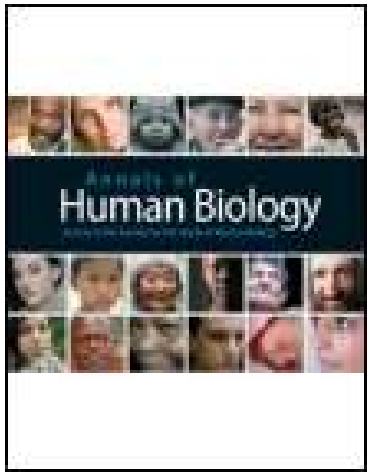
Il genoma

Il genoma Umano

- Nel corpo umano ci sono circa 3.72×10^{13} (37200 miliardi) di cellule.
- Ci sono due metri di DNA in ogni cellula umana.
- Ci sono 3 miliardi di nucleotidi x 2 in ciascuna cellula
- Se tutto il DNA di un corpo umano venisse messo in fila coprirebbe la distanza terra-sole 1000 volte.



An estimation of the number of cells in the human body
Annals of Human Biology, Volume 40, Issue 6, 2013

Il genoma Umano

- Il genoma è la lista completa delle informazioni necessarie a creare e far funzionare un individuo.

- La porzione del genoma responsabile della diversità nella popolazione umana è circa 0,1%.

- Il DNA umano è per il 98% identico a quello dello scimpanzé.

- Il 97% del genoma umano non contiene informazione attualmente riconoscibile.

Goals del progetto genoma umano:

-Determinare la sequenza dell'intero genoma umano

-Identificare tutti i geni

-Immagazzinare queste informazioni su database

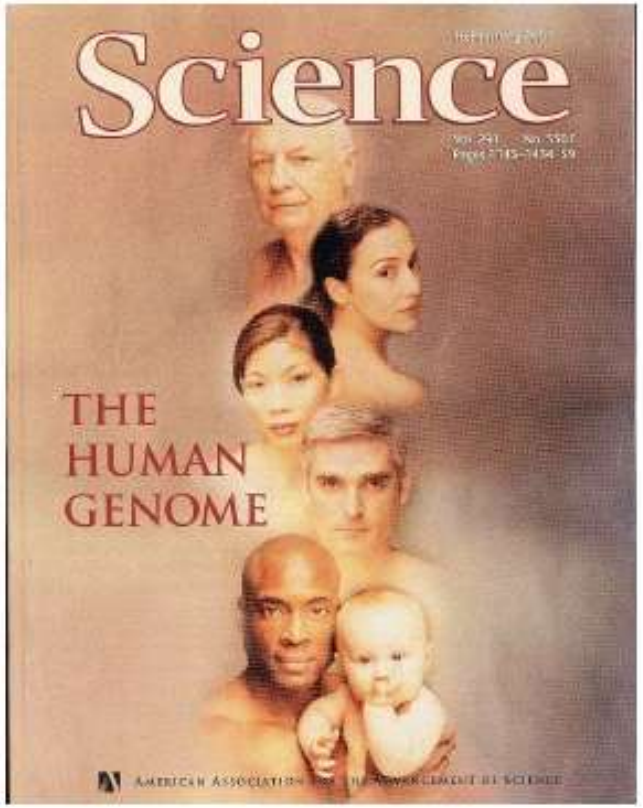
-Sviluppare programmi per l'analisi di questi dati

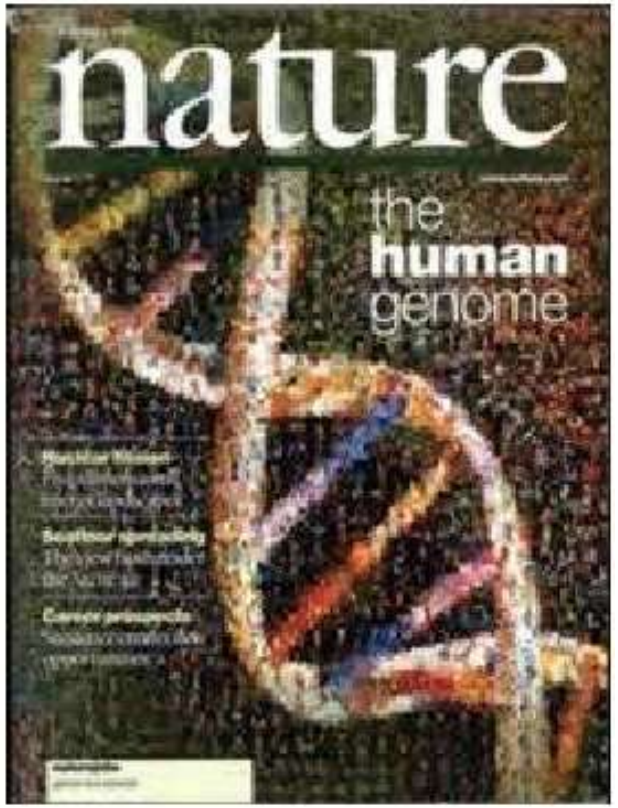
-Stabilire principi etici e legali per l'utilizzo di questi dati

Consorzio Pubblico internazionale HGP

(USA, UK, Francia Germania, Cina e altri)

Celera Genomics di Craig Venter





Nel 2003, dopo 13 anni dalla istituzione del consorzio HGP, si dà l'annuncio ufficiale del sequenziamento del genoma umano

Le sequenze dei cromosomi interi vengono ricostruite a partire dalle sequenze di centinaia di migliaia di frammenti di DNA, normalmente di lunghezza compresa fra 500 e 800 pb

Si usano 2 strategie generali per la frammentazione e la ricostruzione:

Approccio gerarchico

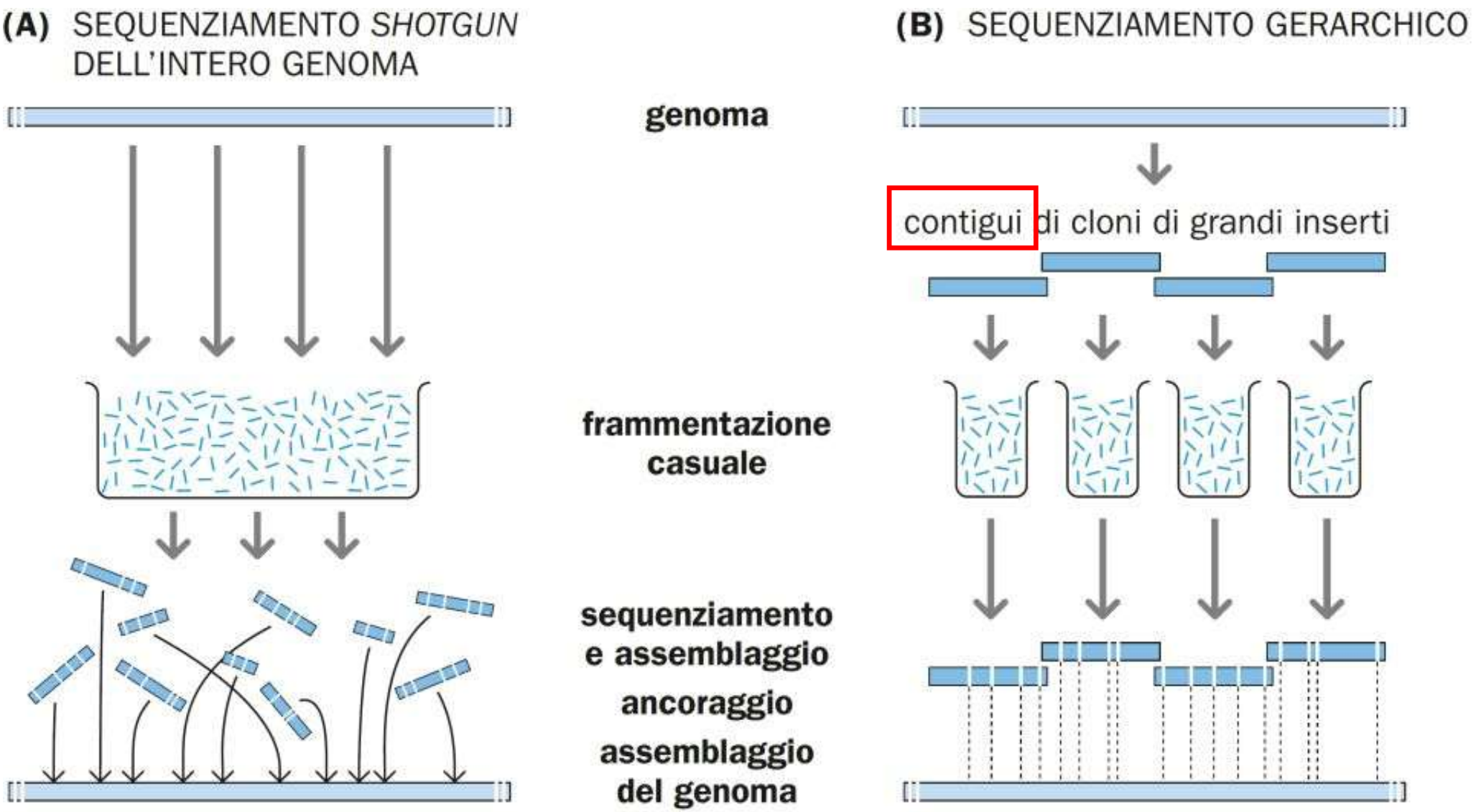
Approccio globale (shotgun)

Le due strategie sono reciprocamente complementari

Differenze:

Con il metodo shotgun (seguito dalla Celera Genomics) si frammenta semplicemente il genoma in piccole unità sequenziabili e si affida ad algoritmi del computer la ricostruzione dell'ordine dei frammenti.

Nel sequenziamento gerarchico (seguito dal consorzio HGP) il primo passaggio è quello di creare grandi pezzi di DNA il cui ordine è già stabilito da una precedente mappatura. La sequenza è il secondo passaggio.


(A) SEQUENZIAMENTO *SHOTGUN* DELL'INTERO GENOMA
(B) SEQUENZIAMENTO GERARCHICO
genoma
contigui di cloni di grandi inserti
frammentazione casuale
sequenziamento e assemblaggio
ancoraggio
assemblaggio del genoma

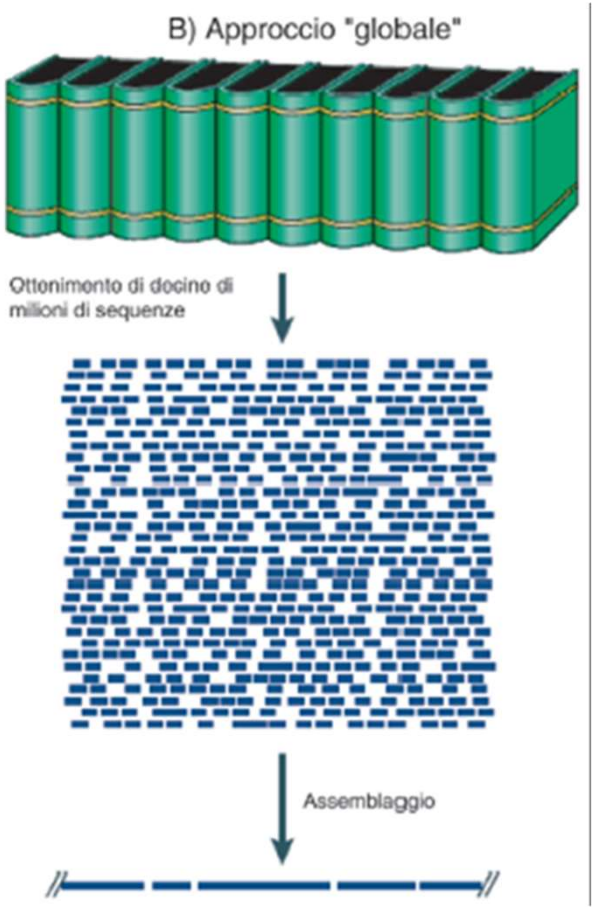
*Contig:

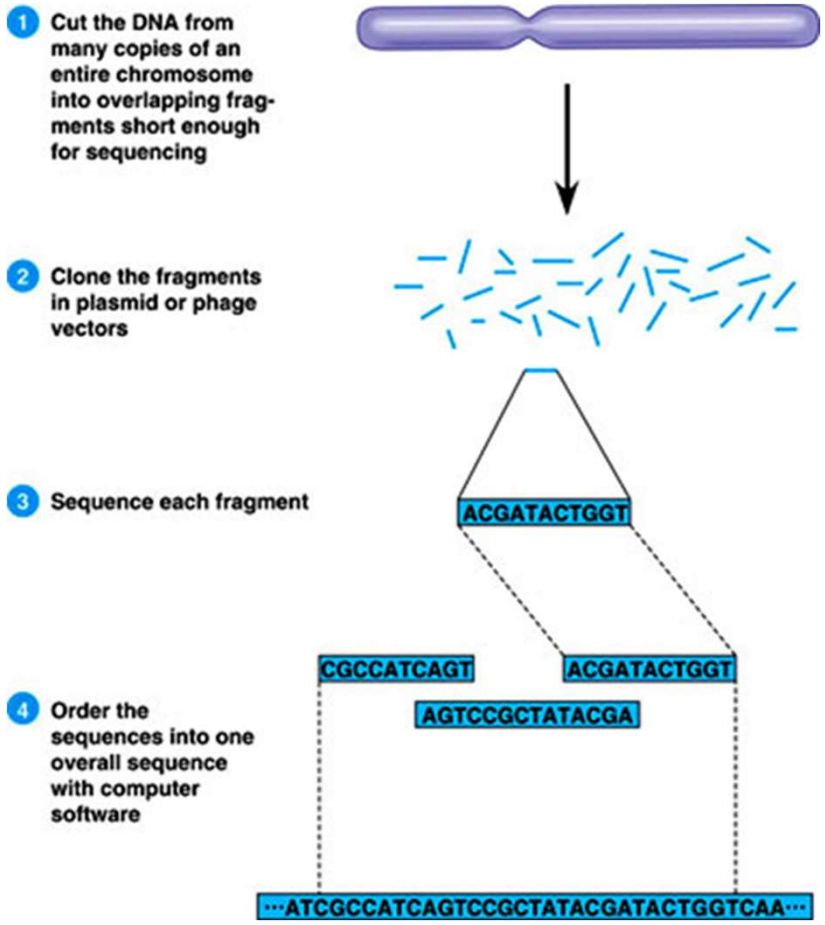
Un contig è un gruppo di frammenti di DNA parzialmente sovrapposti che insieme vanno a costituire una regione più o meno ampia del genoma.

Nell'approccio globale il Contig è il risultato della sovrapposizione di pezzi di DNA sequenziato.

Nell'approccio gerarchico il Contig è il risultato della sovrapposizione di frammenti tramite mappatura. Il sequenziamento è successivo.

Approccio globale (shotgun)

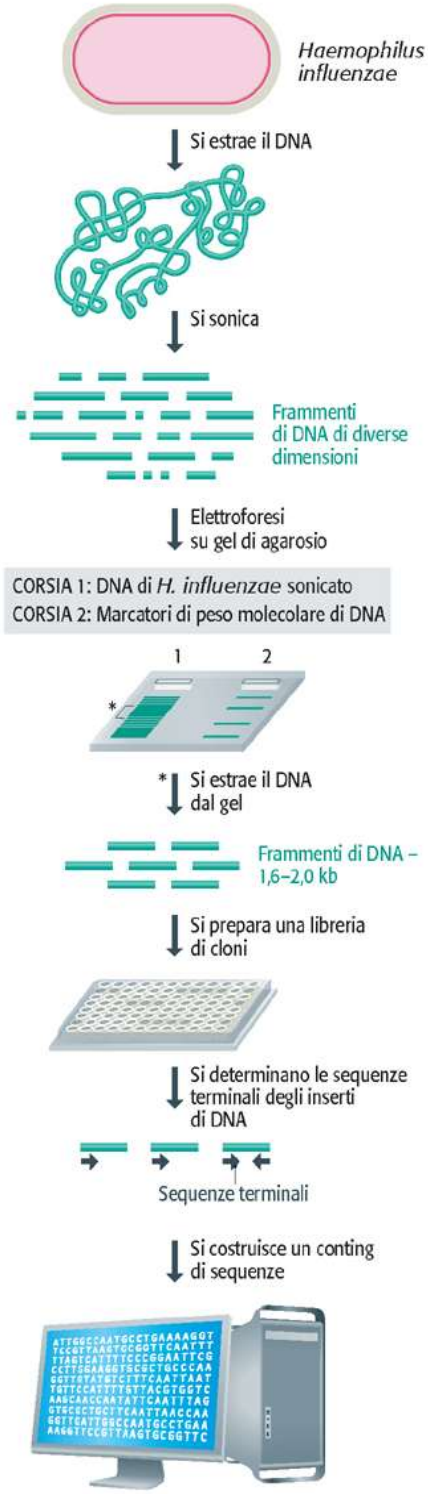




Il DNA è frammentato ed ogni frammento è sequenziato. L'intera sequenza è assemblata cercando sovrapposizioni tra le singole sequenze.

Il potenziale dell'approccio shotgun è stato confermato dal sequenziamento di Haemophilus influenzae

Nel 1995 è stata pubblicata la sequenza di 1830 kb del genoma del batterio H. Influenzae.



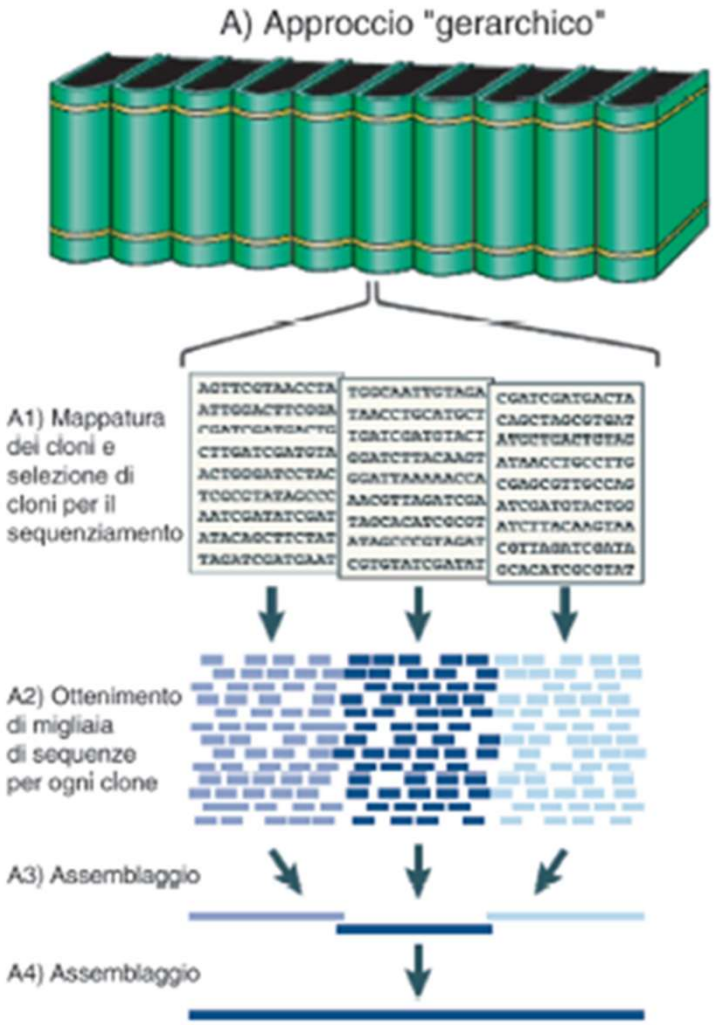
28643 reazioni di sequenza sono state effettuate da 8 persone utilizzando in media 14 DNA sequencer al giorno per 3 mesi.

L'assembling di 24304 frammenti in **210 contigs*** ha richiesto **30 ore di processamento** continuo su un computer SPARCenter 2000 con 512 Mb di RAM

Il costo stimato è stato di **0.48 centesimi di dollaro/base** sequenziata.

Se la tecnologia applicata per il sequenziamento del Genoma Umano (2000-2001) venisse di nuovo applicata al genoma dell'*Haemophilus influenzae* il suo genoma potrebbe essere nuovamente sequenziato e assemblato in poche ore.

Approccio gerarchico



Le strategie gerarchiche sono state ideate alla fine degli anni 80 quando i reagenti erano molto costosi e i computer non erano ancora abbastanza potenti per elaborare sequenze intere ottenute con lo shotgun.

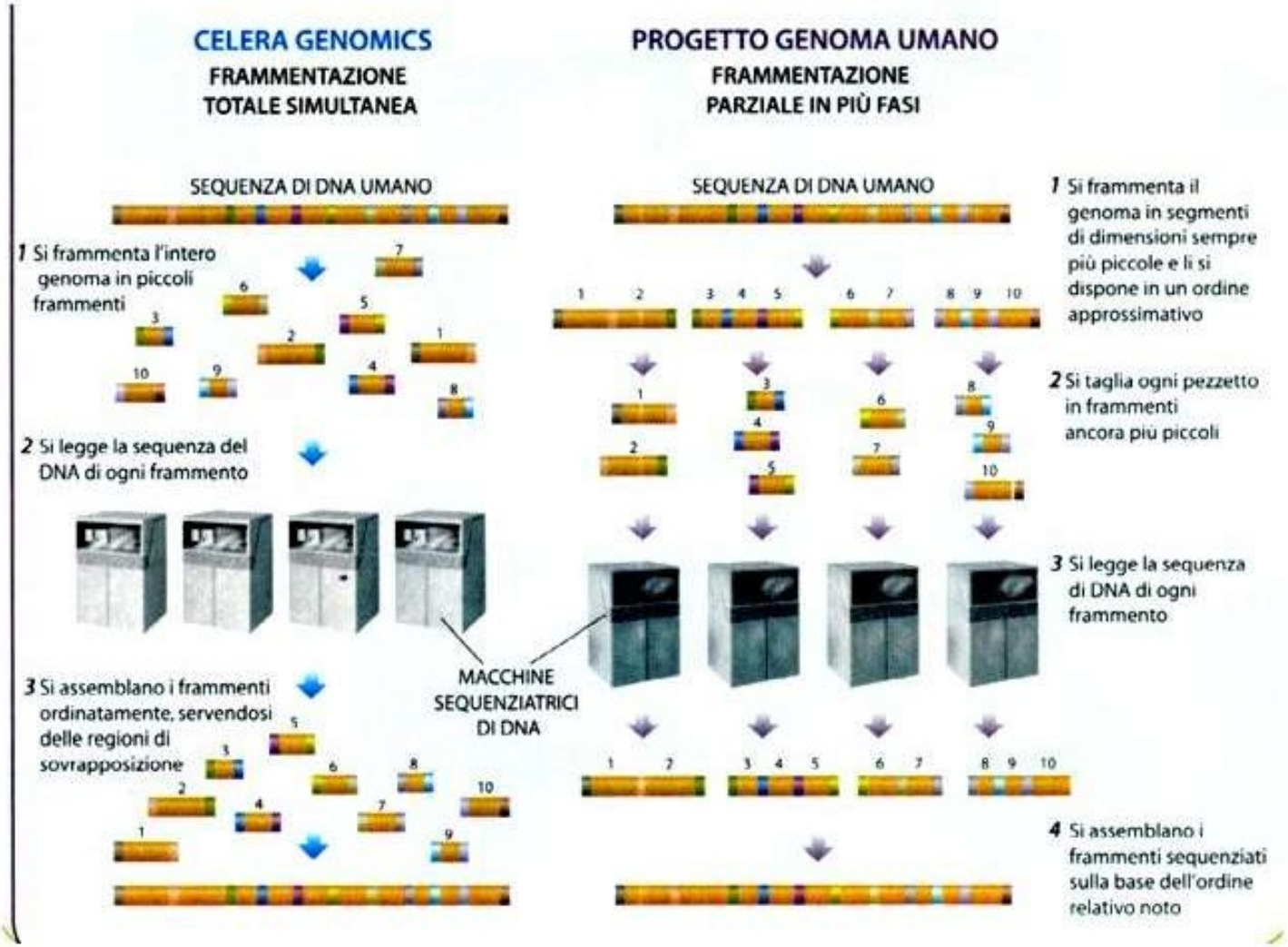
Metodologia: il metodo gerarchico si applica a grossi frammenti di DNA (50-200 Kb) clonati in vettori di derivazione fagica, le cui posizioni relative sono conosciute prima di iniziare il sequenziamento (mappatura).

Vantaggi dell'approccio gerarchico: favorisce la ricostruzione di mappe fisiche e genetiche ad alta risoluzione e permette a gruppi di lavoro di tutto il mondo di formare consorzi e lavorare insieme senza rischiare di ripetere le stesse ricerche
(un gruppo = un cromosoma)

Confronto delle due strategie adottate per il sequenziamento del genoma umano

Le strategie seguite dal Consorzio Pubblico (sequenziamento **clone per clone** e dalla Celera Genomics (sequenziamento **shotgun**) sono diverse e allo stesso tempo complementari.

C. Venter F. Collins



La Celera Genomics ha potuto allineare i differenti scaffolds sia avvalendosi dei dati ottenuti dal consorzio pubblico sia mediante l'uso dei siti di sequenze-etichetta STS (ibrido tra shotgun puro e strategia di mappatura-sequenziamento).

Importanza dei progetti genomici

- Identificazione e Descrizione di ogni singolo gene di un genoma (anche di funzione ignota)
- Identificazione di regioni regolative
- Isolamento e utilizzazione geni (responsabili di malattie ereditarie, prodotti proteici di rilevanza industriale)
- Individuazione di ruoli per quello che chiamiamo DNA non codificante attraverso la comparazione di tali regioni del DNA in genomi provenienti da organismi diversi (genomica comparata)

Applicazioni della ricerca genomica

- Medicina molecolare
  - Diagnosi delle malattie
  - Individuazione della predisposizione genetica per una malattia
  - Creazione di farmaci specifici basati sul profilo genetico del paziente
  - Terapia genica
  - Creazione di farmaci basati su informazioni molecolari

- Bisogna PERO' tenere a mente che sebbene sia prassi comune parlare della sequenza del genoma umano ci sono in realtà molte sequenze perché

ogni individuo, eccetto i gemelli identici, ha la propria versione.

- Le differenze tra individui diversi sono dovute principalmente a

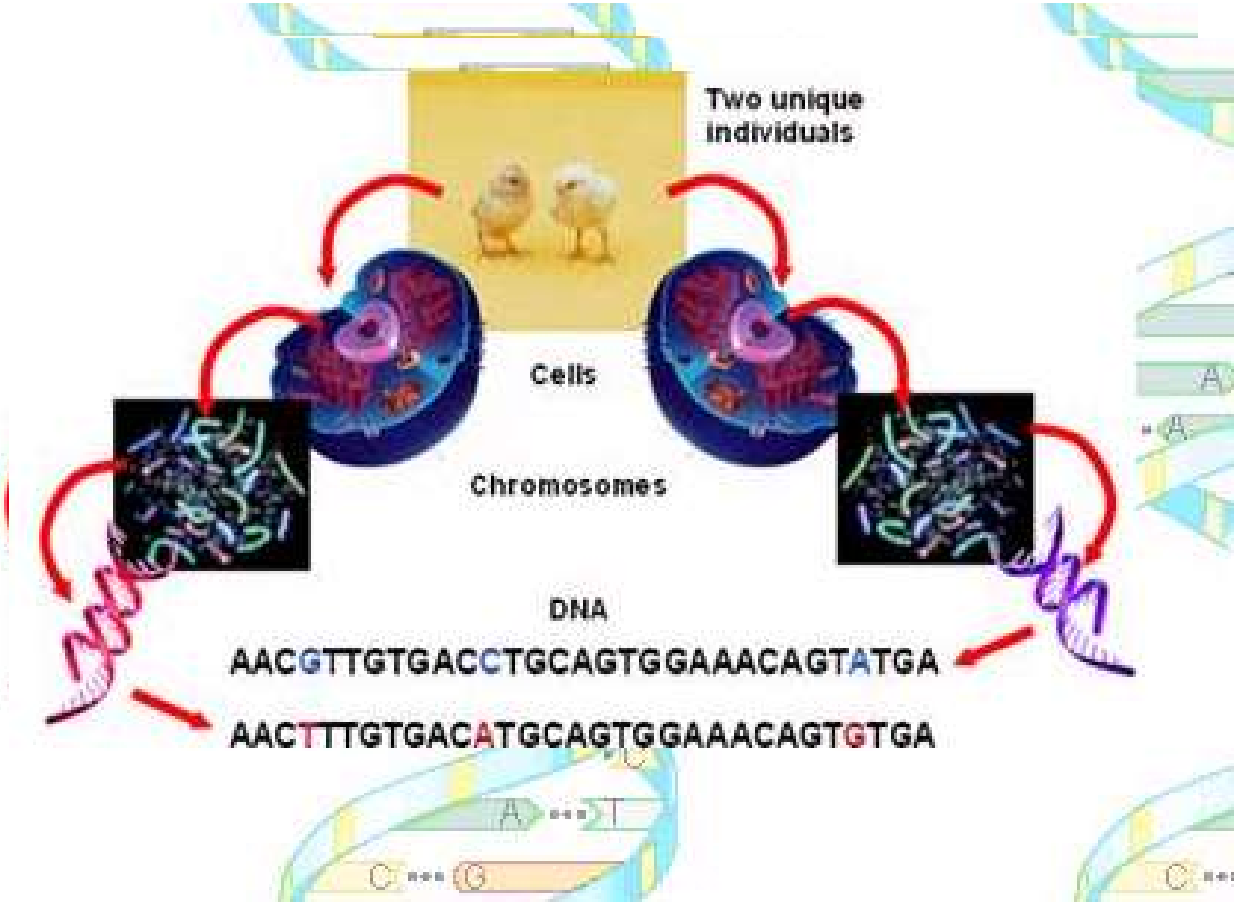
polimorfismi di singola base (SNPs)

- Tipi di variazione genetica

1) **SNP**

2) **CNV** = variazione nel numero di copie di un gene o di sequenze ripetute (satelliti)

3) **Indel** = inserzioni/delezioni di nucleotidi, raramente a livello di geni (e.f.), più spesso in zone intergeniche (e.g.);

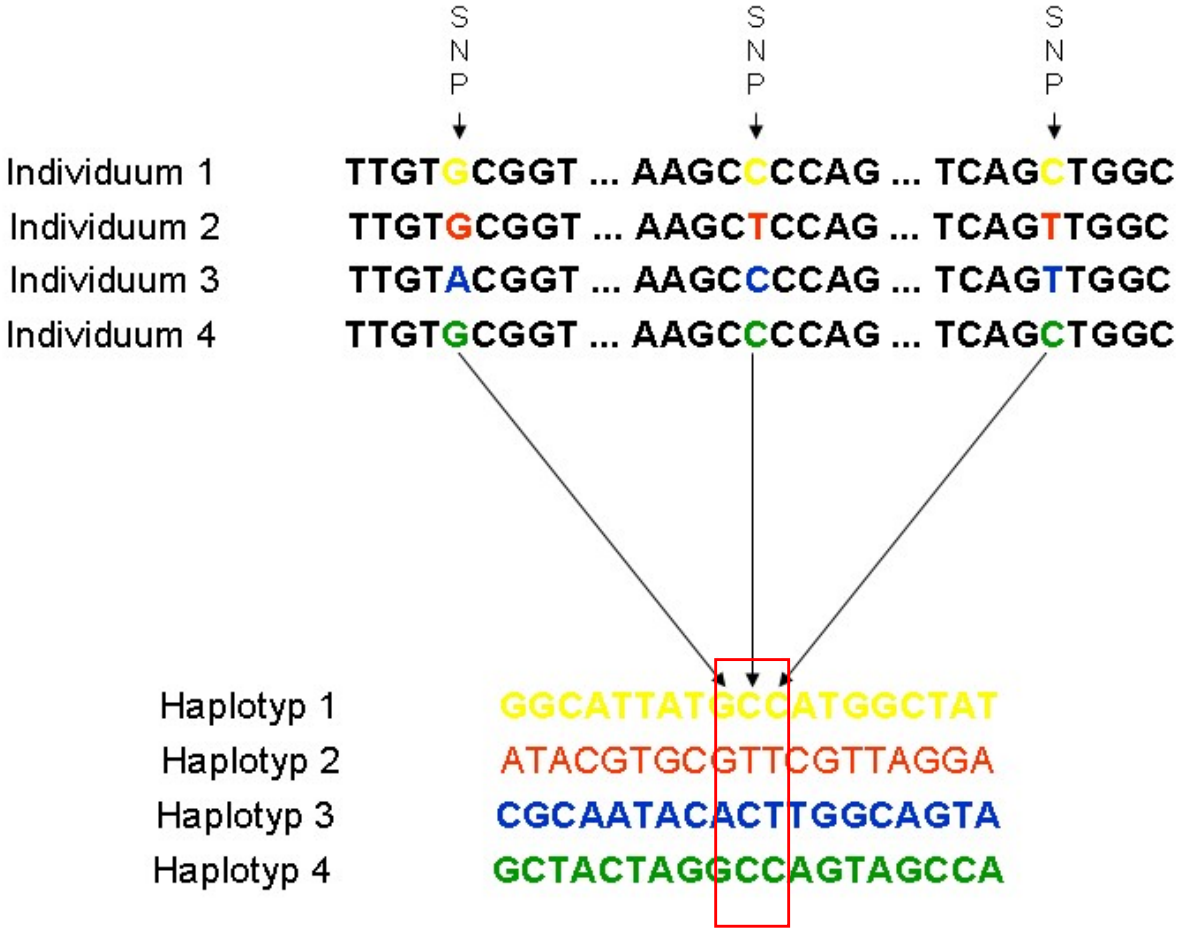


Sono stati identificati più di **3 milioni di SNPs**, una media di circa 1 ogni 1000 coppie di basi (alcuni autori valutano le differenze 1/300).

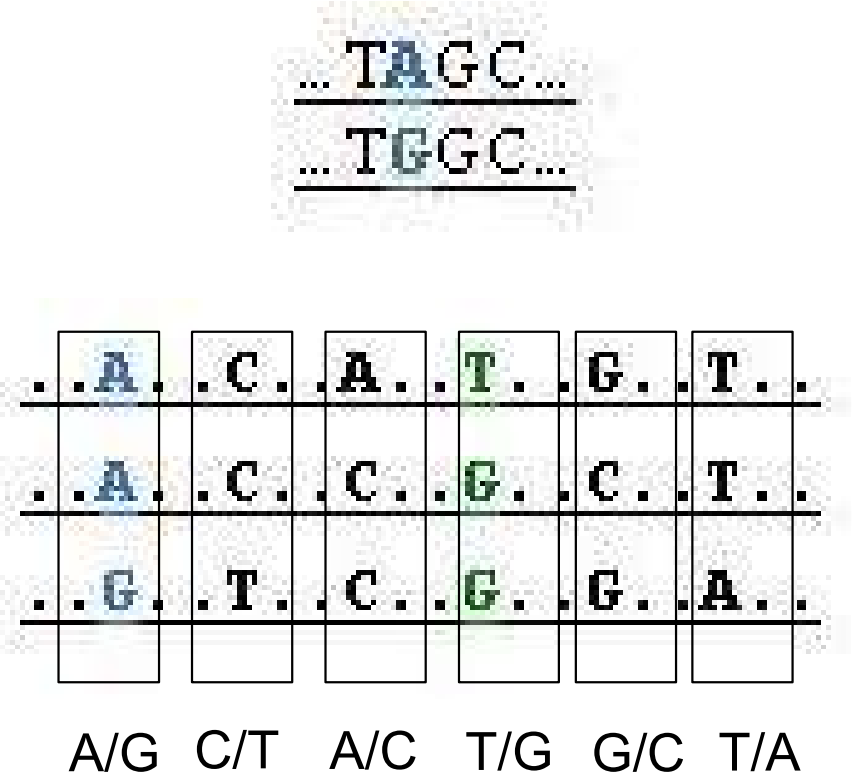
Essi costituiscono il 90% delle variazioni genetiche individuali.

La combinazione di un certo numero di snp costituisce un **aplotipo**.

Con il termine **aplotipo** si definisce la combinazione di varianti alleliche lungo un cromosoma o segmento cromosomico contenente loci in linkage disequilibrium, cioè strettamente associati tra di loro e che, in genere, vengono ereditati insieme.



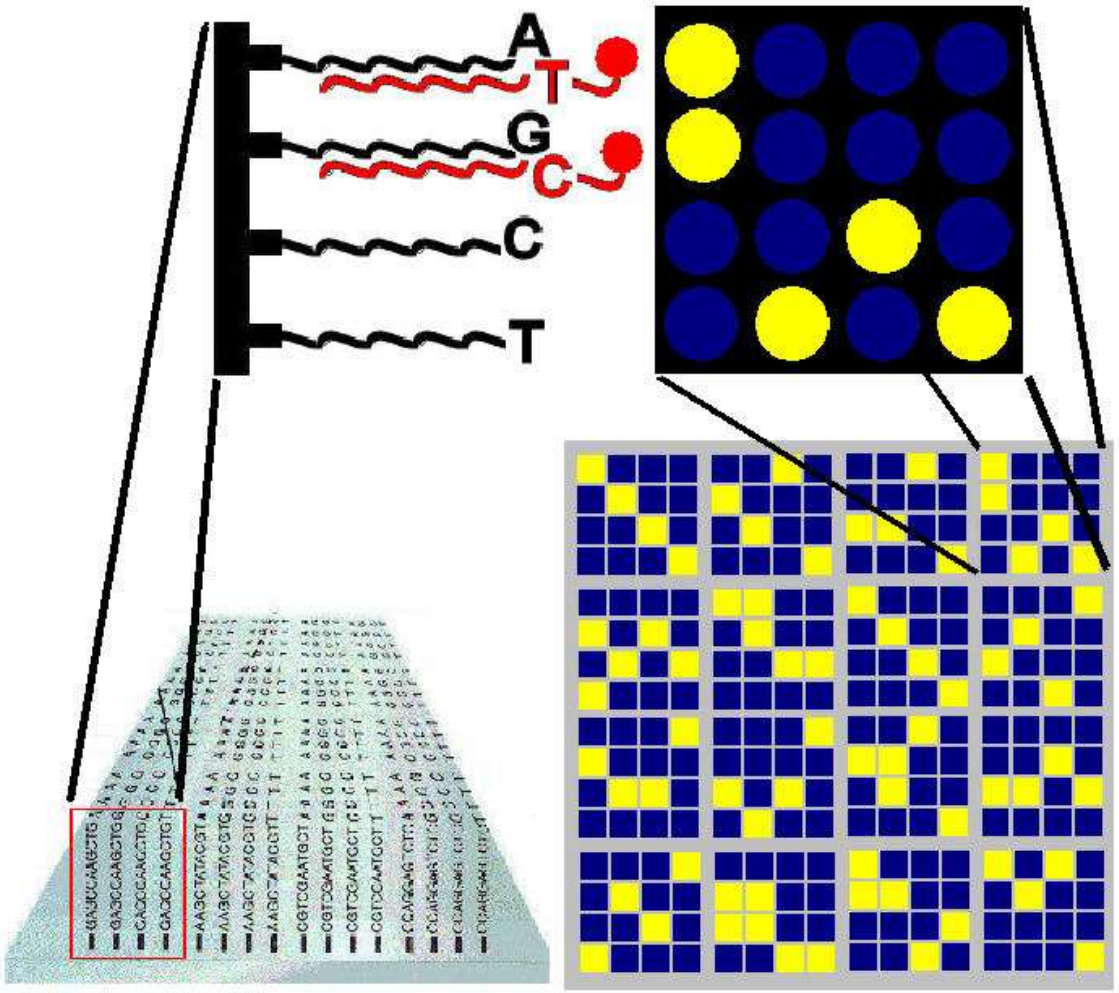
Ogni snp ha solo due forme alleliche



3 diversi **aplotipi**.

Diversi aplotipi sono relativi a corrispondenti regioni cromosomiche nella popolazione

Più che di aplotipi semplici oggi possiamo parlare di veri e propri **profili di SNPs** grazie allo sviluppo di nuove tecnologie per la miniaturizzazione e l'automatizzazione di questo tipo di analisi basata sugli esperimenti con "microchips" di DNA che permettono l'analisi simultanea di milioni di SNPs.



Le sonde sul supporto sono allele-specifiche e le loro posizioni sono stabilite

Nella figura sono riportate le analisi su Chip relativi a 4 diversi snp (linee verticali)*

*ogni sonda allele/specifica (disposta sulla linea verticale) termina con uno dei 4 nt anche se ogni snp ha solo due alleli

Il genotipo di un organismo potrebbe non definire in modo univoco il suo aplotipo

(nell'esempio è riportata l'identificazione di un aplotipo costituito da due snp: A/T e G/C)

	AA	AT	TT
GG	AG AG	AG TG	TG TG
GC	AG AC	AG TC or AC TG	TG TC
CC	AC AC	AC TC	TC TC

L'individuazione precisa dell'aplotipo nei doppi eterozigoti è possibile solo con il sequenziamento

Molti SNPs non hanno effetto sulla funzionalità del genoma mentre altri sì.

Per esempio 60.000 SNPs si trovano all'interno di geni ed hanno un impatto sulla loro attività, determinando quelle differenze che rendono ognuno di noi un organismo unico.

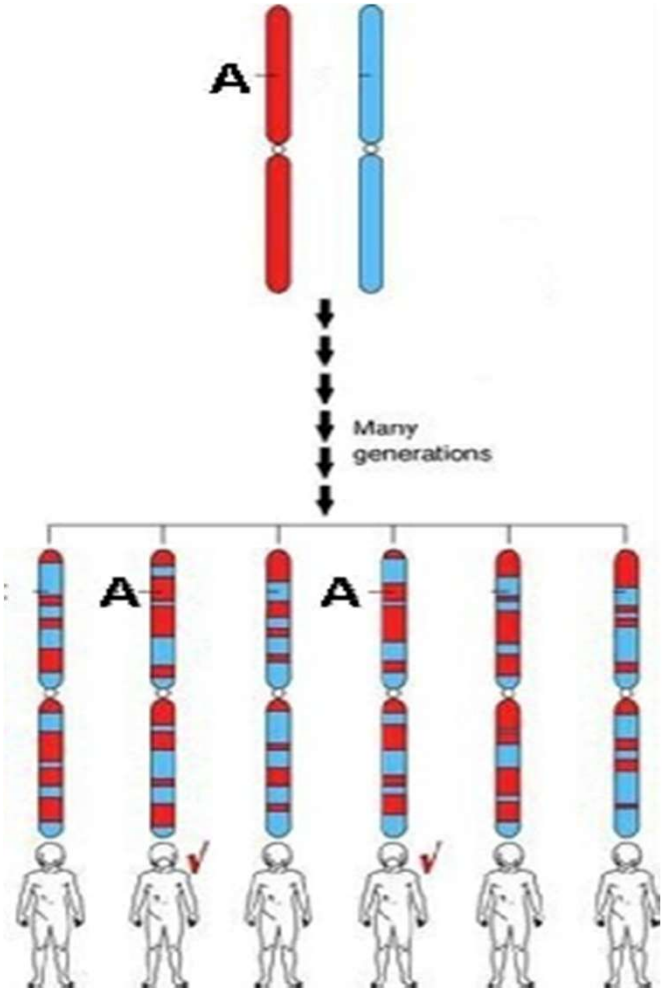
Gli SNPs che non si trovano in sequenza codificante possono avere consequenze sullo splicing o sul legame di fattori di trascrizione e quindi influenzare l'espressione genica

Lo studio degli SNP è molto utile poiché variazioni anche di singoli nucleotidi possono influenzare lo **sviluppo di patologie o la risposta a patogeni, ad agenti chimici, a farmaci**.

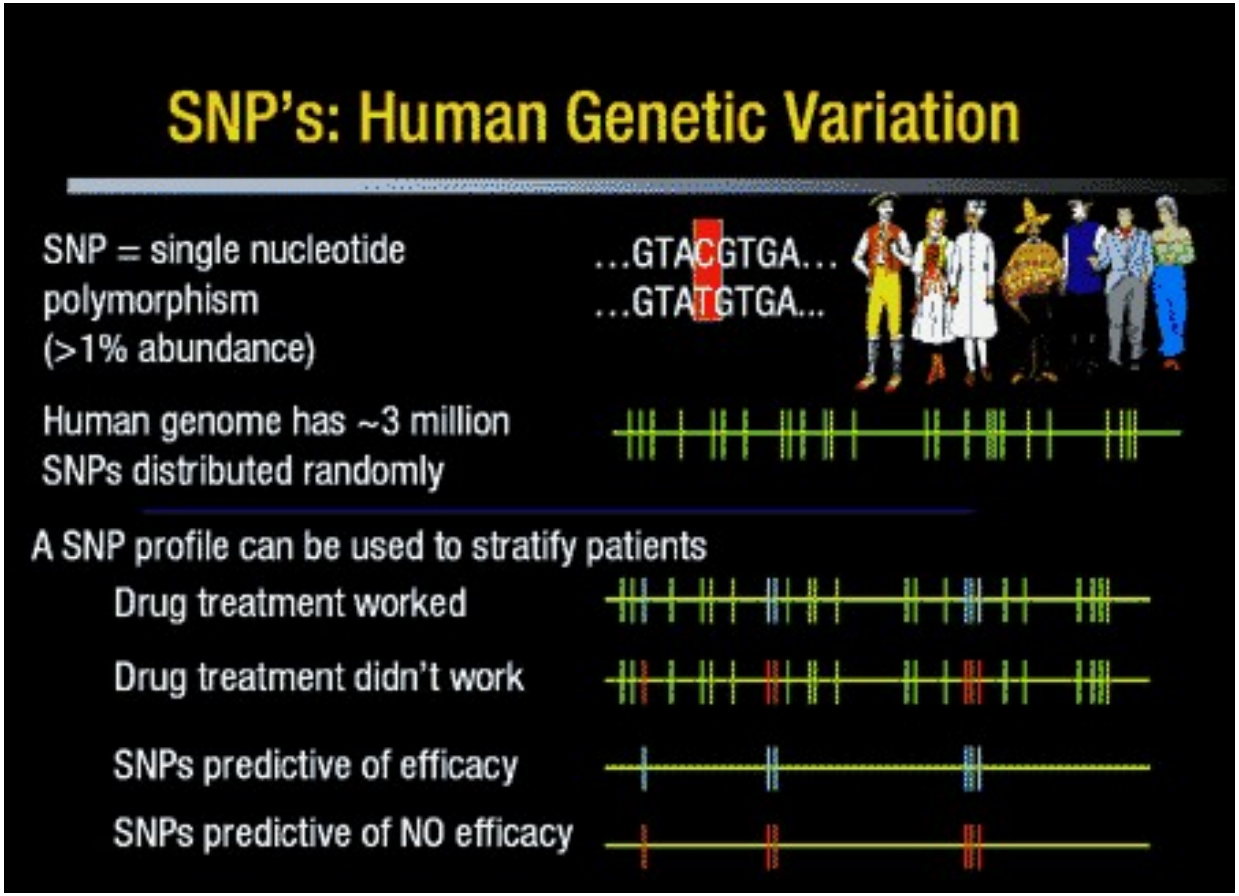
Gli SNPs possono avere una grande importanza nello **sviluppo di nuovi farmaci e nella pianificazione di protocolli terapeutici**: gli SNPs presenti nel gene responsabile della metabolizzazione di un farmaco consentono di predire l'effetto che esso potrà avere su quell'individuo.

Qual è l'origine di un aplotipo?

A partire da due cromosomi originali mediante RICOMBINAZIONE nel corso delle generazioni si otterranno cromosomi differenti.



L'associazione statistica tra loci si manifesta in assenza di ricombinazione tra i loci stessi. Per quanto riguarda gli autosomi (cromosomi non sessuali) e le regioni pseudoautosomiche dei cromosomi sessuali questo può essere dovuto alla **vicinanza fisica** tra i loci considerati e **all'assenza di hot-spot** di ricombinazione tra di loro.
Invece gli alleli della regione non ricombinante del cromosoma Y **(NRY)** sono sempre associati a formare aplotipi, così come gli alleli del genoma mitocondriale **(mtDNA).**
Infatti queste due porzioni del genoma non ricombinano, essendo ereditate con modalità uniparentali, paterna la prima, materna la seconda.
Aplotipi differenti sono generati da un aplotipo ancestrale per effetto della mutazione ai singoli loci.



Individuazione di specifici profili di SNPs in pazienti sensibili o resistenti ad una terapia **(Farmacogenetica)**

Gli SNPs e la GENETICA FORENSE

Normalmente utilizzati i polimorfismi minisatelliti, polimorfismi microsatelliti (STR) e polimorfismi di sequenza (SNPs).

Gli SNPs suscitano un particolare interesse nella genetica forense per varie ragioni:
-alta automatizzazione
- sufficienti amplificati di PCR minori di 100 coppie di base (campioni danneggiati)

Ma:

Gli SNPs sono marcatori di alleli, ovvero esistono solo due alleli possibili per ogni coppia di loci. Pertanto, per conseguire un elevato potere di discriminazione si richiede l'analisi di un maggior numero di marcatori. Si è stimato che è necessario analizzare circa un centinaio di SNPs per ottenere probabilità maggiori di coincidenza con i 13 STRs del CODIS*

*Combined ***DNA*** Index System: un ***DNA*** database creato dall'FBI.

1) La quantità totale di DNA di un organismo non correla con la complessità dell'organismo (paradosso valore C)

2) Il numero di geni di un organismo non correla con la complessità dell'organismo

3) Il numero di geni non correla con la quantità di DNA

1) La quantità totale di DNA di un organismo non correla con la complessità dell'organismo (paradosso valore C)

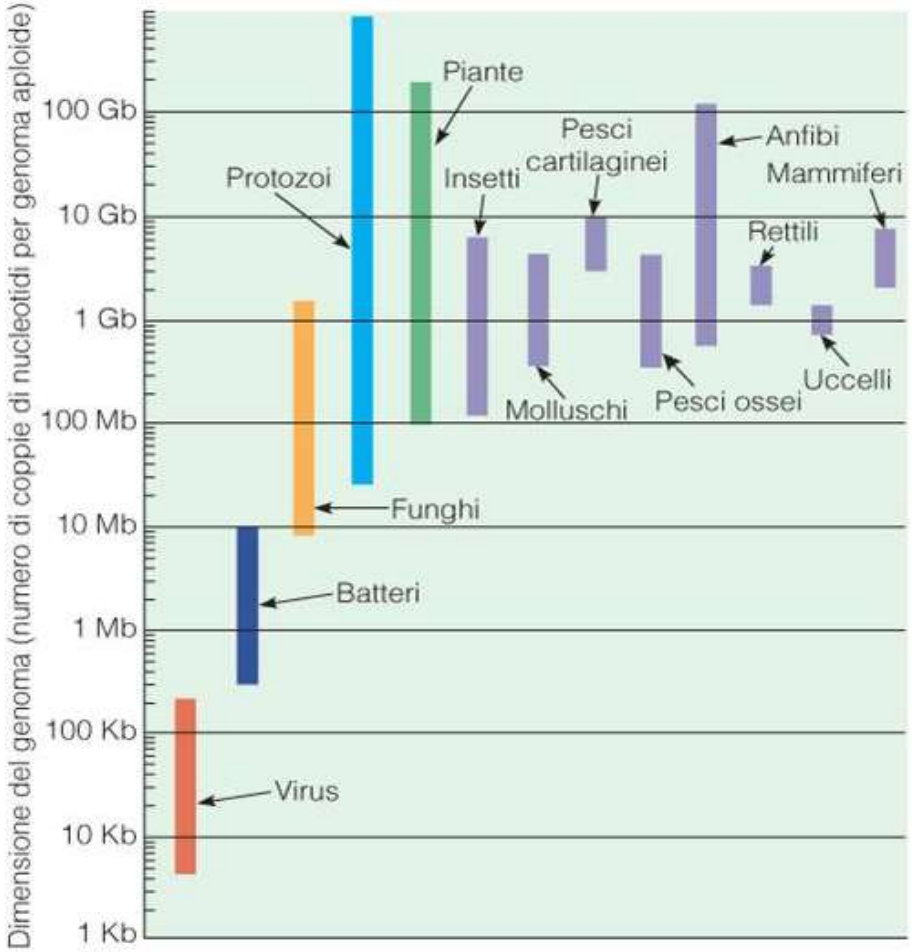


Figura 18-11 **Correlazione fra dimensione del genoma e tipo di organismo.** Per ogni gruppo di organismi mostrato, la barra rappresenta la variabilità approssimativa della dimensione del genoma, misurata come numero di coppie di nucleotidi per genoma aploide. Lo stesso colore (viola) è stato utilizzato per indicare i vari gruppi che appartengono al regno animale.

Dimensioni dei genomi eucariotici

	size (Mb)
Fungi	
Saccharomyces cerevisiae	**12.1**
Aspergillus nidulans	25.4
Protozoa	
Tetrahymena pyriformis	190
Invertebrates	
Locusta migratoria (locust)	5000
Bombyx mori (silkworm)	490
Drosophila melanogaster	**180**
Caenorhabditis elegans	**97**
Vertebrates	
Mus musculus (mouse)	3300
Homo sapiens	**3200**
Takifugu rubripes (pufferfish)	400
Plants	
Fritillaria assyriaca (fritillary)	120 000
***Triticum aestivum* (wheat)**	**16 000**
Pisum sativum (pea)	4800
***Zea mays* (maize)**	**2500**
Oryza sativa (rice)	430
Arabidopsis thaliana (vetch)	125

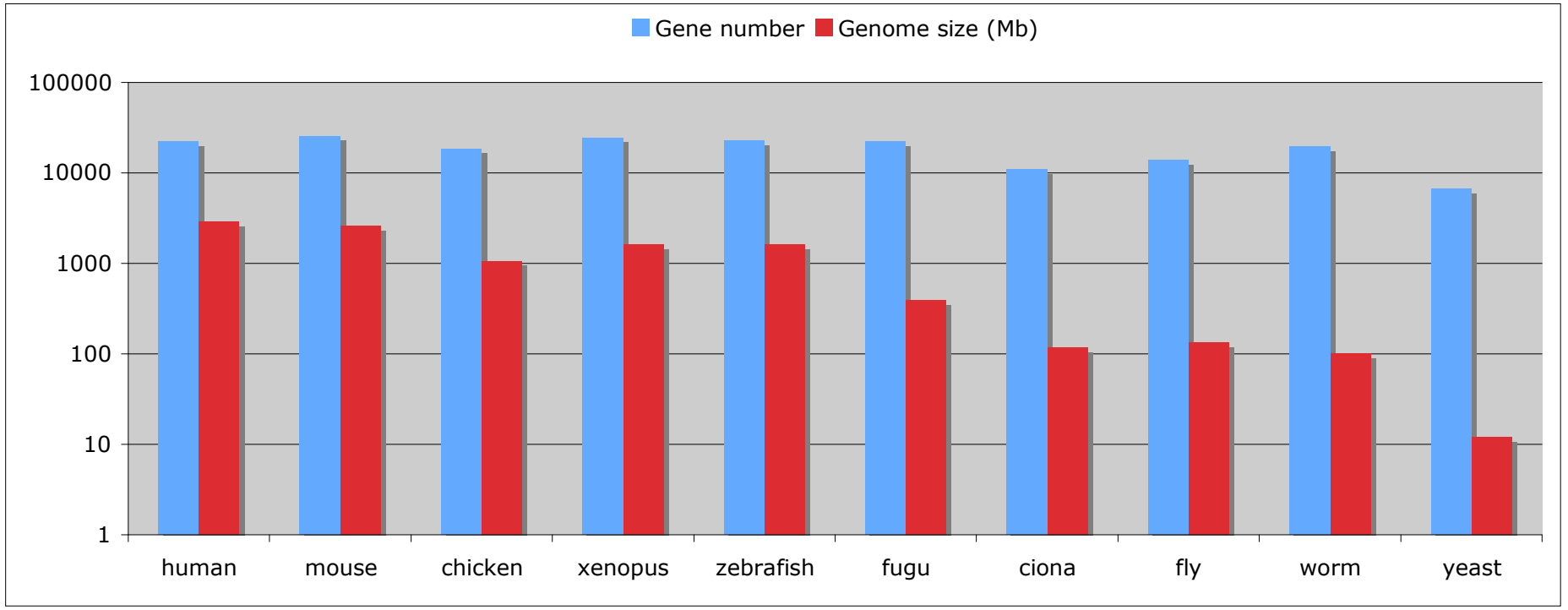
2) Il numero di geni di un organismo non correla con la complessità dell'organismo

Dalle analisi comparative dei genomi non ci sono rivelazioni sorprendenti su cosa rende un uomo diverso da uno scimpanzé

Sulla base del **numero di geni** noi siamo solo 3 volte più complessi del moscerino della frutta e solo due volte più complessi del verme microscopico Caenorhabditis elegans

3) Il numero di geni non correla con la quantità di DNA

Assenza di correlazione tra numero di geni e dimensione del genoma negli eucarioti



Uomo:
Lunghezza del genoma: 3,2 miliardi di paia di basi
Numero di geni: ~20.000–21.000 geni codificanti proteine

Banana:
Lunghezza del genoma: 520 milioni di paia di basi
Numero di geni: 35.000–36.000 geni codificanti proteine

Mela:
Lunghezza del genoma: circa 650–750 milioni di paia di basi
Numero di geni: circa 57.000–63.000 geni codificanti proteine

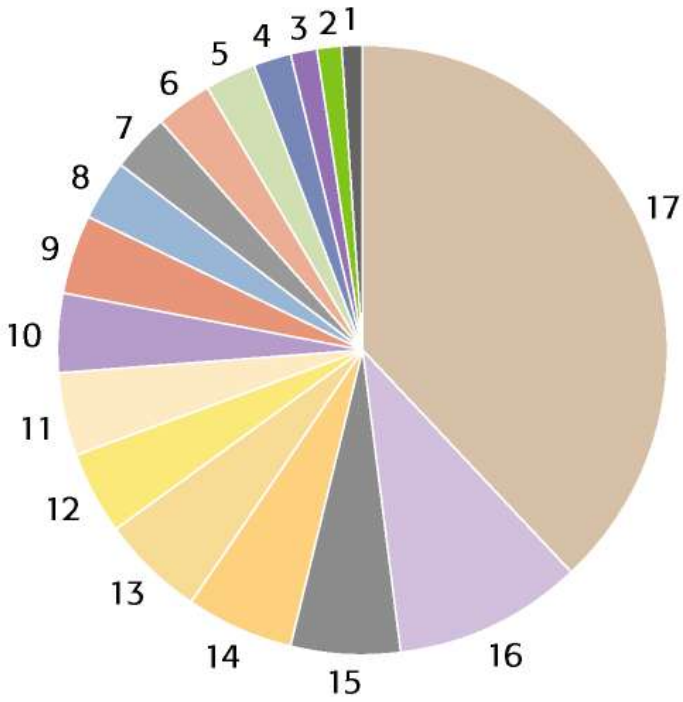
Grano tenero:
Lunghezza del genoma: circa 15–17 miliardi di paia di basi
Numero di geni: Circa 100.000–107.000 geni codificanti proteine

Il genoma dei procarioti

(a) *Escherichia coli* (un batterio comune)

Un cromosoma circolare
Dimensioni del genoma: 4,64 milioni bp
Numero dei geni: 4289
Contenuto in G + C: 51%

(b)



1. Metabolismo degli acidi grassi e dei fosfolipidi
2. Trascrizione, metabolismo dell'RNA
3. Metabolismo dei nucleotidi
4. Funzioni relative a fagi, trasposoni, plasmidi
5. Replicazione, ricombinazione, riparazione di DNA
6. Metabolismo dei composti del carbonio
7. Metabolismo degli amminoacidi
8. Altri geni con funzioni note
9. Funzioni regolative
10. Traduzione, metabolismo delle proteine
11. Metabolismo intermedio centrale
12. Funzioni di adattamento e protezione
13. Componenti strutturali della cellula, della parete, della membrana
14. Metabolismo energetico
15. Enzimi presunti
16. Proteine di trasporto
17. Geni con funzioni sconosciute

- Dimensioni: variabili da 580.000 a 7 milioni coppie di basi
- Numero di geni di solito tra 1000 e 2000 (480 Mycoplasma genitalium, 4.400 E.coli)
- Densità genica costante di circa 1 gene/1000 coppie di basi

(genoma più grande-più geni)

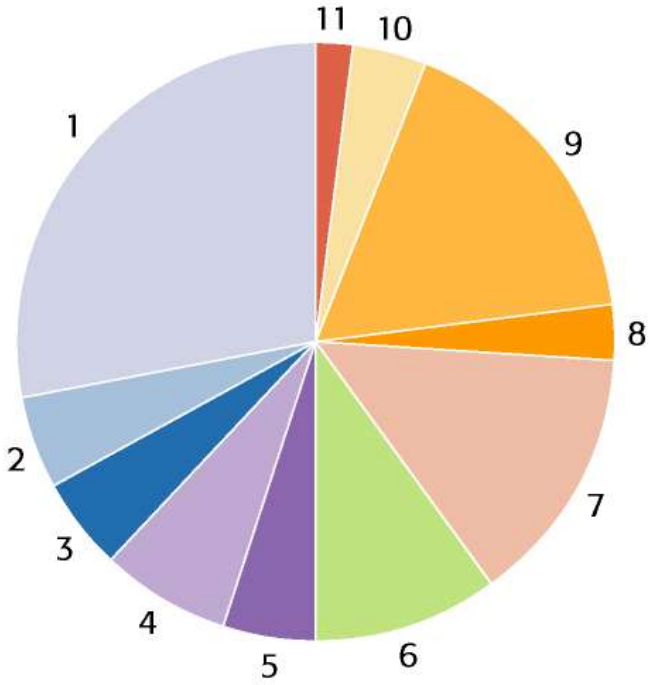
. Geni per trascrizione e traduzione numero simile tra le varie specie (anche se dimensione del genoma variabili)

. Geni per metabolismo numero variabile secondo specie

(a) *Saccharomyces cerevisiae* (lievito)

16 coppie di cromosomi lineari
Dimensioni del genoma: 12,1 milioni bp
Numero dei geni: 6100
Contenuto in G + C: 38%

(b)



1. Organizzazione cellulare e biogenesi
2. Trasporto intracellulare
3. Trasporto facilitato
4. Destinazione delle proteine
5. Sintesi proteica
6. Trascrizione
7. Crescita cellulare, divisione cellulare e sintesi del DNA
8. Energia
9. Metabolismo
10. Salvataggio cellulare
11. Trasduzione del segnale

- 12,1 milioni di coppie di basi
- 6.100 geni di cui 5.900 codificano per proteine
- Densità genica: 1 gene/2000 coppie di basi
- 239 introni
- 30% dei geni in due o più copie (ridondanza)

Caenorhabditis elegans
(verme)

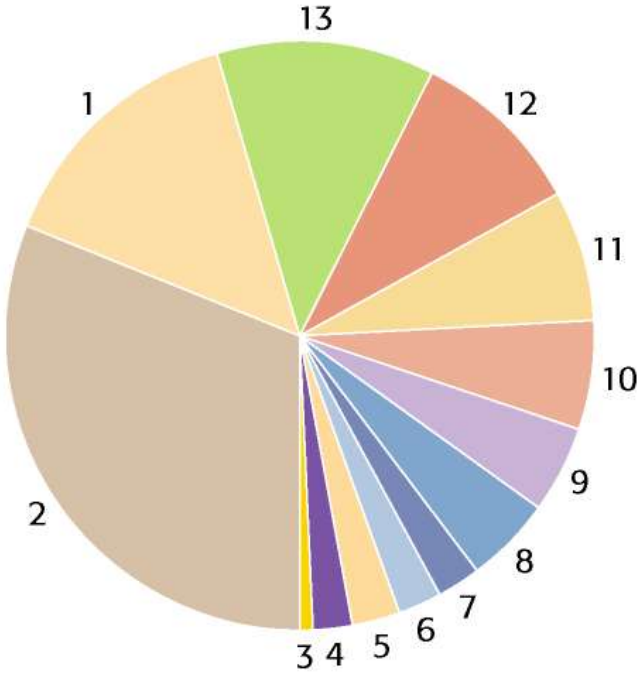
Sei coppie di cromosomi lineari
Dimensioni del genoma: 97 milioni bp
Numero dei geni: 18 266
Contenuto in G + C: 49%

- 97 milioni di coppie di basi
- 18.000 geni codificanti per proteine
- Densità 1 gene/5000 coppie di basi

(a) ***Arabidopsis thaliana*** **(erba infestante della famiglia della senape)**

Cinque paia di cromosomi lineari
Dimensioni del genoma: 167 milioni bp
Numero dei geni: 25 706
Contenuto in G + C: 47%

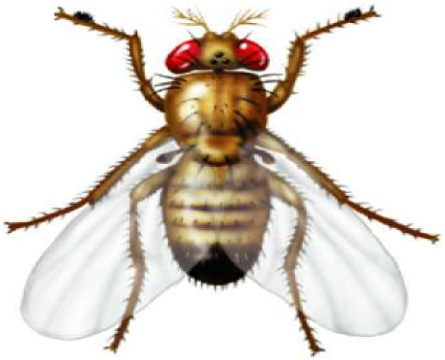
(b)



1. Metabolismo
2. Non classificati
3. Omeostasi degli ioni
4. Sintesi proteica
5. Energia
6. Trasporto facilitato
7. Biogenesi cellulare
8. Trasporto intracellulare
9. Destinazione delle proteine
10. Comunicazione cellulare e trasduzione del segnale
11. Salvataggio cellulare, difesa, morte cellulare, invecchiamento
12. Crescita cellulare, divisione cellulare e sintesi del DNA
13. Trascrizione

- 167 milioni di coppie di basi
- 25.700 geni presunti
- Densità 1 gene/6.500 coppie di basi
- 60% del genoma formato da segmenti replicati (poliploidia) (duplicazione genica importante nell'evoluzione)
- 17% geni disposti in tandem (crossing over ineguale)

Drosophila melanogaster (moscerino della frutta)



Quattro paia di cromosomi lineari

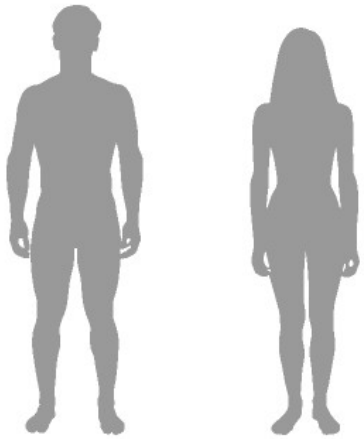
Dimensioni del genoma:
180 milioni bp

Numero dei geni: 13 338

Contenuto in G + C: 41%

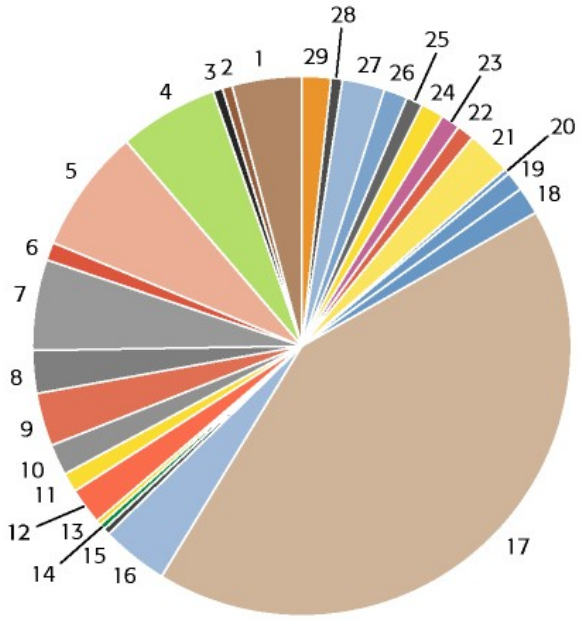
- 180 milioni di coppie di basi
- 13.000 geni
- Densità 79 geni/Megabase

(a) *Homo sapiens* (uomo)



23 paia di cromosomi lineari
Dimensioni del genoma: 3,4 miliardi bp
Numero dei geni: ~32 000
Contenuto in G + C: 41%

(b)



1. Miscellanea
2. Proteina virale
3. Proteina di trasporto/ vettore
4. Fattore di trascrizione
5. Enzima per gli acidi nucleici
6. Molecola segnale
7. Recettore
8. Cinasi
9. Molecola regolatrice
10. Trasferasi
11. Sintasi e sintetasi
12. Ossidoriduttasi
13. Liasi
14. Ligasi
15. Isomerasi
16. Idrolasi
17. Funzione molecolare sconosciuta
18. Trasportatore
19. Trasportatore intracellulare
20. Proteina legante il calcio
21. Protooncogene
22. Proteina strutturale muscolare
23. Funzione motoria
24. Canale ionico
25. Immunoglobulina
26. Matrice extracellulare
27. Proteina strutturale del citoscheletro
28. Chaperone
29. Adesione cellulare

3,2 miliardi di coppie di basi

25% trascritto in RNA 2% codifica per proteine

Un gene umano medio 27 Kb con circa 9 esoni

Splicing alternativo 1 gene 2-3 proteine 20.000 geni 96.000 proteine

Densità genica diversa nei diversi cromosomi (cr. 1 circa 3000 geni, cr. Y circa 231), in media 11 geni/Megabase

1,5-10 milioni SNP (Single Nucleotide Polimorphism)

Sequenze ripetute almeno il 50% del totale

I Genomi degli Eucarioti
1) Compattezza

I genomi degli eucarioti hanno una densità genica molto ridotta.
Geni per proteine: 2-4% dell'intero genoma.
La struttura discontinua dei geni, con introni che nei mammiferi possono raggiungere dimensioni intorno a 20-30 kb (ed oltre) e la presenza di elementi ripetuti sono alla base della scarsa compattezza.

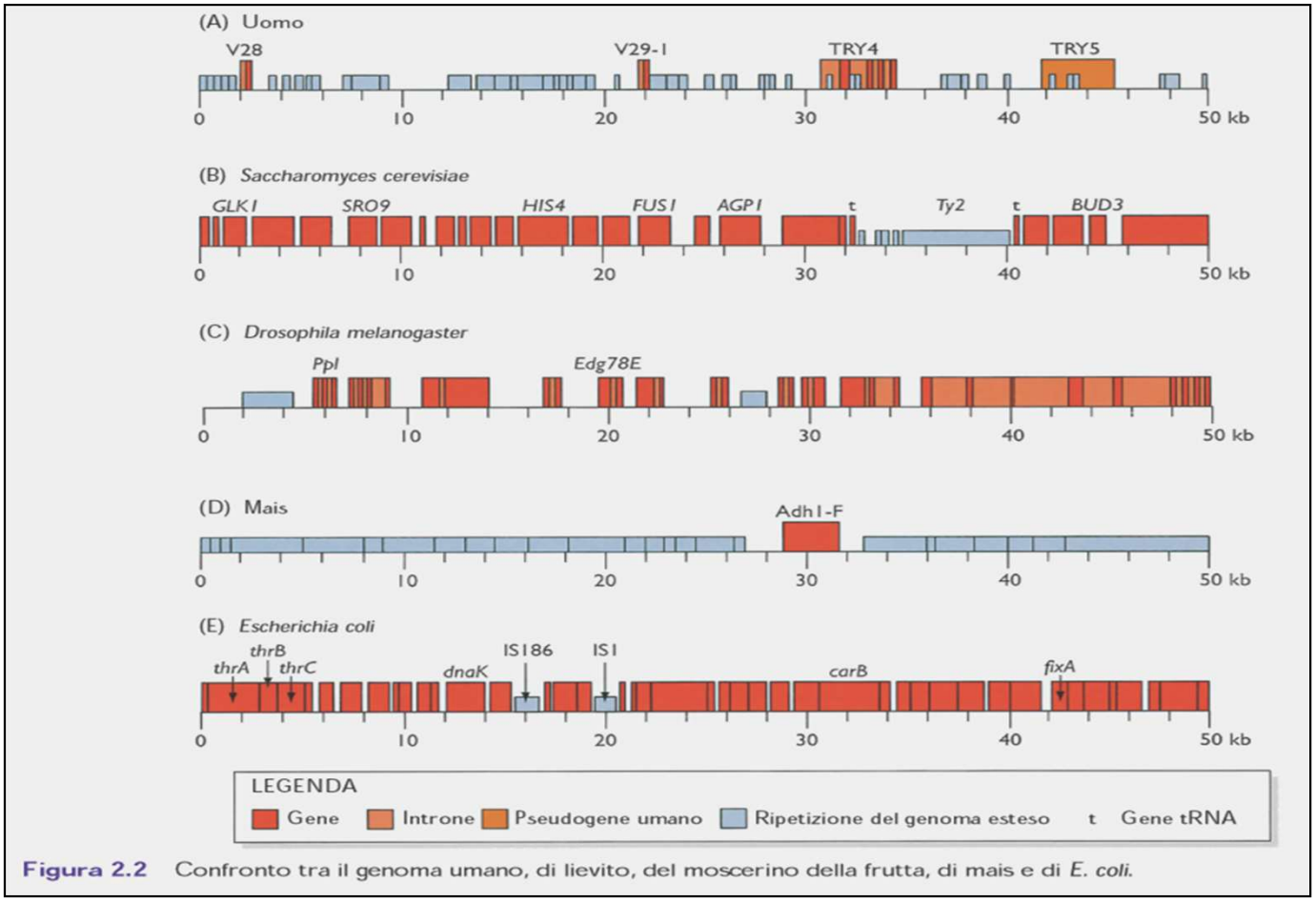


Figura 2.2 Confronto tra il genoma umano, di lievito, del moscerino della frutta, di mais e di *E. coli*.

Densità genica in un segmento di 50 kbp

Compattezza di alcuni genomi eucariotici

Proprietà del genoma	***S.cerevisiae***	***D.melanogaster***	***H. sapiens***
Densità genica (numero medio di geni per Mb)	479	79	11
Introni per gene (media)	0,04	3	9
% del genoma occupata dalle ripetizioni intersperse	3,4%	12%	44%

2) Composizione in basi dei genomi eucariotici e variazione intra-genomica della composizione in basi

Contenuto in G+C di alcuni genomi nucleari eucariotici

Higher taxon	Species	G+C%
Mammalia	*H. sapiens*	41
	M. musculus	42
Plants	*A. thaliana*	36
	O. sativa	44
Nematoda	*C. elegans*	36
Fungi	*S. cerevisiae*	38
	S. pombe	36

Rispetto ai genomi procariotici, negli eucarioti si osserva una marcata **variazione intra-genomica della composizione in basi**.

Negli eucarioti superiori e nei vertebrati a sangue caldo, sono presenti regioni genomiche a composizione in basi omogenea.

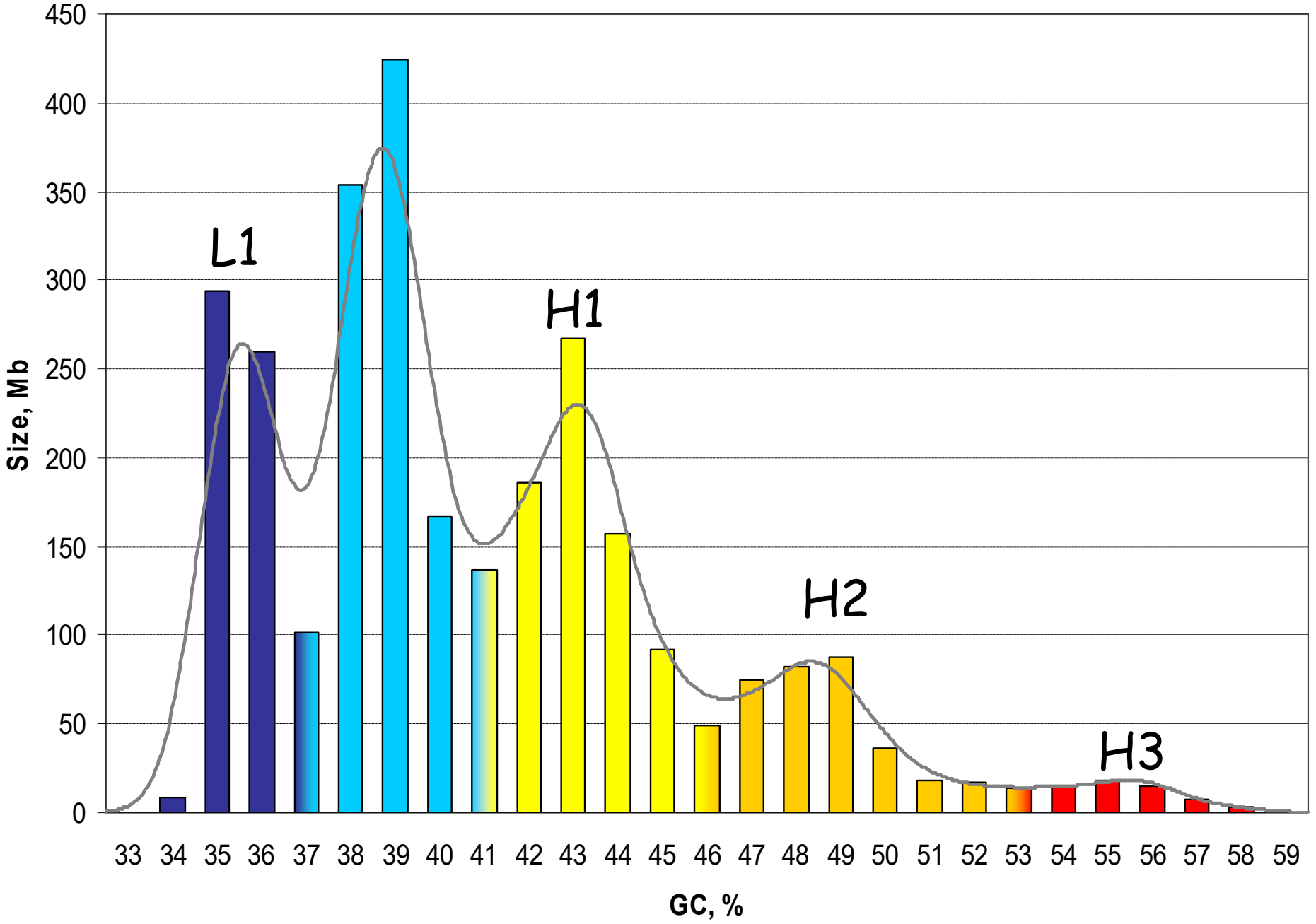
Modello delle Isocore

Secondo il **modello delle isocore** (Bernardi et al., 1985), il genoma dei vertebrati è un mosaico di segmenti di DNA, chiamati **isocore** (>>300 kbp), ciascuno caratterizzato da una propria ed omogenea composizione in basi.

Nei vertebrati a sangue caldo (mammiferi, uccelli) si osservano 5 classi differenti:
- **L1 e L2**: isocore povere in *GC* (circa il 60% del genoma)
-**H1, H2, H3**: isocore ricche in *GC*

La struttura del genoma ad isocore è correlata ad alcune proprietà del genoma nucleare

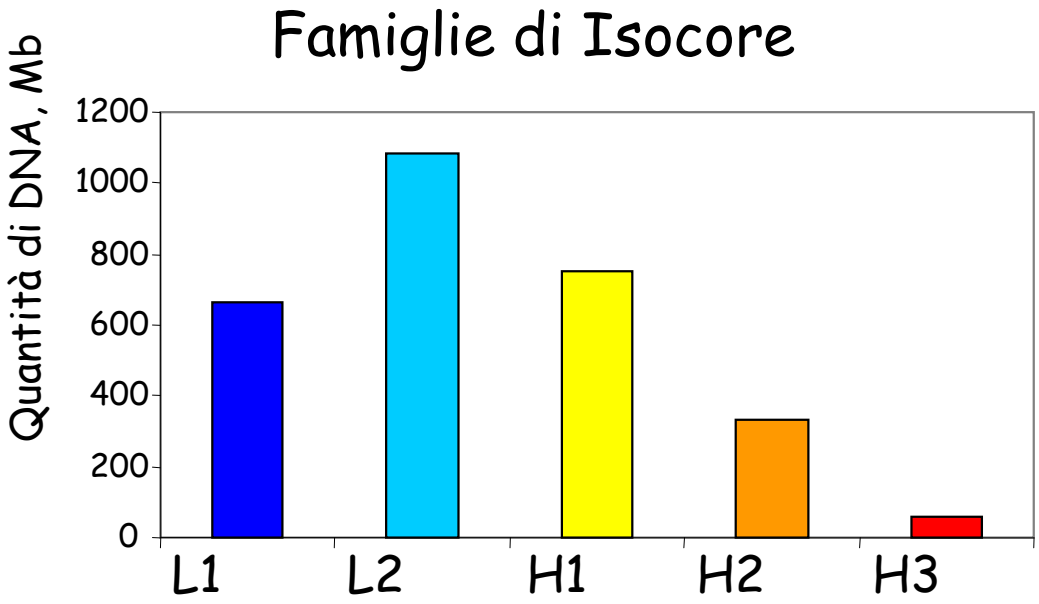
Modello delle Isocore

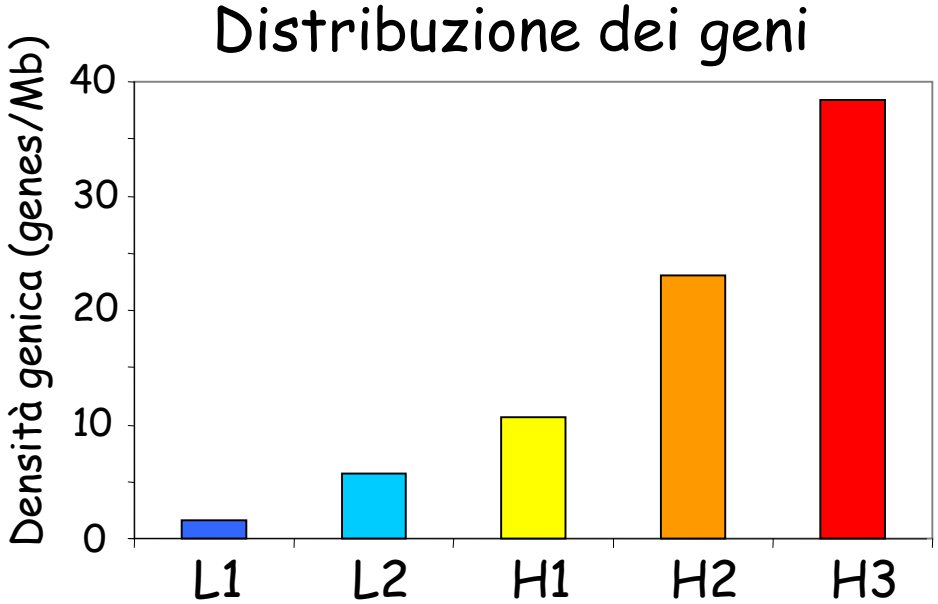


Maria Costantini et al. Genome Res. 2006; 16: 536-541

Correlazione tra isocore e proprietà del genoma

La maggior parte del genoma è costituita da isocore leggere (L1, L2). Al contrario la maggior parte dei geni è localizzata nelle isocore pesanti (H1, H2 e H3).





Nel *genome core* costituito dalle isocore H2 e H3 (12% del genoma) la densità dei geni è molto alta (un gene per 5-15kb), mentre nel cosiddetto *empty space* formato dalle isocore di tipo L e H1 (88% del genoma) la densità genica è molto bassa (un gene per 50-150kb).

Ipotesi della maggiore stabilità termodinamica delle sequenze codificanti, importante negli animali a sangue caldo

3) Numero di geni

Ma che cosa è un GENE?

Se non ci si accorda sulla definizione di gene, non è possibile determinarne il numero, anche assumendo di disporre della annotazione completa del genoma.

L'avvento dell'era genomica ha messo in crisi la tradizionale definizione di **GENE**, tuttora molto dibattuta.

Lewin B. Il Gene VIII
Segmento di DNA coinvolto nella produzione di una catena polipeptidica, comprende regioni (leader e coda) che precedono e seguono la regione codificante, oltre alle sequenze intercalate (introni) tra i singoli elementi codificanti (esoni).

Brown T.A. Genomi 2
Un segmento di DNA contenente informazioni biologiche, che codifica per una molecola di RNA e/o proteina.

Ambedue queste definizioni non possono essere considerate corrette alla luce delle attuali conoscenze.

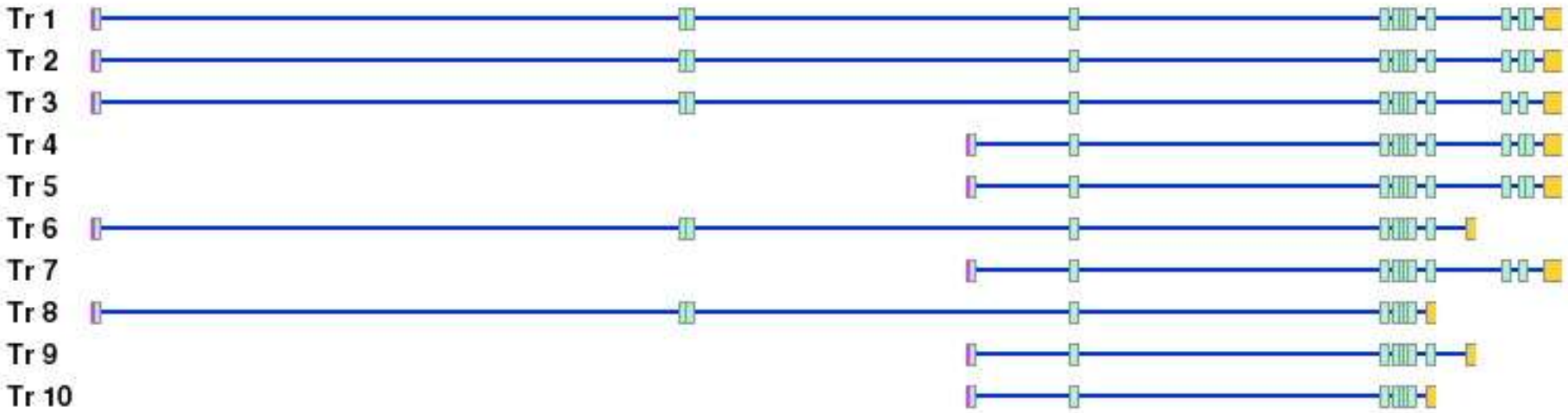
Un gene può avere più inizi, più terminazioni e più processamenti del suo RNA

Per giungere ad una definizione il più possibile corretta di **GENE** è necessario conoscerne le caratteristiche principali.

- Un gene può utilizzare diversi promotori

- La trascrizione di un gene si può arrestare in corrispondenza di diversi terminatori

- I trascritti espressi da un gene possono subire splicing alternativo che generano trascritti che differiscono sia nelle regioni non tradotte (5' e 3'UTR) che nella regione codificante



Il gene per tp73L (fattore trascrizionale della famiglia p53)
codifica per **10 trascritti alternativi,**
utilizza **2 promotori** e **3 diversi terminatori della trascrizione**
(predizione ottenuta dal programma ASPIC).

Quindi uno stesso gene con tanti promotori e tanti terminatori.

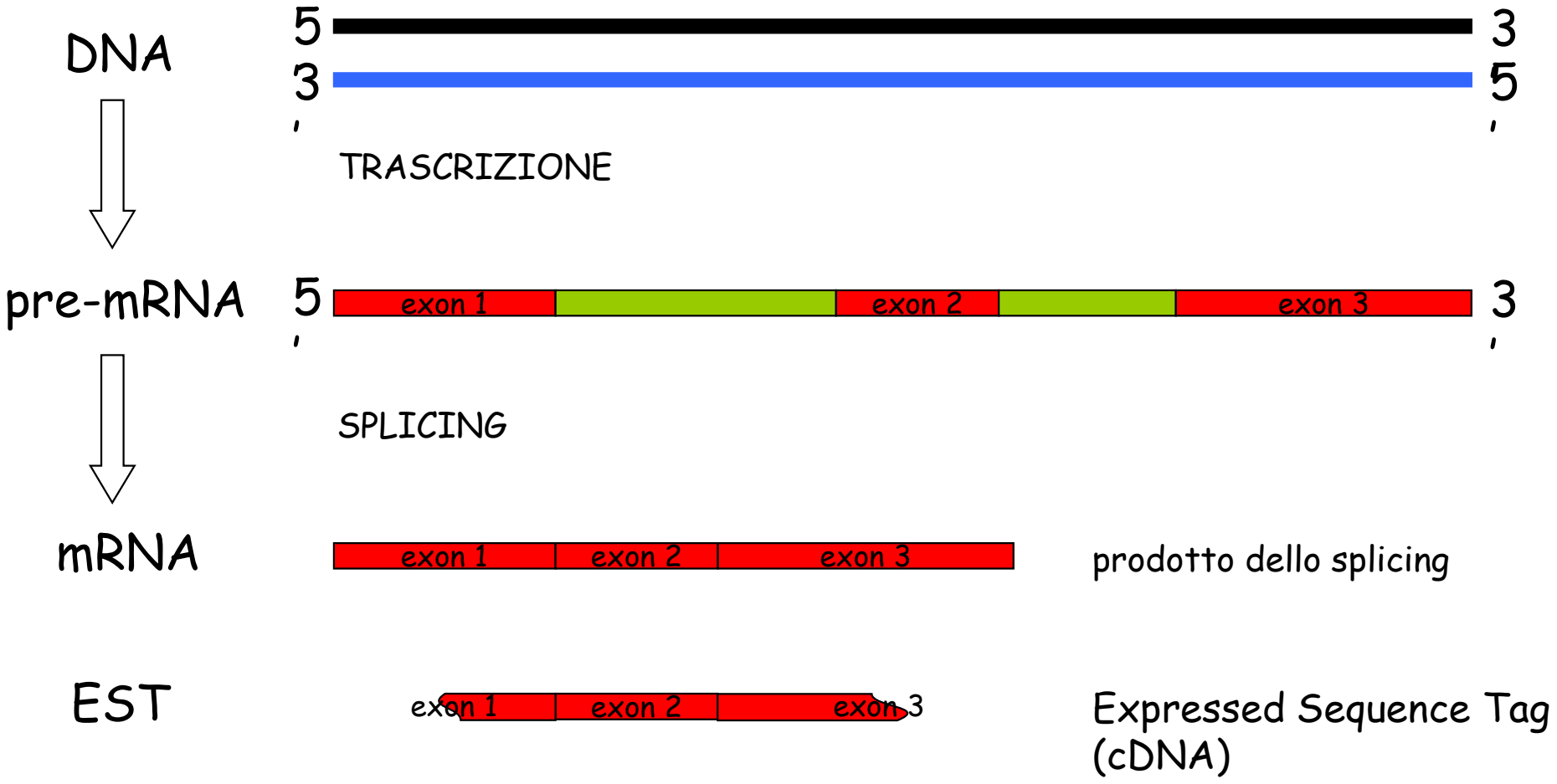
Ma un gene con lo stesso promotore e lo stesso terminatore può dare comunque trascritti diversi: lo splicing alternativo

ASPic (Alternative Splicing Prediction)

è uno strumento informatico che compara la sequenza genomica di un gene con quella delle ESTs correlate.

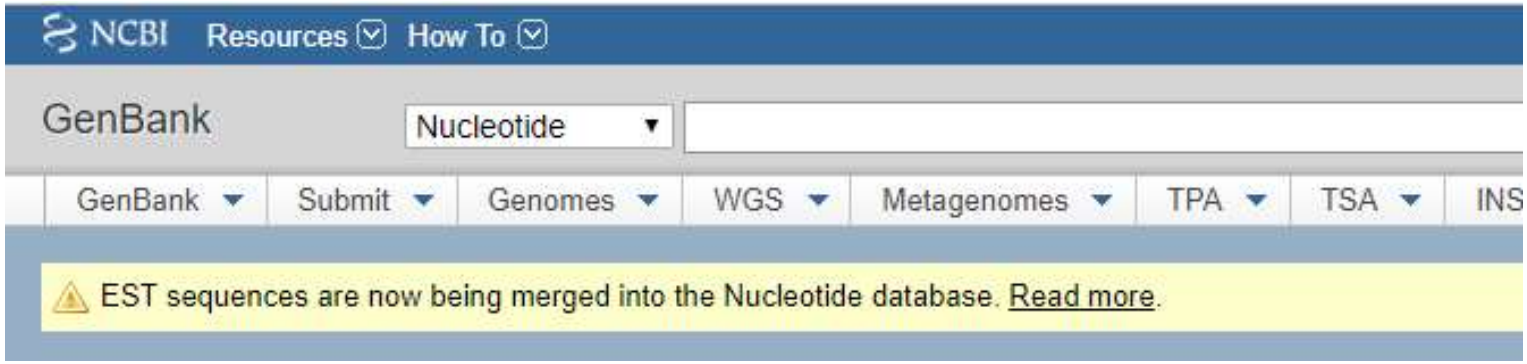
Il programma è basato sull'identificazione di sequenze introne/esone di un gene.

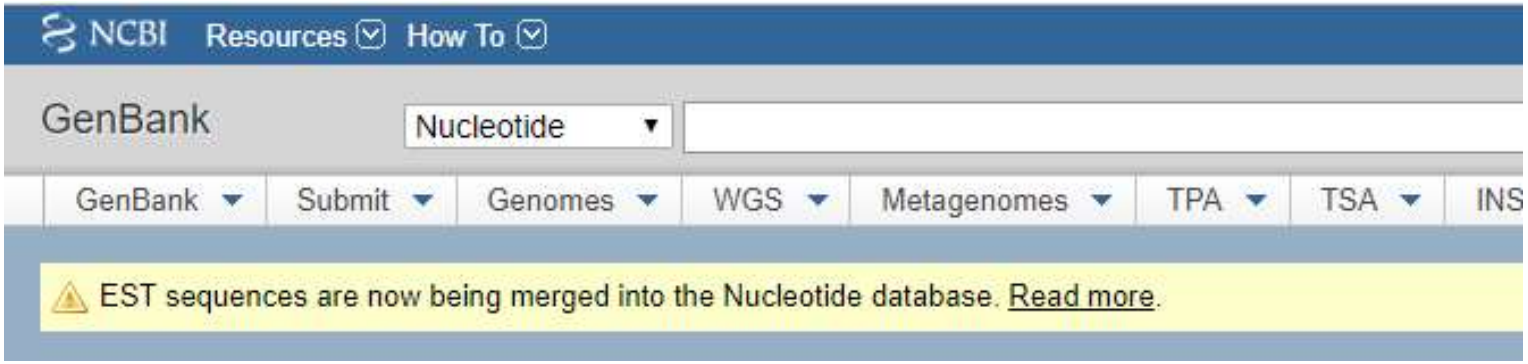
Bonizzoni P, Rizzi R, Pesole G. *ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences*. BMC Bioinformatics. 2005 Oct 5;6:244.



Le **ESTs** (Expressed Sequence Tag) sono frammenti di sequenze di mRNA ottenute mediante sequenziamento di cloni di librerie di cDNA selezionati randomicamente.

A gennaio 2013 in database pubblici risultavano archiviati 74.2 milioni di ESTs relativi a tutte le specie (tra queste sono comprese molte sequenze ridondanti o che mappano in punti diversi dello stesso mRNA, il che spiega l'enorme abbondanza di EST in rapporto al numero di geni).


NCBI
Resources
How To
GenBank
Nucleotide
GenBank
Submit
Genomes
WGS
Metagenomes
TPA
TSA
INS
EST sequences are now being merged into the Nucleotide database. Read more.



dbEST release 103103

Summary by Organism - October 31, 2003

Number of public entries 18,971,362

Homo sapiens (human)	5,427,521
Mus musculus + domesticus (mouse)	3,915,334
Rattus sp. (rat)	538,251
Triticum aestivum (wheat)	500,902
Ciona intestinalis	492,488
Gallus gallus (chicken)	451,565
Zea mays (maize)	383,759
Danio rerio (zebrafish)	362,445
Hordeum vulgare + subsp. vulgare (barley)	348,233
Xenopus laevis (African clawed frog)	344,747
Glycine max (soybean)	341,578
Bos taurus (cattle)	329,387
Drosophila melanogaster (fruit fly)	261,414
Oryza sativa (rice)	260,890
Saccharum officinarum	246,301
Caenorhabditis elegans (nematode)	215,200
Silurana tropicalis	209,240
Arabidopsis thaliana (thale cress)	190,732
Medicago truncatula (barrel medic)	187,763
Sus scrofa (pig)	171,920

dbEST release 130101

Summary by Organism - 01 January 2013

Number of public entries 74,186,692

Homo sapiens (human)	8,704,790
Mus musculus + domesticus (mouse)	4,853,570
Zea mays (maize)	2,019,137
Sus scrofa (pig)	1,669,337
Bos taurus (cattle)	1,559,495
Arabidopsis thaliana (thale cress)	1,529,700
Danio rerio (zebrafish)	1,488,275
Glycine max (soybean)	1,461,722
Triticum aestivum (wheat)	1,286,372
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,480
Oryza sativa (rice)	1,253,557
Ciona intestinalis	1,205,674
Rattus norvegicus + sp. (rat)	1,162,136
Drosophila melanogaster (fruit fly)	821,005
Panicum virgatum (switchgrass)	720,590
Xenopus laevis (African clawed frog)	677,911
Oryzias latipes (Japanese medaka)	666,891
Brassica napus (oilseed rape)	643,881

EST: stime aggiornate ad aprile 2026

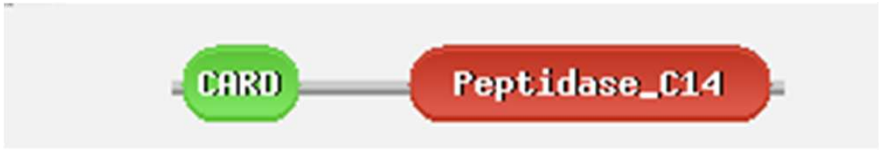
Categoria	Numero Approssimativo
Totale sequenze EST in dbEST	~82.500.000
EST Umane (Homo sapiens)	~9.200.000
EST di Topo (Mus musculus)	~5.100.000
Specie rappresentate	> 2.500

La sezione EST non cresce più velocemente come un tempo. Il motivo è tecnico: Avvento di TSA (Transcriptome Shotgun Assembly): Oggi i ricercatori preferiscono assemblare i trascritti completi derivati da RNAseq piuttosto che depositare singoli frammenti (EST). Le sequenze annotate nei database TSA superano ormai di gran lunga le vecchie EST in termini di utilità informativa

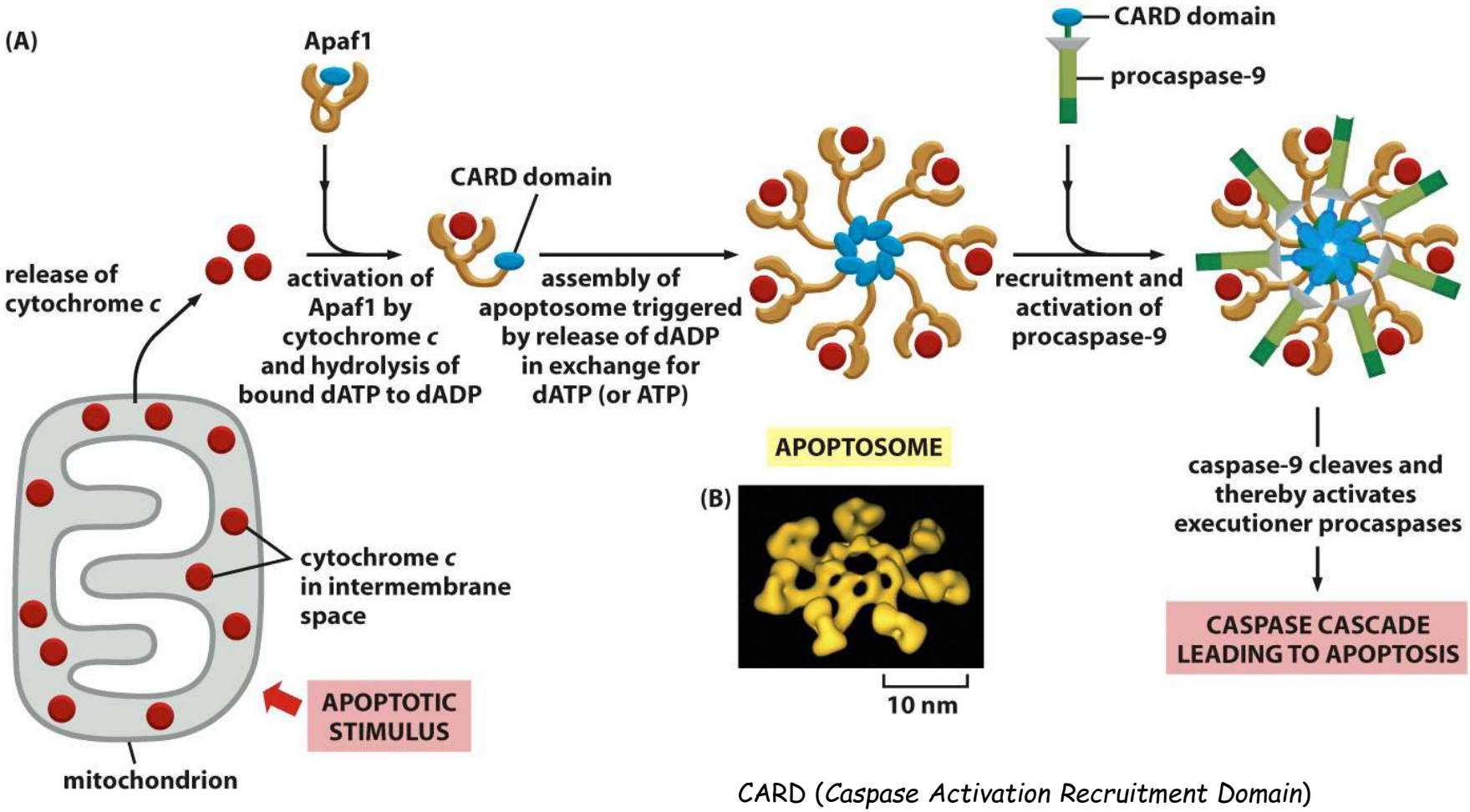
Splicing alternativo (1)

Uno stesso gene può esprimere proteine con funzioni opposte

La forma costitutiva della proteina (CASP9, 9 esoni, 416 aa) induce apoptosi. Essa contiene un Caspase recruitment domain (CARD) e un dominio caspasi Peptidase_C14.



L'isoforma più corta della proteina (CASP9S, 5 esoni, 266 aa) contiene un dominio Caspase recruitment domain (CARD) e un dominio tronco della Peptidase_C14. Questa isoforma è priva dell'attività proteasica e agisce da inibitore dell'apoptosi.



CARD (*Caspase Activation Recruitment Domain*)

APAF-1 (*Apoptotic Protease Activating Factor 1*)

Apoptosoma: complesso multi-molecolare con una caratteristica struttura a ruota, formato da sette dimeri di caspasi 9 che interagiscono con sette monomeri di Apaf-1.

Il citocromo C si lega attivandola alla proteina Apaf1 che cambia conformazione e espone il dominio CARD utile alla creazione di aggregati (apoptosoma): ai domini CARD dell'apoptosoma si legano i domini omonimi sulle pro-caspasi 9 che vengono reclutate e attivate sull'apoptosoma.

Attivazione di una Caspasi esecutrice

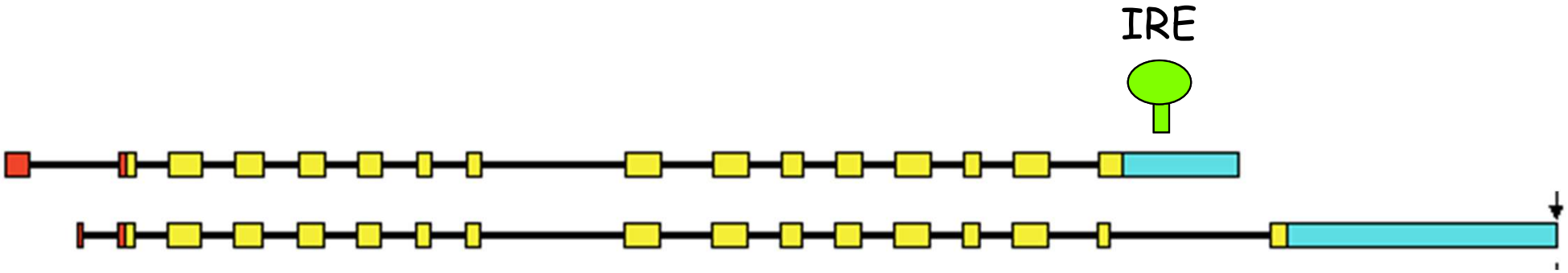
L'isoforma senza dominio caspasico, risultato di uno splicing alternativo, funziona da **Dominante Negativo**.

La forma tronca satura l'apoptosoma impedendo l'attacco della forma lunga della proteina Casp9.

Splicing alternativo (2)

Uno stesso gene può codificare trascritti soggetti ad un diverso meccanismo di regolazione post-trascrizionale

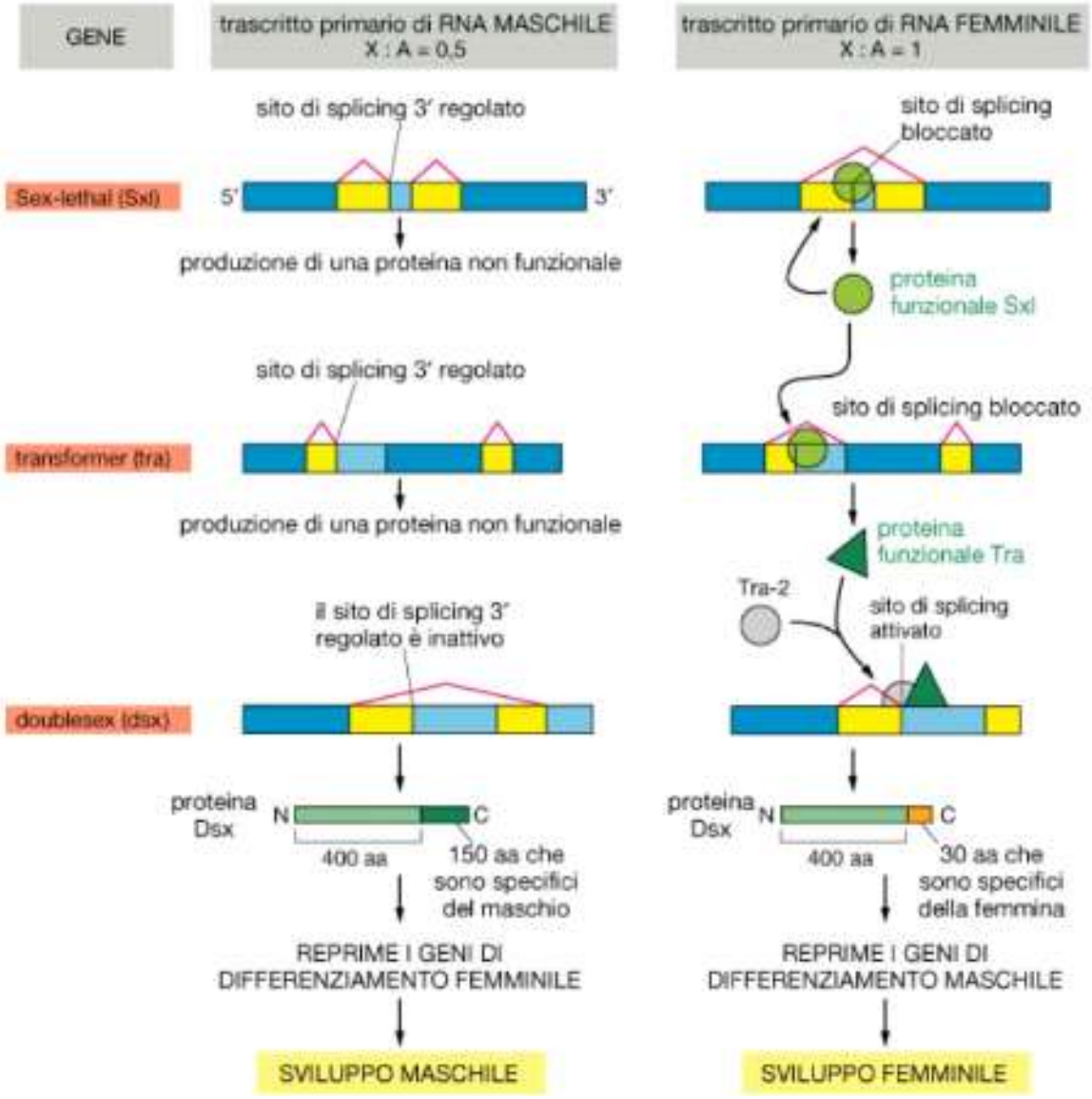
Il **gene SLC11A2** (trasportatore di cationi bivalenti) codifica per (almeno) due diverse isoforme, solo una delle quali risponde alla concentrazione del ferro (i.e. i livelli della proteina aumentano sensibilmente in seguito alla carenza di ferro). L' "Iron Responsive Element (IRE)" nella regione 3'UTR è presente solo in una delle due isoforme.



Nell'uomo il trascritto contenente l'IRE (16 exons) codifica per una proteina di 561 aa (NM_000617). Il trascritto privo di IRE (17 exons) codifica per una proteina di 568 aa.

Splicing alternativo (3)

Determinazione del sesso in drosofila

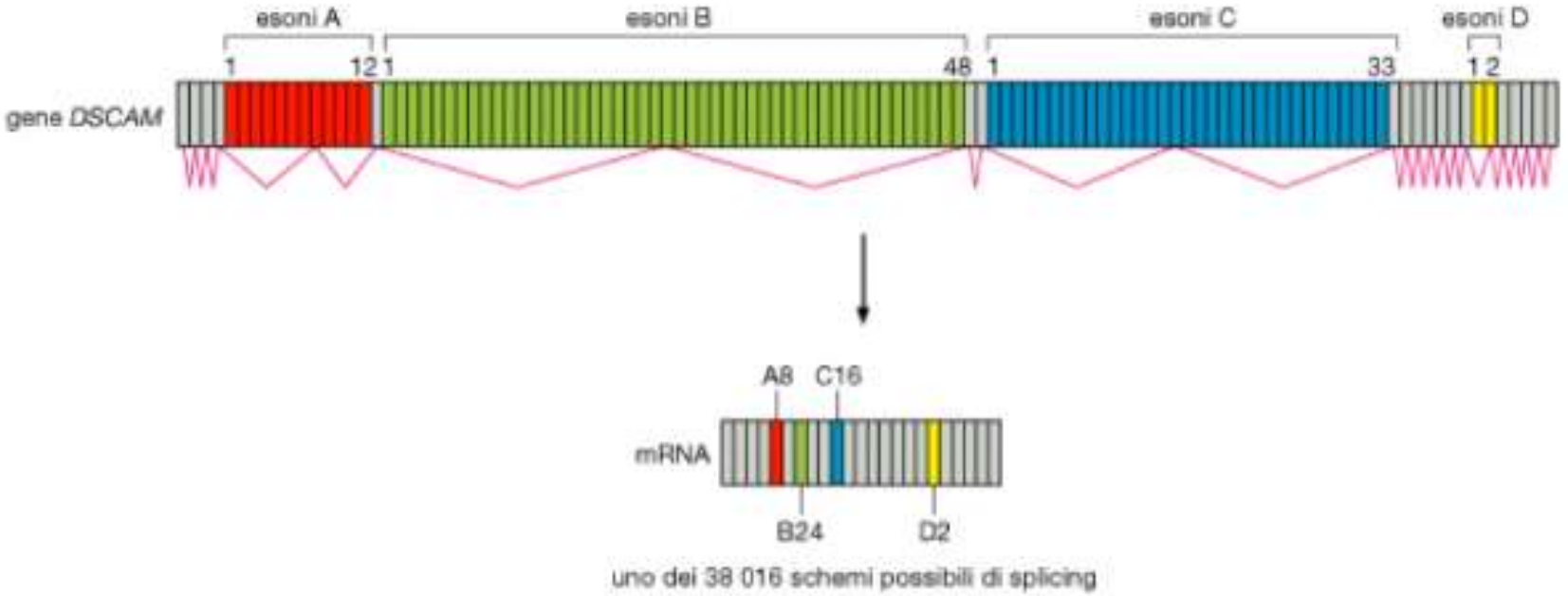


Splicing alternativo (4)

Sviluppo del Sistema nervoso in drosofila

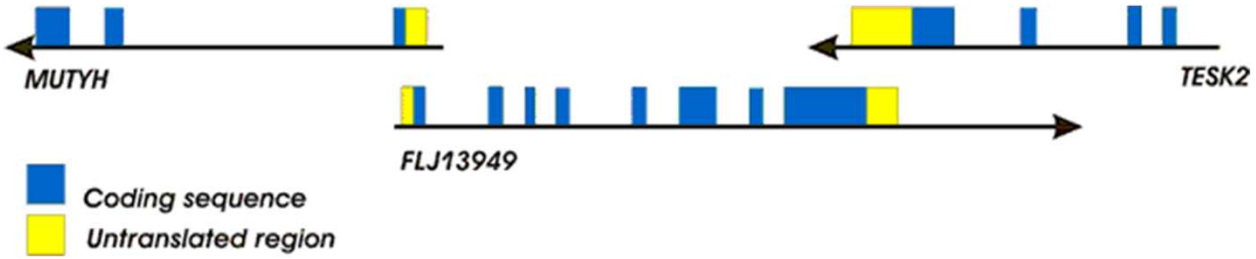
Caso estremo di splicing alternativo: il gene DSCAM codifica per una proteina transmembrana che guida la formazione di sinapsi tra neuroni che esprimono la stessa isoforma.

Ogni neurone sceglie stocasticamente solo una splicing.



I geni possono essere sovrapposti

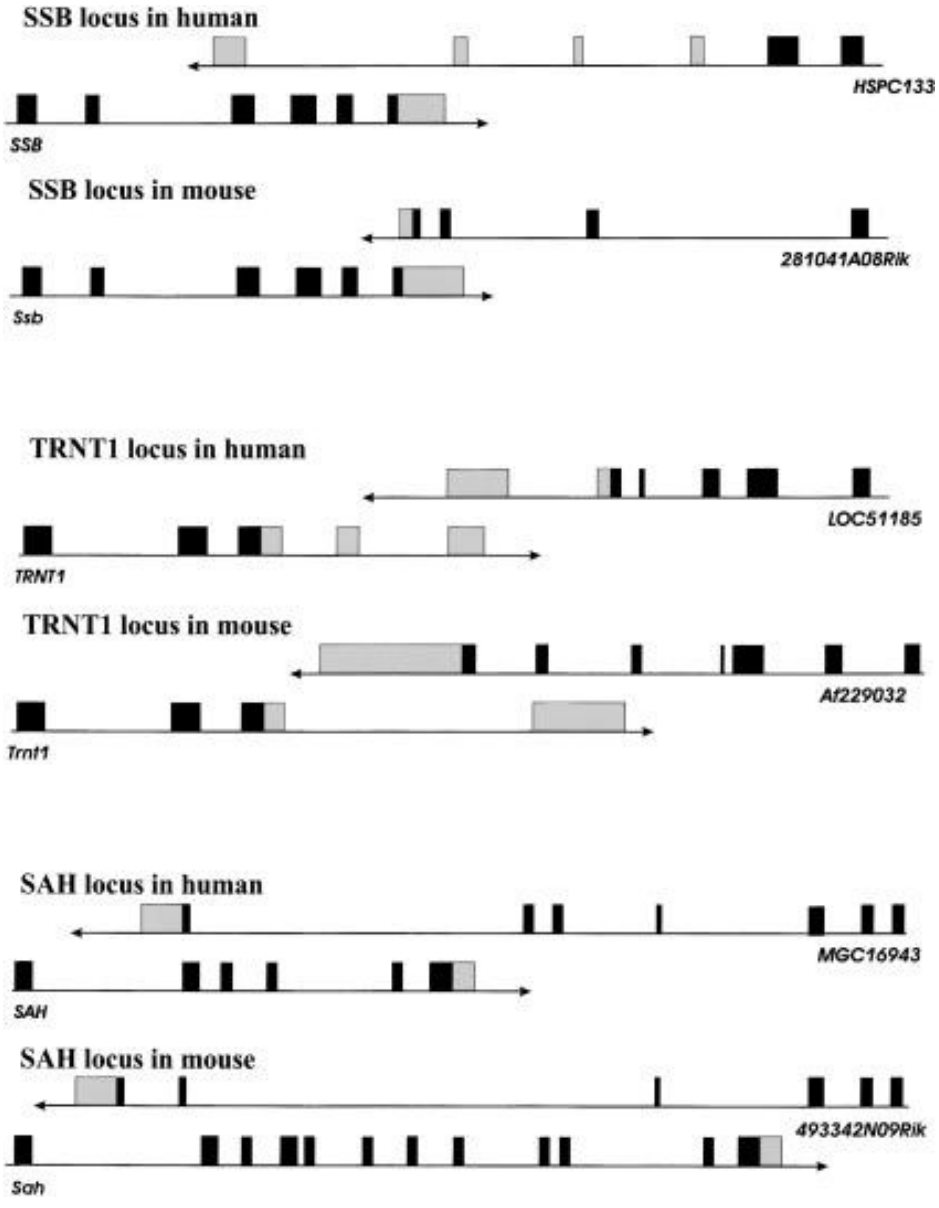
I geni possono essere sovrapposti tra loro, nello stesso orientamento o in orientamento opposto.



FLJ13949 codifica per un lncRNA che lavora in cis inibendo, o meglio armonizzando*, la trascrizione dei due geni per ingombro sterico nei confronti dell'apparato trascrizionale

*i due geni possono essere inibiti in maniera diversa e coordinata

vedi: http://posnania.biotec.psu.edu/research/overlapping_genes.html

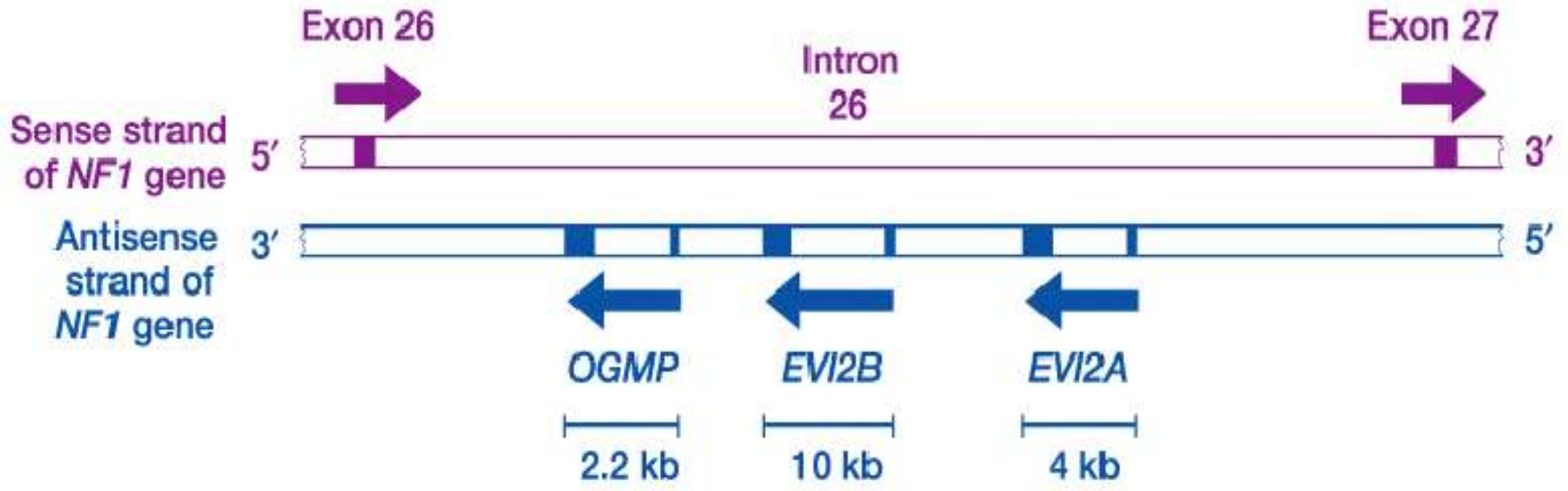


Esempi di geni con differenti pattern di sovrapposizione in uomo e topo (geni ortologhi).

Box neri: sequenze codificanti
Box grigi: sequenze non tradotte

I geni possono stare dentro altri geni

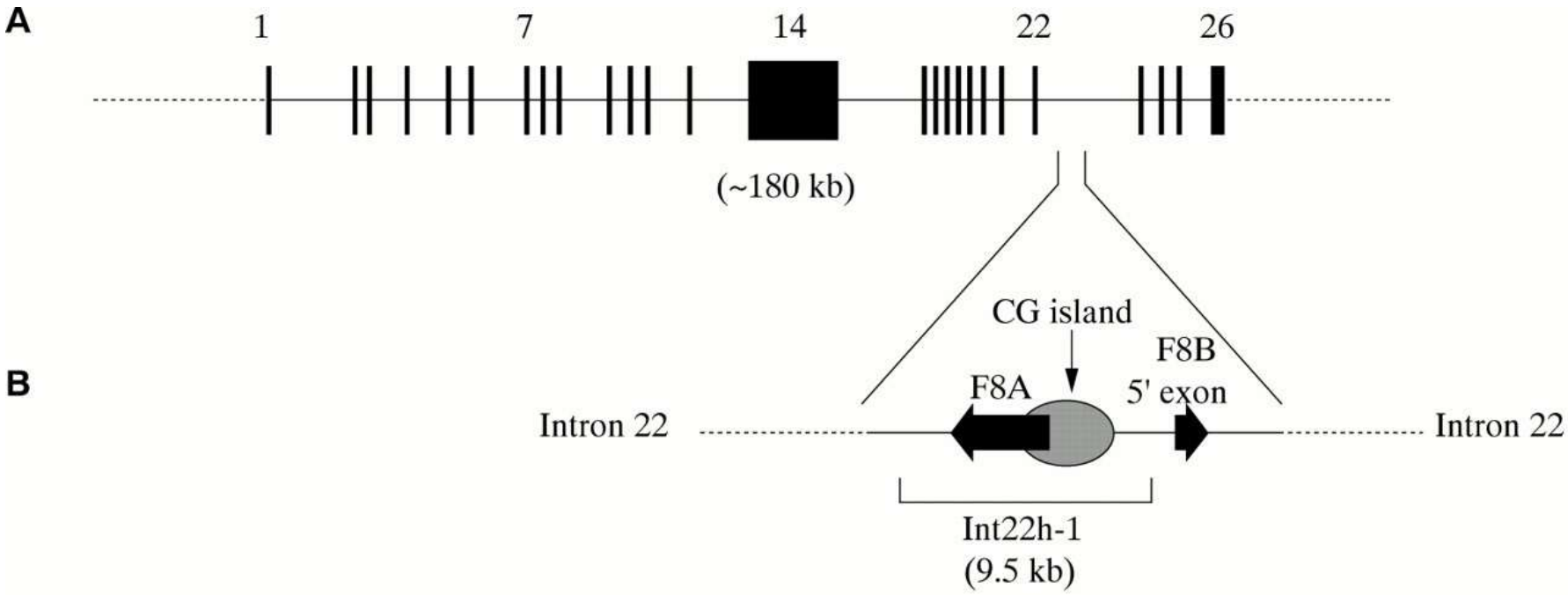
L'introne 26 del gene *neurofibromatosis type I* (NF1) contiene 3 geni diversi nell'orientamento opposto (*OMGP*, *EVI2A*, *EVI2B*).



Mol Cell Biol. 1991 Feb;11(2):906-12.

The gene encoding the oligodendrocyte-myelin glycoprotein is embedded within the neurofibromatosis type 1 gene.

Viskochil D[1], Cawthon R, O'Connell P, Xu GF, Stevens J, Culver M, Carey J, White R.



Fattore VIII della coagulazione: l'introne 22 contiene due geni che utilizzano la stessa isola CpG nelle due direzioni. Il gene F8A rimane nell'introne 22 e viene abbondantemente trascritto in molti tipi cellulari ed utilizzando il filamento opposto a F8; è molto conservato (funzione correlate a quella dell'hungtintina).

F8B sintetizza un corto mRNA che ha un esone nuovo + gli esoni dal 23 al 26 di F8 (proteina più corta dell'F8 canonico)

I geni che codificano per i microRNAs sono spesso dei geni nei geni

La loro sequenze codificanti occupano spesso gli introni di geni codificanti per polipeptidi

Più del 50% dei geni per i microRNAs sono in cluster e possono essere trascritti in un unico RNA policistronico successivamente processato

Più geni possono produrre un unico trascritto: i trascritti chimerici

- Alcuni trascritti vengono originati dalla ligazione di diverse molecole di RNA attraverso il meccanismo del transplicing (cis-splicing è il processamento canonico)
- Si possono formare trascritti chimerici in seguito alla co-trascrizione di geni disposti in tandem

Abstract ▾ Send to: ▾

Genes Chromosomes Cancer. 2014 Dec;53(12):963-71. doi: 10.1002/gcc.22207. Epub 2014 Aug 11.

Chimeric RNAs generated by intergenic splicing in normal and cancer cells.

Jividen K[1], Li H.

⊕ **Author information**

Abstract

A hallmark of many neoplasias is chromosomal rearrangement, an event that commonly results in the fusion of two separate genes. The RNA and protein resulting from these gene fusions often play critical roles in cancer development, maintenance, and progression. Traditionally, these fusion products are thought to be produced solely due to DNA level changes and are therefore considered unique to cancer. Recent advances in microarray and deep-sequencing have revealed many more fusion transcripts. Surprisingly, some are without detectable rearrangement at the DNA level. Reports have demonstrated that at least some of these chimeric RNAs are generated via intergenic splicing. In this review, we highlight three examples of these noncanonical chimeric transcripts that are formed by trans-splicing or cis-splicing of adjacent genes and summarize the knowledge we have regarding these noncanonical fusions. We discuss the implications of the chimeric RNAs in both cancer and normal physiology, as some of these fusion transcripts are found in normal, noncancerous cells with sequences identical to those generated by canonical chromosomal translocation found in cancer cells. Finally, we present methods that are currently being used to discover additional chimeric RNAs.

PMID: 25131334 [PubMed - in process]

Utilizzo di diversi siti di traduzione:

Uno stesso gene può codificare per proteine indirizzate a diversi compartimenti cellulari

Il gene per la proteina NFS1 (coinvolta nella formazione delle proteine ferro-zolfo) presenta **siti di inizio alternativi della traduzione** per generare una isoforma mitocondriale ed una isoforma citoplasmatica. La selezione del sito di inizio della traduzione è regolata dal pH citosolico.

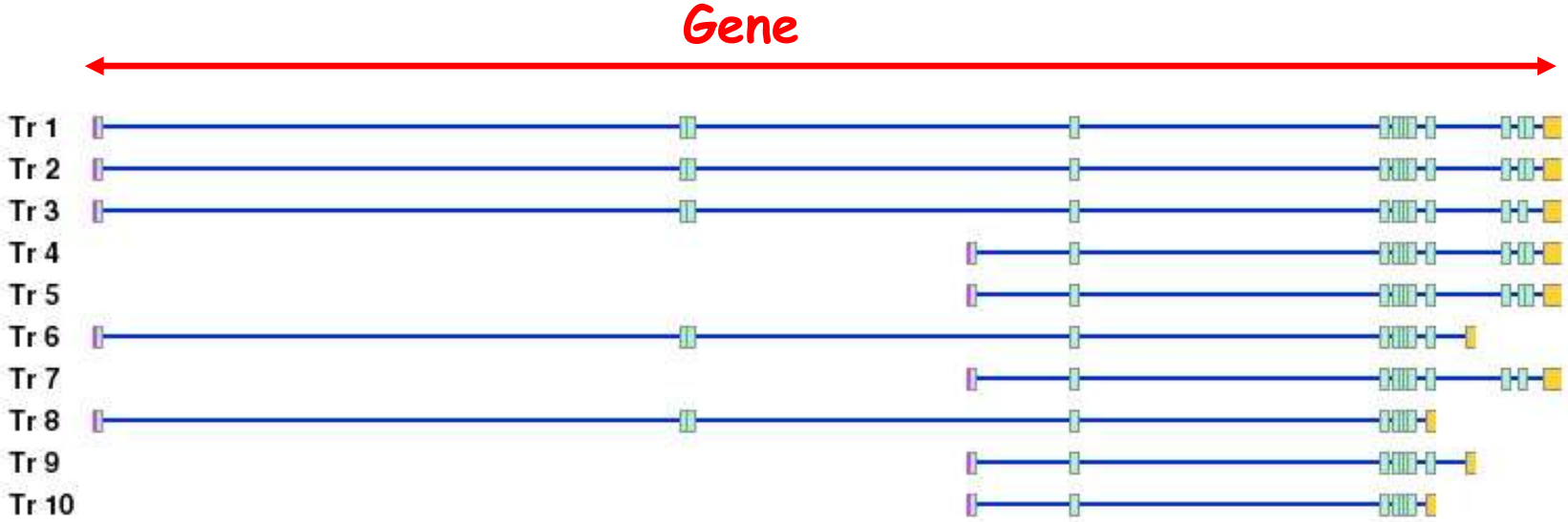
L'isoforma che codifica per la proteina mitocondriale (457 aa) contiene un peptide segnale e un dominio aminotransferasico.

L'altra isoforma, che deriva da un sito di inizio alternativo della traduzione codifica per una proteina più corta (397 aa) priva del peptide segnale ma contenente il dominio aminotransferasico.

NUOVA DEFINIZIONE di GENE

Una specifica regione di DNA, la cui trascrizione è regolata da uno o più promotori e altri elementi di controllo trascrizionale che contiene l'informazione per la sintesi di proteine e RNA non codificanti funzionali, tra loro correlati per la condivisione di informazione genetica (con un tratto di sequenza genomica in comune) a livello dei prodotti finali (proteine o ncRNA).

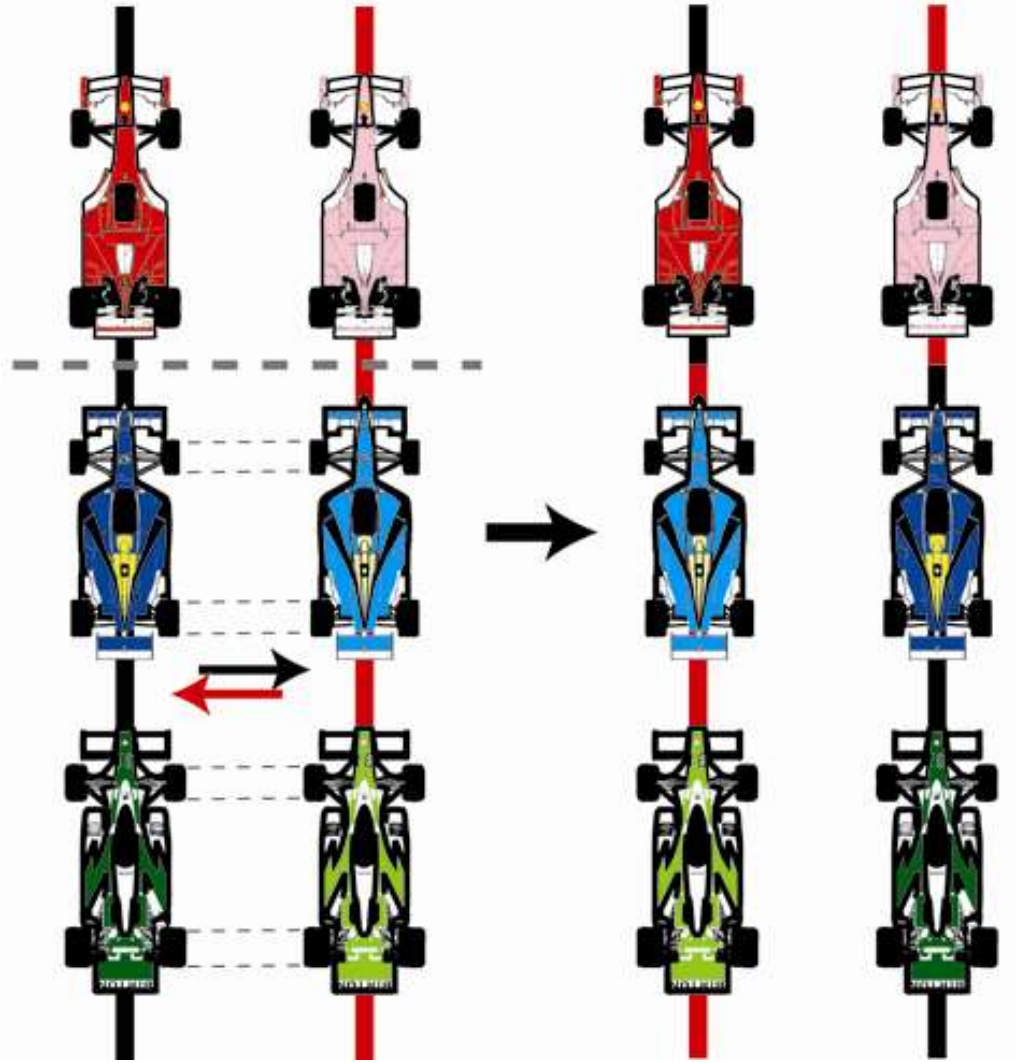
In questo modo è possibile associare al gene specifiche coordinate genomiche che coincidono con il sito di inizio della trascrizione più a monte e il sito di terminazione più a valle.



Nel genoma ci sono anche geni finti, gli Pseudogeni

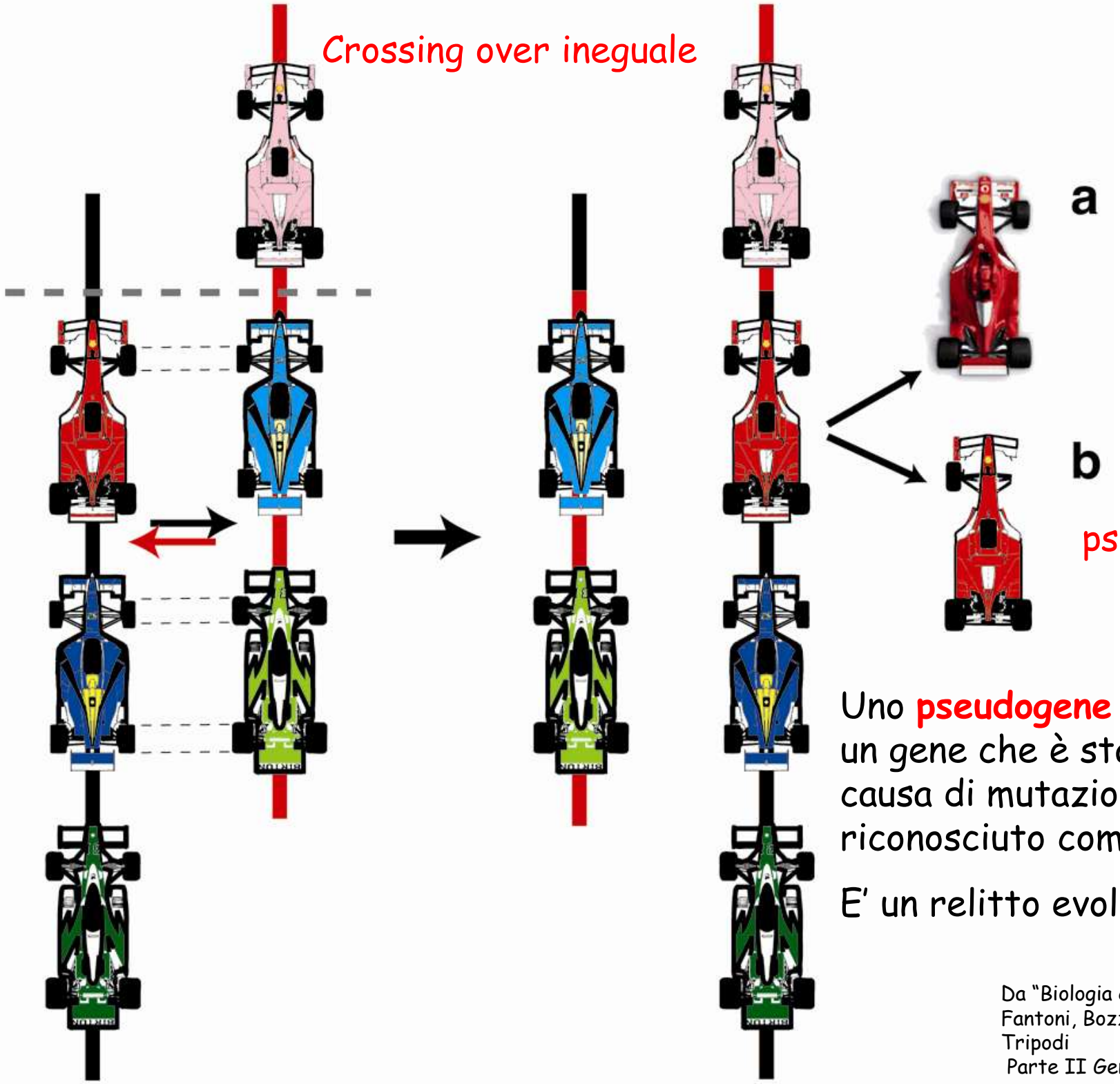
(Pseudogeni duplicati o non processati e Pseudogeni retrotrasposti o processati)

Gli **pseudogeni duplicati, o non processati,** (circa 11000 nel genoma umano) derivano da duplicazione in tandem o da crossing-over ineguale



Crossing over normale

Da "Biologia cellulare e genetica"
Fantoni, Bozzaro, Del Sal, Ferrari, Tripodi
Parte II Genetica, Piccin



pseudogene

Uno **pseudogene convenzionale** è un gene che è stato inattivato a causa di mutazioni (non viene più riconosciuto come gene).

E' un relitto evolutivo.

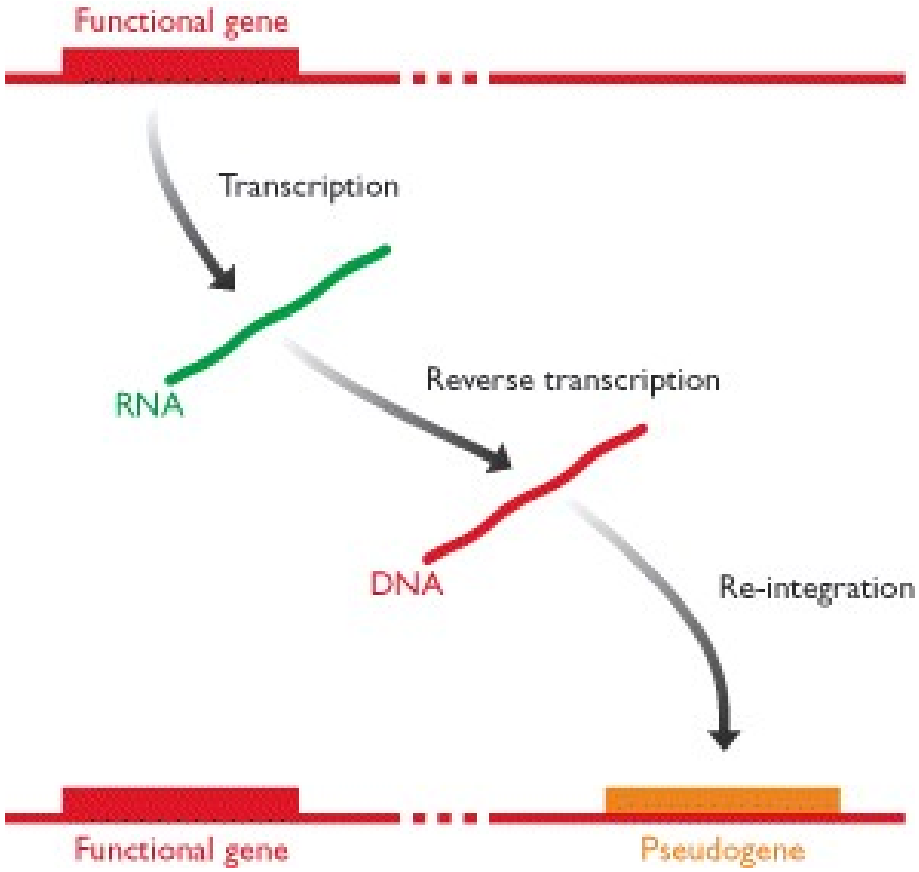
Da "Biologia cellulare e genetica"
Fantoni, Bozzaro, Del Sal, Ferrari, Tripodi
Parte II Genetica, Piccin

Uno **pseudogene retrotrasposto, o processato,**

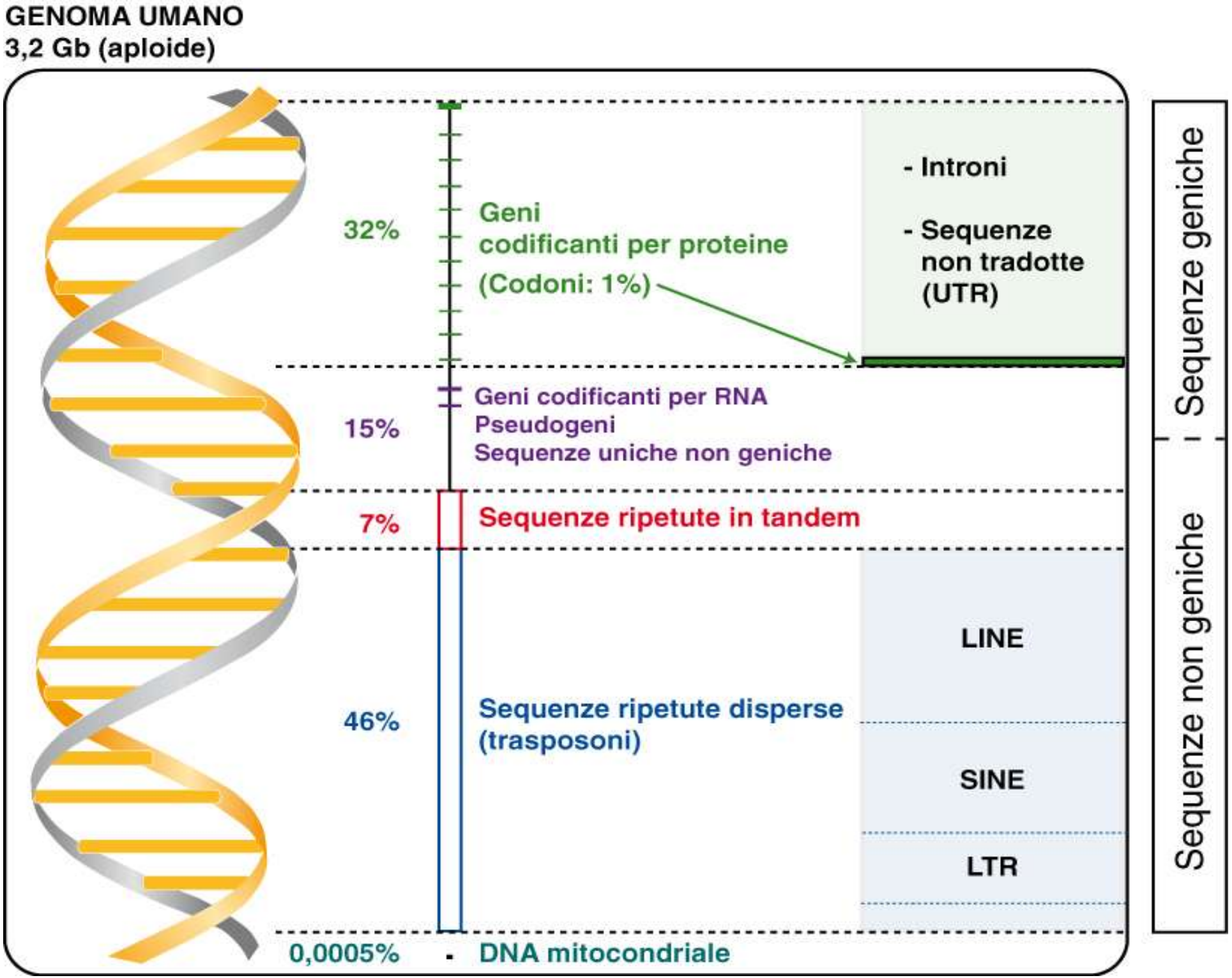
deriva dall'mRNA di un gene su cui viene sintetizzata una copia di DNA che successivamente viene reinserita nel genoma.

Uno pseudogene processato non contiene le sequenze introniche e le sequenze 5-UTR che regolano l'espressione del gene.

Uno pseudogene processato è inattivo.



4) Sequenza ripetute

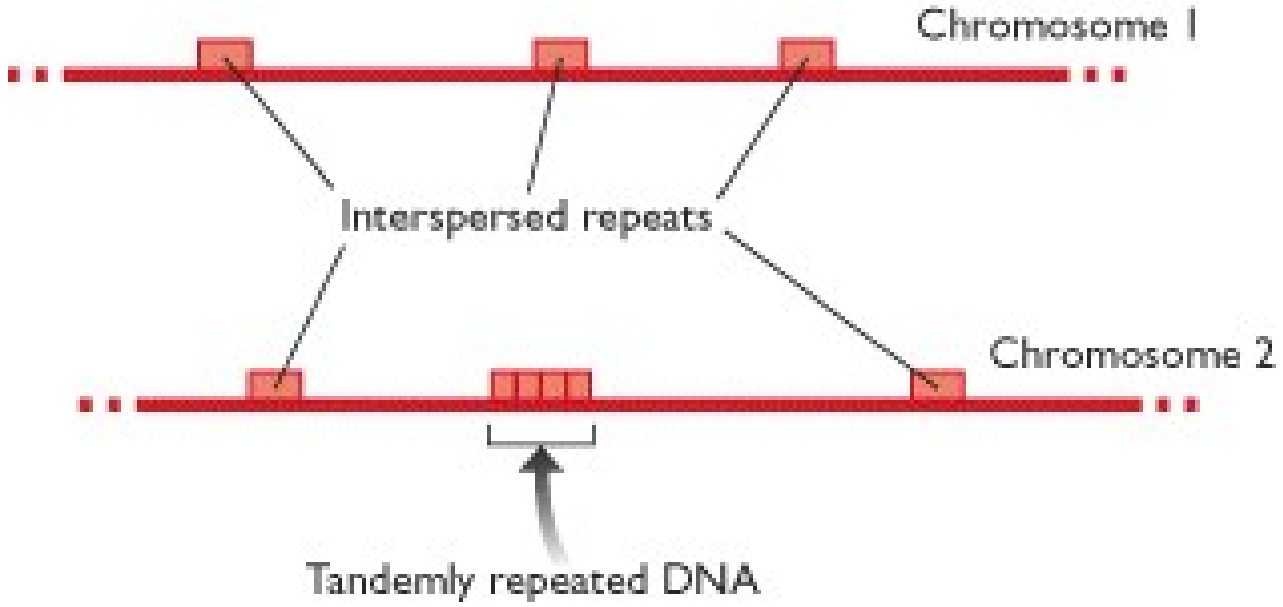


Da "Biologia cellulare e genetica" Fantoni, Bozzaro, Del Sal, Ferrari, Tripodi Parte II Genetica, Piccin

Il **DNA ripetitivo** può essere distinto in due categorie:

- Ripetizioni intersperse: le cui unità sono distribuite nel genoma in modo casuale e occupano il 46% del genoma

- Sequenze ripetute in tandem: le cui unità sono disposte in serie l'una vicina all'altra ed occupano il 6% del genoma



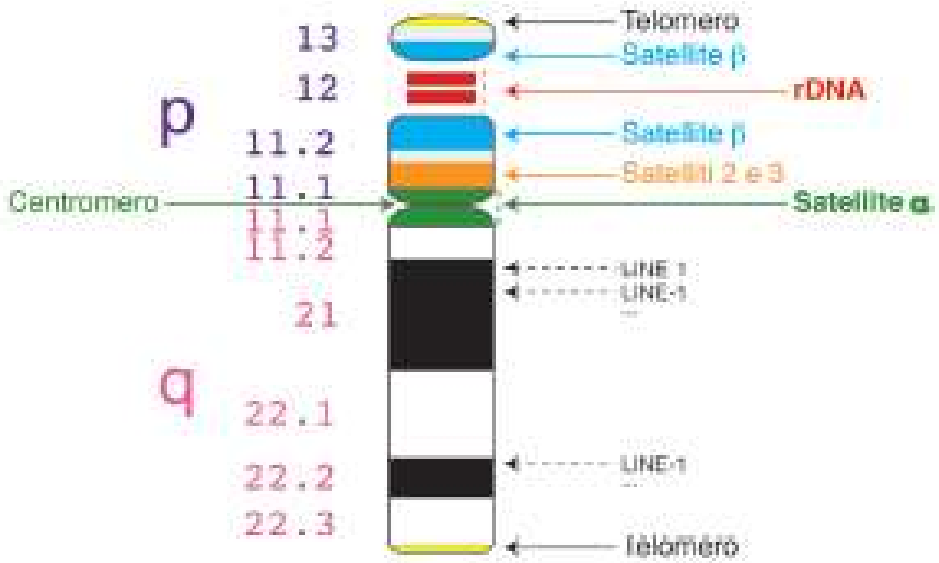
Sequenze ripetute a tandem

DNA satellite: costituisce la maggior parte delle regioni di eterocromatina, in particolare di quella pericentrica e centromerica. Unità ripetute di 5-171 nt.

DNA minisatellite: ipervariabile (in dimensioni), telomerico. L'unità ripetuta può essere lunga fino a 64 nt.

DNA microsatellite: STR= Short Tandem Repeats (1-6 bp che si susseguono nell'ambito di piccoli blocchi di lunghezza inferiore a 150 bp).
Uniformemente interdispersi in tutti i cromosomi.
Variabilità molto elevata nel numero di unità ripetute (VNTR=Variable Number of Tandem Repeats) riscontrabile tra diversi individui.

Cromosoma 21: localizzazione delle sequenze ripetute



I satelliti sono principalmente centromerici e telomerici.

Le sequenze iterdisperse si trovano principalmente nelle bande G

Nb: nel secondo restringimento situato a livello del braccio corto dei cromosomi acrocentrici risiede il cluster dei geni ribosomiali

Sequenze ripetute a tandem: funzione?

- Si sa che derivano da errori nel processo di copiatura del genoma durante la divisione cellulare mitotica (scivolamento replicativo) o meiotica (crossing over ineguale) e potrebbero essere prodotti inevitabili della replicazione del genoma

- si ritiene che la loro funzione sia di natura strutturale e che quindi abbiano un qualche ruolo nell'organizzazione dei cromosomi.

Ripetizioni Intersperse

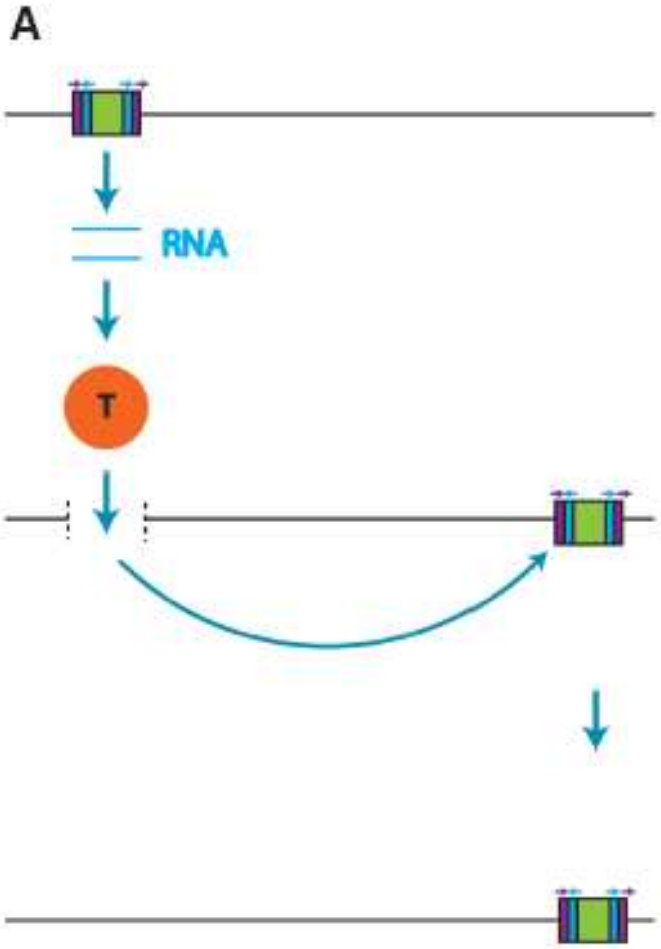
L'elemento di sequenza che ricorre più volte è distribuito nel genoma in tante localizzazioni diverse.

Non dipendono da errori di copiatura o di ricombinazione del DNA: sono degli **elementi trasponibili o trasposoni.**

Trasposoni a DNA: tratti di DNA escissi e reinseriti in altri punti del genoma. Possono essi stessi codificare per enzimi che li rendono autonomi nel processo di trasposizione. Si tratta di una trasposizione non replicativa.

Retrotrasposoni: derivano da trascritti di RNA copiati dalla Trascrittasi Inversa e integrati nel genoma. Si tratta di una trasposizione replicativa.

TRASPOSONI a DNA



La sequenza del trasposone viene trascritta e tradotta in una proteina ad attività enzimatica (trasposasi). Questa rientra nel nucleo e rimuove la sequenza del trasposone dalla sua localizzazione originaria e la inserisce in una nuova. L'interruzione nel DNA viene riparata.

Sono elementi mobili del genoma che non creano sequenze ripetute e che quindi non sono causa di espansione del genoma.

Attenzione: sequenze di trasposoni a DNA occupano circa occupano circa il 3% del genoma umano

MA

L'ultima volta che un trasposone a DNA ha "saltato" attivamente nel genoma dei nostri antenati (i primati antropomorfi) è stato circa **37-40 milioni di anni fa**. Da allora, sono rimasti bloccati nella posizione in cui si trovavano.

Attenzione: L'ultima volta che un trasposone a DNA ha "saltato" attivamente nel genoma dei nostri antenati (i primati antropomorfi) è stato circa **37-40 milioni di anni fa**. Da allora, sono rimasti bloccati nella posizione in cui si trovavano.

I trasposoni a DNA sono stati tra i più potenti **motori dell'evoluzione** del genoma umano.

1) Molti **promotori** e **enhancer** dei geni «moderni» derivano da rimaneggiamenti evolutivi di sequenze trasposoniche.

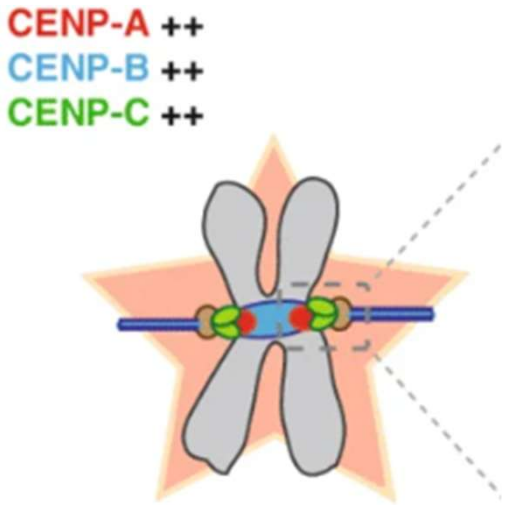
Quando un trasposone si fermava vicino a un gene, portava con sé dei siti di legame per fattori di trascrizione. Molti di questi sono diventati enhancer o promotori che oggi accendono o spengono geni cruciali durante lo sviluppo embrionale o nel cervello. Circa il 10-20% dei nostri elementi regolatori ha origini derivate dai trasposoni.

I trasposoni a DNA sono stati tra i più potenti **motori dell'evoluzione** del genoma umano.

2) Circa 500 milioni di anni fa, un trasposone a DNA è "saltato" nel genoma di un nostro antenato vertebrato.
La sua trasposasi è diventata la **proteina RAG1**, che oggi i nostri linfociti usano per tagliare e rimescolare i geni degli anticorpi.

I trasposoni a DNA sono stati tra i più potenti **motori dell'evoluzione** del genoma umano.

3) La proteina **CENP-B**, una componente essenziale del cinetocore, deriva da un'antica trasposasi



RETROTRASPOSONI

- **LINE (Long Interdispersed Nuclear Elements) (elementi autonomi)**

- **SINE (Short Interdispersed Nuclear Elements) (elementi non autonomi)**

(sequenze *Alu* = 1.1 milione di copie nel genoma)

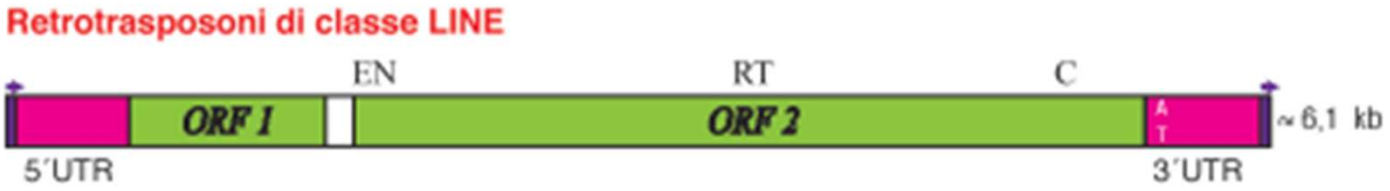
- Trasposoni con **elementi LTR (Long Terminal Repeats) (elementi autonomi):** derivano da retrovirus.

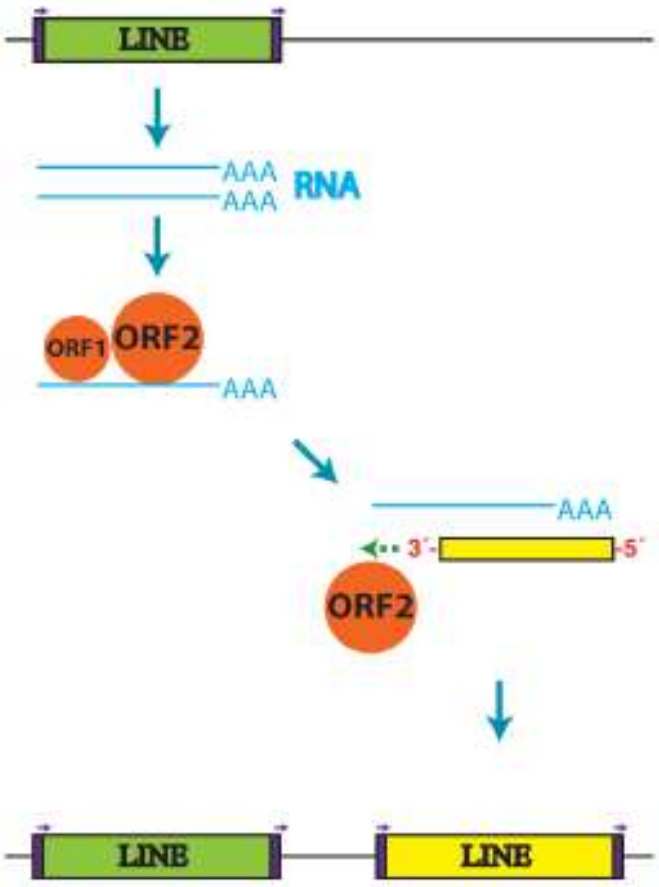
- LINE (Long Interdispersed Nuclear Elements) (elementi autonomi)

LINE-1 (L1): È l'unico retrotrasposone **autonomo** rimasto.

Nel genoma ci sono circa 500.000 copie di LINE-1, ma solo circa **80-100 copie** sono ancora "intatte" e capaci di saltare.

Queste poche copie producono le proteine necessarie per muovere se stesse e anche altri elementi.





La sequenza del trasposone viene trascritta e tradotta nelle proteine orf1 e orf2.

L'RNA rientra nel nucleo associato alle proteine a cui ha dato origine. La proteina orf2 ha attività di trascrittasi inversa e di endonucleasi per cui converte l'RNA in cDNA e lo reinserisce in una nuova localizzazione.

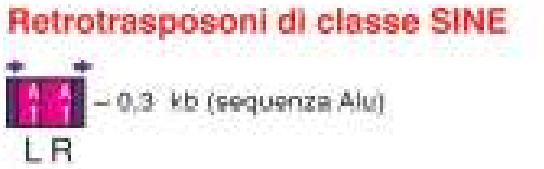
Le attività enzimatiche di orf2 possono essere utilizzate anche da trasposoni che non codificano per essi (elementi non autonomi).

*La retrotrascizione viene attivata sul DNA dall'azione dell' endonucleasi

- SINE (Short Interdispersed Nuclear Elements) (elementi non autonomi)

Sequenze Alu: Sono elementi **non autonomi**. Non sanno saltare da soli perché non hanno geni. Tuttavia, rubano le proteine prodotte dai LINE-1 per farsi copiare e incollare altrove.

Sono estremamente efficienti: ci sono più di un milione di copie di *Alu* nel genoma umano.



- Trasposoni con **elementi LTR (Long Terminal Repeats) (elementi autonomi):** derivano da retrovirus.

•**LTR (Long Terminal Repeats):** segnali di "inizio" e "fine"per la trascrizione.
•***Gag*:** Gene che codifica per proteine strutturali (simili al capside virale).
•**Pol: Trascrittasi Inversa** (per convertire l'RNA in *DNA*) e **Integrasi** (per inserire il nuovo DNA nel genoma).

Nel genoma umano, quasi tutti i retrotrasposoni LTR sono classificati come **HERV** (*Human Endogenous Retroviruses*).

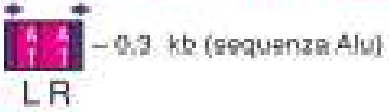
Costituiscono circa l'**8%** del nostro genoma (una quota enorme!).

Sono il risultato di infezioni virali avvenute milioni di anni fa nelle cellule germinali (ovociti o spermatozoi) dei nostri antenati.

Stato attuale: La maggior parte degli HERV nell'uomo è "difettosa" (piena di mutazioni) e non può più produrre virus infettivi o saltare

Retrotrasposoni di classe LINE

Retrotrasposoni di classe SINE



Retrotrasposoni con LTR (retrovirus endogeni)

Trasposoni a DNA

Patologie umane correlabili al «salto» di retrotrasposizioni.

- **Distrofia Muscolare di Duchenne**: Causata dall'inserzione di elementi **Alu** o **LINE-1** nel gene della Distrofina.

- **Tumore al Colon (rari casi)**: Inserzione di un LINE-1 nel gene oncosoppressore **APC**, che disattiva le difese della cellula contro il cancro.

- **Neurofibromatosi di tipo 1**: Inserzione di un elemento Alu che altera lo splicing del gene NF1.

Dove so inseriscono o retrotrasoposoni?
Caso o Destino?

-Più il gene è grande più ha probabilità di essere bersagliato da un retrotrasposone.

-Più la cromatina è aperta più può essere bersagliata.

-I retrotrasposoni saltano di più dove le difese della stabilità del genoma sono più basse. Ad esempio:

Nelle cellule germinali (per garantire l'ereditarietà).

Nei neuroni progenitori, dove il rimodellamento della cromatina è talmente frenetico che i LINE-1 ne approfittano per inserirsi.

Ogni volta che un bambino nasce, c'è una probabilità (circa 1 su 20 neonati) che una nuova inserzione di un retrotrasposone sia avvenuta da qualche parte nel suo genoma. La maggior parte finisce in zone "deserto" e non fa nulla.

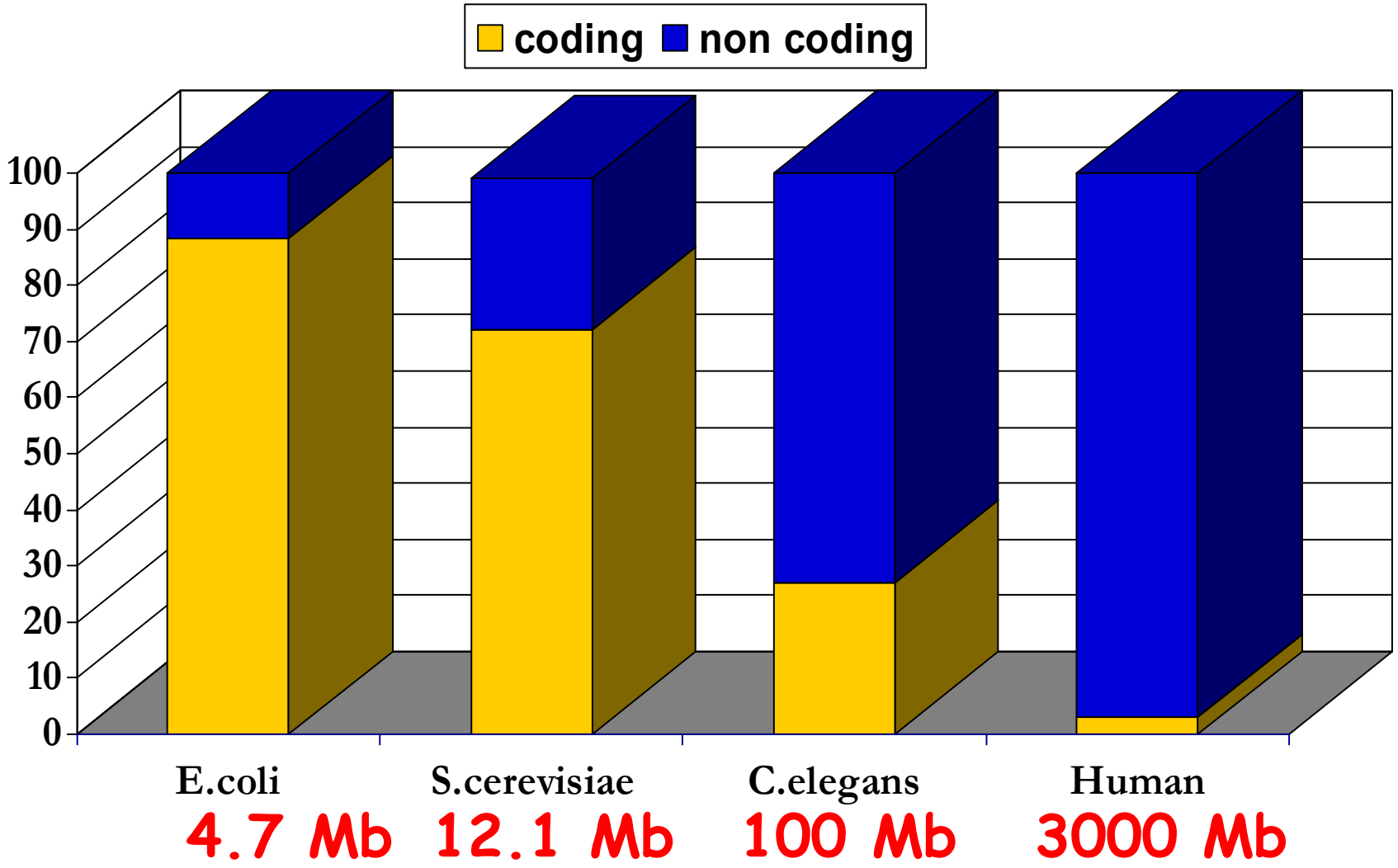
Quasi metà del genoma umano deriva da elementi trasponibili

Nel genoma umano la quantità di DNA derivante da elementi trasponibili è **20 volte** la quantità di DNA che codifica per tutte le proteine umane

Ishihara's test per il daltonismo.
Siamo talmente impegnati nella ricerca di nuovi geni che non riusciamo a vedere ciò che gene non è!

Non possiamo ignorare le sequenze non codificanti alle quali dobbiamo attribuire importanti funzioni regolatorie.

La porzione non codificante dei genomi eucariotici



L'annotazione funzionale delle porzioni non-codificanti del genoma è una delle sfide principali dell'era post-genomica.

Ipotesi sull'origine e la funzione del DNA non codificante

Queste regioni potrebbero essere rimasugli di pseudogeni, che nel corso dell'evoluzione avrebbero perso la loro funzione, anche a causa di eventuali frammentazioni della sequenza codificante.
Il *DNA non codificante* potrebbe avere una funzione protettiva nei confronti delle regioni codificanti. Dal momento che il DNA è continuamente esposto a danni casuali da parte di agenti esterni, infatti, una tanto alta percentuale di DNA non codificante permette di pensare che le regioni ad essere statisticamente più danneggiate siano in realtà non codificanti.
Il DNA non codificante potrebbe anche essere una sorta di *riserva di sequenze* al momento non codificate, ma dalle quali potrebbe emergere un qualche gene in grado di conferire vantaggio all'organismo. Da questo punto di vista, dunque, tali regioni costituirebbero le vere basi genetiche dell'evoluzione.
Parte del *DNA non codificante* è ritenuto essere, più semplicemente, un elemento spaziatore tra geni. In questo modo gli enzimi che hanno rapporti con il materiale genetico avrebbero la possibilità di complessare più agevolmente il DNA. Il DNA non codificante così potrebbe avere una funzione fondamentale pur essendo composto di una sequenza assolutamente casuale.
Alcune regioni di DNA non codificante potrebbero avere una funzione regolatoria sconosciuta: potrebbero ad esempio controllare l'espressione di alcuni geni o lo sviluppo di un organismo dallo stato embrionale fino a quello adulto.
Nel *DNA non codificante* potrebbero essere contenute numerose sequenze trascritte in non coding RNA (si ritiene possano essere molti di più di quelli attualmente noti).

Alcune teorie puntano invece a confermare che tale DNA non abbia in effetti alcuna funzione. In un recente esperimento è stata rimossa una quantità di *DNA non codificante* dal genoma murino pari all'1%. I topi sottoposti al trattamento non hanno mostrato alcun fenotipo. Ciò può comunque essere interpretato in due modi: il DNA non codificante non ha effettivamente nessuna funzione, oppure i ricercatori non sono stati in grado di sviluppare un metodo di rilevazione tale da osservare cambiamenti fenotipici nei topi.

The ENCODE Project Consortium

La "Encyclopedia of DNA Elements (ENCODE)" è un progetto di ricerca pubblico promosso da US National Human Genome Research Institute (NHGRI) nel settembre 2003.

Il progetto ENCODE punta ad identificare tutti gli elementi funzionali del genoma umano, al di là dei geni codificanti per proteine.

Il progetto coinvolge un consorzio mondiale di gruppi di ricerca.

I dati ottenuti sono accessibili attraverso database pubblici.

Obiettivo del progetto è l'identificazione del cosiddetto

Reguloma, cioè di quella varietà di elementi del DNA (promotori, enhancer, silencer, regioni della cromatina suscettibili di intense modificazioni epigenetiche, geni per trascritti regolatori…) che possono regolare l'espressione dei geni codificanti proteine.

La disfunzione del reguloma può essere alla base di molte patologie alle quali ancora non è stata ancora attribuita una base genetica e molecolare.