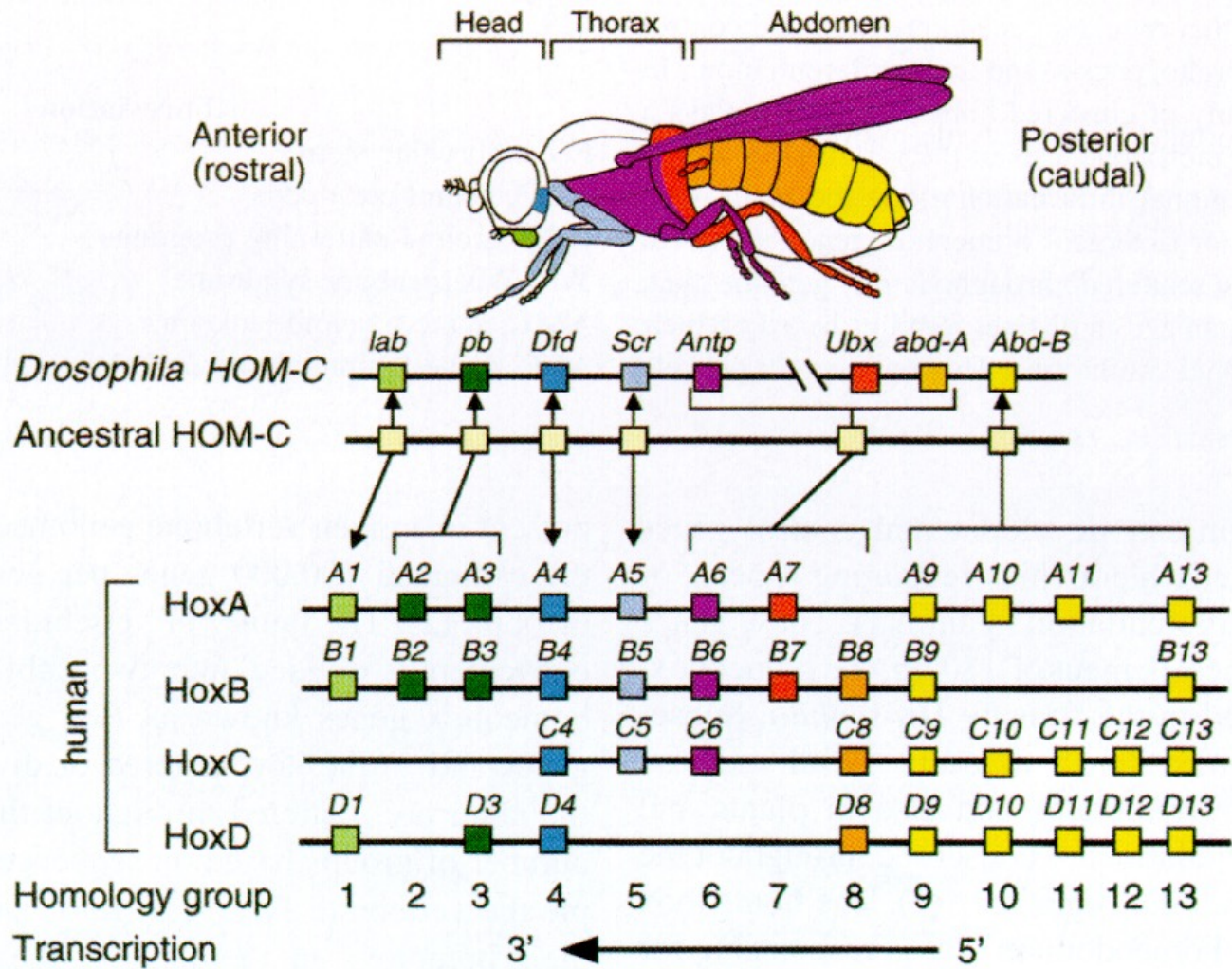
**Drosophila**

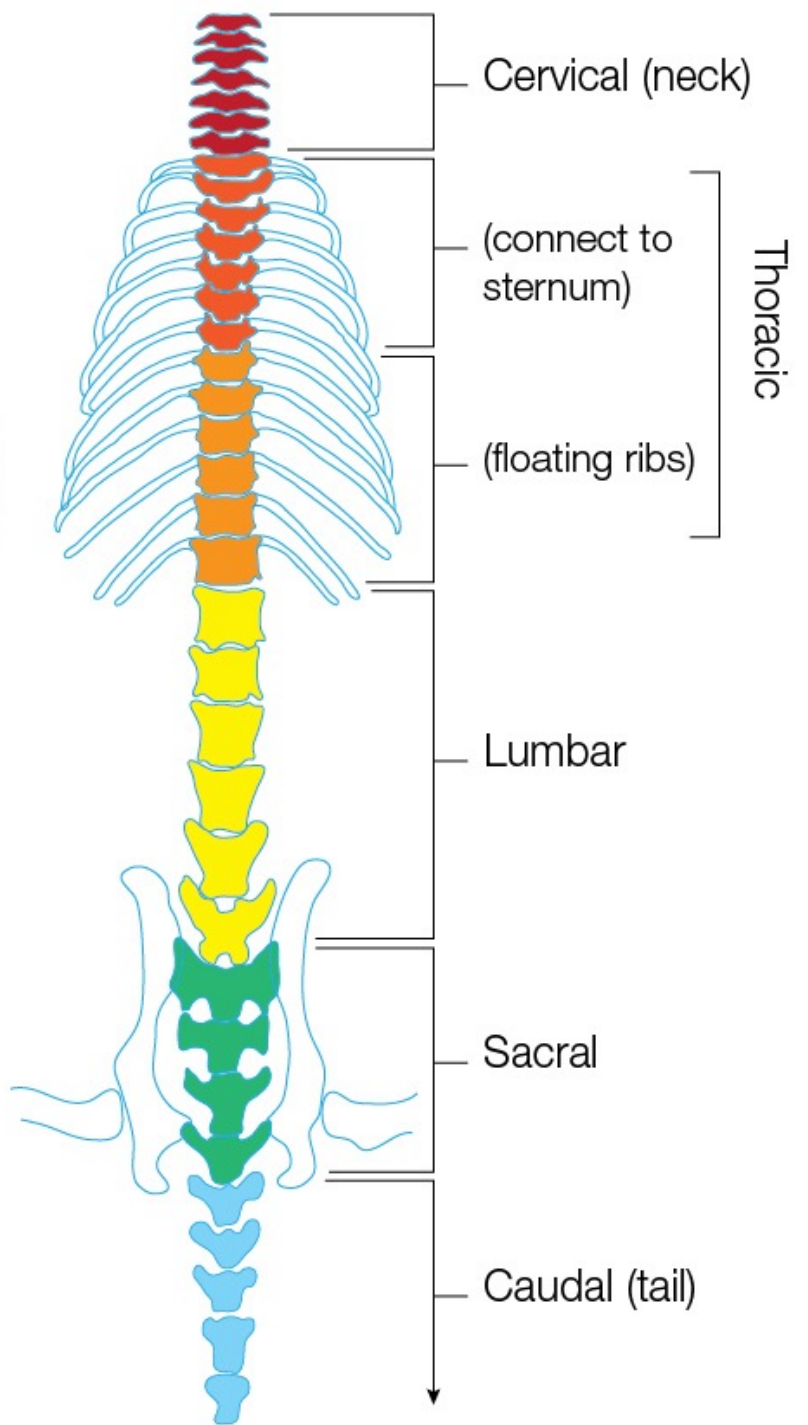
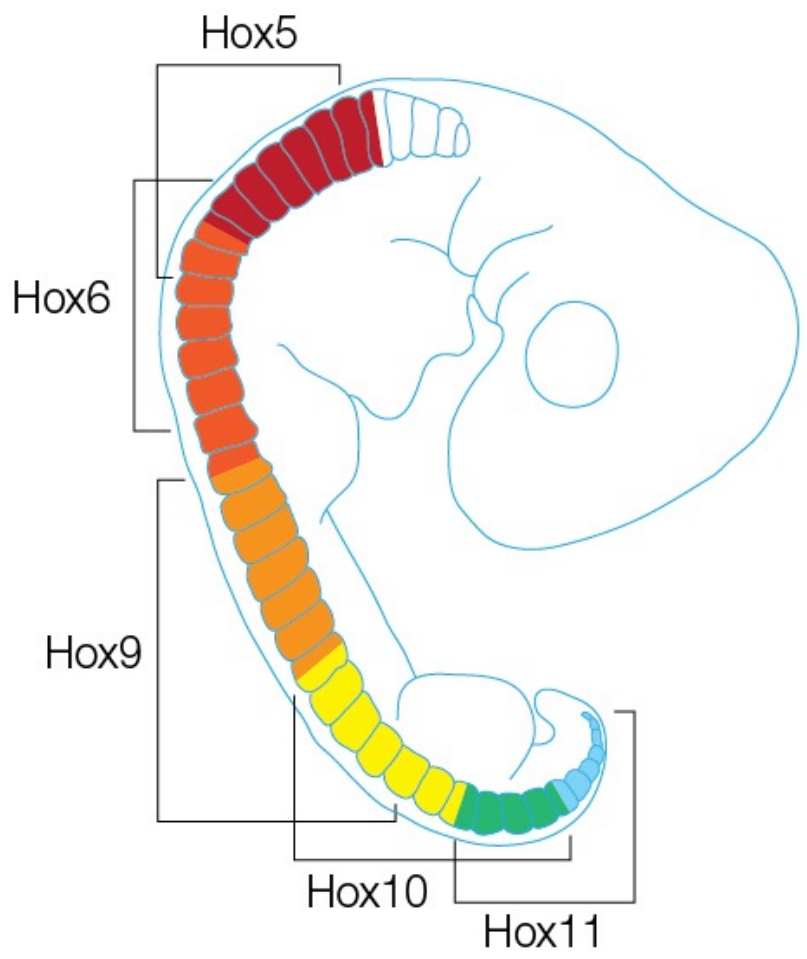


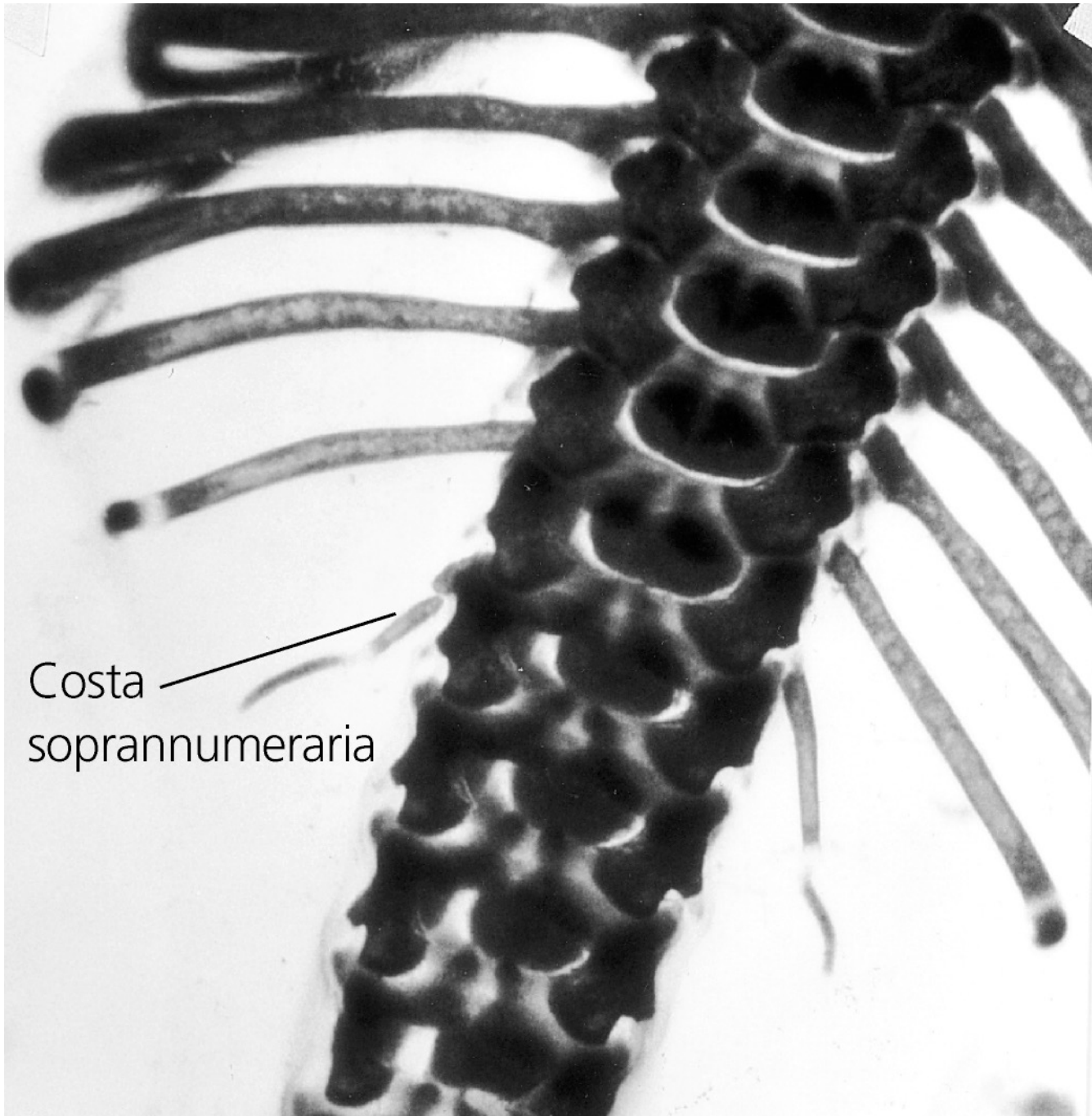
**Mouse embryo**











Costa  
soprannumeraria

# Cosa genera le famiglie geniche

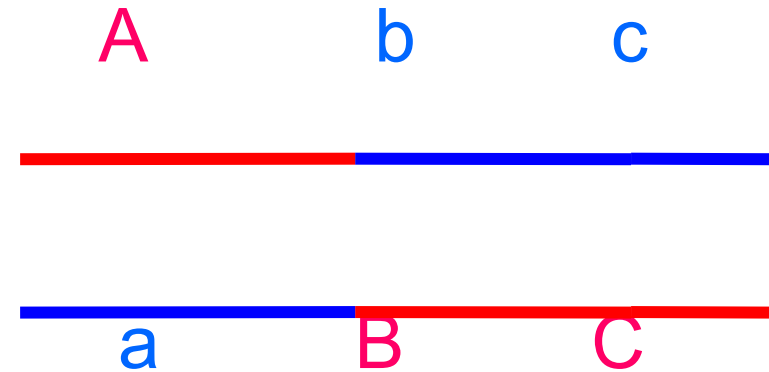
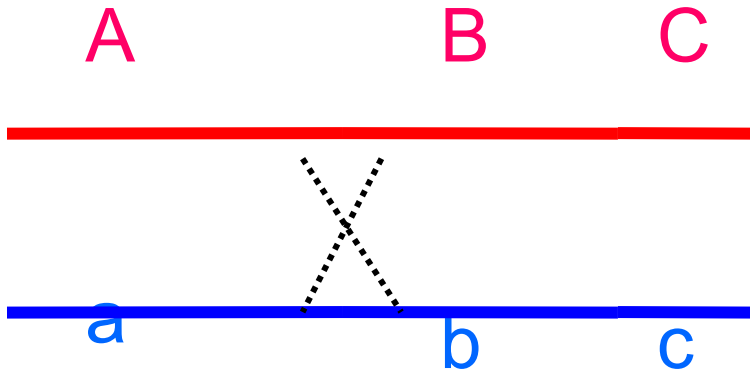
- Duplicazione genica e mutazioni puntiformi and point mutation
  - La duplicazione di geni aumenta la quantità di materiale genico su cui l'evoluzione puo' agire ed e' considerata molto importante per l'evoluzione biologica.
- Natural selection and random drift

# Cosa genera le famiglie geniche

- Crossing over ineguale e conversione genica
  - **Crossing over ineguale** : e' un evento di crossing-over che causa una delezione o duplicazione di una regione di DNA in seguito a spostamento dell'appaiamento cromosomico.
  - Il disappaiamento e scambio di DNA tra regioni cromosomiche geneticamente simili ma non omologhe causa la duplicazione o delezione di materiale genetico nelle cellule figlie.
  - Questo scambio e' favorito in regioni contenenti regioni ripetute in tandem e genera coromosomi di lunghezza ineguale.
  - **Gene conversion**: evento in cui una regione di DNA diventa identica a quella di una regione omologa mediante appaiamento cromosomico.



# Crossing Over of Chromosomes in Meiosis



# Unequal Crossing-Over

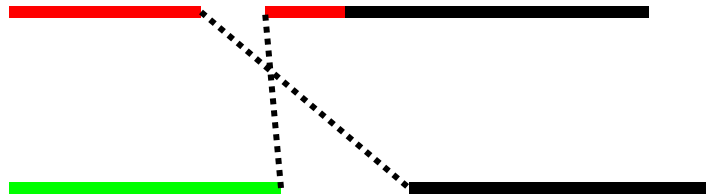
It's not likely that an existing gene can adopt a new function while continuing to serve its old one. Fundamentally the source of new genes is the *duplication* of existing genes and their *divergence* in function.



Homologous  
chromosome lineup



Chromosome Breaks



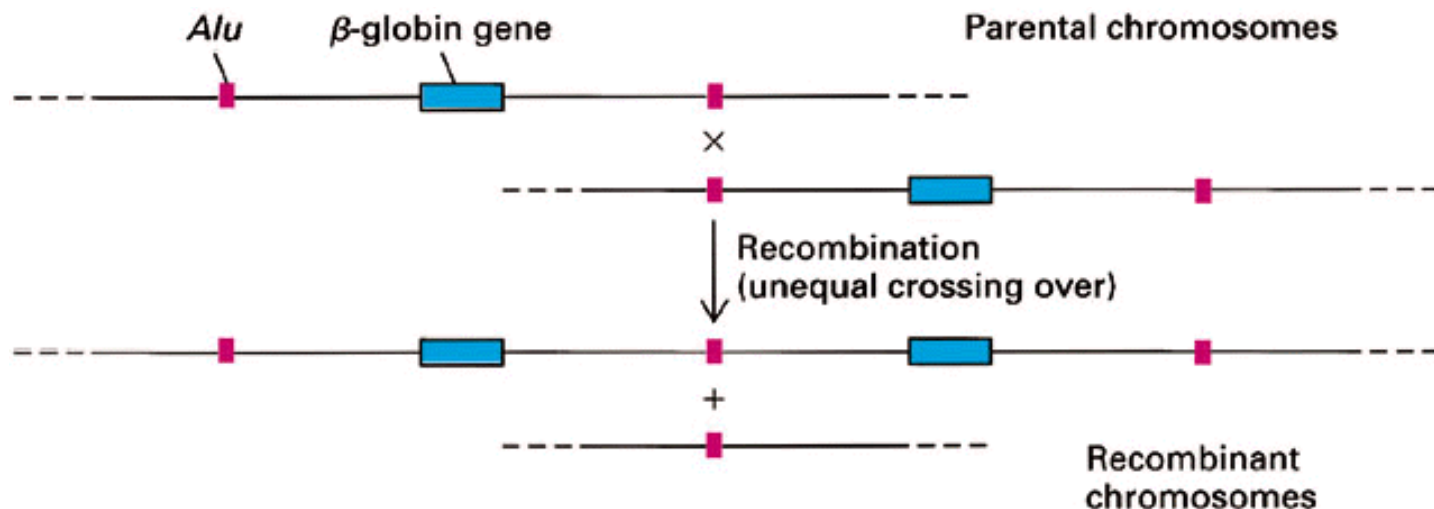
Realignment



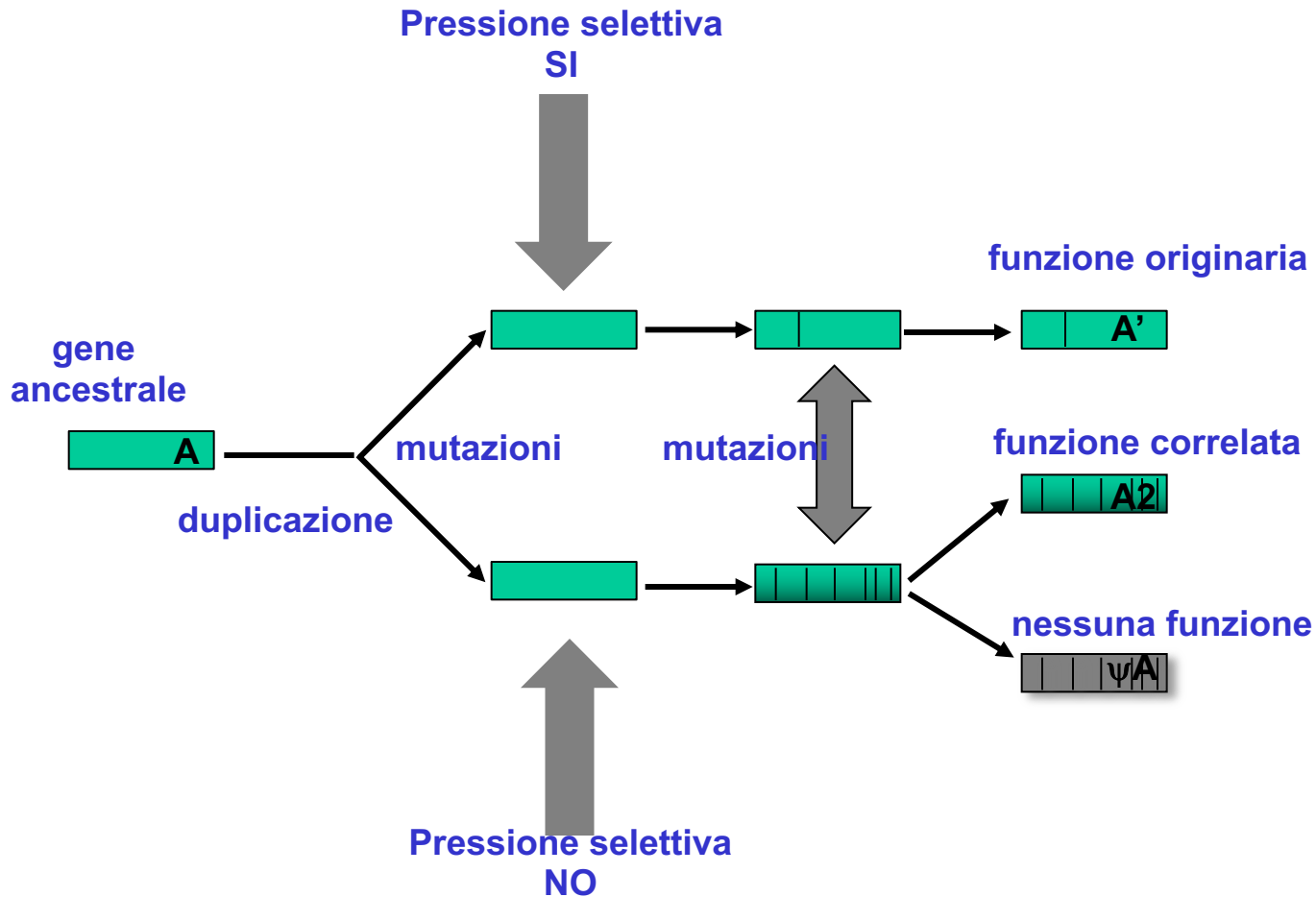
One truncate and one  
elongated chromosome

# Come avviene il crossing-over ineguale?

- Un Crossing-over ineguale durante la meiosi genera un cromosoma con due copie di un gene, e un altro con nessuna copia.
- La ricombinazione può avvenire tra corte sequenze di DNA ripetuto (elementi Alu)



# Duplicazione genica e: mantenimento della stessa funzione, acquisizione di nuove funzioni o perdita funzionale



# Pseudogeni I

- **Pseudogeni non processati: convenzionali ed espressi**

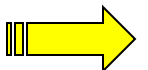
☞ Copie non funzionali del DNA genomico di un gene.

Contengono esoni, introni e spesso le sequenze fiancheggianti.

Data la loro somiglianza nell'organizzazione genomica, la loro natura non funzionale può essere riconosciuta, a livello di sequenza, dalla presenza di codoni di stop nella regione corrispondente alla porzione codificante del gene funzionale o dalla presenza di un'elevato numero di mutazioni ognuna delle quali originerebbe una molecola mutante.

Sono comuni nelle famiglie di geni raggruppati

■ **Talvolta possono venir espressi a livello di RNA o addirittura come polipeptide, che non viene utilizzato nella molecola funzionale: gene della globina  $\theta$ , sicuramente viene espresso, ma non se ne riscontra la presenza nell'emoglobina funzionale**



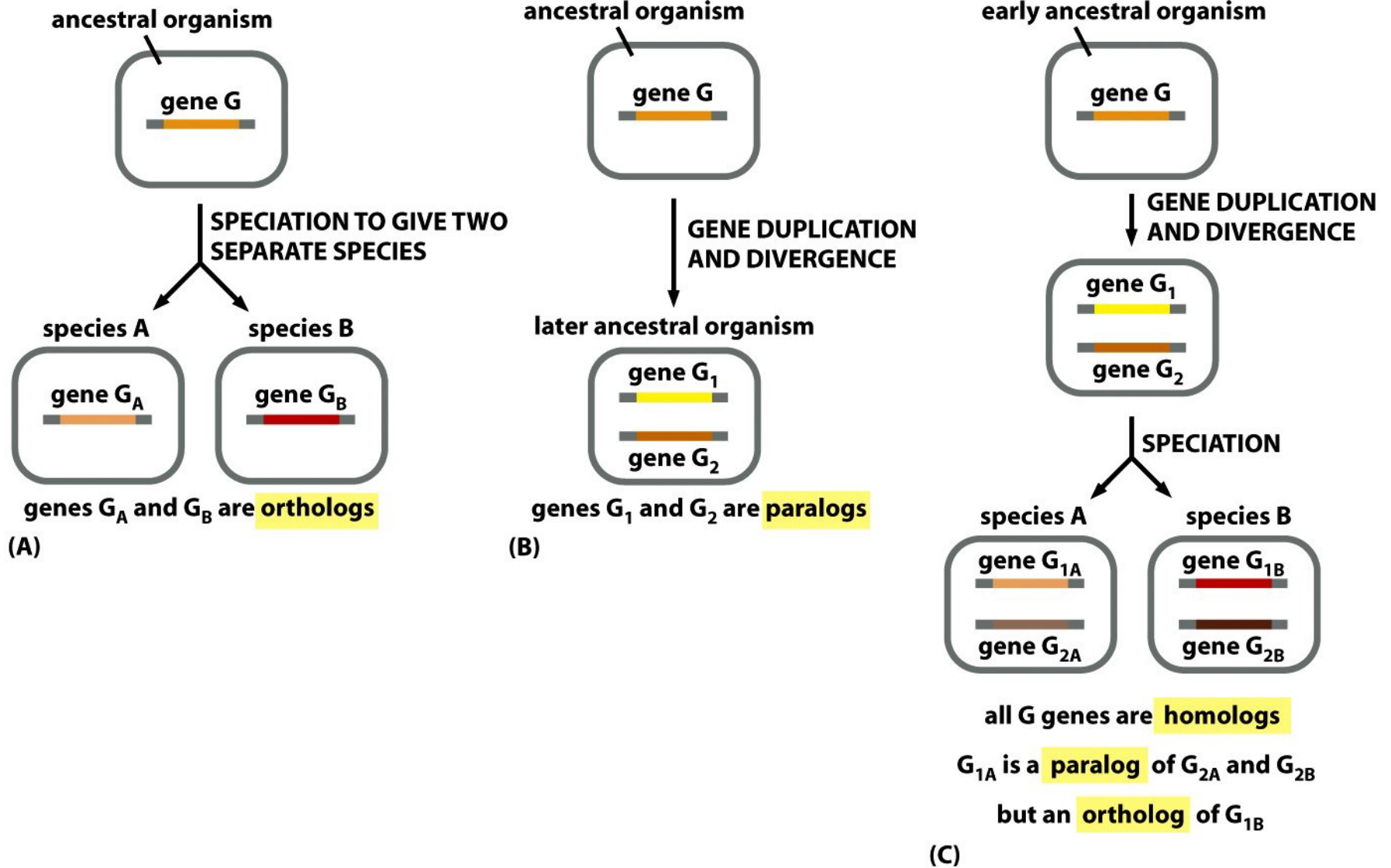


Figure 1-25 *Molecular Biology of the Cell*, Fifth Edition (© Garland Science 2008)

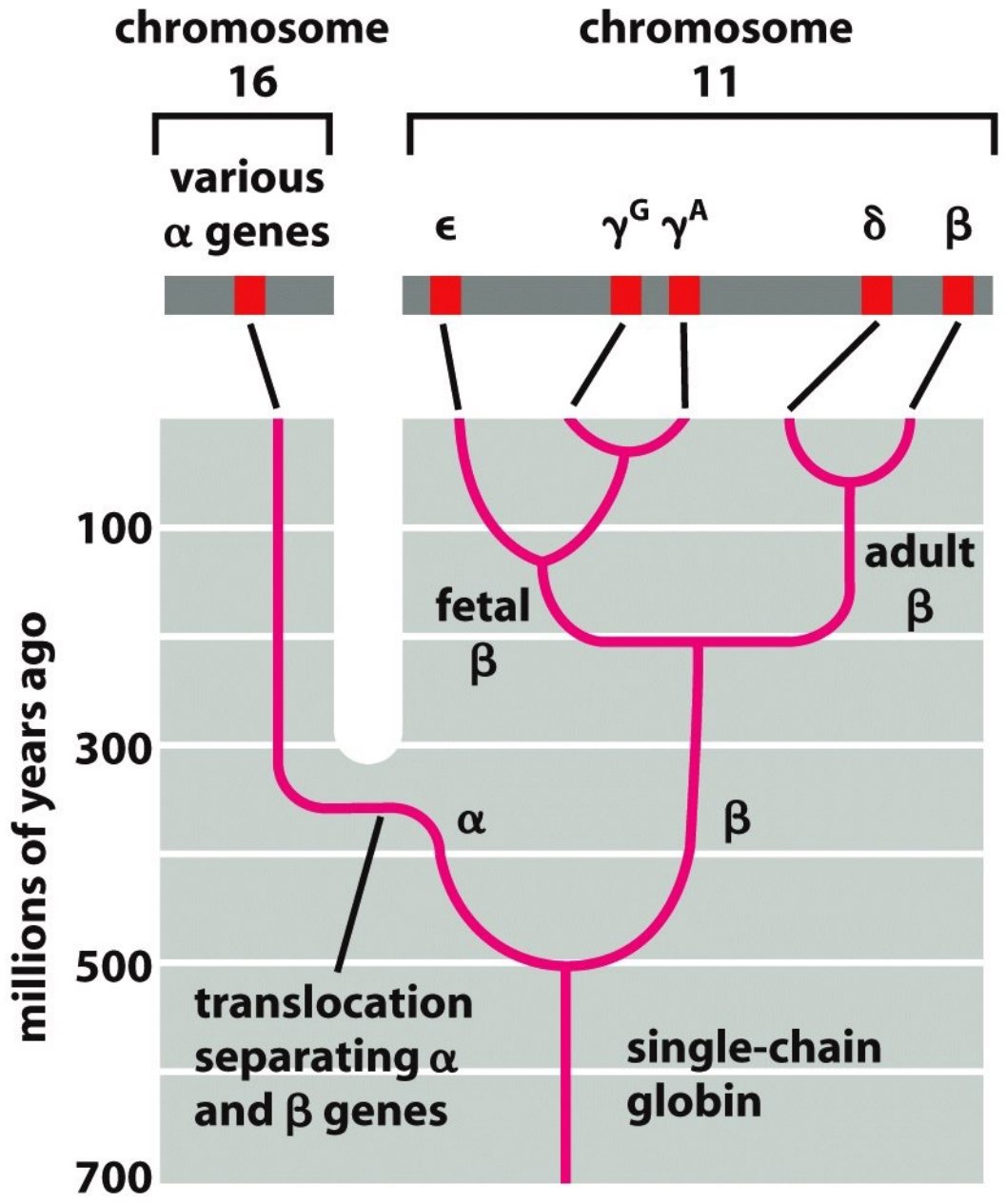
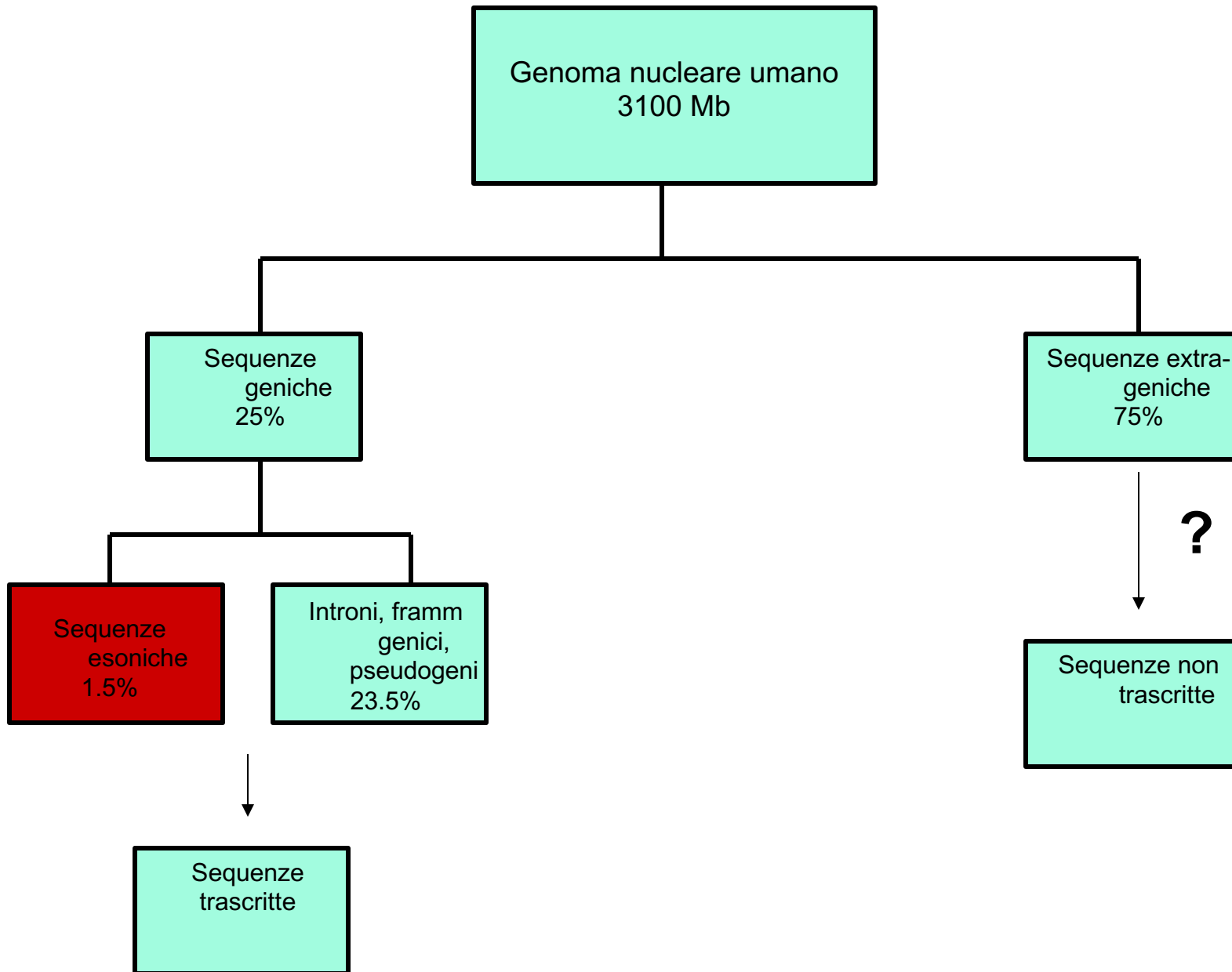


Figure 4-87 *Molecular Biology of the Cell* (© Garland Science 2008)

# Quindi, considerando l'organizzazione dei geni ....





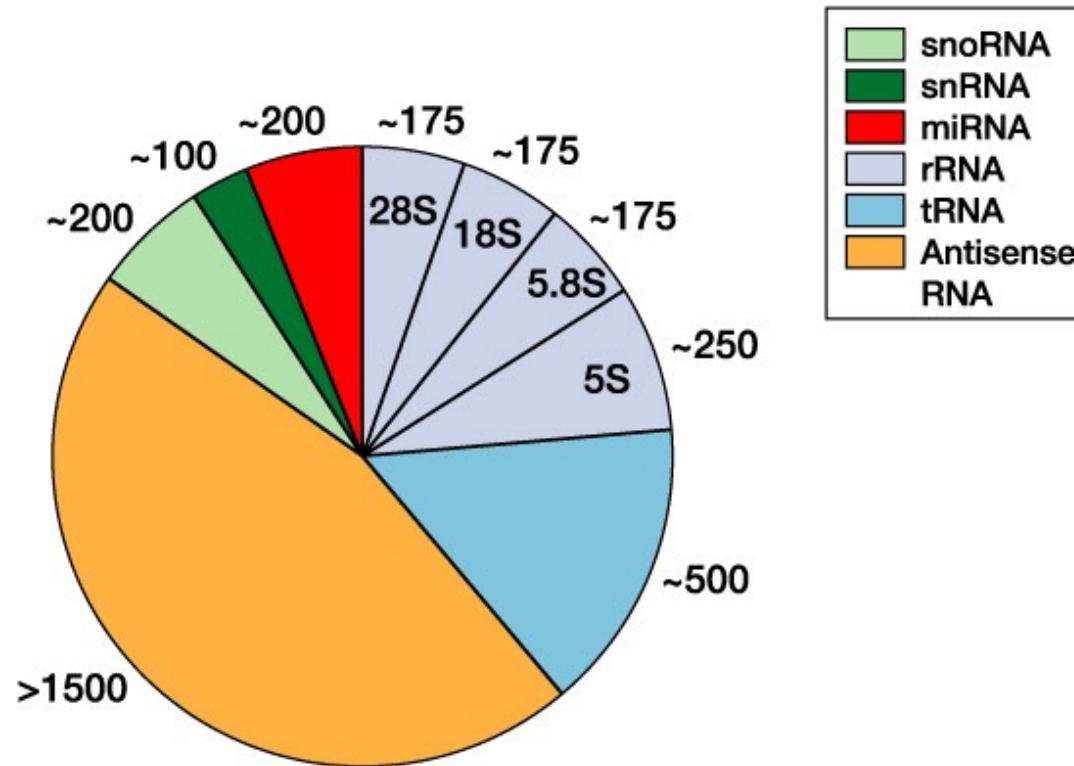
A remarkable discovery of whole-genome sequencing studies was that protein coding exons account for less than 2% of mammalian DNA.

Subsequent largescale transcriptome studies have established that at **least two-thirds of the genome is nonetheless transcribed into RNA**, exposing novel and exciting layers of genetic regulation.

Mammals, genomic regions previously believed to be so-called “junk DNA” are now known to be rich with transcripts encoding a large and diverse population of RNAs without protein-coding potential, referred to as noncoding RNAs (ncRNAs).

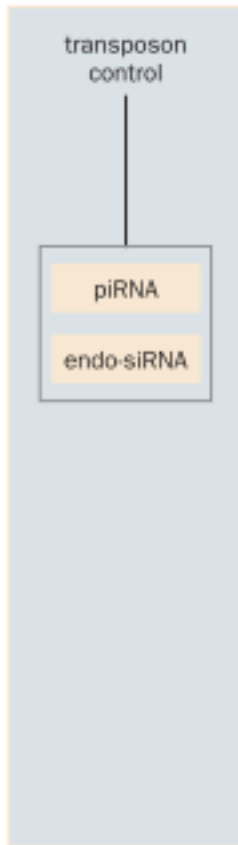
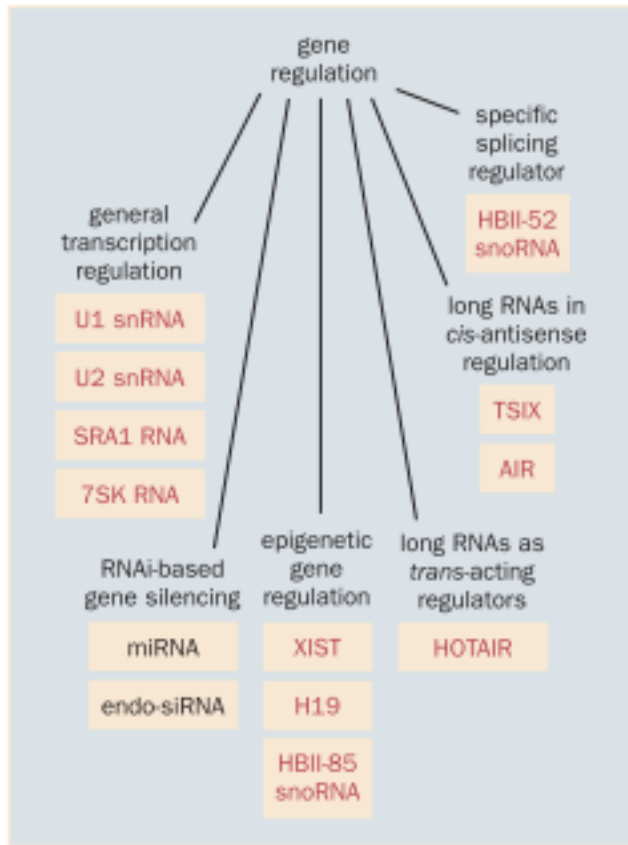
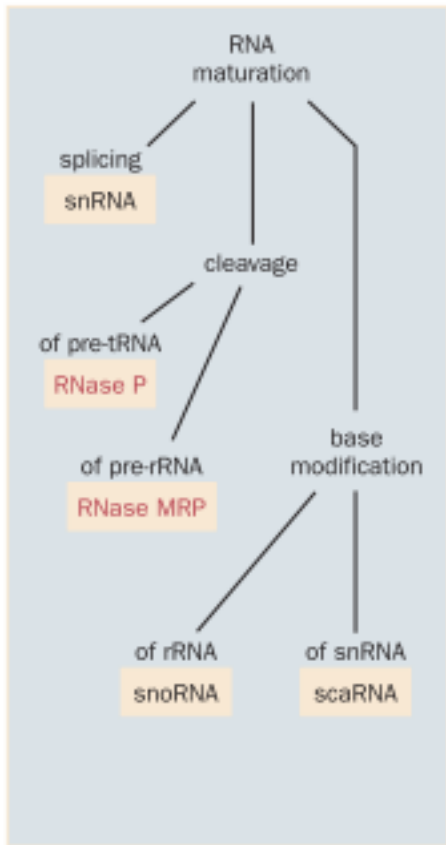
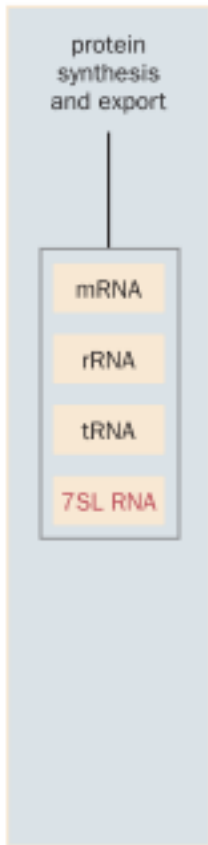
# Organizzazione dei Geni

## RNAs genes



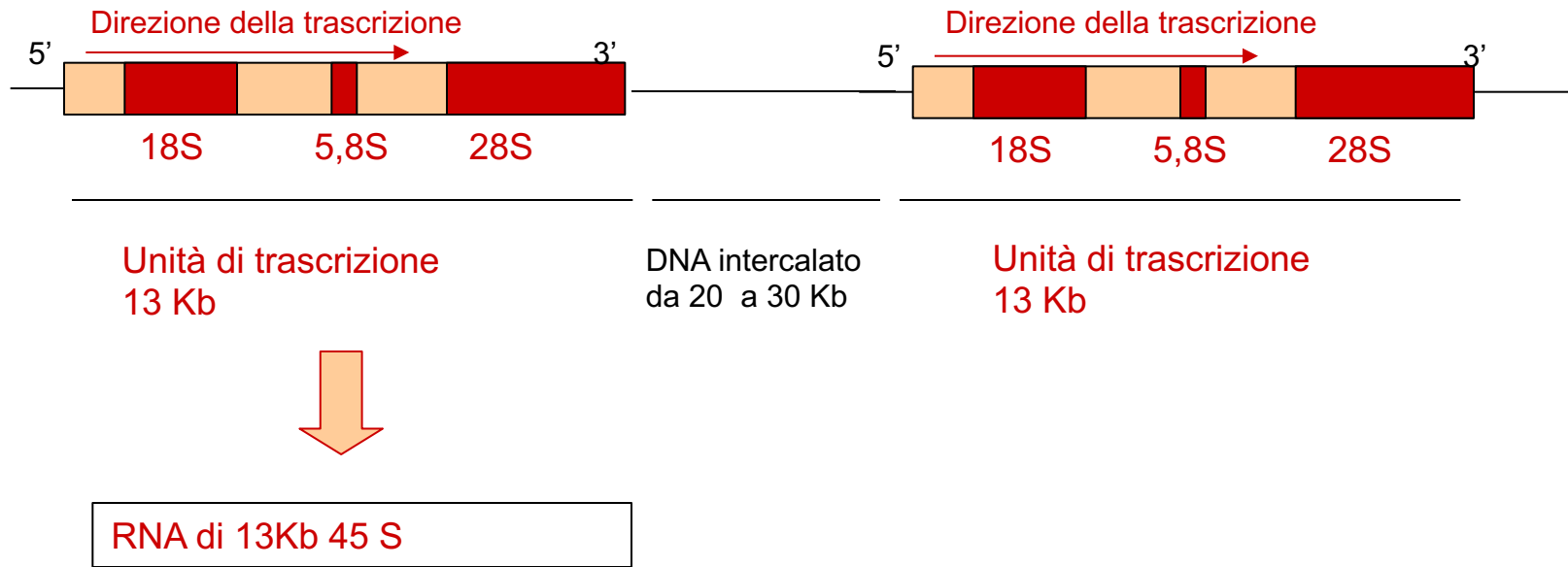
Circa il 5% dei geni nucleari sono “RNA-genes”

Attualmente se ne contano circa 6000



# I geni per gli rRNA-eucariotici

cluster di geni  
ripetuti



RNA di 13Kb 45 S

# Richiesta di grandi quantità di prodotto:

## Famiglia di geni ripetuti

L'unità di trascrizione è **ripetuta in tandem, identica**, circa 250 volte in cinque raggruppamenti (**clusters**), localizzati sul braccio corto dei cromosomi acrocentrici (50 unità ripetute in tandem per raggruppamento)

Le sequenze codificanti sono

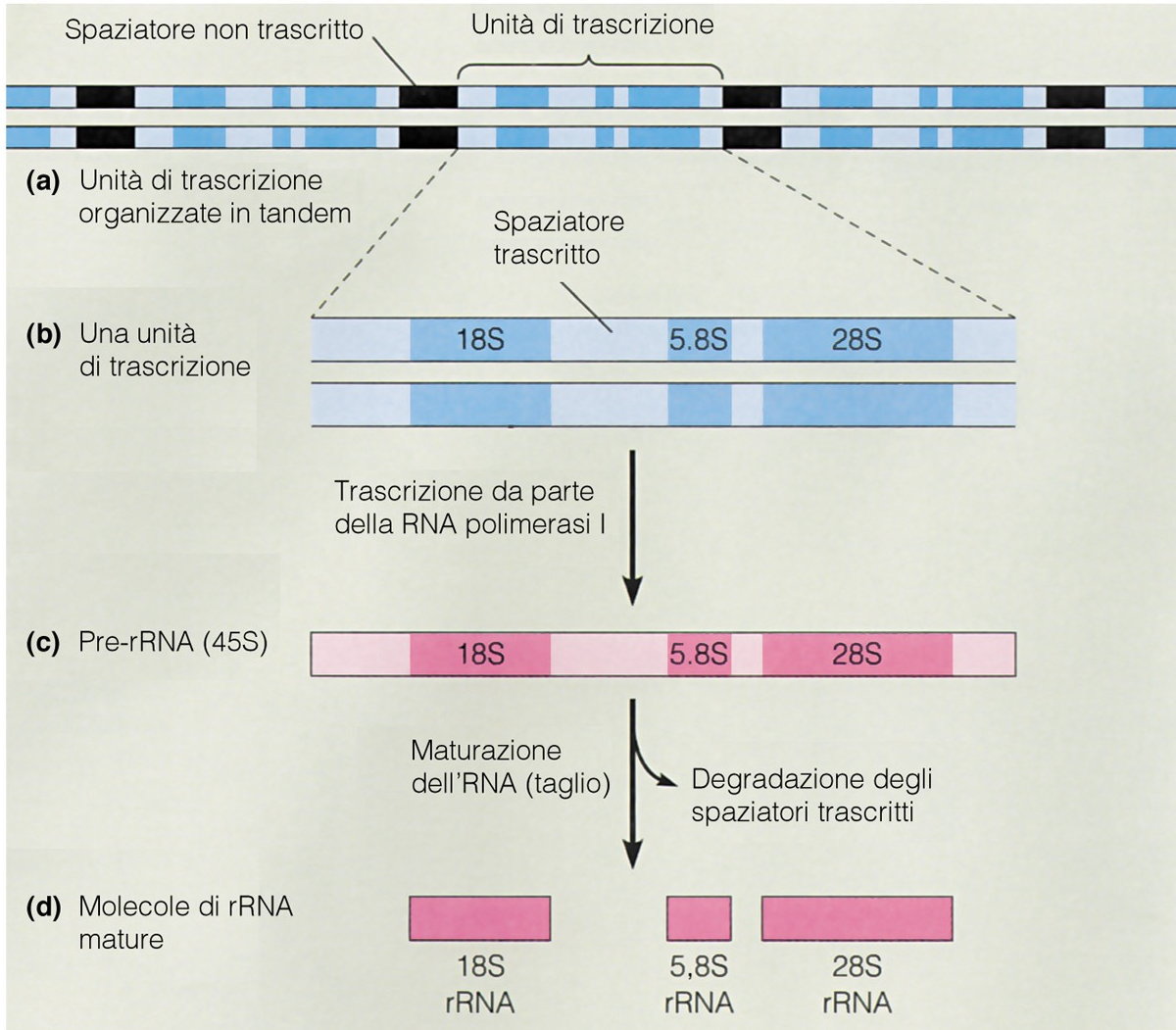


Figura 19-16

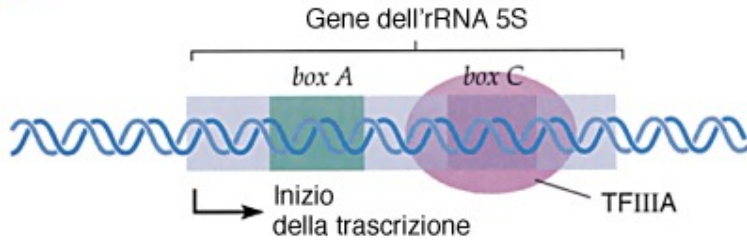
# I geni ribosomali

Organismo	n° ripetizioni	Organizzazione
Lievito	100-200	raggruppati su un singolo cromosoma
drosofila	130-250	due raggruppamenti su due cromosomi
Xenopus	400-600	1 cluster
uomo	300	5 cluster

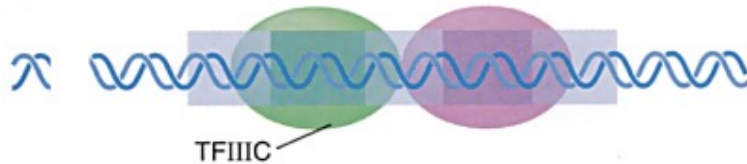
**TABLE 9-2** Effect of Gene Copy Number and Loading with RNA Polymerase on Rate of Pre-rRNA Synthesis in Human Cells

Copies of Pre-rRNA Gene	RNA Polymerase Molecules per Gene	Molecules of Pre-rRNA Produced in 24 Hours
1	1	288
1	≈250	≈70,000
100	≈250	≈7,000,000

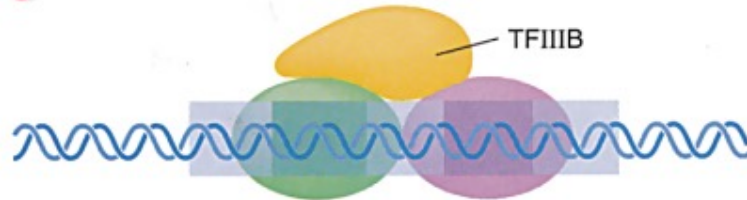
1 TFIIIA si lega al *boxC*



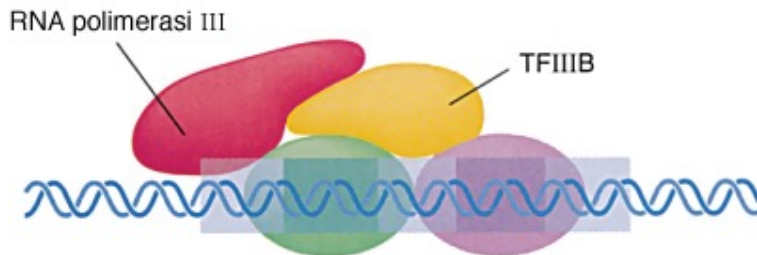
2 Viene facilitato il legame di TFIIIC al *boxA*



3 TFIIIB si lega agli altri TF, ma non al DNA



4 TFIIIB posiziona la RNA polimerasi III sul gene

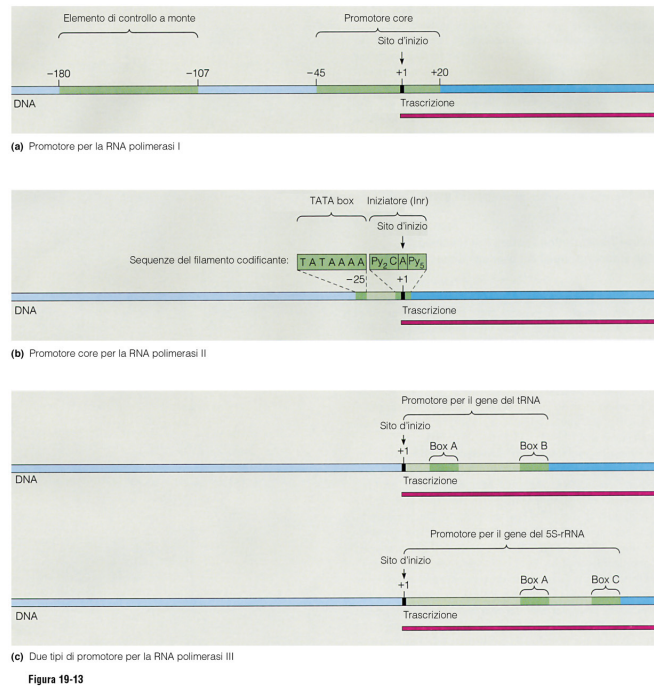


I geni per l'rRNA 5S, per i tRNA e per alcuni snRNA hanno una struttura simile.

Il promotore è interno all'unità di trascrizione.

Questi geni sono trascritti dalla RNA polIII

# I geni per i tRNA-umani



Circa 500 geni per tRNA, classificabili in 49 famiglie sulla base della specificità dell'anticodone.  
Sono dispersi nel genoma ma con circa il 50% presente sul cromosoma 1 o sul 6.



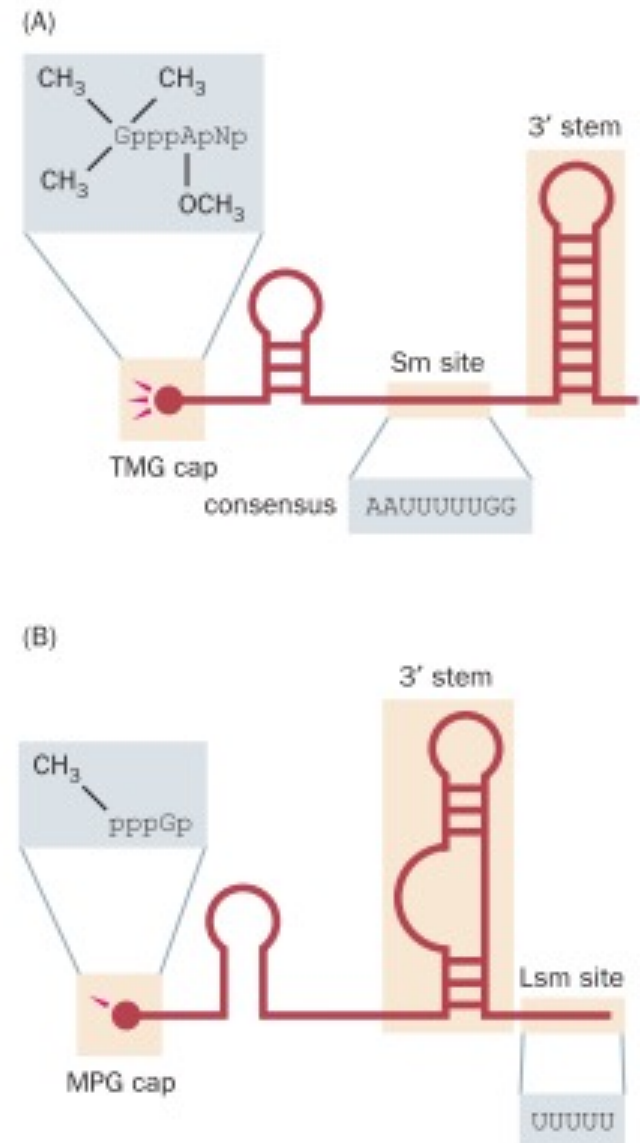
# Small nuclear RNA (snRNA) genes

Due sottoclassi:

a) Spliceosomal family (Sm and Lsm), nove differenti tipi-U1, 2, 4, 5 e 6 che processano gli introni standard GU-AG e U4atac, U6atac, U11 e U12 che processano gli introni rari AU-AC.

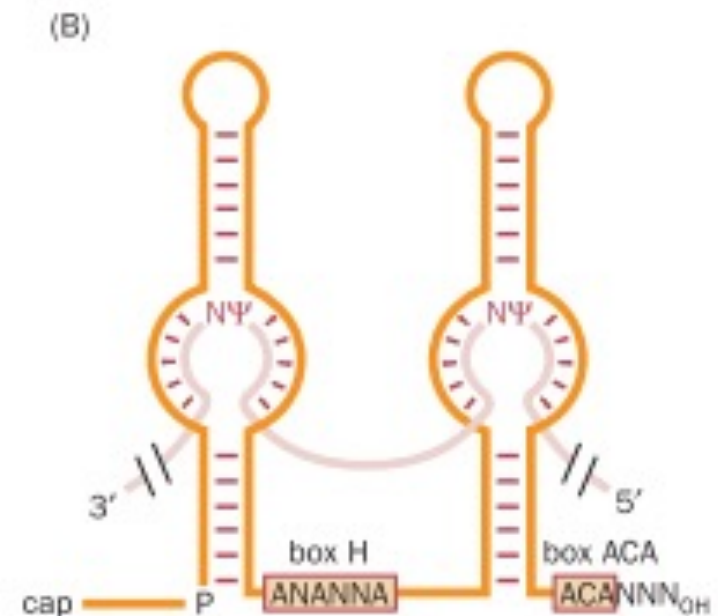
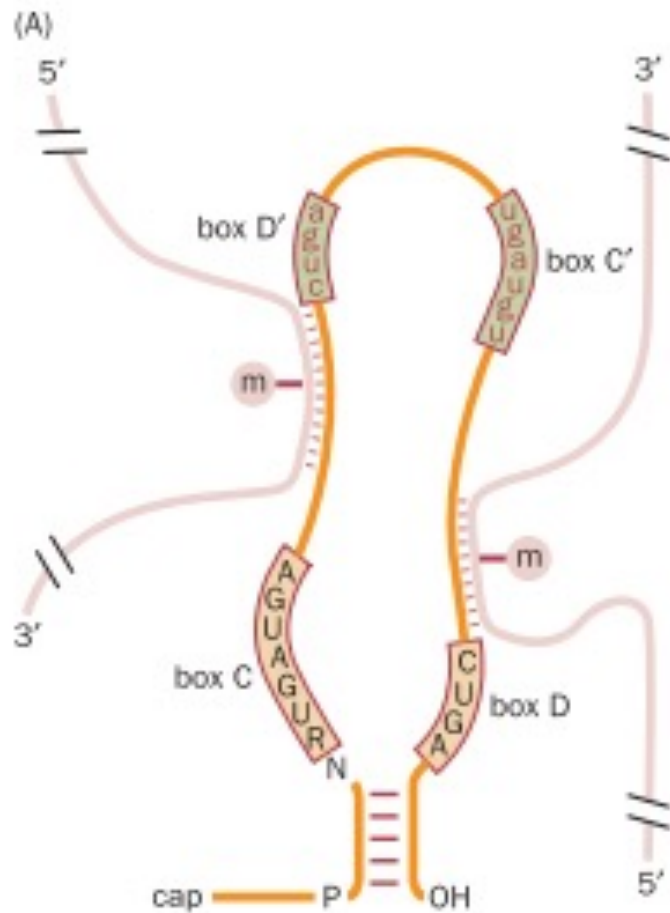
b) Non-spliceosomal snRNAs, ne esistono diversi tipi, con funzioni diverse (regolazione trascrizione, replicazione DNA, cell proliferation)

Per lo più geni funzionali a singola copia

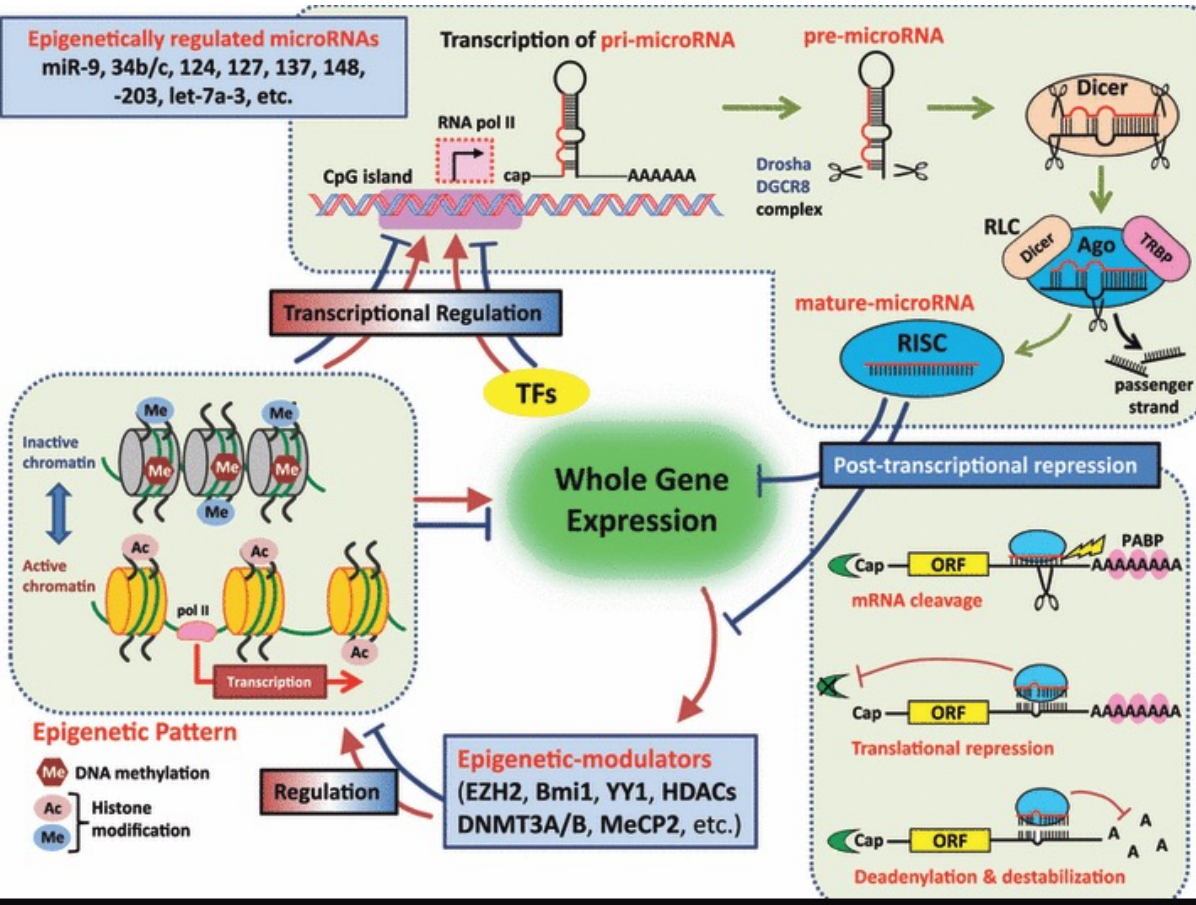


## Small nucleolar RNA (snoRNA) genes

Lunghi tra i 60 e i 300 nt, appartengono a due sottofamiglie, implicati nella maturazione degli rRNA. Di solito i geni si trovano in introni di protein-coding genes, in più localizzazioni cromosomiche. In alcuni casi sono in copie multiple clusterizzate.



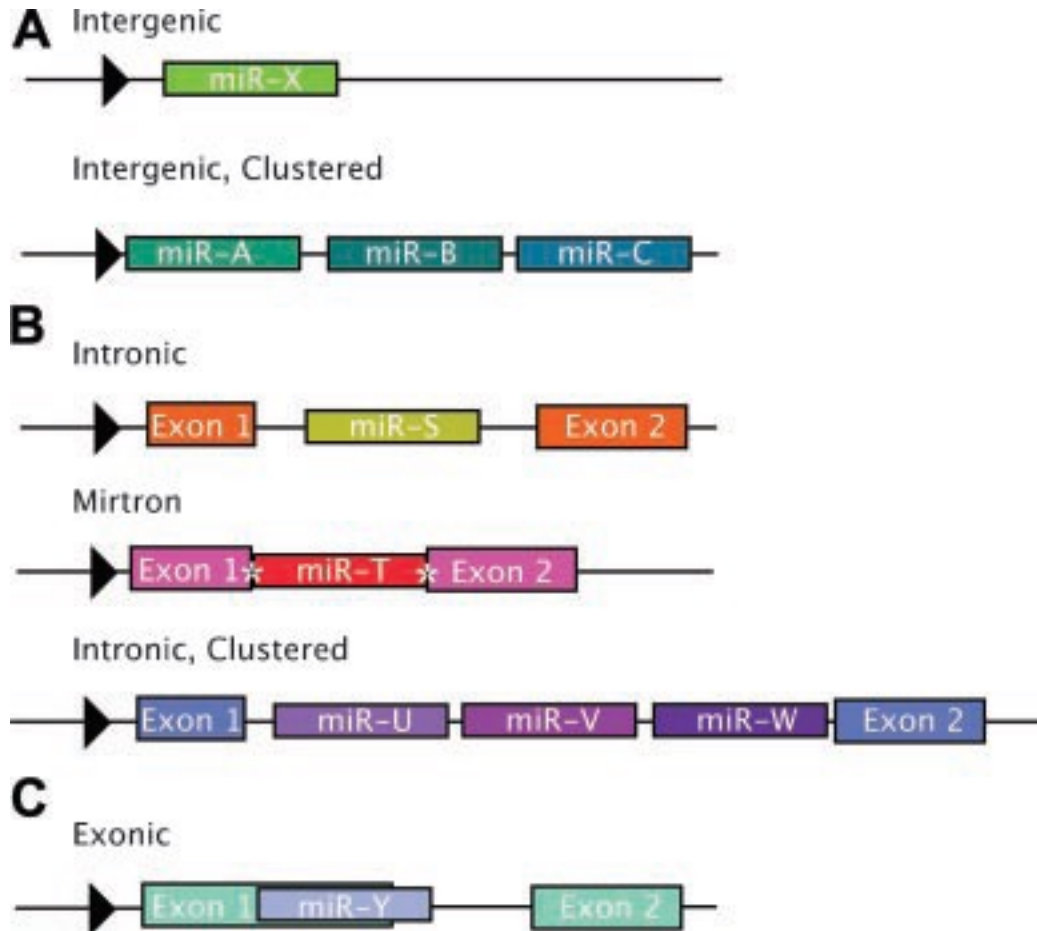
# micro RNA



In the nucleus, mainly RNA pol II initially transcribes miRNAs into long segments of RNA, known as pri-miRNAs, which are usually capped and polyadenylated.

MicroRNAs (miRNAs) comprise species of short noncoding RNA that regulate gene expression post-transcriptionally.

# Genomic location of miRNAs



**Intergenic miRNAs:** la maggioranza ha un trascritto primario di 3-4 kb, delimitato da un transcription start site e da un segnale di poliadenilazione.

Il 36% di questi miRNAs sono stati trovati in clusters.

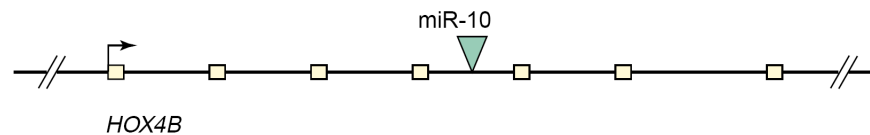
Questi clusters frequentemente contengono miRNA rappresentativi di diverse famiglie, che quindi possono targettare multipli differenti mRNA.

Basandosi sull'analisi dei network di interazioni proteiche è stato proposto che mRNA codificanti per proteine che interagiscono tra loro sono tipicamente targettati da miRNA derivanti dallo stesso cluster policistronico.

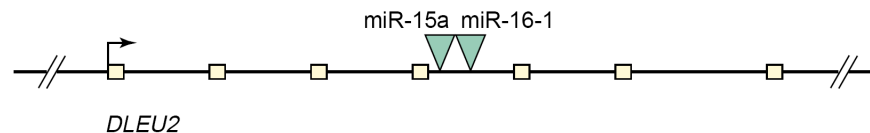
## miRNA codificati all'interno di unità di trascrizione:

possono essere intronici, esonici o una combinazione dei due. Quelli esonici sono una minoranza e solitamente sono all'interno di trascritti non codificanti. miRNA intronici, in orientamento senso, sono i più comuni. Generalmente si osserva una espressione coordinata tra il gene ospite e i miRNA. Si possono trovare sia all'interno di protein-coding genes sia di long non-coding RN,

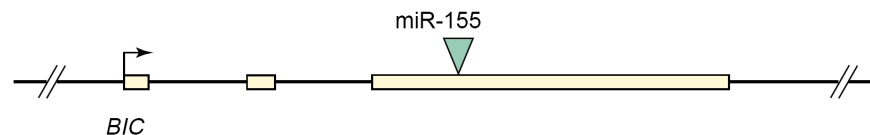
(a) Intronic miRNA in protein-coding TU (61%)



(b) Intronic miRNA in non-coding TU (18%)



(c) Exonic miRNA in non-coding TU (20%)



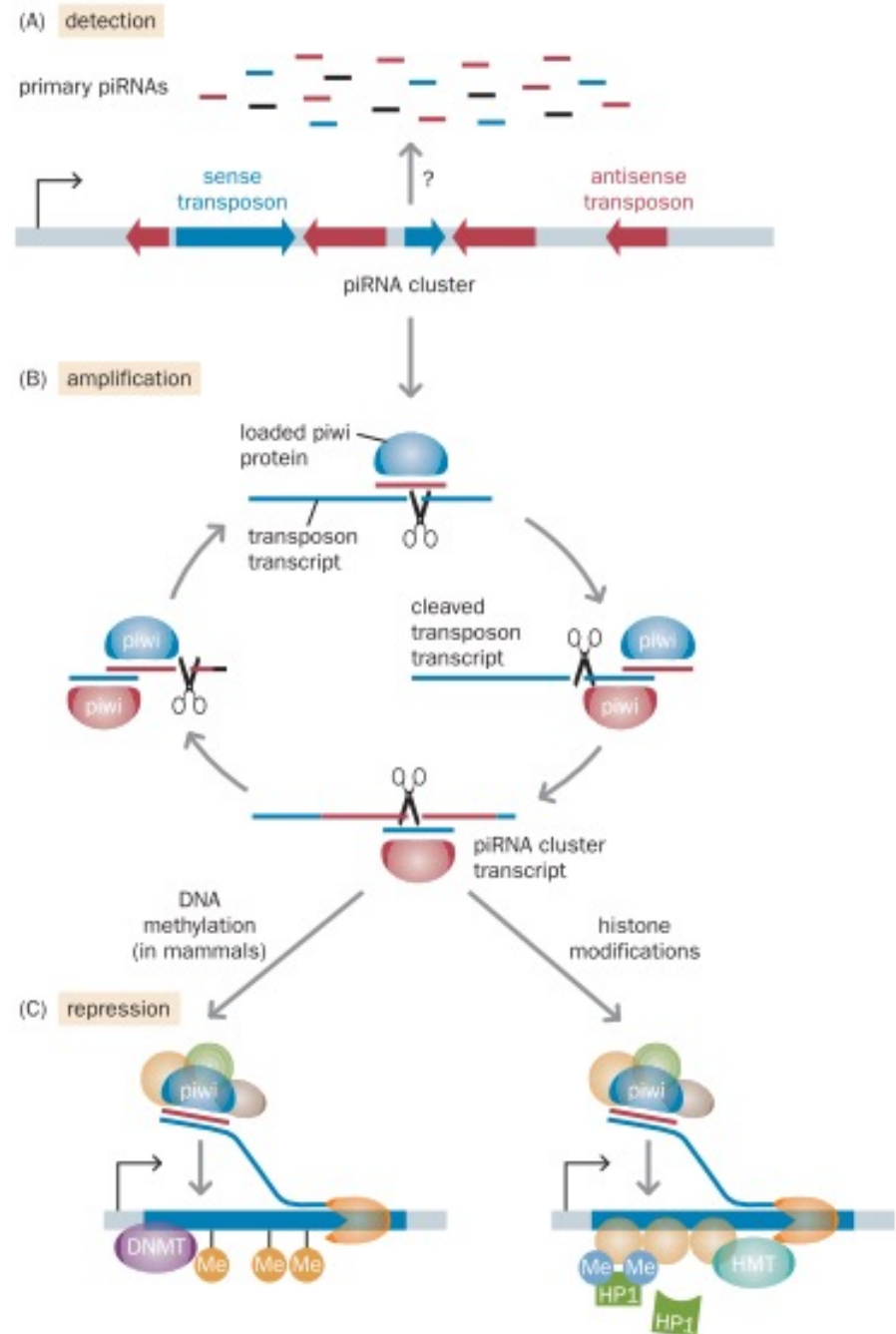
# piRNAs

## Piwi-protein-interacting RNAs

More than 15,000 different human piRNAs have been identified, that map back to 89 genomic intervals of about 10-75 kb.

Primary piRNAs are 24-31 nt long and are processed from long RNA precursors transcribed from defined loci called piRNAs clusters.

They are expressed in germ-line cells in mammals, where they have a major role in limiting transposition by retrotransposons





Less than 2% of the mammalian genome carries protein-coding potential, yet 70 to 90% is transcribed at some point during development to produce a large transcriptome of long noncoding RNA (lncRNA, defined as RNA > 100 nucleotides in length). Some estimate total membership to exceed 200,000, whereas others suggest fewer than 10,000.

## LNC RNAs

lncRNAs range in size from 200 nucleotides to beyond 10 kilobases. They can be polyadenylated or not, can be nuclear or cytoplasmic, and are generally expressed at lower levels compared with protein-coding mRNAs.

Similar to protein-coding genes, many lncRNA genes are bound by essential cell-type-specific nuclear factors, transcribed from what appear to be conventional promoters, and are spliced.

Other lncRNAs, both polyadenylated and nonpolyadenylated, arise from enhancers.

It is estimated that many thousands of lncRNAs are encoded in the human



<b>RNA</b>	<b>Size</b>	<b>Gene location</b>	<b>Gene organization</b>	<b>Function</b>
<b>XIST</b>	19.3 kb	Xq13	6 exons	Regulator of X-chromosome inactivation
<b>TSIX</b>	37.0 kb	Xq13	1 exon	Antisense regulator of XIST
<b>H19</b>	2.3 kb	11p15	5 exon	Involved in imprinting. Associated to Beckwith-Wiedemann syndrome

The story of lncRNAs first debuted in the phenomena of genomic imprinting and X-chromosome inactivation (XCI).

To balance X-chromosome gene expression between males and females, the X-inactivation center (Xic) on the mammalian X chromosome controls the initiation steps of XCI through a series of RNA-based switches. Today, the Xic serves as a model for understanding epigenetic regulation by lncRNA.

By chance una copia di cDNA può integrarsi in una regione cromosomica adiacente ad un promotore, che può quindi dirigerne l'espressione. Molte di queste copie provengono dal cromosoma X ed hanno un pattern di espressione testis-specifico. Questo può essere servito per superare il problema dell'inattivazione del cromosoma X (nel body XY) durante la meiosi maschile.

Esempi di retrogeni privi di introni e del loro gene omologo (con introne)		
<b>Retrogene</b>	<b>Gene omologo</b>	<b>Prodotto</b>
GK2 4q13	GK1 Xp21	Glycerol kinase
PDHA2 4q22	PDHA1 Xp22	Pyruvate dehydrogenase
PGK2 6p12	PGK Xq13	Phosphoglycerate kinase
TAF1L 9p13	TAF1 Xq13	TATA box binding protein associated factor, 250 kD
MYCL1 1p34	MYCL2 Xq22	Homolog of v-Myc oncogene
GLUD1 10q23	GLUD2 Xq25	Glutamate dehydrogenase
RBMXL 9p13	RBMX Xq26	RNA-binding protein important in brain development

Bioinformatic methods have identified thousands of lncRNAs, and an interesting debate has emerged regarding how many of these actually exert important biological functions.

For example, enhancers frequently produce lncRNA transcripts. Some of these may simply represent nonfunctional byproducts of transcription factor chromatin occupancy, whereas others are apparently required in cis for the regulatory activity of the enhancer.

Understanding the functions of lncRNAs is also complicated by their interesting and unusual evolutionary properties. For example, **lncRNA promoters and loci show greater sequence conservation than background DNA, although it appears that many lncRNAs detected in one species might not even be transcribed in another species**. Whether this implies lack of function or an ability to gain and lose function more rapidly during evolution than protein-coding genes requires further study.

# Non Coding RNAs:

‘Diversity of life’

## Box 1 | Abbreviations for different classes of non-coding RNA

- fRNA  
Functional RNA — essentially synonymous with non-coding RNA<sup>104</sup>
- miRNA  
MicroRNA — putative translational regulatory gene family
- ncRNA  
Non-coding RNA — all RNAs other than mRNA<sup>13</sup>
- rRNA  
Ribosomal RNA
- siRNA  
Small interfering RNA — active molecules in RNA interference
- snRNA  
Small nuclear RNA — includes spliceosomal RNAs
- snmRNA  
Small non-mRNA — essentially synonymous with small ncRNAs<sup>14</sup>
- snoRNA  
Small nucleolar RNA — most known snoRNAs are involved in rRNA modification
- stRNA  
Small temporal RNA — for example, *lin-4* and *let-7* in *Caenorhabditis elegans*
- tRNA  
Transfer RNA

rRN  
tRN

Val  
Y R  
7St

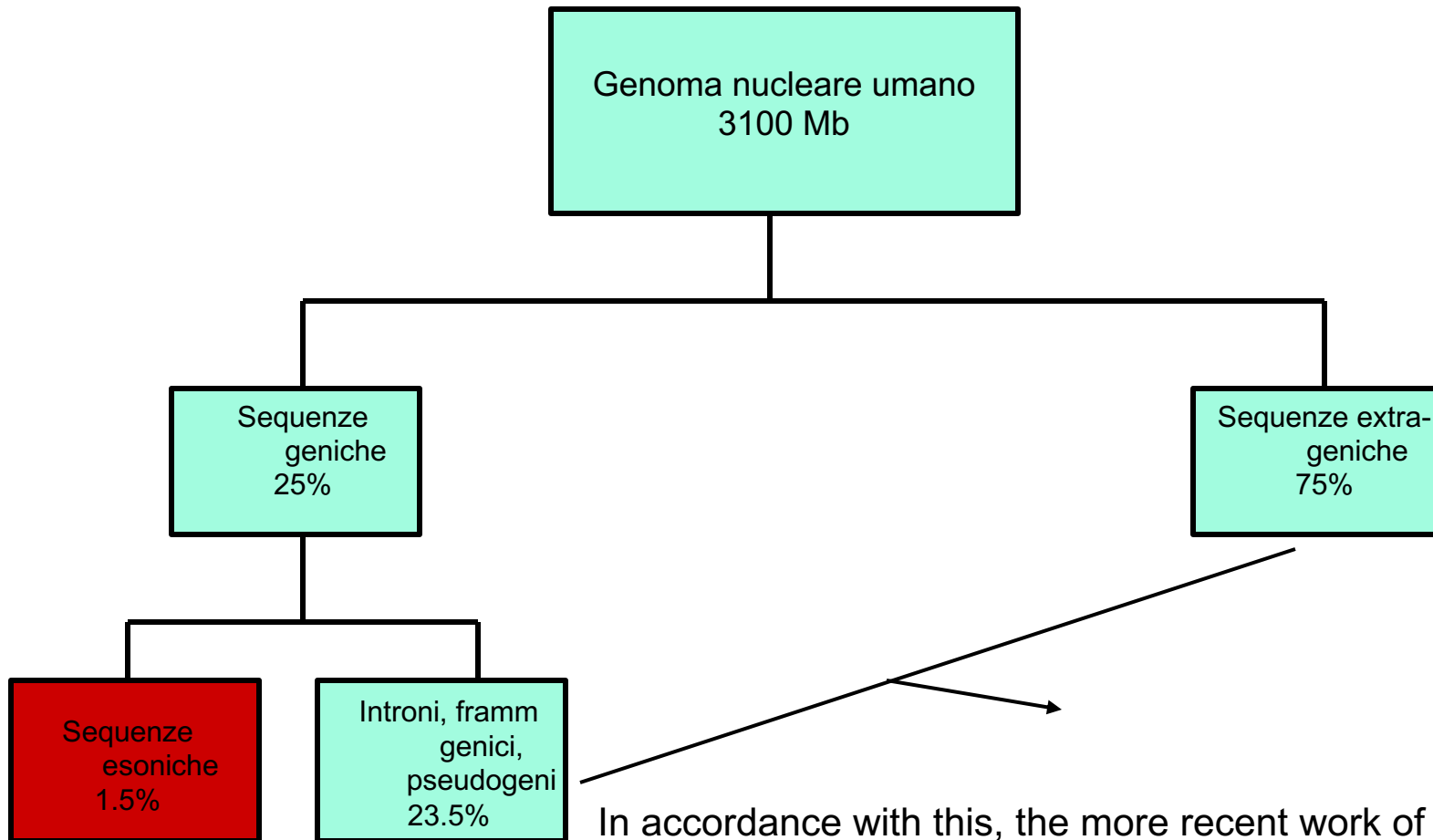
Xist,

ase

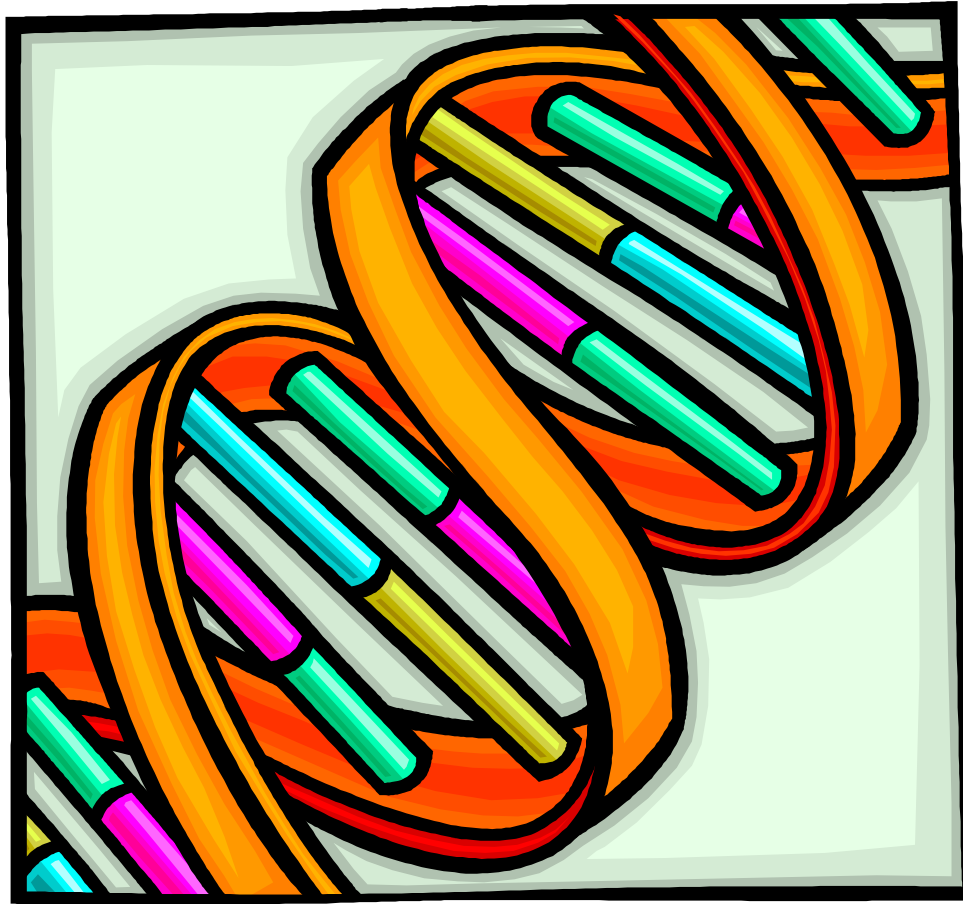
s

s  
sons

# Quindi, considerando l'organizzazione dei geni ....



In accordance with this, the more recent work of the ENCODE Consortium has indicated that at least 80% of the human genome is transcribed, resulting in the generation of a large number of non-coding RNA (ncRNA) transcripts. Although ncRNA was believed to be non-functional 'junk' or simply sequencing artefacts, various analyses have now highlighted that ncRNA species may have key roles in the regulation of gene expression.



**Famiglie di DNA ripetuto non genico**

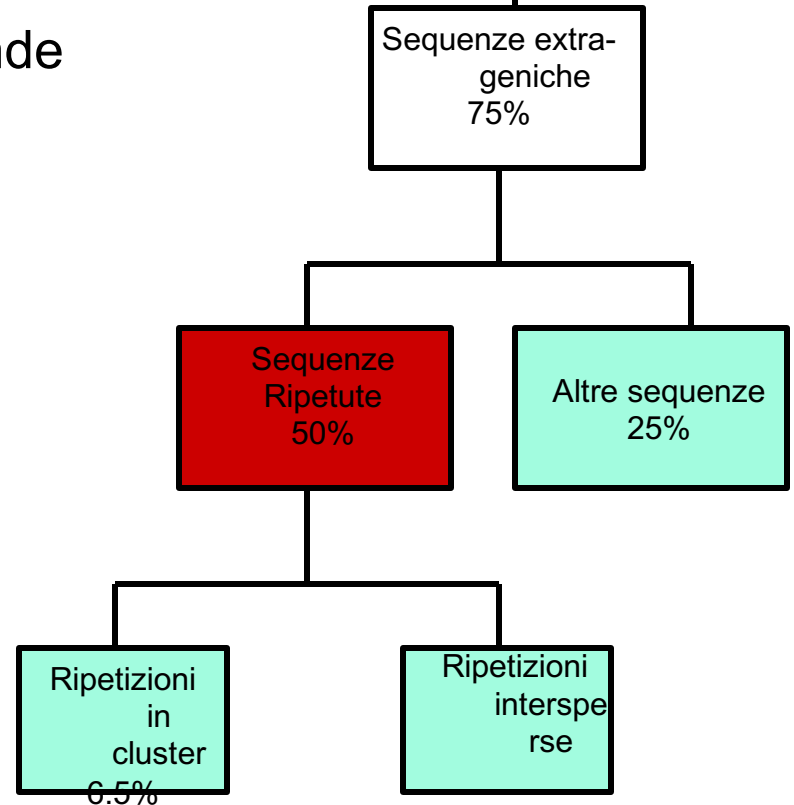
# Le sequenze extrageniche

Genoma nucleare umano  
3100 Mb

Il genoma nucleare contiene una grande  
quantità di sequenze ripetute



A. DNA RIPETUTO IN TANDEM  
B. DNA RIPETTUTO  
INTERSPERSO



# DNA altamente ripetuto in tandem

L'eterocromatina costitutiva rappresenta circa il 6.5% del genoma umano (200 Mb). E' maggiormente distribuita nelle regioni centromeriche (megabasi) e in minore estensione al telomero di tutti i cromosomi. Inoltre è presente in gran parte del cromosoma Y e sul braccio corto dei cromosomi acrocentrici.

E' principalmente costituita da lunghi tratti in tandem di sequenze di DNA ripetute ad alto numero di copie.

## DNA satellite

Sequenze semplici; unità di sequenza di 2 - 200 bp

Parecchie famiglie per ciascun organismo.

Poco conservate evolutivamente



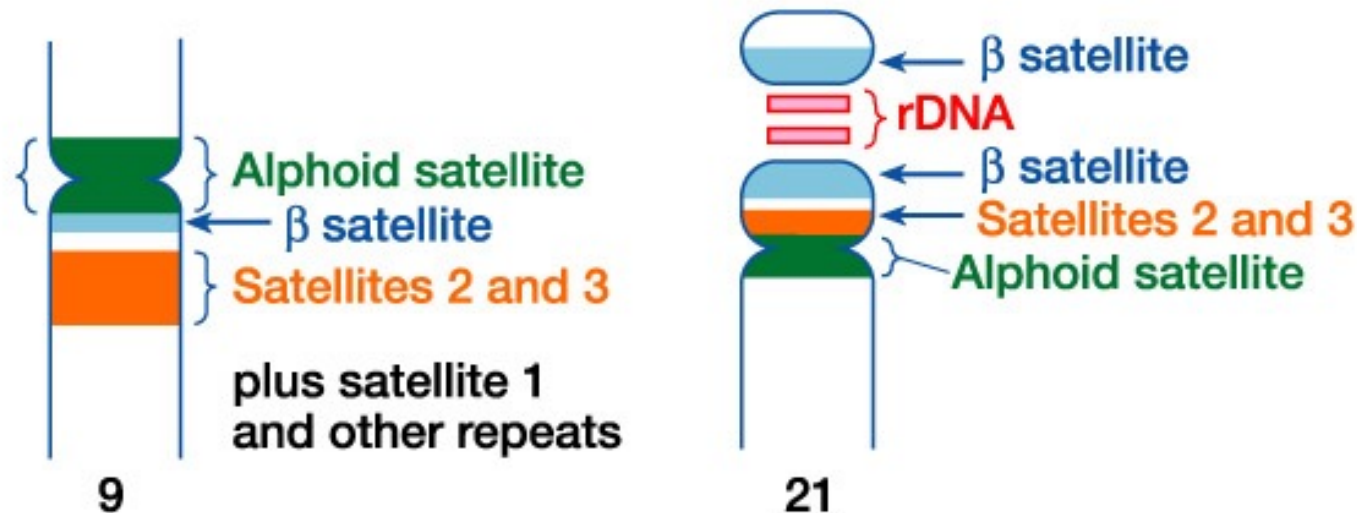
# Classi principali del DNA ripetuto in tandem

**Tabella 8.3** Classi principali di DNA umano ripetuto in tandem

Classe	Dimensioni dell'unità ripetuta (bp)	Principale localizzazione cromosomica
<i>DNA satellite</i> (spesso blocchi lunghi da 100 a diverse Mb)		
Satelliti 2 e 3	5	La maggior parte dei cromosomi (probabilmente tutti)
Satellite 1 (ricco in AT)	25-48	Eterocromatina centromerica della maggior parte dei cromosomi e altre regioni eterocromatiche
$\alpha$ (DNA alfoide)	171	Eterocromatina centromerica di tutti i cromosomi
$\beta$ (famiglia <i>Sau3A</i> )	68	In particolare l'eterocromatina centromerica dei cromosomi 1, 9, 13, 14, 15, 21, 22 e Y
<i>DNA minisatellite</i> (spesso blocchi lunghi da 0,1 a 20 kb)		
Famiglie telomeriche	6	Tutti i telomeri
Famiglie ipervariabili	9-24	Tutti i cromosomi, spesso vicino ai telomeri
<i>DNA microsatellite</i> (spesso blocchi inferiori alle 150 bp)	1-4	Tutti i cromosomi

Lunghi tratti di eterocromatina sono normalmente composti da un mosaico di differenti sequenze di DNA satellite, occasionalmente interrotte da sequenze tipo transposone, ma prive di geni.

La maggior parte del DNA eterocromatico non è stata sequenziata, con l'eccezione dell'eterocromatina del cromosoma Y

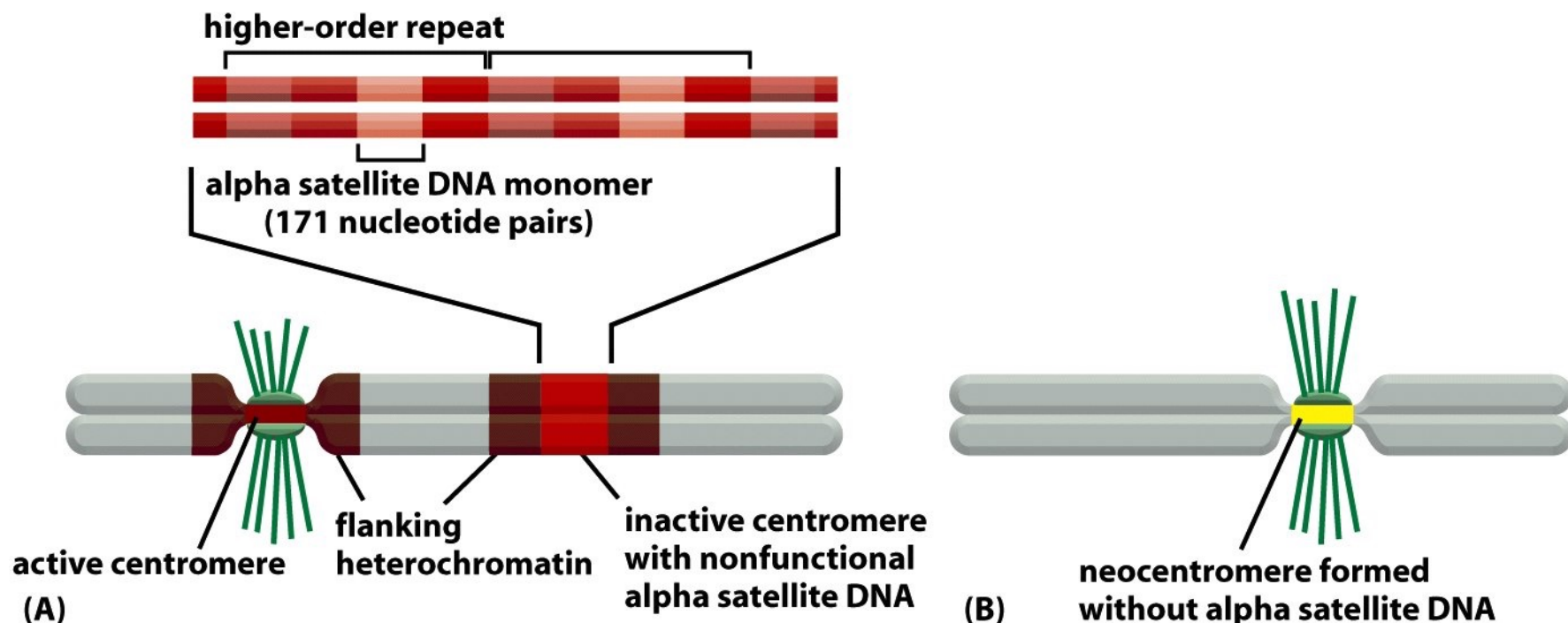


## Organizzazione dei DNA satelliti nei centromeri

I DNA centromerici sono un puzzle di sequenze appartenenti a diverse famiglie

La funzione centromerica dipende dalla struttura della cromatina più che dalle sequenze di DNA.

È un dominio definito **epigeneticamente**: una volta stabilito deve essere stabilmente mantenuto attraverso le molteplici divisioni cellulari.



Delle varie famiglie di DNA satellite associate ai centromeri solo il DNA  $\alpha$  satellite è presente in tutti i centromeri. La sua unità ripetuta spesso contiene un sito di legame per la proteina centromerica CENPB.

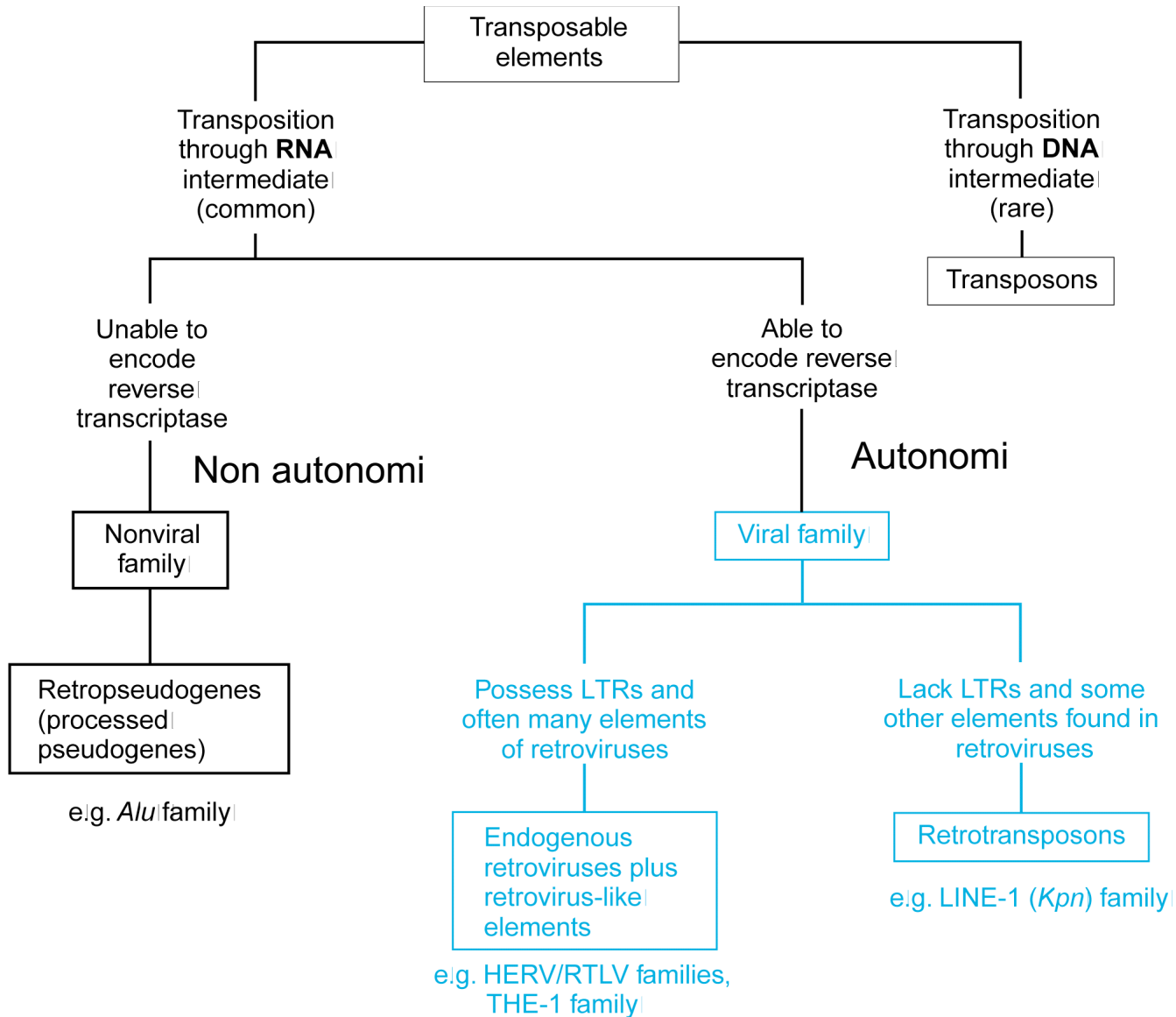
# Telomeri

➡ **COMPOSTI DA ETEROCROMATINA**

➡ **HANNO UNA SEQUENZA MINISATELLITE COMUNE (in uomo TTAGGG), CHE RAGGIUNGE UNA LUNGHEZZA DI POCHE KILOBASI.**

 ESSENZIALI NELLA  
STABILITA' DEI CROMOSOMI

# DNA ripetuto intersperso



# DNA RIPETUTO INTERSPERSO

Quasi tutte le sequenze ripetute intersperse presenti nel genoma umano derivano da (elementi trasponibili)

:

**SINE**-Short Interspersed Nucleotide Elements;

**LINE**-Long Interspersed Nucleotide Elements

**LTR**-Long Terminal Repeats

Classe	Famiglia	Dimensioni unità ripetuta	N° copie	% Genoma
<b>SINE</b>	<i>Alu</i>	0,3 kb lunghezza completa	1.200.000 ca	10,7% ca
	MIR	Dimensione media 0,13 kb	450.000 ca	2,5% ca
<b>LINE</b>	LINE-1 ( <i>Kpn</i> )	6,1 kb lunghezza completa, ma le dimensioni medie sono 0,8 kb	2600.000 ca	17,3%
	LINE -2	Dimensione media 0,25 kb	370.000	3,3%
<b>LTR</b>	ERV	Dimensione media 1,3 kb	240.000	4,7%
<b>Trasposoni a DNA</b>	MER-1 (Charlie)	Dimensione	213.000	3,0%
	MER-2	media 0,25 kb ca		

**SINE** la famiglia più rappresentata è quella *Alu*. Probabilmente la sequenza *Alu* deriva per retrotrasposizione dal gene dell'RNA 7SL, trascritto dalla RNA polIII.

*Alu* è specifica dei primati

Le ripetizioni *Alu* hanno un elevato contenuto in GC e sono localizzate nelle bande G chiare

La tipica sequenza *Alu* è circa 280 bp, e consiste di due sequenze ripetute in tandem; nella regione terminale è presente una corta sequenza ricca in residui A

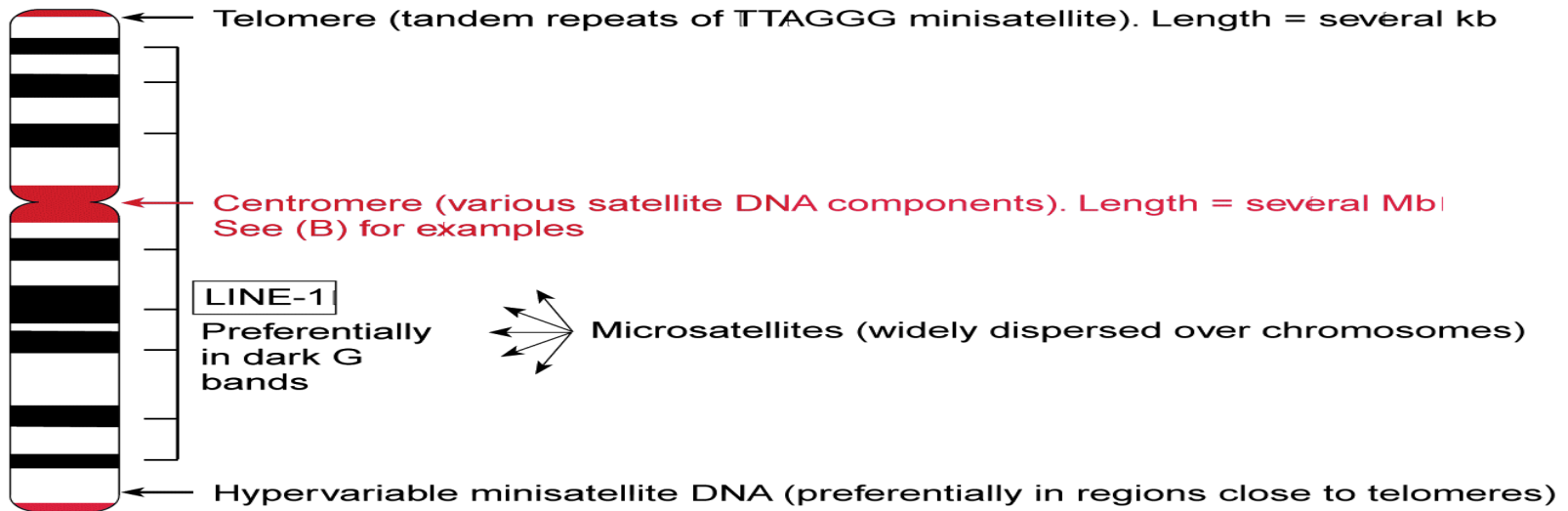


Tra i due dimeri vi è una asimmetria dovuta all'inserzione di un elemento di 32 nt all'interno della seconda ripetizione.

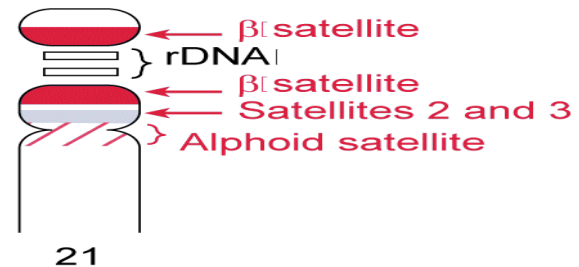
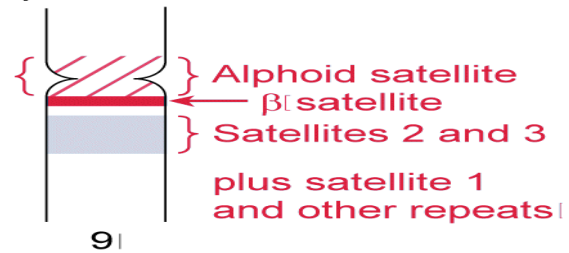
Si traspone parassitando il macchinario delle LINE

# Localizzazione genomica

(A)



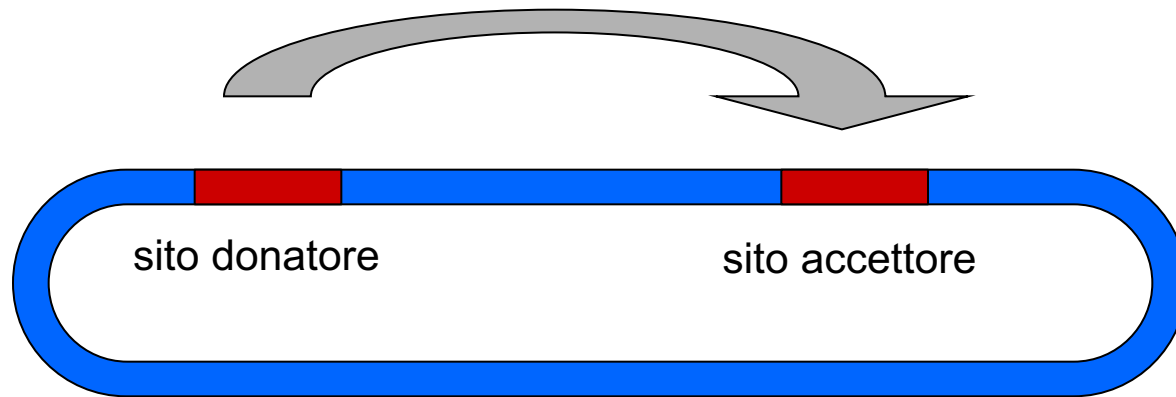
(B)



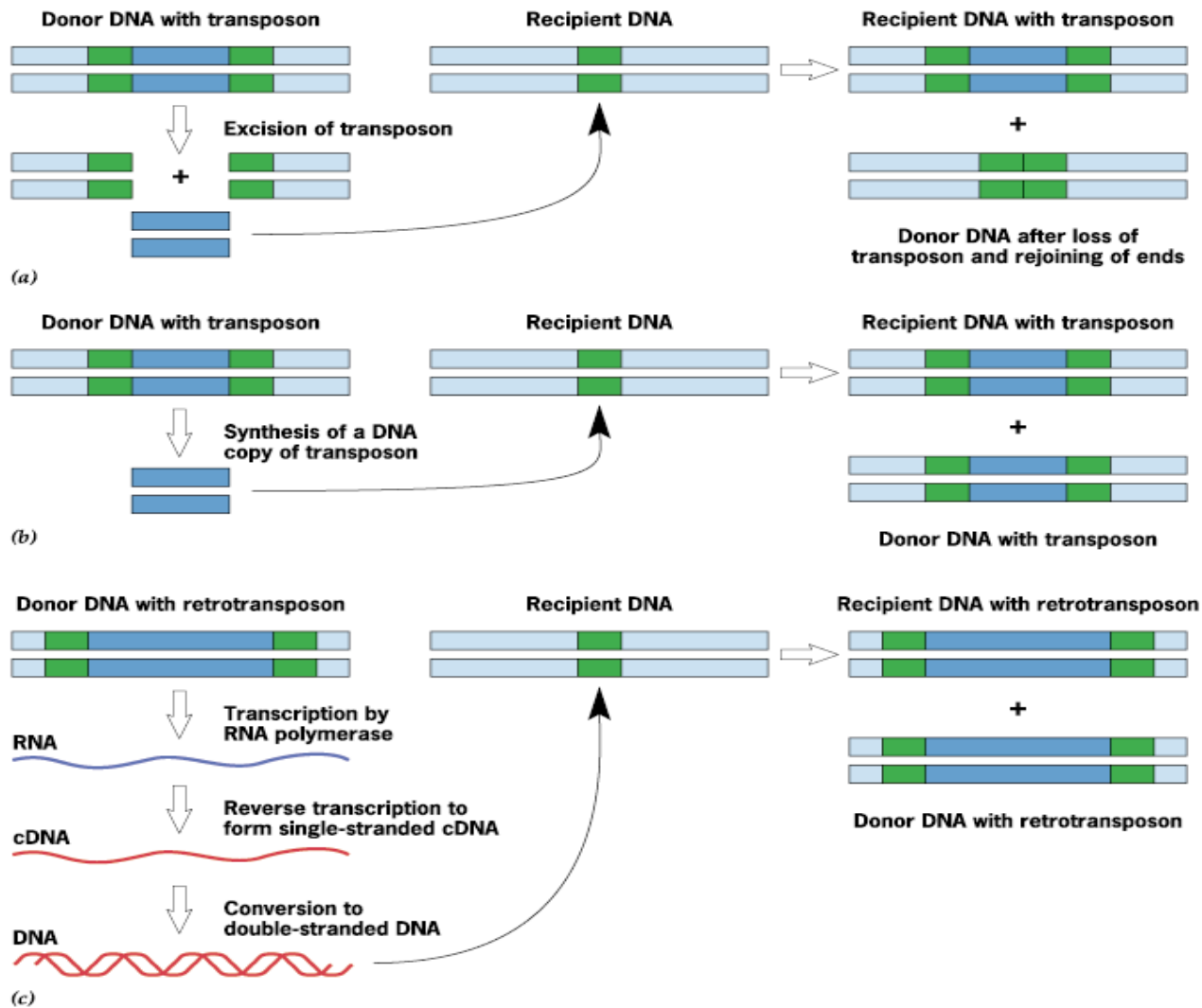


## Elementi trasponibili

Sequenze che hanno la capacità di muoversi da un sito all'altro del cromosoma

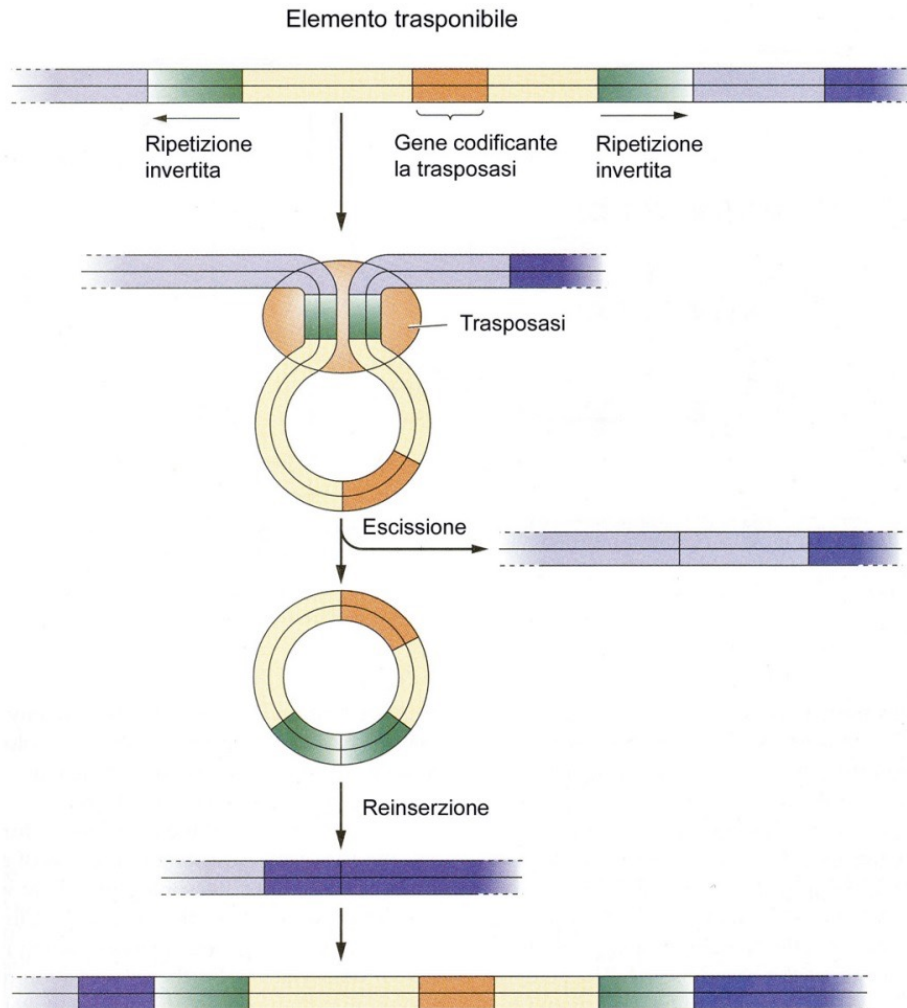


Diversamente dagli altri processi coinvolti nella ristrutturazione genomica, per la trasposizione non c'è nessuna omologia tra le sequenze del sito donatore e accettore



**Figure 10.30** Three alternate pathways by which transposable elements move from place to place within the genome.

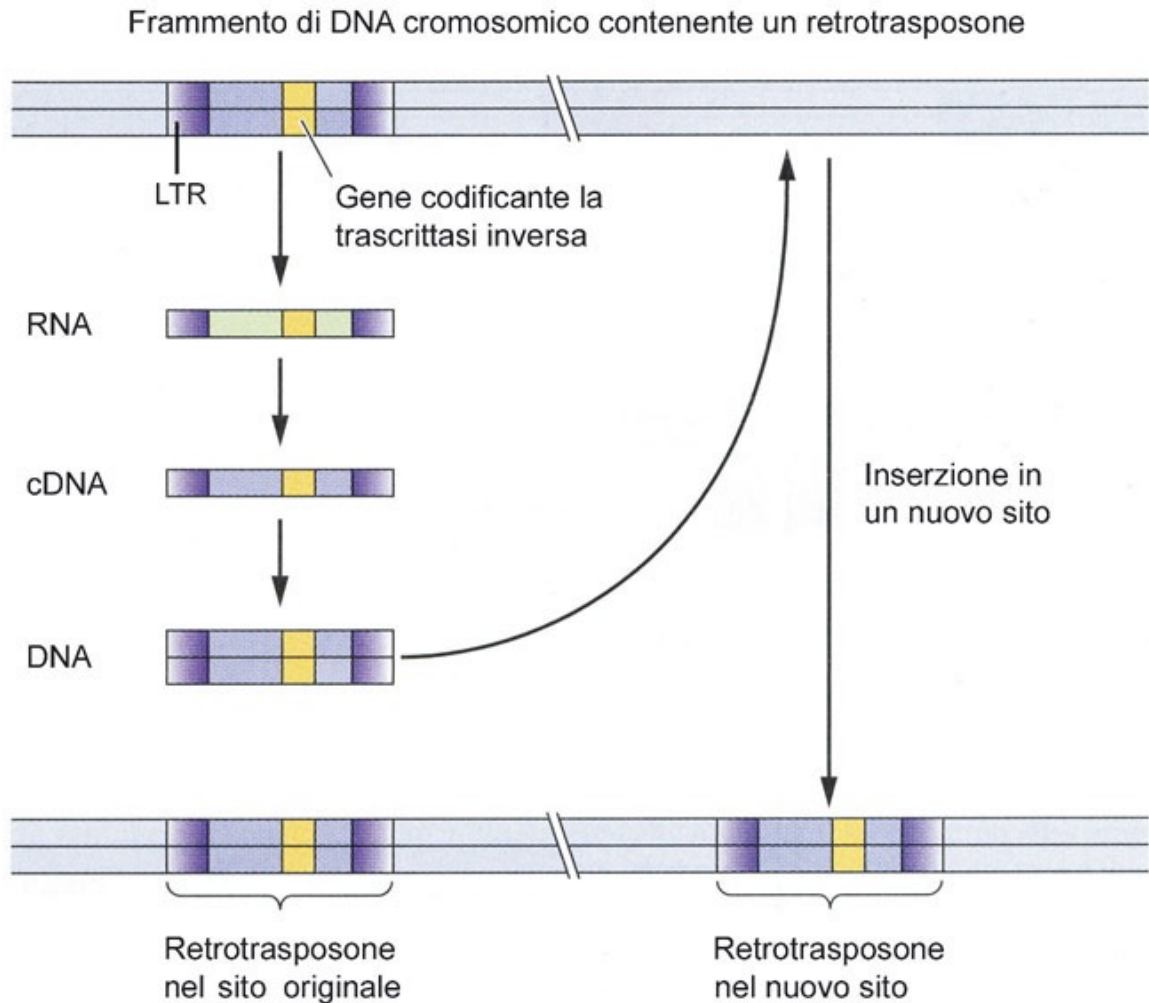
## 7.13 ELEMENTI GENETICI MOBILI O ELEMENTI TRASPONIBILI



**Figura 7.58**

Struttura e trasposizione di un elemento trasponibile: l'enzima trasposasi riconosce le sequenze ripetute invertite che fiancheggiano l'elemento trasponibile, forma una struttura ad anello inducendo l'escissione dell'elemento stesso che in seguito può inserirsi in un diverso sito genomico (da: B.B. Bunhanan *et al.* 2000, modificata).

## 7.13 ELEMENTI GENETICI MOBILI O ELEMENTI TRASPONIBILI



**Figura 7.59**  
Struttura e trasposizione di un retrotrasposone: l'enzima trascrittasi inversa produce una copia di DNA usando come stampo un RNA intermedio (da: B.B. Buchanan *et al.* 2000, modificata).

# Viral retrotransposons

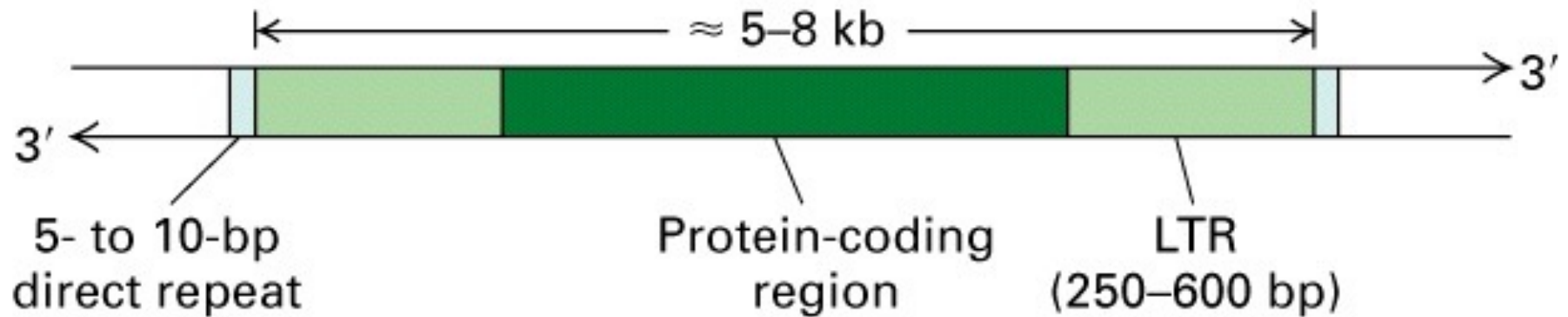
Includono elementi simil-retrovirali a:  
trasposizione autonoma. Sequenze Retrovirali Endogene (HERV). Contengono i geni *gag* e *pol* che codificano per una proteasi, una trascrittasi inversa, RNAse H e integrasi.

Circa 240.000 sequenze nel genoma (4.6%)

- trasposizione non autonoma. Mancano del gene *pol* e spesso anche del gene *gag*. 4% del genoma

# Viral retrotransposons contain LTRs and behave like retroviruses in the genome

Viral retrotransposons

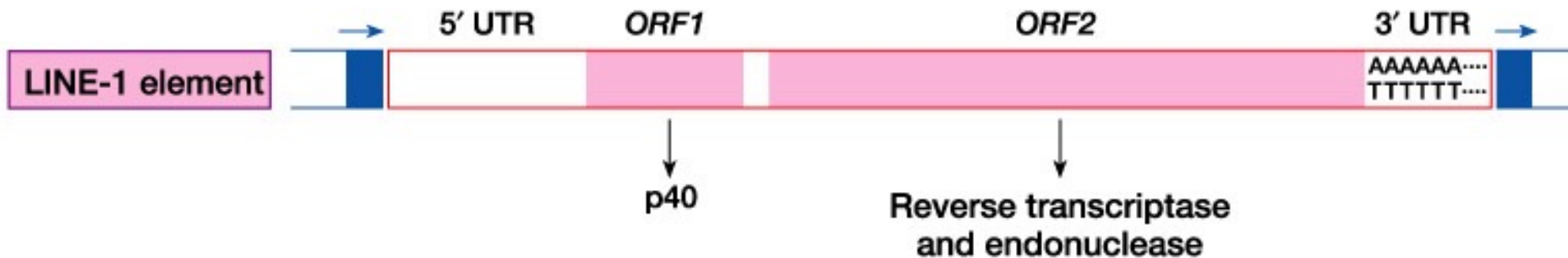


# Nonretroviral retrotransposons

**LINE:** I membri di tale famiglia sono elementi trasponibili autonomi, elementi di DNA instabili che migrano in regioni differenti del genoma.

Consiste di tre famiglie di cui la LINE-1 è la più rappresentata ed è l'unica ad avere dei membri con attività di trasposizione.

(A)



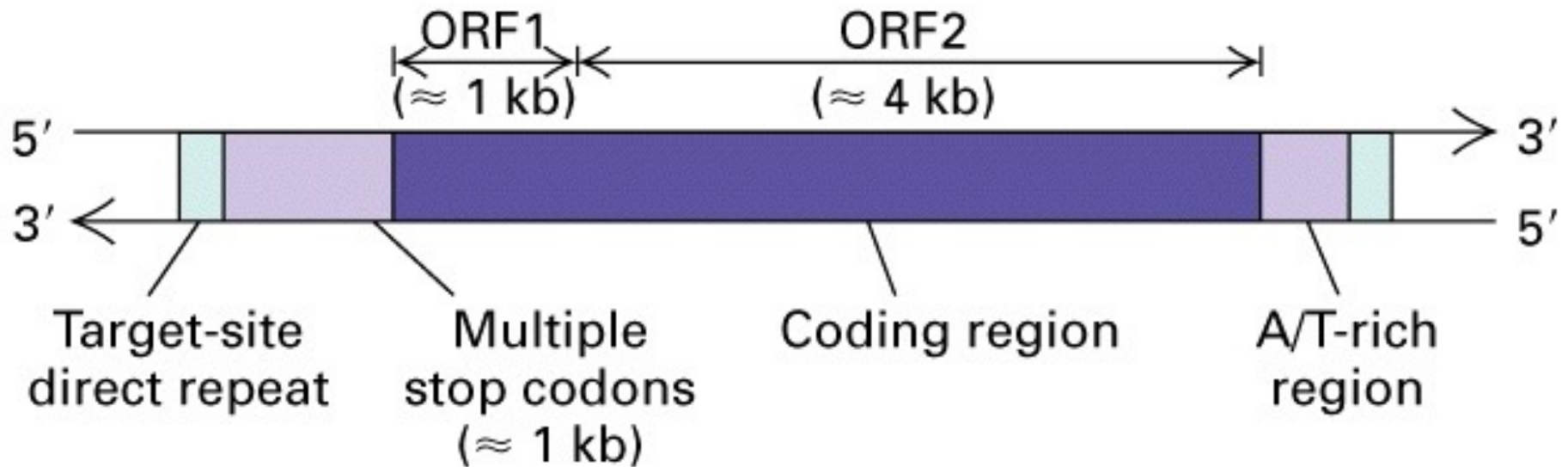
L'elemento consenso della famiglia *Kpn* (6,1 kb) possiede due ORF:

**ORF 1:** codifica per p40, una RNA-binding protein

**ORF2:** codifica per una proteina con un'attività endonucleasica e di trascrittasi inversa.

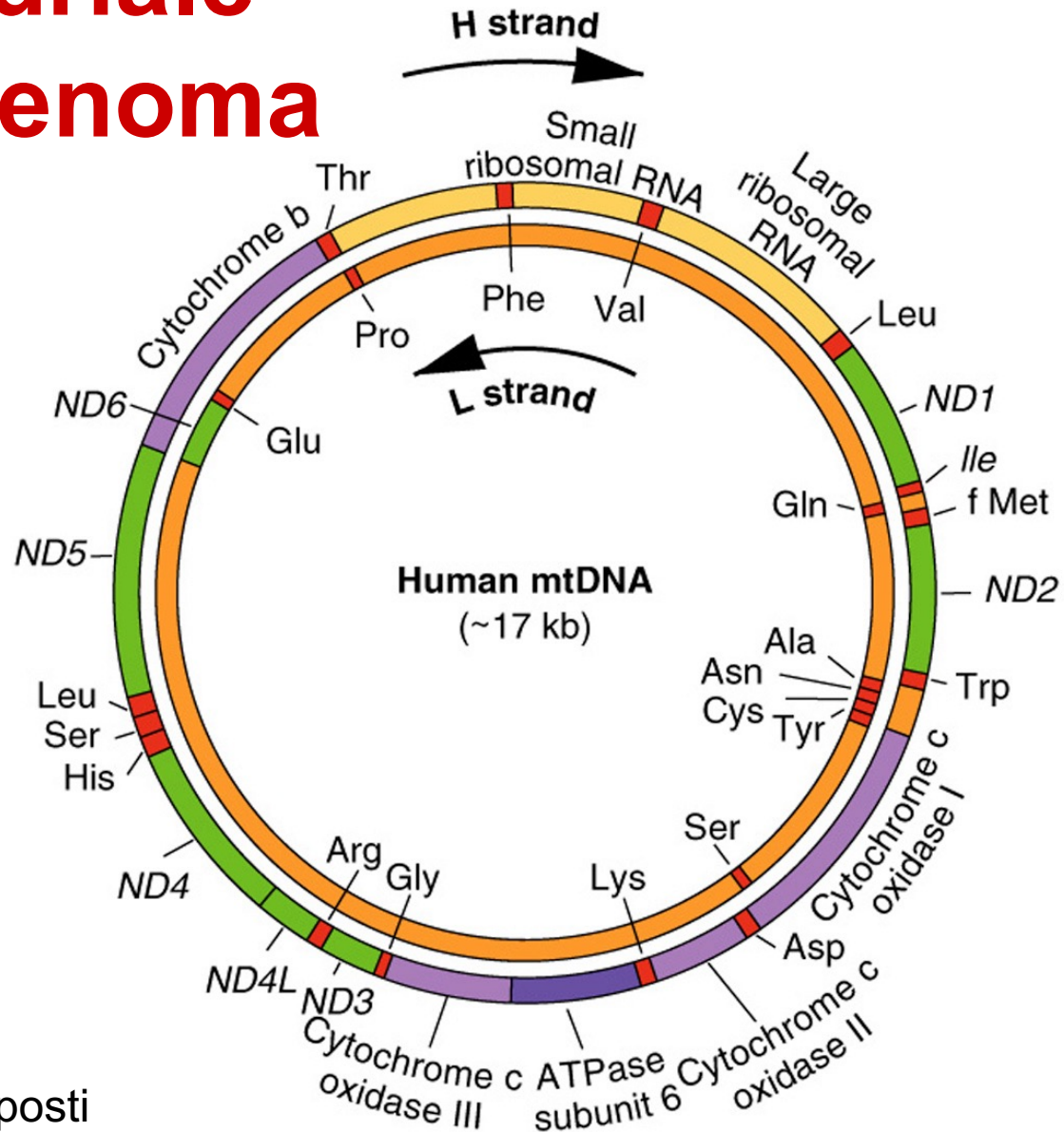
La sequenza completa è tuttavia rara.

# Nonretroviral retrotransposons lack LTRs





# Il DNA mitocondriale 0.0005% del genoma umano



**37 geni:**  
13 codificano per polipeptidi del mitocondrio (concetto di semiautonomia);  
24 codificano per prodotti maturi ad RNA, necessari per l'espressione del genoma mitocondriale.

I geni sono estremamente compatti: privi di introni e parzialmente sovrapposti

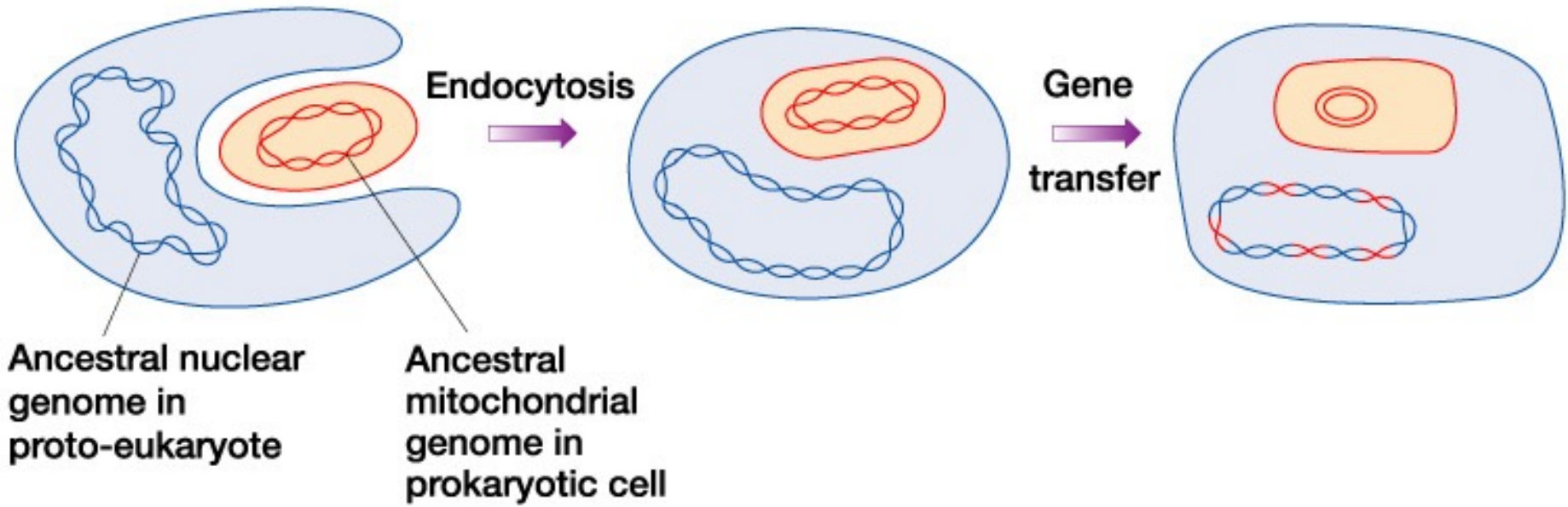


Figure 12-8 Human Molecular Genetics, 3/e. (© Garland Science 2004)

# Caratteristiche del mtDNA

- Poliplasmia: in ogni cellula sono presenti numerosi mitocondri ed ogni mitocondrio contiene multiple copie di mtDNA
- Eteroplasmia: coesistenza nella stessa cellula o nello stesso tessuto di mtDNA “wild type” e mutante
- Effetto soglia: una quantità minima di mtDNA mutante é necessaria per causare un deficit e, quindi, espressione fenotipica della mutazione
- Segregazione mitotica: ad ogni divisione cellulare la proporzione di mtDNA mutato può cambiare influenzando il fenotipo.
- Trasmissione matrilineare: al momento della fertilizzazione tutto il mtDNA deriva dall'oocita