22.
L’organizzazione del genoma:
Eucarioti

II. EUKARYOTIC GENOME ORGANIZATION

- Multiple genomes: nuclear, mitochondria, chloroplasts
- Plastid genomes resemble prokaryotic genomes
- Multiple linear chromosomes, total size 5-15,000 MB,
- 5000 to 50,000 genes
- Mono-cistronic transcription units
- Discontinuous coding regions (introns and exons)
- Large amounts of non-coding DNA
- Transcription and translation take place in different compartments
- Variety of RNAs:
  - Coding (mRNA, rRNA, tRNA),
  - Non-coding (snRNA, snoRNA, microRNAs, etc).
- Often diploid genomes and obligatory sexual reproduction
- Standard mechanism of recombination: meiosis.

Hierarchy of gene organization

Gene – single unit of genetic function

Operon – genes transcribed in single transcript

Regulon – genes controlled by same regulator

Modulon – genes modulated by same stimulus

Element – plasmid, phage, chromosome,

Genome

** **order of ascending complexity**

Quali informazioni ha fornito il sequenziamento dei genomi?

Grande genoma: molti geni?

- La dimensione del genoma è correlata al numero di geni che esso contiene?

Grande genoma = Molti geni?

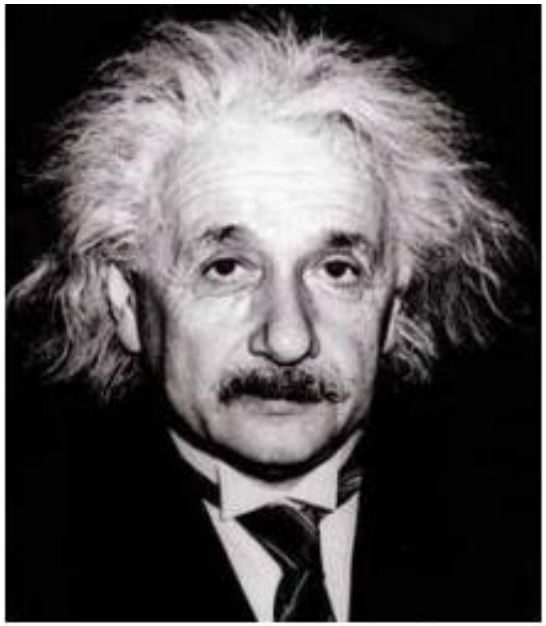
Non è detto!

Saccharomyces cerevisiae **(lievito)**	***Homo sapiens*** **(uomo)**
Genoma: 12 Mb (1/260 la grandezza del genoma umano)	**Genoma: 3200 Mb**
Geni: 5800	**Geni: 25.000**

- Se dimensione e numero di geni fossero correlate,
  - il lievito dovrebbe contenere, in proporzione a quello umano, solo 100 geni,
  - mentre ne contiene 5800!

La grandezza del genoma non è correlata alla complessità dell'organismo

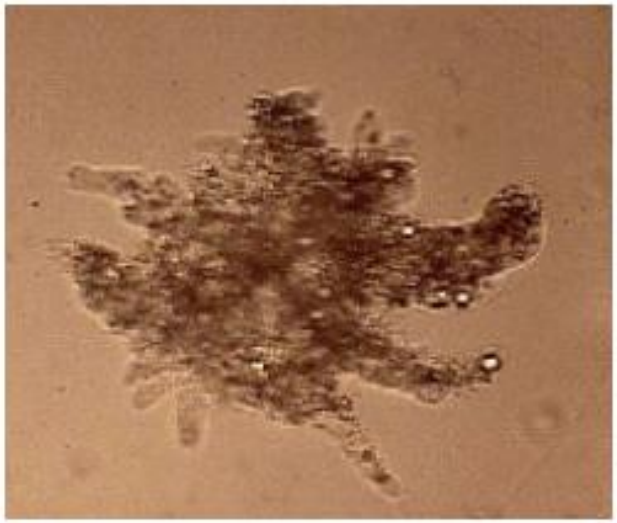
Paradosso del valore C



Homo sapiens $3x10^9$ pb



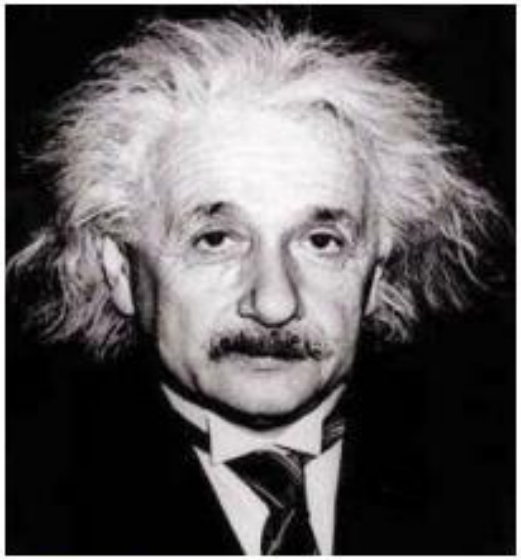
Allium cepa $1.5x10^{10}$ pb



Amoeba dubia $6,7x10^{11}$ pb

QUINDI:

- La complessità dell'organismo umano non è correlata al numero di geni
- Paradosso del valore N



~25.000 geni



~22.000 geni

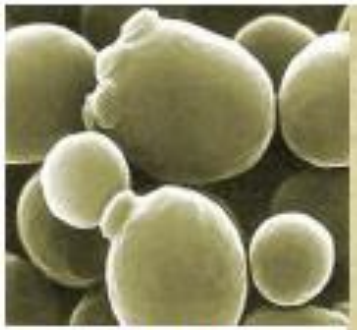


~25.000 geni

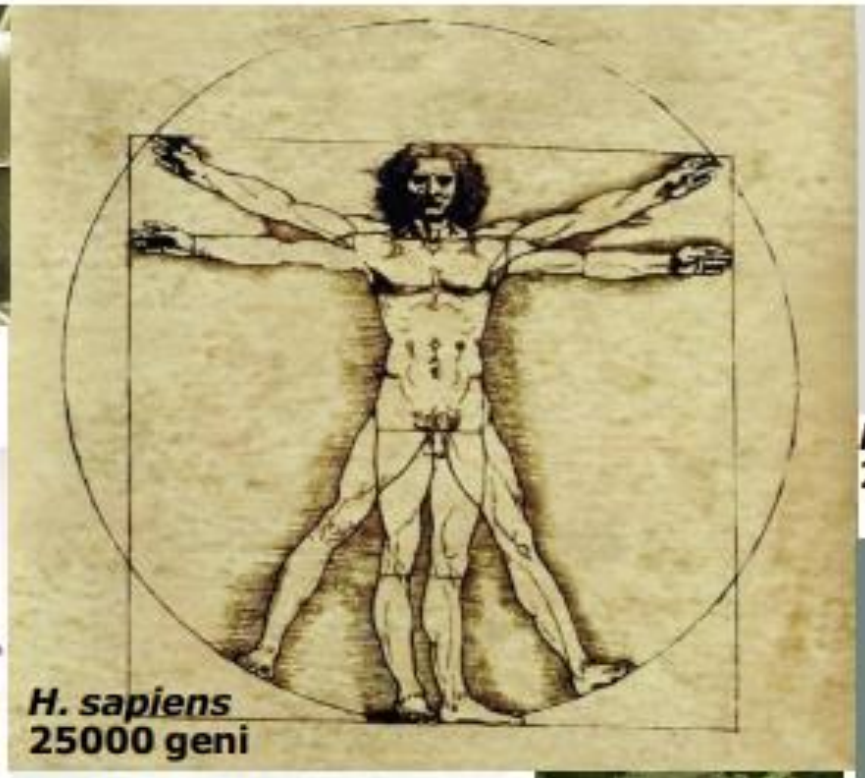


~60.000 geni

Il numero dei geni umani codificanti per proteine è di gran lunga inferiore rispetto a quanto atteso



S. cerevisiae
6000 geni



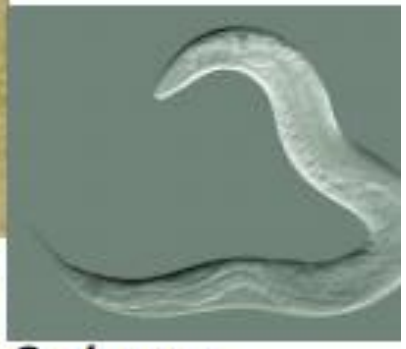
H. sapiens
25000 geni



M. musculus
22000 geni



D. melanogaster
13000 geni



C. elegans
21000 geni



D. rerio
26000 geni

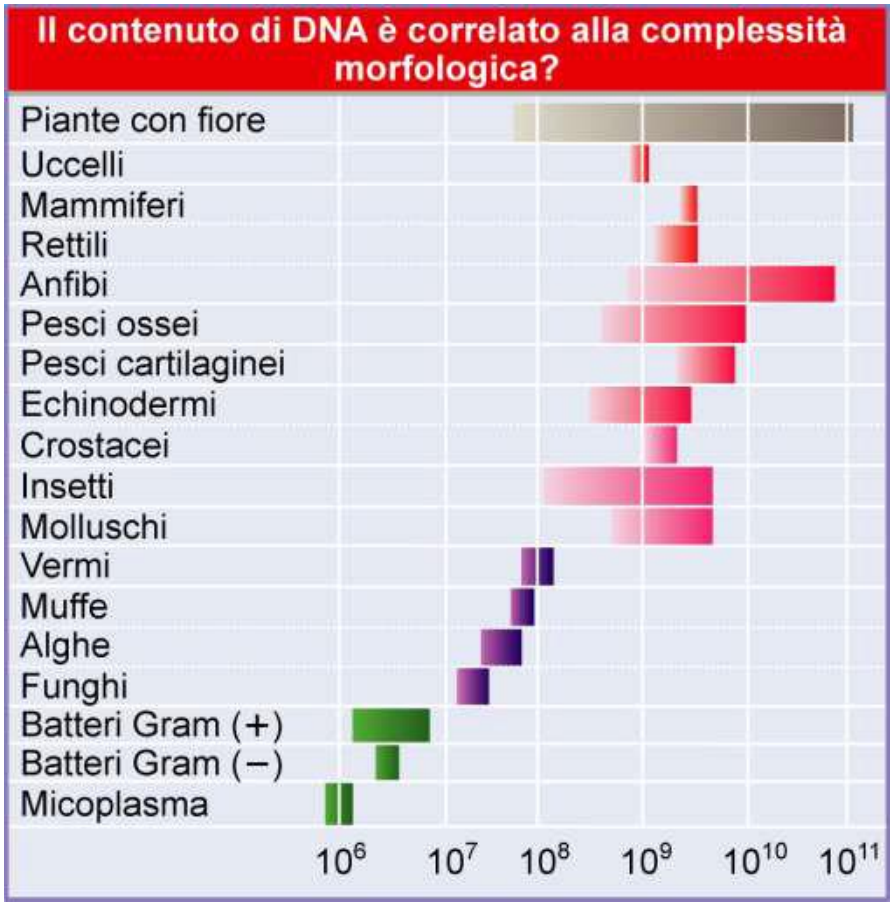


Z. mays
26000 geni

Perchè alcuni genomi sono così grandi?

➢ La **quantità di DNA contenuta nel genoma aploide** è una **caratteristica propria di ciascuna specie vivente** ed è definita come **valore C**.

- Il valore C varia enormemente, da <10^6 pb per un micoplasma a >10^{11} pb per alcune piante o anfibi.
- La **dimensione minima del genoma cresce con la complessità dello organismo.**
- Ma in alcuni gruppi c'è una enorme variabilità nel contenuto di DNA di specie simili tra loro
  - per esempio, anfibi e piante con i fiori



➢ Il **paradosso del valore C:**

- **si riferisce alla mancanza di correlazione tra le dimensioni del genoma e la complessità genetica.**

Solutions ?

Solution 1 to the N-value paradox:

Many protein-encoding genes produce more than one protein product (*e.g.*, by alternative splicing or by RNA editing).

Solution 2 to the N-value paradox:

We are counting the wrong things, we should count other genetic elements (*e.g.*, small RNAs).

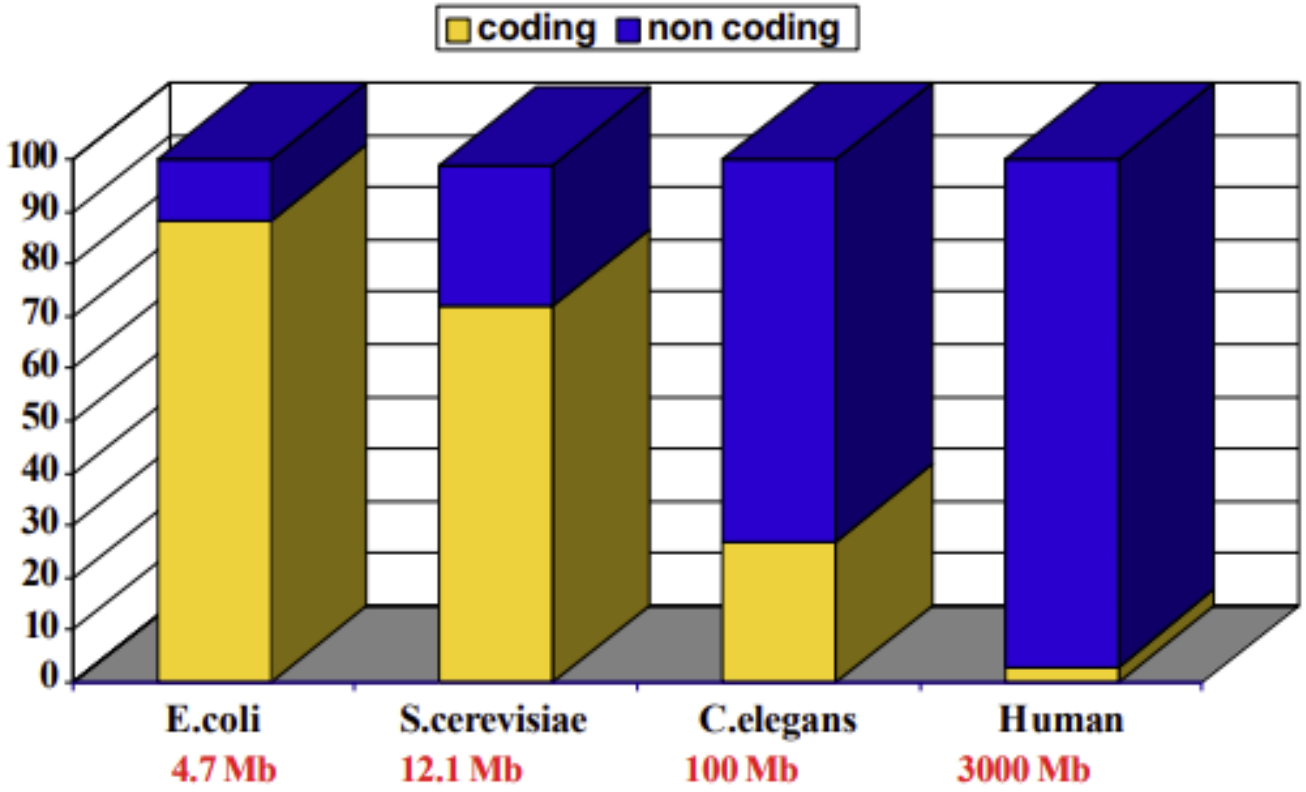
Solution 3 to the N-value paradox:

We should look at connectivity rather than at nodes.

These should be exciting and should stimulate the next generation of genomic investigation.

La frazione non codificante del genoma umano è > 98%

La sfida post-genomica



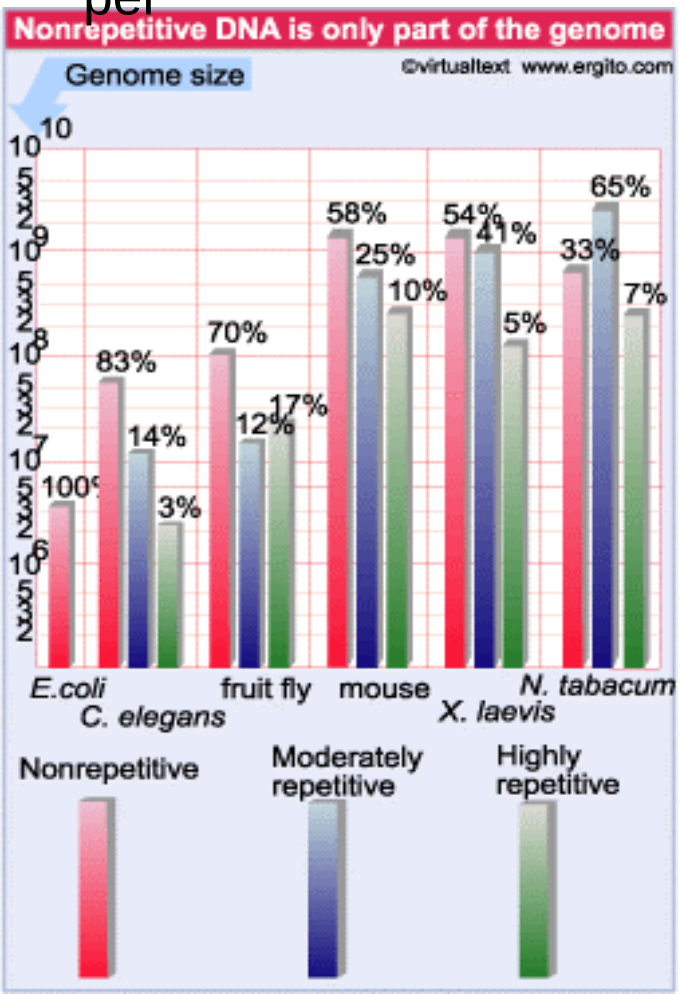
- L'annotazione funzionale del genoma delle porzioni non codificanti del genoma è una delle principali sfide dell'era post-genomica

La maggior parte del DNA in genomi grandi non è codificante

- Genomi complessi hanno circa da **10x a 30x più DNA** rispetto a quello richiesto per codificare tutti gli RNA o proteine di un organismo.
- Contribuiscono al **DNA non-codificante:**
  - **Introni in geni**
  - **Elementi regolatori di geni**
  - **Copie multiple di geni, inclusi:**
    - **pseudogeni**
    - **sequenze intergeniche**
    - **ripetizioni intersperse**

I genomi eucariotici contengono sia sequenze uniche che DNA ripetitivo

- **DNA non ripetitivo:** formato da sequenze uniche, una copia nel genoma aploide.
- **DNA ripetitivo**: formato da sequenze che sono presenti in più di una copia per genoma aploide



- **DNA moderatamente ripetitivo**

  sequenze relativamente corte presenti tra le **10 e 1000 volte**; sono intersperse nel genoma

- **DNA altamente ripetitivo**

  corte sequenze (tipicamente <100 paia di basi), presenti **molte migliaia di volte.**

  - La **percentuale di DNA ripetitivo nei genomi dei singoli organismi varia moltissimo**

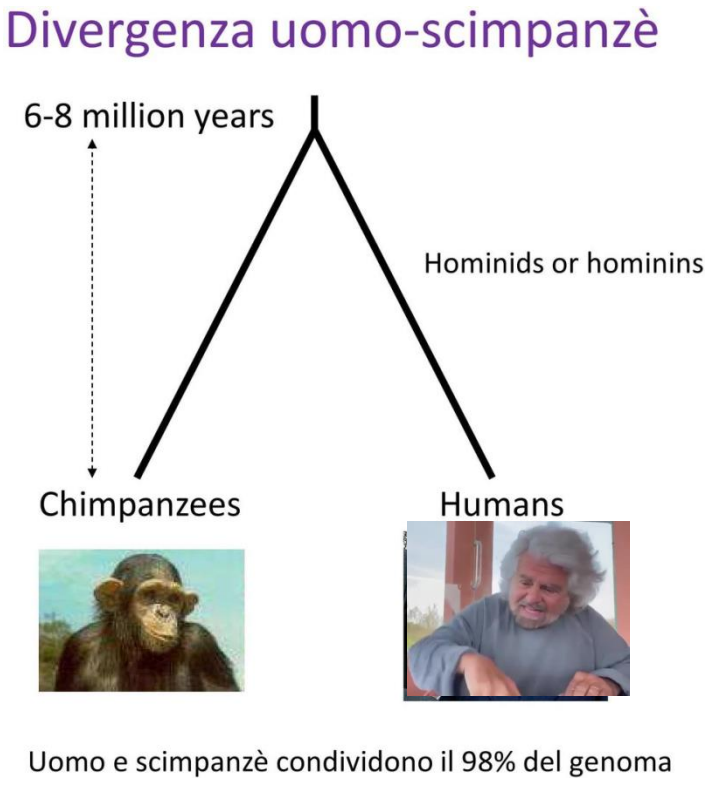
- i procarioti contengono solo DNA non ripetitivo
- negli eucarioti inferiori il DNA ripetitivo è 20%
- nelle cellule animali, fino a metà del DNA è ripetitivo
- nelle piante e negli anfibi il DNA ripetitivo può rappresentare fino all'80% del totale.

Componenti distinte in genomi complessi

- **DNA altamente ripetitivo**
  - *R* (frequenza di ripetizione) ≥100.000
  - Quasi nessuna informazione, bassa complessità
- **DNA moderatamente ripetitivo**
  - 10<*R*<10.000
  - Scarsamente informativo, moderata complessità
- **DNA "Singola copia"**
  - *R*=1 or 2
  - Molto informativo, alta complessità

I genomi dell'uomo e dello scimpanzé sono uguali all'incirca al 98,5%





Uomo e scimpanzè condividono il 98% del genoma

- Ancor oggi l'uomo è più scimmia di qualsiasi scimmia (Nietzsche)

- 87% dei geni umani ha un omologo nello scimpanzé
- 30% delle proteine omologhe sono identiche

Da cosa dipende la differenza?





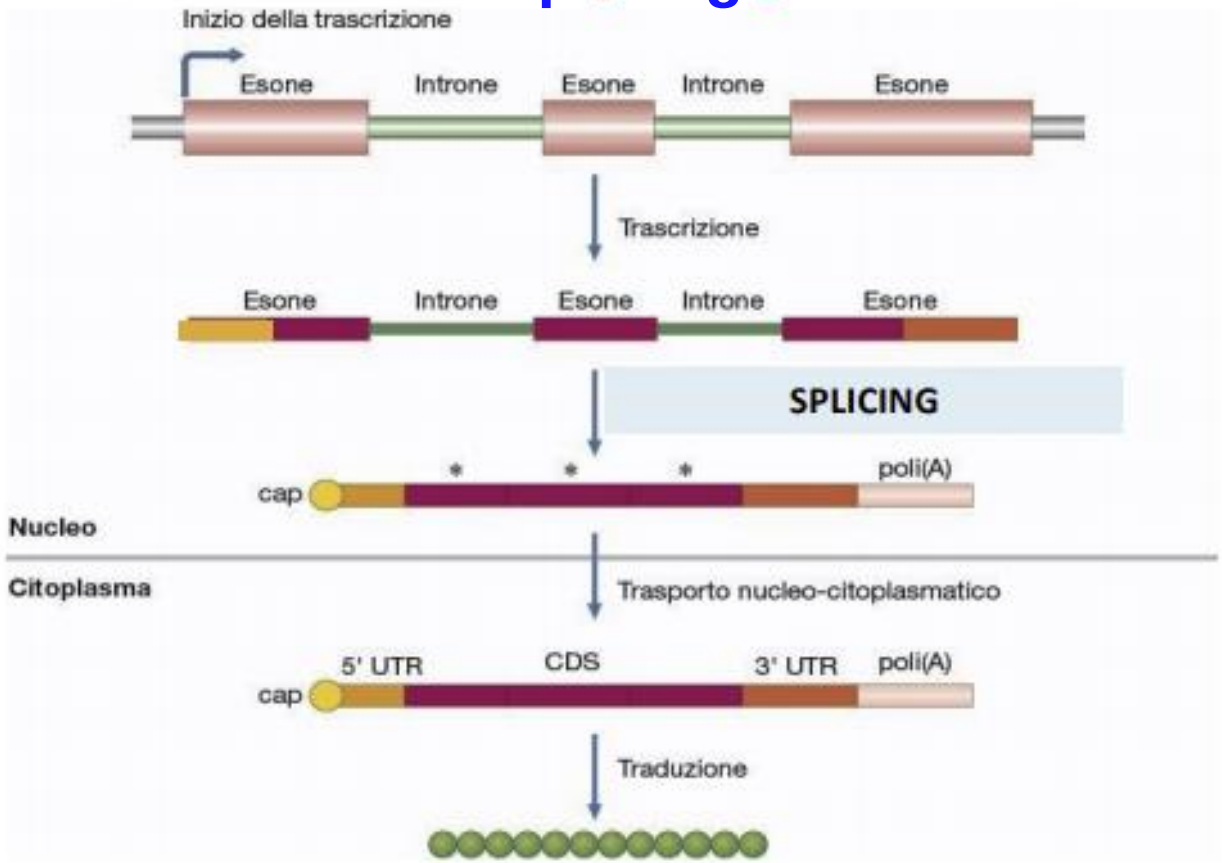


- E’ evidente che
  - **geni simili** e
  - interazioni fra gruppi di geni simili

    possono generare **‘’programmi**’’molto diversi

Da cosa dipendono le differenze?

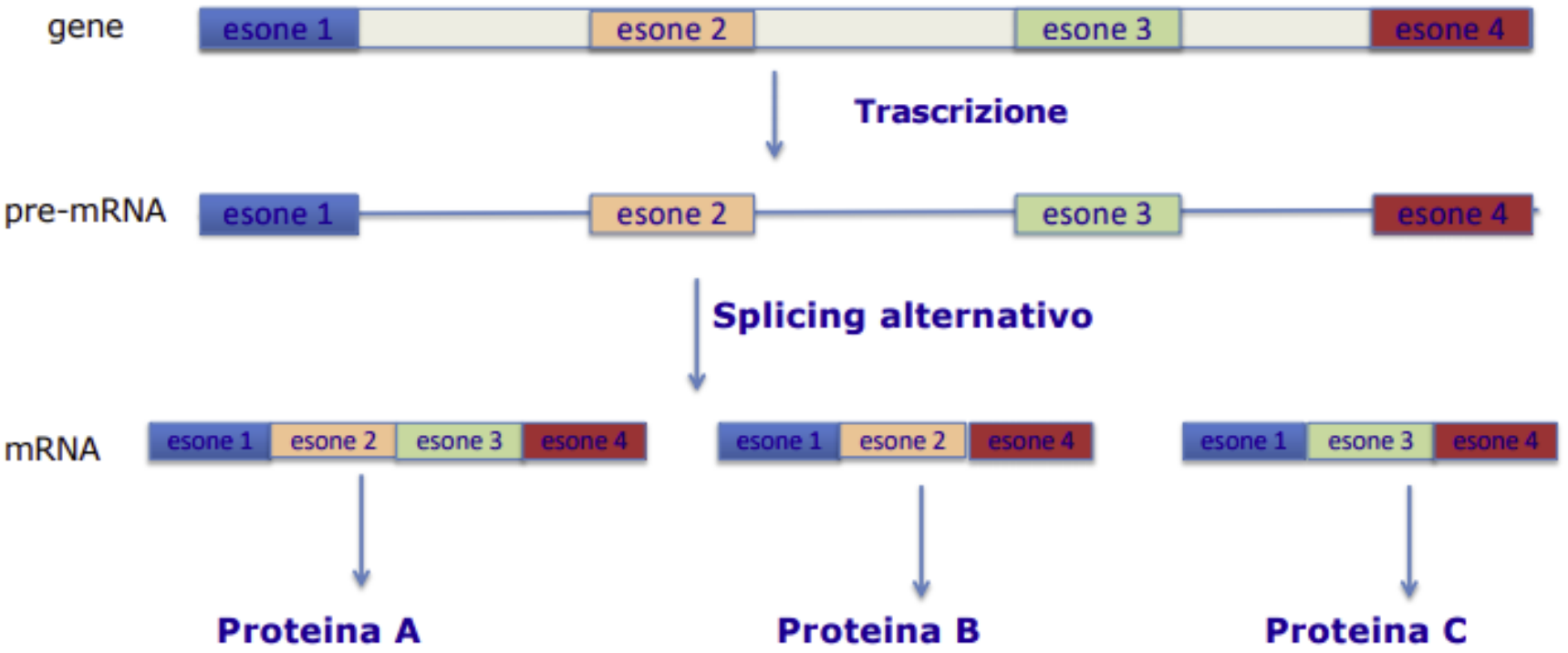
1. Lo splicing



- I geni espressi sotto forma di **lunghi RNA precursori** nel nucleo sono sottoposti
  - ad un processo di maturazione questo
  - comporta la rimozione di tratti polinucleotidici interni **chiamati introni**
  - associazione delle restanti regioni chiamate **esoni (meccanismo di splicing**).
- L'mRNA maturo è quindi trasportato nel citoplasma per essere tradotto in proteina

2. Lo splicing alternativo

- E' un meccanismo attraverso il quale uno stesso pre-mRNA può subire eventi di splicing alternativo differenti con la produzione di mRNA maturi alternativi che **possono dare origine a proteine diverse.**
- **Le proteine alternative generate dallo splicing (isoforme**) possono avere funzioni
  - **simili,**
  - **distinte**
  - o persino **antagoniste.**



Lo splicing alternativo contribuisce all'espansione delle capacità di espressione del genoma umano

- Oltre il 95% dei geni umani è espresso mediante splicing alternativo
- Il numero medio trascritti prodotti da un gene è di circa 9-10

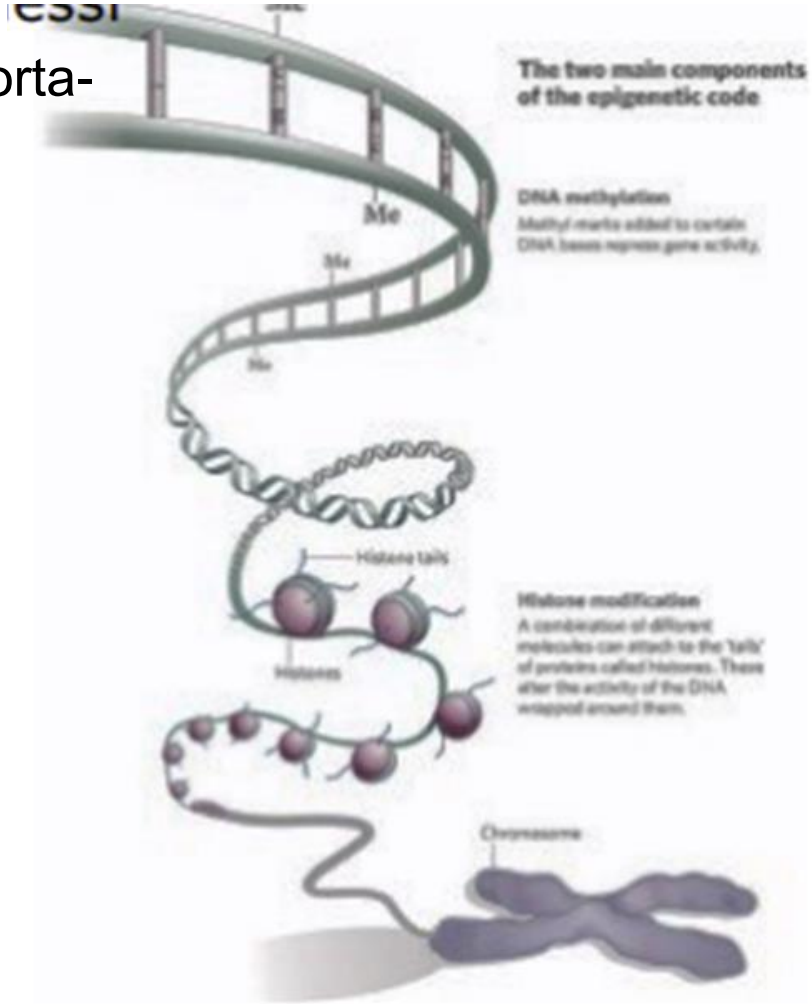
Numero di trascritti umani è maggiore di 200.000

➢ **Aumento della complessità del**
- **trascrittoma e del proteoma**
- **di almeno un ordine di grandezza**

3. Regolazione epigenetica

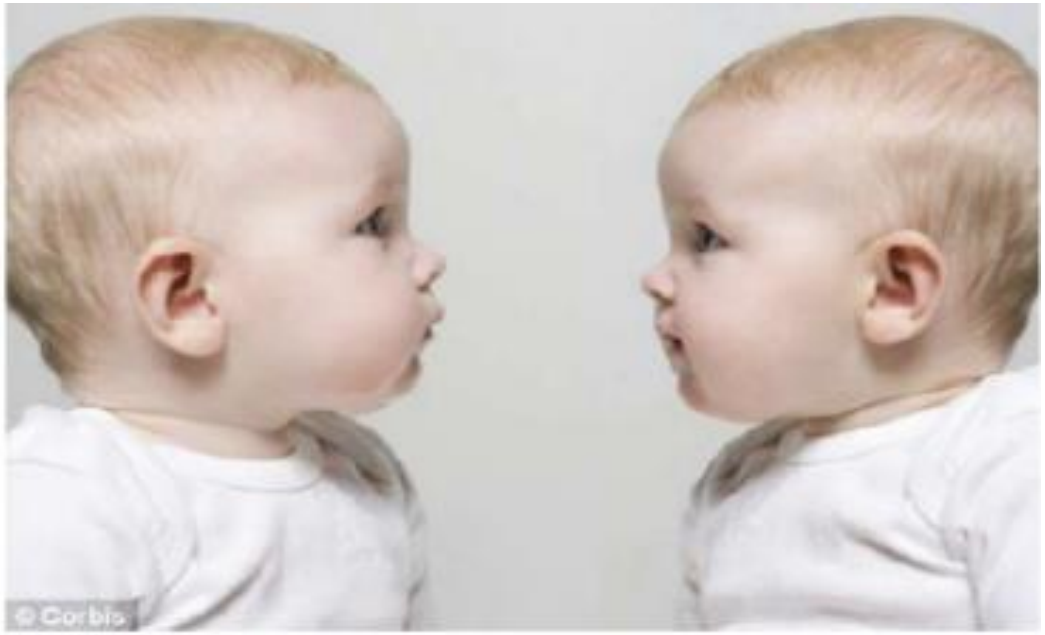
- La **regolazione epigenetica** indica un meccanismo di controllo **dell'espressione dei geni che non è dipendente dalla sequenza del DNA**.
- Ciò' significa che il **fenotipo di una cellula è determinato non solo dal** genoma ereditato, quanto anche da una sorta di ’impronta" che ne influenza il comportamento funzionale.
- Cambiamenti di questi caratteri possono avvenire attraverso le divisioni cellulari.

- I principali meccanismi alla base dell’ereditarietà epigenetica sono:
  - **metilazione del DNA;**
  - **rimodellamento della struttura della cromatina** (modificazione dei nucleosomi)



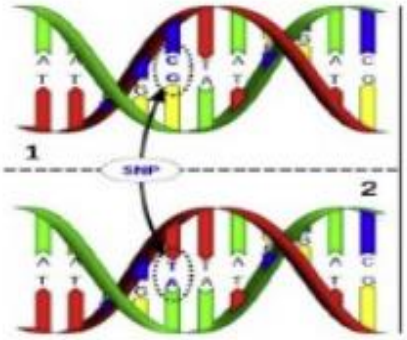
Regolazione epigenetica nell'uomo

- I gemelli omozigoti hanno identico corredo genetico.
- Ma due gemelli omozigoti non sono identici dal punto epigenetico !
  - Due gemelli geneticamente predisposti all'insorgenza di una malattia, potranno o non sviluppare entrambi la malattia.
  - In tal caso entrano in gioco **i fattori epigenetici** che determinano l'insorgenza della malattia in uno e non nell'altro.



4. La variabilità genetica

- La maggior parte delle differenze tra individui sono rappresentate da **cambiamenti di singole basi nella sequenza del DNA chiamati**
  singoli polimorfismi nucleotidici (SNPs)



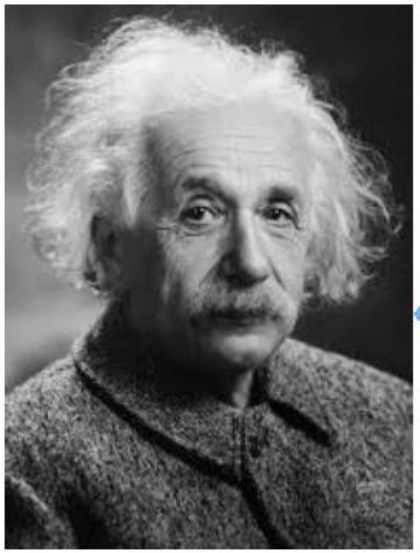


Variabilità genetica

- Il 99.9% del DNA è identico in tutti gli individui
- Lo 0.1% del DNA mostra variabilità



La variabilità genetica









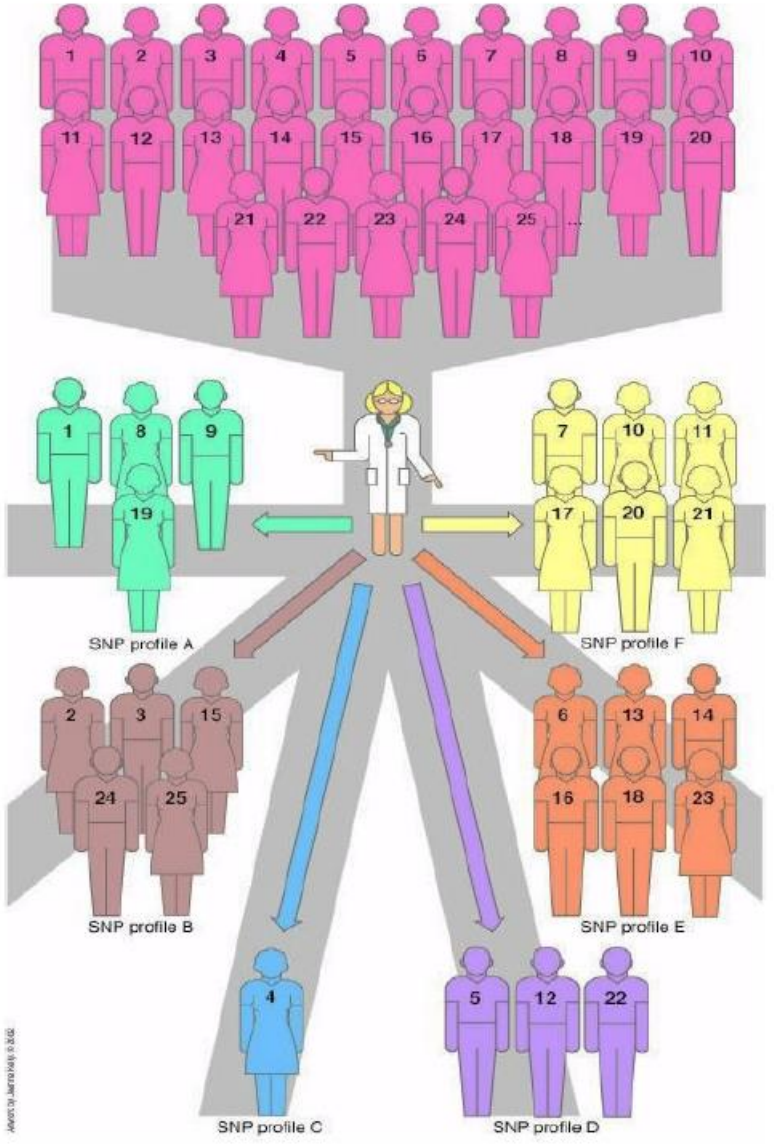
- La **componente genetica individuale** è alla base della:
  - **Variabilità fenotipica** (diverse altezza, colore degli occhi e capelli, ecc)
  - **Risposta individuale** all'ambiente;
  - Diversa **predisposizione allo sviluppo** di malattie complesse;
  - Diversa risposta ai farmaci

Profilo individuale degli SNPs

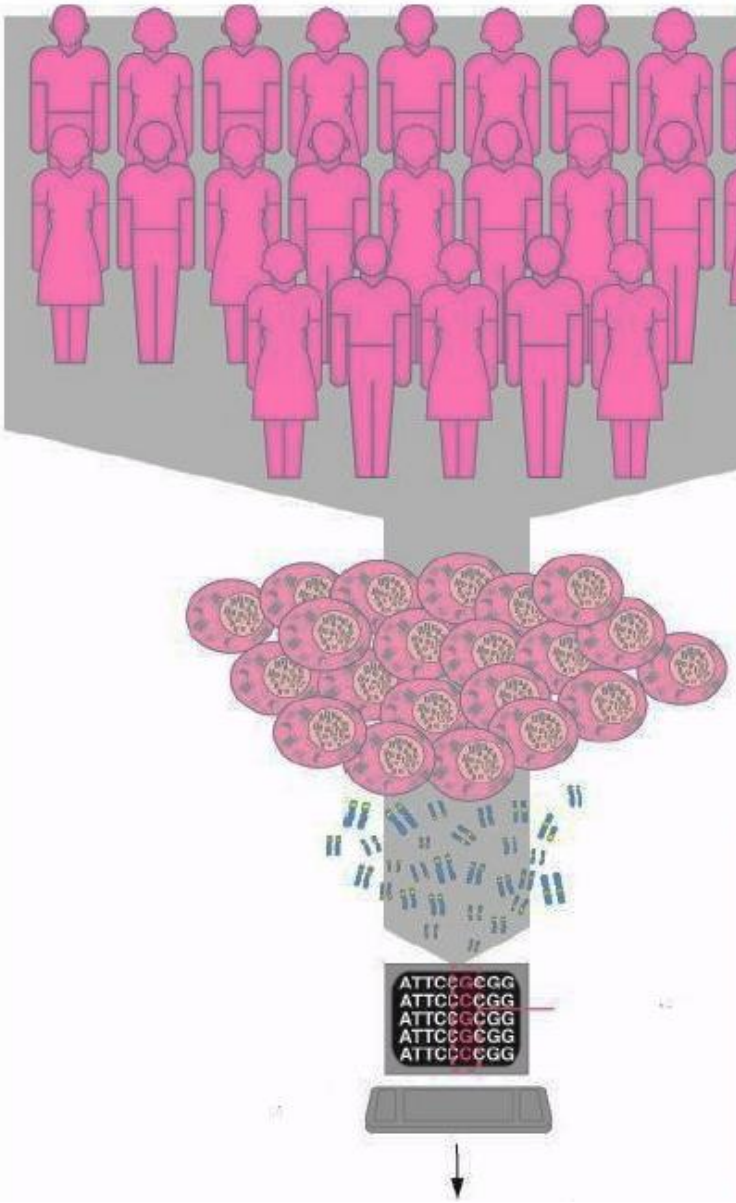
- **Il genoma di ciascun individuo contiene un assortimento diverso di SNPs**

- Gli individui possono essere riuniti in classi in base al loro profilo di SNPs
- **Il profilo di SNPs è importante** per identificare la
  - **risposta ad un farmaco**
  - **predisposizione ad una malattia**

- Si possono individuare delle correlazioni tra il profilo di SNPs e la risposta ad uno specifico trattamento farmacologico

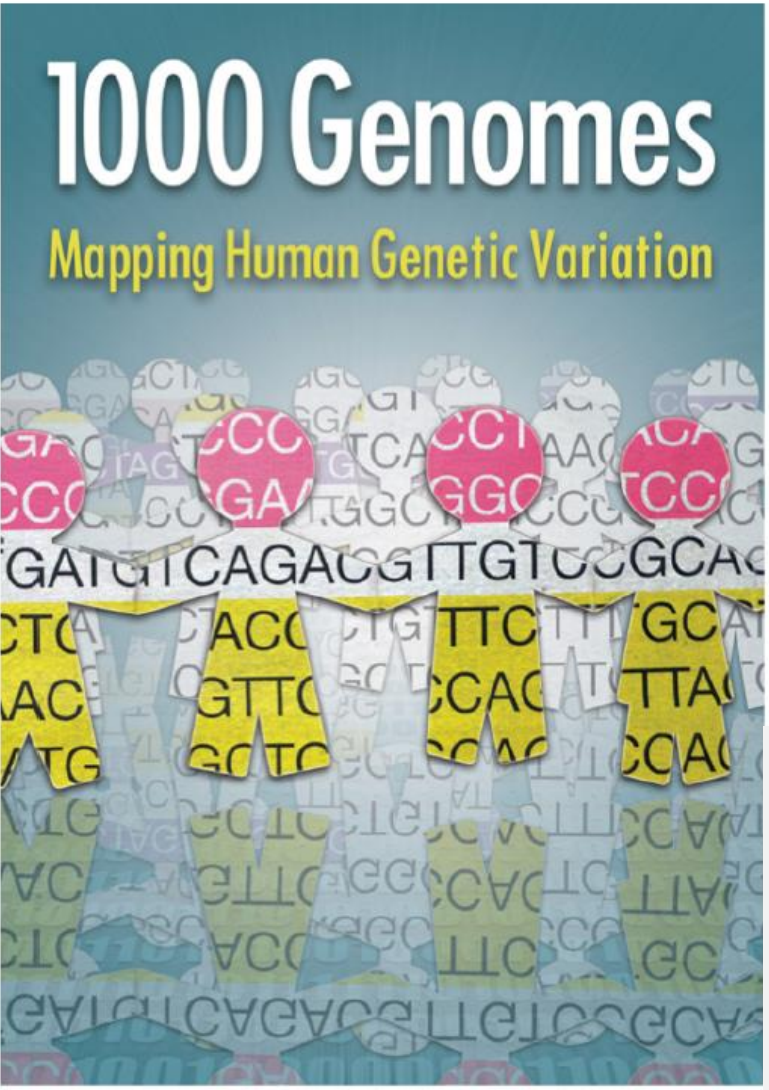


Identificazione degli SNPs



Sequenziamento del genoma

Mappatura degli SNPs:
SNP #1 nella posizione x sul cromosoma n;
SNP #2 nella posizione y sul cromosoma m;
ecc..



2008

1000 Genomes Project

Scopo: identificare le varianti nel genoma umano con una frequenza di almeno 1% nella popolazione

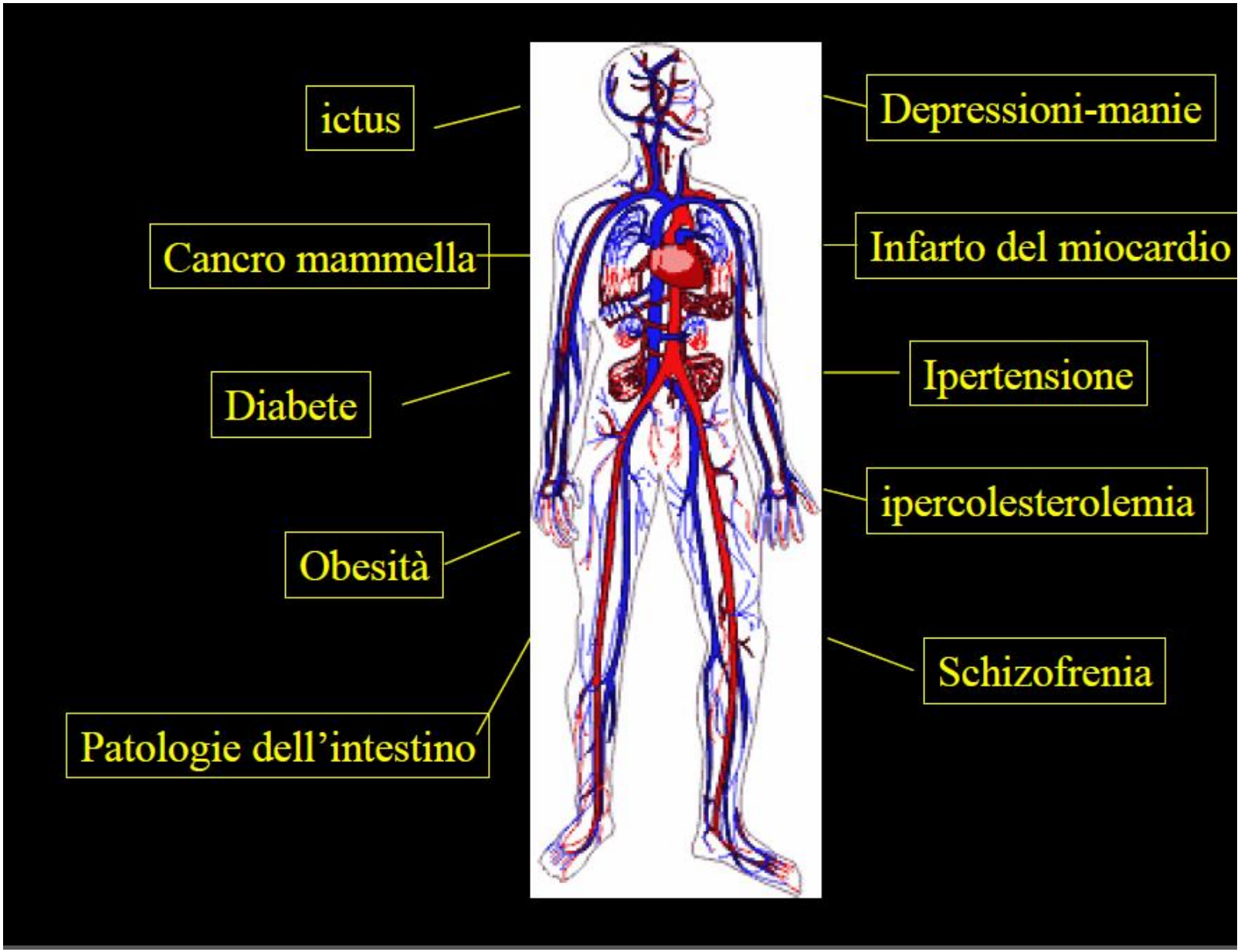
Sequenziati i genomi di 2500 individui appartenenti a 26 popolazioni diverse

I numeri

- **88 milioni varianti SNP**
- **4-8 SNPs presenti in ogni gene**
- **300 - 1000 bp separano ogni SNP**
- **150.000 SNP non sinonimi**
- **6% SNP rari (< 20%)**
- **53% SNP comuni (> 20%**

Ad alcuni loci le persone hanno sequenze nucleotidiche differenti




ictus
Cancro mammella
Diabete
Obesità
Patologie dell'intestino
Depressioni-manie
Infarto del miocardio
Ipertensione
ipercolesterolemia
Schizofrenia


HEART DISEASE
CANCER
STROKE

I Genomi degli Eucarioti: Caratteristiche

- I genomi dei mammiferi, e l'uomo ne è un tipico esempio, eucarioti hanno **una densità genica molto ridotta.**
- In media, i **geni codificanti** per proteine occupano solo il 1-3% dell'intero genoma.
- La scarsa compattezza del genoma nucleare è **dovuta alla struttura discontinua dei geni**, **con introni che nei mammiferi possono** raggiungere dimensioni intorno a 20-30 kb (ed oltre)

- **presenza di elementi ripetuti** (44% del genoma).
- I geni eucariotici sono monocistronici, tuttavia strutture simili agli operoni batterici sono state descritte in C. elegans.

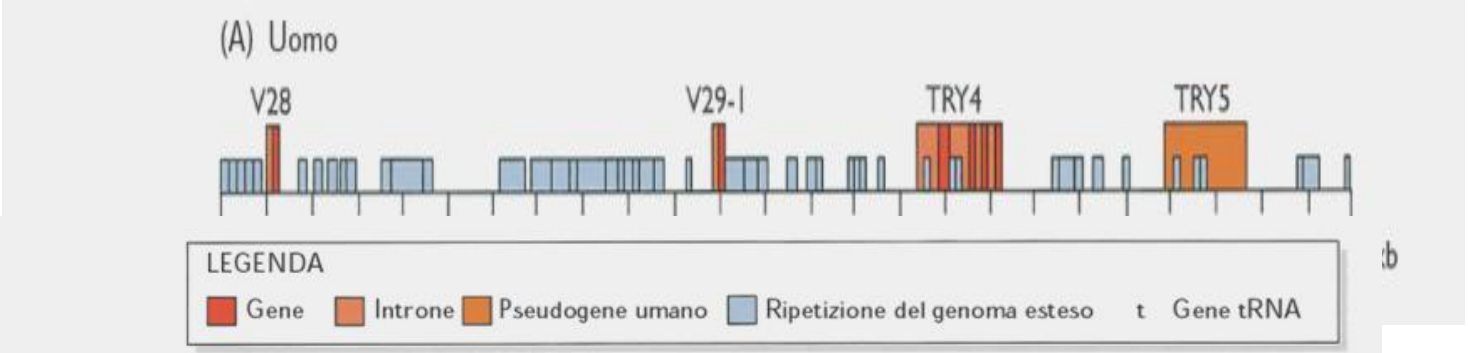


Figura 2.2 Confronto tra il genoma umano, di lievito, del moscerino della frutta, di mais e di *E. coli*.

I geni distribuiti sono uniformemente sui cromosomi?

- Alcuni cromosomi sono relativamente poveri di geni, ovvero **hanno >25% delle loro sequenze come "deserte"**:

  regioni più lunghe di 500 kb dove non ci sono geni.



I Genomi degli Eucarioti:Compattezza

- Nelle piante infine troviamo delle situazioni ancora più spinte verso la **diminuzione della compattezza**

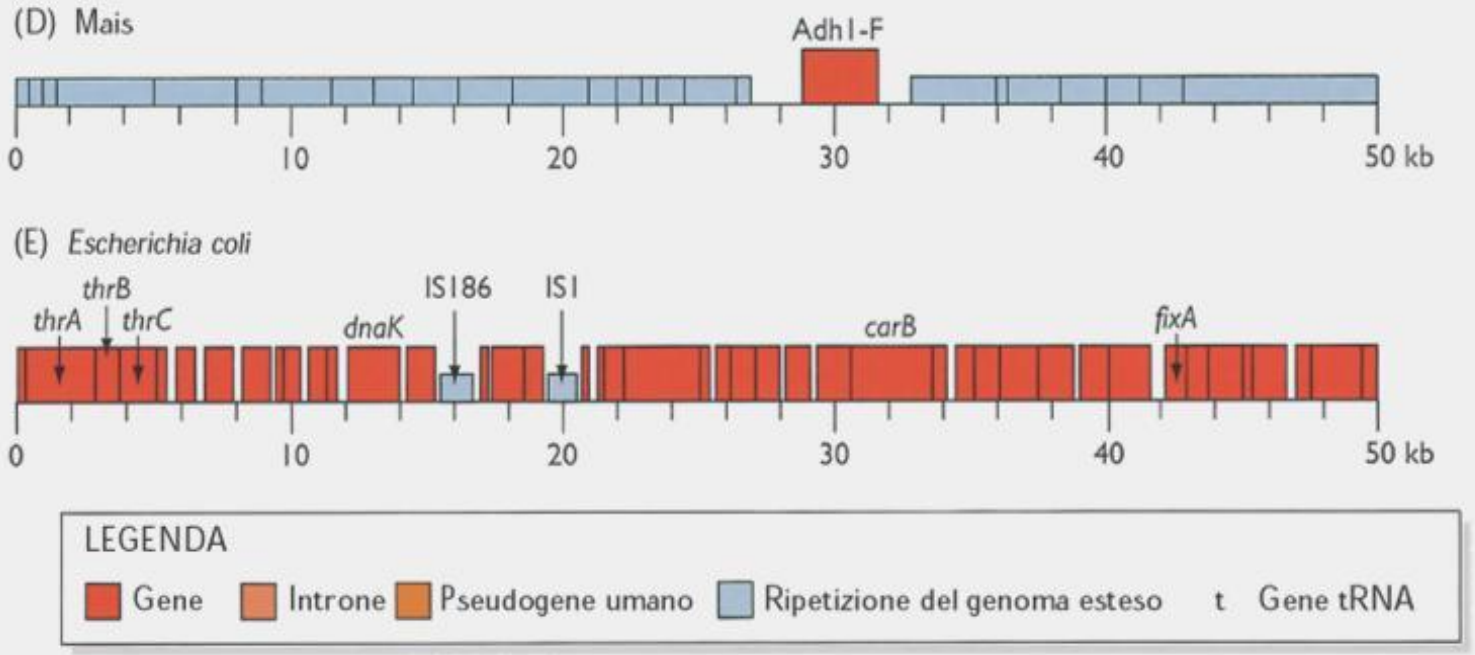


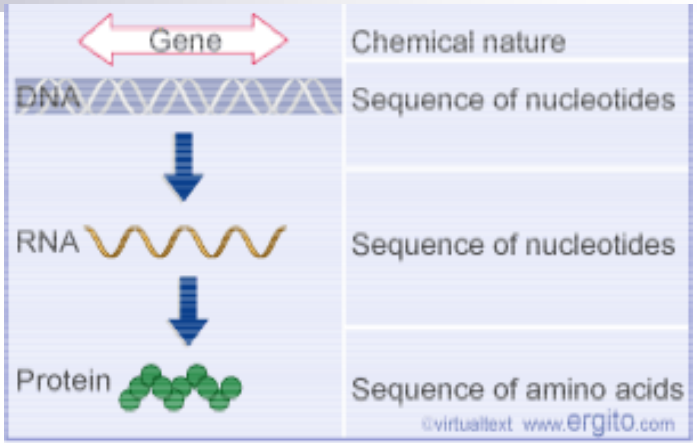
Figura 2.2 Confronto tra il genoma umano, di lievito, del moscerino della frutta, di mais e di *E. coli*.

1. Geni che codificano proteine

- Tali geni variano ampiamente nelle dimensioni e organizzazione
- **Assenza di introni**: Alcuni geni non hanno alcun introne
  Un esempio molto noto sono i **geni degli istoni.**

- **Alcuni geni sono molto grandi:** distrofina (associata con la distrofia muscolare di Duchenne) è pari a 2.4 Mbp ed impiega ~16 hr per formare il trascritto.
  Più del 99% di questo gene è formato da introni (79 introni in totale).
- Geni che sono altamente espressi generalmente hanno introni piccoli.
- La maggior parte degli esoni sono corti: in media 200 bp.
- Le dimensioni degli introni variano ampiamente, da 10 a milioni di bp.

I geni eucariotici sono interrotti



- Fino a quando i geni eucariotici non sono stati caratterizzati a livello molecolare, si pensava che essi avessero la stessa organizzazione dei geni procariotici:
  un frammento di DNA colineare con una proteina.

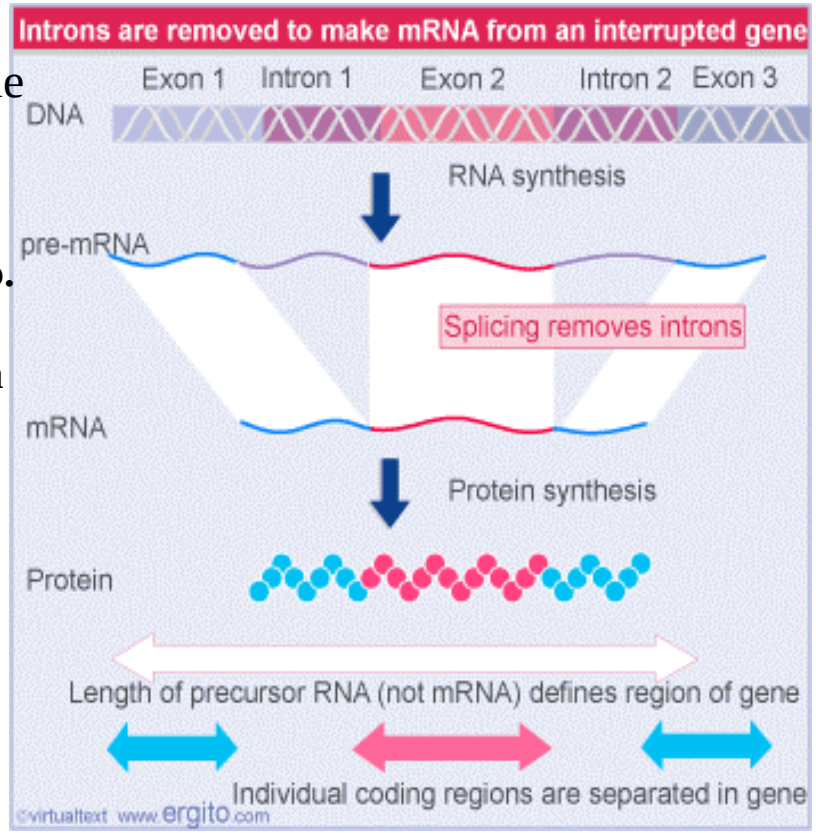
•I geni contengono sequenze addizionali che si trovano tra le sequenze codificanti e interrompono le sequenze che rappresentano le proteine.

•Esoni

- **sono le sequenze presenti nell'RNA maturo. Per definizione, un gene comincia e finisce con esoni, che corrispondono alle estremità 5' e 3' dell'RNA.**

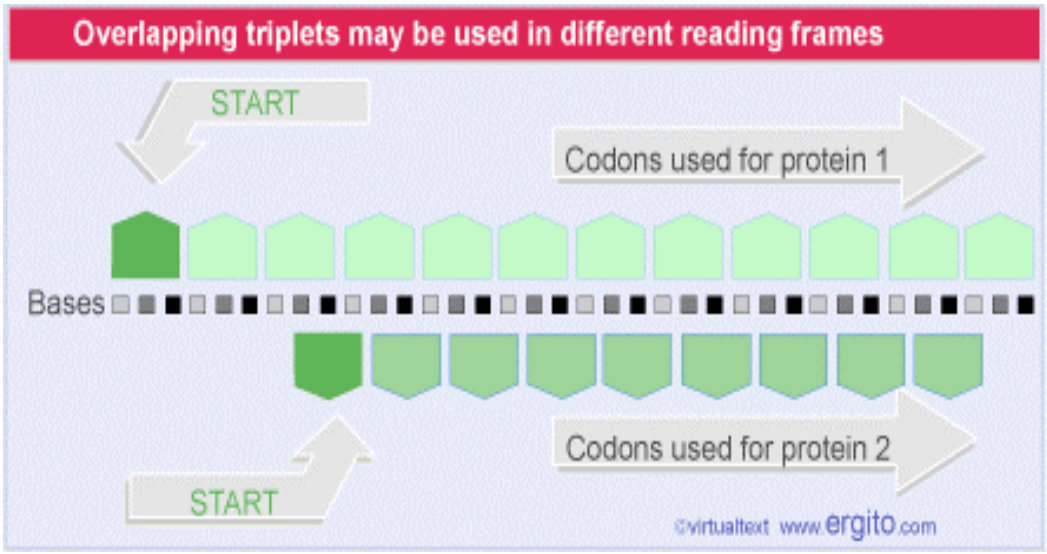
•Introni

- sono rimossi dal trascritto primario:
  il processo è chiamato **montaggio dell'RNA (splicing)**.



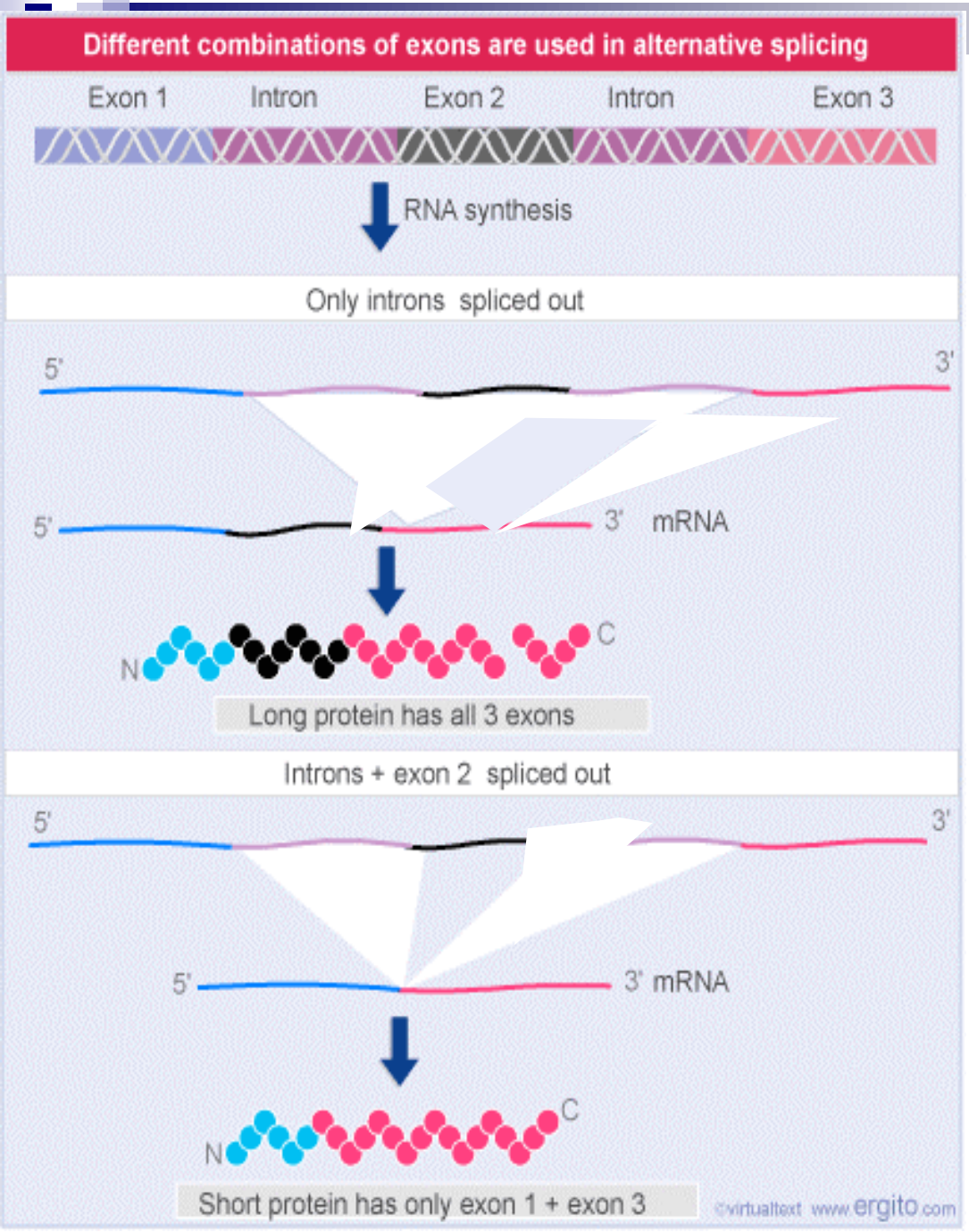
Geni sovrapposti

- In alcuni casi due geni si **sovrappongono** in maniera più complessa:
  sono stati trovati geni in cui la stessa sequenza di DNA è divisa tra due proteine *non omologhe*.



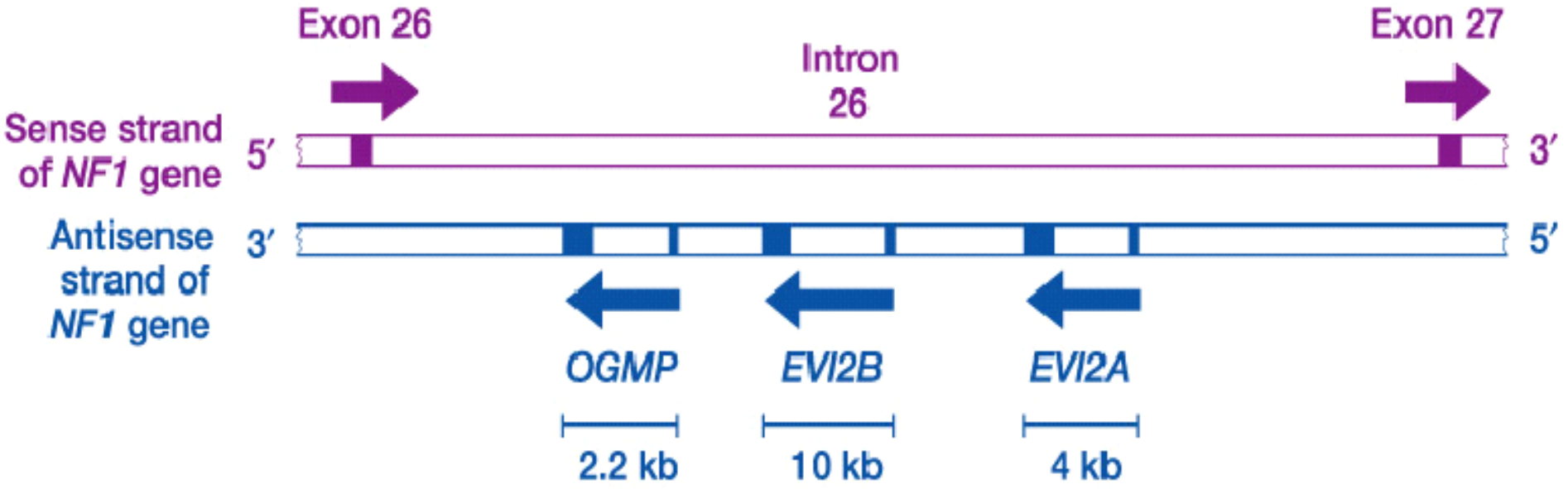
- **Ciò si ottiene leggendo la sequenza del DNA in due moduli di lettura (*reading frame*) diversi.**
- All'interno del primo modulo di lettura aperto (ORF) si trova un altro modulo di lettura.
- **I due moduli di lettura coesistono nella stessa sequenza di DNA.**
- Il tratto di sovrapposizione varia ampiamente, ma di solito è relativamente corto.
- Organizzazioni come queste si ritrovano in alcuni geni del DNA dei mitocondri (organelli subcellulari).

Splicing alternativo



- Un **singolo tipo di trascritto può essere giuntato (spliced) in maniera alternativa**.
- In una modalità i due introni sono rimossi e i tre esoni vengono uniti assieme.
- Nella modalità alternativa, il secondo esone non viene riconosciuto e viene rimosso un unico introne formato da introne 1, esone 2 e introne 2 – esone 2 è stato considerato parte di un singolo introne.
- Le due proteine sono identiche alle estremità, ma una delle due ha una sequenza addizionale nel mezzo.

Genes within genes



- **Neurofibromatosis gene (NF1) intron 26 encode :**
  - **OGMP (oligodendrocyte myelin glycoprotein),**
  - **EVI2A and EVO2B (homologues of ecotropic viral integration sites in mouse)**

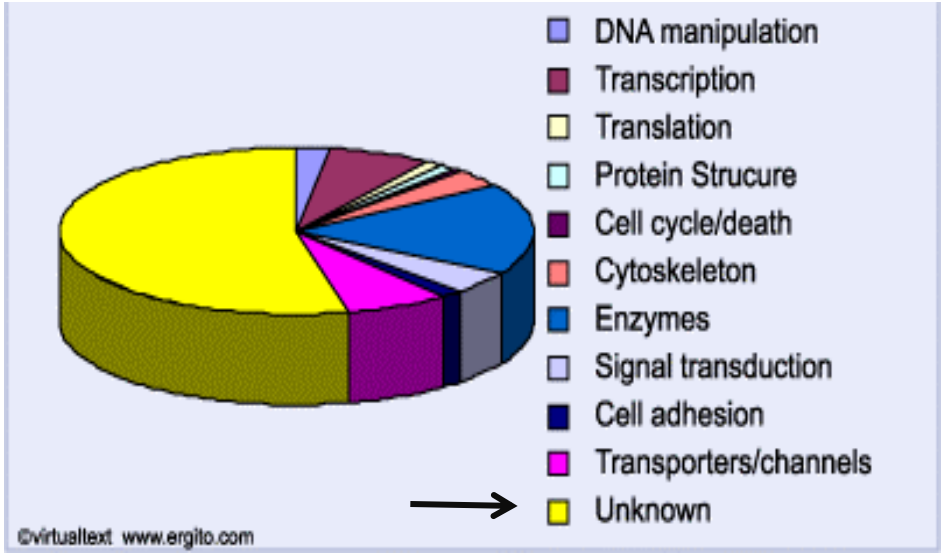
Qual'è la funzione dei vari geni?

- Analizzando il genoma di **Drosophila**, possiamo cercare di capire quanti geni siano dedicati a ciascuna funzione.

Tra i geni la cui funzione è stata identificata ci sono:

- **2.500 enzimi;**
- **≈ 750 fattori di trascrizione;**
- **≈ 700 proteine con funzione di trasporto oppure canali transmembrana per gli ioni;**
- **≈ 700 proteine con funzione di trasduzione del segnale;**

- **NON SI CONOSCE ANCORA LA FUNZIONE DI PIU' DI META' DEI GENI.**



2. Gene families

- A gene family is a **group of genes that share important characteristics.**
- These may be
  - **Structural:**

    similar sequence of DNA building blocks (nucleotides).

    Their products (such as proteins) have a similar structure or function.
  - **Functional:**

    proteins, produced from these genes, **work together as a unit or participate in the same process**

Multigene families

- **Some genes are Transcribed (But don't Make Proteins)**
- The entire family of genes probably **evolved from a single ancestral gene**.
  - Famous examples: **rRNA, globin genes**
  - Four different pieces of **rRNA** are used to make up a ribosome: **18S, 5.8S, 28S, and 5S.**
  - It turns out that three of these **rRNAs (18S, 5.8S, 28S,**) occur in the genome as a gene (**on chrom 13, 14, 15, 21, 22**) & transcribed together. (**one 5S on chrom. 1**)
  - The entire multigene family is **repeated nearly** 300 times in clusters on five different chromosomes!
    - It makes sense to have many repeats of this multigene family because each cell needs many ribosomes for protein synthesis

Multigene family: rRNA Genes

- **RNA polymerase I** synthesizes 45S which matures into 28S, 18S and 5.8S rRNAs
- **RNA polymerase II** synthesizes mRNAs and most snRNA and microRNAs.
- **RNA polymerase III** synthesizes tRNAs, rRNA 5S and other small RNAs found in the nucleus and cytosol.

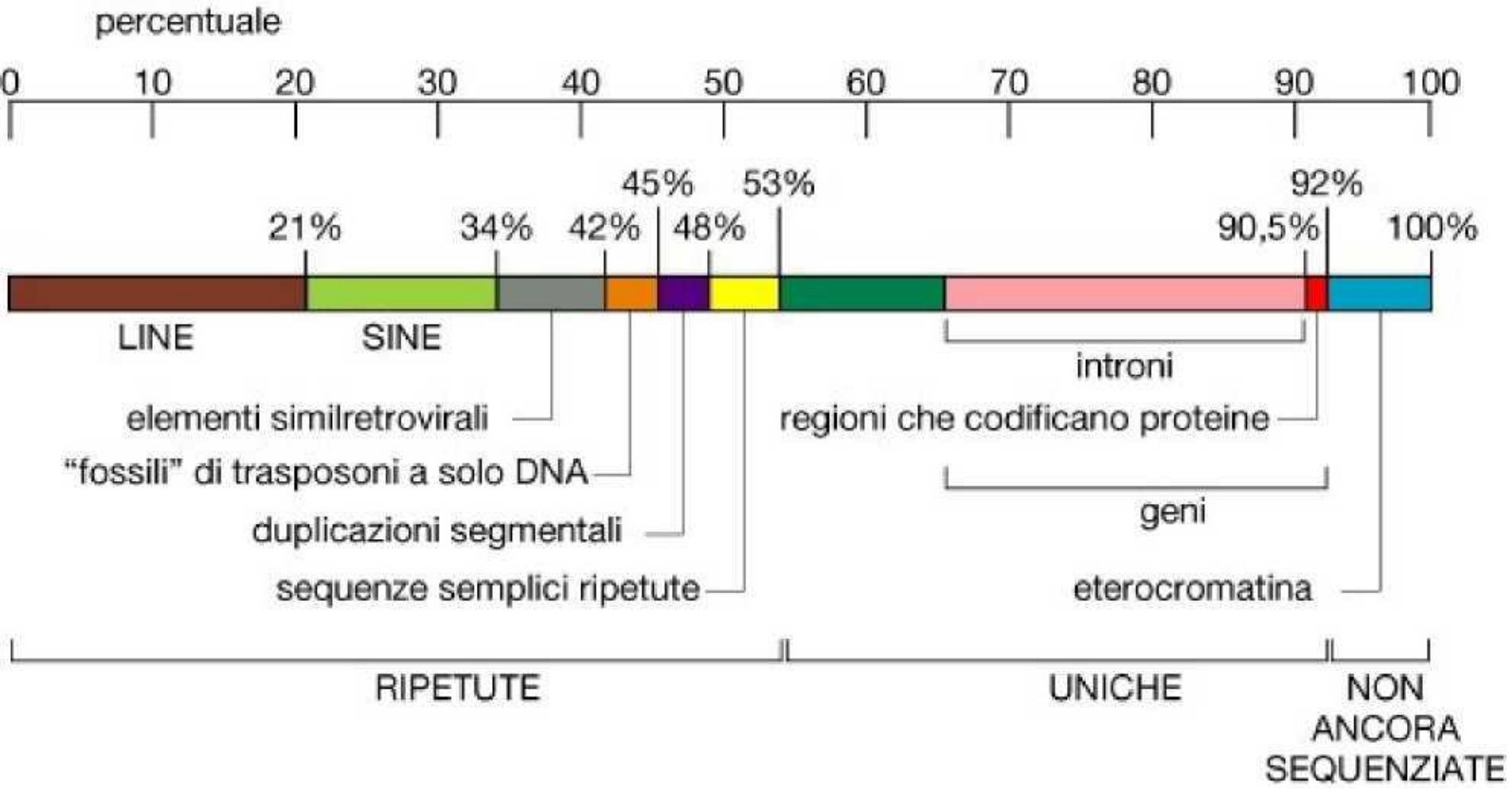
tRNA genes
(497 nuclear genes + 324 putative pseudogenes)

- Humans have fewer tRNA genes that the worm (584), but more than the fly (284);
- Frog (*Xenopus laevis*) has thousands of tRNA genes;
- **Number of tRNA genes correlates with size of the oocytes;**

  In large oocytes lots of protein needs to be sythesized simultaneously.

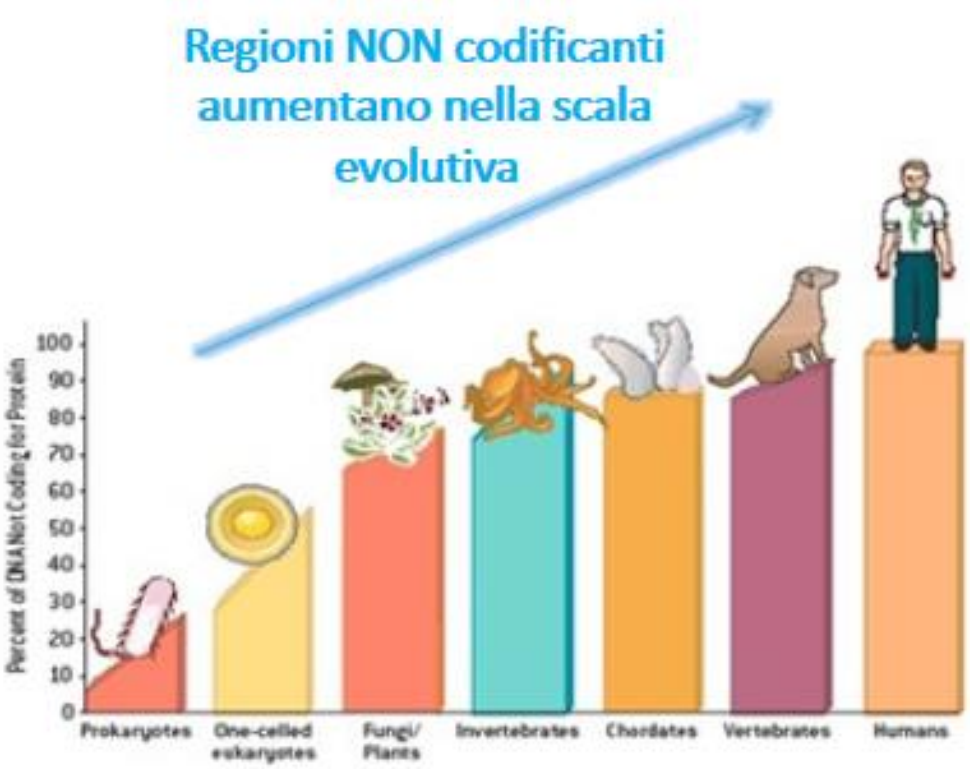
Il genoma è fatto solo di geni?

Anatomia del Genoma Umano'



3. Le regioni NON codificanti hanno un ruolo chiave nella complessità degli organismi





La maggior parte del DNA in genomi grandi non è codificante

- Genomi complessi hanno circa da 10x a 30x più DNA rispetto a quello richiesto per codificare tutti gli RNA o proteine di un organismo.
- Contribuiscono al DNA non-codificante:
  - **Introni in geni**
  - **Elementi regolatori di geni**
  - **Copie multiple di geni, inclusi:**
    - **pseudogeni**
    - **sequenze intergeniche**
    - **ripetizioni intersperse**

Non-coding RNAs

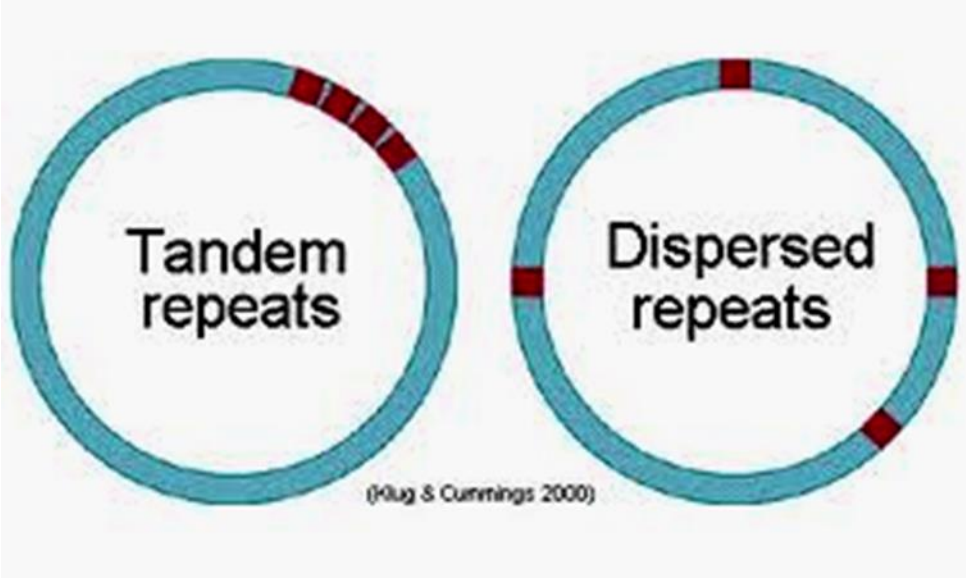
- **tRNA & rRNA** **4.5S & 7S RNA** (Signal Recognition Particles)
- **snRNA** – Pre-mRNA splicing
- **snoRNA** – rRNA modification
- **siRNA** – small interfering RNA
- **gRNA** – guide RNA in RNA editing
- **Telomerase** RNA – primer for telomeric DNA synthesis
- **tmRNA** is a hybrid molecule, half tRNA, half mRNA
- **Xist**: The X chromosome silencing is mediated by Xist – a 16,000 nt long ncRNA
- **shRNA** (small heterochromatic RNAs): expresses only one allele while other is silenced
- **LNA** Locked Nucleic Acid
- **piRNA** Piwi-interacting RNA

Il genoma è fatto solo di geni?

3. Il DNA NON CODIFICANTE

3. DNA RIPETITIVO

- Two types
  - Tandemly repetitive
  - Interspersed repetitive



3.1. Tandem repeats

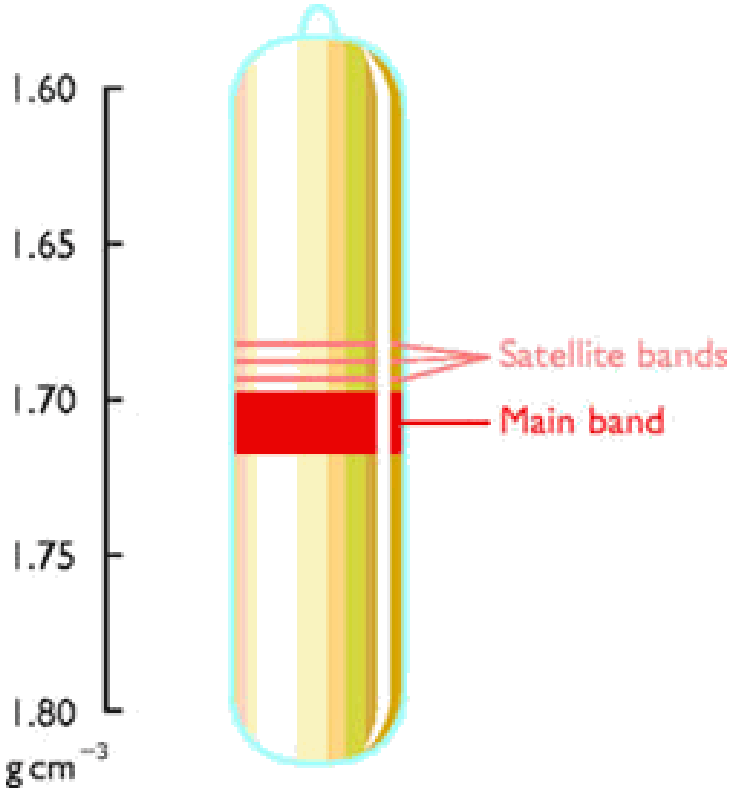
- **Tandem repeats** occur in DNA when a pattern of two or more nucleotides is repeated and the repetitions are **adjacent** to each other
- Form different density band on density gradient centrifugation (from bulk DNA) -**satellite**

Example:

A-T-T-C-G-A-T-T-C-G-A-T-T-C-G

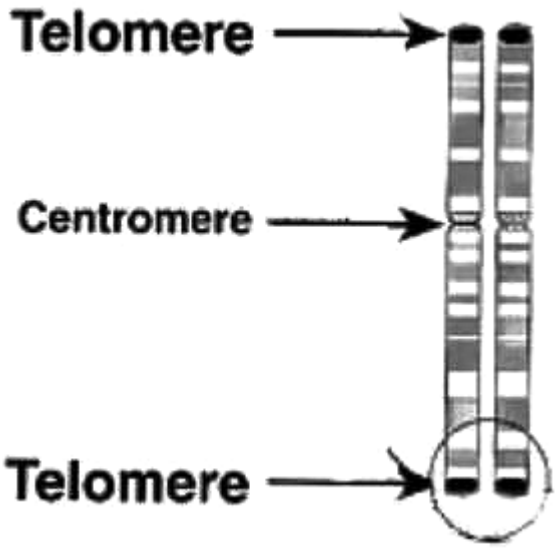
Tandem repeats:

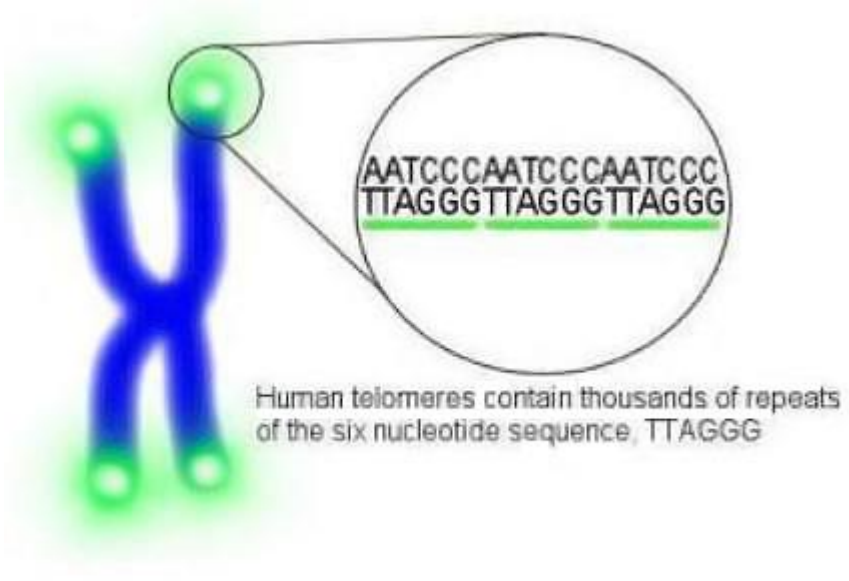
- **Satellite DNA:**
- **Microsatellite:**
- **Minisatellite:**



a. Satellite DNA

- Unit - 5-300 bp depending on species.
- Repeat - 10^5- 10^6 times.
- Location - Generally heterochromatic.
- Examples - Centromeric DNA, telomeric DNA. There are at least 10 distinct human types of satellite DNA.





b. Microsatelliti e c. Minisatelliti

- **B**) I **microsatelliti** sono costituiti da unità di ripetizione lunghe da **1 a 10 pb**, ripetute in tandem 10-20 volte, che formano raggruppamenti molto corti, <150pb, di tipo (A)n, (CA)n, (CGG)n, ecc.
- Sono anche detti **SSR (simple sequence repeats)**. Le ripetizioni possono essere perfette o presentare piccole variazioni.

- **C)** I **minisatelliti** sono costituiti da unità più lunghe (da **11 a 100pb**) ripetute in tandem 20-50 volte che formano raggruppamenti di lunghezza fino a 20kb

- Gli SSR costituiscono circa il 3% del genoma umano. Sono molto importanti nello studio delle malattie genetiche in quanto mostrano un elevato grado di polimorfismo nella popolazione umana.

Da: Lander et al. Nature 2001, 409: 860

Table 14 SSR content of the human genome

Length of repeat unit	Average bases per Mb	Ave rage n u mber of SSR elements per Mb
1	1,660	36.7
2	5,046	43.1
3	1J013	11 8
4	3.383	32.5
5	2,686	17.6
6	1,376	152
7	906	8.4
8	1,139	11.1
9	900	8.6
10	1,576	8.6
11	770	8.7

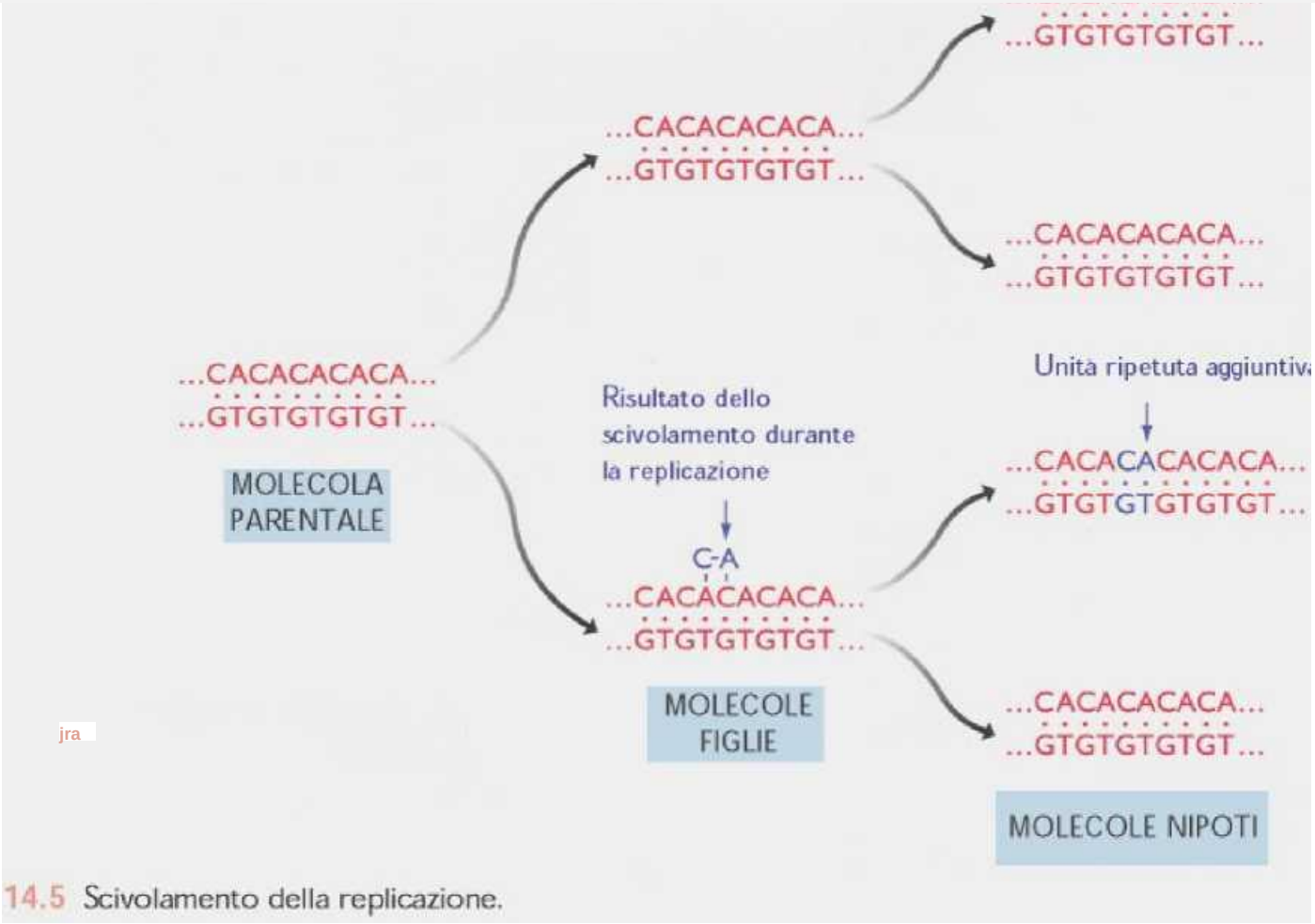
SSRs were identified by using the computer program Tandem Repeat Finder with the following parameters: match score 2, mismatch score 3, indel 5, minimum alignment 50, maximum repeat length 500, minimum repeat fength 1.

Table 15 SSRs by repeat unit

Repeat unit	Number of SSRs per M
AC	27
AT	19
AG	8.2
GC	0.1
AAT	4.1
AAC	2.6
AGG	1.5
AAG	1.4
ATG	0.7
CGG	0.6
ACC	0.4
AGC	0.3
ACT	0.2
ACG	0.0

SSRs were identified as in Table 14.

Gli SSR possono formarsi attraverso un meccanismo di scivolamento della replicazione



14.5 Scivolamento della replicazione.

- Gli SSR sono presenti con una frequenza di almeno uno ogni circa 2 kb del genoma
- Si originano da vari meccanismi tra cui il più importante è lo scivolamento della DNA polimerasi durante la replicazione.

Microsatelliti: Genetic Fingerprint

❑ Caratteristiche degli SSRs

- Polimorfismo di lunghezza: DNA fingerprinting
- Spesso adoperati come marcatori genetici per la mappatura di geni associati a patologie.

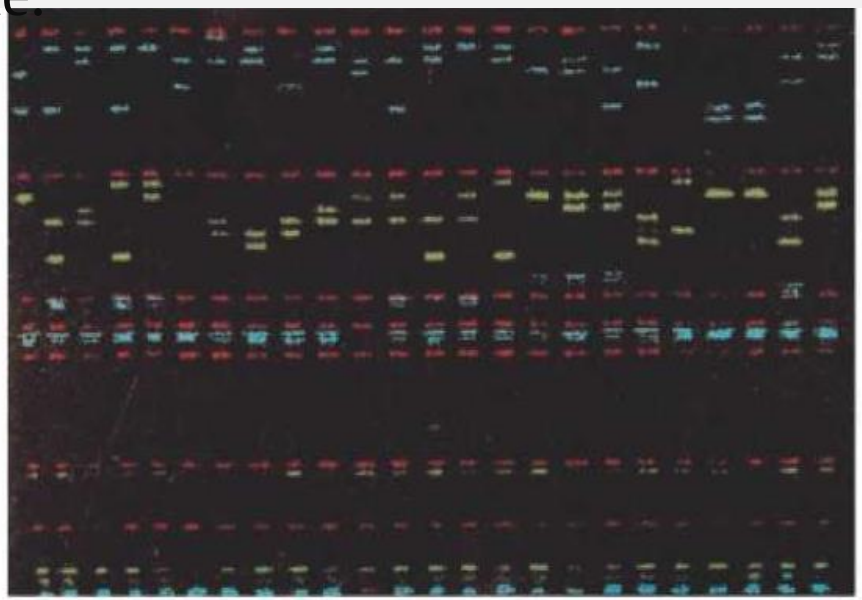


Figura 2.25 L'uso dell'analisi dei microsatelliti per la determinazione del profilo genetico.

T.A. Brown
Genomi
EdiSES

Tandemly Repetitive DNA Can Cause Diseases:

- **Fragile X Syndrome**
  - "CGG" is repeated hundreds or even thousands of times creating a "fragile" site on the X chromosome.
  - It leads to mental retardation.
- **Huntington's Disease**
  - "CAG" repeat causes a protein to have long stretches of the amino acid glutamine.
  - Leads to a neurological disorder that results in death

Microsatelliti e malattie genetiche

- I microsatelliti, ed in particolare le **ripetizioni di triplette** sono associati a varie malattie genetiche

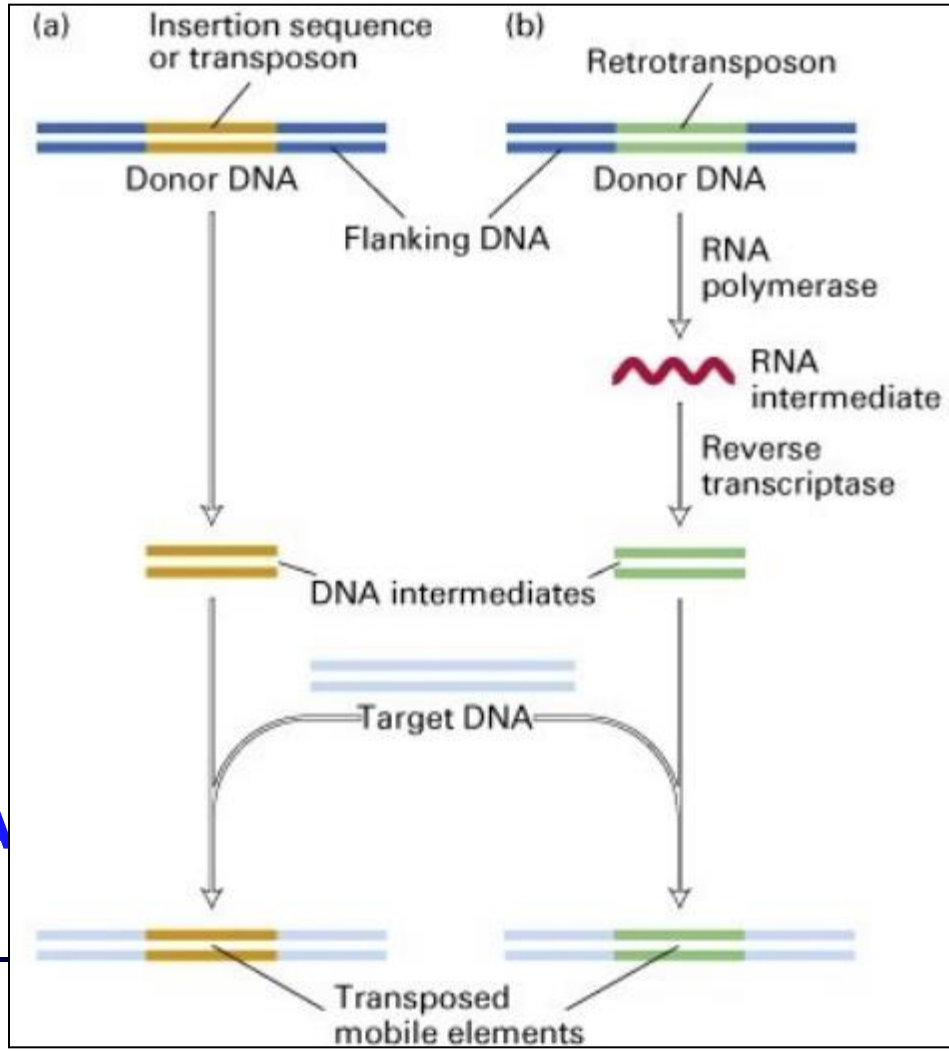
TABLE 1 Summary of trinoceotide repeat disorders

					Repeat Size		
Disease	Gene	Locus	Protein	Repeat	Normal	Disease	Repeat location
Fragile X syndrome	WRf(FRAXA)	Xg27.3	FMRd protein (FMRP)	CGG	6-53	60-200 (pre) >230 (full)	5'-UTR
Fragile XE syndrome	*FMR2* (FRAXE)	Xg28	FMR-2 protein	GCC	6-35	61-200 (pre) >200 (fuH)	5‘-UTR
Friedreich ataxia	*X25*	9q13-21 1	Frataxin	GAA	7-34	34—80 (pre) >160 (full)	Intron 1
Myotonic dystrophy	*DMPK*	19g13	Myotonic dystrophy protein kinase (DMPK)	CTG	5-37	50-thou sands	3‘ UTR
Spinobulbar muscular atrophy (Kennedy disease)	*AR*	Xg 13-21	Androgen receptor (AR)	CAG	9-36	36-62	Coding (amino terminal)
Huntington disease	*HD*	4p16.3	Huntington	CAG	6-35	36-121	Coding (amino terminal)
Dentatorubral-pallidoluysian atrophy (Haw River syndrome)	*DRPLA*	12p13.31	Atrophin-1	CAG	3 35	49-66	Coding [amino terminal)
Spinocerebellar ataxia type 1	*SCA1*	6p23	Ataxin-1	CAG	6-44	39-82	Coding (amino terminal)
Spinocerebellar ataxia type 2	*SCA2*	12g24.1	Ataxin-2	CAG	15-31	36-63	Coding (amino terminal)
Spinocerebellar ataxia type 3 [Machado-Joseph disease)	*SC A3 [MJD1]*	14g 32.1	Ataxin-3	CAG	**12-46**	55-84	Coding (carboxy terminal)
Spinocerebellar ataxia type 6	*SCA6*	19p13	Cf_{1A}-Voltage-dependent calcium channel subunit	CAG	i 18	21-33	Coding (carboxy terminal)
Spinocerebellar ataxia type 7	*SCA7*	3p12-13	Ataxin-7	CAG	4-35	37-306	Coding (an-ino terminal)
Spinocerebellar ataxia type 8	*SCAB*	13g21	None	CTG	16-37	110 to <250?	3-Terminal exon (antisense?)
Spinocerebellar ataxia type 12	*SCA12*	5q31-33	PP2A-PR55p(PPP2R2B)	CAG	7-28	66-78	ff-UTR

3.2. Ripetizioni intersperse

- Costituite da sequenze di DNA ripetute, disperse in tutto il genoma.
- Sono definite anche **Elementi mobili del DNA**, perché derivano da elementi trasponibili(sequenze di DNA che si muovono o sono duplicate da una posiszione ad un'altra del genoma

- **Classe I o Retrotrasposoni** si originano per eventi di retrotrasposizione, attraverso un intermedio ad RNA
  - **elementi LTR**
  - **LINEs**: long interspersed nuclear elements
  - **SINEs**: short interspersed nuclear elements
- **Classe II o Trasposoni a DNA** si originano attraverso un intermedio a DNA, secondo meccanismo di trasposizione conservativa o replicativa



a.Retrotrasposoni

- ✓ La caratteristica di tutti i retrotrasposoni è la **presenza di brevi ripetizioni dirette** alle estremità 3’ e 5’ , copia della sequenza del sito d’integrazione.

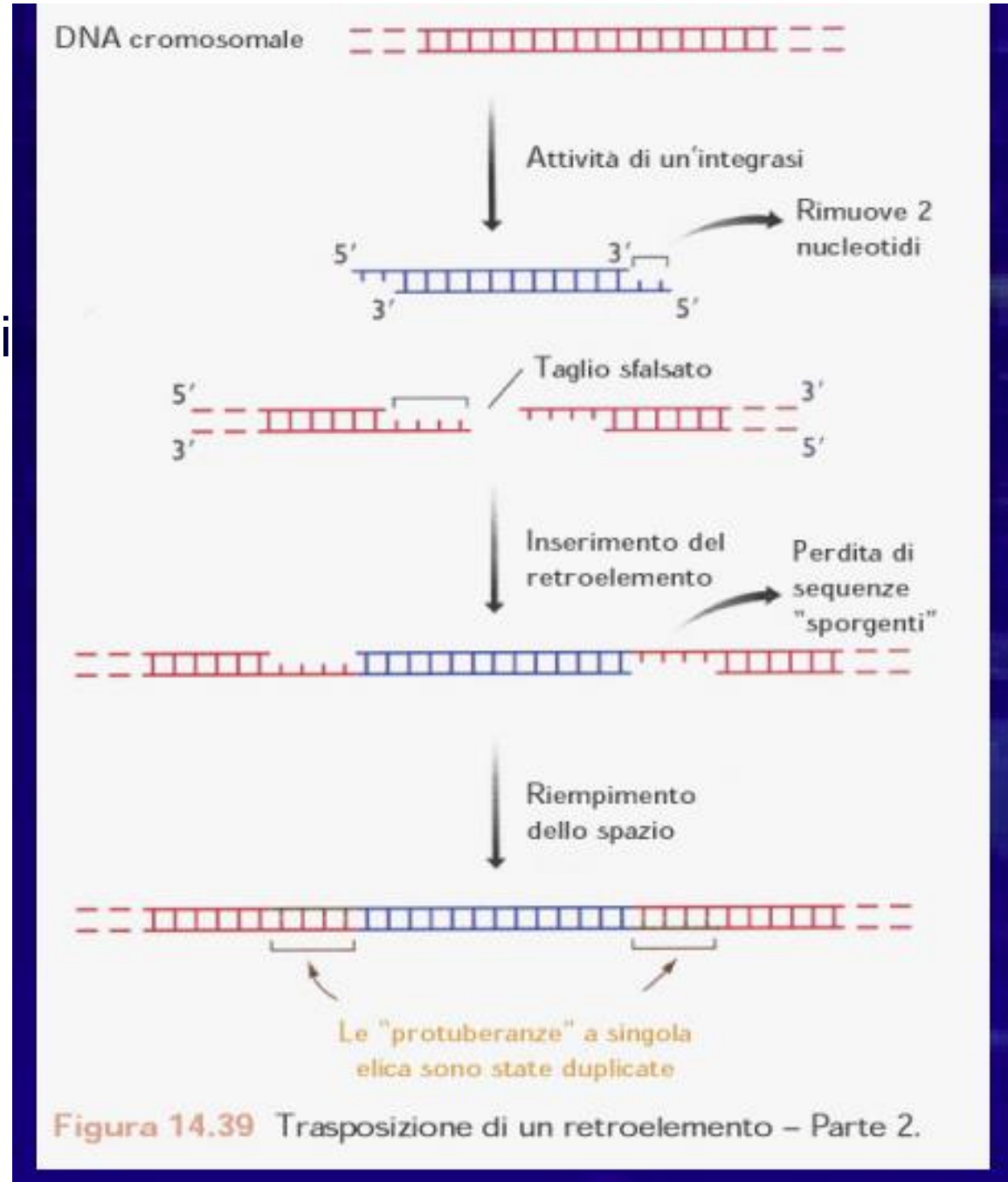


Figura 14.39 Trasposizione di un retroelemento – Parte 2.

Ripetizioni Intersperse nel Genoma Umano

- Gli elementi ripetuti interspresi costituiscono circa il 45% del genoma umano.

- **LINE (Long interspersed nuclear elements)**

- L1, L2, L3 LINE (~21% del genoma, ~100,000 copie)

- **SINE (Short interspersed nuclear elements**)

- *Alu* (~10,7% del genoma, ~1,200, 000 copie)
- MIR, MIR3 (~3% del genoma, ~500,000 copie)

- **Elementi LTR (Long Terminal Repeats)**

- ERV, MalR (8% del genoma, ~500,000 copie)

- **Transposoni a DNA**

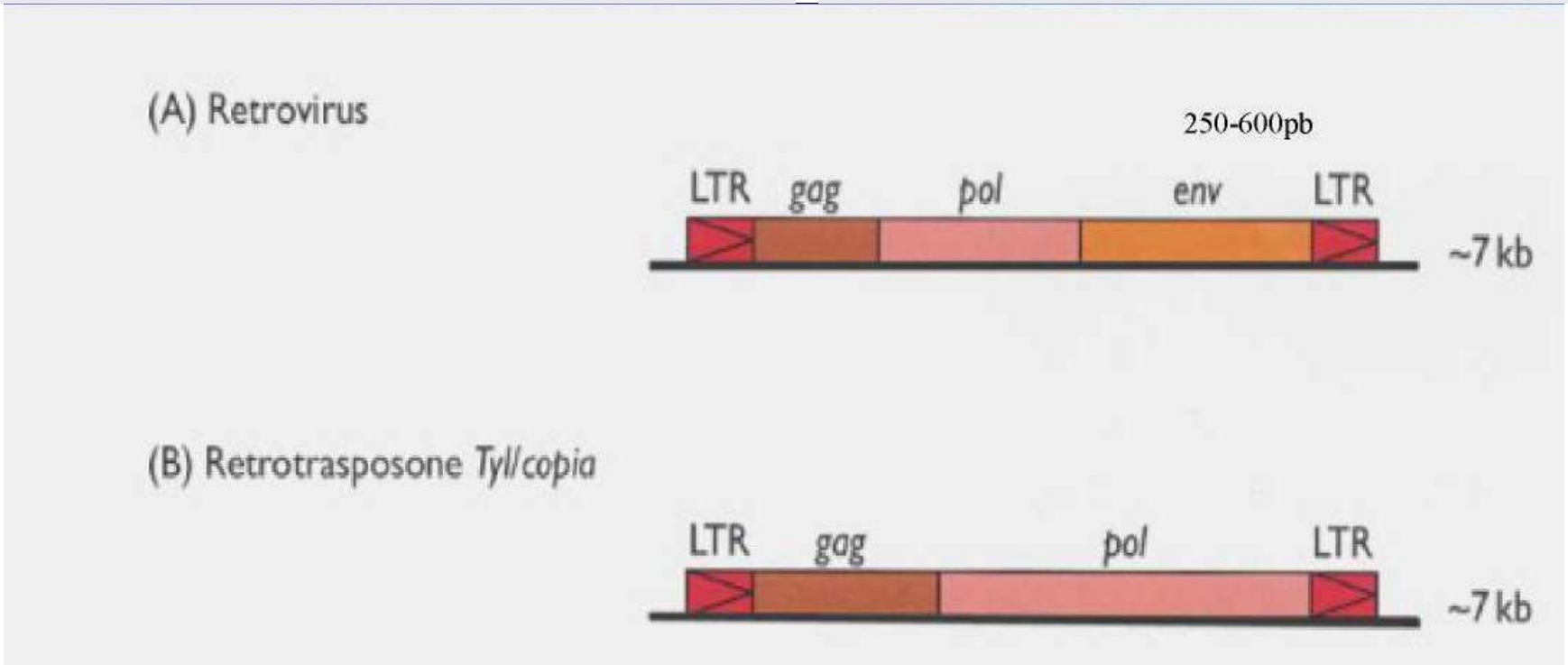
- MERI (Charlie), MER2 (Tigger), others (2,8% del genoma, ~350, 000 copie)

Elementi LTR

- Gli **elementi LTR** o **retrotrasposoni virali** (6-7kb) presentano analogie con i retrovirus.
- Caratteristici degli invertebrati (piante, funghi, insetti) dove sono presenti in gran numero di copie

- **E**lementi *Ty* in *S. cerevisiae*
- Elementri *Copia* in *Drosophila*

- mancano del gene **env**
- e non possono formare particelle virali



B, Trasposoni a DNA

- I Trasposoni a DNA sono elementi mobili distinti in due categorie
  - **Trasposoni a DNA che si spostano replicandosi:** una copia rimane nel sito originale, mentre la nuova copia si inserisce altrove nel genoma
  - **Trasposoni a DNA che si spostano in maniera conservativa** da un sito all'altro del genoma senza aumentare il numero di copie

- Sono caratterizzati da una sequenza codificante la trasposasi contenente introni, fiancheggiata da ripetizioni terminali invertite, simili a quelle dei trasposoni batterici.

- Sono meno comuni negli eucarioti (3% nel genoma umano, raggruppati in 7 classi principali) rispetto ai retrotrasposoni.
- I più noti sono gli Elementi Ac e Ds del mais, i primi elementi mobili identificati negli i anni 50 da B. McClintock e gli elementi P di Drosophila.
- Traspongono mediante il meccanismo di trasposizione conservativa

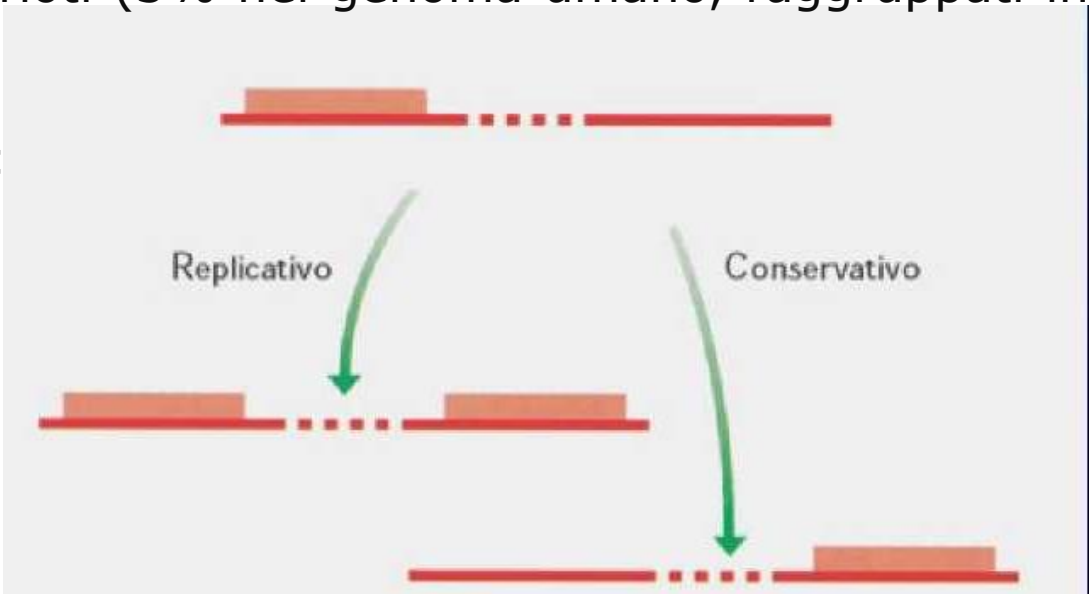


Figura 2.28 Due meccanismi di trasposizione utilizzati dai trasposoni a DNA

Funzione degli elementi ripetuti

- **Punti caldi** per ricombinazione (duplicazioni, inversioni, traslocazioni; creazione di nuovi geni per shuffling esonici)

- **Alterazione della espressione genica** in quanto portatori di segnali trascrizionali (es. promotori e enhancer di LTR; promotori di Alu; siti di terminazione deboli della trascrizione di elementi L1; segnali di poliadenilazione)

- **Presenza in geni per proteine** (Le Alu contengono siti criptici di splicing; fonte di domini proteici; contributo a variabilità delle proteine)

- **Reclutamento come elementi regolatori** (es. BC200 di primati deriva da Alu monomerica)

- **Fonte di pseudogeni processati** (ritorno in vita come lunghi esoni? Come nuovi geni?)

- **Fonte di plasticità del genoma** e quindi ruolo attivo nel rimodellamento genomico (riarrangiamenti cromosomici, reshuffling di geni, etc)

Identificazione Bioinformatica di elementi ripetuti



RepeatMasker Web Server

RepeatMasker screens DNA sequences in FASTA format against a library of repetitive elements and returns a masked query sequence ready for database searches. RepeatMasker also generates a table annotating the masked regions.

Reference: A.F.A. Smit, R. Hubley & P. Green, unpublished data. Current Version: open-3.1.6

Check Current Queue Status

www.repeatmasker.org

Basic Options

Sequence: Sfoglia... or

Select a sequence file to process or paste the sequences(s) in FASTA format. Large sequences will be queued, and may take a while to process.

Search Engine: cross_match wublast

Select the search engine to use when searching the sequence. Cross_match is slower but often more sensitive than WUBlast.

Speed/Sensitivity: rush quick default slow

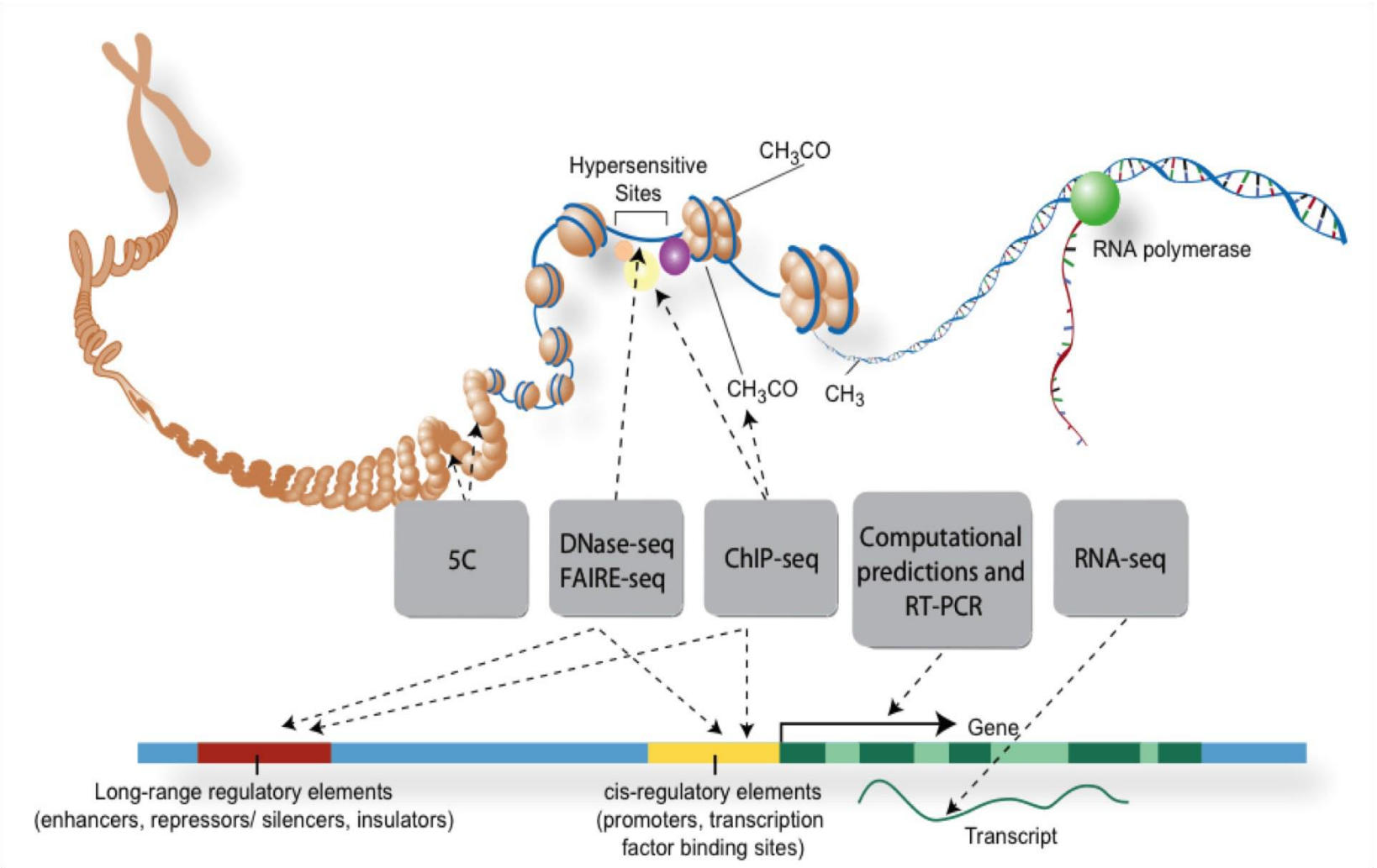
Select the sensitivity of your search. The more sensitive the longer the processing time.

DNA source: Human

Select a species from the drop down box or select "Other.." and enter a species name in the text box. Try the protein based repeatmasker if the repeat database for your species is small.

RepBase colleziona gli elementi ripetuti noti in una serie di organismi

The ENCODE project (Encyclopedia of DNA Elements)



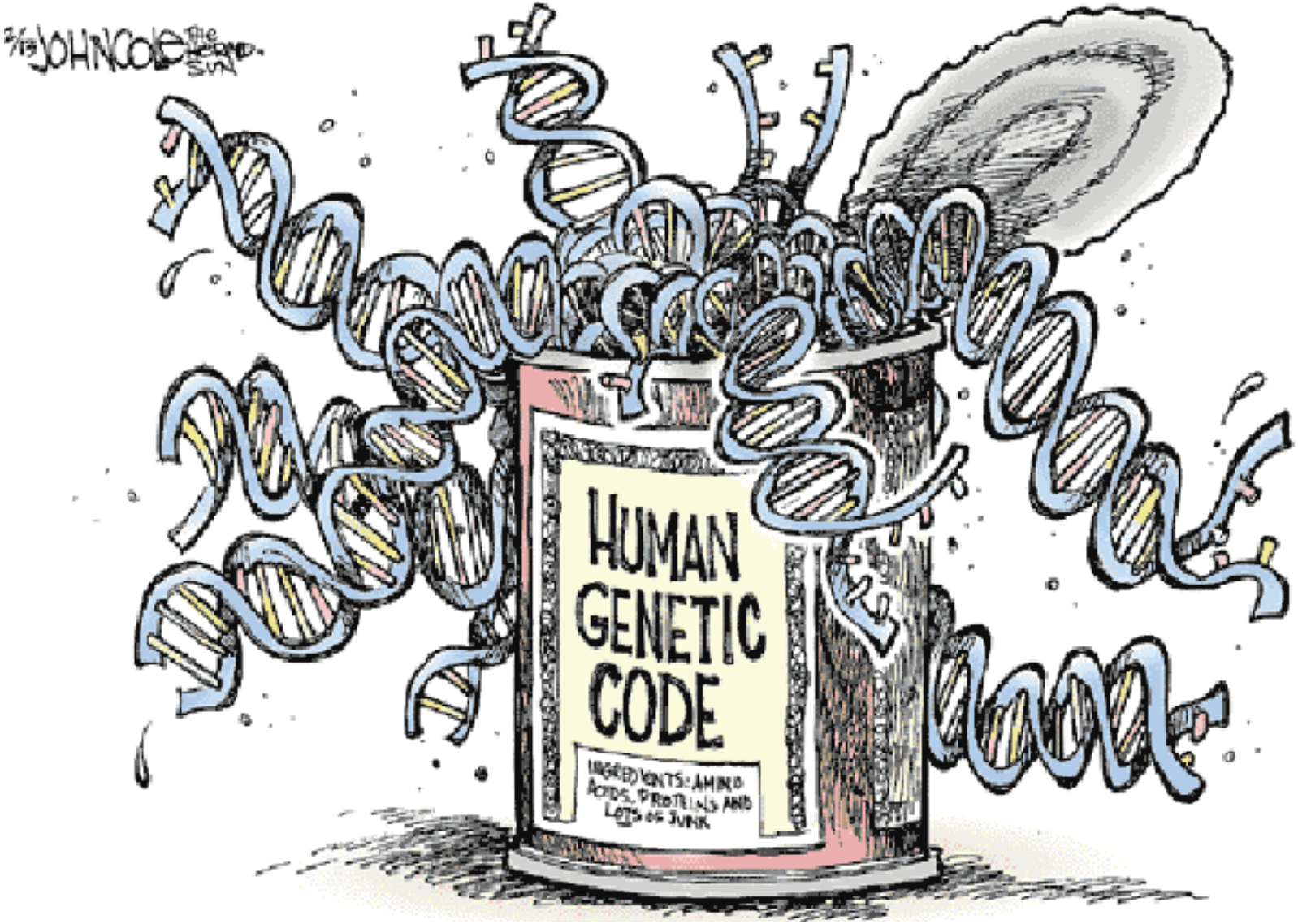
The ENCODE project

(Encyclopedia of DNA Elements)

- Goal of ENCODE: build a list of all sequence- based functional elements in human DNA.

This includes:

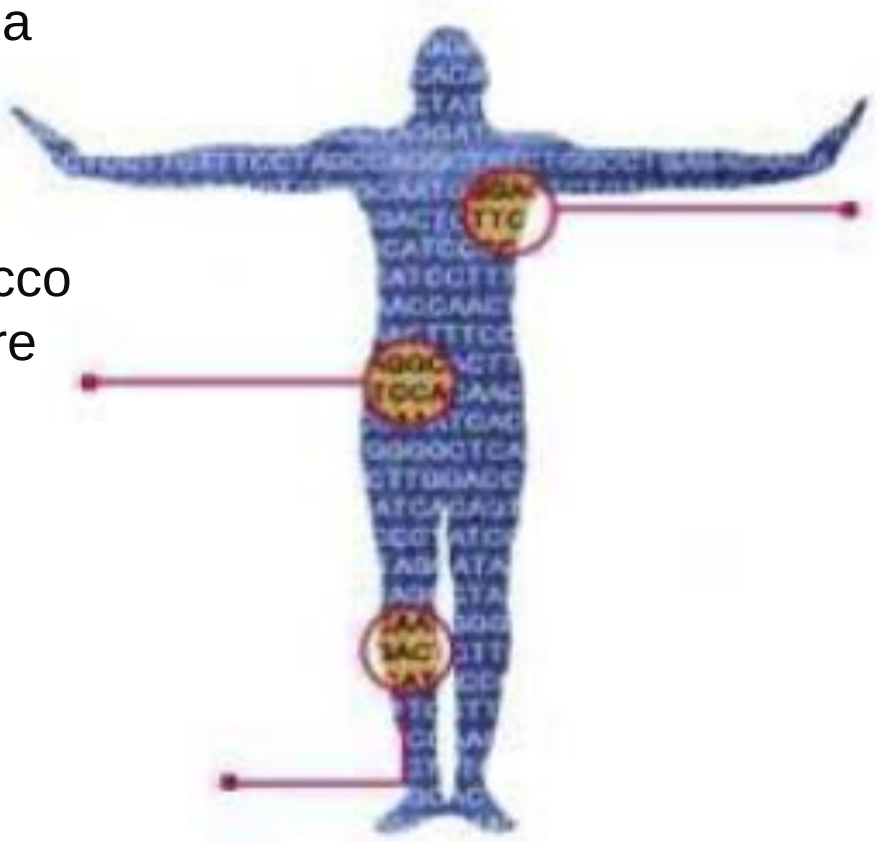
- protein-coding genes
- non-protein-coding genes
- regulatory elements involved in the control of gene transcription
- DNA sequences that mediate chromosomal structure and dynamics


JOHN COLE
HUMAN GENETIC CODE

Il Progetto Genoma Umano

- La conoscenza di tutta l'informazione racchiusa nel genoma è la condizione necessaria per comprendere la completa biologia di un determinato organismo, cioè significa comprendere il **"segreto della vita".**

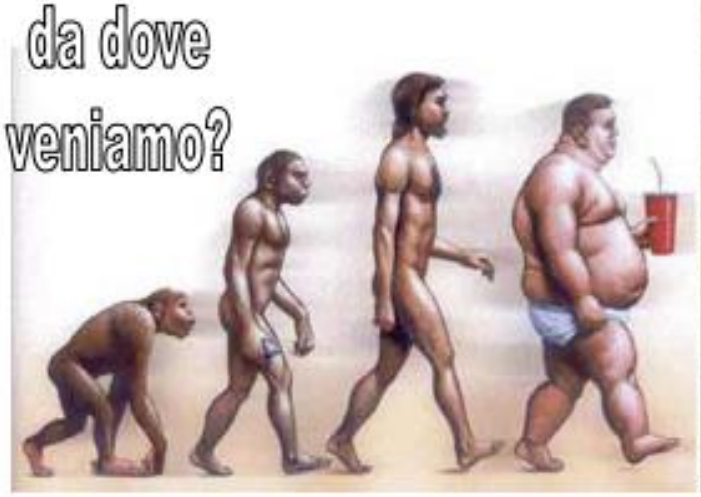
- Conoscere l'intera sequenza del genoma umano è come conoscere tutte le pagine del manuale necessario per costruire il corpo umano.
- Nel 1986 il premio Nobel Renato Dulbecco e Leroy Hood lanciano l'idea di sequenziare l'intero genoma umano



Perché studiamo il genoma umano











Il Progetto Genoma Umano

Perché?

- La sequenza del genoma umano mira a:

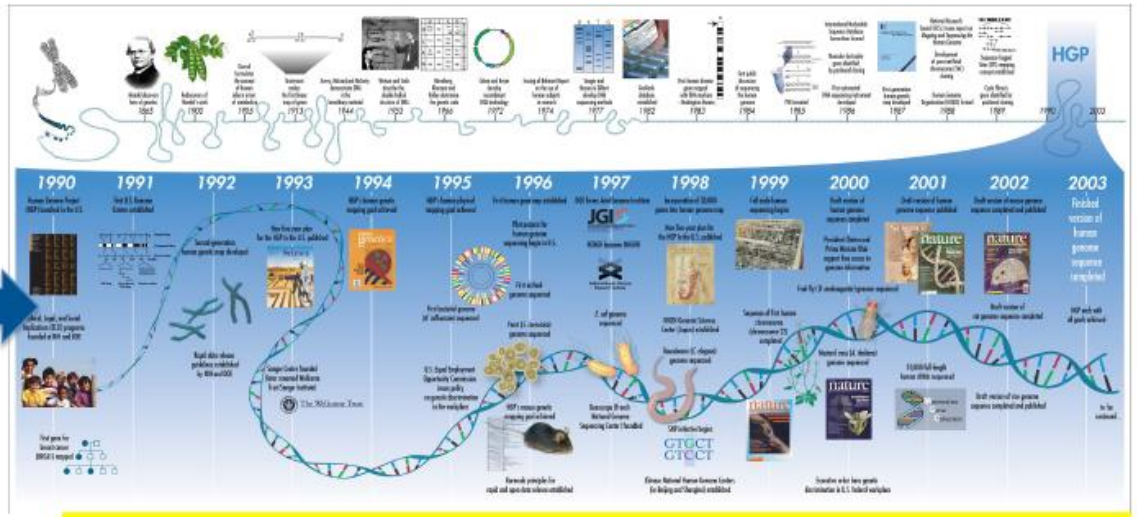


- identificare **tutti i geni umani**
- determinare la **struttura dei geni**
- definire le funzioni dei geni
- identificare i **geni responsabili delle malattie** mendeliane
- determinare le **regioni non codificanti con funzione regolatoria**
- scoprire **l'inatteso**

Il sequenziamento del genoma umano oggi



1953 – DNA structure



Human Genome Project – 13 years - > 3 billion dollars

Next generation sequencing (NGS) technologies

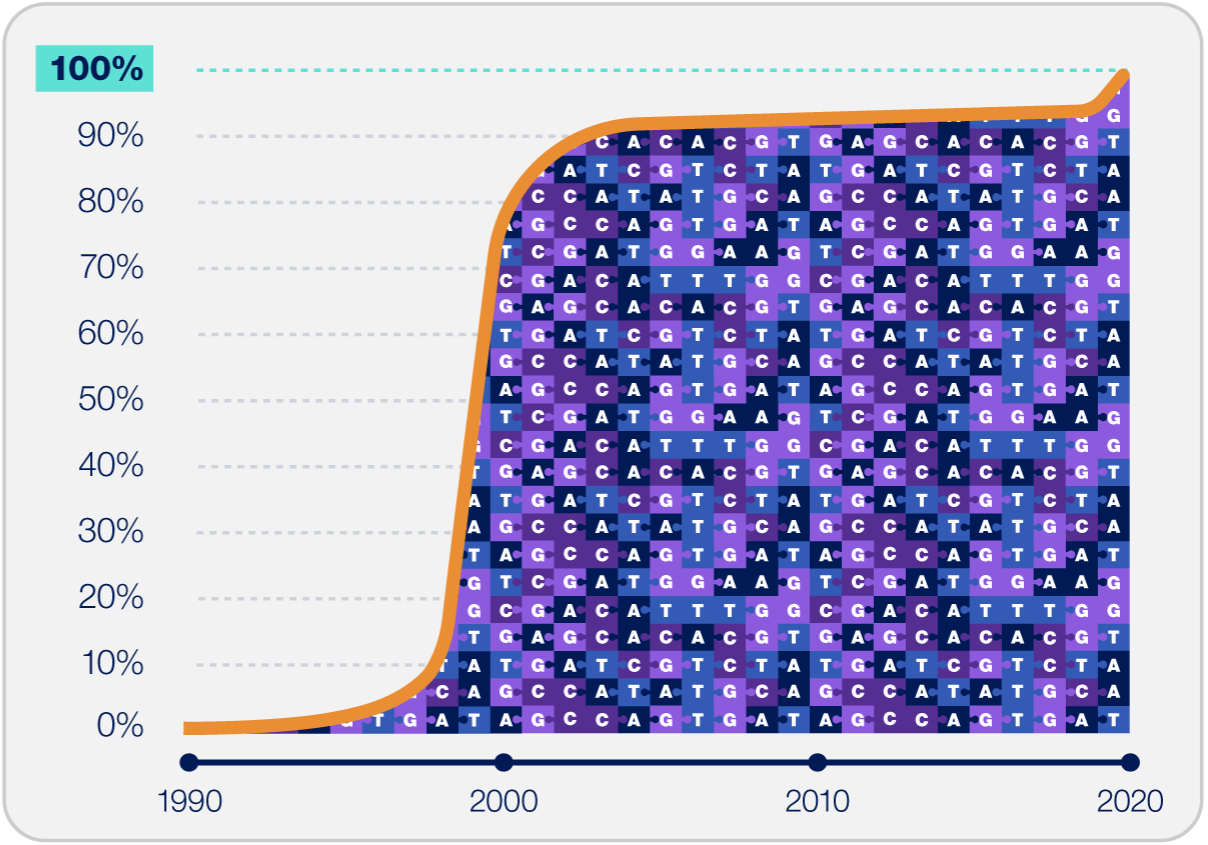


Human genome - 2016 – few days - 1000 dollars !!

Researchers generate the first complete, gapless sequence of a human genome.

Science 376, 1 April 2022

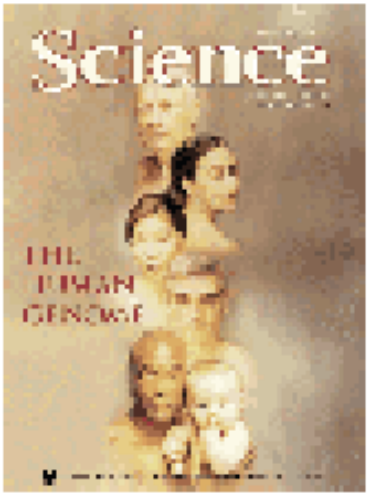




Sequenza genoma umano

Venter et al., The sequence of the human genome. 2001 Science

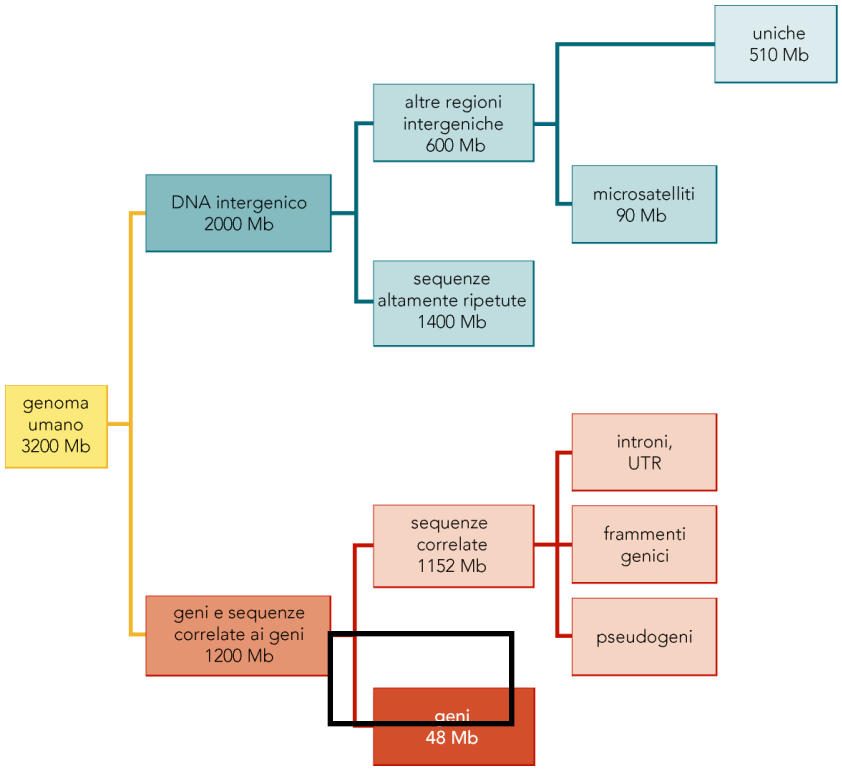
Size of the genome (including gaps)	**2.91 Gbp**
Size of the genome (excluding gaps)	2.66 Gbp
Percent of A+T in the genome	**54**
Percent of G+C in the genome	**38**
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	**26,383**
Percent of annotated genes with unknown function	**42**
Number of genes (hypothetical and annotated)	**39,114**
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	**Chr. 19 (23 genes/Mb)**
Least gene-rich chromosomes	**Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)**
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)



➢ Risultati

❑ Genoma umano: 3.2 Gb

- La mappa 2004 contiene 2.85 Gb corrispondente al 90% del genoma.
- L' errore è del 0.001% e la sequenza ottenuta presenta 341 interruzioni (DNA altamente ripetuto o Z-DNA). Pubblicazione su *Nature.*
- Le interruzioni sono state ridotte a 200 nel 2009.

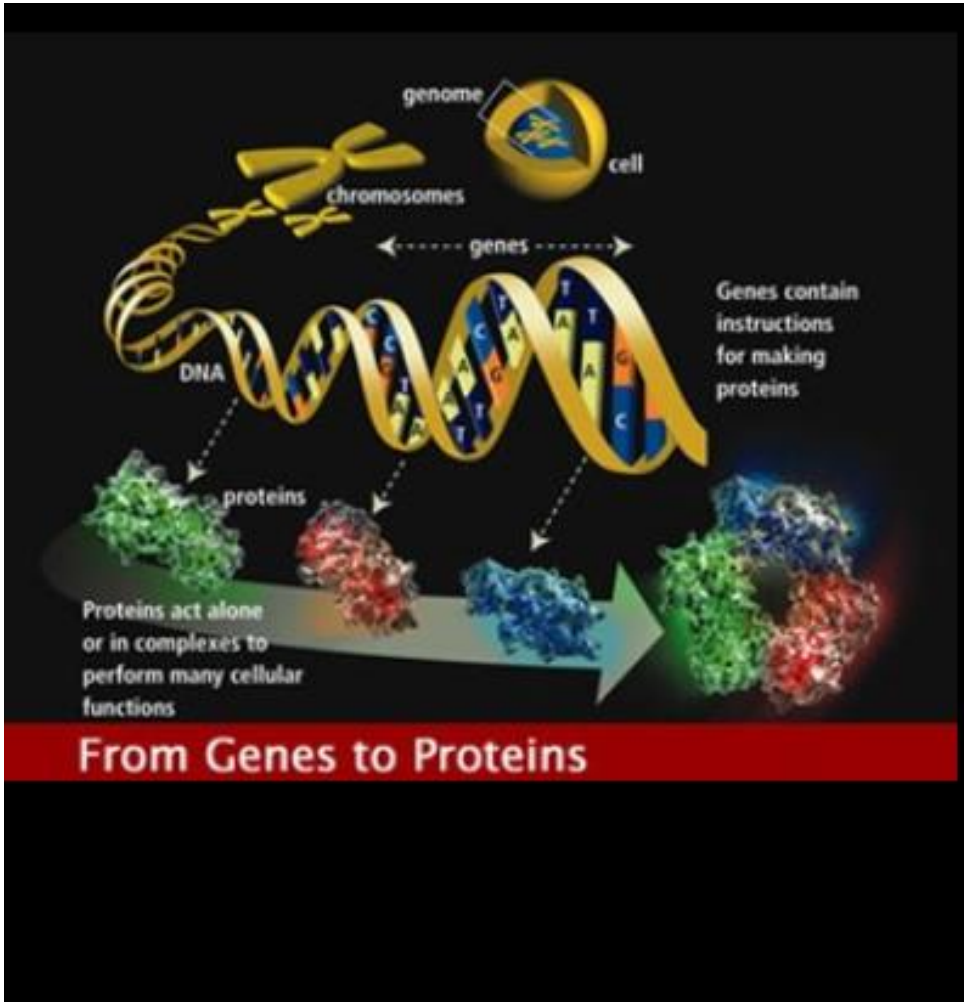


- Meno del **~2 % del genoma** sono **esoni**.
- Sono presenti circa **21500 geni codificanti proteine** e **4800 geni codificanti RNA** (rRNA, tRNA, snRNA e microRNA Il numero di geni stimato era 32-35000.
- **23%** è costituito da **introni.**
- **75%** è DNA intergenico.
- Identificati circa **3 milioni di Single Nucleotide Polymorphisms (SNPs).** Pubblicazione su *Nature* nel 2005 e 2007.
- Circa **1%** di questi **SNPs** può avere un impatto su funzioni proteiche.

- Nel 2007-08 inizia l'era del sequenziamento genomico personalizzato.
- I primi 2 DNA sequenziati sono stati quelli di J. Watson e C. Venter.
- Sequenziamento di 1000 individui appartenenti a diversi gruppi etnici.

Il genoma umano

- Contiene circa 3 miliardi di paia di basi organizzato in:
  - **GENI:** di circa 25,000 geni (2% dell'intero genoma), sono sequenze codificanti per le proteine
  - SEQUENZE CODIFICANTI che generano RNA funzionali
  - ma che non codificano per proteine (es. RNA ribosomiale)
  - SEQUENZE NON CODIFICANTI
    - Funzione strutturale
    - Funzione regolatrice della trascrizione

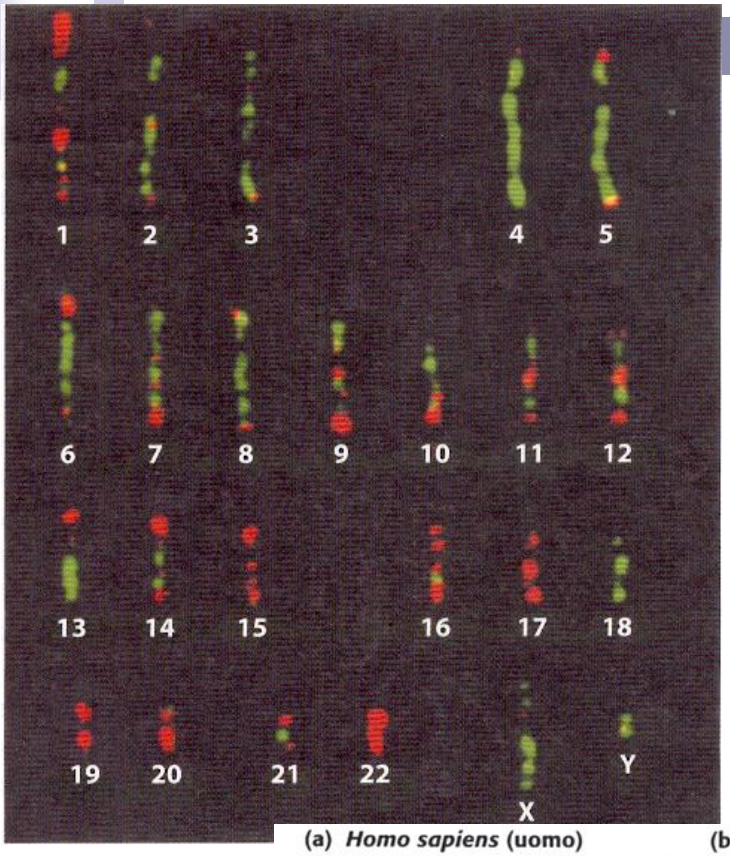


Caratteristiche dei geni umani

- Elevata variabilità nel numero e lunghezza degli introni.
- Geni codificanti proteine possono arrivare fino a >2.4Mbp di lunghezza (distrofina)
- Lunghezza media dei geni: ~ 8,000 bp
- Numero medio di esoni per gene: 9.8
- Numero massimo di esoni per gene: 363
- Lunghezza media degli esoni: ~300 bp
- Lunghezza media degli introni: ~2,000 bp
- Geni monoesonici: ~8%

Dimensione dei geni umani

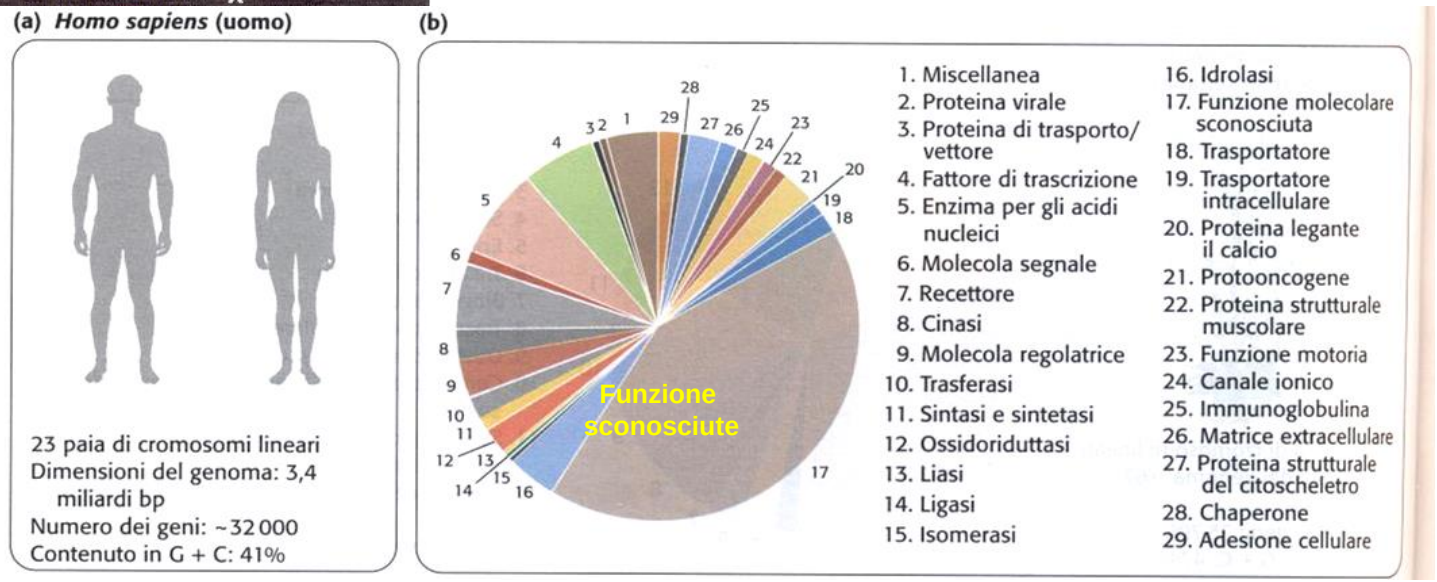
Human protein	Size of protein (no. of amino acids)	Size of gene (kb)	No. of exons	Coding DNA (%)	Average size of exon (bp)	Average size of intron (bp)
SRY	204	0.9	1	94	850	–
b-Globin	146	1.6	3	38	150	490
Type VII collagen	2928	31	118	29	77	190
p53	393	39	10	6	236	3076
Huntingtin	3144	189	67	8	201	2361
CFTR	1480	250	27	2.4	227	9100
Dystrophin	3685	2400	79	0.6	180	30,770



Mappa genica umana

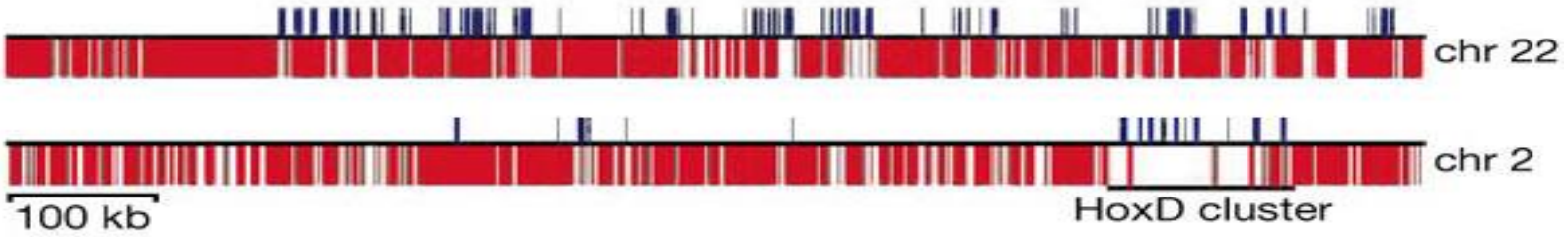
- Le regioni in rosso indicano porzioni dei cromosomi ad <u>alta densità genica</u> (ad esempio i cromosomi 15, 16, 17, 19, 20 e 22).

- Altri cromosomi come 4, 18, X e Y mostrano una <u>colorazione rossa molto debole e sono poveri di geni.</u>



Dati cromosoma 22

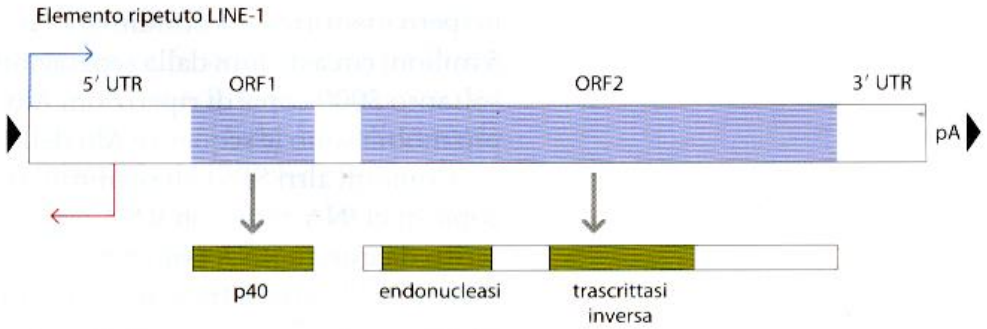
- Un cromosoma ha una lunghezza media di ~150 Mb. La sequenza del cromosoma 22 (braccio lungo) è nota per il 97% (margine errore 1:50,000).
- Interruzioni tra i diversi contigs si hanno nelle regioni centromeriche e telomeri-che (eterocromatina).
- Pur essendo uno dei cromosomi a più alta densità genica, le regioni codificanti occupano solo una piccola frazione del cromosoma 22.
- Sono stati **identificati 679 geni: 247 geni noti, 150 geni correlati** le cui sequenze sono omologhe a geni noti umani o di altre specie, 148 predetti, 134 pseudo-geni che a causa di mutazioni non sono espressi).
- Presenza di numerosi geni ortologhi (uomo-topo).



Barre rosse → sequenze ripetute

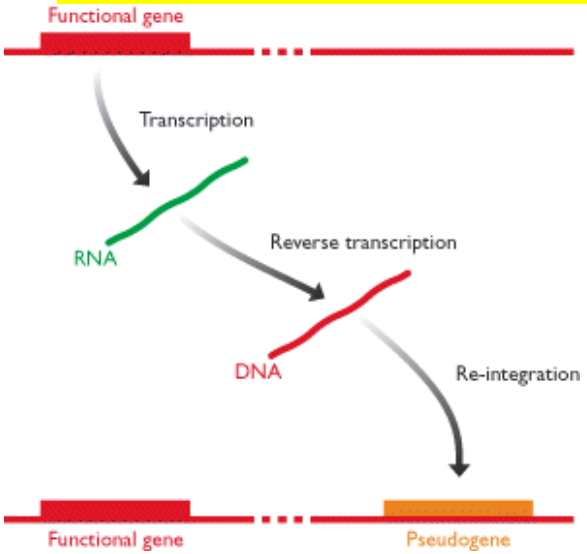
Barre blu → sequenze corrispondenti ad esoni

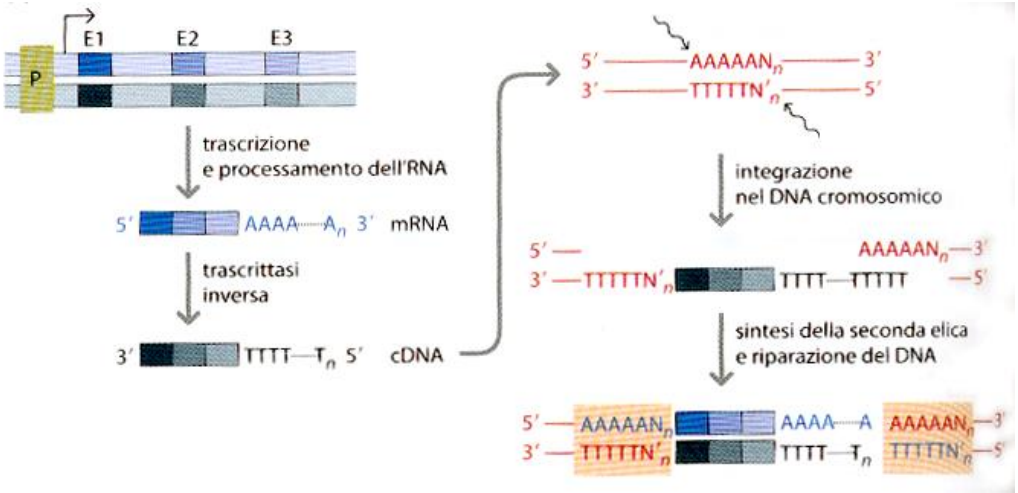
- Circa 850,000 Long Interspersed Elements **LINES** (6-8 kb elementi completi, 1-2 kb elementi tronchi) pari al 21% dell'intero genoma.
- Circa 1,5 milioni di copie di Short Interspersed Elements **SINES** (100-300 bp) pari al 13% dell'intero genoma. LINES e SINES sono dei retrotrasposoni.
- Retrovirus-trasposoni-simili (450,000 copie, 8% del DNA totale).
- Trasposoni 300,000 copie (3% dell'intero genoma).



- **La proteina p40 (ORF1) è una chaperonina.**
- La ORF2 specifica per una proteina che ha attività sia di endonucleasi che di trascrittasi inversa.

Meccanismo che porta alla formazione di uno pseudo-gene





- La endonucleasi taglia il DNA nelle regioni ricche in AT, in particolare nella sequenza TTTT↓A.
- La trascrittasi inversa spesso non procede fino all'estremità 5' dell'RNA formando elementi tronchi e quindi inattivi.
- Il macchinario molecolare dei LINES è responsabile della retrotrasposizione dei SINES e della formazione di pseudo-geni.

Il genoma umano a confronto con altri organismi

Organismo	Lunghezza (Mb)	Numero di cromosomi	Numero di geni
Uomo	3230.54	46	25 000
Lievito	12.16	16	6 300
E.Coli	4.64	1	4 498

Homo sapiens http://www.ncbi.nlm.nih.gov/genome/51?genome_assembly_id=249422

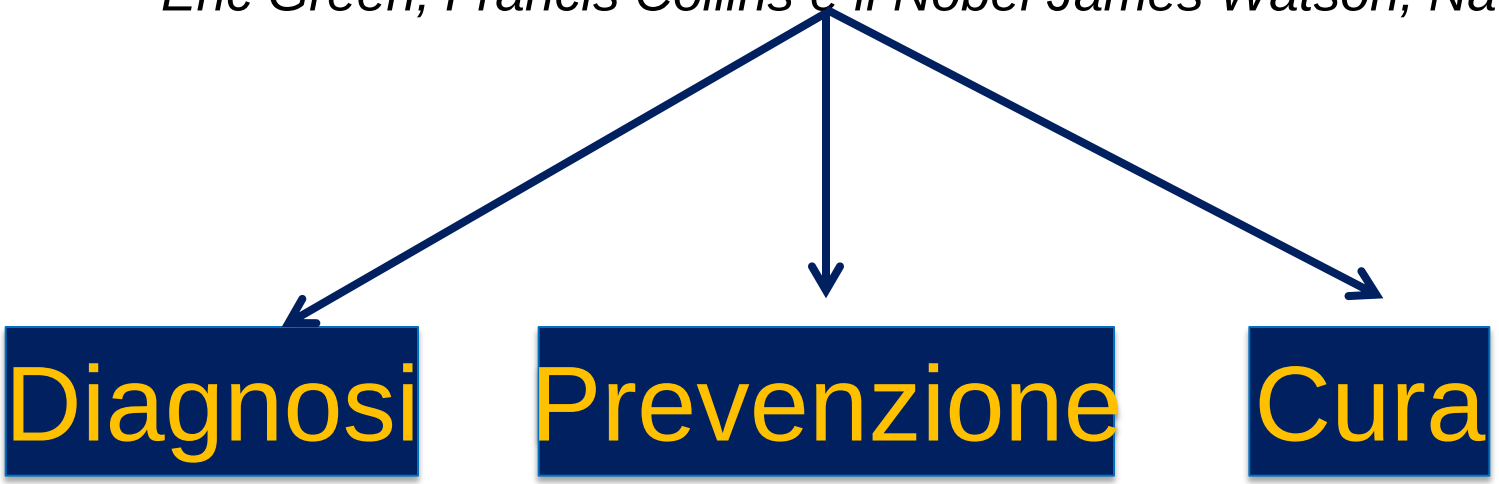
Saccharomyces Cerevisae http://www.ncbi.nlm.nih.gov/genome/15?genome_assembly_id=22535

Escherichia coli *http://www.ncbi.nlm.nih.gov/genome/167?genome_assembly_id=161521*

	Numero stimato dei geni umani
Prima del sequenziamento del Genoma	70 000 – 100 000
Dopo la prima bozza del Genoma	30 000 – 40 000
Al termine del sequenziamento	< 25 000
Ad oggi	20 500

"Il Progetto Genoma Umano ha cambiato il modo di fare scienza"

Eric Green, Francis Collins e il Nobel James Watson, Nature



1000 Genomes

A Deep Catalog of Human Genetic Variation

Home | About | Data | Analysis | Participants | Contact | Browser | Wiki | FTP search | Search

PROJECT OVERVIEW

Recent improvements in sequencing technology ("next-gen" sequencing platforms) have sharply reduced the cost of sequencing. The 1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation.

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied. This goal can be attained by sequencing many individuals lightly.

2001- 03





2010-12

Nature 2012

http://www.nature.com/nature/journal/v491/n7422/pdf/nature11632.pdf



ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

nature biotechnology

ARTICLES

Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases

Rong Chen[1,2,12], Lisong Shi[1,2,12], Jörg Hakenberg[1,2], Brian Naughton[3,11], Pamela Sklar[1,2,4], Jianguo Zhang[5], Hanlin Zhou[5], Lifeng Tian[6], Om Prakash[7], Mathieu Lemire[8], Patrick Sleiman[6], Wei-yi Cheng[1,2], Wanting Chen[5], Hardik Shah[1,2], Yulan Shen[5], Menachem Fromer[1,2,4], Larsson Omberg[9], Matthew A Deardorff[6], Elaine Zackai[6], Jason R Bobe[1,2], Elissa Levin[1,2], Thomas J Hudson[8], Leif Groop[7], Jun Wang[10], Hakon Hakonarson[6], Anne Wojcicki[3], George A Diaz[1,2], Lisa Edelmann[1,2], Eric E Schadt[1,2] & Stephen H Friend[1,2,9]

Genetic studies of human disease have traditionally focused on the detection of disease-causing mutations in afflicted individuals. Here we describe a complementary approach that seeks to identify healthy individuals resilient to highly penetrant forms of genetic childhood disorders. A comprehensive screen of 874 genes in 589,306 genomes led to the identification of 13 adults harboring mutations for 8 severe Mendelian conditions, with no reported clinical manifestation of the indicated disease. Our findings demonstrate the promise of broadening genetic studies to systematically search for well individuals who are buffering the effects of rare, highly penetrant, deleterious mutations. They also indicate that incomplete penetrance for Mendelian diseases is likely more common than previously believed. The identification of resilient individuals may provide a first step toward uncovering protective genetic variants that could help elucidate the mechanisms of Mendelian diseases and new therapeutic strategies.

Che cosa ci dicono i risultati del Progetto Genoma Umano?

· L'organizzazione del genoma umano

1. le regioni ricche di geni sono ricche in G e C
2. le regioni povere di geni sono ricche in A e T
3. Il cromosoma 1 ha il maggior numero di geni (2968), il cro mosoma Y il minor numero (231)

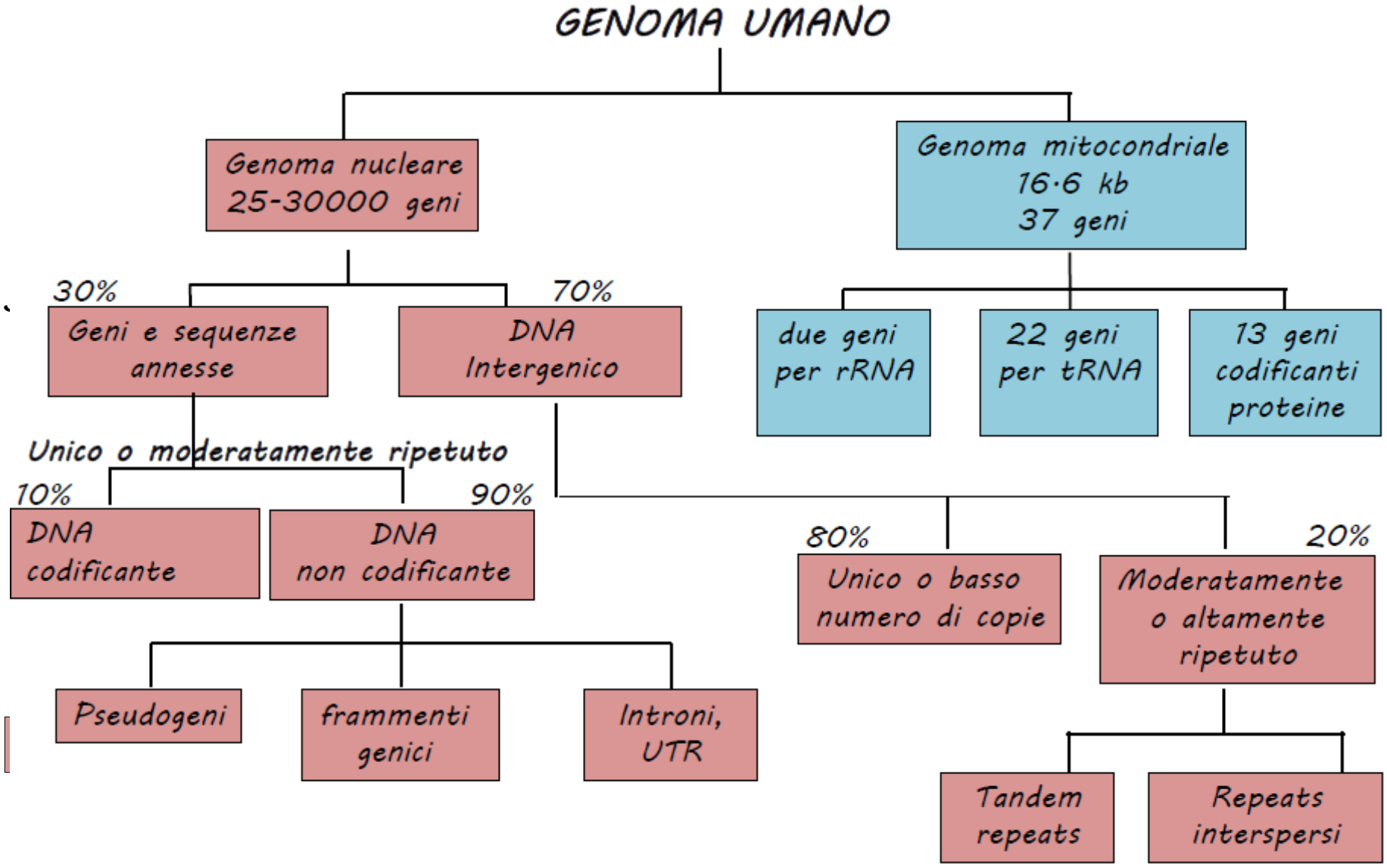
- Meno del 2% del genoma codifica per proteine

- Piu' del 50% del genoma e' costituito da sequenze ripetute con funzione sconosciuta

ORGANIZZAZIONE DEL GENOMA UMANO

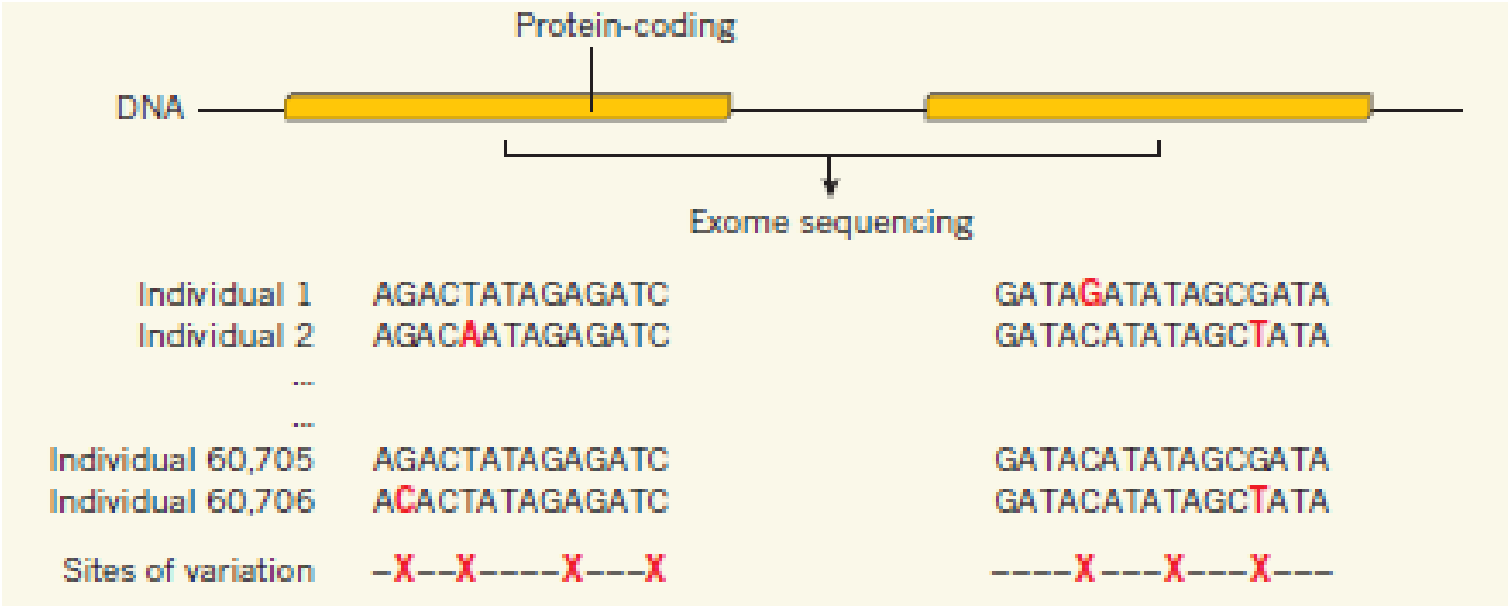
- Nuclear genome
  - 3200 Mb (3Mb frazione eucromatica + 200Kb eterocromatina)
  - 23 (XX) or 24 (XY) linear chromosomes
  - 2% coding
  - 25,000 protein coding genes
  - 1 gene/40kb
  - Introns
  - Repetitive DNA sequences (45%)
  - Recombination
  - Mendelian inheritance (X + auto, paternal Y)
  - Metilation

Organizzazione del genoma umano



Analysis of protein-coding genetic variation: Exome aggregation

- The Exome Aggregation Consortium (ExAC) reanalysed the raw DNA-sequencing data from the protein-coding part of the genome, known as the exome, of 60,706 individuals, aggregated from 14 distinct studies.
- Genetic variants (red) are compared to produce a data-base of all sites of variation between the individuals.



- This catalogue of human genetic diversity contains **an average of one variant every eight bases of the exome,** and provides direct evidence for the presence of widespread mutational recurrence.
- This catalogue was used to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation;
- identifying 3,230 genes with near-complete depletion of predicted protein-truncating variants**, with 72% of these genes having no currently established human disease phenotype**

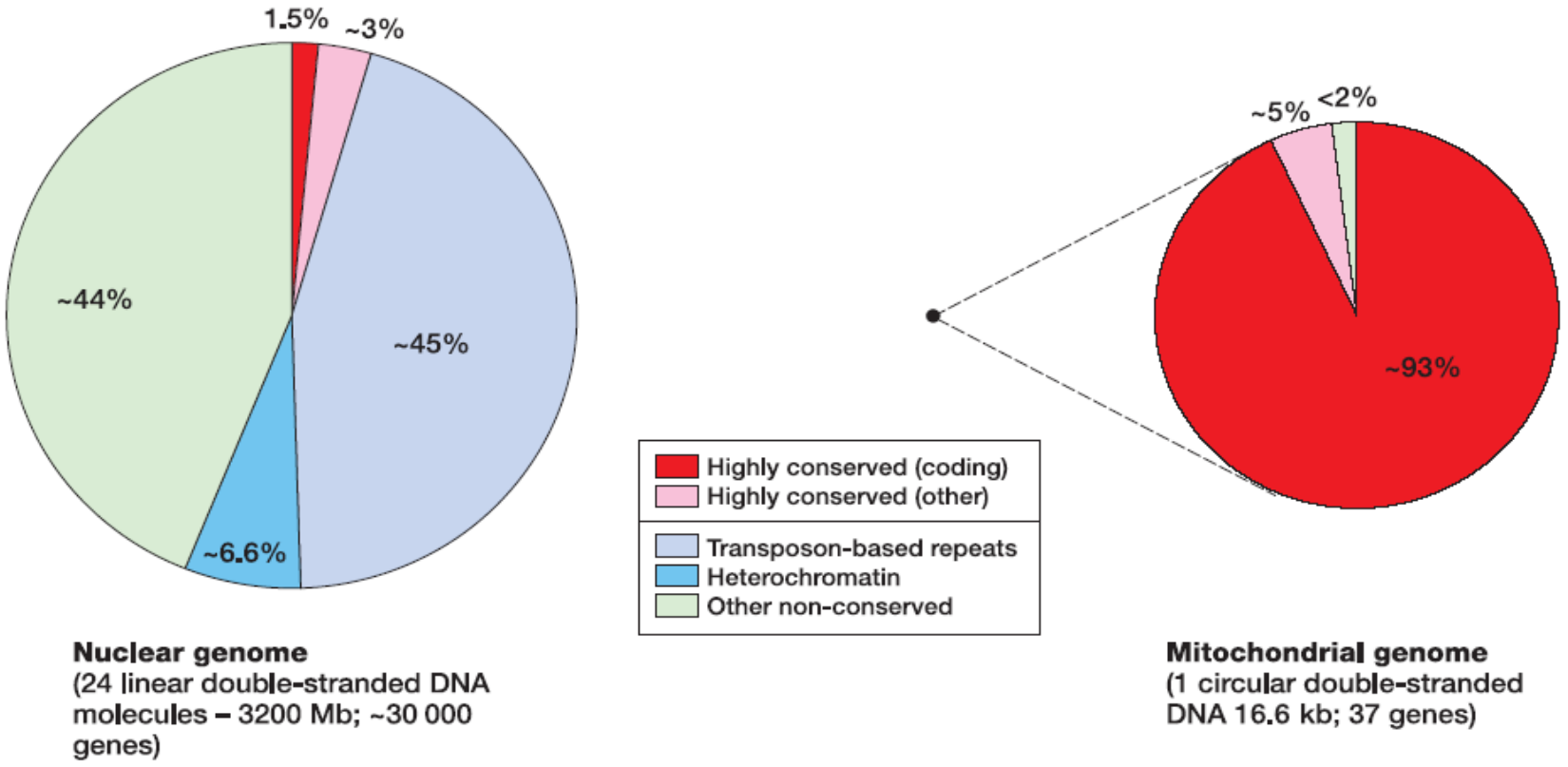
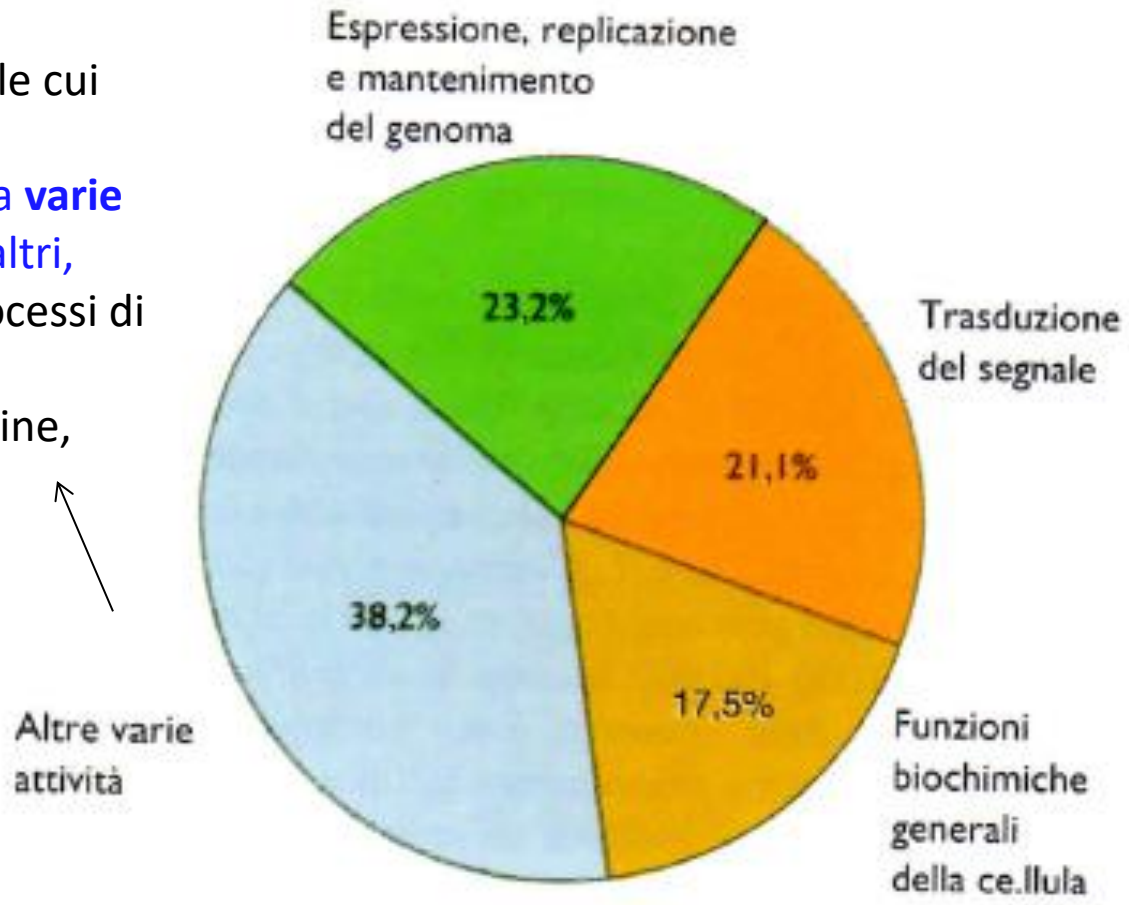


Figure 9.1: Organization of the human genome.

The dot in the middle represents the relative size of the mitochondrial genome using the same scale as for the nuclear genome. ***Note*** also the profound difference between the two genomes in the extent of highly conserved DNA (coding sequence, regulatory sequence etc.) and the fraction of highly repetitive noncoding DNA.

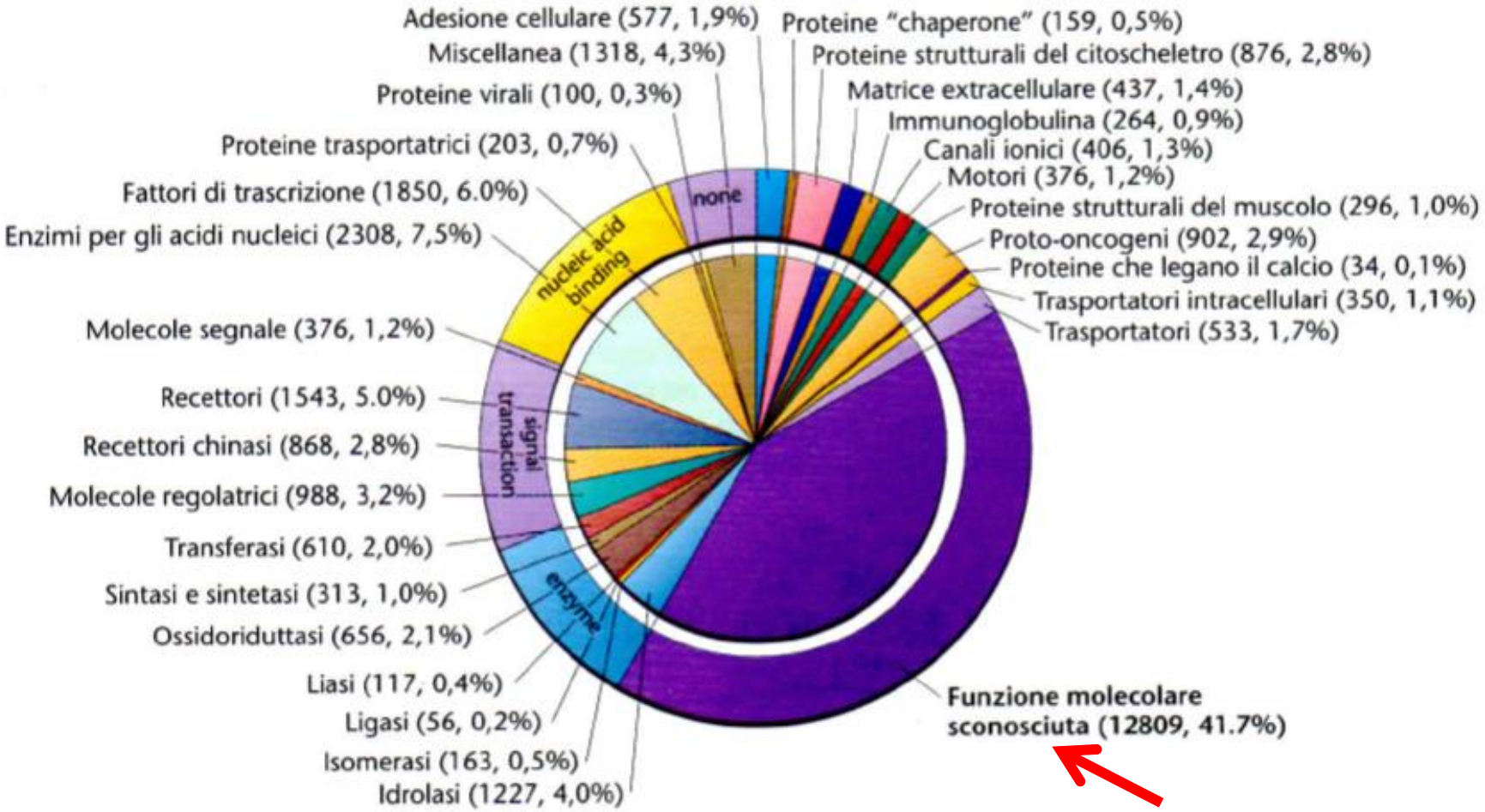
Classificazione del catalogo dei geni umani.

- **Geni umani codificanti proteine finora identificati.**
- Mancano appr. 13.000 geni, le cui funzioni sono sconosciute.
- La porzione corrispondente a **varie** altre attività include, tra gli altri,
  - le proteine coinvolte nei processi di trasporto biochimico,
  - nel ripiegamento delle proteine,
  - proteine immunologiche
  - e proteine strutturali.



Trasporto in-out cellula, corretto ripiegamento proteine,
Risposta immunitaria, sintesi proteine strutturali
(citoscheletro, muscoli)

Lista di funzioni assegnate ai 26 588 geni (presunti) dell' 1,5% del genoma umano sulla base della similarità con proteine a funzione nota



Tra i più comuni ci sono i geni che codificano per proteine coinvolte nel metabolismo degli acidi nucleici (7,5% di tutti i geni identificati), fattori di trascrizione (6,0%), recettori (5%), proteine chinasi (2,8%)
e proteine strutturali del citoscheletro (2,8%). Un totale di 12 809 proteine predette (41%) hanno una funzione sconosciuta: ciò riflette il lavoro che è ancora necessario per decifrare completamente il nostro genoma.

Non si può capire come è fatto il genoma umano e come funzionano i nostri geni se non li confrontiamo con quelli dei principali organismi modello:

Gli altri progetti genoma

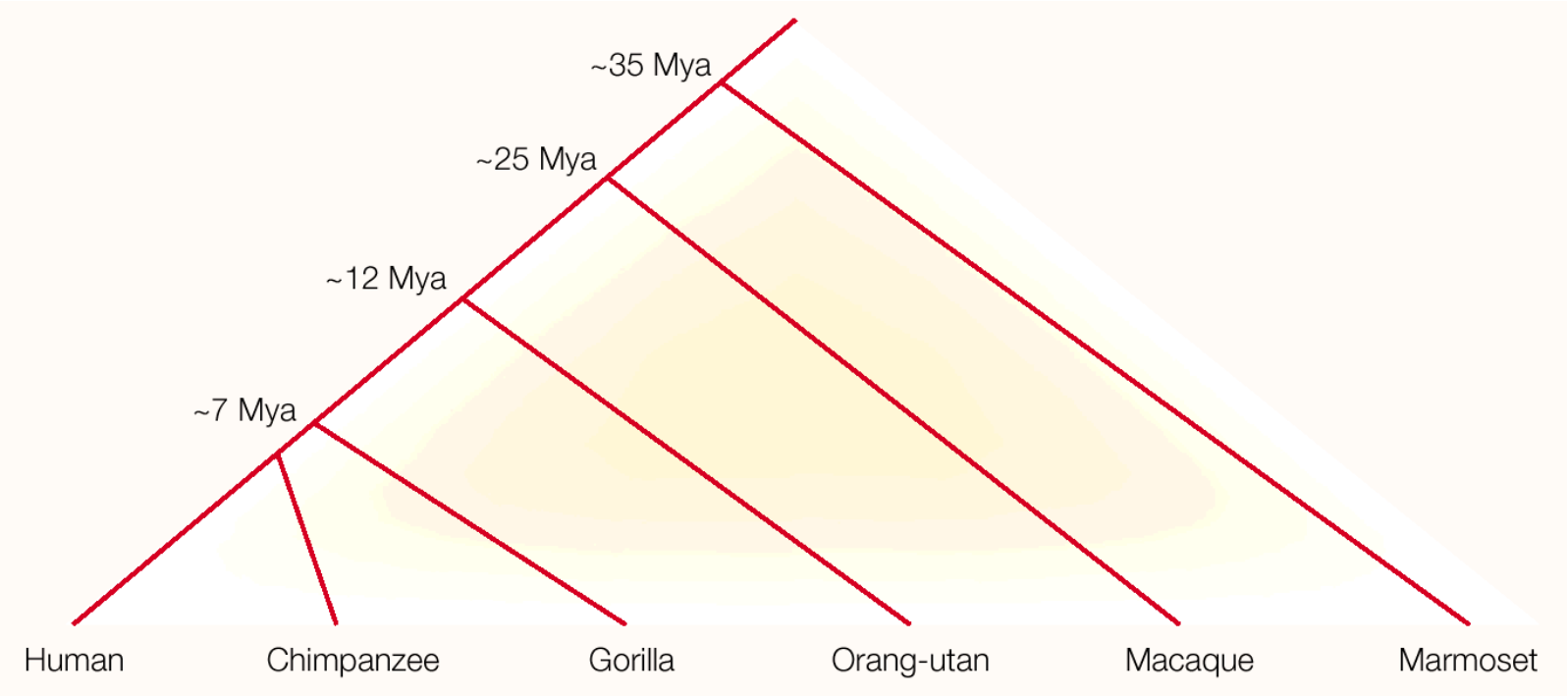
- *E. Coli*
- *S. Cerevisiae*
- *Drosophila Melanogaster*
- *Danio Rerio (zebrafish)*
- *Mus musculus*
- *Rattus Norvegicus*
- Primati non umani

Genomica comparativa e organismi modello

- **geni responsabili di malattie umane identificati mediante genomica comparativa con il topo**

Malattia	Fenotipo
Sindrome di Meckel	Difetti del sistema nervoso centrale
Difetto del tubo neurale	Difetti del tubo neurale spinale
Sindrome di Wolf-Hirschhorn	Ritardo di accrescimento fisico e ritardo mentale
Spinulosa decalvans	Ispessimento della pelle, perdita dei capelli
Convulsioni familiare benigna	Attacchi convulsivi dopo i 6 mesi di vita
Atassia cerebellare a esordio infantile	Perdita progressiva dell'equilibrio, sordità improvvisa
Neuromiopatia vacuolare	Debolezza muscolare progressiva
Tilosi con tumore dell'esofago	Ispessimento della pelle, sviluppo del tumore dell'esofago nell'adulto

Genome Sequence Divergence agrees with Paleontological and Evolutionary Record.



98.7%

98.3%

97.0%

94.5%

91.0%

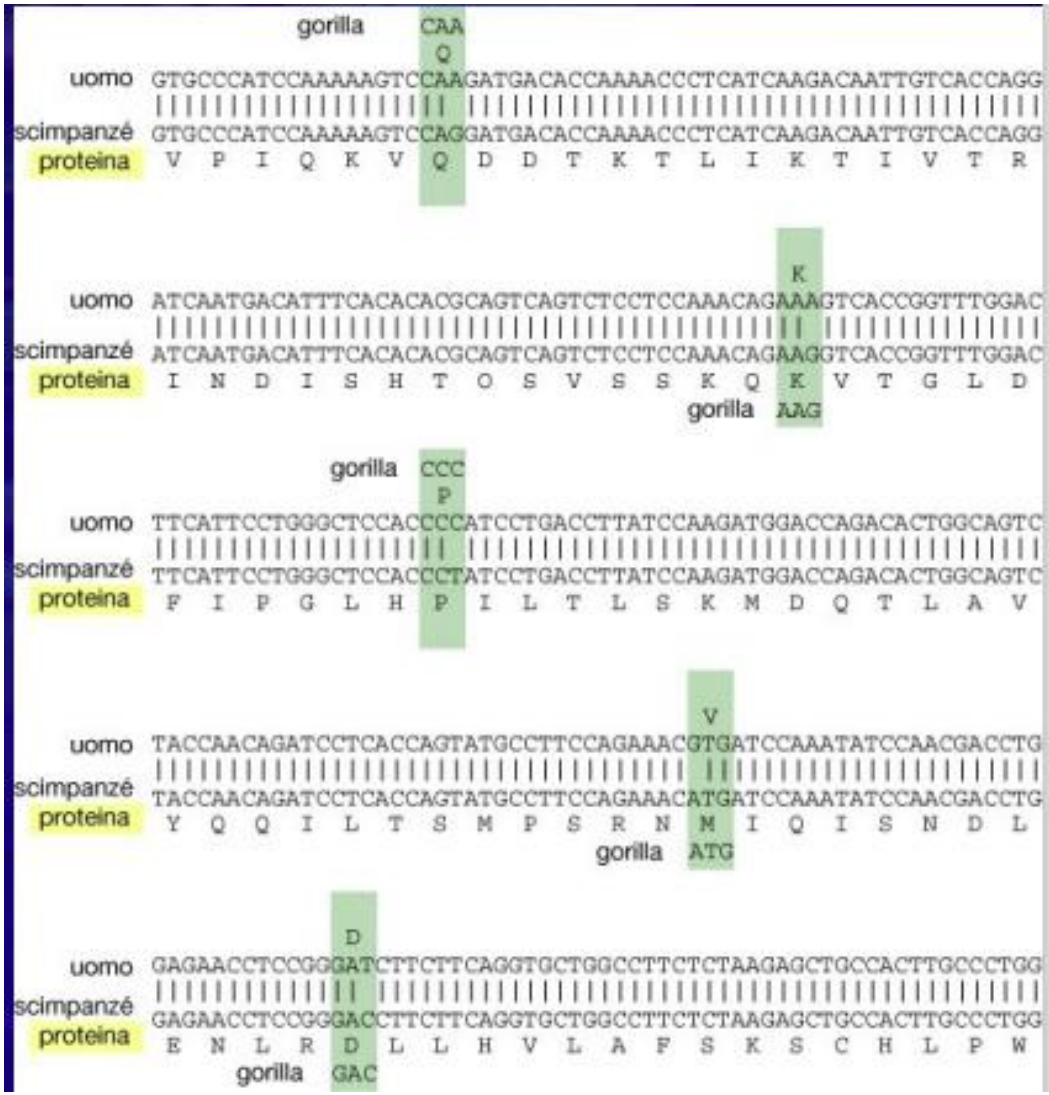
Origine ed evoluzione dei genomi

IL GENOMA UMANO: GLI ULTIMI 5 MILIONI DI ANNI

Uomo - Scimpanzè= 98,5% di omologia

- Usando una statistica corretta, considerando il numero di misure fatte, la similitudine si riduce al 96%.

- Recenti studi indipendenti hanno ricalcolato queste percentuali, ottenendo un range di omologia compreso tra il 66 ed il 76%
(in base al cromosoma)



IL GENOMA UMANO: GLI ULTIMI 5 MILIONI DI ANNI

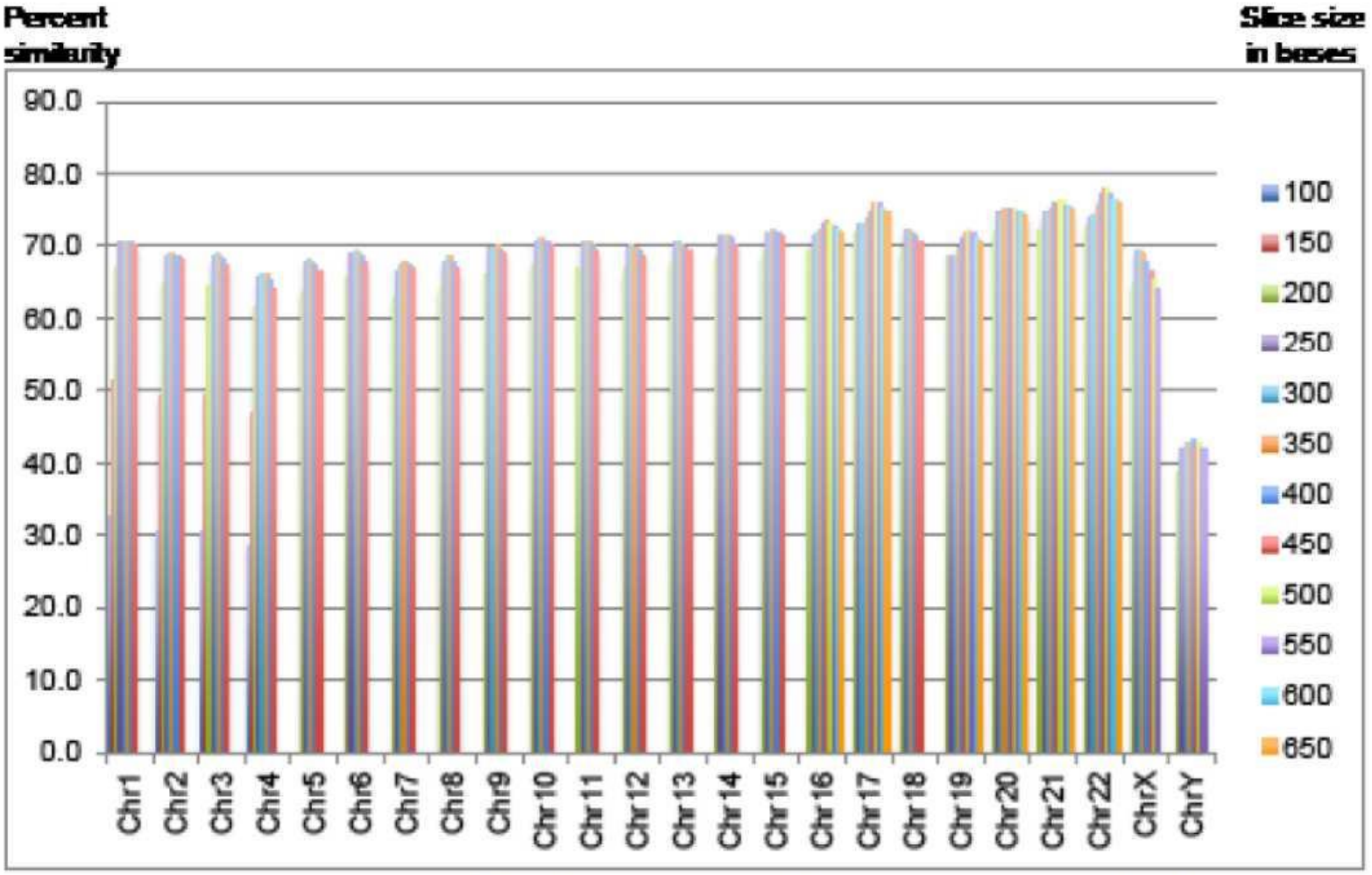


Fig. 1. Percent of chimp sequence aligned using optimized sequence slices sorted by chromosome.

Recenti studi indipendenti hanno ricalcolato queste percentuali, ottenendo un range di omologia compreso tra il 66 e il 76% (in base al cromosoma)

Origine ed evoluzione dei genomi

IL GENOMA UMANO: GLI ULTIMI 5 MILIONI DI ANNI

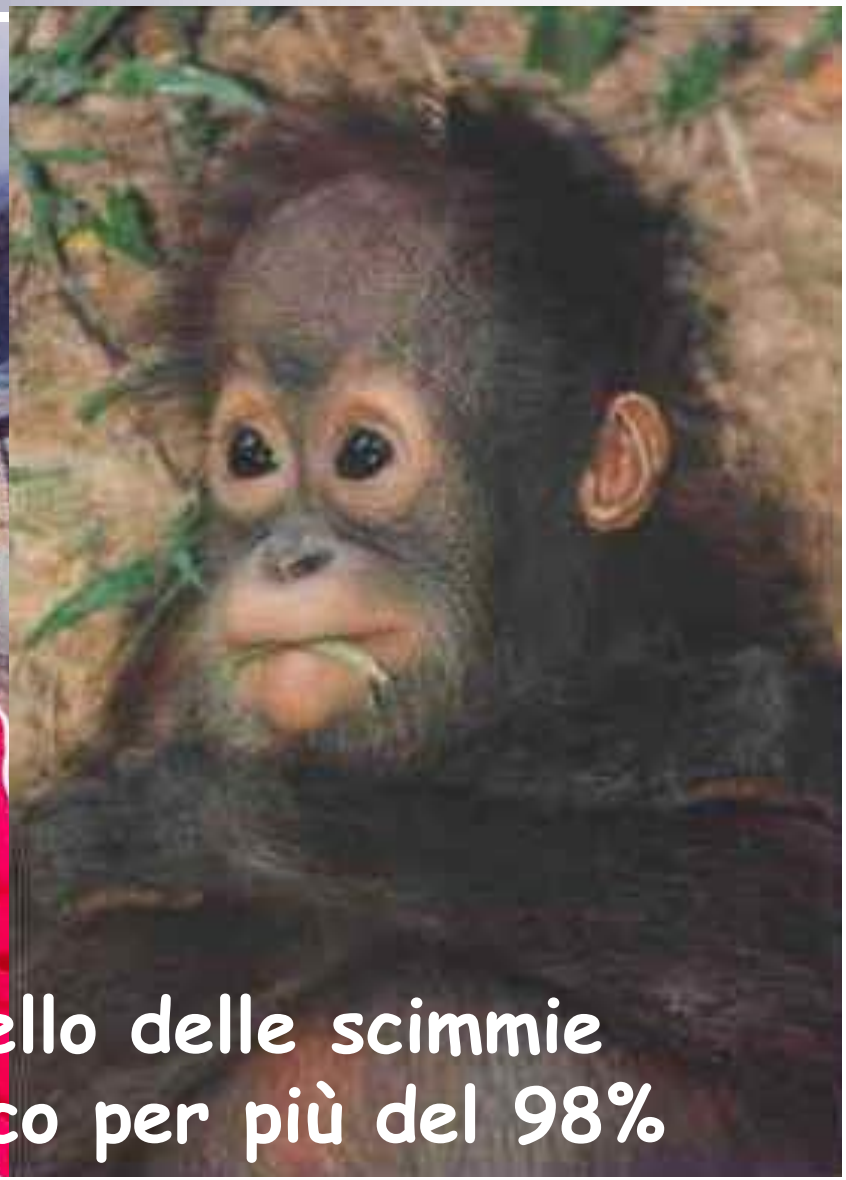
Che cosa ci rende diversi dalle scimmie?



Some candidate genes and gene families that may contribute to phenotypic differences between humans and apes

Gene(s)	Gene product(s)	Unusual hominid or human-specific features	Potential relevance to the human condition
Individual genes			
FOXP2	Putative transcription factor with polyglutamine tract and forkhead DNA binding domain	Two human-specific amino acid changes	Mutant humans have motoric speech disorder (developmental verbal dyspraxia). Region positively selected and fixed in humans <200,000 years ago (Enard et al. 2002b; Zhang et al. 2002)
MYH16	Myosin heavy chain 16	Human-specific 2-bp deletion causing frameshift—predicted 76-kD unstable head domain	Claimed to be cause of reduction in the type II fibres of human jaw muscle. (Stedman et al. 2004; Perry et al. 2005)
CMAH	CMP-Neu5Ac hydroxylase	92-bp deletion of exon 6 causing frameshift and inactive enzyme. Fixed in modern humans	Absence of sialic acid Neu5Gc. Change in resistance or susceptibility to pathogens. Loss of ligand for some Siglecs. Dated ~2.5–3 Mya. Dietary Neu5Gc in meat became foreign antigen (Chou et al. 1998, 2002; Irie et al. 1998; Hayakawa et al. 2001)
MAOA	Monoamine oxidase A	Human-specific nonconservative change Glu151Lys in active site	Substitution affects protein dimerization according to a 3D structural model and predicts functional change (Andres et al. 2004)
ASPM	Modulator of mitotic spindle in neural progenitors?	Accelerated evolution in ape and human lineages	Deletions in ASPM lead to microcephaly. Presumed to be related to increased brain size and/or other features of human brain (Zhang 2003; Dorus et al. 2004; Evans et al. 2004; Kouprina et al. 2004b; Mekel-Bobrov et al. 2005)
MCPH1	Microcephalin	As above	As above (Dorus et al. 2004; Evans et al. 2004, 2005)
TTR	Transthyretin	Decreased expression in humans in blood and brain	May be related to altered thyroid hormone metabolism in humans versus chimpanzees (Gagneux et al. 2001)
ST6GAL1	Alpha 2–6 sialyltransferase	Apparent human-specific up-regulation on epithelia	Can explain relative resistance of chimpanzees to human influenza A virus. Other consequences unknown. (Gagneux et al. 2003)
EMR4	EGF-TM7 receptor family	Human-specific deletion in exon 8. Frameshift	Predominantly expressed by immune system cells. Functional significance unknown (Hamann et al. 2003)
PCDH11Y	Protocadherin XY	Duplicated onto Y in Yp11.2/Xq21.3 pseudoautosomal region only in humans	Expressed from Y and escapes X-inactivation? Y copy has undergone structural changes. Selectively expressed in brain. Probable adhesion molecule. Significance unknown, hypothesized to be involved in brain development, lateralization and schizophrenia risk (Ross et al. 2003; Blanco-Arias et

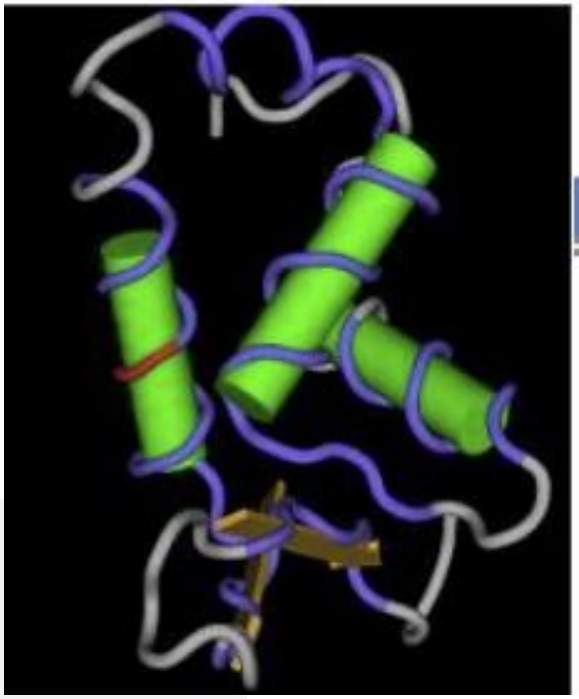




Il genoma umano e quello delle scimmie antropomorfe è identico per più del 98%

La Proteina FOXP2

- Il gene *Foxp2* (cromosoma 7) --> **Proteina Foxp2 :**
  - costituita da 715 aa
  - fattore di trascrizione della famiglia forkhead/wind-helix dei FT regola espressione genica legandosi al DNA.
  - espressa nel cervello fetale ed adulto ed altri organi (cuore polmoni e sistema digestivo).
  - contiene una regione ricca di poli-glutammine
  - Si lega direttamente a circa 300-400 promotori del genoma umano che regola l'espressione di una varietà di geni
  - è tra il 5% delle proteine più conservate
  - Presente in molti vertebrati , dove ha importanti ruoli nel suono degli uccelli e eco-localiizzazione nei pipistrelli.
  - mostra particolari similitudini in tutti i primati



- **In rosso è mostrata la posizione della mutazione**
- **dell' aa 325**

Questa è dovuta ad una sostituzione di un aa che inibisce il dominio di legame della proteina FOX2 al DNA

Il gene *Foxp2* svolge un ruolo essenziale nello sviluppo della comunicazione sociale.

- ***FOXP2:*** un gene presente anche nei topi e negli scimpanze';
- tale gene ha subito una **mutazione recente nell'evoluzione** (circa 120-200 mila anni fa) producendo una nuova sequenza genica che si e' fissata nella nostra specie



- Grazie alle proteine prodotte dalla nuova sequenza genica,
  - **bocca e laringe si sono perfezionate**
  - Ciò permettere l'articolazione di suoni complessi

- Le persone con una sola copia di questo gene funzionante hanno problemi:
  - **articolare il linguaggio,**
  - **seguire le regole grammaticali**
  - **muovere i muscoli del viso.**

Origine ed evoluzione dei genomi

IL GENOMA UMANO: GLI ULTIMI 5 MILIONI DI ANNI

- **Che cosa ci rende diversi dalle scimmie?**
  - ✓ Quello che ci rende umani probabilmente non è il genoma umano di per sé,
  - ✓ ma il **modo in cui il genoma funziona**
  - ✓ Sottili cambiamenti nei profili di espressione dei geni coinvolti in processi di sviluppo e nella specificazione delle interconnessioni all'interno del sistema nervoso

  - ✓ Si ipotizza che
  - **geni simili**
  - **le interazioni fra gruppi di geni simili**

    possono generare 'programmi' molto diversi.

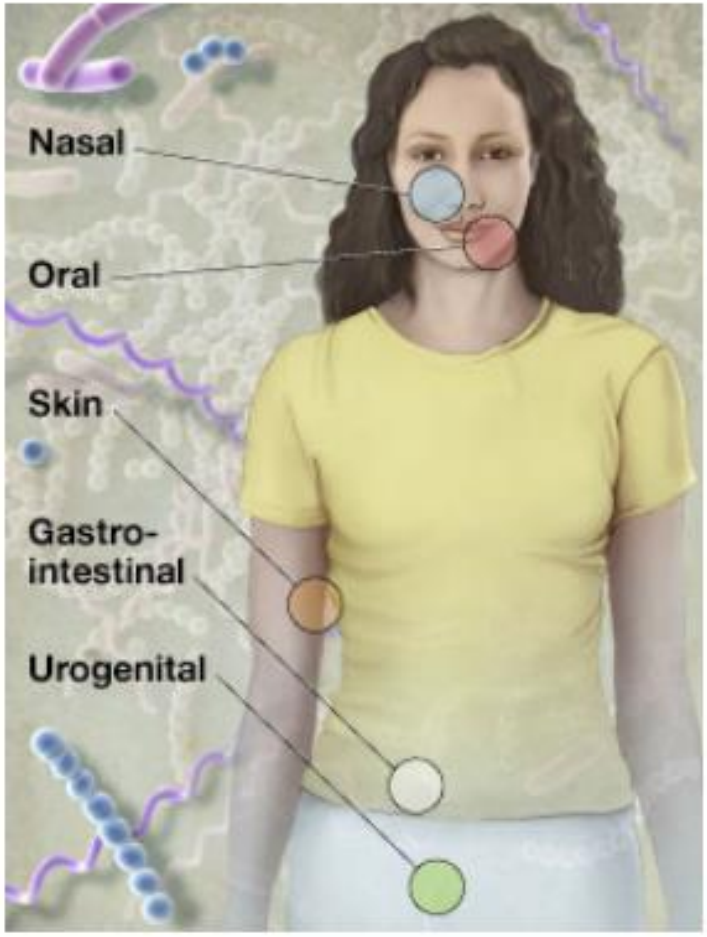
- **Perché l'uomo ha un alto rischio di cancro, sebbene lo scimpanzé sviluppa raramente tale patologia?**

- Gli scienziati hanno analizzato campioni di cervello di ambedue le specie.
- Essi hanno riscontato che **differenze nella metilazione del DNA** possono contribuire a **indurre variazioni fenotipiche.**
- Inoltre i risultati evidenziano che la **metilazione del DNA**, svolge un ruolo significativo per alcune malattie correlate ai fenotipi nell'uomo, inclusi cancro e autismo.



Il nostro secondo genoma, il microbioma

Microbioma: l'insieme del patrimonio genetico della totalità dei microrganismi di un ambiente definito



Corpo umano

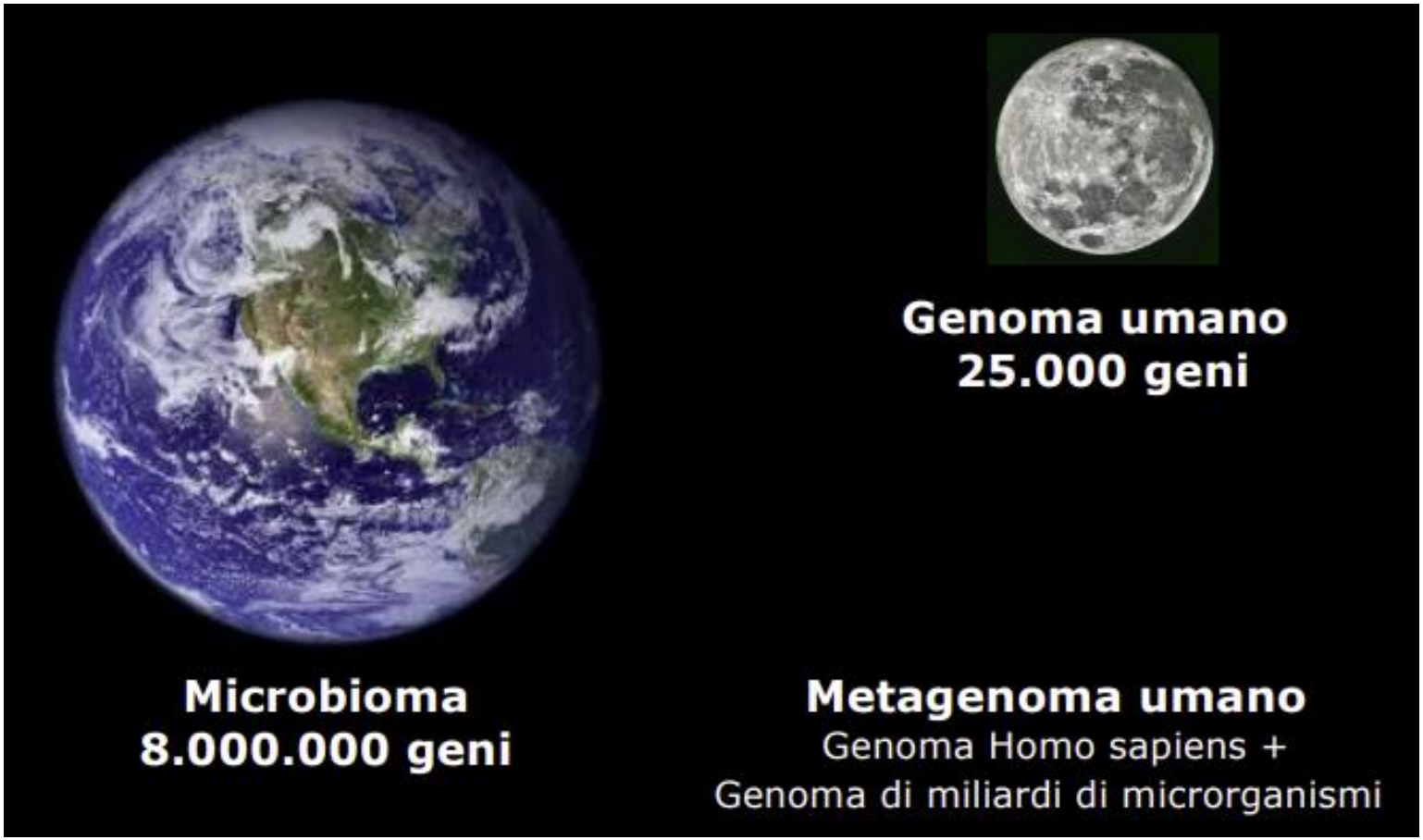
$10x10^{12}$ cellule

$100x10^{12}$ cellule batteriche **(microbiota)**

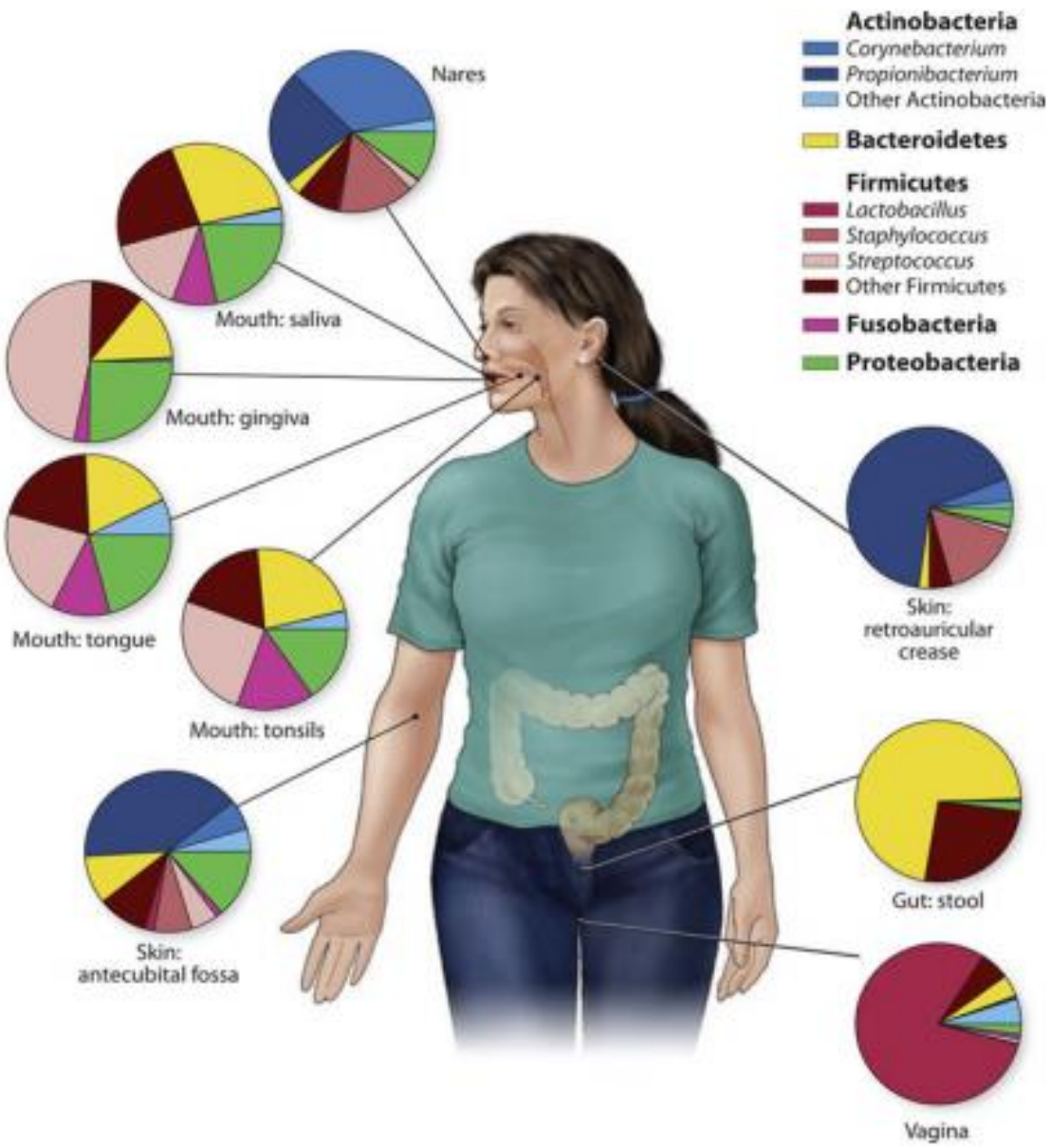
Per 1 cellula umana vi sono 10 cellule microbiche

SUPER-ORGANISMO

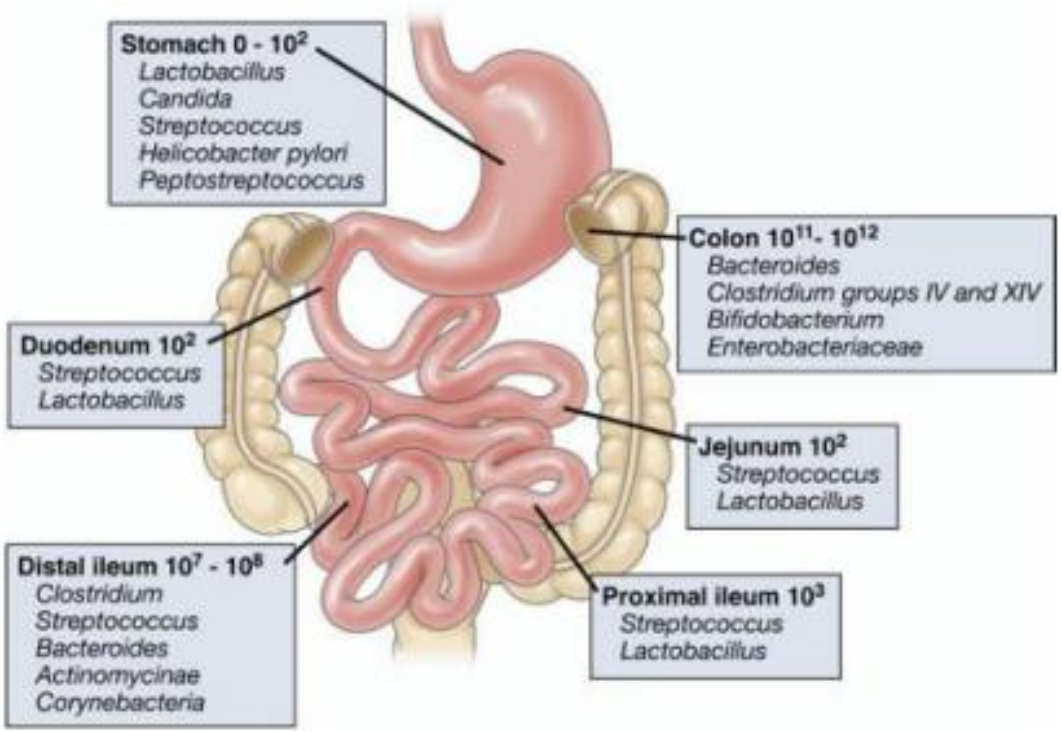
Il nostro secondo genoma, il microbioma



Il microbiota umano



Il microbiota intestinale



Microbiota intestinale

–2 kg di microorganismi

–interagiscono su 300mq di intestino (quasi un campo da tennis)

- Inizia a formarsi alla nascita
- Il parto e la nutrizione del neonato influiscono sullo sviluppo iniziale del microbioma
- L'ambiente e la dieta influenzano il microbioma maturo
- La composizione del microbioma cambia con l'età

Quali sono i benefici del microbiota intestinale?

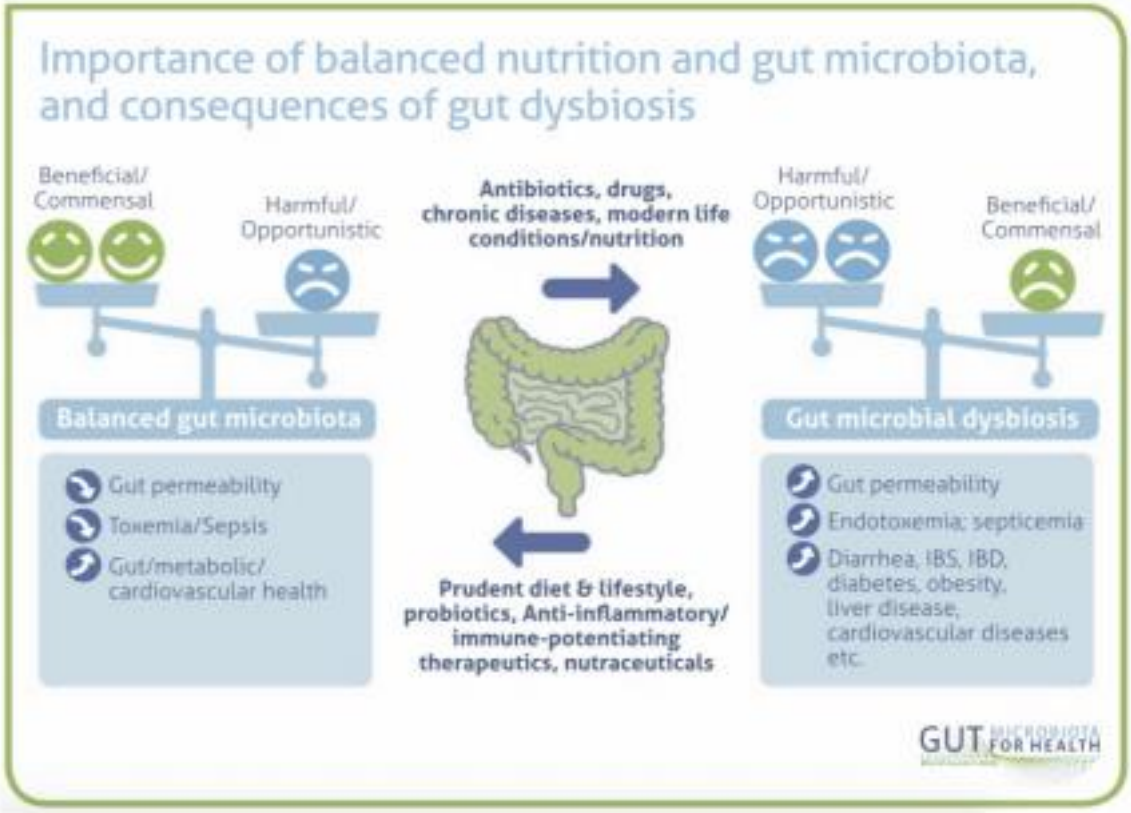
- Influenza lo sviluppo delle cellule della mucosa intestinale
- Serve come barriera contro i microrganismi patogeni (possibili fonti di malattie)
- Stimola le difese immunitarie
- E' fonte di energie e di sostanze nutritive (es. vitamine, prodotti di fermentazione delle fibre)
-



Importanza del microbiota intestinale

L'alterazione qualitativa e quantitativa della microflora intestinale (*disbiosi)* è legata allo sviluppo di diverse patologie, quali:

- infiammazioni croniche intestinali
- Obesità
- Diabete di tipo I
- Cancro del colon-retto
- Malattie neurodegenerative (es. Alzheimer, Parkinson, ….)
- ……



Metagenomica

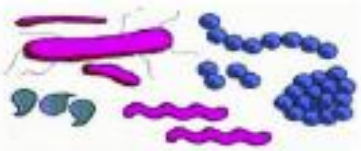
La **metagenomica** è una nuova disciplina che studia l'insieme dei diversi materiali genetici (detto **metagenoma**) derivanti dalle diverse specie presenti in un determinato ambiente.
Ad esempio il materiale genetico dell'uomo e quello derivante dal suo microbiota (detto microbioma), rappresenta il metagenoma umano.

L'**approccio metagenomico** è applicato con successo per rivelare la diversità genetica e tassonomica delle comunità microbiche.

Non richiede l'isolamento e la messa in coltura di ogni singola specie, e la sua identificazione su basi biochimiche o morfologiche.

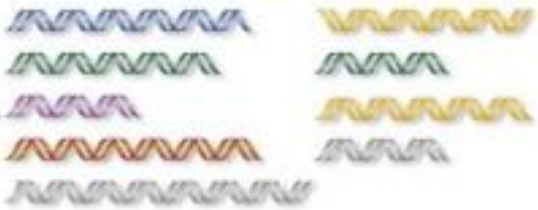


Sequenziamento Metagenomico





Campioni clinici o ambientali



Estrazione DNA

Approccio Metabarcoding

Approccio Shotgun



(e.g. 16S rRNA, ITS, Cox1, ..)

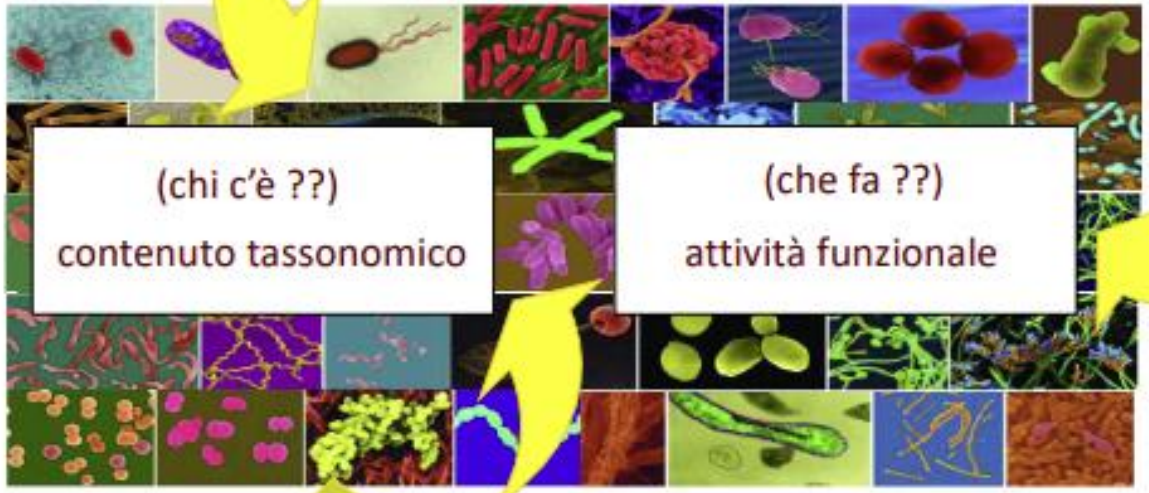
sequenziamento NGS
analisi dei dati NGS



Sequenziamento Metagenomico

Sequenziamento Shotgun

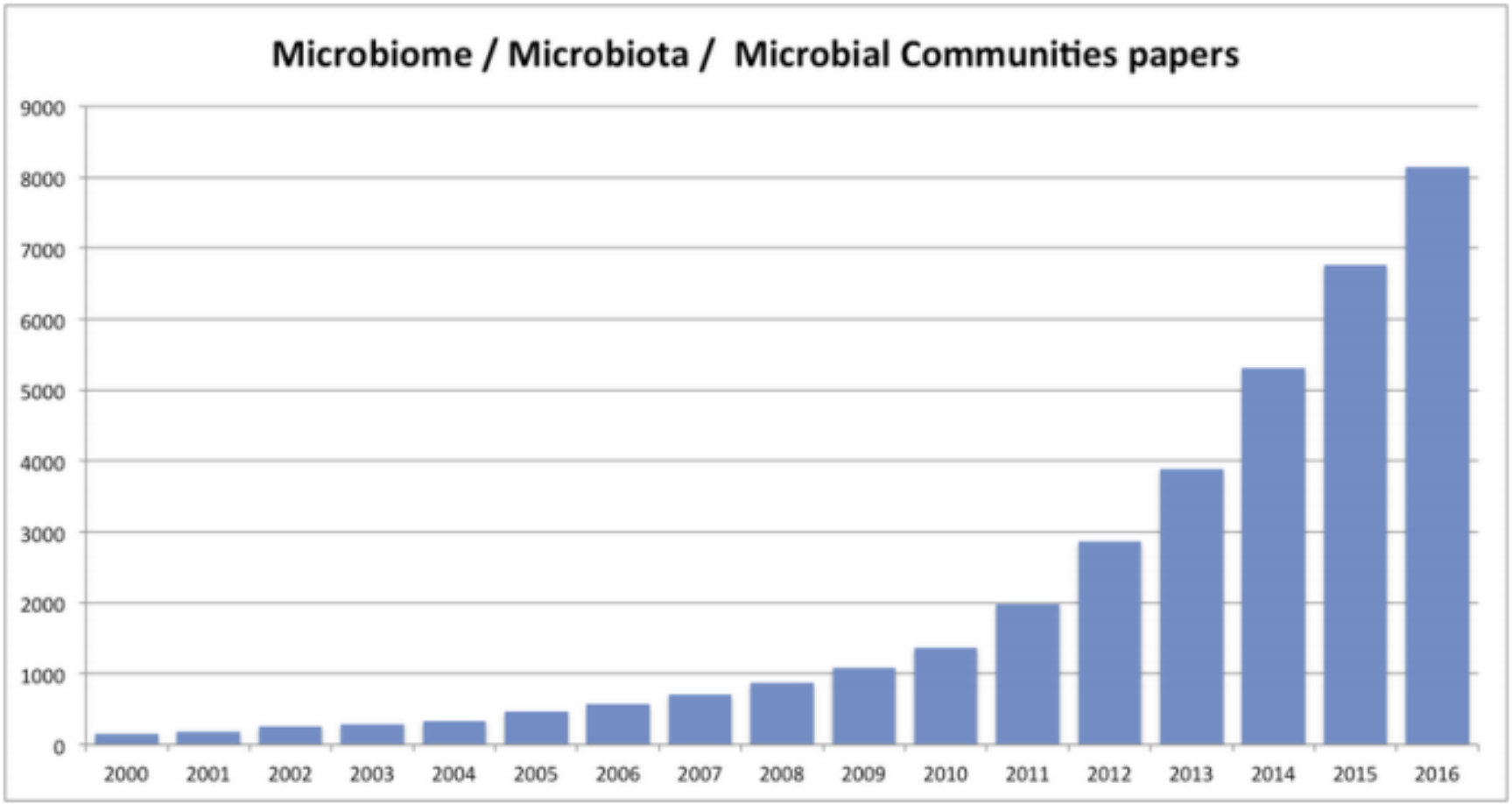
- identifica specie, geni e attività funzionali di comunità microbiche complesse
- coverage limitato per rilevare specie rare
- costoso sia in termini di sequenziamento che di analisi computazione



Sequenziamento del Metabarcoding

- elevata sensibilità nella identificazione di specie
- poco costoso in termini di sequenziamento che di analisi computazione
- basato su PCR con primer universali
- richiede appropriato database di riferimento (es. RDP per 16S, ITSoneDB per ITS1)
- possibili distorsioni dovute a diversa efficienza di amplificazione in diversi gruppi tassonomici
- nessuna informazione funzionale

Microbioma: un universo in via di esplorazione



Fonte:PubMed

Il Progetto Genoma Umano….. dal 2003 ad oggi



Le sfide future

Cosa ancora non sappiamo ……

- L' esatto numero di geni presenti nel genoma umano
- La funzione di circa il 50% dei geni
- La regolazione dell'espressione dei geni
- L' informazione contenuta nella porzione non codificante e la sua funzione
- I geni coinvolti nelle malattie complesse
- La correlazione tra variazioni della sequenza del genoma tra singoli individui e la suscettibilità a specifiche malattie
- La relazione tra microbiota e organismo
- ecc

Concludendo....

L'analisi integrata di dati molecolari (genoma, trascrittoma, proteoma, microbioma), metabolici, funzionali, ecc. può essere utile per fornire le informazioni necessarie per la comprensione di tutti meccanismi alla base dei processi fisiologici e patologici dell'uomo e quindi a far luce sul segreto della vita, ma……..

La strada è ancora lunga!!

