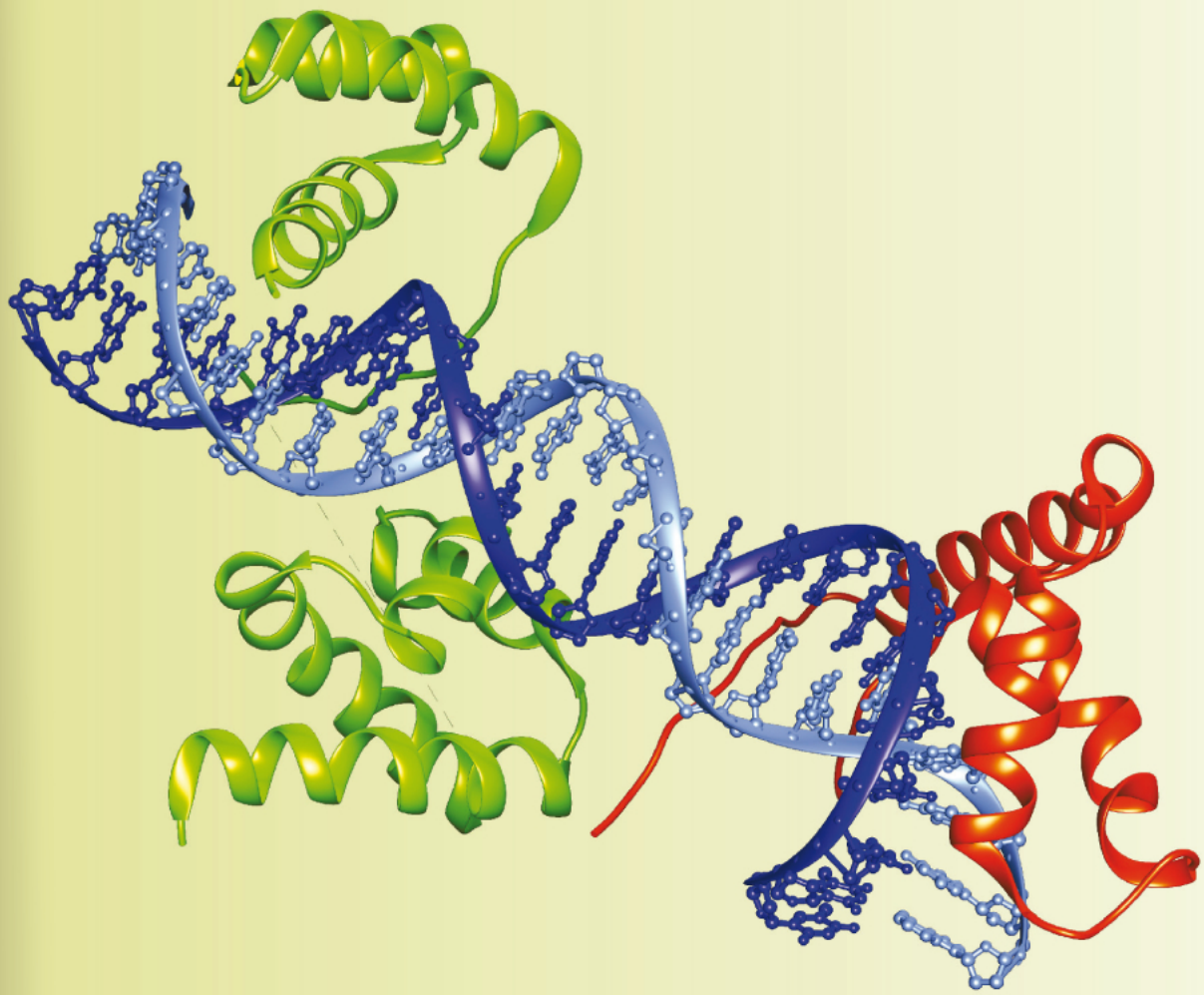
Accedi all'ebook

Biologia Molecolare

G. Capranico
E. Martegani
G. Musci
G. Raugei
T. Russo
N. Zambrano
V. Zappavigna



EdiSES

Capitolo

[4]

Geni e genomi: organizzazione e funzione

SOMMARIO

[4.1] LA FUNZIONE BIOLOGICA: IL PROGRAMMA DELLA VITA

Il genoma di un organismo è definito come **l'intero patrimonio genetico aploide contenuto nelle cellule**, e fisicamente corrisponde all'intera sequenza di nucleotidi delle molecole di DNA (o di RNA nel caso di alcuni virus) presenti nelle cellule. Il termine **genoma** è stato usato per la prima volta dal botanico tedesco Hans Winkler nel 1920, e deriva dalla fusione dei due termini *gene* e *chromosome*. Il genoma può essere relativamente semplice o molto complesso, ma in tutti i casi contiene le istruzioni o informazioni necessarie per eseguire e regolare le molteplici funzioni delle cellule. Quindi, la sua funzione principale e universale è quella di costituire un "deposito" di informazioni che sia trasmissibile con poche eventuali alterazioni (mutazioni) alla progenie. Con l'avanzare delle conoscenze, però, si è compreso che questo "deposito" non è da intendersi come qualcosa di statico, che immagazzina e all'occorrenza fornisce le istruzioni necessarie, quanto piuttosto a una fonte di informazioni dinamica che può e a volte deve modificarsi per regolare le funzioni dell'organismo.

L'importanza fondamentale del genoma per la vita è evidenziata dall'impatto che ha avuto sulle nostre conoscenze e tecnologie il sequenziamento del genoma umano. Partito nei primi anni '90 del secolo scorso, il sequenziamento del genoma umano è stato completato all'inizio di questo secolo. Due consorzi di ricercatori e enti di ricerca, uno pubblico e uno privato, hanno pubblicato nel 2001 la sequenza di circa l'85% del genoma umano su *Nature* e *Science*, due riviste scientifiche molto prestigiose (**Figura 4.1**). In questa versione preliminare, mancavano regioni ripetute povere di geni, che non erano allora facilmente sequenziabili per ragioni tecniche. Il genoma umano nucleare è di circa 3,3 miliardi di coppie di basi, quindi ciò vuol dire che in una cellula somatica umana l'intero contenuto (diploide) di DNA nucleare corrisponde a 6,6 miliardi di coppie di basi. Oltre il DNA nucleare, il genoma di una cellula eucariotica comprende anche il DNA mitocondriale e dei cloroplasti, poiché questi organelli contengono un proprio DNA. In questi casi, comunque, la lunghezza della molecola di DNA è molto inferiore di quello nucleare, per esempio il genoma mitocondriale umano è di 16.571 coppie di basi.

Un numero così grande di nucleotidi del genoma umano è difficile da comprendere subito e pone delle difficoltà nuove per riuscire a capire le sue funzionalità. Per fare un'analogia consideriamo questo libro che contiene circa un 1 milione di caratteri alfanumerici; quindi, se volessimo scrivere l'intera sequenza di basi del genoma aploide umano ci vorrebbero più o meno 3300 libri della dimensione di questo. Ecco perché comprendere e interpretare pienamente il genoma umano è un'impresa enorme, e infatti, il 2001 ha segnato uno spartiacque tra un'era pre- e un'era post-genomica. Da quel momento il genoma di un organismo è divenuto il punto di riferimento di molte indagini scientifiche, e in buona misura una finalità importante e comune a molti ricercatori è la definizione dei meccanismi e delle funzionalità complesse dei genomi. Inoltre, il sequenziamento dei genomi è stato reso possibile dagli enormi avanzamenti tecnologici e informatici (si veda il paragrafo successivo). Questi avanzamenti tecnologici hanno rivoluzionato il modo di acquisire conoscenze anche in altri settori della biologia molecolare, per esempio la trascrizione e le interazioni proteina-acido nucleico. In effetti, l'era post-genomica ha rivoluzionato tutta la biologia e ha già permesso ulteriori avanzamenti delle conoscenze in diversi campi della biologia e delle biotecnologie. Anche più inatteso è forse l'impatto sulla vita quotidiana di ognuno di noi che la conoscenza dei genomi ha avuto, e progressivamente avrà sempre di più, grazie allo sviluppo delle biotecnologie in numerosi settori della vita economica e sociale delle popolazioni umane.

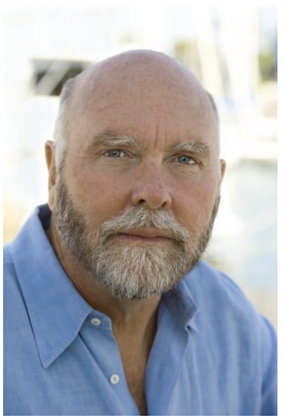
FIGURA 4.1 ▶ Le copertine di *Nature* e *Science* che celebrano la pubblicazione nel 2001 della prima versione del genoma umano da parte del Consorzio di Istituzioni di ricerca Internazionali e della Celera Genomics.

I genomi non sono entità statiche e immutabili, al contrario sono strutture dinamiche che vanno incontro a modificazioni anche importanti nella trasmissione ereditaria alla progenie. Un importante dato che emerge dall'enorme mole di dati genomici attualmente presente nelle banche dati è la straordinaria conservazione a livello molecolare tra le specie viventi. La diversità delle forme viventi sulla terra sono unificate da temi comuni che riguardano sia processi molecolari che comportamenti biologici. L'analisi e il confronto di molti genomi evidenziano come molti geni sono simili tra varie forme viventi così come altri tipi di sequenze genomiche. Questo significa che anche i processi di espressione dell'informazione genetica in forme funzionali sono simili tra i vari organismi. Il confronto dei genomi pertanto permette indagini approfondite dell'evoluzione della vita sulla terra e di stabilire le relazioni evolutive tra le specie con la costruzione di **alberi filogenetici**. Nello stesso tempo, queste osservazioni forniscono una solida base razionale e scientifica per l'utilizzo di **organismi modello** non solo per la ricerca di base dei processi molecolari ma anche per la ricerca applicata volta, per esempio, alla scoperta dei meccanismi delle patologie o per lo sviluppo di nuovi agenti terapeutici per malattie umane.

[4.2] SEQUENZIAMENTO E ANNOTAZIONE DEI GENOMI

Il progetto **Genoma Umano** è partito nel 1990 su iniziativa pubblica dei *National Institutes of Health* (**NIH**) e del US *Department of Energy* del governo statunitense, che hanno costituito un apposito consorzio (*International Human Genome Sequencing Consortium*) che includeva altre 20 istituzioni scientifiche degli USA, Gran Bretagna, Francia, Germania, Cina e Giappone. Il coordinamento del progetto fu affidato al Centro di Ricerca genomica dell'NIH diretto da Francis Collins (**Figura 4.2b**), e all'inizio l'impresa sembrava titanica. Il progetto Genoma Umano funzionò subito come un catalizzatore di progressivi avanzamenti tecnologici che hanno reso sempre più economico e semplice sequenziare una lunga e complessa sequenza di nucleotidi. Il progetto subì quindi una forte accelerazione fino a tal punto che fu terminato con due anni di anticipo rispetto al termine previsto inizialmente. I progressi tecnologici inoltre stimolarono la nascita nel 1997 di un'impresa privata, la *Celera Genomics* di J. Graig Venter (**Figura 4.2a**), con l'obiettivo di sequenziare l'intero genoma umano in tempi più rapidi. I due progetti pubblicarono la prima bozza del genoma umano contemporaneamente nel 2001, e infine la sequenza completa nel 2004 nota come **NCBI *Human Build* 35 (May 2004)**.

L'approccio metodologico utilizzato dal consorzio pubblico per il sequenziamento del Genoma umano prevedeva di frammentare il genoma con enzimi di restrizione e quindi clonare frammenti lunghi in vettori **BAC** (*Bacterial Artificial Chromosome*) e **YAC** (*Yeast Artificial Chromosome*) (**Figura 4.3**). Questi vettori permettono di clonare inserti di DNA molto lunghi rispettivamente in *E. coli* e in *S. cerevisiae*. Quindi, si procedeva a identificare i cloni che erano sovrapposti attraverso cross-ibridazione e altri metodi, e che venivano quindi allineati in una serie di **contig**, tratti cromosomali contigui. I vari contig contenevano uno o più marcatori cromosomali precedentemente caratterizzati, come le **STS** (*Sequence Tagged Sites*) o geni espressi e identificati con le **EST** (*Expressed Sequence Tags*). Questi **marcatori** di posizione cromosomale sono stati essenziali per la costruzione dei contig e per la fase successiva di **as-**



a)

b)

FIGURA 4.2 ◀ Craig Venter (**a**) e Francis Collins (**b**) leader scientifici del Progetto Genoma Umano, rispettivamente, della *Celera Genomics* e del Consorzio di Istituzioni di ricerca Internazionali.

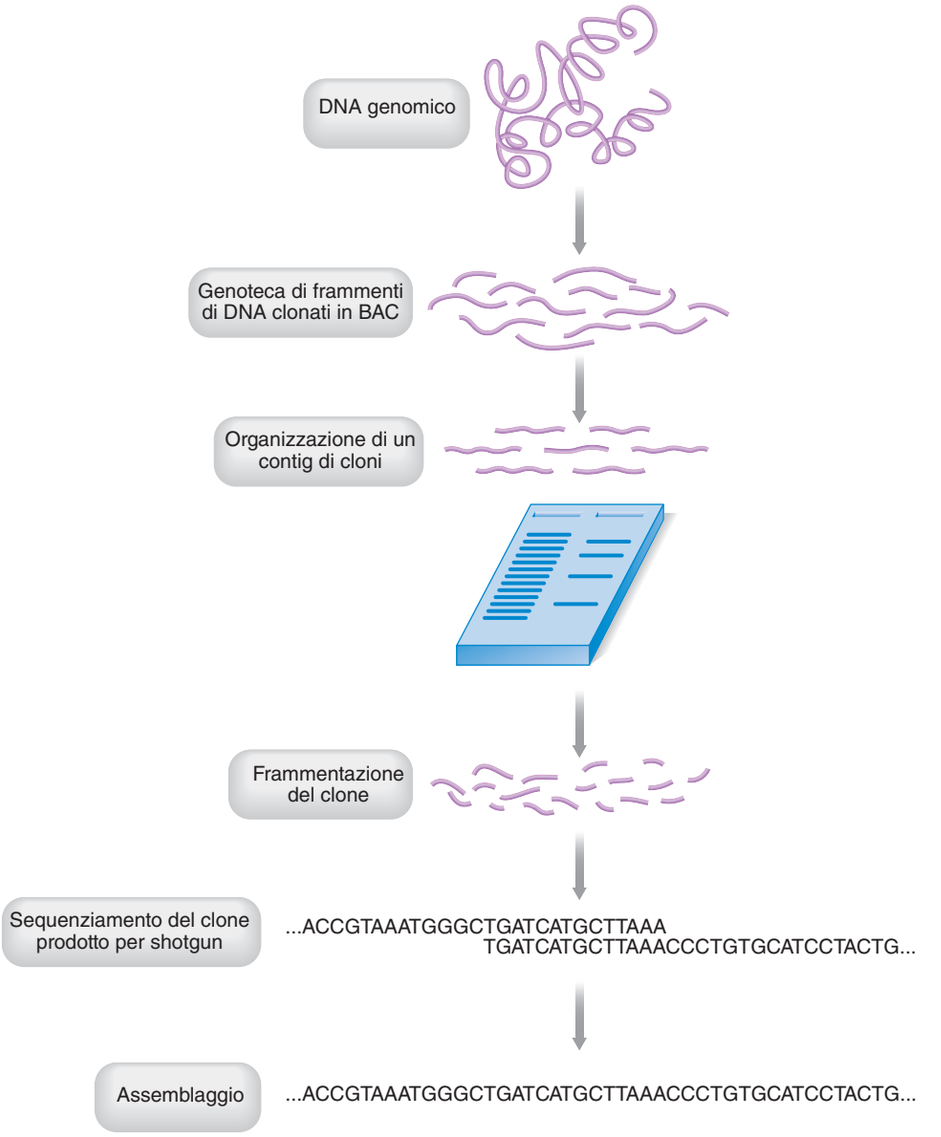
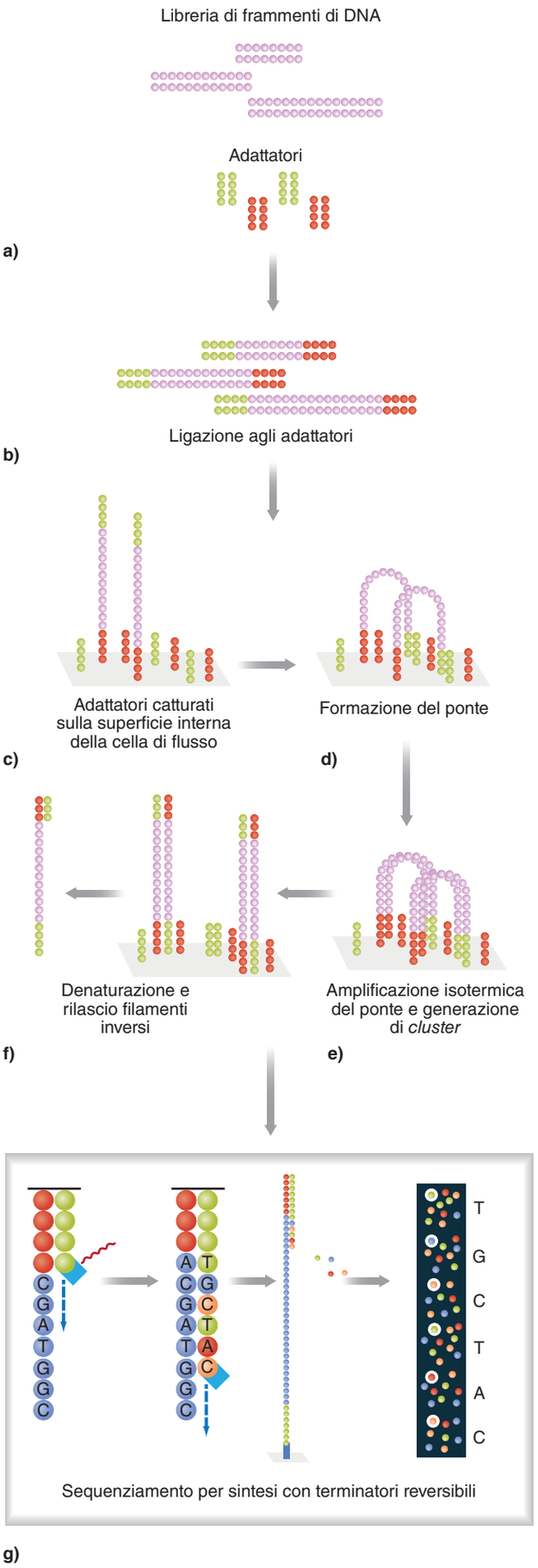


FIGURA 4.3 ▲ Schema delle fasi sperimentali utilizzate nel Progetto Genoma Umano. Per sequenziare il genoma umano, si partì dalla costruzione di una libreria di cloni contenenti grandi frammenti di tutto il genoma umano; nella illustrazione sono mostrati i vettori BAC (*Bacterial Artificial Chromosome*). Quindi, i grandi pezzi clonati venivano allineati in una mappa fisica sulla base di vari marcatori genomici noti. I cloni BAC erano quindi sequenziati tramite la metodica *shotgun*. Le sequenze di DNA erano quindi assemblate per ricostruire la sequenza dei contig, e anche sulla base della mappa fisica, dell'intero genoma.

semblaggio delle sequenze genomiche. I vari contig furono quindi assegnati alle varie istituzioni del consorzio per il sequenziamento (**Figura 4.3**).

Molti dei cloni BAC o YAC erano più lunghi di 100.000 coppie di basi e poiché le tecniche di sequenziamento alla fine del secolo scorso coprivano al massimo 600-700 basi alla volta, ciascun clone fu sequenziato a pezzi. La procedura seguita fu denominata ***shotgun***, che utilizzava sequenziatori automatici e altri strumenti robotizzati per il sequenziamento di frammenti casuali del clone (**Figura 4.3**). Successivamente, le varie sequenze casuali erano assemblate in sequenze più lunghe sulla base dell'identificazione di sequenze identiche in comune tramite metodologie bioinformatiche. Il sequenziamento ripetuto almeno quattro volte di ciascun clone assicurava una maggiore precisione della sequenza ottenuta. I notevoli progressi ottenuti soprattutto a livello bioinformatico, stimolarono la nascita di *Celera Genomics* che utilizzò il **sequenziamento shotgun dell'intero genoma umano** eliminando del tutto la costruzione



dei BAC e YAC, e sequenziando frammenti casuali di tutto il genoma clonati in vari tipi di vettori. Questo approccio è molto più veloce ma rende la fase di ricostruzione della sequenza (assemblaggio) più complessa e quindi richiede strumenti bioinformatici molto più sofisticati. Ovviamente possono rimanere dei buchi più o meno estesi in funzione soprattutto del contenuto di sequenze ripetute. Lo sviluppo di nuove piattaforme di sequenziamento **NGS** (*Next Generation Sequencing*) permette attualmente di sequenziare direttamente frammenti di DNA o RNA evitando completamente le complesse fasi di clonaggio e rendendo ancora più veloce e meno costoso l'intero processo. Nel 2010, un consorzio internazionale guidato dal Centro di Ricerca genomica di Shenzhen, Cina, ha pubblicato il genoma del panda gigante, *Ailuropoda melanoleura*, il primo a essere stato interamente sequenziato tramite la tecnologia NGS di Illumina (**Figura 4.4**).

Con il sequenziamento del genoma umano sono stati sequenziati genomi di altre specie, soprattutto di organismi modello utilizzati nei laboratori. Il primo genoma di un organismo vivente sequenziato completamente è stato quello del batterio *Haemophilus influenzae* nel 1995, e il primo eucariote è stato il lievito *Saccharomyces cerevisiae* nel 1996. Il genoma del batterio forse più utilizzato nei laboratori di ricerca, *Escherichia coli,* è stato invece pubblicato dopo, nel 1997. In seguito, il numero di genomi sequenziati è cresciuto di molto e alla data del 5 maggio 2015, i genomi sequenziati con il metodo shotgun (quello utilizzato per il genoma umano) e resi disponibili pubblicamente sono quasi 4000 includendo eucarioti, *Archaea* e batteri. Le informazioni aggiornate si possono trovare sul sito europeo dell'**EMBL-European Bioinformatics Institute**, **www.ebi.ac.uk/genomes**. Tuttavia, le metodologie di sequenziamento sono ormai numerose e i progetti che coinvolgono il sequenziamento sia del DNA che dell'RNA, purificati da una varietà di fonti biologiche (individui diversi, singole cellule o campioni ambientali), sono ormai decine di migliaia. Tutta questa mole di dati e informazioni richiede uno sforzo particolare di standardizzazione

FIGURA 4.4 ◀ Fasi della tecnologia ILLUMINA di sequenziamento massivo del DNA. Le fasi illustrate nella figura sono: **a)** il DNA è frammentato in pezzi di 200-600 basi; **b)** corti oligonucleotidi, chiamati *adapters*, sono legati ai frammenti di DNA; **c)** il DNA è denaturato e i frammenti di DNA a singola catena sono posti su una superficie dove sono presenti dei *primer* complementari agli *adapters*, che quindi si appaiano tra loro; **d)** il DNA viene fatto replicare per formare dei *cluster* di frammenti unici sulla superficie; **e)** le doppie eliche sono quindi denaturate per formare clusters di filamenti singoli di DNA; **f)** *primer* specifici vengono quindi aggiunti che si appaiano ai framenti singoli di DNA e il sequenziamento avviene per sintesi del DNA con l'uso di nucleotidi fluorescenti (illustrato nella parte in basso della figura); **g)** la sequenza viene rilevata automaticamente da un computer collegato a un fotorilevatore.

dei protocolli sperimentali e analitici e di modalità omogenee di gestione e organizzazione delle banche dati dedicate. Va in questa direzione il ***Genomes Online Database*-GOLD** (https://gold.jgi-psf.org) che è una risorsa pubblica per l'accesso completo e integrato alle informazioni genomiche, post-genomiche e metagenomiche degli studi condotti in tutto il mondo. In questo database sono conservate e rese disponibili informazioni genomiche che riguardano quasi 70.000 specie o campioni biologici.

Nell'era post-genomica, le indagini scientifiche e tecnologiche sono rivolte principalmente alla comprensione delle funzioni del genoma. La **genomica** è dunque un nuovo campo di indagini scientifiche e tecnologiche che si dedica allo studio della struttura e funzioni dei genomi e degli acidi nucleici, a livello cellulare (un sito generale di riferimento è **www.ncbi.nlm.nih.gov**/genome dell'americano **NCBI – *National Center for Biotechnology Information*** o genome.**ucsc.edu UCSC *Genome Browser*** dell'Università della California a Santa Cruz (si veda la **Figura 4.5** per un esempio di un browser genomico). Come sottolineato nei paragrafi precedenti, i progressi in questo nuovo campo dipendono sostanzialmente dagli avanzamenti tecnologici di sequenziamento (*high-throughput sequencing*) quali NGS e di metodologie bioinformatiche di analisi e di organizzazione in banche dati dei risultati ottenuti. Lo sviluppo di banche dati dedicate è crescente in termini numerici e di modalità organizzative sempre migliori e integrate. Infatti, una sequenza di genoma è semplicemente una stringa di 4 caratteri, A, T, G e C, ma il suo valore e significato biologico dipendono quasi esclusivamente dalle modalità organizzative attraverso le quali la sequenza è depositata. Queste modalità devono pertanto prevedere tutte le informazioni disponibili di ciascuna sequenza o regione genomica relativamente alle funzioni e alle strutture. L'aggiunta di ulteriori informazioni funzionali alle sequenze genomiche è dunque un processo molto importante ai fini della possibilità di studiare e conoscere il genoma, ed è generalmente indicato come **annotazione genomica.**

L'annotazione genomica generalmente parte con l'identificazione di tutti i geni presenti nel genoma. Per gene, qui intendiamo una regione del DNA che codifica per una proteina o per un RNA non codificante (*non-coding RNA*) come tRNA o rRNA. Per definire la posizione dei geni in un genoma vengono cercate sequenze con particolari caratteristiche. Un primo, classico metodo è la ricerca di sequenze **ORF** (*Open Reading Frame*) che individua un potenziale gene codificante per una proteina. Con il progressivo ampliarsi delle conoscenze di geni di molti organismi diversi, un altro metodo è la ricerca di sequenze simili a geni noti di altri organismi con metodologie bioinformatiche (si vedano anche i paragrafi successivi). La conoscenza del gene e delle sue funzioni in un organismo permette infatti l'individuazione di nuovi geni in altri genomi e caratterizzarli dal punto di vista funzionale. Identificare tutti i geni di un organismo e assegnare a essi una funzione è uno degli aspetti iniziali dell'annotazione e rappresenta anche il più impegnativo visto che ancora oggi una frazione importante di geni non ha una funzione conosciuta.

È stata sviluppata una modalità standard per definire e assegnare le funzioni ai prodotti genici: il sistema di annotazione ***Gene Ontology*** (***GO annotation***). GO descrive le funzioni dei prodotti genici in modo completo considerando tre aspetti: il processo cellulare (*biological process*), la localizzazione cellulare (*cellular component*) e la funzione molecolare (*molecular function*). L'annotazione quindi consiste nell'associare al gene tre attributi (*GO attributes*) che

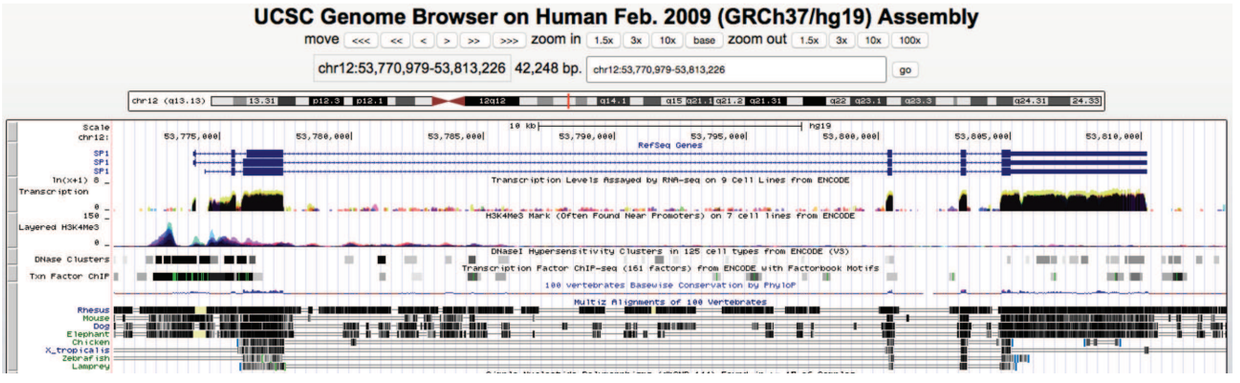


FIGURA 4.5 ▲ Esempio di una finestra con un browser genomico (https://genome.ucsc.edu/) che mostra il gene umano *SP1*.

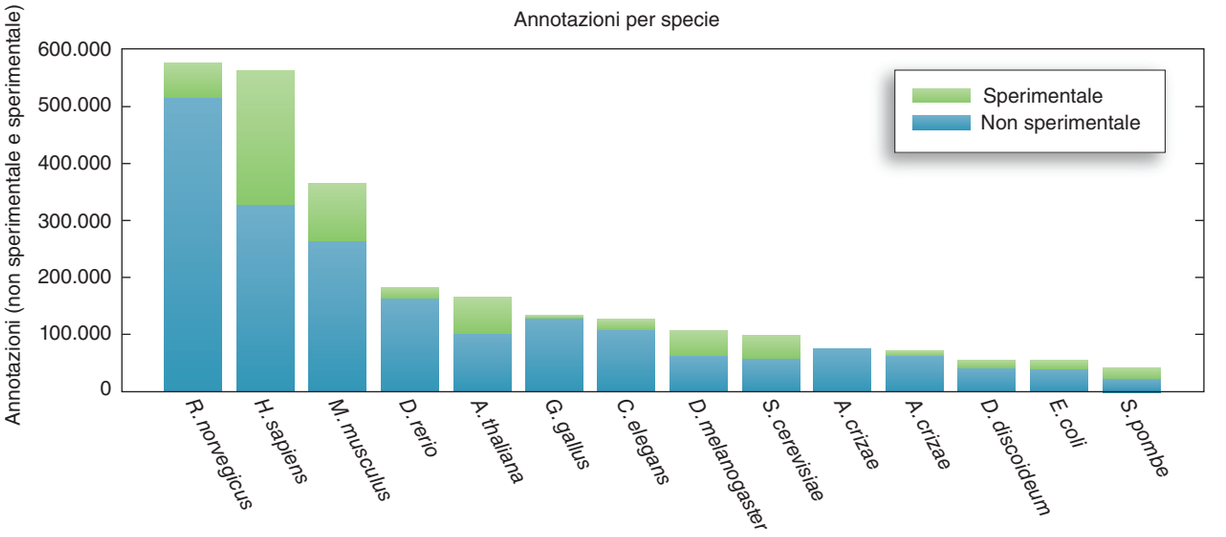


FIGURA 4.6 ▲ L'annotazione dei genomi si arricchisce di nuove informazioni continuamente. Questo era il numero di annotazioni GO, verificate sperimentalmente (verde) o solo bioinformaticamente (blu) a Gennaio 2016. Le info aggiornate possono essere ottenute visitando il sito http://geneontology.org/page/current-go-statistics. *R. norvegicus*, *H. sapiens* e *M. musculus* sono i genomi con maggiori annotazioni.

descrivono tre caratteristiche del prodotto genico, rispettivamente: il processo metabolico in cui è coinvolto, la sua localizzazione nei comparti intracellulari e la sua attività o funzione a livello molecolare. La struttura di GO è complessa in quanto in ognuna delle tre categorie gli attributi sono organizzati gerarchicamente per permettere una descrizione da molto generale a molto specifica delle tre caratteristiche del gene. Un Consorzio internazionale gestisce GO e il suo sito di riferimento (http://geneontology.org) e, ad oggi, il numero maggiore di annotazioni è per *R. norvegicus* (**Figura 4.6**). Un pregio del sistema GO è che l'annotazione funzionale dei geni è indipendente dalla specie e riguarda soltanto le funzioni fisiologiche dei prodotti genici, non includendo quindi eventuali "funzioni" patologiche. Un tutorial è disponibile online per approfondimenti su GO (**http://www.geneontology.org/teaching_resources/tutorials/2003_MBL_jblake.pdf**).

La disponibilità di numerosi genomi di specie diverse permette lo studio delle similitudini strutturali dei genomi, un'operazione denominata **genomica comparata**. Questi studi principalmente confrontano la struttura primaria di DNA, RNA e proteine, e si possono successivamente estendere anche al confronto delle loro strutture secondarie. Una significativa similitudine di sequenza tra due geni spesso implica una relazione evolutiva che consiste nell'avere un comune gene ancestrale da cui sono derivati entrambi nel corso dell'evoluzione. In questi casi, i due geni sono detti **omologhi** anche se sono solo parzialmente correlati dal punto di vista funzionale. Due geni omologhi con stessa funzione in specie diverse sono detti **ortologhi**, mentre due geni omologhi ma distinti per funzione e/o struttura presenti all'interno dello stesso genoma sono denominati **paraloghi** (**Figura 4.7**). Geni paraloghi sono spesso presenti nei genomi degli eucarioti e si formano per eventi di duplicazione genica durante l'evoluzione. Geni ortologhi spesso conservano la stessa funzione in specie diverse e sono di grande aiuto quindi nell'annotazione di nuovi genomi sequenziati (si veda sopra). Il confronto tra genomi di specie molto distanti è spesso più difficoltoso, mentre il confronto tra genomi di specie evolutivamente vicine permette anche un'utile analisi di regioni genomiche molto più lunghe di un gene. In effetti, confrontando per esempio i genomi umano e murino è stato possibile stabilire che spesso l'ordine dei geni lungo il genoma si è conservato durante l'evoluzione. La **sintenia**, cioè la conservazione dell'ordine dei geni, fornisce un altro elemento a sostegno della relazione di omologia tra geni che si trovano in posizioni simili in genomi di specie diverse. La genomica comparativa può fornire pertanto importanti elementi strutturali e funzionali per una dettagliata annotazione dei genomi. L'annotazione dei genomi è in continuo aggiornamento e include anche altre informazioni funzionali, relativamente alle strutture proteiche e della cromatina, alle funzioni trascrizionali e replicative di specifiche regioni, e a altri aspetti funzionali sia fisiologici che patologici.

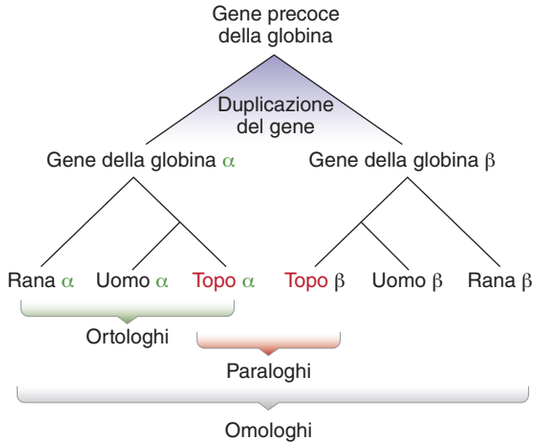


FIGURA 4.7 ◀ La formazione i nuovi geni avviene spesso in due passaggi: duplicazione del gene; successiva diversificazione per cambiamenti della sequenza nucleotidica. Questo genera geni *paraloghi* (geni con funzioni diversificate nella stessa specie) e geni *ortologhi* (gene con stessa funzione in specie diverse). Tutti questi geni sono tra loro *omologhi* (geni con funzioni simili che sono derivati evolutivamente da un comune gene ancestrale).

[4.3] ANATOMIA DEL GENOMA PROCARIOTICO

[4.3.1] Struttura dei genomi procariotici

I genomi procariotici sono distinti da quelli degli eucarioti per una serie di caratteristiche. In generale, la lunghezza del genoma dei procarioti è molto inferiore a quello eucariotico anche se vi è una qualche sovrapposizione tra il genoma procariotico più grande e quello eucariotico più piccolo. Altre differenze evidenti tra i due tipi di genoma sono il numero dei geni e l'organizzazione fisica. Attualmente, le migliaia di genomi sequenziati di *Archaea* e batteri e le annotazioni degli stessi permettono di avere un quadro più completo dell'organizzazione dei genomi procariotici. Nello stesso tempo, tutte le nuove informazioni accumulate finora hanno cambiato la visione tradizionale che, nell'era pre-genomica, era basata sostanzialmente sul genoma di *E. coli* (**Figura 4.8**). Il ceppo K12 di questo batterio, usato in laboratorio e non patogenico per l'uomo, ha un genoma costituito da un singolo DNA circolare (a volte chiamato **cromosoma batterico**) di circa 4,64 Mbp con 4466 geni (per aggiornamenti si veda il sito **UCSC *Genome Browser*** http://microbes.ucsc.edu/cgi-bin/hgGateway?db=eschColi_K12). Il DNA di *E. coli* ha una sola origine di replicazione da cui parte la sintesi del DNA prima della divisione cellulare. Inoltre, le cellule di *E. coli* possono contenere dei **plasmidi**, DNA circolari indipendenti dal genoma che si replicano autonomamente. I plasmidi contengono geni non essenziali per la cellula ma che possono essere utili in determinate circostanze ambientali, per esempio per la presenza di geni che conferiscono resistenza ad antibiotici. Sia il cromosoma batterico che i plasmidi sono contenuti in una zona della cellula, chiamata nucleoide, dove la lunga molecola di DNA è organizzata in anse (*loops*) di 100 kbp circa da superavvolgimenti della doppia elica e dalle proteine HU (strutturalmente distinte dagli istoni eucariotici) che hanno un ruolo architetturale e permettono una struttura più compatta della molecola circolare di DNA.

La conoscenza di molti altri genomi batterici ha dimostrato che il genoma è sostanzialmente diverso tra specie batteriche distinte e spesso non è costante all'interno di una singola specie. Per esempio, il ceppo O157 di *E. coli,* decisamente patogeno per l'uomo, ha un genoma di 5,53 Mbp quindi più lungo del ceppo non patogeno K12 (**Tabella 4.1**). Il DNA aggiuntivo di O157 è distribuito in quasi 200 regioni diverse del genoma, denominate "isole O" che contengono 1387 geni specifici, molti dei quali codificano per tossine o proteine coinvolte nella virulenza del ceppo O157. Il ceppo K12, a sua volta, contiene 528 geni specifici distribuiti in "isole K", la cui assenza nel ceppo O157 potrebbe contribuire alla sua virulenza. Pertanto, sia il ceppo K12 che O157 contengono geni specifici, e la lunghezza totale delle isole K e O è, rispettivamente, di 0,53 e 1,34 Mbp che corrispondono a un 11,4% e 24,2% dei rispettivi genomi. Queste percentuali sono molto al di sopra delle variazioni accettate per individui della stessa specie di organismi superiori e quindi richiedono una nuova definizione di cosa sia una specie nel caso dei microrganismi.

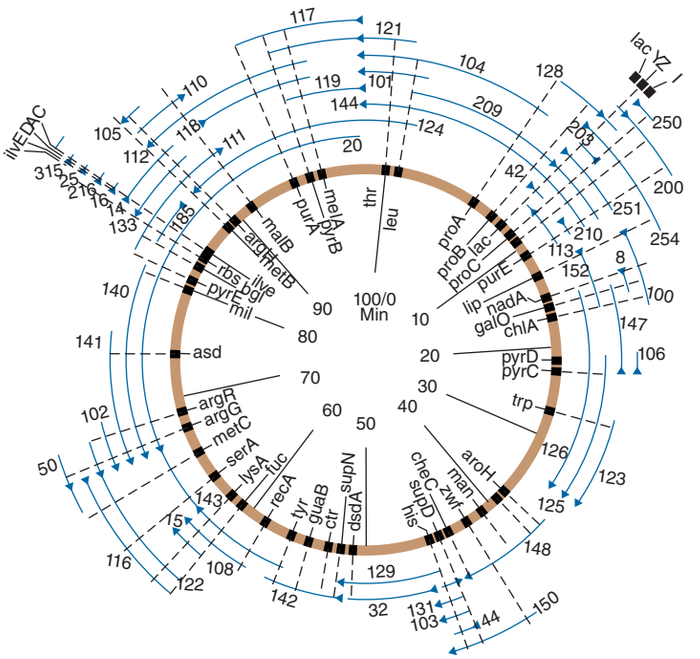


FIGURA 4.8 ▶ Mappa del genoma di *E. coli*. È mostrato il genoma del ceppo K-12 comunemente usato in laboratorio. L'utilizzo del tempo in minuti come unità di misura del genoma (all'interno del cerchio) è dovuto al tempo impiegato (100 minuti circa) per il trasferimento dell'intero cromosoma batterico dalla cellula donatrice a quella ricevente nella coniugazione batterica. In questa mappa, l'origine di replicazione è a 85 minuti circa; con la sigla e il colore blu sono indicati alcuni geni. Esternamente, i numeri e le frecce indicano specifici ceppi donatori Hfr. Per ogni ceppo Hfr, l'apice della freccia indica il punto iniziale del trasferimento di DNA. Il frammento del genoma batterico presente nel plasmide F′ è indicato dall'arco corrispondente al segmento della mappa circolare di *E. coli*.

TABELLA 4.1 ▼ **Organizzazione e dimensioni di alcuni genomi procariotici**

Specie	*Tipologia*	*Cromosomi*	*Plasmidi*	*Lunghezza totale media (Mbp)*	*Geni codificanti proteine*	*GC mediana (%)*	*Codice RefSeq*
Aquifex aeolicus	Ipertermofilo	1 (circolare)	1	1,59	1526	43,3	NC_000918.1 NC_001880.1
Bifidobacterium bifidum	Probiotico	1 (circolare)	–	2,21	1705	62,7	NC_014638.1
Borreliella burgdorferi	Patogeno (malattia di lyme)	1 (lineare)	fino a 20	1,21	942	28,3	NC_001318.1 NC_001850.1
Agrobacterium fabrum	Patogeno vegetale	1 (lineare) 1 (circolare)	2	5,67	5355	59,1	NC_003062.2 NC_003063.2 NC_003065.3 NC_003064.2
Deinococcus radiodurans	Radio-resistente	2 (circolare)	2	3,24	3022	66,8	NC_001263.1 NC_001264.1 NC_000958.1 NC_000959.1
Vibrio cholerae	Patogeno (colera)	2 (circolare)	–	4,02	3590	47,5	NC_002506.1 NC_002505.1
Escherichia coli K-12	Enterico	1 (circolare)	–	4,64	4140	50,8	NC_000913.3
Escherichia coli O157	Enterico patogeno	1 (circolare)	2	5,60	5292	50,5	NC_002695.1 NC_002127.1 NC_002128.1
Mycobacterium tuberculosis	Patogeno (tubercolosi)	1 (circolare)	–	4,41	3906	65,6	NC_000962.3
Rickettsia prowazekii	Patogeno (tifo)	1 (circolare)	–	1,11	850	29	NC_000963.1
Neisseria meningitidis	Patogeno (meningite)	1 (circolare)	–	2,16	2050	51,8	NC_003112.2
Mycoplasma genitalium	Patogeno (infezioni genitali e respiratorie)	1 (circolare)	–	0,5797	484	31,7	NC_000908.2

La situazione diventa anche più complessa se consideriamo altri batteri e *Archaea* (**Tabella 4.1**). È sempre più chiaro che i plasmidi o altri tipi di molecole di DNA, che una volta si pensavano essere non indispensabili per la vita della cellula, in alcuni casi appaiono far parte del patrimonio genetico del microrganismo. Il batterio *Deinococcus radiodurans* è caratterizzato da un alto grado di resistenza alle radiazioni ionizzanti e presenta i suoi geni essenziali distribuiti tra due cromosomi circolari e due plasmidi. Anche *Vibrio cholerae*, l'agente causale del colera, ha due molecole di DNA circolari relativamente grandi (2,96 e 1,07 Mbp) che contengono i geni essenziali per le funzioni fondamentali della cellula e di patogenicità. Un'analisi più approfondita ha però evidenziato che il cromosoma più piccolo contiene delle caratteristiche proprie dei plasmidi, come la presenza di regioni di DNA e geni codificanti per proteine che permettono al plasmide di catturare geni da altri plasmidi o da batteriofagi. Quindi, il secondo cromosoma di *V. cholera*e appare come un superplasmide acquisito durante l'evoluzione del batterio da antichi elementi di DNA e altri plasmidi. In altri tipi di batteri, il genoma appare essere costituito da cromosomi lineari e decine di plasmidi sia circolari che lineari (si veda la **Tabella 4.1** per degli esempi).

Queste novità, emerse dal sequenziamento di molti genomi batterici, rientrano in un quadro di complessità maggiore rispetto alla visione dell'era pre-genomica ma probabilmente indicano anche la necessità di un cambio di paradigma per capire meglio i genomi dei microrganismi.

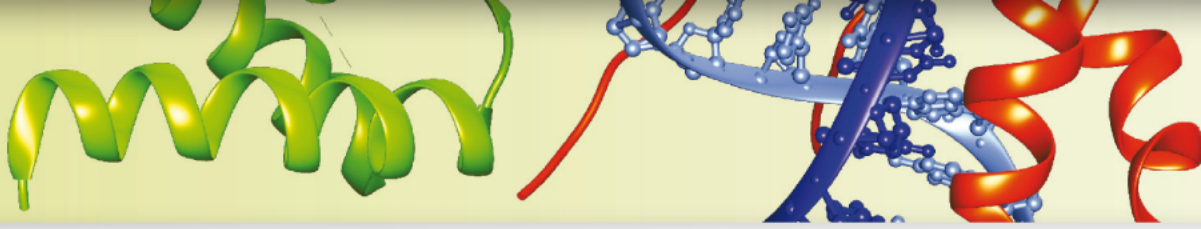
Anche se era già noto da osservazioni dell'epoca pre-genomica, l'entità del **trasferimento genico orizzontale o laterale (LGT)** evidenziata dalle sequenze dei genomi è molto maggiore rispetto alle attese (**Figura 4.9**). Nei microrganismi, il patrimonio genetico non si trasferisce solo da una cellula madre alle cellule figlie (**trasferimento genico verticale**) ma anche tra cellule di una popolazione batterica e anche tra cellule di specie diverse (trasferimento genico orizzontale).

Questo è radicalmente diverso dai meccanismi del flusso genico negli organismi superiori, dove il trasferimento orizzontale è molto raro. Le modalità del trasferimento orizzontale sono tre: **trasformazione**, **coniugazione** e **trasduzione**. Nella trasformazione, il batterio riceve DNA dall'ambiente quando questo si sia reso strutturalmente compatibile con il trasferimento stesso; nella coniugazione, il trasferimento è attivo e mediato da certi plasmidi, e avviene in seguito a un contatto tra due cellule batteriche; nella trasduzione, il trasferimento genico è mediato da batteriofagi infettanti la cellula batterica. La coniugazione in *E. coli*, ben conosciuta anche nell'era pregenomica, evidenzia bene come i meccanismi di LGT determinano fortemente la plasticità dei genomi batterici. In *E. coli* esiste il plasmide F che attivamente induce l'evento di coniugazione tra una cellula donatrice F^+ (quella ospitante il plasmide F) ed una ricevente F^- (che è priva del plasmide F).

Il plasmide F può anche esistere come DNA integrato nel cromosoma batterico e il sito di integrazione varia molto tra i vari ceppi batterici. Il plasmide F integrato produce ceppi di *E. coli* donatori particolarmente capaci di avviare la coniugazione e ricombinare il DNA, noti come ceppi **Hfr** (*High-Frequency Recombination*). Le cellule Hfr avviano il trasferimento partendo dal DNA del plasmide F integrato e possono pertanto trasferire anche parte del cromosoma batterico, che potrà ricombinarsi con il genoma della cellula ricevente (**Figura 4.8**). I plasmidi F′ sono plasmidi F che contengono frammenti di cromosoma batterico. L'insieme di queste nuove conoscenze ha dunque evidenziato una straordinaria plasticità del genoma batterico per cui le dimensioni possono variare in modo significativo tra individui della stessa specie, e dunque a una parte comune del genoma si aggiunge una parte specifica e variabile tra ceppi diversi.

Inoltre, poiché i geni possono essere trasmessi anche orizzontalmente tra microrganismi, questo implica che la storia evolutiva di un qualunque singolo gene potrebbe differire significativamente dalla storia evolutiva dell'intero organismo. Ciò ovviamente ha delle conseguenze nella costruzione di alberi filogenetici dei microrganismi che deve inevitabilmente tener conto del genoma nel suo complesso.

Una modalità organizzativa molto particolare di certi geni sono le **sequenze di inserzione** (**IS**) che sono dei veri e propri **elementi mobili di DNA**. È come se questi geni si fossero organizzati durante l'evoluzione per muoversi da un posto all'altro del genoma e perfino da un genoma all'altro di specie diverse. Una IS è costituita da una o due ORF codificanti per trasposasi/resolvasi e fiancheggiate a monte e a valle da piccole ripetizioni invertite (**IR**, *small terminal Inverted Repeats*). Le trasposasi sono enzimi codificati dalla IS e in grado di tagliare la stessa IS e inserirla in un altro punto del genoma interagendo specificamente con le sequenze



Biologia Molecolare

Accedi all'**ebook** e ai **contenuti digitali** › **Espandi** le tue risorse › con un libro che **non pesa** e si **adatta** alle dimensioni del tuo **lettore**

All'interno del volume il **codice personale** e le istruzioni per accedere alla versione **ebook** del testo e agli ulteriori servizi. L'accesso alle risorse digitali è **gratuito** ma limitato a **18 mesi dalla attivazione del servizio**.

EdiSES
www.edises.it

ISBN 978-88-7959-916-0

€ 55,00