Dal *Progetto Genoma Umano* al *third generation sequencing:* la nascita del *genome wide* e le sue ricadute nella pratica clinica

Martina Collotta*
Candiolo Cancer Institute - FPO, IRCCS, Candiolo, Torino

Secondo articolo della serie *dove sta andando la ricerca in campo biomedico* per saperne di più sulle “Scienze Omiche”, dalle loro origini fino alle sempre più frequenti e importanti applicazioni nella pratica clinica. Come nel precedente articolo, un Glossario ci aiuterà a fissare concetti e definizioni che per il clinico possono essere poco consueti.

Sequenziamento e scienze omiche: l’approccio *genome wide*

Quando, negli anni ’90, venne ideato il Progetto Genoma Umano (*Human Genome Project,* HGP), non ci si immaginava nemmeno che, poco più di vent’anni dopo, il sequenziamento di un intero genoma avrebbe richiesto meno di un giorno.
Ed è proprio la velocità di sequenziamento, allora impensabile, che ha permesso la nascita delle scienze omiche. Non solo è diventato possibile conoscere la sequenza di basi che costituiscono gli acidi nucleici, ma questo procedimento è ora relativamente veloce ed economico.
Gli approcci *omici*, che partono dall’assunto che un sistema complesso possa essere meglio compreso quando analizzato nel suo insieme, necessitano di una mole di dati riguardanti l’intero genoma, forniti, tra le altre tecniche, proprio dal sequenziamento, di cui è debitore l’approccio omico *genome wide*.

Nascita ed evoluzione del sequenziamento

Sappiamo che, nella cellula eucariotica, sono presenti due genomi: il ben noto genoma *nucleare* e il genoma *mitocondriale*, per lo più codificante proteine coinvolte nella respirazione cellulare che avviene in essi.
Il primo genoma umano di cui si ottenne la sequenza completa fu proprio quello *mitocondriale*, di circa 16 mila coppie di basi. Era il 1981, e questo rappresentava un risultato straordinario. Ma il genoma *nucleare* umano, con i suoi 3 miliardi di paia di basi, è circa 200.000 volte più grande!
Questo ha richiesto sviluppi nelle tecnologie di sequenziamento del DNA che hanno portato dal *sequenziamento Sanger* (vedi oltre), alle tecniche di *seconda* e *terza generazione*.
Parallelamente agli avanzamenti nelle tecnologie sperimentali di sequenziamento, è avvenuto il grande sviluppo della *bioinformatica,* che ha permesso di analizzare l’enorme mole di dati provenienti dal sequenziamento dei genomi.

Il *Progetto Genoma Umano* (HGP)

La genomica moderna ha inizio con l’HGP, cominciato nel 1991 e con una durata prevista di 15 anni.
L’HGP fu condotto da un consorzio internazionale (*International Human Genome Sequencing Consortium,* IHGSC), comprendente istituzioni di Stati Uniti, Gran Bretagna, Francia, Germania e Cina. Accanto a questo consorzio pubblico, la Celera Genomics, con i suoi laboratori privati, iniziò la medesima impresa.
Scopo dell’HGP era mettere a disposizione, attraverso banche dati pubbliche, la sequenza di basi che costituiscono il genoma umano, identificandone i geni; scopo raggiunto con due anni di anticipo rispetto al previsto, grazie allo sviluppo di nuovi mezzi tecnici.
Nel 2000 furono presentati, congiuntamente, i primi risultati dell’IHGSC e della Celera Genomics; nel 2003 venne presentata la versione “definitiva”, con un tasso di errore inferiore a 1/10.000 paia di basi e una copertura del genoma del 99%.

* Martina Collotta, laureata in Medicina e Chirurgia con lode e menzione speciale presso l’Università di Milano con una tesi sperimentale nel settore dell’epigenetica, si dedica allo studio della biologia molecolare. Attualmente, dopo l’esperienza acquisita nel campo della genetica delle malattie neuromuscolari, frequenta il Master in Molecular Biotechnology presso l’Università di Torino e svolge attività di ricerca in oncologia molecolare presso l’IRCCS FPO di Candiolo (Torino).

Dalla *genomica strutturale* alla *genomica funzionale*

Di *definitivo*, ovviamente, non c'è nulla! La nostra conoscenza del genoma è in continua crescita, *le annotazioni al DNA* sono un continuo *work in progress.*

Annotare, significa in questo caso, segnalare per ciascun gene la struttura, le mutazioni note, la proteina da esso codificata, i percorsi *(pathway)* in cui essa è coinvolta.

Non basta, infatti, fermarsi al sequenziamento e alla mappatura genetica, ambiti della *genomica strutturale*, ma è necessario arrivare alla comprensione delle modalità con cui i geni dirigono lo sviluppo e il funzionamento del nostro organismo (*genomica funzionale*).

Ad esempio, il progetto ENCODE (*ENCyclopedia Of Dna Elements*), del 2007, ha permesso di identificare i geni codificanti e non codificanti proteine, e gli altri elementi funzionali contenuti nella sequenza del DNA.

La genomica funzionale ha quell'approccio olistico proprio delle scienze omiche: lo studio *simultaneo* di geni e prodotti genici che, talvolta, porta a seguire un *percorso inverso* rispetto a quello tradizionale.

"Inverso" in quanto l'ipotesi non viene verificata *dopo* essere stata formulata a priori, ma i dati vengono raccolti e analizzati e, dall'osservazione globale dei risultati, vengono generate ipotesi di correlazioni, da validare attraverso nuovi dati o esperimenti.

È possibile, ad esempio, cercare un gene "sconosciuto" partendo da ciò che si sa circa un gene analogo presente in organismi differenti, comunemente usati in laboratorio (dalla *Drosophila melanogaster*, il moscerino della frutta, al topo). Confrontando la sequenza del gene noto con la sequenza di basi del genoma umano, è possibile identificare dei *geni candidati* che potrebbero, nell'uomo, svolgere un'analoga funzione (ad esempio la codifica di una proteina coinvolta nel processo metabolico in esame). Sono questi "geni candidati" a costituire l'ipotesi da verificare sperimentalmente, per arrivare a trovare il gene codificante la proteina di interesse, o il gene coinvolto in una certa funzione cellulare o, ancora, implicato nella patogenesi di una malattia.

Infatti, per raggiungere lo scopo della genomica funzionale, ovvero fornire le *annotazioni funzionali* che vanno ad arricchire i dati a disposizione circa il genoma umano, spesso i ricercatori si avvalgono dell'ausilio della *genomica comparativa*, branca della genomica funzionale che confronta genomi di diverse specie.

Anche questo approccio è stato permesso dall'abbattimento dei costi e dei tempi necessari per il sequenziamento degli acidi nucleici: se, infatti, il solo HGP aveva richiesto anni di sforzi internazionali, come si sarebbe potuto pensare di sequenziare genomi di altre specie, al solo scopo di effettuare confronti?

L'evoluzione delle tecnologie di sequenziamento, dunque, ha importanti ricadute per la ricerca di base e clinica, proprio attraverso la genomica funzionale, che non si ferma alla conoscenza della funzione dei geni, ma s'interroga anche sul ruolo del loro malfunzionamento nell'indurre uno stato patologico.

Sequenziamento dei genomi: dal *Sanger* al *third generation sequencing*

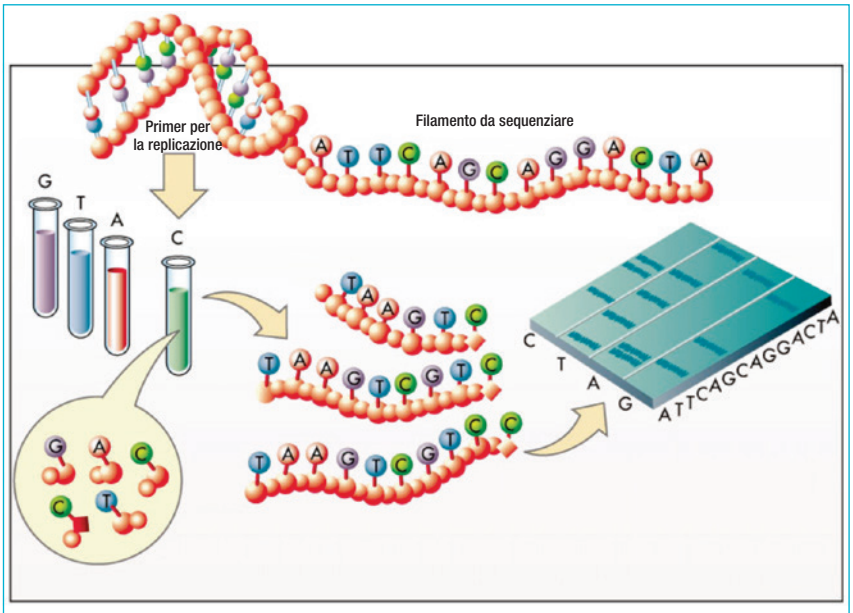
Due diversi approcci sono stati utilizzati nel corso dell'HGP per sequenziare il genoma umano.

L'approccio dell'IHGSC prevedeva la suddivisione del genoma in diversi segmenti, minimizzandone la sovrapposizione, in base a una mappa predisposta inizialmente. I frammenti venivano poi clonati in batteri vettori capaci di replicare il DNA per poterne avere quantità sufficienti per il sequenziamento. L'allineamento delle sequenze avveniva poi per mezzo di *software* dedicati.

La seconda tecnica, quella utilizzata dalla Celera Genomics, il *whole-genome shotgun*, prevedeva di frammentare il genoma in modo casuale, con parziali sovrapposizioni, in modo da permettere la ricostruzione della sequenza con analisi informatiche.

Ma come fare per conoscere, base per base, la sequenza nucleotidica?

FIGURA 1.
Sequenziamento Sanger.



Il DNA a singolo filamento viene mixato con un primer e diviso in 4 aliquote, ciascuna contenente la DNA polimerasi (enzima che sintetizza un filamento di DNA complementare a partire da un filamento stampo, permettendo la replicazione del DNA), il primer (una breve sequenza nucleotidica che permette alla polimerasi di iniziare la sintesi del filamento complementare), i 4 desossiribonucleotidi trifosfati (marcati radioattivamente o per fluorescenza) e un terminatore della replicazione (un nucleotide modificato in modo da impedire il legame fosfodiesterico con il nucleotide successivo). Ciascuna reazione di replicazione procede fino a che il nucleotide terminatore viene aggiunto e questo avviene, casualmente, in differenti momenti, in modo da generare filamenti complementari di lunghezza diversa. Le aliquote vengono poi caricate su un gel per elettroforesi in 4 corsie, ciascuna corrispondente a ognuno dei 4 nucleotidi marcati; il DNA può essere così visualizzato sotto forma di bande a distanze differenti secondo la diversa lunghezza.

La metodica di sequenziamento più utilizzata nel passato è stata quella di *Sanger* (Fig. 1), basata sul *sequencing-by-synthesis*, ovvero sul sequenziamento attraverso la sintesi di un filamento di DNA complementare a quello in esame, con nucleotidi capaci di interromperne la sintesi, marcati per poter essere "letti" da appositi strumenti e identificare così, base per base, al livello dell'interruzione, la sequenza dell'acido nucleico.

Il grande passo avanti fu fatto nel 2005 con il *Next Generation Sequencing* (*NGS*, noto anche come *Second Generation Sequencing*), attraverso cui è diventato possibile sequenziare un intero genoma umano in circa una settimana, a un costo di poche migliaia di dollari.

Tuttavia, la metodica NGS ha il difetto di avvalersi di piattaforme che si basano su sistemi estremamente complessi, con costi di esercizio ancora elevati (seppur di molto inferiori a quelli del sequenziamento Sanger), tempi di analisi ancora troppo lunghi e tassi di errore non trascurabili.

Nei sistemi attuali *(Third Generation Sequencing)* si è cercato di eliminare o ridurre le fasi intermedie di manipolazione degli acidi nucleici, causa di imprecisione nei risultati, puntando su tecnologie a elevata sensibilità che consentano di analizzare singole molecole di DNA o RNA.

Il *Third Generation Sequencing* include tecnologie capaci di rilevare un segnale elettronico invece che luminescente per identificare le singole basi, sistemi a *nanopori* attraverso cui passa una sola molecola di DNA alla volta, sistemi di sequenziamento di singola molecola e di sequenziamento in tempo reale *(real time)*, il tutto a costi relativamente contenuti (l'obiettivo è il sequenziamento di un genoma a 100 dollari).

Idealmente una metodica di sequenziamento perfetta dovrebbe permettere di sequenziare direttamente anche una sola molecola di acido nucleico, utilizzando quantità anche minime di DNA o RNA e con tassi di errore esigui. Costi contenuti e accessibilità delle tecnologie, infine, sono altri requisiti che non possono essere trascurati.

Sequenziamento degli acidi nucleici: l'importanza clinica

La velocità di sequenziamento ora disponibile e i costi abbattuti, permettono di condurre studi di popolazione su larga scala, includendo l'analisi del genoma per individuare le associazioni genotipo-fenotipo.

Nascono così gli studi di associazione *genome wide* (GWAS, *Genome Wide Association Studies*), in cui vengono ricercati i *polimorfismi di singolo nucleotide* (SNP, *Single Nucleotide Polymorphism*), ovvero quelle variazioni da individuo a individuo, che riguardano una singola base del DNA e che sono associate al fenotipo malattia in maniera più complessa rispetto, ad esempio, all'ereditarietà mendeliana autosomica dominante (Fig. 2).

Si tratta, infatti, di una relazione causale che necessita di altri fattori per dare luogo al fenotipo malattia: concomitanti alterazioni in geni differenti o concause ambientali e comportamentali.

È opportuno dunque parlare di una *predisposizione* su base genetica alla malattia, piuttosto che di una causa del tipo "tutto o nulla".

Molte delle malattie croniche che sono tutti i giorni sotto gli occhi del clinico, dal diabete tipo 2 alle malattie cardiovascolari, possono essere associate a polimorfismi del DNA identificati proprio attraverso studi GWAS.

È evidente che per poter associare uno SNP a una patologia, è necessario avere un campione numeroso di cui si abbia una caratterizzazione molecolare ben precisa (precisa al singolo nucleotide!) ottenuta attraverso il sequenziamento.

Ad esempio, il progetto HapMap, è nato nel 2002 proprio con lo scopo di identificare geni associati a patologie croniche e alla variazione individuale nella risposta ai rischi ambientali e ai farmaci.

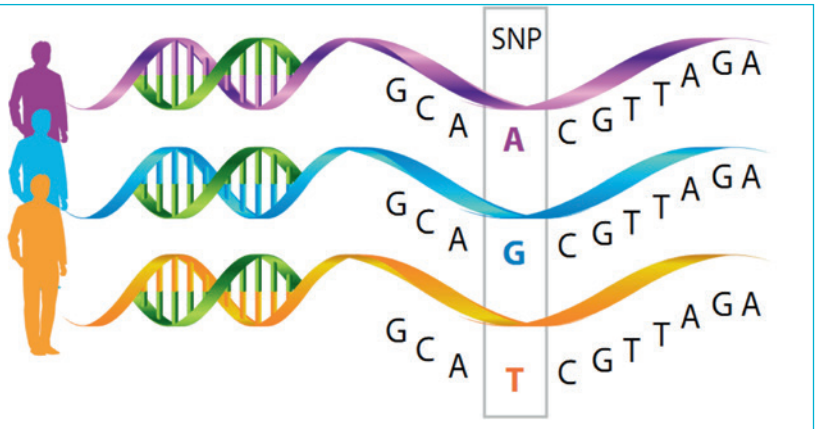
Nello studio i soggetti presentavano molte differenti sfumature di patologia, frequenze dei polimorfismi non elevate e, soprattutto, di singola base!

Non è dunque possibile prescindere né dalla velocità, né dall'accuratezza e sensibilità delle tecniche di sequenziamento, se si vuole condurre questo tipo di studi su malattie poligeniche e multifattoriali.

Anche il sequenziamento di acidi nucleici della *singola cellula* ha in sé potenzialità rilevanti per la clinica. Pensiamo, ad esempio, all'eterogeneità di un tumore: i cloni cellulari che lo compongono sono differenti tra loro dal punto di vista fenotipico, proprio perché differiscono a livello molecolare.

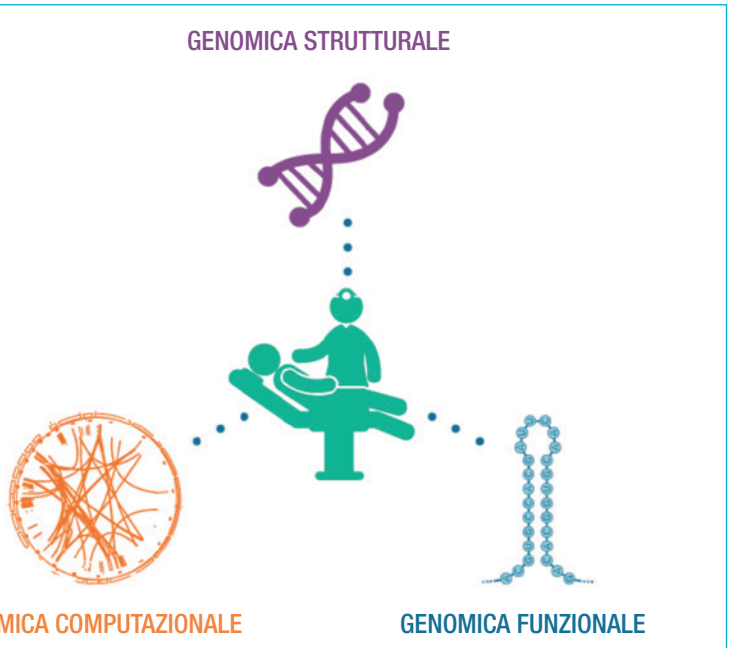
Il fenotipo "risposta alla terapia" o il fenotipo "aggressività", dipendono da mutazioni del DNA o da alterazioni dell'espressione genica che possono essere identificate attraverso le tecniche di sequenziamento di terza generazione,

FIGURA 2.
Single Nucleotide Polymorphism.



Gli SNP *(Single Nucleotide Polymorphism)* sono delle variazioni della sequenza del DNA dei diversi individui della popolazione, a carico di un singolo nucleotide. Nell'esempio della Figura, il nucleotide alla quarta posizione, varia tra i diversi soggetti, i quali presentano, rispettivamente, un'adenina, una guanina e una timina. Tali variazioni possono essere associate a fenotipi di rilevanza clinica, permettendo così di stabilire associazioni con la predisposizione allo sviluppo di patologie.

FIGURA 3.
La genomica al servizio del paziente.



Le tre branche della genomica, così come si è evoluta grazie alle nuove tecniche di sequenziamento, possono aiutare a comprendere meglio le patologie di cui sono affetti i singoli pazienti, dare a esse una caratterizzazione molecolare, sviluppare terapie individualizzate. Dalla genomica strutturale, che descrive come è fatto il genoma, in termini di sequenza nucleotidica e geni, passiamo al come funziona la genomica funzionale, grazie ai dati elaborati dalla genomica computazionale (vedi glossario), che individua correlazioni tra il genoma del paziente, il genoma delle cellule malate (ad es. tumorali) e il genoma di riferimento, al fine di generare ipotesi da verificare sperimentalmente o validare ipotesi precedentemente formulate.

capaci di descrivere le sottopopolazioni che compongono un medesimo tumore, attraverso la caratterizzazione molecolare cellula per cellula.

Prospettive future

Quali, dunque, le prospettive future?
Non è più considerato così utopico avere a disposizione il genoma del singolo paziente per ottenere il quadro molecolare della sua patologia o per individuare precocemente, sulla base del suo genoma, la sua predisposizione a sviluppare una certa malattia (con tutti i problemi etici annessi) e, ancora, grazie alla farmacogenomica, a calibrare la terapia sul singolo paziente.
L'interdisciplinarietà sembra essere il futuro della genomica: la biologia ha incontrato l'informatica e l'ingegneria, per fornire risposte sempre più accurate e con implicazioni cliniche sempre più rilevanti. Tanto più veloci, a basso costo e automatizzate, diventeranno le tecniche di sequenziamento, tanto più la diagnostica molecolare prenderà spazio nella pratica clinica, portando i risultati della ricerca al letto del malato.
Appuntamento alla prossima puntata!

Bibliografia di riferimento

Dulbecco R. *A turning point in cancer research. Sequencing the human genome*. Science 1986;231:1055-6.

McPherson JD, Marra M, Hillier L, et al. *A physical map of the human genome*. Nature 2001;409:934-41.

Sachidanandam R, Weissman D, Schmidt SC et al. *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms*. Nature 2001;409:928-33.

Venter JC, Adams MD, Myers EW, et al. *The sequence of the human genome*. Science 2001;291:1304-51.

GLOSSARIO

Genoma di riferimento	Sequenza di DNA identificata dal *Progetto Genoma Umano*, su cui sono stati mappati i geni e attraverso cui è possibile confrontare ogni altro genoma umano sequenziato.
Genoma mitocondriale	DNA contenuto nei mitocondri; nell'uomo il DNA mitocondriale consta di circa 16.000 paia di basi e 37 geni (che codificano per 13 polipeptidi sintetizzati dal ribosoma mitocondriale, 22 tRNA e 2 rRNA), coinvolti nella produzione di proteine necessarie alla respirazione cellulare.
Genoma nucleare	DNA contenuto nel nucleo della cellula; nell'uomo è costituito da circa 3 miliardi di paia di basi di DNA e 20.000-25.000 geni.
Genome wide	"Di ampiezza genomica", "esteso all'intero genoma". È il "punto di vista" delle scienze omiche, non più limitato a studiare gene per gene la struttura e la funzione del DNA, ma capace di guardare complessivamente l'intero patrimonio genetico della cellula o dell'organismo.
Genomica comparativa	Branca della biologia molecolare che studia il DNA umano, confrontandolo con informazioni ricavate dallo studio del genoma appartenente ad altre specie.
Genomica computazionale	Studio del genoma attraverso l'uso di analisi statistiche e computazionali per decifrare la biologia dalle sequenze di DNA e RNA, per comprendere, ad esempio, la funzione e l'espressione dei geni o l'associazione tra un determinato polimorfismo del DNA e una patologia (Fig. 3).
Genomica funzionale	Studio delle modalità con cui i geni dirigono lo sviluppo e il funzionamento della cellula o dell'organismo e di come il loro malfunzionamento possa essere causa di malattia.
Genomica strutturale	Studio degli aspetti strutturali dei geni, come sequenziamento e mappatura (localizzazione nel genoma, localizzazione cromosomica).
GWAS (*Genome Wide Association Study*)	Studio di epidemiologia genetica in cui vengono indagati i geni di diversi individui di una popolazione per determinare le variazioni geniche tra essi. In seguito si tenta di associare le differenze osservate con alcuni tratti fenotipici, in particolare patologie.
Next generation sequencing* (NGS) – *Second generation sequencing	Serie di tecnologie che permettono di sequenziare grandi genomi in un tempo inferiore rispetto al sequenziamento Sanger. Comprende mezzi ottici capaci di "leggere" la fluorescenza dei diversi nucleotidi marcati, come il *pirosequenziamento*.
Sequenziamento	Processo attraverso cui viene determinato l'ordine dei nucleotidi che costituiscono l'acido nucleico (DNA o RNA).
Sequenziamento *Sanger*	Vedi Figura 1.
SNP (*Single Nucleotide Polymorphism*)	Vedi Figura 2.
Third Generation Sequencing	Comprende le tecnologie più recenti (e ancora in fase di sviluppo) per sequenziare il DNA. Il *third generation sequencing* comprende mezzi tecnici in capaci di sequenziare single molecole di acidi nucleici, tecnologie in grado di leggere sequenze più lunghe rispetto al NGS, e con velocità superiori.