Genoma Umano, Polimorfismi e Mutazioni

Il GENOMA: cos'è ?

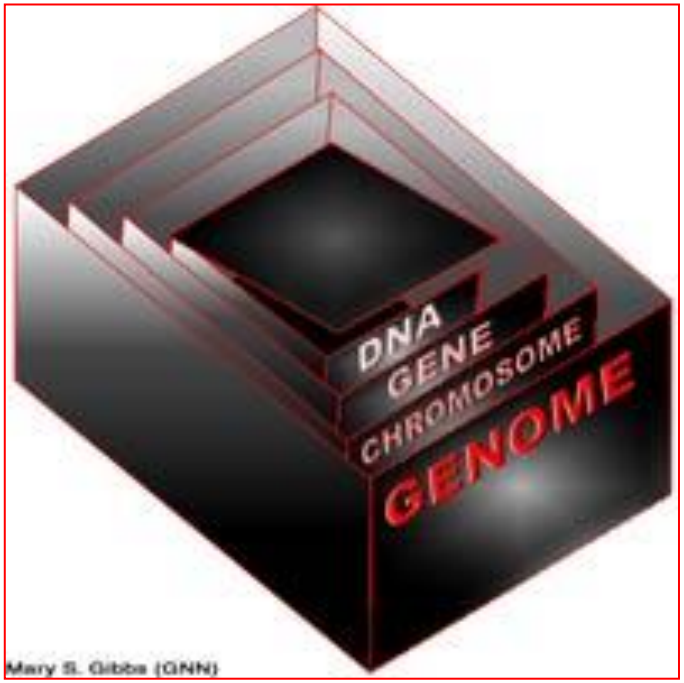
- Il genoma è il materiale genetico presente in qualsiasi organismo vivente.
- Costituisce l'intera serie di istruzioni ereditarie per la costruzione, la gestione e il mantenimento di un organismo e la trasmissione della vita alla generazione successiva.
- Nella maggior parte degli esseri viventi, il genoma è costituito da una sostanza chimica chiamata DNA.
- Anche l'RNA è usato nei genomi di alcuni organismi

- Il genoma contiene i geni, cioè le unità funzionali che codificano per tutte le caratteristiche specifiche dell'organismo.

Il GENOMA: cos'è ?

A volte, come nell'uomo,
il genoma è diviso in cromosomi,
i cromosomi contengono geni
i geni sono segmenti di DNA

Ogni specie vivente ha il proprio genoma distintivo: il genoma del cane, il genoma del grano, il genoma umano e così via



Rivelare la sequenza nucleotidica di un genoma può aiutare a comprendere come funziona

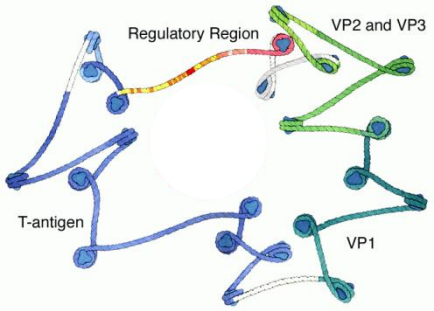
SV40 genome

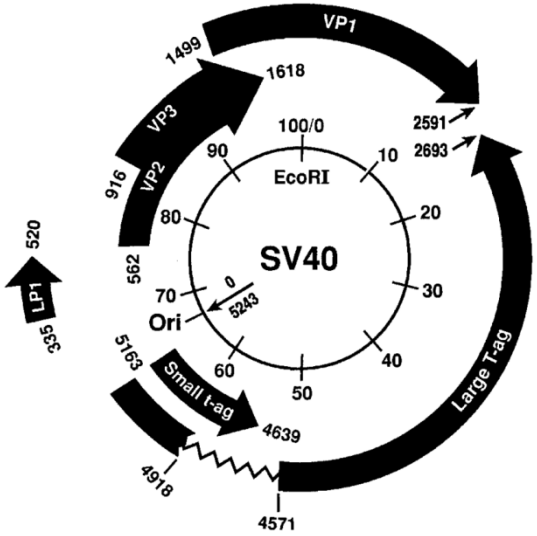
- SV40 (simian virus 40) è un virus della famiglia polyomavirus.
- Ospite naturale è la scimmia, ma può moltiplicarsi anche in cellule umane

SV40 genome

- SV40 (simian virus 40) è un virus della famiglia polyomavirus.
- Ospite naturale è la scimmia, ma può moltiplicarsi anche in cellule umane

- Il genoma di SV40 è costituito da una molecola di DNA circolare di circa 5000 nucleotidi
- Il sequenziamento ha consentito di generare una mappa trascrizionale funzionale: (i) una regione precoce regolatoria contenente i geni codificanti per *antigene T* ed *antigene t*; (ii) una regione tardiva codificante per le proteine strutturali *VP1, VP2*, *VP3*.

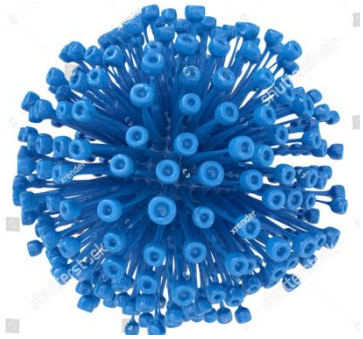




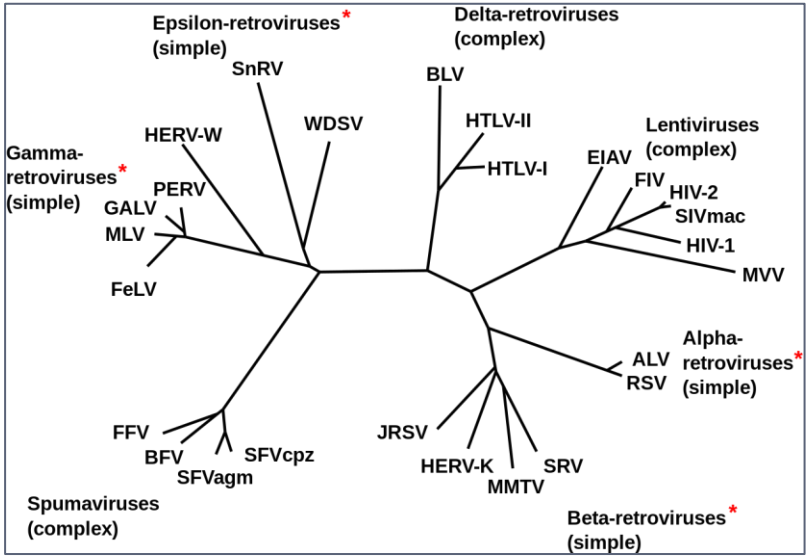
MAPPA FUNZIONALE DEL GENOMA DI SV40

Retrovirus genome

- I retrovirus costituiscono una larga famiglia di virus con genoma a RNA a singola catena.



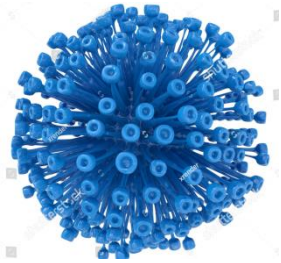
- I diversi membri di questa famiglia sono stati identificati in tutte le specie di mammifero dove sono stati studiati

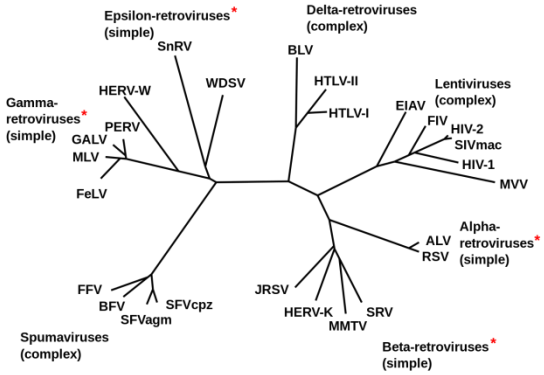


Retrovirus genome

- Caratteristica dei retrovirus è di replicarsi attraverso un intermedio a DNA che si integra nel genoma della cellula ospite, divenendone parte integrante.

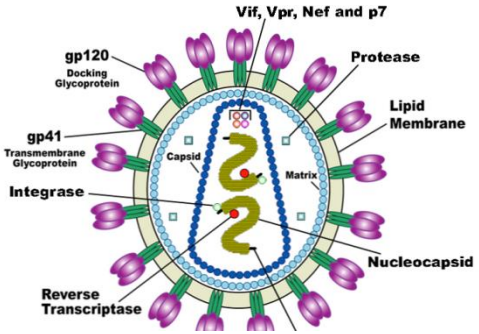
Trascrittasi inversa		*RNA Pol*	
RNA	→	**DNA**	→ **RNA**





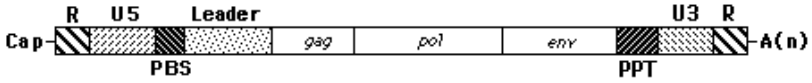
- Il genoma di un retrovirus è costituito da una molecola di RNA di circa 10 kb, che codificano per:

- Gag: proteine del capside
- Pol: trascrittasi inversa
- Env: proteine di membrana



- Presenta inoltre regioni regolatrici alle estremità: i segmenti LTR (long terminal repeats) , indispensabili per regolare l'attività trascrizionale e l'integrazione nel DNA della cellula ospite

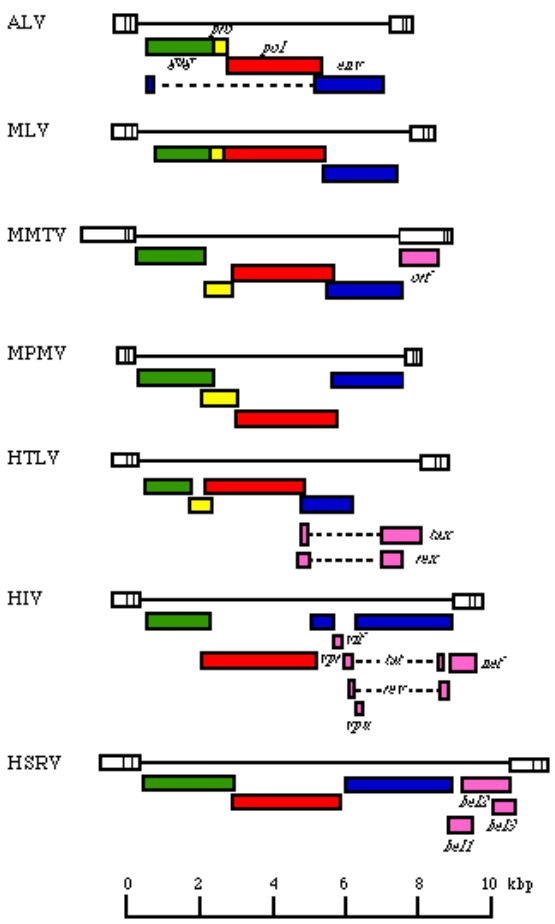
5' - gag - pol - env - 3'



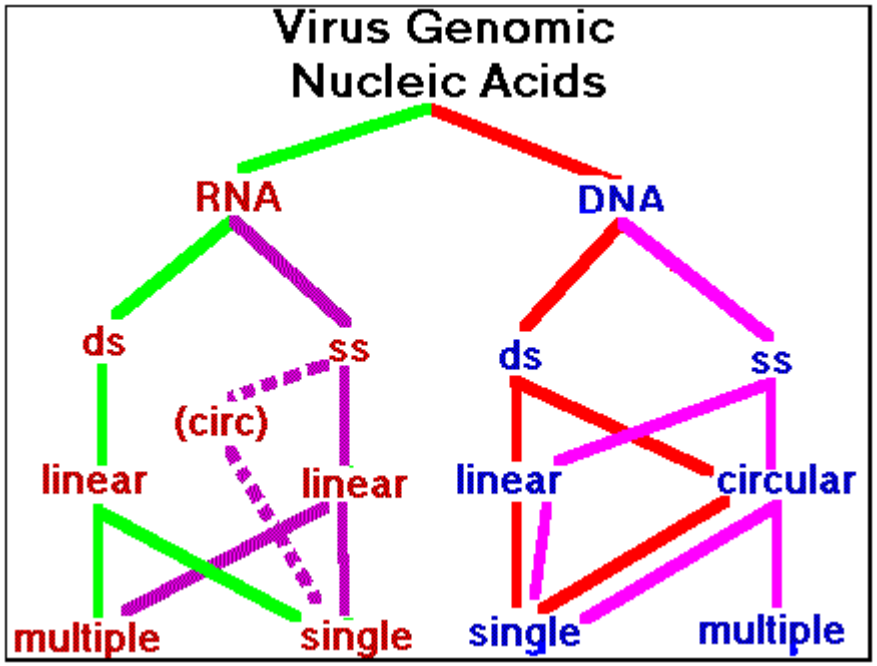
Retrovirus genome: variabilità

- Aspetti che emergono dall'analisi dei genomi dei retrovirus sono:
  - una **struttura generale comune conservata**
  - una notevole **variabilità di sequenze tra le diverse specie** virali

- Per la loro capacità di duplicarsi e produrre rapidamente **nuove generazioni**, i virus rappresentano l'esempio vivente di **come i genomi si possano modificare per meglio adattarsi all'ambiente**, cioè agli organismi **ospiti** che infettano



I Genomi Virali



La capacità di adattamento dei virus ha portato nel corso dell'evoluzione a sfruttare ogni tipo e formato di acido nucleico nella costituzione del proprio genoma:

DNA o RNA; a singola o doppia catena; molecole lineari o circolari; costituito da uno o più segmenti

Indagini genetiche e sequenziamento del genoma hanno consentito di definire una mappa funzionale, in cui è possibile riconoscere:

- i geni codificanti proteine
- gli elementi funzionalmente rilevanti per la regolazione dell'espressione genica
- gli elementi funzionalmente rilevanti per la duplicazione del genoma

Comprendere il genoma umano: primo, determinare la sua sequenza nucleotidica

- Così come indicato dall'analisi dei genomi virali, possiamo dedurre che il sequenziamento di un genoma, compreso quello umano, sia un passo importante verso la sua comprensione
- Il genoma umano è costituito da DNA organizzato in 22 autosomi e due cromosomi sessuali, X e Y.
- Conoscerne la sequenza nucleotidica può fornire informazioni utili per la sua caratterizzazione funzionale.
- Ciò può aiutare anche a comprendere le basi genetiche delle malattie umane

Il Progetto Genoma Umano

Inizio del sequenziamento in 6 laboratori USA nel 1990

Feb ‘96, 97, 98 - Bermuda Meetings → Coordinamento Internazionale

- Coordinamento
- Tecnologia
- Costi

- Rilascio dei dati
- Qualità dei dati
- Quantità dei dati

Sequenze primarie del genoma umano rilasciate rapidamente in database pubblici accessibili.

BCM-HGSC

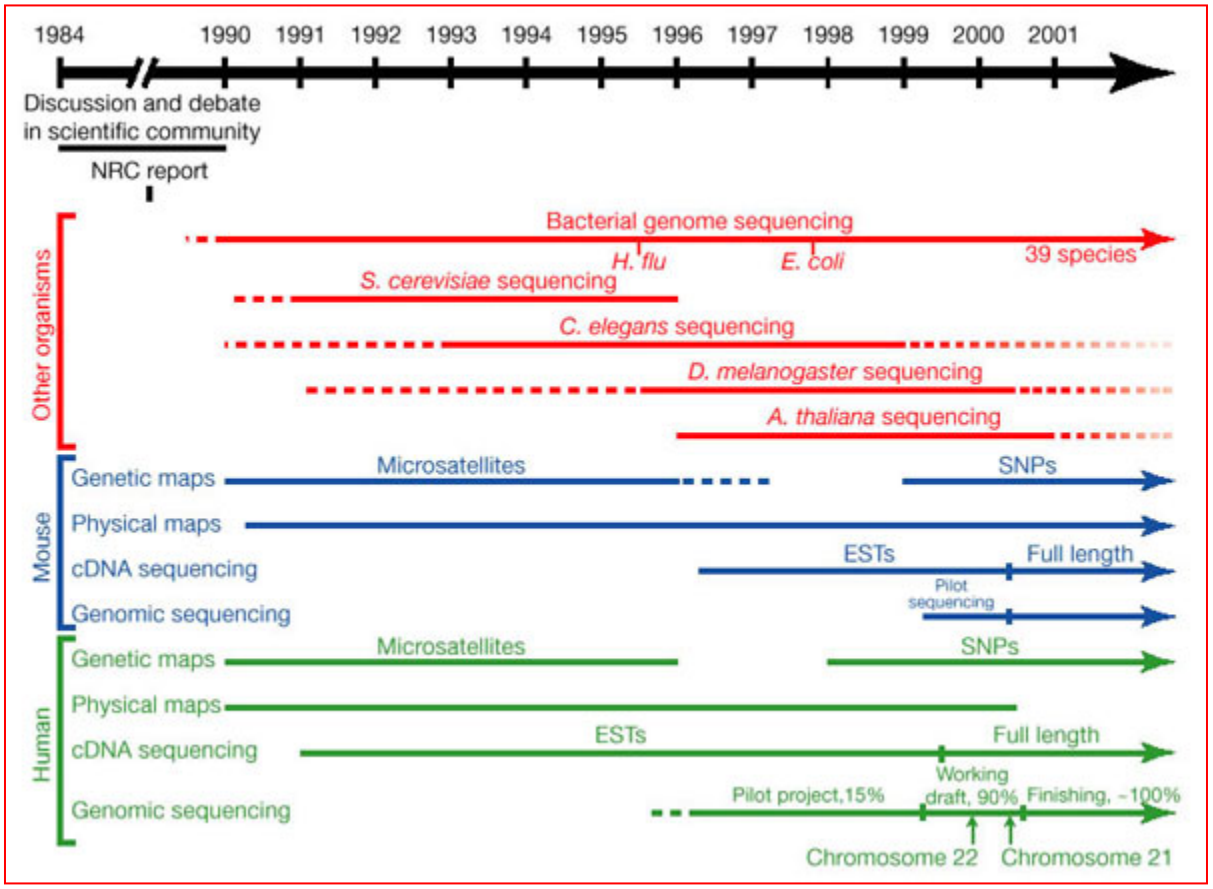
Il Progetto Genoma Umano

Per la sua dimensione,

il sequenziamento del primo genoma umano ha richiesto uno **sforzo a livello internazionale enorme** ed un ingente investimento di risorse economiche e finanziarie.

- Human 3.0×10^9
- Mouse 3.0×10^9
- Drosophila 1.1×10^8
- Nematodes 1.0×10^8
- Dictyostellium 3.4×10^7
- Yeast 1.2×10^7
- Bacteria 5.0×10^6
- Virus $10^3 .. 10^5$

Il Progetto Genoma Umano: Timeline



Il progetto prevedeva anche il sequenziamento del genoma di altri organismi eucarioti e procarioti

Genoma Umano: il release del 2004

NCBI Assembly 35
July 2004

	Length size	Known genes	Novel genes	miRNA	rRNA	Misc RNA
chromosome 1	245.522.847	2062	140	40	52	92
chromosome 2	243.018.229	1295	164	14	24	73
chromosome 3	199.505.740	1076	66	19	22	67
chromosome 4	191.411.218	762	90	9	13	58
chromosome 5	180.857.866	867	86	14	22	70
chromosome 6	170.975.699	1070	69	8	17	57
chromosome 7	158.628.139	957	133	24	14	63
chromosome 8	146.274.826	710	72	11	14	39
chromosome 9	138.429.268	802	91	32	10	48
chromosome 10	135.413.628	777	79	4	17	45
chromosome 11	134.452.384	1294	89	21	19	48
chromosome 12	132.449.811	1022	63	14	15	66
chromosome 13	114.142.980	334	53	16	9	34
chromosome 14	106.368.585	651	60	32	14	39
chromosome 15	100.338.915	610	89	11	6	35
chromosome 16	88.827.254	864	83	5	13	31
chromosome 17	78.774.742	1129	92	26	10	52
chromosome 18	76.117.153	278	35	8	5	21
chromosome 19	63.811.651	1365	83	24	6	18
chromosome 20	62.435.964	583	36	7	8	34
chromosome 21	46.944.323	247	20	8	3	6
chromosome 22	49.554.710	483	59	10	2	20
chromosome X	154.824.264	807	98	32	20	48
chromosome Y	57.701.691	76	28	0	6	2
	3.076.781.887	**20.121**	**1.878**	**389**	**341**	**1.066**
		23.795				

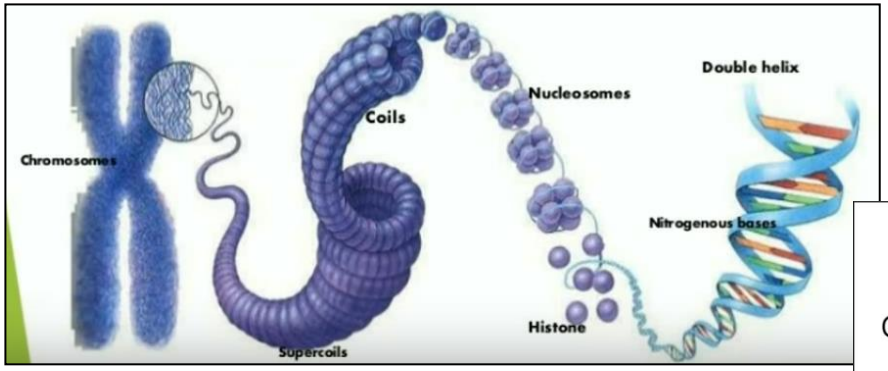
Formalmente chiuso in Aprile 2003, sequenza del genoma umano con un'accuratezza del 99,99%, è stata riportata nel Maggio 2004. Ulteriori analisi continuano ancora oggi.

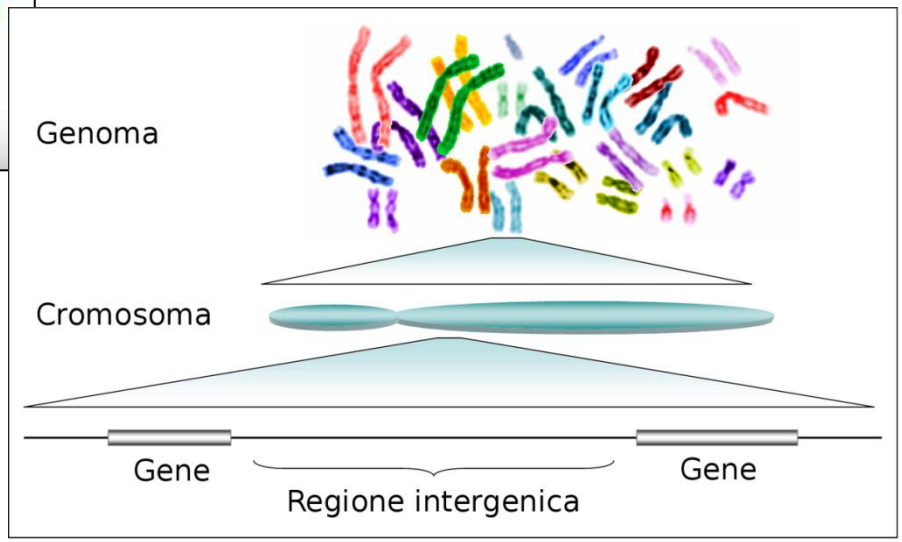
Ancora nel 2019, esistono decine di problematiche di sequenza non risolte.

il Progetto Genoma Umano non ha sequenziato tutto il genoma, ma le porzioni eucromatiche, mentre quelle eterocromatiche relative ai centromeri e telomeri non sono state sequenziate nell'ambito del progetto

Genoma Umano: generalità

- Ogni cromosoma contiene geni che sono trascritti in RNA, i quali a loro volta sono tradotti in proteine o funzionano come RNA (rRNA, tRNA, microRNA ed altri non-coding RNA)





Genoma Umano: generalità

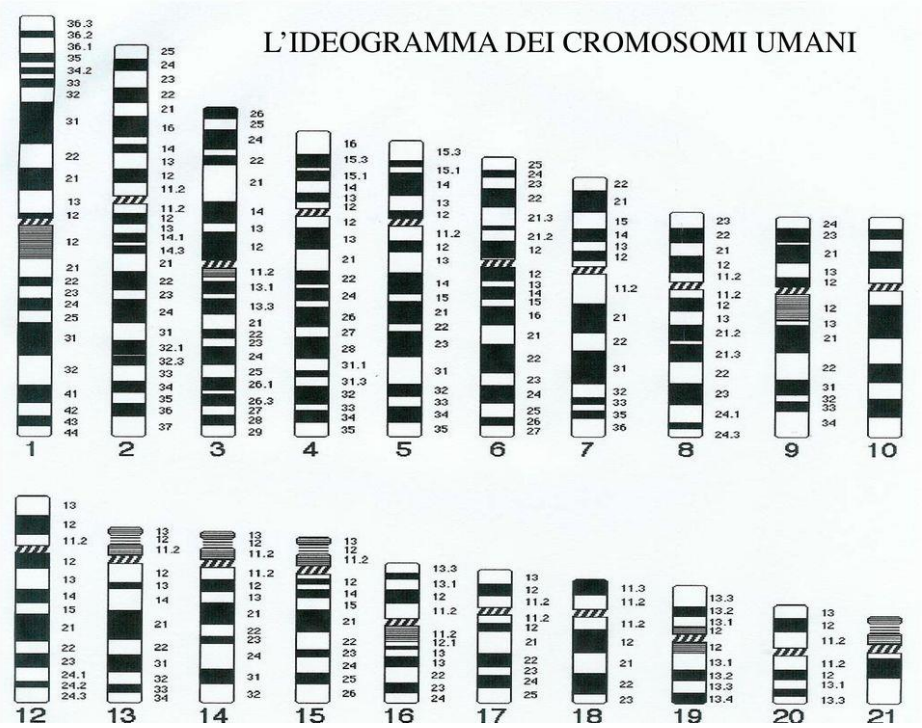
Caratteristiche	Genoma nucleare
Dimensione	3,2 Gb
Numero di differenti molecole di DNA	23 (in cellule XX) o 24 (in cellule XY)
Numero di molecole di DNA per cellula	46 in cellule diploidi, 23 nei gameti (aploidi)
Proteine associate	Proteine istoniche e non istoniche
Numero di geni codificanti per proteine	~ 20.000
Numero di geni per ncRNA	~ 3.500 (considerando solo sequenze validate sperimentalmente); ~ 18.000 se si considerano tutte le sequenze annotate in NCBI Gene
Densità genica	~ 1/80 kb
Numero di pseudogeni	~ 15.000
Sequenze ripetute	~ 53%

G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

- Il numero di geni che codificano proteine nel genoma umano è circa 20,000
- La sequenza dei centromeri e dei telomeri è in gran parte costituita da sequenza ripetute
- Per quanto riguarda i segmenti di DNA compresi tra centromero e telomeri:
- 96.5% sono regioni non codificanti, introni e sequenze intergeniche, di cui gran parte costituite da elementi genetici trasponibili e da DNA derivato da residui fossili di elementi trasponibili

Genoma Umano: cromosomi

- Il genoma umano è costituito da DNA: circa $3x10^9$ coppie di basi di DNA, separate in 22 autosomi e due cromosomi sessuali, X e Y.



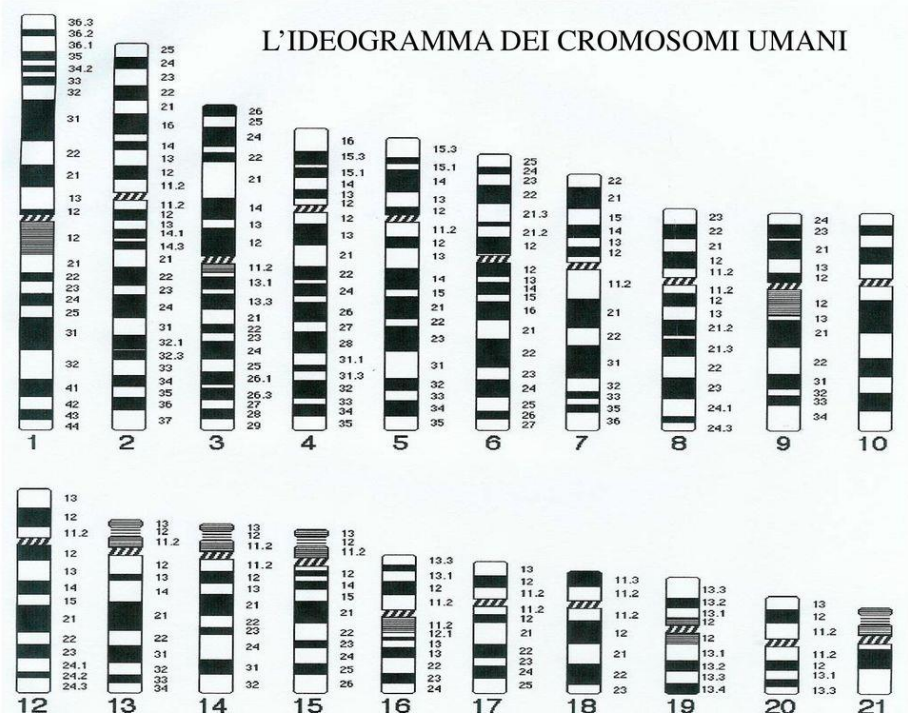
Genoma umano: segmenti specializzati rilevanti

- Ogni cromosoma contiene segmenti specializzati:

  - un centromero
  - due telomeri alle estremità

che servono al mantenimento della struttura ed alla trasmissione corretta da cellula a cellula;

  - numerose origini di replicazione

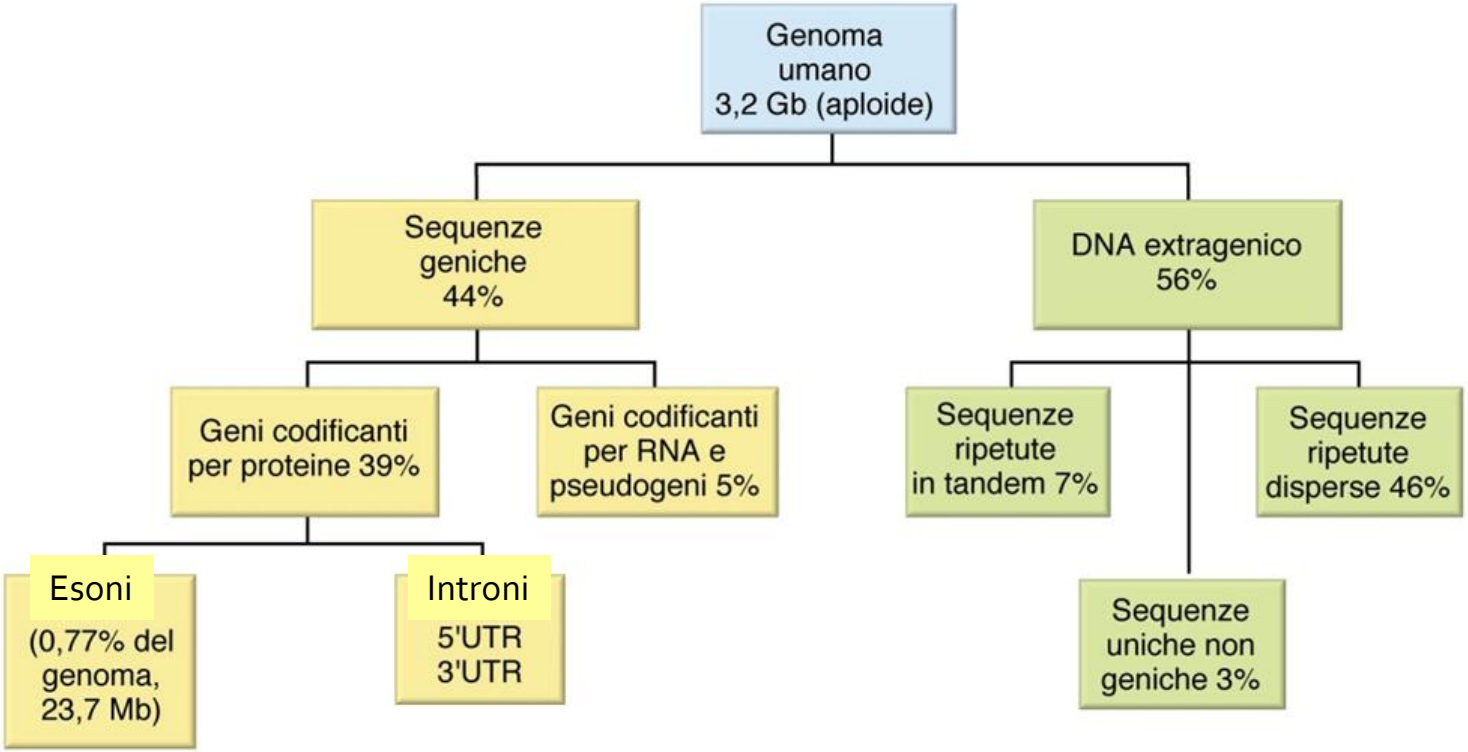


Classi e sottoclassi	Dimensione o sequenza della unità ripetuta	Principali localizzazioni cromosomiche
DNA SATELLITE ("blocco" > 100 kb)		
α satellite	171 bp	Eterocromatina centromerica di tutti i cromosomi
β satellite	68 bp	Eterocromatina centromerica dei cromosomi 1, 9, Y e dei cinque cromosomi acrocentrici (13, 14, 15, 21, 22)
Satellite 1	25-48 bp (sequenza ricca in AT)	Eterocromatina centromerica della maggior parte dei cromosomi e altre regioni eterocromatiniche
Satellite 2	Forme divergenti da ATTCC/GGAAT	Tutti i cromosomi
Satellite 3	ATTCC/GGAAT	Bracci corti dei cinque cromosomi acrocentrici e eterocromatina dei cromosomi 1q, 9q e Yq12
DNA MINISATELLITE ("blocco" 0,1-20 kb)		
Minisatellite telomerico	TTAGGG	Tutti i telomeri
Minisatelliti ipervariabili	9-64 bp	Regioni sub-telomeriche eucromatiniche di tutti i cromosomi
DNA MICROSATELLITE ("blocco" < 100 bp)		
–	1-4 bp	Disperse lungo tutto i cromosomi
Garrido-Ramos MA. Genes (Basel). 2017 Sep; 8(9): 230. Satellite DNA: An Evolving Topic.		

TABELLA 5.4 Principali classi e sottoclassi di sequenze ripetute in tandem.

Organizzazione generale del genoma umano

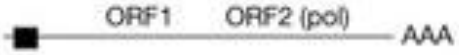
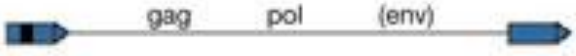
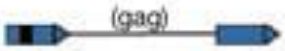
Se consideriamo centromeri, telomeri, sequenze intergeniche ed introni, il genoma umano è in gran parte costituito da DNA non codificante, spesso formato da sequenze ripetute



G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

Sequenze altamente ripetute intersperse nel genoma umano

Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous		6–8 kb	850,000	21%
SINEs	Non-autonomous		100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous		6–11 kb	450,000	8%
	Non-autonomous		1.5–3 kb		
DNA transposon fossils	Autonomous		2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

Gran parte delle sequenze ripetute intersperse nel genoma umano sono costituite da elementi genetici trasponibili o da «rottami» frammentati di elementi trasponibili non più attivi, quale possibile retaggio del processo evolutivo in cui questi elementi hanno avuto ed ancora hanno un ruolo importante

Sequenze altamente ripetute intersperse nel genoma umano

Classe	Famiglia	Dimensioni dell'unità ripetuta	N° copie	% genoma
SINE	*Alu* Altre	Lunghezza completa 0,3 kb Dimensione media 0,13 kb	1.300.000 ca 500.000 ca	10,7% ca 2,5% ca
LINE	LINE-I (*Kpn*) Altre	Lunghezza completa 6,1 kb, ma le dimensioni medie sono 0,8 kb	900.000 ca 370.000 ca	17,3% ca 3,3% ca
LTR	ERV Altre	Dimensione media 1,3 kb	240.000 ca 280.000 ca	4,7% ca 3,8% ca
Trasposoni a DNA	MER-I (Charlie) Altre	Dimensione media 0,25 kb ca	213.000 ca 130.000 ca	1,5% ca 1,4% ca

Sequenze altamente ripetute del genoma umano: elementi ALU

Alu (SINE: short interspersed elements)

http://alugene.tau.ac.il/

GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAG
GTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCC
GGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCG
GGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCC
GTCTCAAAAAAAA

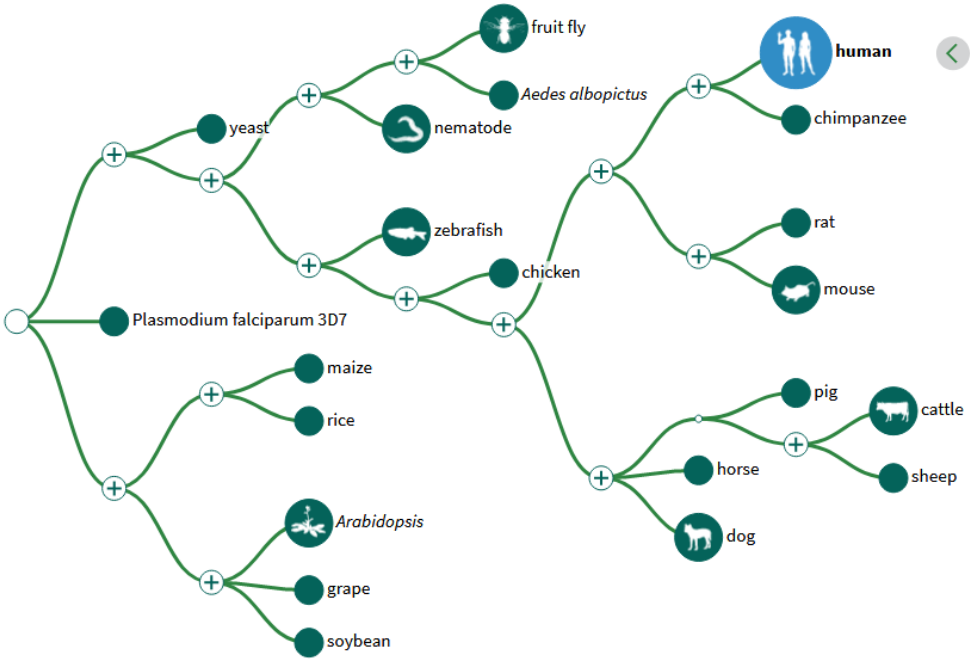
There are about 1.5 million of 300 bp Alu sequences interspersed throughout the human genome, and it is estimated that about 13% of the human genome consists of Alu sequences.

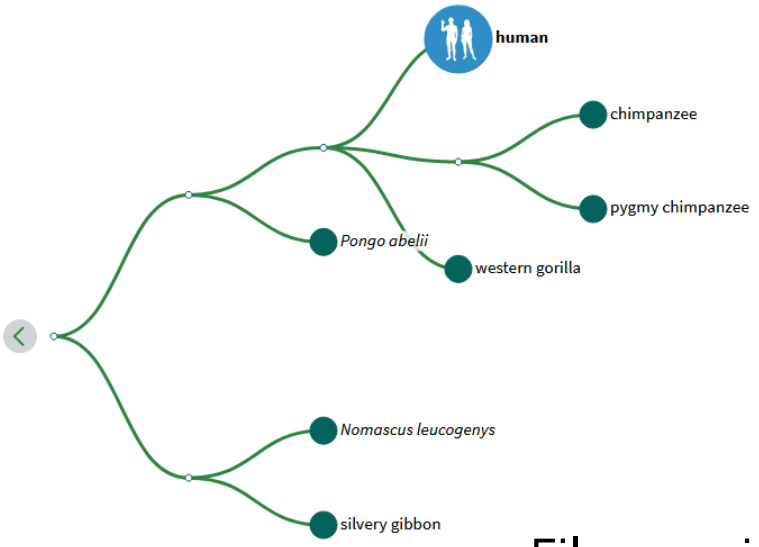
Alu elements are found exclusively in primate species.

Most human Alu sequence insertions can be found in the corresponding positions in the genomes of other primates, but about 2,000 Alu insertions are unique to humans

Genomica comparativa: l'albero filogenetico dei genomi

La filogenesi genomica riflette la già descritta distanza tra gli organismi viventi



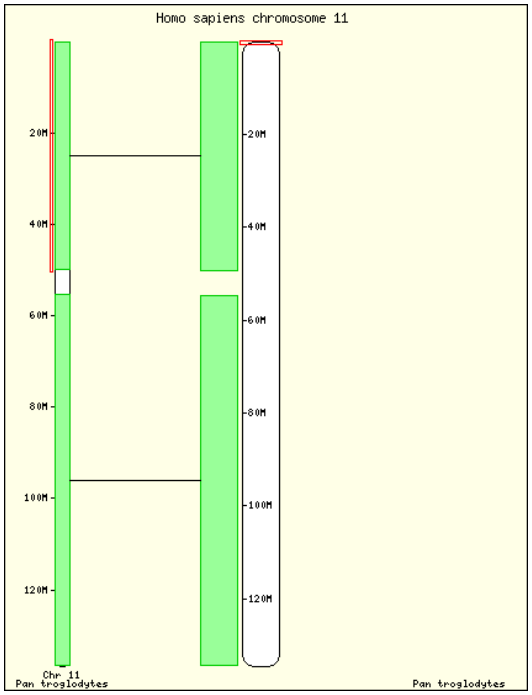


Filogenesi genomica tra i primati

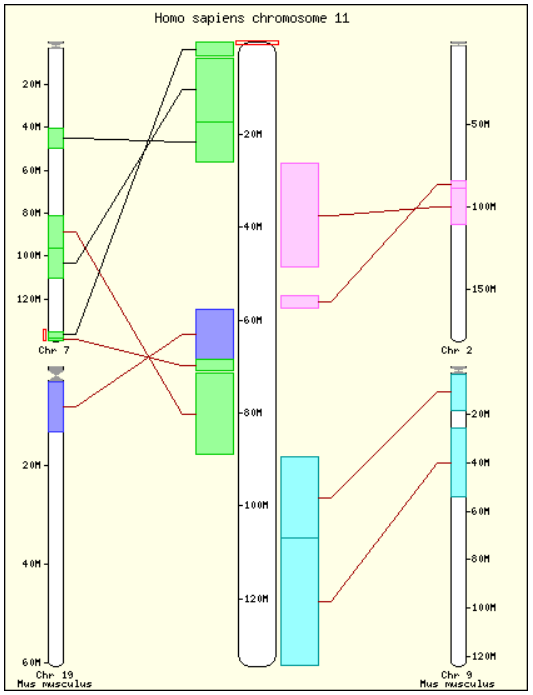
Evoluzione attraverso il «rimpasto» di lunghi segmenti genomici (sintenia)

Lunghi segmenti cromosomici sono sintenici con segmenti cromosomici di altre specie

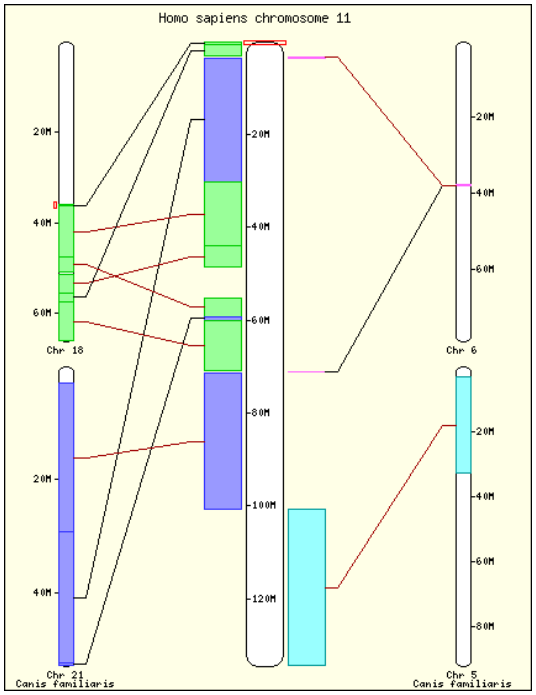
Sintenia uomo-scimpanzè



Sintenia uomo-topo

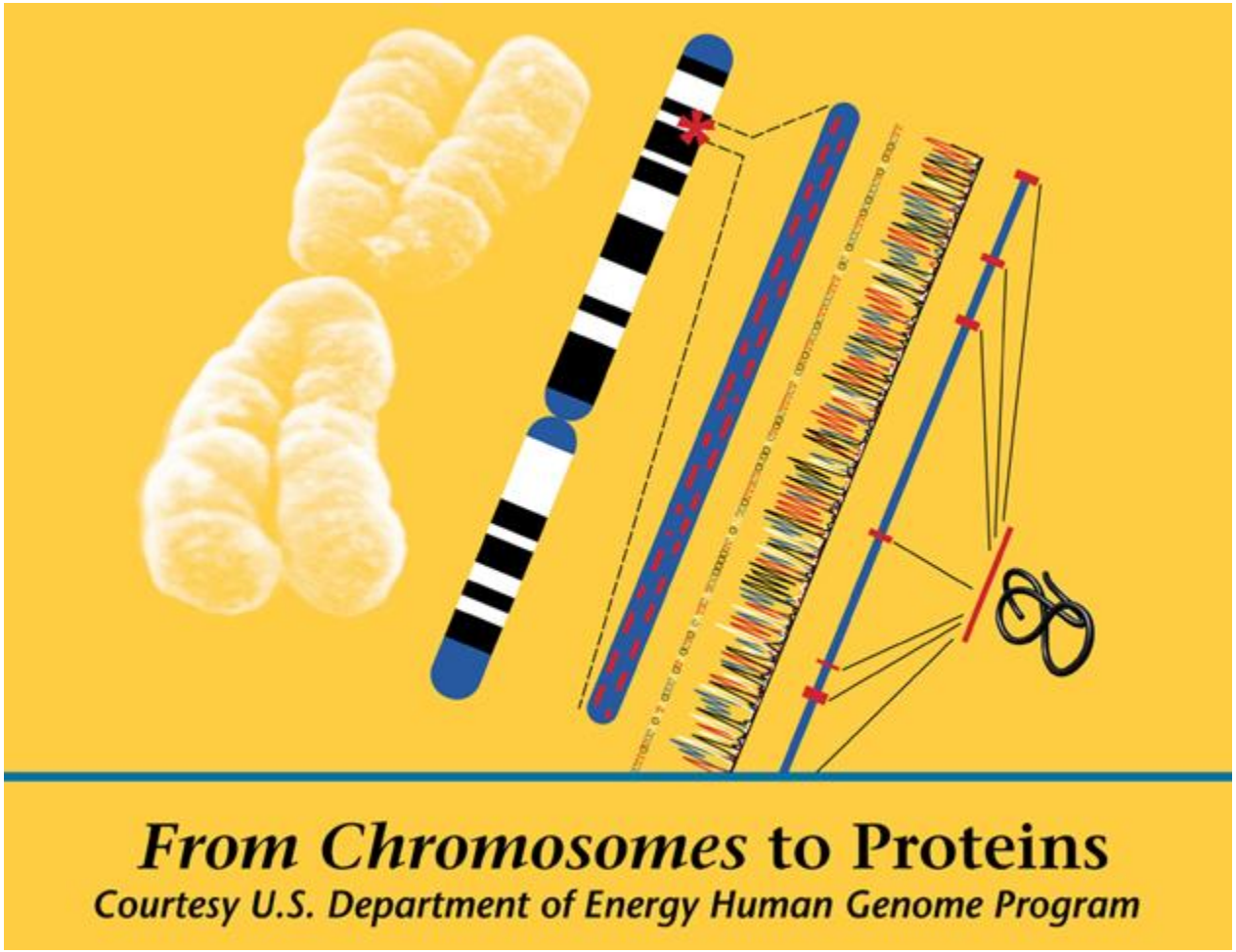


Sintenia uomo-cane



http://www.ensembl.org/Homo_sapiens/index.html

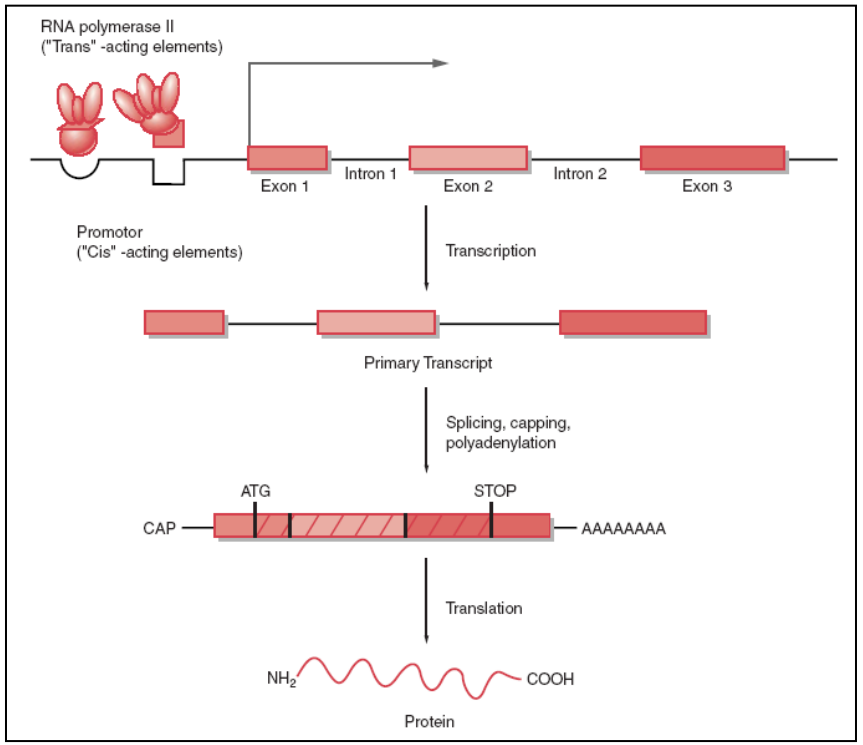
Genoma Umano: la porzione codificante



Geni Codificanti Proteine: Struttura e RNA splicing

- La maggior parte dei geni codificanti proteine sono organizzati in segmenti presenti nel RNA maturo (**esoni**) separati da segmenti assenti nel RNA finale (**introni**).

- Gli esoni sono riuniti nel RNA maturo attraverso un meccanismo di ***splicing***

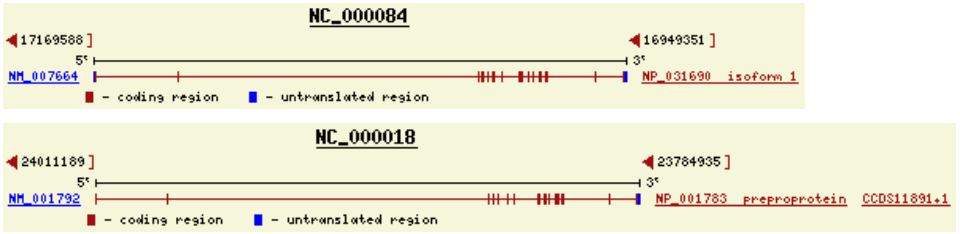


I geni umani presentano struttura esonica filogeneticamente conservata

La struttura genica e l'omologia di sequenza è conservata tra specie

Mus musculus

Homo sapiens

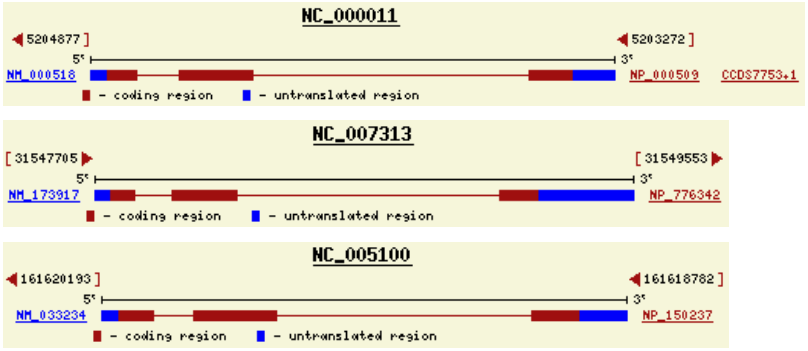


CDH2 – Cadherin 2

Homo sapiens

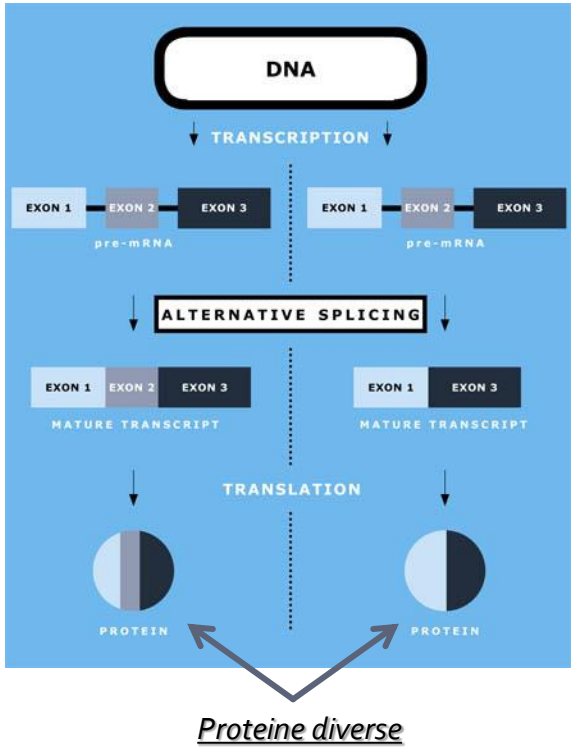
Bos taurus

Rattus norvegicus



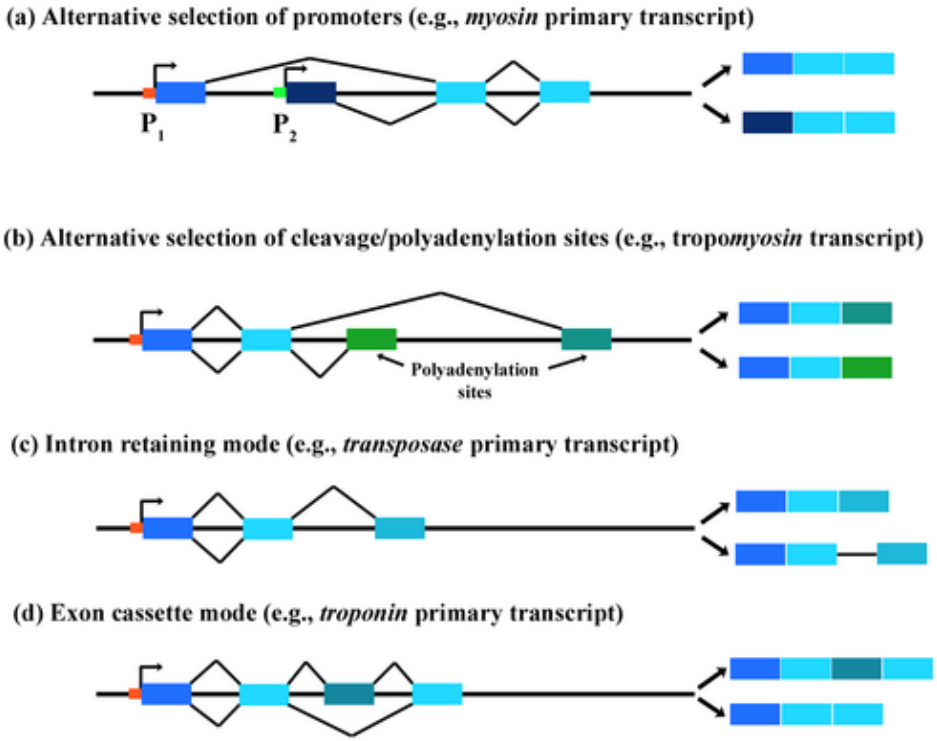
HBB - Beta globin gene

Splicing alternativo



Proteine diverse

Variazioni sul tema

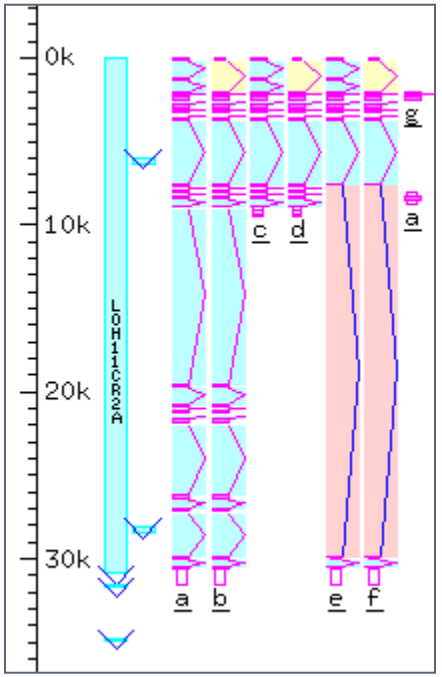


Multipli trascritti e proteine possono essere associate a ciascun gene

Un gene → Multipli trascritti e proteine

- La maggior parte dei geni codifichi per una proteina prevalente.
- Tuttavia, la maggior parte dei geni presentano, a causa dello splicing alternativo, trascritti che codificano per proteine parzialmente diverse.
- L'abbondanza relativa di questi trascritti può modificare la funzione associata al gene. Gran parte di questi aspetti sono oggetto di studio.

Assembly:	NCBI 35, July 2004
Genebuild:	12-Dec-05
Database version:	36.35i
Known genes:	23,071
Novel genes:	1,878
Pseudogenes:	1,947
RNA genes:	1,015
Genscan gene predictions:	68,101
Gene exons:	245,231
Gene transcripts:	52,097
Base Pairs:	3,272,187,692



LOH11CR2A

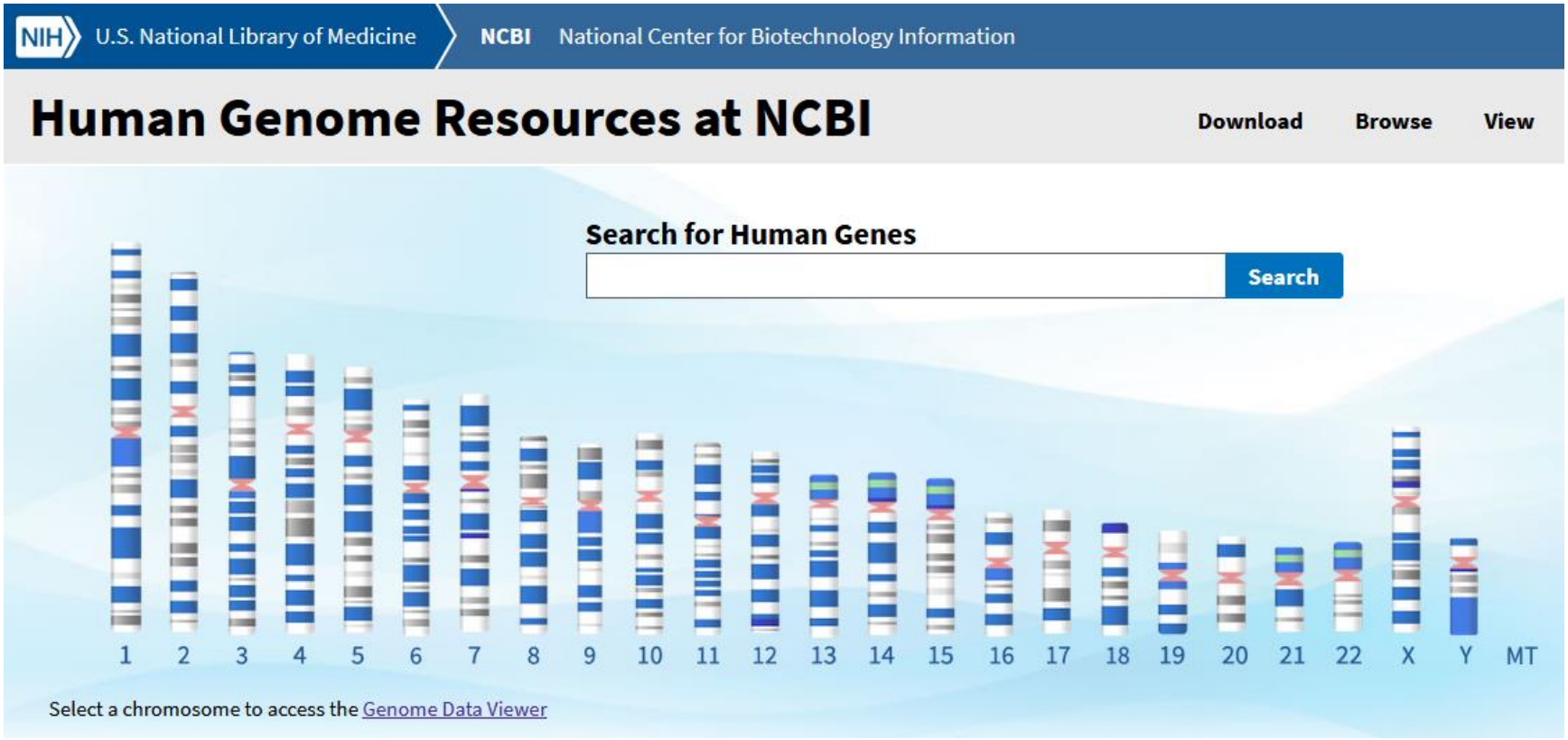
Siti web per conoscere organizzazione del genoma umano

Attività	Risorse	Indirizzo web
Annotazioni di sequenze	GENCODE NCBI Annotation	https://www.gencodegenes.org/ https://www.ncbi.nlm.nih.gov/genome/
Sequenze nucleotidiche e amminoacidiche di riferimento	RefSeq RefSeqGene	https://www.ncbi.nlm.nih.gov/refseq/ https://www.ncbi.nlm.nih.gov/refseq/rsg/
Analisi di sequenze proteiche	UniProt InterPro	https://www.uniprot.org/ https://www.ebi.ac.uk/interpro/
Genome browsers	Ensembl NCBI Genome data Viewer UCSC Genome Browser	http://www.ensembl.org https://www.ncbi.nlm.nih.gov/genome/gdv/ https://genome.ucsc.edu/
Confronto tra sequenze	BLAST BLAT	https://blast.ncbi.nlm.nih.gov/Blast.cgi https://genome.ucsc.edu/cgi-bin/hgBlat?command=start

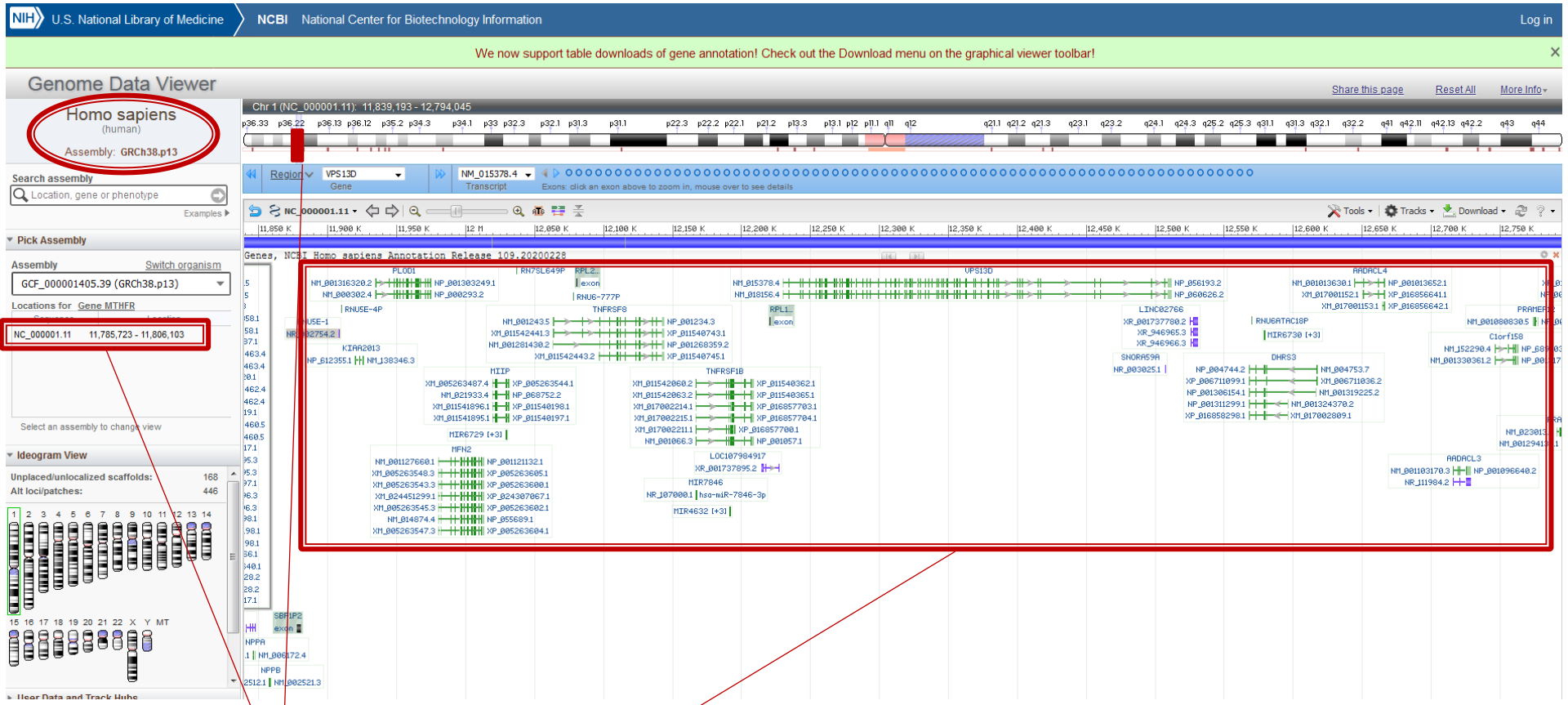
TABELLA I.5.1.1 Banche dati e strumenti bioinformatici.

G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

Human Genome Resources @ NCBI



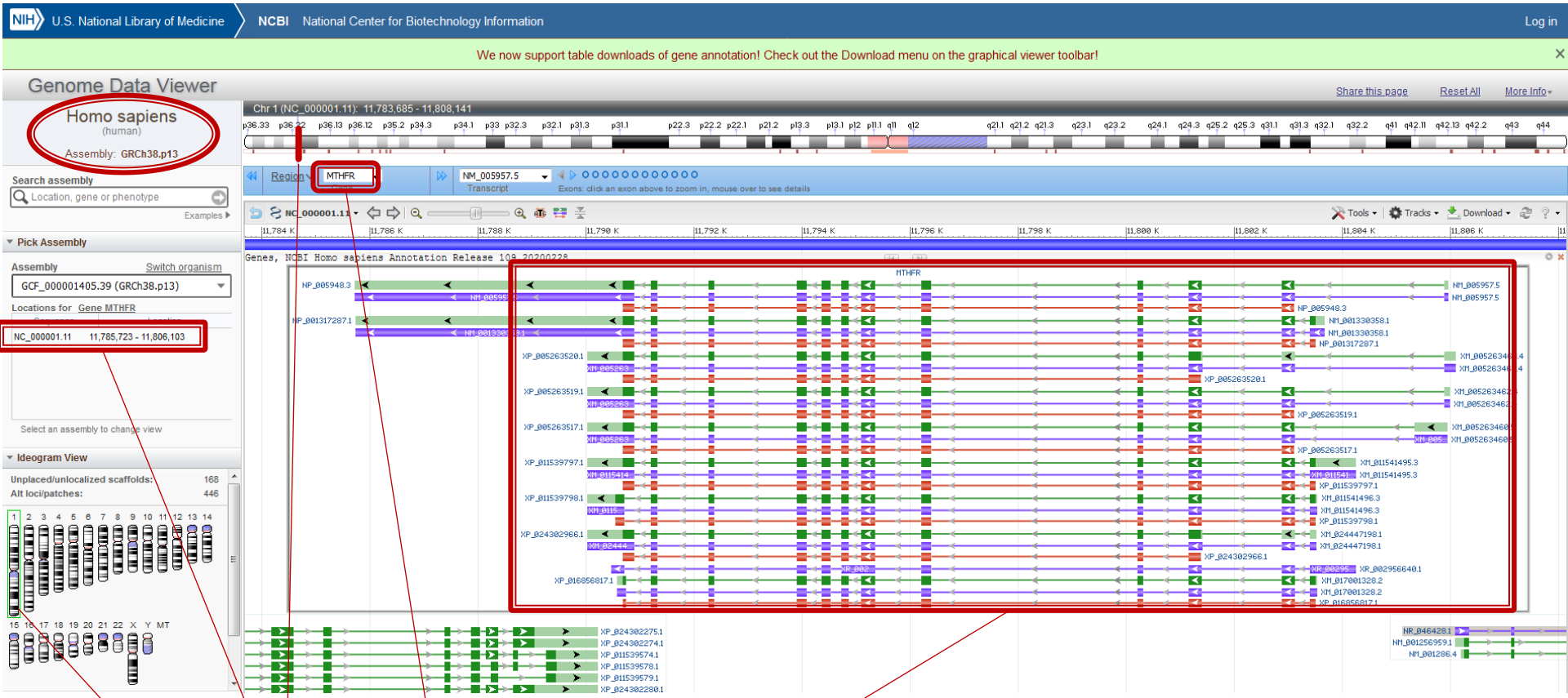
Genoma umano: *Data viewer*



Regione
Cromosoma

Trascritti della
Regione Cromosomica

Genoma umano: *Data viewer*



Cromosoma

Regione Cromosoma

Gene

Trascritti e regioni codificanti del Gene

I geni codificanti per proteine sono la porzione decodificata del genoma

- La porzione del genoma umano codificante per proteine (esoni) rappresenta circa il 2% di tutto il genoma
- La funzione del resto del genoma è in gran parte da decifrare

- Una porzione funzionalmente importante è costituita da geni per RNA non codificanti, la cui funzione è solo in minima parte chiarita
- È plausibile che dalle analisi di questa porzione del genoma emergeranno aspetti inattesi legati a funzioni e disfunzioni riconducibili alla genetica umana.

Classi di ncRNA	Principali famiglie di ncRNA	Funzioni
RNA ribosomale (rRNA)	rRNA 28S, 18S, 5,8S, 5S	Traduzione
RNA transfer (tRNA)	48 tRNA con diversi anticodoni	Traduzione
Altri ncRNA		
Corti ncRNA	snRNA (small nuclear RNA)	Splicing
	snoRNA (small nucleolar RNA)	Maturazione degli rRNA
	scaRNA (small Cajal body RNA)	Maturazione dei snRNA
	miRNA (microRNA)	Regolazione della espressione genica (regolazione post-traduzionale sequenza-specifica)
	piRNA (piWi protein-interacting RNA)	Limitano la possibilità di mobilizzazione dei trasposoni nelle linee germinali
	siRNA (small interfering RNA)	Limitano la possibilità di mobilizzazione dei trasposoni nelle linee germinali
Lunghi ncRNA	lncRNA (long non-coding RNA) nucleari	Regolazione della espressione genica (regolazione pre-trascrizionale, in particolare controllo della conformazione della cromatina; regolazione trascrizionale)
	lncRNA citoplasmatici	Regolazione dell'espressione genica (regolazione traduzionale, produzione della proteina; regolazione post-traduzionale, controllo della localizzazione della proteina o della sua attività)

Hombach S. et al. Adv Exp Med Biol. 2016; 937: 3-17. doi: 10.1007/978-3-319-42059-2_1. Non-coding RNAs: Classification, Biology and Functioning.

TABELLA I.5.3.1 **Descrizione delle principali classi di ncRNA presenti nelle cellule umane.**

G. De Leo, S. Fasano, E. Ginelli
Biologia e Genetica, IV ed.
EdiSES Università

Genoma Umano nel 2018

Chromosome	Base pairs	Protein-coding genes	Pseudo-genes	long ncRNA	small ncRNA	miRNA	rRNA	snRNA	snoRNA	Misc ncRNA
1	248.956.422	2.058	1.220	1.200	496	134	66	221	145	192
2	242.193.529	1.309	1.023	1.037	375	115	40	161	117	176
3	198.295.559	1.078	763	711	298	99	29	138	87	134
4	190.214.555	752	727	657	228	92	24	120	56	104
5	181.538.259	876	721	844	235	83	25	106	61	119
6	170.805.979	1.048	801	639	234	81	26	111	73	105
7	159.345.973	989	885	605	208	90	24	90	76	143
8	145.138.636	677	613	735	214	80	28	86	52	82
9	138.394.717	786	661	491	190	69	19	66	51	96
10	133.797.422	733	568	579	204	64	32	87	56	89
11	135.086.622	1.298	821	710	233	63	24	74	76	97
12	133.275.309	1.034	617	848	227	72	27	106	62	115
13	114.364.328	327	372	397	104	42	16	45	34	75
14	107.043.718	830	523	533	239	92	10	65	97	79
15	101.991.189	613	510	639	250	78	13	63	136	93
16	90.338.345	873	465	799	187	52	32	53	58	51
17	83.257.441	1.197	531	834	235	61	15	80	71	99
18	80.373.285	270	247	453	109	32	13	51	36	41
19	58.617.616	1.472	512	628	179	110	13	29	31	61
20	64.444.167	544	249	384	131	57	15	46	37	68
21	46.709.983	234	185	305	71	16	5	21	19	24
22	50.818.468	488	324	357	78	31	5	23	23	62
X	156.040.895	842	874	271	258	128	22	85	64	100
Y	57.227.415	71	388	71	30	15	7	17	3	8
mtDNA	16.569	13	0	0	24	0	2	0	0	0
total	3.088.286.401	20.412	14.600	14.727	5.037	1.756	532	1.944	1.521	2.213

Genoma umano: numero di geni inconsistente nei diversi database

- Ancora oggi, il numero esatto di geni nel genoma umano non è del tutto chiarito.
- Ciò riguarda particolarmente geni per RNA non codificanti.
- Invece, il numero di geni codificanti proteine è meglio noto (varia dai 20.376 in Ensembl ai 20.345 in RefSeq).
- Sebbene altri database indichino numeri variabili da 19.901 (Gencode) a 21.306 (CHESS).
- Questa variabilità nei numeri non deve sorprendere, in quanto molti di questi geni sono solo previsti tramite algoritmi informatici e presentano *open reading frame* brevi che possono o meno codificare per proteine funzionali