Trascrittomica: la differenza che *conta*

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

In questo numero affronteremo il tema dell’analisi dei trascrittomi, un utile strumento per la descrizione del comportamento molecolare di cellule e tessuti, in condizioni fisiologiche e/o patologiche.
L’analisi dell’*espressione genica differenziale*, infatti, risulta essere un valido ausilio per la pratica clinica, per identificare cambiamenti patologici a livello molecolare, ancor prima che si manifestino segni clinici, per predirre l’evoluzione della patologia, per avere un riscontro obiettivo della risposta o meno alla terapia.
Se gli aspetti che affronteremo vi sembreranno eccessivamente tecnici, sappiate che essi sono indispensabili per comprendere la letteratura medico-scientifica che approccia la clinica a partire da una prospettiva molecolare (*genomica*, *trascrittomica*, *proteomica*).
Conoscere il “come” dell’ottenimento dei risultati *omici*, permetterà di valutare gli stessi criticamente, oltre che di comprenderne maggiormente le importanti ricadute cliniche che riguardano un futuro immediato, se non, addirittura, il presente della clinica medica.

Contare la differenza che conta

Dal box di presentazione di questo articolo, avete certamente intuito che la branca delle scienze *omiche* che andremo ad analizzare è, come le altre, di non trascurabile rilevanza clinica.
Di fronte a una condizione patologica qualsiasi, quello che permette di definirla tale è proprio la *differenza* rispetto alla fisiologia, alla norma, all’ambito di valori di riferimento (per dirla con la nuova medicina statistica).
Quando si tratta di definire in *cosa* consista questa differenza, tuttavia, non è facile a dirsi.
Allo stato attuale delle cose, non possiamo, da clinici, limitarci all’oggettività dei segni (o all’oggettivizzazione, spesso forzata, dei sintomi), ma dobbiamo necessariamente ricorrere al dato, al *quantum*.
Le analisi di laboratorio, gli esami clinici strumentali corredati di parametri numerici oggettivi, l’*imaging* capace di quantificare l’intensità di un segnale… ma non solo. La medicina contemporanea, grazie allo sviluppo scientifico avvenuto in campo biomedico, permette di descrivere *differenze a livello molecolare* e, perfino, di *quantificare* tali differenze, ovvero *contare* la differenza!
Nel campo delle scienze omiche, lo abbiamo visto con la genomica, possiamo parlare di differenze in termini di mutazioni puntiformi, inserzioni e delezioni o, ancora, di modifiche del DNA che non riguardano la sequenza nucleotidica, ma l’epigenoma (modificazioni delle code istoniche, stato della cromatina ecc.).
Facciamo ora un passo avanti, ricordando il *dogma centrale della biologia molecolare*: guardiamo al *trascrittoma*, per capire come esso sia, almeno per ora, il migliore strumento attraverso cui descrivere e quantificare quella differenza tra fisiologico e patologico, tra il prima e il dopo (nelle fasi del ciclo cellulare, dello sviluppo o pre- e post-terapia), che deve essere oggettivata per raggiungere quella soglia di evidenza che permetta il *consensus* della comunità scientifica.

Un passo avanti nello studio delle *omiche*: la *trascrittomica*

Lo scopo dell’approccio olistico delle scienze *omiche*, ricordiamo, è quello di conoscere principi operativi biologici complessi, a partire da un approccio integrato, in cui i fattori vengono studiati nel loro insieme quali *pool* di molecole biologiche che fanno parte di una cellula o di un sistema cellulare o tissutale.
L’integrazione tra le scienze e le tecnologie omiche, nella cosiddetta *biologia dei sistemi complessi*, permette così di migliorare la comprensione del *sistema*, considerato come insieme delle molecole biologiche che lo compongono.
Trattandosi di una scienza *omica*, la *trascrittomica* affronta l’analisi degli RNA collettivamente, prendendo in considerazione tutto il *pool* di RNA cellulari.
Dal DNA, trascritto in RNA messaggeri (mRNA), giungiamo alle proteine attraverso il processo di traduzione. Fermandoci al primo *step*, quello della trascrizione, guardiamo al *pool* degli RNA intracellulari.
Il *trascrittoma* comprende l’insieme delle molecole di RNA presenti in una cellula di un dato tessuto in un dato momento. Possiamo guadare solo agli *mRNA*, che contengono in sé l’informazione genetica codificante le proteine, o anche a tutte le altre forme di RNA intracellulari che svolgono, per lo più, funzioni regolatorie (dagli RNA di trasporto o *tRNA*, a quelli ribosomiali o *rRNA*, fino ai forse meno noti *micro-RNA* o ai *long non-coding RNA*).
Possiamo dire che il trascrittoma costituisce una peculiarità della singola cellula in un determinato momento o condizione. L’espressione dei trascritti, infatti, si modifi-

ca a seconda delle condizioni dell'ambiente extra- e intra-cellulare.
L'oggetto di studio della *trascrittomica* è proprio questo: il trascrittoma cellulare o tissutale, insieme con le variazioni che esso subisce tra cellula e cellula o tessuto e tessuto, in seguito a mutamenti delle condizioni in cui la cellula si trova.

L'analisi dell'espressione genica differenziale

L'espressione genica tessuto-specifica determina il fenotipo morfo-funzionale dei tipi cellulari e tissutali.
In ogni cellula differenziata e in ogni particolare momento dello sviluppo è attivo solo un sottoinsieme di geni. In tutti gli organismi viventi, infatti, le informazioni contenute nel genoma non si esprimono contemporaneamente, ma sono finemente regolate.
I geni possono essere divisi, semplicisticamente, in tre categorie: quelli a espressione costitutiva (*housekeeping genes*), quelli a espressione *condizionale* (inducibili o reprimibili), geni *specializzati* (tessuto-specifici, stadio-specifici, che a loro volta possono essere costitutivi o condizionali).
L'attivazione o inattivazione dell'espressione genica negli eucarioti, avviene in base al differenziamento cellulare durante lo sviluppo, alla regolazione del ciclo cellulare, alla risposta a mediatori esterni (ormoni, fattori di crescita, citochine ecc.).
Per definire un gene *differenzialmente espresso* si tiene conto se la sua espressione genica si discosta dalla situazione di uguale espressione nei due stati in modo significativo, confrontando, ad esempio, con un valore soglia per definire se sovra- o sotto-espressi rispetto a tale valore di espressione.
Differenti trascrittomi non significano altro se non differenti *pool* di geni attivati e attivamente trascritti (per poi essere tradotti in proteine o per "restare" RNA a funzione regolatoria).
Quantificare il trascrittoma permette di comprendere quali geni siano attivati nelle diverse fasi del ciclo cellulare, dello sviluppo o in risposta a determinati segnali provenienti dalle altre cellule e dall'ambiente extracellulare.
Dall'approccio *one-gene* (di singolo gene), con le scienze *omiche*, siamo giunti all'approccio di tipo *large-scale* (su larga scala). Nel primo caso, ci si limitava unicamente a valutare se il gene di interesse era espresso o meno in un tessuto in un dato momento dello sviluppo e quanto era attivo dal punto di vista trascrizionale; nel secondo caso si considerano gli stessi parametri, ma prendendo in considerazione più geni simultaneamente, attraverso lo studio dei profili di espressione del genoma, o trascrittomi.
La *trascrittomica quantitativa* si occupa di effettuare un'analisi differenziale dell'espressione genica ottenuta confrontando i profili trascrizionali di due o più tessuti o del medesimo tipo cellulare in condizioni o momenti diversi (es. diversi per fase di sviluppo, fase del ciclo cellulare, salute o malattia).
L'analisi dei trascrittomi viene effettuata attraverso l'utilizzo di due differenti tecnologie: la prima, basata sull'ibridazione (*microarray*), la seconda, basata su tecnologie NGS (*Next-Generation Sequencing*).
Nel 1995 vennero sviluppati i primi *microarray* basati su *spotting* di molecole di cDNA, nel 2002 si passò a utilizzare i cosiddetti *high density oligo microarrays*; dal 2008 abbiamo a disposizione anche la tecnica di *RNA-Seq* che permette il sequenziamento degli RNA messaggeri attraverso tecniche di tipo NGS.

I *microarray*

Per quanto da molti i *microarray* (Fig. 1) siano considerati una tecnica quasi obsoleta, alla luce delle nuove tecnologie di RNA-Seq in tutte le loro possibili varianti, riteniamo utile presentare questa tecnica di ibridazione per facilitare la comprensione del concetto di analisi differenziale e per comprendere il perché del salto qualitativo (o meglio, quantitativo, e vedremo poi perché) permesso dal NGS applicato all'RNA.
I *microarray* sono costituiti da un supporto solido su cui sono "ancorate" delle sonde di DNA (*probe*) che occupano specifiche posizioni all'interno dell'*array*, in modo da poter essere identificate da coordinate spaziali. Un *microarray* di DNA (o *chip* a DNA) è un insieme di microscopiche sonde di DNA attaccate a una superficie solida (vetro, plastica o chip di silicio) formanti un *array* (matrice). Tali *array* permettono di esaminare simultaneamente la presenza di moltissimi geni all'interno di un campione di DNA.
Ogni sonda, presente in numerose copie all'interno dello stesso *spot* dell'array (il "punto" dove fisicamente la sonda è posizionata), codifica un gene (o una parte di esso) di cui s'intende valutare l'espressione.
L'RNA che costituisce il trascrittoma cellulare, viene trasformato in DNA (detto cDNA), tracciato con marcatori fluorescenti e ibridato con il *microarray*.
L'ibridazione consiste nell'appaiamento del cDNA con il DNA della *probe* ancorato all'*array*; la fluorescenza emessa in seguito all'appaiamento dei nucleotidi complementari dei due filamenti viene dunque quantificata: tanto maggiore la fluorescenza emessa, tanto maggiore è l'espressione del gene corrispondente alla *probe*, nella cellula in esame (quella da cui proviene il cDNA, ottenuto dalla retrotrascrizione dell'RNA in esame).
I limiti di questa tecnica sono numerosi: prima di tutto è necessario conoscere la sequenza dei geni di cui valutare l'espressione, al fine di disegnare le *probe* da caricare sul *microarray*; l'ibridazione è inoltre influenzata da "rumore di fondo" dato da appaiamenti incompleti, imprecisi, che riducono la specificità della tecnica; infine, e questo è il punto cruciale che ha fatto sorgere la necessità di passare a tecnologie NGS, la quantificazione è imprecisa, espressa con un valore continuo.

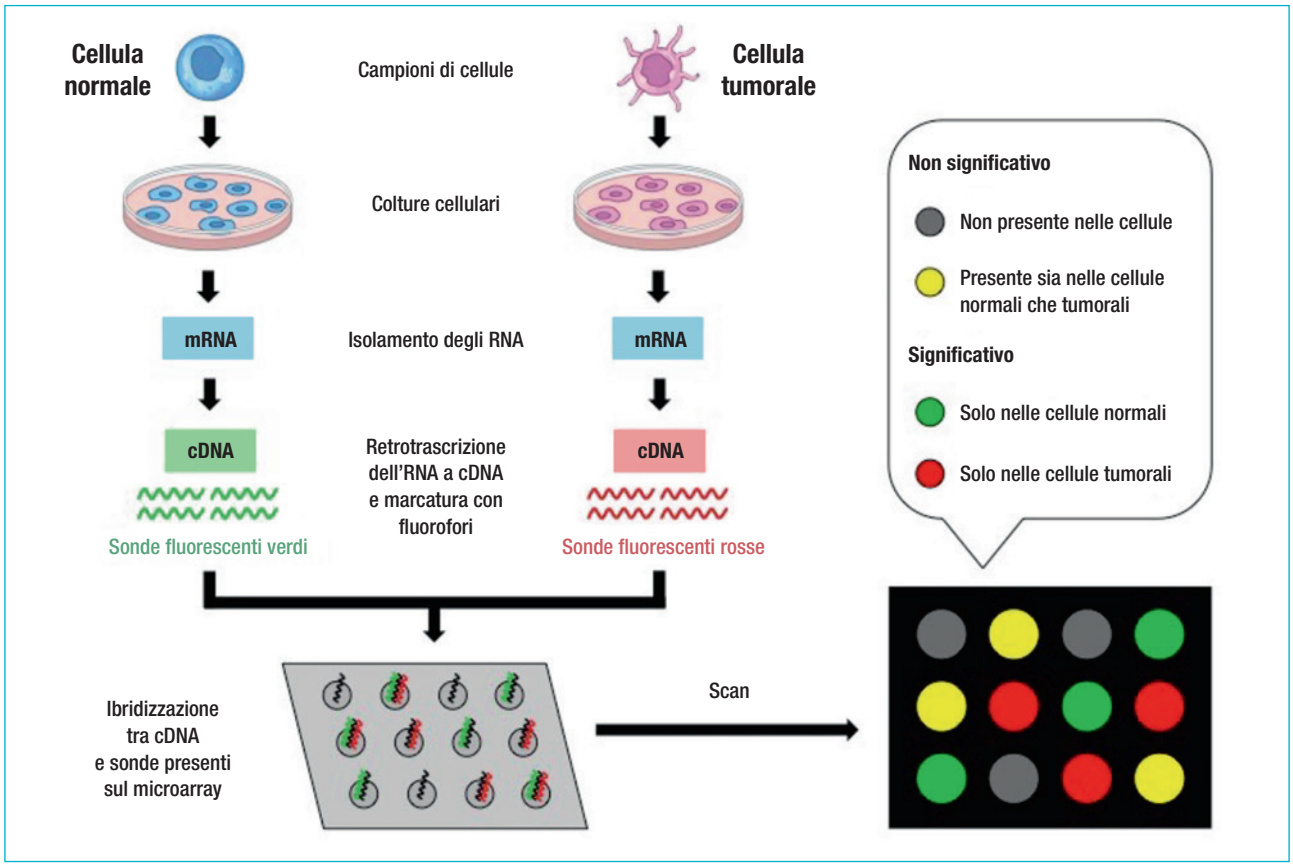
RNA-Seq

Con le tecniche NGS applicate all'RNA (qui vedremo la tecnica nota come *RNA-Seq*), invece, è possibile analizzare il trascrittoma cellulare, utilizzando come *input* per la strumentazione, l'acido nucleico da sequenziare e quantificare.
Gli RNA cellulari, dopo essere stati estratti dalla cellula, vengono retrotrascritti in DNA, sequenziato attraverso tecniche NGS. Vengono così ottenute delle sequenze di DNA di lunghezza variabile in base alla tecnologia di sequenziamento utilizzata, definite *reads*. Tali sequenze (*reads*) vengono poi mappate su un genoma o un trascrittoma di riferimento per identificare i geni espressi nel campione in esame.
La tecnica è quantitativa: tanto maggiore è il numero di molecole di trascritto (RNA), tanto maggiore sarà il numero di *read* prodotte attraverso il sequenziamento. Il totale delle

FIGURA 1.

Microarray.

Lo strumento per l'analisi del trascrittoma è il microarray di DNA, che è costituito da una collezione di microscopiche sonde di DNA attaccate a una superficie solida (vetro, plastica o chip siliconici), in modo da formare una matrice. Nel caso in cui si vogliano studiare gli mRNA, questi debbono essere estratti dalle cellule, convertiti in DNA complementare (cDNA, ossia un DNA a doppia elica ricostruito a partire dall'mRNA attraverso un enzima chiamato trascrittasi inversa) e marcati con una sonda fluorescente. Quando si fa avvenire l'ibridazione fra la sonda presente sulla matrice e il cDNA bersaglio, quest'ultimo rimarrà legato alla sonda e potrà essere identificato semplicemente rilevando la posizione in cui è rimasto legato. Se, come nella figura, il cDNA proveniente da cellule sane e da cellule malate è marcato con diversi fluorofori, è possibile comparare l'espressione genica in condizioni fisiologiche e patologiche. Da notare che, in alcuni casi, si ottengono risultati non significativi: quando, infatti, sia ha un'equivalente espressione del medesimo gene nella cellula sana e in quella malata, il dato non ha rilevanza sperimentale. Importanti, invece, i dati di espressione specifici per un tipo cellulare.



read allineate su un gene (o un trascritto, o un esone), detto *count*, è una unità di misura dell'espressione del gene stesso, interpretabile attraverso modelli statistici (Fig. 2).

I vantaggi della tecnologia RNA-Seq rispetto a quella dei *microarray*, sempre meno utilizzati, sta essenzialmente nella possibilità di essere utilizzata anche quando non è nota la sequenza del gene in esame (quella che sarebbe necessaria per disegnare le *probe*) e nella maggiore precisione e più alta risoluzione della quantificazione delle misure.

Infatti, i dati di RNA-Seq, i *count*, sono una misura discreta, mentre nel caso dei *microarray* si tratta di dati analogici, ovvero del valore continuo dell'intensità luminosa.

Nel campo dell'analisi dell'espressione differenziale, cioè l'identificazione di geni che presentano significative differenze del loro livello di espressione tra due o più condizioni sperimentali (interne o esterne alla cellula). Si valuta cioè, se le differenze osservate tra i *count* delle diverse condizioni sperimentali siano o meno statisticamente significative.

Analisi dei dati: *clusterizzazione* e PCA

Come in ogni studio basato su tecnologia *omica*, la quantità di dati generata è enorme, dunque richiede l'utilizzo di software di bioinformatica e statistica per elaborare le informazioni e creare modelli multidimensionali in grado di spiegare le interazioni all'interno del sistema o tra vari sistemi.

L'analisi dei dati di trascrittomica, avviene in generale attraverso l'utilizzo di metodi di *clustering* (Fig. 3), metodi di statistica multivariata che raggruppano unità statistiche sulla base di misure di similarità o dissimilarità.

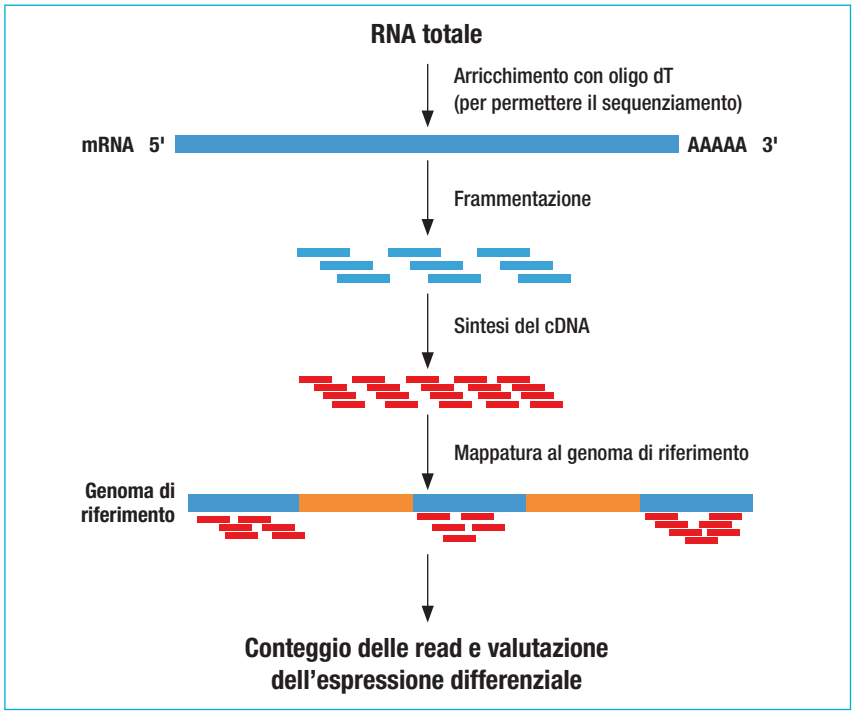
Simili rispetto a cosa? Si considerano i geni come punti nello spazio tra cui si misura la distanza: punti vicini sono raggruppati (clusterizzati) insieme.

Gli *algoritmi di clustering* possono essere di tipo *gerarchico* o *non gerarchico*, laddove i primi non necessitano di informazioni a pri-

FIGURA 2.

RNA-Seq.

L'RNA-Seq è una tecnica per l'analisi del trascrittoma e la sua quantificazione basata sulle tecnologie di Next-Generation Sequencing. I sequenziatori forniscono una valutazione dell'espressione genica attraverso le read, ovvero le sequenze che identificano l'ordine in cui si susseguono le basi azotate che compongono il gene; il numero di read per ciascun gene (mappato su genoma o trascrittoma di riferimento) viene detto count e costituisce una misura dell'espressione genica. Nella figura le read vengono visualizzate come frammenti rossi allineati al genoma di riferimento.



ori e lavorano bottom-up, mentre i secondi sono utili nel caso si voglia definire in anticipo un numero di classi in cui raggruppare i dati gli oggetti (geni).

Un'altra tecnica molto utilizzata è quella dell'analisi delle componenti principali (PCA, *Principal Component Analysis*), che permette di ridurre il numero di variabili causali che descrivono un fenomeno. L'obiettivo è identificare un sottoinsieme di variabili causali dalle quali dipende la maggiore varianza (variabilità) del fenomeno, trovando relazioni non precedentemente sospettate tra le variabili.

Applicazioni cliniche

Il trascrittoma deve essere considerato come una struttura molto complessa e dinamica, sensibile all'ambiente in cui si trovano le cellule e i tessuti in cui viene espresso, sensibile al tempo, alla fase del ciclo cellulare o dello sviluppo del tipo cellulare in esame, sensibile agli effetti di fattori esterni che possono influenzare i processi trascrizionali e post-trascrizionali (ad es. condizioni patologiche).

Ecco allora che possiamo comprendere che quanto accade fisiologicamente e che abbiamo fino a ora descritto, accade anche nelle situazioni patologiche: il trascrittoma di una cellula "malata" è differente da quello di una cellula sana, dunque, a partire da informazioni sui trascritti di soggetti sani o malati, è possibile rilevare quali geni sono espressi in maniera differente tra i due gruppi e, quindi, di evidenziare le conseguenze molecolari della condizione patologica, aiutando così a comprendere l'eziopatogenesi della malattia in esame.

Questo tipo di indagine ha notevoli implicazioni in campo medico-clinico, fornendo un mezzo efficace per la prognosi e la diagnosi precoci, e per lo sviluppo di strategie terapeutiche mirate. Attraverso il monitoraggio del trascrittoma è possibile caratterizzare le basi molecolari di uno stato patologico ed elaborare protocolli terapeutici mirati e con minori effetti indesiderati.

Tale analisi può essere applicata anche sul medesimo soggetto-paziente lungo i diversi step del *follow-up*, con il fine di verificare l'efficacia o meno di un trattamento o la recidiva di malattia.

L'analisi dei trascrittomi, in clinica, può avvenire a partire da differenti campioni biologici (dal siero, alle urine, fino a liquor, saliva e tessuti).

I cambiamenti dell'espressione genica possono essere associati a patologie come causa genetica multifattoriale. Infatti, alla luce delle nuove conoscenze omiche, non possiamo più pensare al concetto di malattia genetica (quella essenzialmente monogenica) che avevamo in passato: le patologie, infatti, sembrano essere causate non solo dal singolo gene difettoso, ma anche dalle interazioni che esso instaura con altri geni e diversi trascritti o diverse proteine, al cambiamento dei *pool* intracellulari di macromolecole, fino a parlare di malattie *monotrascrittomiche* o *monoproteomiche*.

Vediamo solo alcuni esempi.

Fasi precise della *vita prenatale* costituiscono infatti il momento di insorgenza di determinate patologie genetiche e neurologiche di alcuni bambini, aprendo scenari sull'applicazione delle tecnologie omiche nella diagnosi pre-natale, per comprendere determinate malattie che in fasi successive dello sviluppo daranno i primi sintomi, modificheranno le loro caratteristiche fenotipiche, per rivelarsi come forme sindromiche complesse.

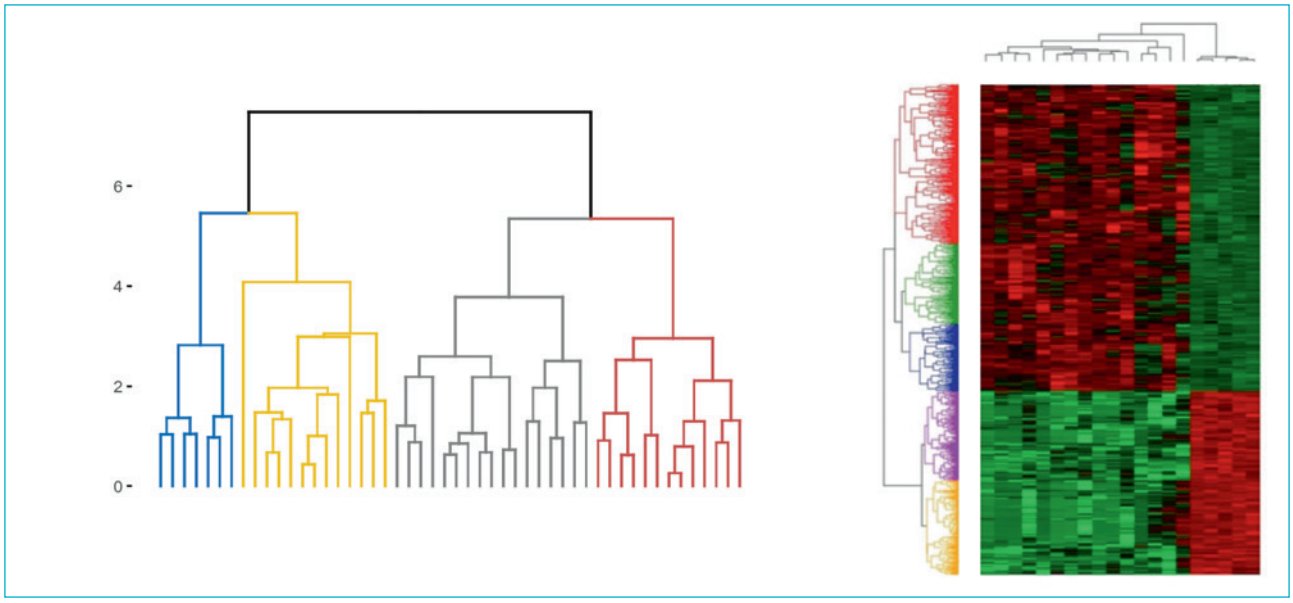
Nella *sclerosi multipla* a esordio nell'età adulta, una delle maggiori limitazioni sta nel fatto che le alterazioni fisiopatologiche iniziali si siano verificate anche diversi anni prima dell'inizio delle manifestazioni cliniche. Studiare dunque più precocemente i fenomeni che si verificano in questo gruppo di patologie, sia a livello molecolare che clinico, non avrebbe unicamente una finalità di comprensione eziopatogenetica, ma anche terapeutica, permettendo di anticipare le cure (anche prima che i sintomi si siano manifestati) e rallentare il decorso di malattia.

I dati di trascrittomica sono stati inoltre spesso utilizzati per identificare le differenze di espressione geniche nelle *cellule*

FIGURA 3.

Clusterizzazione ed espressione genica differenziale.

Il clustering gerarchico permette di raggruppare geni con pattern di espressione simili, visualizzandoli attraverso dei dendogrammi (il grafico nella figura). Maggiore è la distanza tra i punti collegati da linee, maggiore è la differenza di espressione. I punti collegati tra loro da linee più brevi rappresentano geni con comportamento via via più simile, partendo dall'alto del dendogramma (la radice dell'albero), fino alla parte bassa (le foglie dell'albero). Nella parte sinistra della figura, il dendrogramma mostra quattro cluster in quattro diversi colori. Nella figura accanto, al dendrogramma viene associata una rappresentazione grafica definita heat map, che spesso utilizza colori analoghi a quelli delle sonde del microarray, per facilitare l'interpretazione dei dati. Da notare che il cluster di geni che occupa la parte alta dell'heat map, ha un comportamento opposto al cluster che occupa la parte bassa: geni attivi e inattivi, espressi o non, ovvero verdi e rossi, seguono andamento opposto.



tumorali. Gli scopi di tali studi vanno dal cercare di ottenere una migliore classificazione dei tipi di tumori e di identificare i tipi cellulari da cui i tumori provengono, alla caratterizzazione dei profili di espressione che possono aiutare a prevedere la risposta alla terapia, raggruppare i geni per formulare ipotesi riguardanti il loro meccanismo d'azione nella cancerogenesi, identificare nuovi bersagli genici per la chemioterapia.
Le modificazioni del trascrittoma *anticipano* le modificazioni istopatologiche riconoscibili solo successivamente e questo vale per un gran numero di patologie, da quelle neoplastiche a quelle neurodegenerative, fino a quelle cardiovascolari e al diabete. È questa la chiave dell'applicazione clinica dell'analisi dei trascrittomi: *prevedere, anticipare, trattare* ecc. prima dell'esordio clinico.
Infine, i profili di espressione sono utilizzati anche per valutare la *risposta ai farmaci*. È possibile valutare la sensibilità a un certo gruppo di farmaci, identificando anche i geni candidati coinvolti nella risposta ai farmaci.

Bibliografia di riferimento

Mortazavi A, Williams BA. *Mapping and quantifying mammalian transcriptomes by RNA-Seq*, Nat Methods 2008;5:621-8.

Velculescu VE, Vogelstein B. *Serial analysis of gene expression*, Science 1995;270:484-7.

Wang Z, Gerstein M. *RNA-Seq: a revolutionary tool for transcriptomics*, Nat Review Genet 2009;10:57-63.

OMICS Publishing group [online]. http://omicsonline.org.

Per coloro che desiderassero approfondire, basta provare a inserire il nome della patologia di interesse (ad es. *autism*) nella banca dati di PubMed, insieme alle parole chiave *transcriptome* o *gene expression* (o anche *micro-RNA* per chi volesse scendere ancor più nel dettaglio): è interessante vedere quanti studi su specifiche patologie siano stati condotti per identificare i cambiamenti di espressione genica e correlarli alla clinica (dall'eziopatogenesi, alla prognosi, alla risposta alla terapia).