1

Struttura e funzione dei geni e del genoma umano

P. Chiurazzi, G. Marangi

ACIDI NUCLEICI E INFORMAZIONE GENETICA

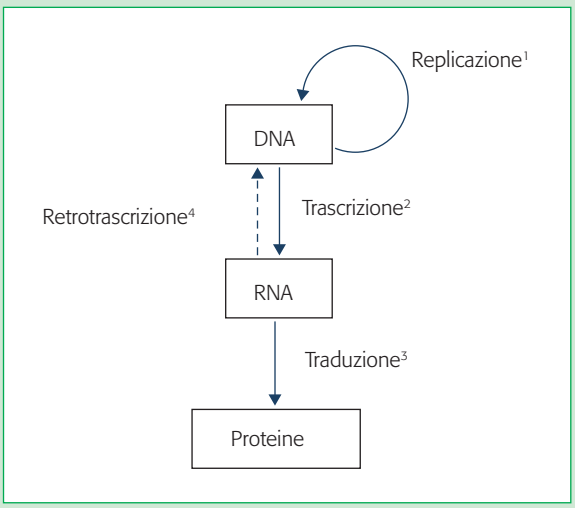
Ogni organismo vivente, unicellulare o multicellulare, è caratterizzato dalla presenza di specifiche molecole di acido nucleico (**DNA** o **RNA**) in grado di produrre una copia identica di sé stesse. Esclusi alcuni virus a RNA, l'informazione genetica degli organismi viventi è contenuta nelle molecole di DNA.
Il patrimonio genetico di un organismo è detto **genoma**. Il genoma umano è organizzato in 46 **cromosomi** lineari, 23 trasmessi con l'ovocita materno e 23 con lo spermatozoo paterno, per un totale di circa 6 (3 x 2) miliardi di desossiribonucleotidi (le unità base del DNA). Inoltre, ogni cellula umana contiene tante piccole molecole di DNA circolare (lunghe circa 16500 nucleotidi) quanti sono i suoi mitocondri. Il DNA genomico deve essere duplicato prima di ogni divisione cellulare per assicurare una copia completa delle informazioni necessarie alla sintesi delle proteine e degli RNA che presiederanno al funzionamento delle cellule figlie (Box 1.1). L'essenza biologica della vita è riconducibile, in ultima analisi, alla continua replicazione e trasmissione degli acidi nucleici da una generazione all'altra di cellule.

Struttura degli acidi nucleici

Gli acidi nucleici sono molecole polimeriche composte da unità dette **nucleotidi** (Fig. 1.1). Nel caso dell'acido desossiribonucleico (DNA), ogni nucleotide è formato da una **base azotata** variabile (adenina, citosina, guanina, timina, abbreviate in A, C, G, T), unita al carbonio 1′ del 2′-desossiribosio (uno **zucchero** a 5 atomi di carbonio o pentoso) a sua volta

BOX 1.1 IL DOGMA CENTRALE DELLA GENETICA MOLECOLARE

La sequenza dei nucleotidi del DNA contiene l'informazione necessaria alla codifica delle proteine di ogni cellula. A seconda delle necessità della cellula, il DNA dei geni viene trascritto in mRNA, che a sua volta dirige la sintesi delle corrispondenti proteine. Il cosiddetto dogma centrale della genetica molecolare è rappresentato dunque dal flusso dell'informazione che dal DNA, attraverso l'RNA, arriva alle proteine.
Nella figura sono indicati con le *frecce continue* i principali processi del flusso di informazione. La replicazione del DNA (1), a opera della DNA polimerasi, avviene prima di ogni divisione cellulare per assicurare il mantenimento dell'informazione genetica in tutte le cellule; la trascrizione del DNA in RNA (2) avviene a opera delle diverse RNA polimerasi e genera diversi tipi di RNA, codificanti e non codificanti; infine, la traduzione (3) degli mRNA consente la sintesi proteica, che avviene sui ribosomi nel citoplasma.
Il dogma centrale è stato spesso ridotto all'equazione «1 gene = 1 mRNA = 1 proteina», semplificando la complessità della sintesi degli RNA e delle proteine.
La *freccia tratteggiata* evidenzia invece un processo non considerato nella formulazione originaria del dogma centrale: la retrotrascrizione (4) dell'RNA in DNA, a opera della trascrittasi inversa, che sta alla base della moltiplicazione degli elementi trasponibili del genoma e della genesi degli pseudogeni processati.



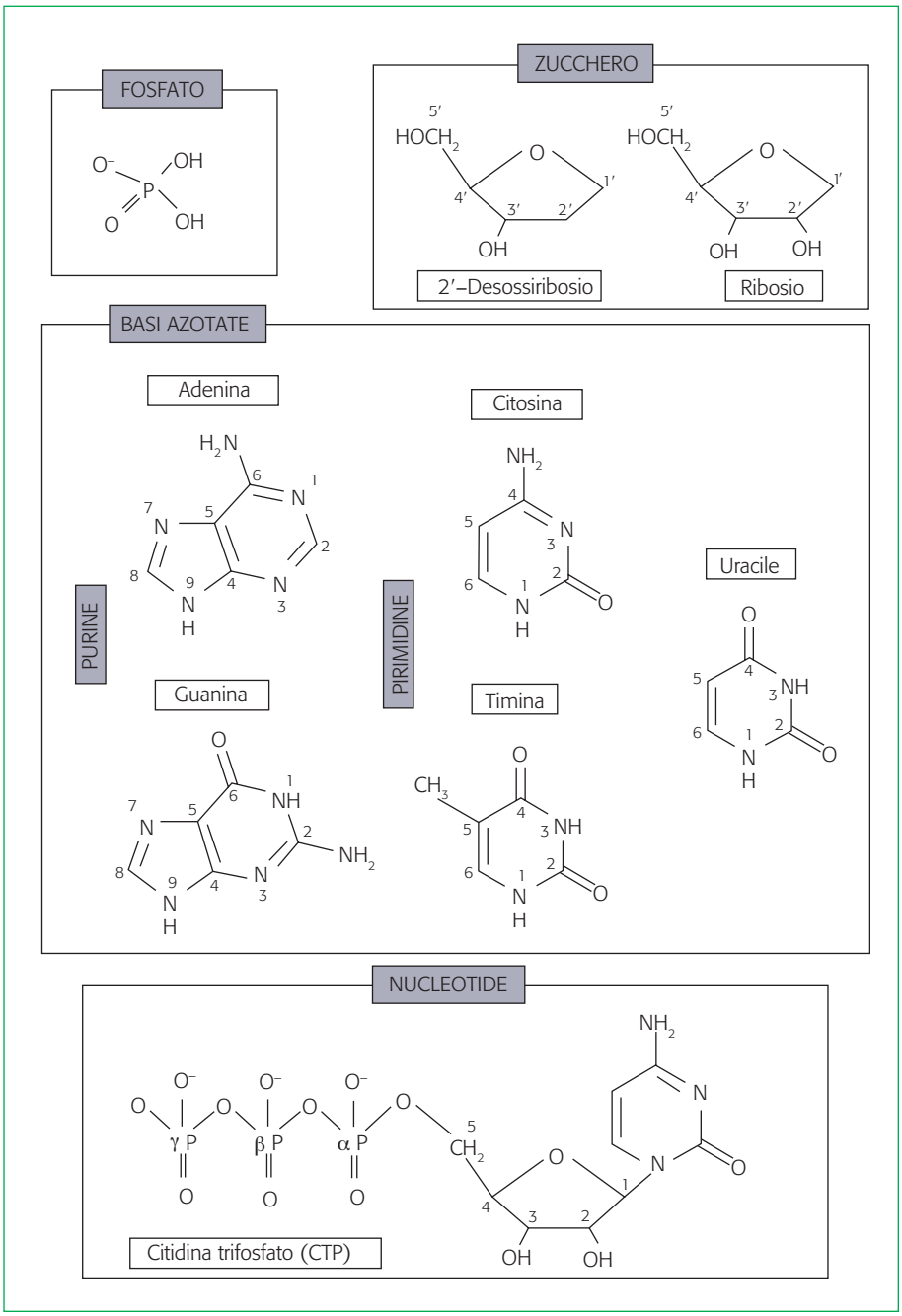


FIG. 1.1 Componenti dei nucleotidi: gruppi fosfato, zuccheri (ribosio e 2′-desossiribosio) e basi azotate (purine e pirimidine).

esterificato con una molecola di **acido fosforico** al carbonio 5′. Ogni nucleotide è unito al precedente mediante il gruppo fosforico che viene esterificato anche al carbonio 3′ del desossiribosio appartenente al nucleotide precedente (Fig. 1.2). Nel caso dell'acido ribonucleico (RNA), lo zucchero pentoso è il ribosio (invece del 2′-desossiribosio) e al posto della timina si trova una base azotata molto simile, l'uracile.

Inoltre, a differenza dell'RNA, il DNA è quasi sempre costituito da due filamenti che si avvolgono uno intorno all'altro, come in una doppia elica, e sono uniti da deboli legami elettrostatici (**legami idrogeno**) fra le basi azotate, dovuti all'appaiamento preciso e complementare dell'adenina con la timina e della guanina con la citosina (Fig. 1.3), secondo il modello proposto da Watson e Crick nel 1953. Considerata la struttura a doppia elica del DNA, la lunghezza di una molecola di DNA è espressa in termini di paia di basi (azotate) o *base pair* (bp). Se un frammento di DNA misura 1000 o 1000000 bp, si parlerà rispettivamente di una **kilobase** (**kb**)

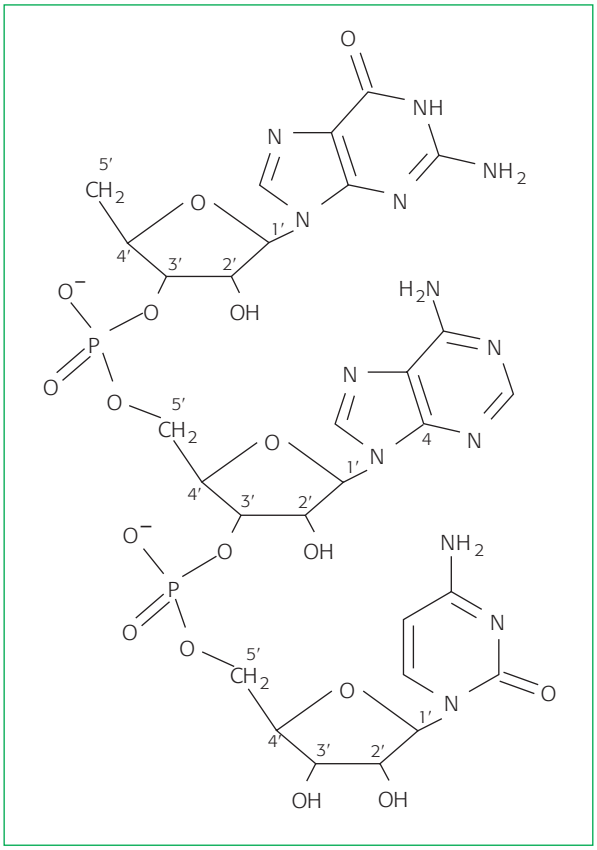


FIG. 1.2 Tre nucleotidi uniti dal gruppo fosforico interposto tra le molecole di ribosio.

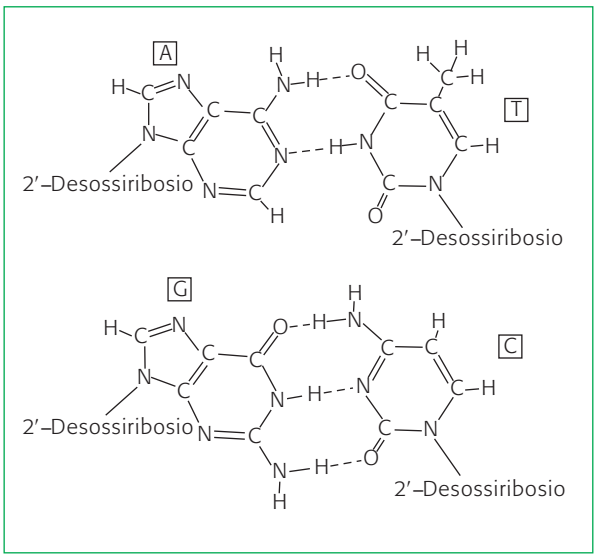


FIG. 1.3 Il riconoscimento dei filamenti complementari degli acidi nucleici: due legami idrogeno fra A e T (U nell'RNA) e tre legami idrogeno fra C e G.

o di una **megabase** (**Mb**). È importante ricordare che i due filamenti del DNA sono **antiparalleli**, ovvero se la catena di acido fosforico-desossiribosio procede nel senso 5′-3′ su un filamento, sarà nel senso 3′-5′ sul filamento complementare (Fig. 1.4). La struttura del DNA schematizzata in Figura 1.4 è quella della più frequente conformazione B, caratterizzata dalla formazione di un'elica destrorsa (che si avvolge in senso orario), anche se in tratti ad alto contenuto di guanina e citosina (G + C) è possibile riscontrare un'elica sinistrorsa (conformazione Z). La doppia elica, nella conformazione B, ha un diametro di circa 2 nm (10^{-9} m) e la distanza fra una coppia di basi e la successiva è di circa 3,4 Å (10^{-10} m).

Il DNA all'interno delle cellule è estremamente compattato; infatti, se i 3 miliardi di nucleotidi del genoma umano fossero distesi linearmente avrebbero una lunghezza di circa 1 m. Il compattamento del DNA si realizza mediante una stretta associazione a proteine basiche dette **istoni**, organizzate in ottameri (costituiti da 2 copie degli istoni H2A, H2B, H3 e H4). Il **nucleosoma** rappresenta l'unità base della **cromatina**, ovvero l'insieme del DNA e delle proteine a esso associate, è largo 11 nm ed è costituito da 146 bp avvolte per quasi due giri intorno a un ottamero di proteine istoniche (Fig. 1.5). Un quinto istone (H1) blocca invece la doppia elica del DNA come un "fermaglio" sull'ottamero istonico.

Nella maggior parte del genoma la distanza media fra un nucleosoma e l'altro è di 200 bp, considerando le 146 bp avvolte al nucleosoma e circa 50 bp di DNA interposto.

Al microscopio elettronico è inoltre possibile visualizzare una **fibra cromatinica** di circa 30 nm di spessore (detta *fibra "a solenoide"*), costituita da 6 nucleosomi che si succedono in modo elicoidale per ogni giro.

Questi due gradi di avvolgimento del DNA consentono di ridurre di circa 40-50 volte la lunghezza totale del genoma, riducendola a 2-3 cm. È evidente che sono necessari ulteriori gradi di compattamento della cromatina per consentire all'intero genoma di trovare spazio all'interno dei pochi micrometri di diametro del nucleo cellulare. Infatti, fra una divisione cellulare e l'altra (**interfase**), la cromatina appare come un gomitolo più o meno compatto. Si distinguono due tipi di cromatina: l'**eterocromatina**, che rimane altamente condensata durante la maggior parte del ciclo cellulare ed è trascrizionalmente inattiva, e l'**eucromatina**, che è meno compatta e consente sia la trascrizione del DNA in RNA sia la duplicazione del DNA durante la fase S. Piccole modificazioni covalenti degli istoni, come l'aggiunta di gruppi metile, acetile o fosfato, correlano con lo stato più o meno attivo della cromatina (Box 1.2). Durante la divisione cellulare (**mitosi**), il grado di compattamento della cromatina è massimo e i singoli cromosomi si rendono finalmente visibili (Capitolo 4).

Replicazione del DNA

La complementarità dei filamenti del DNA suggerisce anche il meccanismo di sintesi (**replicazione**). Gli stessi Watson e Crick hanno fin dal 1953 ipotizzato che ciascun filamento di DNA funzioni da stampo per la sintesi di filamenti complementari. Da una doppia elica del DNA si generano quindi due doppie eliche, ciascuna contenente un filamento che

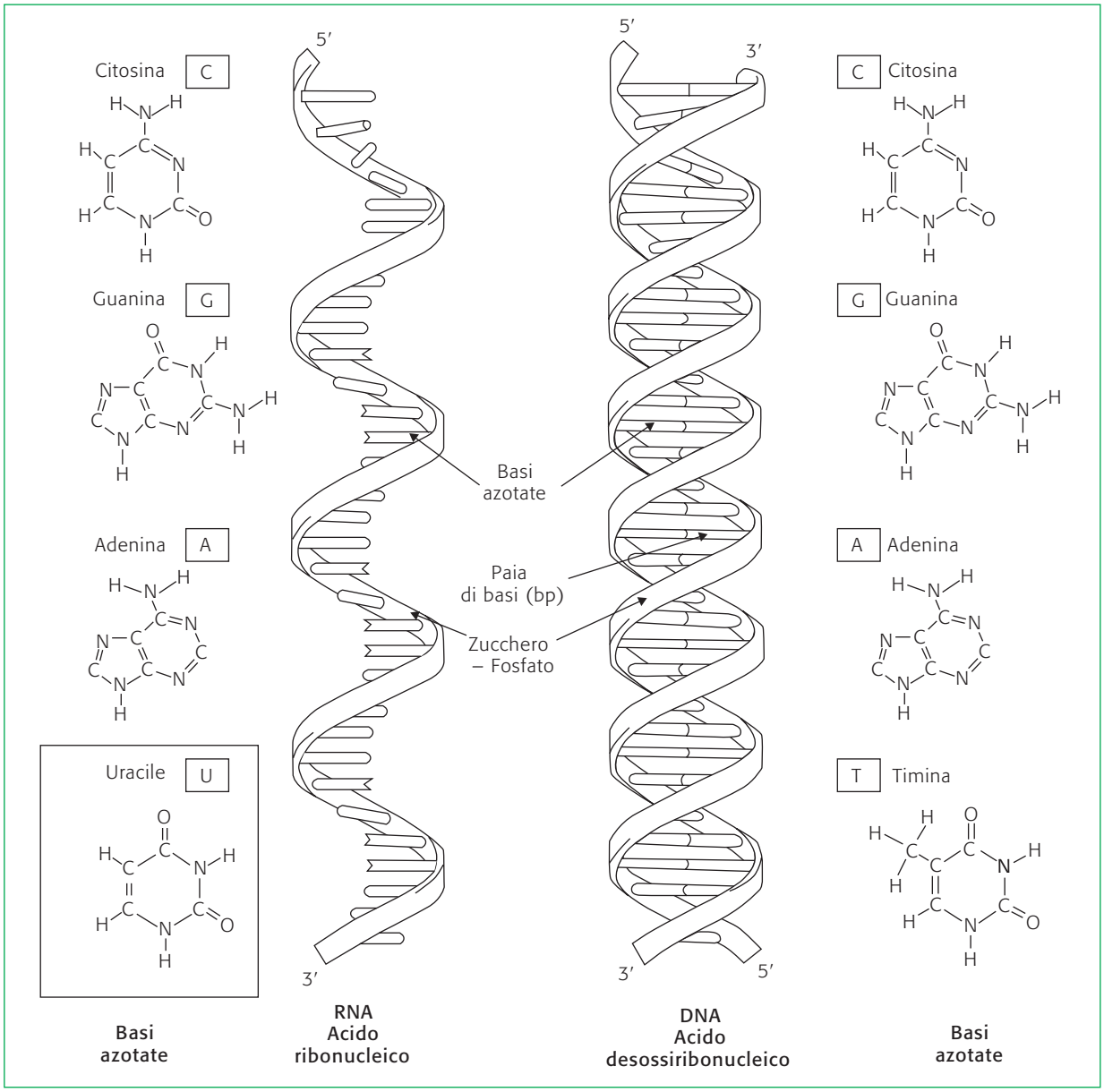


FIG. 1.4 Struttura del DNA e dell'RNA. Entrambe le molecole sono polimeri di nucleotidi, uniti dal legame fra gruppi fosfato e gli zuccheri pentosi 2′-desossiribosio (DNA) o ribosio (RNA). L'RNA è normalmente presente come singolo filamento, mentre il DNA è formato da due filamenti antiparalleli e complementari, appaiati dal riconoscimento di adenina e timina (A:T) e di guanina e citosina (G:C). La doppia elica del DNA compie un giro in senso orario ogni 10 bp (corrispondenti a 34 Å o 3,4 nm) nella conformazione B, come mostrato nella figura.

apparteneva alla molecola parentale e un filamento neosintetizzato. Per questo motivo, il processo replicativo è detto **semiconservativo**.

La replicazione del DNA inizia in punti specifici, detti **origini di replicazione**, nei quali i due filamenti della molecola di DNA parentale vengono separati da enzimi, detti **elicasi**, formando una **bolla di replicazione**. L'enzima **DNA polimerasi** catalizza la sintesi dei nuovi filamenti di DNA utilizzando come precursori nucleotidici i quattro desossinucleotidi trifosfati (dATP, dCTP, dGTP e dTTP). Poiché i due filamenti del DNA parentale sono antiparalleli, ne consegue che i due nuovi filamenti vengano sintetizzati in direzioni opposte (Fig. 1.6). Poiché la DNA polimerasi catalizza l'allungamento delle catene esclusivamente in direzione 5′-3′, su un filamento, detto **filamento veloce** (*leading strand*), la sintesi può procedere direttamente in questo senso, mentre sull'altro, detto **filamento lento** (*lagging strand*), la sintesi 5′-3′ procede in direzione opposta a quella in cui si muove la forcella di replicazione. La sintesi del filamento lento avviene perciò copiando una serie progressiva di piccoli frammenti lunghi circa 100-1000 nucleotidi, detti **frammenti di Okazaki**, che vengono successivamente uniti per mezzo dell'enzima **DNA ligasi**. Poiché

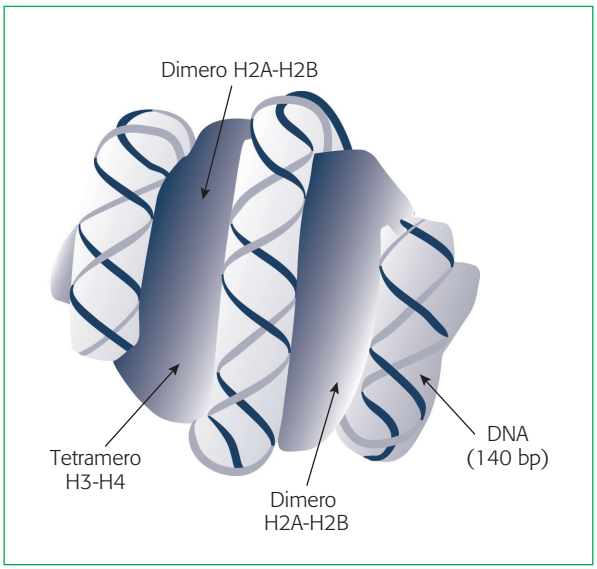


FIG. 1.5 Un nucleosoma, composto da 8 proteine istoniche (doppia copia di H2A, H2B, H3, H4) e circa 140 bp di DNA avvolte agli istoni con doppio giro. L'istone H1 (*non mostrato*) lega invece il DNA che si trova fra un nucleosoma e il successivo.

solo il filamento veloce viene sintetizzato in modo continuo, si dice che la sintesi dei filamenti di DNA è **semidiscontinua**. Dai numerosi punti di origine la replicazione procede in maniera bidirezionale, formando bolle di replicazione che infine si fondono. La completa replicazione del DNA delle cellule umane in coltura richiede mediamente 8 ore.

Le estremità di DNA dei cromosomi (**telomeri**) richiedono un meccanismo diverso di replicazione perché manca lo stampo per l'allungamento del filamento lento. La struttura dei telomeri eucariotici è particolare in quanto consiste in una lunga sequenza di unità ripetute. Nell'uomo la sequenza ripetuta è l'esanucleotide TTAGGG. Queste ripetizioni, oltre a permettere la replicazione, hanno anche la funzione di proteggere le estremità dei cromosomi dalla degradazione. Un enzima specifico, la **telomerasi**, utilizzando un RNA stampo, estende il filamento veloce, che a sua volta funziona successivamente da stampo per la sintesi del filamento lento. La telomerasi è particolarmente attiva nelle cellule germinali, ma lo è molto meno nella maggior parte dei tessuti somatici, determinando così l'accorciamento dei telomeri con i cicli successivi di divisione cellulare.

BOX 1.2 MODIFICAZIONI EPIGENETICHE

La sequenza del DNA è il primo ma non l'unico determinante del flusso di informazione che modula le attività cellulari. Infatti, le modificazioni "epigenetiche" sono fondamentali per regolare la trascrizione del DNA e quindi i livelli di RNA necessari a seconda del momento e del tipo cellulare. Le modificazioni epigenetiche sono reversibili e consistono nell'aggiunta (e rimozione) di gruppi metile, acetile, fosfato agli istoni (Turner BM. Cellular Memory and the Histone Code. Cell 2002;111:285-91) o nella metilazione (e demetilazione) di alcune citosine nel DNA. Il 99,98% delle citosine metilate nelle cellule somatiche appartiene a dinucleotidi CG, ovvero è seguito da una guanina, ma nelle cellule staminali embrionali il 25% delle citosine metilate non è seguito da guanina (Lister et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315-22). La tabella elenca alcune delle principali modificazioni epigenetiche della cromatina, classificandole a seconda della loro presenza nella cromatina attiva (eucromatina) o inattiva (eterocromatina). È importante ricordare che le modificazioni epigenetiche vengono aggiunte o eliminate grazie a specifici enzimi (metilasi/demetilasi, acetiltransferasi/deacetilasi, fosfochinasi/fosfatasi ecc.) e che il loro effetto è quasi sempre mediato da proteine dotate di specifici domini in grado di riconoscere gli istoni o il DNA modificati. Per l'importanza delle modificazioni epigenetiche in patologia, si veda il Capitolo 15.

Modificazioni associate a eterocromatina (inattiva)	**Modificazioni associate a eucromatina (attiva)**
Metilazione di citosine (DNA)	Assenza di metilcitosine (DNA)
Deacetilazione delle lisine (K) 5, 8,12 e 16 (istone H4)	Acetilazione delle lisine (K) 5, 8,12 e 16 (istone H4)
Deacetilazione delle lisine (K) 9,14,18 e 23 (istone H3)	Acetilazione delle lisine (K) 9,14,18 e 23 (istone H3)
Demetilazione della lisina (K) 4 (istone H3)	Metilazione della lisina (K) 4 (istone H3)
Demetilazione della lisina (K) 36 (istone H3)	Metilazione della lisina (K) 36 (istone H3)
Metilazione della lisina (K) 79 (istone H3)	Demetilazione della lisina (K) 79 (istone H3)
Metilazione della lisina (K) 9 (istone H3)	Demetilazione della lisina (K) 9 (istone H3)
Metilazione della lisina (K) 27 (istone H3)	Demetilazione della lisina (K) 27 (istone H3)
Deubiquitinazione della lisina (K) 119 (istone H2A)	Ubiquitinazione della lisina (K) 119 (istone H2A)

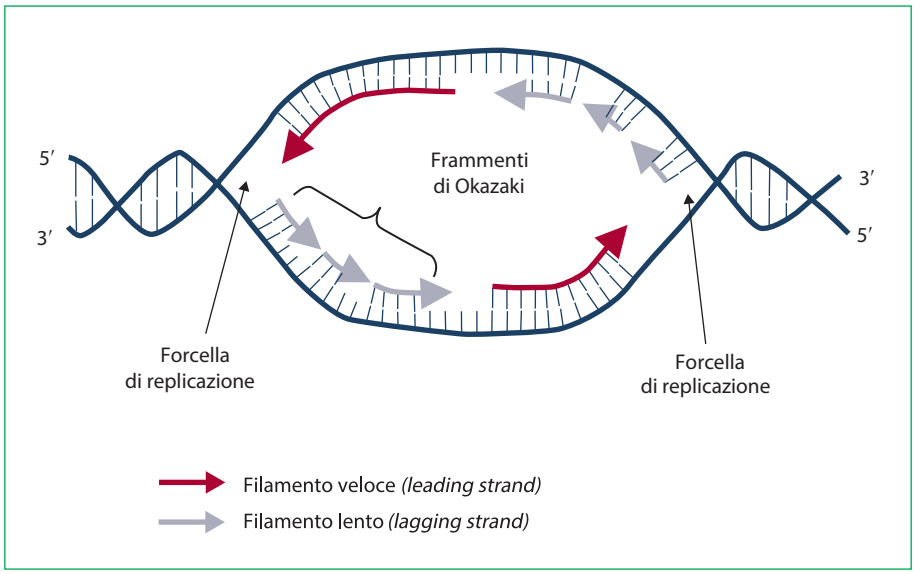


FIG. 1.6 Sintesi del DNA: il filamento veloce (*leading strand*) può essere sintetizzato in modo continuo dato che la doppia elica del DNA viene svolta davanti a esso dalle elicasi (*frecce rosse*). Il filamento lento (*lagging strand*) deve essere sintetizzato in modo discontinuo (sempre in direzione 5′-3′) a partire da piccoli primer di RNA ed è composto da frammenti detti di Okazaki (*frecce grigie*), successivamente uniti dalla DNA ligasi.

Trascrizione e maturazione degli RNA

L'informazione genetica contenuta nelle sequenze del DNA viene trasferita all'RNA ed eventualmente al polipeptide (proteina) corrispondente. Il flusso dell'informazione genetica è pertanto unidirezionale (DNA → RNA → proteine). Questo processo è anche noto come **dogma centrale** della biologia molecolare (si veda Box 1.1).

Il trasferimento dell'informazione genetica dal DNA all'RNA è detto **trascrizione**. Durante questo processo un complesso proteico, comprendente l'enzima **RNA polimerasi**, sintetizza le molecole di RNA usando come stampo le sequenze di DNA che costituiscono le **unità di trascrizione**. La RNA polimerasi si lega al sito di inizio della trascrizione richiamata da altre proteine, dette **fattori di trascrizione**. Questi fattori, mediante l'interazione con brevi sequenze di DNA presenti nella regione a monte dell'inizio della trascrizione (**promotore**), servono a posizionare la RNA polimerasi nel sito giusto e a separare i due filamenti del DNA per formare la **bolla di trascrizione**. L'enzima usa come stampo uno dei due filamenti di DNA (filamento **antisenso**) e sintetizza una molecola di RNA (sempre in direzione 5′-3′), catalizzando il legame fosfodiestere tra il gruppo ossidrilico legato al carbonio 3′ del ribonucleotide precedente e il fosfato del successivo ribonucleotide. Il processo continua finché la polimerasi non incontra una sequenza di arresto. A questo punto, si stacca e libera la catena di RNA, mentre la bolla di trascrizione si richiude e il DNA riassume la sua conformazione a doppia elica. L'RNA neosintetizzato ha la sequenza di basi identica a quella di uno dei due filamenti di DNA (il filamento **senso**), anche se la timina è sostituita dall'uracile. Da uno stesso gene possono essere trascritte consecutivamente numerose copie di RNA e il livello di trascrizione dipende da complessi meccanismi (si veda *Regolazione della trascrizione*).

È importante ricordare che le cellule eucariotiche possiedono tre tipi di RNA polimerasi: la **RNA polimerasi I** trascrive i geni degli RNA ribosomiali 18S, 5.8S e 28S (dove S indica per Svedberg ed è una misura di sedimentazione in ultracentrifuga proporzionale alla massa); la **RNA polimerasi II** trascrive i geni che codificano proteine sintetizzando i precursori degli **RNA messaggeri** (mRNA) e anche alcuni piccoli RNA; infine la **RNA polimerasi III** trascrive i geni di tutti gli RNA transfer (tRNA), l'RNA ribosomiale 5S e altri piccoli RNA.

I precursori degli mRNA, sintetizzati dalla RNA polimerasi II (**trascritti primari**), devono subire una serie di modificazioni prima di essere trasferiti nel citoplasma per venire tradotti sui ribosomi. Questo processo di **maturazione** degli mRNA include le seguenti modificazioni:

1) Aggiunta all'estremità 5′ di un **cap** o cappuccio. Al primo nucleotide all'estremità 5′ della molecola di RNA nascente viene rimosso il fosfato terminale e viene aggiunta una molecola di guanosina monofosfato (GMP) metilata in posizione 7′. Il **capping** serve per proteggere il trascritto dall'attacco delle **esonucleasi**, che lo degraderebbero, e per facilitare il trasporto dal nucleo al citoplasma.
2) Rimozione di alcune sequenze che non vengono tradotte (**splicing**). Quasi tutti i geni eucariotici sono divisi in sequenze codificanti, chiamate **esoni**, e sequenze non tradotte, dette **introni**. Questi ultimi vengono rimossi dai trascritti primari mediante il processo di splicing. Gli introni sono quindi sequenze di DNA, situate fra due esoni, le quali sono trascritte ma non tradotte. Salvo rare eccezioni, gli introni cominciano con i nucleotidi GT e terminano con i nucleotidi AG (*regola GT-AG*). Nel processo di splicing si verifica prima il taglio al 5′ dell'introne (sito *donatore*), poi l'estremità libera si ripiega su sé stessa formando una struttura simile a un laccio e legandosi all'interno dell'introne (sito di *branching*), quindi avviene il taglio al 3′ dell'introne (sito *accettore*) e i due esoni si uniscono mentre l'introne va perso. La struttura macromolecolare che promuove e controlla le reazioni dello splicing è detta **spliceosoma**. Questo complesso comprende varie **subunità di piccole particelle ribonucleoproteiche** (snRNP), ciascuna costituita da una o due **molecole di piccoli RNA nucleari** ricche di uridine (U snRNA) e da una serie di proteine specifiche.

3) Aggiunta all'estremità 3′ di una **coda poli-A**. La maggior parte delle unità di trascrizione ha una breve sequenza (AATAAA) che specifica il termine della trascrizione. Circa 15-30 nucleotidi a valle di questo sito, l'RNA neosintetizzato viene tagliato da un'**endonucleasi** e alla molecola di RNA vengono aggiunti circa 200 residui di adenosina monofosfato (AMP). La coda poli-A ha lo scopo di stabilizzare le molecole degli mRNA maturi e di facilitare il loro trasporto dal nucleo al citoplasma. Gli mRNA codificanti gli istoni "canonici" (ovvero H2A, H2B, H3, H4 e H1) fanno eccezione a questa regola e, pur essendo trascritti dalla RNA polimerasi II, sono privi della coda poli-A.

Altre importanti classi di RNA non vengono tradotte in proteina (Tab. 1.1). Tali RNA, fondamentali per la sopravvivenza della cellula, includono: gli **RNA ribosomiali** (rRNA), gli **RNA transfer** (tRNA), i **piccoli RNA nucleari** (snRNA) e **citoplasmatici** (scRNA) e altre molecole di **RNA non codificante** (ncRNA).

Tab. 1.1 Il genoma umano*

Cromosoma	Lunghezza (bp)	Geni codificanti proteine	Densità geni codificanti proteine (per megabase)	Geni codificanti RNA non tradotti	Pseudogeni
1	248.956.422	2.061	8,28	2.263	1.300
2	242.193.529	1.299	5,36	1.914	1.080
3	198.295.559	1.081	5,45	1.393	803
4	190.214.555	757	3,98	1.183	757
5	181.538.259	882	4,86	1.379	741
6	170.805.979	1.051	6,15	1.265	833
7	159.345.973	1.010	6,34	1.189	908
8	145.138.636	701	4,83	1.173	641
9	138.394.717	778	5,62	926	692
10	133.797.422	730	5,46	1.069	599
11	135.086.622	1.317	9,75	1.238	839
12	133.275.309	1.037	7,78	1.353	655
13	114.364.328	322	2,82	711	395
14	107.043.718	821	7,67	969	530
15	101.991.189	617	6,05	1.129	529
16	90.338.345	863	9,55	1.168	509
17	83.257.441	1.186	14,24	1.331	548
18	80.373.285	266	3,31	685	264
19	58.617.616	1.476	25,18	1.001	527
20	64.444.167	546	8,47	681	265
21	46.709.983	221	4,73	447	185
22	50.818.468	495	9,74	595	345
X	156.040.895	859	5,50	717	893
Y	57.227.415	63	1,10	140	397
Totale	**3.088.269.832**	**20.439**	**6,62**	**25.919**	**15.235**
Mitocondrio	16.569	13	784,60	24	-

**Il genoma umano è ripartito in 23 coppie di cromosomi, contenenti un numero variabile di geni. Accanto al numero di ciascun cromosoma sono riportati la lunghezza in paia di basi (bp), il numero di geni codificanti proteine identificati su ciascuno di essi, la densità genica (numero di geni codificanti per proteine per megabase), il numero di geni codificanti RNA non tradotti. L'ultima riga riporta i dati relativi al genoma mitocondriale, piccolo e circolare, presente in numero variabile di copie a seconda dell'abbondanza di mitocondri nella cellula. I dati sono aggiornati a giugno 2023 e tratti dalla banca dati EnsEMBL (si veda Box 1.3).*

Sintesi delle proteine e codice genetico

Gli mRNA maturi escono dal nucleo nel citoplasma, dove dirigono la sintesi delle proteine. Essi si legano ai **ribosomi**, costituiti da due complessi ribonucleoproteici, detti *subunità ribosomiali*. La subunità grande (60S) contiene circa 50 proteine e tre tipi di rRNA: 28S, 5.8S e 5S. La subunità piccola (40S) è composta dall'rRNA 18S e da circa 30 proteine.
La sequenza aminoacidica della proteina nascente è determinata dalla sequenza nucleotidica dell'mRNA grazie al **codice genetico** universale basato su triplette, ovvero gruppi di tre nucleotidi, detti **codoni** (unità codificanti), corrispondenti a singoli aminoacidi (Fig. 1.7). Le molecole di tRNA veicolano gli aminoacidi ai ribosomi; ciascuna molecola di tRNA ha una struttura simile a un trifoglio e lega con l'estremità 3′ uno specifico aminoacido, mentre sul versante opposto presenta una specifica tripletta di basi, detta **anticodone**. L'anticodone è complementare e, quindi, si lega alla tripletta del codone presente sull'mRNA. La sintesi del polipeptide ha inizio con l'interazione dell'anticodone del primo tRNA, al quale è legata la metionina, col codone d'inizio (AUG) sull'mRNA. In particolari condizioni però è stata osservata la traduzione di peptidi che non utilizzano il codone di start canonico (AUG) come nel caso della *Repeat-Associated Non-AUG mediated translation* (si veda Box 16.1). L'aminoacido successivo viene aggiunto mediante un legame peptidico tra il gruppo carbossilico del primo aminoacido e il gruppo aminico del successivo, la cui formazione è catalizzata dalla subunità grande del ribosoma che agisce come un enzima (ribozima). Nello stesso modo vengono incorporati nella catena polipeptidica nascente i successivi aminoacidi, finché si incontra un codone di terminazione (**codone nonsenso** o di **stop**) che non ha un tRNA corrispondente. Dato che in ognuna delle tre posizioni di un codone è possibile che ci siano 4 basi diverse, il codice genetico ha 64 codoni; 3 di questi sono codoni di stop (UAA, UAG, UGA), mentre gli altri 61 specificano i 20 diversi aminoacidi. Poiché

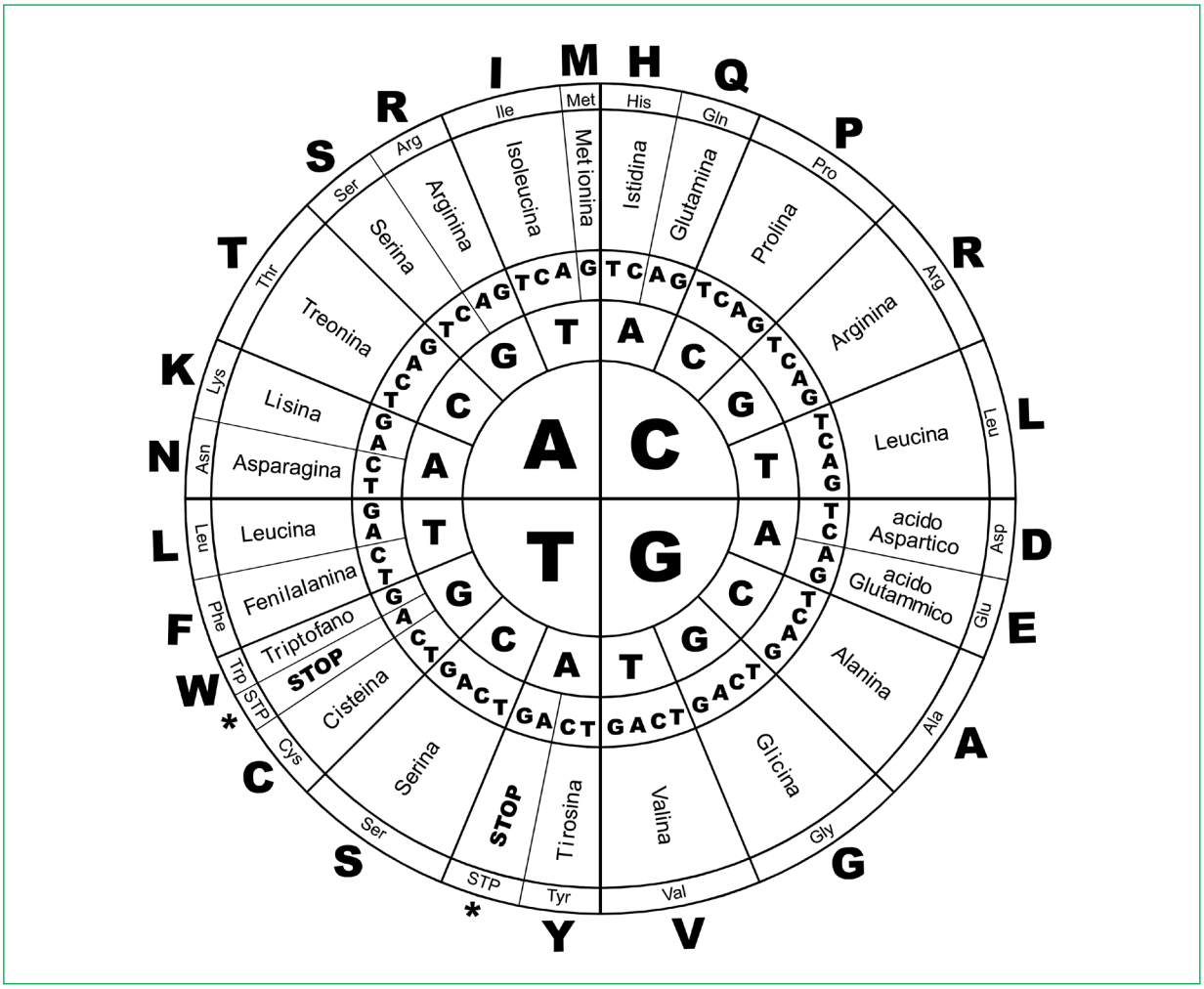


FIG. 1.7 Il "codice genetico", ovvero la corrispondenza fra trinucleotidi codificanti (codoni) e aminoacidi utilizzata nella sintesi proteica. Il primo nucleotide del codone va scelto nel quadrante interno, quindi il secondo e il terzo nucleotide nelle circonferenze più esterne. Le abbreviazioni a tre lettere e a una lettera dei 20 aminoacidi sono le seguenti: Alanina (Ala-A); Arginina (Arg-R); Asparagina (Asn-N); Acido aspartico (Asp-D); Cisteina (Cys-C); Acido glutamico (Glu-E); Glutamina (Gln-Q); Glicina (Gly-G); Istidina (His-H); Isoleucina (Ile-I); Leucina (Leu-L); Lisina (Lys-K); Metionina (Met-M); Fenilalanina (Phe-F); Prolina (Pro-P); Serina (Ser-S); Treonina (Thr-T); Triptofano (Trp-W); Tirosina (Tyr-Y); Valina (Val-V).

i singoli aminoacidi possono essere specificati da più codoni (si veda Fig. 1.7), si dice che il codice genetico è **degenerato**. Anche il numero di molecole di tRNA è inferiore a 61 (sono circa 30), tutti i 61 codoni possono essere riconosciuti perché l'anticodone del tRNA si appaia specificamente soltanto con le prime due basi del codone sull'mRNA, mentre la terza base può legarsi anche ad altre basi.

STRUTTURA ED ESPRESSIONE DEI GENI

Dal punto di vista della genetica molecolare, per **gene** si intende una sequenza di DNA che può essere trascritta in un RNA **funzionalmente** attivo. Tale RNA può svolgere direttamente una funzione **strutturale** e/o **catalitica** (rRNA, tRNA ecc.) oppure trasportare l'informazione per la sintesi di una proteina (mRNA). Nel genoma umano si stima che siano presenti quasi 21000 geni codificanti proteine e oltre 24000 geni che vengono trascritti in **RNA non codificanti** (ncRNA), che svolgono un ruolo importante nella regolazione della conformazione della cromatina e della trascrizione degli stessi geni codificanti proteine (si veda Tab. 1.1).

Il promotore

La regione a monte del sito di inizio della trascrizione è detta **promotore** (Fig. 1.8). La numerazione dei nucleotidi del promotore inizia da −1, che corrisponde al nucleotide che precede il sito di inizio della trascrizione (indicato con +1). In questa regione, di lunghezza variabile, si trova una serie di brevi sequenze (elementi *cis*) che viene riconosciuta e legata da **fattori di trascrizione** (elementi *trans*). I fattori di trascrizione reclutano l'RNA polimerasi II al sito d'inizio della trascrizione grazie all'interazione col ***mediator complex***, un complesso multiproteico contenente circa 30 proteine. Tale complesso è assolutamente necessario per l'inizio della sintesi dell'RNA, ma anche per le fasi successive della trascrizione. I geni che presentano elevati livelli di trascrizione, come gli istoni o la β-globina, hanno elementi del promotore che includono sempre un TATA box (TATAAA o una sua variante) circa 25 bp a monte del sito di inizio della trascrizione. I promotori di molti altri geni, tra cui i cosiddetti *geni housekeeping*, ossia geni necessari al funzionamento generale della cellula, non hanno TATA box, ma presentano spesso altri elementi come i GC box (GGGCGG).
Tra gli elementi frequentemente presenti nel promotore ricordiamo il CAAT box (a circa −80 bp dal sito di inizio della trascrizione). Oltre a sequenze comuni a molti promotori vi sono elementi che vengono riconosciuti da fattori di trascrizione **tessuto-specifici**. Anche geni che mostrano un'espressione tessuto-specifica vengono spesso trascritti a livelli molto bassi in tutte le cellule (**trascrizione illegittima** o **ectopica**).
Vi sono altre sequenze specifiche che vengono riconosciute da fattori di trascrizione quali gli elementi di risposta (**ER**), localizzati nel promotore o nella regione 5′ del gene, e gli elementi ***enhancer*** (intensificatori), che servono per aumentare i livelli basali della trascrizione e sono localizzati a distanza variabile dal gene (anche >50 kb), talvolta anche a valle del sito di inizio della trascrizione, all'interno della regione trascritta.

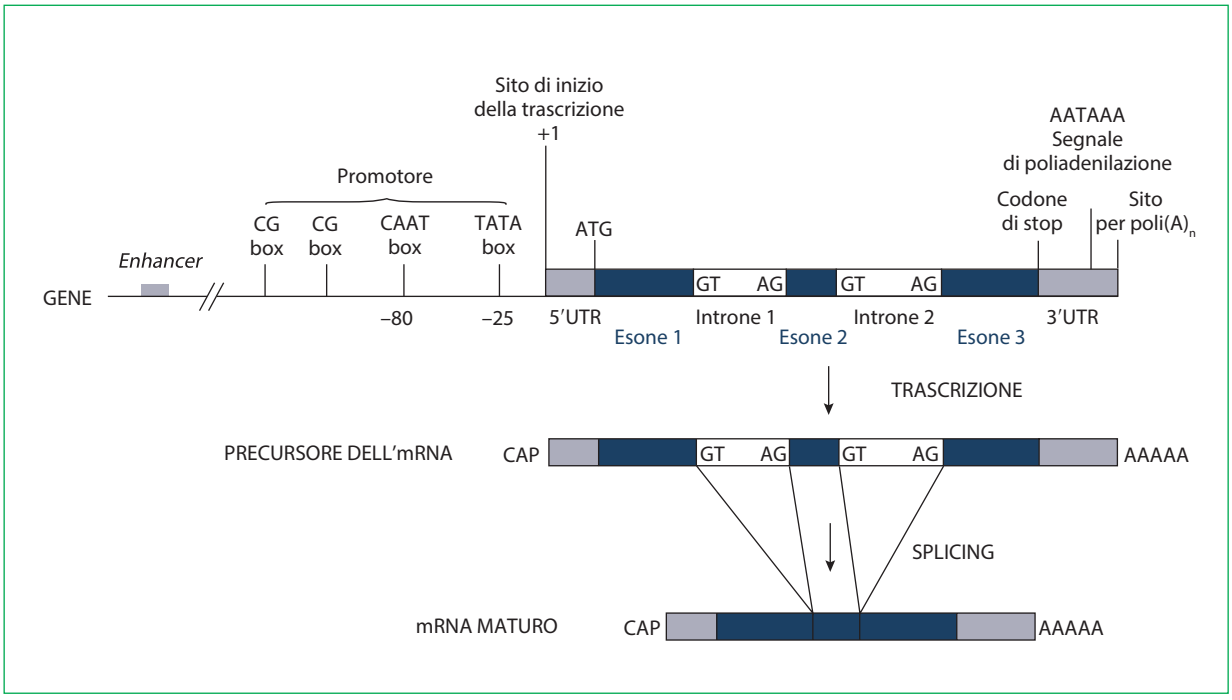


FIG. 1.8 Schema di un gene codificante proteina costituito da 3 esoni. Sono indicati alcuni elementi del promotore e i siti importanti per la trascrizione e la maturazione dell'RNA. Sono rappresentati anche il trascritto primario (precursore dell'mRNA) e l'mRNA maturo. Le dimensioni degli introni non sono in scala rispetto a quelle degli esoni.

Regolazione della trascrizione

La trascrizione del DNA in RNA è regolata in modo fine e complesso. Due sono le condizioni perché si abbia un'efficace trascrizione:

1) una conformazione della cromatina del gene (o della sequenza) "aperta", ovvero con i nucleosomi non compattati nella fibra a solenoide da 30 nm e possibilmente col DNA del promotore libero dagli istoni;
2) la presenza nella cellula di specifici fattori di trascrizione che, come accennato sopra, interagiscono con brevi sequenze nel promotore del gene e con sequenze *enhancer* e consentono l'assemblaggio del complesso di trascrizione (inclusa la RNA polimerasi II).

Il controllo dell'espressione genica mediante il legame di fattori proteici con le sequenze di regolazione è estremamente elaborato e coinvolge numerosi fattori che possono essere distinti in ubiquitari e tessuto-specifici. Tra quelli *ubiquitari* si trova il fattore Sp1 che interagisce con i GC box presenti nei geni *housekeeping*. L'interazione di *fattori specifici* con gli elementi *enhancer* è importante per l'espressione genica tessuto-specifica. Esempi di fattori tessuto-specifici sono NF-E2 e MyoD, che agiscono rispettivamente nelle cellule eritroidi e nei mioblasti. Vi sono, inoltre, numerosi altri meccanismi da cui dipende il controllo dell'espressione genica, come l'attivazione dei fattori di trascrizione in seguito al legame di un ligando, per esempio un ormone, con uno specifico recettore sulla superficie cellulare. Infatti, l'espressione di molti geni è controllata da un ormone, da un fattore di crescita o da una molecola di segnale intracellulare (come, per esempio, l'adenosina monofosfato ciclico, cAMP) che, legandosi a recettori specifici, determinano l'attivazione o l'inattivazione di determinati fattori di trascrizione (per esempio, mediante la loro fosforilazione o defosforilazione). I fattori attivati interagiscono, quindi, con delle specifiche sequenze del promotore, detti elementi di risposta (ER), inducendo l'espressione del gene corrispondente.

Notevoli progressi sono stati recentemente compiuti nell'identificazione di modificazioni della cromatina associate a una conformazione aperta (**attiva**) o compatta (**inattiva**). Tali modificazioni sono dette **epigenetiche**, perché non modificano la sequenza primaria del DNA (si veda Box 1.2).

In genere, la cromatina trascrizionalmente attiva è caratterizzata da un alto livello di **acetilazione degli istoni** H3 e H4 a livello dei residui di lisina delle code N-terminali. Tali code sporgono dal nucleosoma e normalmente legano il DNA grazie alla carica positiva dei residui di lisina (e arginina); l'acetilazione dei gruppi aminici delle lisine neutralizza tale carica e diminuisce quindi l'affinità degli istoni H3 e H4 per il DNA, facilitandone il distacco. Al contrario, gli istoni associati alla cromatina inattiva (non trascritta) sono praticamente tutti deacetilati. L'acetilazione e la deacetilazione degli istoni è determinata da due classi di enzimi: le **acetiltransferasi** (HAT, *Histone AcetylTransferase*) e le **deacetilasi** (HDAC, *Histone DeACetylase*). Un'altra modificazione, la **metilazione** del residuo di lisina in posizione 4 dell'istone H3 (**H3-K4**) è tipicamente presente nella cromatina attiva, mentre la **metilazione** del residuo di lisina in posizione 9 (**H3-K9**) è tipica della cromatina compatta e inattiva. Queste modificazioni vengono catalizzate da specifiche **metiltransferasi** (HMT, *Histone MethylTransferase*). Altre proteine sono poi in grado di riconoscere e legare le forme diversamente modificate degli istoni, collaborando allo svolgimento o al compattamento del DNA.

La modificazione epigenetica più studiata riguarda lo stesso DNA e consiste nella **metilazione delle citosine**, tipica delle regioni non trascritte e dei promotori di geni trascrizionalmente inattivi (*silenziati*). Specifiche **DNA metiltransferasi** (DNMT) sono in grado di aggiungere un gruppo metile in posizione 5′ dell'anello pirimidinico della citosina, a patto che questa sia seguita da una guanina, ovvero che faccia parte di un **dinucleotide CG** che viene indicato come CpG. I dinucleotidi CpG sono particolarmente frequenti nella regione 5′ di molti geni, di solito nella regione del promotore. In questo caso, si parla di **isole CpG**. Si stima che oltre il 50% dei geni sia associato a isole CpG. Le citosine metilate vengono riconosciute da proteine, dette MBD (*Methyl-DNA Binding Protein*), fra cui MECP2, determinando il reclutamento di altre proteine che modificano la cromatina e la compattano. Per esempio, MECP2 è in grado di richiamare sul promotore metilato diverse HDAC, che deacetilano ulteriormente gli istoni, facilitando l'inattivazione trascrizionale del gene stesso.

Struttura esoni-introni

Come accennato in precedenza, la sequenza codificante della maggior parte dei geni umani è suddivisa in segmenti detti **esoni**, separati da sequenze interposte non codificanti, dette **introni** (si veda Fig. 1.8). Il primo esone comincia al sito di inizio della trascrizione, ma il primo tratto (lungo in genere 200-300 bp) non è codificante; pertanto questo segmento è trascritto ma non tradotto e viene indicato come **regione 5′UTR** (*UnTranslated Region*). La regione 5′UTR è importante per l'efficienza della traduzione in quanto facilita il legame dell'mRNA ai ribosomi. La regione tradotta inizia di solito col trinucleotide ATG (AUG ovviamente sull'RNA) nel primo o nel secondo esone. Il numero degli esoni presenti nei geni umani è molto variabile; ci sono geni piccoli costituiti da un singolo esone, come i geni per gli istoni, e altri che possiedono più di 100 esoni, come, per esempio, alcuni geni che codificano per le catene del collagene. Il numero medio è circa 9-10 esoni per gene. I singoli esoni sono generalmente piuttosto corti (circa 200 bp), ma esistono alcuni esoni eccezionalmente lunghi che possono superare anche le 5 kb. Al contrario degli esoni, la dimensione degli introni è molto più variabile. In genere, i geni piccoli hanno introni piccoli, mentre in quelli più grandi gli introni possono avere una lunghezza anche di 10-20 kb.

Quasi tutti gli introni cominciano con il dinucleotide **GT** (*sito donatore* di splicing) e terminano con **AG** (*sito accettore* di splicing), come indicato in Figura 1.8. Questi dinucleotidi sono circondati da **sequenze consenso**, altamente conservate nel corso dell'evoluzione. Una terza sequenza consenso si trova nell'introne circa 40 bp a monte del dinucleotide AG. Essa è detta *sito di ramificazione* (*branch site*) e contiene obbligatoriamente una adenina. Tutte e tre le sequenze consenso vengono riconosciute dai fattori di splicing e sono indispensabili per una corretta eliminazione delle sequenze introniche durante la maturazione degli mRNA.

Il processo di splicing deve essere molto preciso, dato che lo spostamento anche di un singolo nucleotide determinerebbe lo slittamento del modulo di lettura e, quindi, la sintesi di una proteina alterata. L'ultimo esone, così come il primo, contiene una sequenza trascritta ma non tradotta, detta **regione 3′UTR** (generalmente molto più lunga della 5′UTR). La regione 3′UTR contiene il segnale di poliadenilazione (AATAAA) e sequenze dalle quali dipende la stabilità delle molecole di mRNA.

Trascritti alternativi e isoforme proteiche

Dai recenti progressi compiuti grazie al **Progetto Genoma Umano** è emerso che il numero dei geni umani codificanti proteine (attualmente stimato in 20400) è solo di poco superiore a quello di altri organismi più semplici, come il moscerino della frutta *Drosophila melanogaster* (13000) o il verme nematode *Caenorhabditis elegans* (18000). Si pensa tuttavia che i geni umani, e quelli dei mammiferi in generale, siano di norma coinvolti in processi biologici più complessi rispetto ai geni omologhi degli organismi più semplici. Infatti, molti geni umani sono in grado di specificare più prodotti proteici. Un tempo si riteneva che ogni gene codificasse per un unico prodotto polipeptidico o molecola di RNA, mentre oggi sappiamo che la maggioranza dei geni umani specifica due o più forme alternative (**isoforme**) proteiche. I meccanismi principali con i quali vengono generate queste diverse isoforme sono: l'uso di promotori alternativi, lo splicing alternativo e la poliadenilazione alternativa (Fig. 1.9).

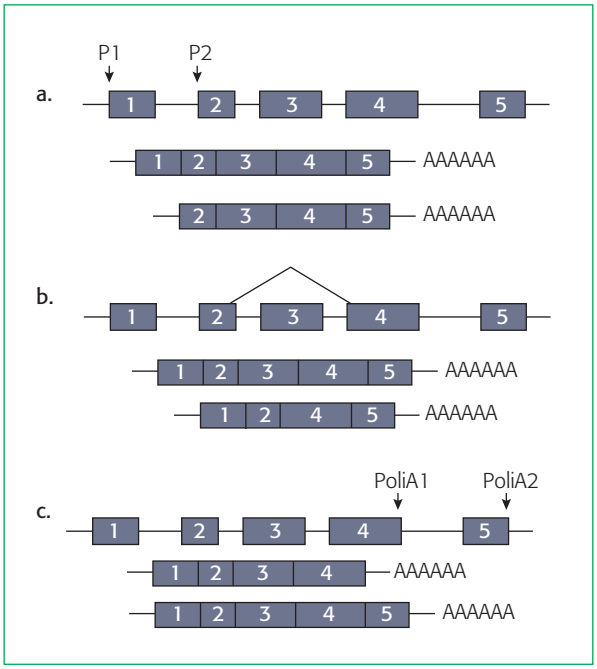


FIG. 1.9 Trascritti alternativi riconducibili a un unico gene. L'utilizzo di promotori diversi **(a)** e l'inclusione o esclusione di alcuni esoni **(b)** consentono la produzione di distinte isoforme, cioè proteine solo in parte simili tra loro. Anche l'impiego di due o più siti di poliadenilazione **(c)** consente la produzione di isoforme diverse che spesso presentano specificità di tessuto.

Si conoscono diversi geni umani che hanno due o più promotori che sono attivi specificamente in determinati tessuti e dirigono la sintesi di isoforme tessuto-specifiche oppure vengono attivati durante un particolare stadio dello sviluppo (Fig. 1.9a). Uno degli esempi più noti è il gene della distrofina, che presenta ben otto **promotori alternativi** tessuto-specifici differenti.

Il meccanismo più frequente col quale si generano delle isoforme diverse è lo **splicing alternativo**, che consiste nell'assemblaggio differenziale di esoni durante la maturazione dell'RNA (Fig. 1.9b). Si stima che oltre il 60% dei geni umani produca due o più proteine mediante questo meccanismo. Molti esoni specificano domini proteici strutturali distinti che possono essere combinati in modo diverso nelle cellule dei diversi tessuti nei quali il gene è espresso. Pertanto, a partire da un singolo gene vengono talvolta generate proteine simili ma non identiche, che possono avere funzioni diverse nei vari tessuti. Per dare un'idea dell'entità del fenomeno, basti osservare che nel database Ensembl sono riportati quasi 200000 diversi trascritti a fronte di 20400 geni codificanti per proteina.

Infine, un altro meccanismo piuttosto comune col quale possono essere generati trascritti diversi è la **poliadenilazione alternativa** (Fig. 1.9c). Molti geni, infatti, presentano nella regione 3'UTR due o più segnali di poliadenilazione che possono essere utilizzati durante la maturazione degli RNA. Anche in questo caso si possono formare isoforme tessuto-specifiche. In molti casi, durante la maturazione di uno specifico RNA, può verificarsi una combinazione di questi meccanismi, che permette la formazione di diverse isoforme proteiche a partire da un singolo gene.

Trascritti non codificanti e antisenso

Il dogma centrale suggerisce una visione semplificata degli eventi molecolari legati al flusso dell'informazione dal DNA alle proteine (si veda Box 1.1). Infatti, potrebbe sembrare che il DNA venga trascritto soltanto dai geni codificanti proteine e solo da un filamento. Inoltre, si potrebbe ritenere che, fatta eccezione per alcuni RNA a funzione strutturale (rRNA, tRNA ecc.), la maggior parte degli RNA abbia una funzione codificante, ovvero che si tratti di RNA messaggeri (mRNA). Entrambe queste supposizioni si sono rivelate errate: i ricercatori coinvolti nel progetto ENCODE (*ENCyclopedia Of DNA Elements*), teso a caratterizzare nel dettaglio la sequenza del genoma umano e dei trascritti da esso derivanti, hanno dimostrato la presenza di innumerevoli noncoding RNA (**ncRNA**) o trascritti non codificanti. Dunque, circa il 27% del genoma umano viene trascritto in mRNA (il 2% di esoni codificanti proteine più il 25% di introni), ma si stima che fino al 70-80% dell'intero genoma umano venga effettivamente trascritto, talvolta da entrambi i filamenti. A fronte dei quasi 21000 geni codificanti proteine si stima che siano prodotti fino a 100000 differenti ncRNA, per lo più non poliadenilati e confinati nel nucleo cellulare. Alcuni ncRNA sono di notevoli dimensioni (*long ncRNA* o **lncRNA**) e possono andare incontro allo splicing, come il trascritto ***XIST*** (*X-Inactivation Specific Transcript*), lungo circa 25 kb e coinvolto nel processo dell'inattivazione del cromosoma X. Molti ncRNA sono invece di piccole dimensioni (*short ncRNA*, tra cui i microRNA o

miRNA), sono ben conservati e hanno un ruolo importante in diversi processi cellulari quali lo splicing degli introni, l'assemblaggio dei ribosomi e il trasporto e la regolazione della stabilità degli RNA messaggeri. I **ncRNA** sono descritti in dettaglio nel Capitolo 3. Inoltre, sembra che la RNA polimerasi II non sia particolarmente specifica e possa iniziare a trascrivere in ogni posizione e direzione, se un tratto di DNA è privo di nucleosomi. Questo comporta anche la produzione di numerosi ncRNA associati ai promotori, che iniziano sia a monte sia a valle del sito di inizio della trascrizione dell'RNA messaggero e si dirigono sia nello stesso senso dell'mRNA sia in senso contrario (**antisenso**). Dunque, i ncRNA sono spesso parzialmente sovrapposti a trascritti codificanti (mRNA), essendo trascritti dal filamento antisenso. Si stima che circa il 60-70% degli mRNA abbia uno o più **trascritti antisenso** non codificanti, almeno parzialmente sovrapposti. Tali ncRNA antisenso svolgono un ruolo importante nella regolazione trascrizionale e post-trascrizionale, mediante il fenomeno della *RNA interference* (**RNAi**), descritto nel Capitolo 3.

Ruolo delle regioni non codificanti del genoma

Con l'introduzione di nuove metodiche *high-throughput* di indagine genetica si sta comprendendo sempre meglio il ruolo e l'importanza non solo degli RNA non codificanti, ma anche delle regioni di DNA conservate ma non trascritte (**CNG**, *Conserved Non-Genic sequences*) che regolano l'espressione genica. Specifiche sequenze nucleotidiche, come sopra descritto, funzionano da *enhancers*, ovvero vengono riconosciute come siti di legame da specifici fattori di trascrizione. È stato però osservato che tali *enhancers* svolgono il loro compito anche quando localizzati a grande distanza (fino a diverse Mb) dal gene controllato. Quindi, per avere una corretta interazione fra i vari elementi coinvolti nella trascrizione, è necessario che si formino delle anse della cromatina che avvicinino *enhancers* e promotori, uniti da ponti molecolari formati da fattori di trascrizione, la stessa RNA polimerasi II, il *mediator complex*, ma anche la proteina CTCF e le coesine. Grazie a metodiche quali la 3C (*chromosome conformation capture*) e l'*high-throughput sequencing* (il sequenziamento massivo parallelo di librerie di DNA ottenute con la 3C), sono stati rilevati diversi tipi di organizzazione tridimensionale della cromatina, con scale di grandezza dell'ordine delle megabasi, definite *topologically associated domains* (**TADs**). Tali domini sono spesso conservati fra le varie specie, fra tessuti e tipi cellulari e creano un'impalcatura che limita i contatti che un *enhancer* può avere, garantendone la specificità. I TADs sembrano avere il compito sia di promuovere contatti al loro interno, sia di evitare quelli con altri domini vicini. L'alterazione delle regioni genomiche che delimitano un TAD può avere conseguenze rilevanti sull'espressione dei geni contenuti nel TAD stesso, risultando in patologie umane. Un esempio riguarda quanto accadde a livello del locus *EPHA4*: una delezione che risparmia il gene ma coinvolge la regione telomerica del TAD che lo contiene causa brachidattilia, mentre inversioni e duplicazioni a carico della porzione centromerica sono associate a una forma complessa di sindattilia. In entrambi i casi, è stato dimostrato che gli *enhancers* che normalmente regolano l'espressione di *EPHA4* a livello dell'abbozzo degli arti vanno ad attivare in maniera aberrante geni diversi, ovvero *PAX3* nella brachidattilia e *WNT6* nella sindattilia.

ANATOMIA DEL GENOMA UMANO

Il Progetto Genoma Umano, iniziato nel 1990 da un consorzio internazionale di laboratori di ricerca, si proponeva nella sua fase iniziale di ottenere il sequenziamento completo dei circa 3 miliardi di nucleotidi che compongono il corredo aploide dei 23 cromosomi umani. Nel febbraio 2001 sono state pubblicate le prime due versioni della sequenza del DNA umano preparate rispettivamente dal consorzio internazionale pubblico IHGSC e dalla Celera Genomics, una compagnia privata che ha sostanzialmente validato i risultati pubblici, utilizzando peraltro le mappe fisiche già disponibili. Le sequenze del consorzio pubblico sono accessibili liberamente mediante apposite banche dati continuamente aggiornate (Box 1.3). I

BOX 1.3 BANCHE DATI ONLINE DI GENOMICA

L'accesso dei ricercatori o anche dei soggetti interessati all'enorme quantità di sequenze del genoma dell'uomo e di altri organismi è oggi possibile mediante il *World Wide Web*. Le principali banche dati dedicate al genoma umano sono:

- Genome Reference Consortium: www.ncbi.nlm.nih.gov/grc/human
- Genbank: www.ncbi.nlm.nih.gov/genbank/
- Ensembl: www.ensembl.org/index.html
- DDBJ (DNA Data Bank of Japan): http://www.ddbj.nig.ac.jp

Ulteriori dati relativi ad altri genomi, a banche dati dedicate alle sequenze degli RNA (trascrittoma) e delle proteine (proteoma), così come strumenti bio-informatici utili per l'interpretazione e l'analisi delle sequenze, sono disponibili ai seguenti siti:

- NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov
- EBI (European Bioinformatics Institute): http://www.ebi.ac.uk
- Sanger Institute: http://www.sanger.ac.uk
- UCSC Genome Bioinformatics (University of California Santa Cruz): http://genome.ucsc.edu

Infine, ogni anno il primo numero della rivista *Nucleic Acids Research*, dedicato alle banche dati di genetica e biologia molecolare presenti sul web, è gratuitamente disponibile sul sito http://nar.oxfordjournals.org/ seguendo il link "Database issue".

dati del 2001 sono stati aggiornati nel 2004 (si veda *Bibliografia essenziale*) fino a includere il 99% dell'eucromatina e circa il 94% del genoma umano, corrispondenti a 2,85 miliardi di nucleotidi sequenziati sul totale di circa 3 miliardi. A ottobre 2009 erano disponibili nelle banche dati genomiche oltre 3,09 miliardi di nucleotidi.

Nei prossimi paragrafi verrà descritta brevemente l'anatomia del genoma umano, classificando le sequenze di DNA in base alla loro ripetitività. In effetti, le sequenze altamente ripetute sono più difficili da collocare nella mappa completa del genoma, non tanto perché sia difficile sequenziarle, quanto perché non è facile ricostruire il numero di copie di repliconi presenti. È probabile che fra qualche anno anche le regioni più ripetitive, come quelle pericentromeriche e subtelomeriche, saranno "risolte". In Figura 1.10 sono riassunte graficamente le diverse classi di sequenze, dettagliandone le sottoparti. È importante ricordare che il Progetto Genoma Umano si è basato sulla ricostruzione della sequenza del DNA di pochi individui (una decina, al massimo), mentre si sa che sono moltissimi i polimorfismi che distinguono i membri di una stessa popolazione; sarà dunque necessario approfondire le variazioni naturalmente presenti fra i diversi individui e le diverse popolazioni con appositi programmi di ricerca dedicati alla genetica di popolazione. Infine, oltre alla conoscenza della sequenza del genoma, è ora di fondamentale importanza caratterizzare tutti gli effettivi trascritti codificanti (mRNA) e non codificanti (ncRNA, snRNA, miRNA ecc.) nei diversi tipi cellulari e correlarli alle diverse proteine presenti in un dato momento in una data cellula. Dopo l'era della "genomica" è dunque iniziata l'era della "proteomica" e della "trascrittomica".

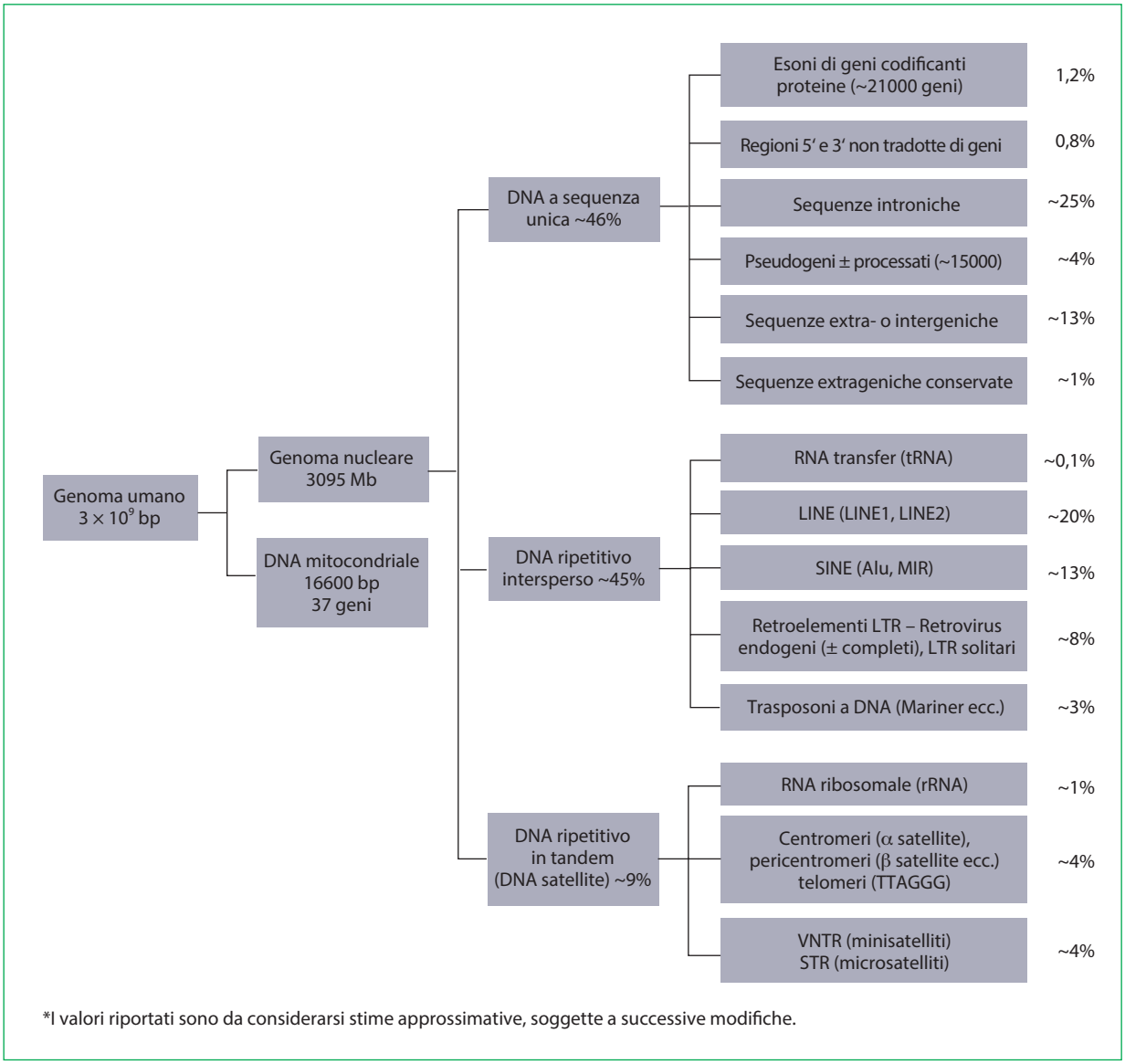


FIG. 1.10 Struttura del genoma umano. Il genoma nucleare è stato ripartito nelle tre componenti principali a seconda del numero di copie presenti (a sequenza unica, ripetitivo intersperso, ripetitivo in tandem). LINE = Long INterspersed Elements; LTR = Long Terminal Repeats; SINE = Short INterspersed Elements; STR = Short Tandem Repeats; VNTR = Variable Number of Tandem Repeats.

DNA a sequenza unica

Circa il 46% del genoma può essere considerato a sequenza unica, ovvero presente in singola copia. La gran parte di questo DNA è rappresentato da **geni codificanti proteine**, che secondo la stima più recente sono circa 20400, come già precisato. Soltanto il 2% dell'informazione genetica del genoma umano è tuttavia rappresentato negli mRNA maturi (circa 1,2% di esoni tradotti e 0,8% di sequenze 5′ e 3′ non tradotte). Infatti, le sequenze introniche ammontano a circa il 25% del genoma. Nella Tabella 1.1 sono indicate le stime attuali del numero di geni codificanti proteine e dei trascritti non tradotti per i singoli cromosomi. Si può notare come la densità dei geni sia estremamente variabile fra i vari cromosomi e nelle diverse porzioni di uno stesso cromosoma. Per esempio, il cromosoma 17 ha quattro volte più geni codificanti proteine del cromosoma 18 (1184 contro 269), pur essendo solo di poco più lungo, mentre il cromosoma 19 contiene più geni del cromosoma 2, che è quattro volte più lungo. È interessante notare che esiste una corrispondenza fra bandeggio cromosomico e densità genica: le regioni cromosomiche scure con il bandeggio G hanno un minore contenuto di guanina e citosina (C+G) e sono povere di geni, mentre le bande chiare hanno un contenuto C+G maggiore e sono più ricche di geni. Nel genoma umano non è raro imbattersi in cluster di geni simili (omologhi), vicini tra loro e derivati per duplicazione da un singolo gene ancestrale. Tali **famiglie di geni** possono anche essere disperse su più di un cromosoma e raggiungere le decine o anche centinaia di copie di geni funzionali, come nel caso dei recettori olfattivi (oltre 190 copie distribuite in almeno cinque cluster su diversi cromosomi).

Nel nostro genoma si trova inoltre un alto numero (circa 15200 secondo le stime più recenti) di copie non funzionali di geni detti **pseudogeni** che si sono originati con due meccanismi diversi: per *duplicazione* o per *retrotrascrizione* e successivo inserimento in altra parte del genoma. Nel primo caso lo pseudogene, completo di esoni e introni, è vicino a un gene omologo funzionante, ma ha accumulato delle mutazioni inattivanti che non consentono la sintesi di una proteina funzionale (**pseudogeni duplicati** o **non processati**). In pratica, questi pseudogeni, che vengono talvolta ancora trascritti, si sono originati in modo analogo ai geni funzionali omologhi a un gene ancestrale, come nel caso dei geni per le catene globiniche γ, β e δ, localizzati uno accanto all'altro sul cromosoma 11. Infatti, uno dei geni del cluster β-globinico è uno pseudogene non processato. La maggioranza degli pseudogeni (~60-70%) sono però **pseudogeni processati**, ovvero privi di introni e derivati dalla retrotrascrizione dell'mRNA della copia funzionale del gene ancestrale e dalla sua inserzione in un altro sito cromosomico (o più di uno). Gli pseudogeni processati, derivando da RNA messaggeri, possono avere una coda di poliA ma sono privi di promotore e quindi sono solo raramente trascritti. Come indicato nella Tabella 1.1, i **geni degli RNA non codificanti proteine** (ncRNA lunghi e corti, microRNA, tRNA, rRNA ecc.) sarebbero circa 24500, ma questa cifra potrebbe aumentare quando le annotazioni delle banche dati registreranno numerosi ncRNA, finora ignorati. Come ricordato nel paragrafo *Trascritti non codificanti e antisenso*, molti di questi ncRNA sono parzialmente sovrapposti a geni codificanti proteine e verrebbero trascritti dal filamento opposto (antisenso). È opportuno qui ricordare che il trascritto codificante per le tre principali molecole di RNA ribosomiale (**rRNA 18S**, **5.8S** e **28S**) è lungo circa 14 kb ed è presente in cluster di decine di copie localizzate in tandem nelle braccia corte dei cromosomi acrocentrici (13, 14, 15, 21 e 22).

DNA ripetitivo intersperso

Poco meno della metà (45%) del nostro genoma è mediamente ripetitivo, essendo composto da numerosissime copie disperse di un numero abbastanza limitato di **elementi trasponibili**. La stragrande maggioranza di questi elementi si è moltiplicata mediante un processo di retrotrascrizione dell'RNA (**retroelementi**), mentre meno del 3% del genoma è rappresentato da **trasposoni** a DNA. Alcuni di questi elementi sono ancora capaci di replicarsi e inserirsi in nuove posizioni del genoma, e questa capacità può essere causa di patologia genetica quando:

- un elemento si inserisce *de novo* all'interno di un gene funzionale e interferisce con la produzione della corrispondente proteina (*mutagenesi inserzionale*);
- due elementi si ricombinano fra di loro e causano la delezione, la duplicazione o l'inversione del tratto intermedio;
- la presenza di un elemento LTR (si veda più avanti) interferisce con la trascrizione di geni vicini.

I cosiddetti **LINE** (*Long INterspersed Elements*) occupano quasi il 20% del genoma totale. Gli elementi **LINE1** (L1) sono i più antichi e se ne contano circa 500000 copie. Solo l'1% di tali sequenze sono complete e, in questo caso, sono lunghe circa 6 kb e contengono un promotore interno per la RNA polimerasi II e due sequenze codificanti rispettivamente una proteina capace di legare lo stesso L1RNA e di traslocarlo nel nucleo e una trascrittasi inversa con attività endonucleasica. Quest'ultima è in grado di sintetizzare una molecola di DNA complementare sullo stampo dell'RNA (per questo motivo è chiamata **trascrittasi inversa**). Molto spesso, però, la trascrizione non è completa e il nuovo elemento LINE risulta troncato e non più in grado di trascriversi. Infatti, nel genoma umano sono presenti soltanto 80-100 elementi L1 completi e capaci di retrotrasposizione, mentre la maggior parte degli elementi LINE è incompleta e ha una lunghezza media di 1 kb. Il cromosoma X è particolarmente ricco di elementi L1 che costituiscono circa il 30% dell'intero cromosoma. Gli elementi LINE attivi sono ritenuti la fonte principale di trascrittasi inversa, responsabile anche della genesi degli pseudogeni processati.

Gli elementi **SINE** (*Short INterspersed Elements*) sono pari al 13% del genoma totale e hanno una lunghezza inferiore ai LINE. I più frequenti di questa classe sono le **sequenze Alu**, presenti in circa 1 milione di copie lunghe mediamente 300 bp (10% del genoma). Le sequenze Alu prendono il nome dall'enzima di restrizione corrispondente (derivato dal batterio *Arthrobacter luteus*), in grado di tagliare la sequenza 5′-AGCT-3′ contenuta nelle Alu. Le Alu derivano da un piccolo RNA denominato 7SL e vengono trascritte grazie a un promotore interno per la RNA polimerasi III; sono dotate di una coda poli-A e non contengono sequenze codificanti, per cui non sono capaci di retrotrasposizione autonoma, ma

possono essere retrotrasposte a opera della trascrittasi inversa degli elementi L1. Rispetto a questi ultimi, le sequenze Alu sono più ricche in C+G e si localizzano prevalentemente nelle regioni cromosomiche a maggiore densità genica, corrispondenti alle bande cromosomiche G-negative.
Circa il 9% del nostro genoma è composto invece da 400000 retroelementi con **LTR** (*Long Terminal Repeat*) e da retrovirus endogeni **HERV** (*Human Endogenous RetroVirus*). Questi ultimi, come molti retrovirus infettivi, sono lunghi circa 9-10 kb e contengono tre geni – *gag*, *pol* ed *env* – rispettivamente codificanti la proteina associata alla particella virale, la trascrittasi inversa e la glicoproteina integrata nella membrana fosfolipidica virale. A monte e a valle di questi tre geni si trovano due sequenze **LTR**, lunghe circa 1 kb, che svolgono la funzione di promotore e di segnale di poliadenilazione. La maggior parte dei retroelementi LTR è incompleta e manca di uno o più geni, anzi spesso è costituita da LTR isolati, probabilmente prodotti per ricombinazione non omologa con eliminazione del tratto interno. Nel caso in cui siano completi, sono spesso presenti numerose mutazioni inattivanti nei geni *gag* e *pol* che impediscono la replicazione. Soltanto la famiglia HERV-K, presente nell'uomo e nelle scimmie antropomorfe e quindi evolutivamente recente, conserva delle copie intere non mutate del genoma virale e rimane potenzialmente infettiva.
Infine, il 2-3% del genoma è composto da sequenze trasponibili (**trasposoni**), dotate di due ripetizioni invertite alle estremità e contenenti un gene codificante una proteina in grado di spostare il trasposone senza doverlo trascrivere in RNA e retrotrascrivere in DNA. Tali elementi sono praticamente tutti silenti nel genoma umano.
La presenza di numerosi elementi trasponibili nel genoma non rappresenta soltanto un rischio di malfunzionamento, ma anche un'opportunità di generare nuovi geni funzionalmente attivi. Per esempio, il gene codificante la sincizina, un'importante proteina placentare espressa a livello del trofoblasto, è chiaramente derivato da un gene *env* retrovirale che codifica una glicoproteina specializzata nella fusione delle membrane cellulari.

DNA ripetitivo in tandem

Circa il 9-10% del genoma umano è composto da regioni di DNA altamente ripetitivo, ossia composto da centinaia (e a volte migliaia) di unità ripetute in tandem di lunghezza variabile. Storicamente, questo DNA è noto come **satellite**, in quanto si dispone in bande distinte da quella principale se il DNA viene centrifugato in un gradiente di densità. Queste regioni, che comprendono i centromeri e i telomeri dei vari cromosomi, sono le più difficili da sequenziare proprio perché la loro natura estremamente ripetitiva rende praticamente impossibile l'allineamento dei cloni. Inoltre, l'aumentata frequenza di ricombinazione non omologa e di conversione genica in queste regioni determina una notevole variabilità fra diversi individui.
Circa 50 Mb di DNA genomico sono occupate dai **centromeri**, composti prevalentemente da DNA α satellite riconosciuto dalle proteine centromeriche che assicurano la corretta segregazione dei cromatidi fratelli durante l'anafase di mitosi e meiosi. Il DNA α satellite è costituito da un'unità base di 171 bp, polimorfica e organizzata in modo distinto sui diversi cromosomi (o gruppi di essi) in blocchi lunghi mediamente 2 Mb. Come accennato nel paragrafo *Replicazione del DNA*, le estremità delle braccia cromosomiche, o **telomeri**, sono composte da un semplice esanucleotide (TTAGGG), ripetuto migliaia di volte per una lunghezza totale di 5-15 kb, inversamente proporzionale all'età delle cellule. Le braccia corte dei cromosomi acrocentrici contengono, invece, oltre 200 copie dei geni codificanti per le subunità 18S, 5.8S e 28S dei ribosomi (**rDNA**), sintetizzati inizialmente come un unico trascritto, anche se il numero totale dei geni per l'rRNA è estremamente variabile.
Come indicato in Figura 1.10, circa il 3% del genoma è rappresentato dai cosiddetti **microsatelliti**, ovvero sequenze di 1-6 bp ripetute decine o anche centinaia di volte. I microsatelliti o **STR** (*Short Tandem Repeat*) sono oltre 1000000 nel genoma umano (quasi 400000 se non si contano quelli contenuti in LINE e SINE e i tratti mononucletidici) e sono abbastanza omogeneamente distribuiti lungo tutto il genoma, per questo vengono ampiamente utilizzati quali marcatori polimorfici per studi di associazione e di linkage (Capitolo 8). Un altro 1% è infine costituito da circa 30000 **minisatelliti**, altrimenti detti **VNTR** (*Variable Number of Tandem Repeats*), la cui unità ripetuta è compresa fra 10 e 100 bp, che storicamente sono stati tra i primi marcatori impiegati in studi di linkage prima di essere sostituiti dai microsatelliti.

VARIABILITÀ DEL GENOMA

Dopo avere brevemente descritto l'anatomia del genoma umano è importante ricordare che il Progetto Genoma non consente di apprezzare appieno la variabilità genetica che distingue un individuo dall'altro. Tale variabilità rappresenta in definitiva la misura della ricchezza genetica della nostra specie. Dobbiamo considerare almeno quattro tipi di polimorfismo genetico: (1) i **polimorfismi di singolo nucleotide** o **SNP** (*Single Nucleotide Polymorphism*); (2) **inserzioni** o **delezioni** (**indel**) di corte sequenze nucleotidiche (che possono risultare in piccole sostituzioni); (3) i **polimorfismi di lunghezza** di unità ripetute come i **microsatelliti** e i **minisatelliti**; (4) le **varianti strutturali,** che includono sia alterazioni bilanciate (con spostamento del materiale genetico) sia alterazioni sbilanciate, che danno luogo alle **varianti di numero di copie** o **CNV** (*Copy Number Variants*).
Gli SNP rappresentano in genere varianti di sequenza limitate a un solo nucleotide sostituito da un altro. Con lo sviluppo di tecnologie di indagine genetica sempre più avanzate, il numero degli SNP censiti nell'apposita banca dati (dbSNP) del *National Center for Biotechnology Information* (NCBI) sta crescendo rapidamente, avendo superato il miliardo nell'ultima versione (Build 155), includendo sia gli SNP presenti in una percentuale della popolazione convenzionalmente superiore all'1% sia quelli più rari (in tal caso si dovrebbe utilizzare il termine generico di SNV, *Single Nucleotide Variant*). Gli SNP sono quindi mutazioni puntiformi che possono anche creare o abolire siti di riconoscimento per enzimi di restrizione del DNA e, in questo caso, corrispondono ai primi marcatori impiegati negli anni Ottanta del secolo scorso per lo studio

del genoma, ovvero i **RFLP** (*Restriction Fragment Length Polymorphisms*). Le **indel** consistono invece nell'inserzione o nella delezione di una serie di nucleotidi e, in questo caso, se contenute all'interno di una sequenza codificante, possono determinare lo slittamento della cornice di lettura (ossia una mutazione *frameshift*). La lunghezza del tratto coinvolto può andare da pochi nucleotidi fino a 1 kb, limite usato per convenzione per distinguerle dalle CNV, sebbene più recentemente sia stato proposto un *cut-off* di 50 nucleotidi.
I **minisatelliti** e i **microsatelliti** potrebbero essere inclusi nella categoria precedente, ma presentano caratteristiche peculiari che saranno descritte a parte (Capitolo 16). Sono rispettivamente decine di migliaia e centinaia di migliaia e la loro variabilità di lunghezza deriva da errori nella replicazione del DNA (***replication slippage***) che determinano l'aggiunta (o la perdita) di alcune unità da un insieme di ripetizioni in tandem. Mentre gli SNP hanno in genere solo due alleli, micro- e minisatelliti possono avere numerosi alleli e sono pertanto molto più informativi per studi di linkage o di associazione.
Infine, le **varianti strutturali** (incluse le **CNV**) sono lunghe almeno 1 kb (o 50 bp, a seconda dei criteri considerati) fino a qualche megabase. In alcuni casi rappresentano il prodotto della ricombinazione non omologa fra sequenze ripetitive intersperse come le sequenze Alu, LINE1 o retroelementi dotati di sequenze LTR, che determinano delle **duplicazioni o delezioni segmentali**.
Le CNV spesso includono geni funzionalmente attivi, per cui possono determinare patologie genetiche dovute ad aploinsufficienza (o all'eccesso di trascrizione) di uno o più geni (disordini genomici), come discusso nel Capitolo 14. Le CNV sono attivamente studiate da diversi anni e il *Database of Genomic Variants* (DGV) documenta quasi un milione di regioni cromosomiche interessate da CNV.
In conclusione, il nostro genoma è estremamente dinamico e ciascuno di noi può potenzialmente differire dagli altri individui a livello di 4-5 milioni di siti differenti (per lo più SNP, indel e microsatelliti), per un totale di circa 20 megabasi. Se si considerano complessivamente le varianti censite da dbSNP (SNP, indel e microsatelliti) e DGV (CNV), più dell'80% del genoma umano è potenzialmente sede di varianti che non compromettono il normale sviluppo dell'individuo.
Le tecnologie di sequenziamento massivo parallelo (discusse nel Capitolo 2), stanno rivoluzionando la genomica e la diagnostica, consentendo il sequenziamento massivo (ed economico) degli acidi nucleici e quindi: (1) l'identificazione di varianti di sequenza presenti soltanto in alcune cellule (mosaicismo tissutale, mutazioni tumorali) e (2) il conteggio di copie di DNA (*Copy Number Variants*) e di RNA (studi dell'espressione genica). Tali tecnologie rappresentano la necessaria premessa per la conoscenza sempre più approfondita del nostro genoma e per l'ideazione di terapie innovative per le patologie genetiche umane.

Bibliografia essenziale

Berretta J, Morillon A. Pervasive transcription constitutes a new level of eukaryotic genome regulation. EMBO Rep 2009;10:973-982.
Carninci P. The long and short of RNAs. Nature 2009;457:974-975.
Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell 2009;136:642-655.
Cecere G. Small RNAs in epigenetic inheritance: from mechanisms to trait transmission. FEBS Lett 2021.
Cedar H, Bergman Y. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 2009;10:295-304.
Conrad DF et al. Origins and functional impact of copy number variation in the human genome. Nature 2010;464:704-712.
Djebali S et al. Landscape of transcription in human cells. Nature 2012;489:101-108.
Green KM, Linsalata AE, Todd PK. RAN translation-What makes it run? Brain Res 2016;1647:30-42.
Hecht A et al. Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res 2017;45:3615-3626.
Iafrate AJ et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949-951.
International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-945.
International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
Lister R et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009;462:315-322.
Matzke M, Bircher JA. RNAi-mediated pathways in the nucleus. Nat Rev Genet 2005;6:24-35.
Mendes Soares LM, Valcarcel J. The expanding transcriptome: the genome as the "Book of Sand". EMBO J 2006;25:923-931.
Mewborn SK, Lese Martin C, Ledbetter DH. The dynamic nature and evolutionary history of subtelomeric and pericentromeric regions. Cytogenet Genome Res 2005;108:22-25.
Morris KV et al. Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. PLoS Genet 2008;4:e1000258.
Richard GF, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 2008;72:686-727.
Roy SW, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 2006;7:211-221.
Schueler MG et al. Genomic and genetic definition of a functional human centromere. Science 2001;294:109-115.
Schwartz JC et al. Antisense transcripts are targets for activating small RNAs. Nat Struct Mol Biol 2008;15:842-848.
Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 2007;14:103-105.
Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome Biol 2003;4(2):R13.
Turner BM. Cellular Memory and the Histone Code. Cell 2002;111:285-291.
Venter JC et al. The sequence of the human genome. Science 2001;291:1304-1351.
Wakatsuki S, Araki T. Novel Molecular Basis for Synapse Formation: Small Non-coding Vault RNA Functions as a Riboregulator of MEK1 to Modulate Synaptogenesis. Front Mol Neurosci 2021;14:748721.

Siti Internet

ClinGen (Clinical Genome Resource): clinicalgenome.org
Database of Genomic Variants: dgv.tcag.ca/dgv/app/home
dbVar,NCBI's database of genomic structural variation: www.ncbi.nlm.nih.gov/dbvar
DDBJ (DNA Data Bank of Japan): www.ddbj.nig.ac.jp
DECIPHER (Mapping the clinical genome): www.deciphergenomics.org
EBI (European Bioinformatics Institute): www.ebi.ac.uk
Encyclopedia of DNA Elements (ENCODE): encodeproject.org
Ensembl Genome Browser: www.ensembl.org
Genbank: www.ncbi.nlm.nih.gov/Genbank
Gene Ontology Resource: geneontology.org
KEGG (Kyoto Encyclopedia of Genes and Genomes): www.genome.jp/kegg
HPO - Human Phenotype Ontology: hpo.jax.org/app
NCBI (National Center for Biotechnology Information): www.ncbi.nlm.nih.gov
Nucleic Acids Research Database Summary Paper Category List: www.oxfordjournals.org/nar/database/c/
OMIM (Online Mendelian Inheritance in Man): www.omim.org
Sanger Institute: www.sanger.ac.uk
SNP database: www.ncbi.nlm.nih.gov/SNP
UCSC (University of California Santa Cruz) Genome Bioinformatics: genome.ucsc.edu
Varsome (The human genetics search engine): varsome.com