Il ruolo dell'ICT nelle scienze omiche *high-throughput*

Lo sviluppo delle nuove scienze "omiche" (genomica, trascrittomica, metabolomica ecc.) e della strumentazione ad alta efficienza (*high throughput*) ha prodotto, in un'area scientifica originariamente legata prevalentemente ad attività di laboratorio, una forte richiesta di supporto dal settore dell'informatica avanzata e del calcolo ad alte prestazioni. L'enorme mole di dati continuamente resa disponibile dagli esperimenti con nuovi strumenti *high throughput* rende indispensabile una forte collaborazione con l'informatica. Nel presente lavoro, vengono passate in rassegna le principali metodologie in uso nella genomica e nella trascrittomica high throughput, mostrate le problematiche computazionali che esse generano ma anche le importanti prospettive di sviluppi di conoscenza che sono in grado di aprire in settori importanti come le biotecnologie e la medicina

Giuseppe Aprea, Giulio Gianese, Marco Pietrella, Vittorio Rosato, Valentina Spedaletti

Introduzione

Il sequenziamento del DNA è considerato uno strumento essenziale per la ricerca biologica per correlare il genotipo di un organismo al fenotipo che ne deriva. Il progetto "Genoma Umano", durato più di 10 anni con investimenti considerevoli, ha consentito di sequenziare l'intero DNA umano ed è stato una pietra miliare per la ricerca sui geni che causano le patologie genetiche che sono oggetto di studio in molti laboratori.

L'approccio al sequenziamento, basato sul tradizionale metodo Sanger, unica metodica disponibile fino a qualche anno fa, ha subito un'improvvisa accelerazione negli ultimi anni con lo sviluppo di metodologie di nuova generazione (NGS) le quali, implementando un sistema dotato di maggiore processività per la lettura delle sequenze in parallelo, sono in grado di sequenziare fino a 600 miliardi di basi (gigabasi o GB)[1] con un ciclo di sequenziamento che dura una decina di giorni, contro le circa 1.000.000 basi/giorno ottenute con il metodo Sanger. Parallelamente a questo straordinario sviluppo in termini di *throughput*, vi è stata una drastica riduzione in termini di costi che ha portato il sequenziamento di una Mb a costare ~1 dollaro contro i ~10.000 dollari necessari per ottenere lo stesso risultato nel 2000. Conseguentemente, il sequenziamento di grandi genomi è diventato accessibile a numerosi laboratori, determinando una vera e propria rivoluzione nel settore. In particolare, la gestione e l'analisi dei dati hanno determinato la nascita e lo sviluppo della bioinformatica, una disciplina che integra le competenze nell'ambito della biologia e dell'ICT.

Giuseppe Aprea, Vittorio Rosato
ENEA, Unità Tecnica Modellistica Energetica e Ambientale

Giulio Gianese, Marco Pietrella, Valentina Spedaletti
Ylichron s.r.l., Laboratorio Genechron, c/o Centro Ricerche Casaccia, Roma

Applicazioni delle tecniche di NGS

I sequenziatori di nuova generazione sono diventati gli strumenti d'elezione per analisi approfondite nel campo della biologia e della medicina. L'ENEA ne possiede di due tipi: uno con tecnologia Roche-454 localizzato nel Centro Ricerche Trisaia ed un altro, con tecnologia di terza generazione Ion Torrent, dislocato nel Centro Ricerche Casaccia. Questi strumenti consentono di approcciarsi in maniera nuova e più globale ad una serie di applicazioni riportate di seguito.

- Sequenziamento de *novo* [1,2]: si applica ad un genoma per il quale non si dispone già di dati di riferimento, e permette di ottenere informazioni essenziali per studi strutturali e funzionali. L'analisi di genomi o trascrittomi complessi è l'ambito che maggiormente ha beneficiato dell'aumento della processività e dell'abbattimento dei costi raggiunti con l'NGS. Basti pensare che oggi il sequenziamento del genoma umano impiegherebbe pochi giorni contro i 10 anni che sono stati necessari per l'ottenimento del primo *draft* [3].
- Sequenziamento del trascrittoma (*RNAseq*) o dei *micro*RNA (piccoli RNA non codificanti ad attività regolativa): si può considerare un altro formidabile strumento d'indagine per lo studio delle funzioni biologiche di una cellula, in quanto fornisce informazioni quantitative sulle differenze nei livelli di espressione dei geni in tessuti diversi o in uno stesso tessuto analizzato in diverse condizioni sperimentali (ad esempio sottoposto a diversi trattamenti) o in differenti stadi di sviluppo [4]. Questo tipo di analisi, inoltre, permette un approccio globale allo studio di eventuali alterazioni strutturali e a problemi biologici complessi come lo studio della regolazione genica.
- *Resequencing*: è il risequenziamento di un intero genoma già sequenziato (o di una sua specifica porzione), che viene effettuato allo scopo di identificare eventuali "difetti" genetici quali mutazioni di un singolo nucleotide (SNP), inserzioni e delezioni di sequenze più o meno lunghe in determinate posizioni del genoma e variazioni nel numero di copie di determinati geni (CNV). Questa tecnica è stata applicata con successo allo studio delle patologie umane, consentendo l'identificazione di nuovi loci genici coinvolti nella loro eziogenesi, come testimonia uno studio sul morbo di Chron, una malattia infiammatoria cronica dell'intestino con una patogenesi poco nota [5]. Le informazioni ottenute non solo possono contribuire alla comprensione dei meccanismi di sviluppo della patologia, ma possono fornire anche informazioni sul trattamento del paziente. Oltre all'intero genoma, le piattaforme NGS possono essere utilizzate per sequenziare selettivamente determinate regioni genomiche o specifici geni (tramite le tecniche di *targeted resequencing* ed *amplicon sequencing*) o per ottenere informazioni sull'intero esoma (l'insieme di tutte le regioni codificanti, tramite il *whole exome-sequencing*). Quest'ultimo approccio si è rivelato particolarmente utile per la ricerca di geni candidati in patologie poligeniche e multifattoriali. Recentemente, infatti, tramite esperimenti di *whole exome-sequencing* sono state identificate delle mutazioni in geni che risultano correlati con i disturbi pervasivi dello sviluppo o dello spettro autistico [6].
- Sequenziamento dei genomi di intere comunità microbiche per studi di metagenomica: è un'analisi che si applica a campioni prelevati direttamente dall'ambiente naturale [7]; ciò permette di ovviare alla necessità, dettata dalle tecnologie tradizionali, di dover coltivare separatamente le specie batteriche per poter procedere al loro sequenziamento. L'impatto di queste nuove metodiche di analisi, in studi di biodiversità microbica, ha permesso di ottenere l'accesso ad informazioni enormemente maggiori di quelle ottenibili in precedenza. Le applicazioni possibili sono innumerevoli e comprendono la caratterizzazione delle comunità microbiche umane a fini clinici e lo studio di popolazioni microbiche ambientali al fine di identificare geni (e specie codificanti) utili per scopi commerciali quali la produzione di biocarburanti o di altri composti a carattere farmaceutico, agrochimico o enzimatico e per *bioremediation* [8,9]. In relazione alle applicazioni in ambito clinico, va menzionato il recente utilizzo delle tecnologie NGS per lo studio della variabilità genomica di virus patogeni, quali HIV e HCV [10]. Per studi tassonomici volti ad identificare i generi

batterici presenti in una comunità microbica e la loro abbondanza relativa, invece, è stato sviluppato un approccio NGS differente basato sul sequenziamento di regioni variabili di RNA ribosomale 16S (rRNA 16S). Tale applicazione può essere condotta con piattaforme a minore throughput pur fornendo un notevole livello di informazione sulla biodiversità del campione analizzato.

- *Chromatin ImmunoPrecipitation-Sequencing* (sequenziamento di cromatina immunoprecipitata) o ChIP-seq: è una tecnica utilizzata per analizzare le interazioni tra DNA e proteine ad attività regolativa. Il *binding* dell'enzima RNA polimerasi, necessario per la trascrizione genica, è regolato da diversi fattori, quali la metilazione del DNA, la sua accessibilità, mediata da proteine note come istoni o la presenza di fattori di trascrizione (TF) in prossimità dei punti di inizio della trascrizione. Qualsiasi alterazione relativa alla sequenza su cui si lega un TF o la RNA polimerasi o riguardante i TF stessi e gli istoni può pregiudicare il corretto funzionamento della trascrizione determinando eventuali stati patologici [11-13]. La tecnica dell'immunoprecipitazione della cromatina consente di selezionare in maniera specifica quelle porzioni di DNA legate da uno specifico TF o istone; il successivo sequenziamento di questi frammenti di DNA permette, quindi, di ottenere un quadro complessivo dell'attività regolativa della proteina analizzata.

Bioinformatica e analisi dei dati: il ruolo dell'ICT nella genomica

L'evoluzione delle strumentazioni per il sequenziamento è illustrata sinteticamente nella tabella 1
I cambiamenti principali sono stati:

- riduzione dei costi (solo dal 2005 al 2009 i costi si sono abbassati di un fattore 10 all'anno);
- riduzione dei tempi;
- aumento del *throughput*;
- migliore qualità delle *read*;
- riduzione della lunghezza delle *read* (ad esclusione della tecnologia Roche 454).

L'unico limite delle piattaforme NGS consiste, escludendo la tecnologia Roche-454, nella minore lunghezza delle read sequenziate. Questo punto è importante poiché la minore estensione delle *read* rende più difficoltoso recuperare la sequenza intera di partenza. Per ovviare a questo problema, sono in continua evoluzione algoritmi in grado di sfruttare gli altri punti di forza delle nuove tecnologie, ossia l'alto *coverage*[2] e la possibilità di sequenziare *read paired-end o mate-pair*, ovvero coppie di read per cui, da protocollo, è nota la distanza (alcune decine di migliaia di basi) sulla sequenza di DNA.
Lo sviluppo dell'ICT ha dunque agevolato quello del *Next Generation Sequencing*; questo è avvenuto sia grazie ai progressi fatti nel tempo dall'hardware e dal software di sistema, sia tramite lo sviluppo di algoritmi

Technology	Sanger Sequencing	Next Generation Sequencing			
Manufacturer	Applied Biosystems	Roche 454	Illumina	Life Technologies	
Model	ABI 3730XL	GS FLX Titanium XL+	HiSeq 2000 dual flow cell	SOLiD 4 System	Ion PGM
Bases per RUN	~ 96 Kb	700 Mb	600 Gb	100 Gb	1 Gb
Time per RUN	2 h	~1 day	~11 days	~14 days	4.5 h
Reads per RUN	96	1 million	6 billions (paired-end)	1.4 billions	5 millions
Reads length	up to 1000 bp	up to 1000 bp (mode 700 bp)	2*100 bp	2*50 bp	up to 400 bp

TABELLA 1 Tabella comparativa tra la precedente e l'attuale generazione di sequenziatori e tra le differenti tipologie di questi ultimi
Fonte: raccolta di specifiche tecniche di strumentazione disponibili online sui siti dei produttori [14-18]

più avanzati che hanno permesso di elaborare in maniera sempre più efficiente i dati e di renderli disponibili alla comunità scientifica.

Lo *storage*

Il punto di partenza per l'analisi di sequenze NGS è senz'altro lo spazio-disco necessario per conservare i dati iniziali e tutte le successive elaborazioni. Si deve considerare, infatti, che ad una base sequenziata possono corrispondere da 8 fino a 16 byte (ma anche oltre, a seconda del tipo di analisi) e che una corsa HiSeq 2000 o SOLiD può portare ad una richiesta di alcuni TB di spazio, trascurando backup o eventuali ridondanze. Per avere un'idea della quantità di dati prodotti da piattaforme NGS, ci si può riferire ai circa 9 petabyte ($18x10^{50}$ byte) generati nel 2010 solamente dal Sanger Institute, uno dei maggiori centri per il sequenziamento al mondo. Naturalmente è importante che questi dati siano adeguatamente accessibili sia per la consultazione sia per il processamento, il che comporta l'utilizzo di sistemi di gestione dati e di rete ad alte prestazioni. L'ENEA implementa già un'infrastruttura di questo tipo, denominata CRESCO, presente nel Centro Ricerche Portici. CRESCO si può considerare una delle primissime infrastrutture in Italia, dotata di un volume di storage GPFS cumulato di oltre 250 TB e collegata, tramite connessione infiniband a 20 Gbit/s, ai circa 3.000 *core* di calcolo.

L'elaborazione

Le applicazioni legate al sequenziamento si basano, per l'ottenimento del risultato finale, sull'assemblaggio e/o sull'allineamento delle read. La quantità già notevole e sempre crescente di dati prodotti pone delle problematiche complesse dal punto di vista sia delle richieste *hardware* (memoria totale richiesta e prestazioni delle *cpu*), sia dell'efficienza degli algoritmi.

L'assemblaggio *de novo*

L'assemblaggio *de novo* ha l'obiettivo di ricostruire una o più sequenze (l'intero genoma, i cromosomi, i geni ecc.) partendo dalle *read* sequenziate che ne costituiscono i frammenti. Poiché la produzione di questi frammenti avviene in modo casuale, derivando da più copie dello stesso DNA o RNA di partenza, si verificano delle sovrapposizioni tra di essi che possono essere sfruttate per riottenere l'intera sequenza originaria. A tal proposito, la possibilità di sequenziare solo delle *read* corte con i sequenziatori di nuova generazione a maggiore *throughput*, rappresenta un problema rilevante. Se già in precedenza i software difficilmente erano in grado di risolvere porzioni genomiche con pattern complessi di sequenze ripetute, attualmente, con la riduzione della lunghezza delle read, tale problematica risulta ancora più evidente. Per completare l'assemblaggio possono essere utili sistemi NGS capaci di produrre read più lunghe, come lo strumento 454 della Roche; questa piattaforma permette, infatti, di ottenere frammenti lunghi fino a 1.000 bp, analogamente al sequenziamento con metodo Sanger. Una strategia alternativa, e largamente utilizzata, è il sequenziamento di librerie *paired-end* o *mate-pair* che, permettendo di determinare quali sequenze si trovano ad una specifica distanza fra loro, semplificano l'assemblaggio anche in presenza di lunghe porzioni altamente ripetitive. In generale, gli algoritmi di assemblaggio, partendo dalla sovrapponibilità (parziale) delle read sequenziate, operano inizialmente la ricostruzione dei primi assemblati o contig, e, successivamente, l'organizzazione di questi ultimi in strutture ordinate e orientate o scaffold. Sfruttando, quindi, l'informazione sulla distanza contenuta nelle librerie *paired-end* o *mate-pair*, si possono determinare in maniera univoca le posizioni reciproche dei *contig* all'interno degli *scaffold*.

I software più utilizzati per l'assemblaggio si possono dividere in due grandi categorie:

- *software* che utilizzano gli algoritmi *Overlap-Layout-Consensus* (OLC) come Newbler [19], Celera Assembler [20] e Arachne [21]. Questi algoritmi costruiscono un grafo in cui i nodi sono le read unite da un link in caso di *overlap* tra loro. Una volta completato, il grafo viene manipolato per estrarre gli insiemi di allineamenti multipli privi di *branch (contigs)* e successivamente viene considerata

l'informazione sulla distanza per ricavare gli scaffold. Questo approccio è più indicato per *read* medio-lunghe e a lunghezza variabile (tipicamente *read* 454 e Ion Torrent).

- *software* basati sui grafi di de Bruijn (DBG) come Euler [22], Velvet [23], Abyss [24] e SOAP2 [25]. In questo caso i nodi rappresentano sottostringhe di lunghezza *k (k-mer)* che va determinata in base alla lunghezza delle *read* disponibili e aggiustata eseguendo più tentativi; un link unisce due *k-mer* nel caso in cui essi si sovrappongano per k-1 basi. Anche in questo caso, dopo aver costruito il grafo relativo ai dati, esso viene manipolato eliminando, ove possibile, gli artefatti *(branch, loop)* per giungere agli *scaffold* finali. Rispetto ai software basati su algoritmi OLC, in questo caso si evita la prima fase di calcolo degli allineamenti *(cpu-demanding)* e risultano più semplici le manipolazioni sul grafo per estrarre i *contig*. Questo approccio è più indirizzato a trattare grandi quantità di *read* medio-corte (tipicamente *read* SOLiD e Illumina).

I problemi tipici che questi algoritmi incontrano sono dovuti ad errori di sequenziamento delle *read* e a porzioni della sequenza da ricostruire che risultano altamente ripetitive; questi elementi causano *loop* o rami ciechi nel grafo, complicando o rendendo impossibile la rilevazione di *contig* e *scaffold*.

In entrambi i casi, l'uso dei grafi consente di moderare la richiesta di memoria che rimane comunque notevole; per dare un'idea più precisa, se si considera un genoma da circa una Gigabase, sequenziato con un *coverage* 20x (come quello di pomodoro, nel cui progetto di sequenziamento l'ENEA è stato tra i capofila), la memoria RAM richiesta è di 60 GB, nel caso di un software come Newbler che sfrutta l'algoritmo OLC, e di circa 10 GB per Abyss, che invece utilizza quello DBG (considerando *read* di lunghezza media pari a 100 nucleotidi e una probabilità d'errore di 0,01 per base).

L'allineamento

L'allineamento delle sequenze è utilizzato in tutte le applicazioni legate al NGS. Nel caso di un genoma eucariote, milioni di *read* devono essere allineate al genoma di riferimento ed i problemi che si possono riscontrare sono: la memoria richiesta, i tempi di calcolo, la gestione delle porzioni di sequenza ripetute sul genoma, gli eventuali errori presenti nelle *read* (SNP) o delezioni e inserzioni presenti nel genoma.

Anche in questo caso si sono affermate due tipologie di software: quelle basate sulle tabelle di *hashing*, e quelle basate sulla trasformata di Burrow-Wheeler (BW).

I software basati sulle tabelle di *hashing* (SSAHA [26], GASSST [27], SHRiMP2 [28]) operano un'indicizzazione[3] della sequenza di riferimento per contenere l'utilizzo di memoria e velocizzare le ricerche. È possibile trattare errori (algoritmo *seed* and *extend*), ma questi non devono essere uniformemente distribuiti sulle *read*, poiché impedirebbero l'ancoraggio di almeno una loro porzione (da utilizzare come *seed*) e di conseguenza il completamento dell'allineamento; la lunghezza del seed è un parametro che va scelto a seconda del caso e degli scopi dell'analisi. Questi *software* hanno problemi di prestazioni nel caso di allineamenti su zone del riferimento con molte ripetizioni.

I software basati sulla trasformata BW (BWA [29], Bowtie [30]), invece, sono molto veloci anche negli allineamenti su porzioni di sequenza con ripetizioni. Al contrario, non esiste ancora una tecnica consolidata in grado di rilevare gli allineamenti in presenza di errori.

Di seguito sono riportati degli esempi di applicazioni NGS per le quali è necessario l'allineamento delle *read*.

Studio dell'espressione differenziale dei geni. Il campione sequenziato è rappresentativo dell'effettivo trascrittoma della cellula, in quanto conserva i rapporti tra le concentrazioni degli mRNA presenti e quindi rende possibile un'analisi differenziale della loro espressione. Una caratteristica dei geni trascritti, che rende più difficoltosa l'analisi, è lo *splicing*. Negli organismi eucarioti la sequenza dell'mRNA "maturo", che viene tradotto in proteina/e, deriva infatti da un processo di rimozione (lo *splicing*) di alcuni segmenti genici (gli introni) dal pre-mRNA e dalla conseguente unione delle porzioni rimanenti e codificanti (gli esoni). Inoltre, uno stesso gene può andare incontro ad un processo di *splicing* alternativo che genera delle "varianti di *splicing*" o isoforme, ovvero mRNA

con sequenza differente sebbene originati dallo stesso gene. La determinazione delle espressioni differenziali di queste isoforme può essere utile per delineare il loro il ruolo funzionale.

È possibile identificare quattro *step* principali nell'analisi dei dati di RNA-seq:

- allineamento: in questo caso il processo di *splicing* rende più difficoltoso l'allineamento perché una *read* potrebbe includere due esoni successivi presenti sulla sequenza genomica di riferimento. Si creano quindi dei *gap* che possono portare ad un fallimento dell'allineamento. Diversi *software* (Mapsplice [31], Tophat [32], GSNAP [33], QPALMA [34]), per rimediare a questo inconveniente, sono in grado di frammentare le read e di allineare i frammenti generati in modo da ricostruire l'allineamento comprendendo i gap; in questo modo, tali algoritmi consentono anche di identificare isoforme sconosciute.
- ricostruzione dei trascritti: i *software* che si occupano di questo *task* si possono dividere principalmente in 2 gruppi: quelli che effettuano una ricostruzione guidata utilizzando il genoma (Scripture [35], Cufflinks [36]) e quelli che invece lavorano senza utilizzare alcun riferimento (Velvet, TransAbyss [37]). Scripture e Cufflinks hanno anche una maggiore sensibilità nell'identificare le isoforme.
- quantificazione dell'espressione: la possibile esistenza di più isoforme per un trascritto crea delle ambiguità nello stabilire da quale di esse abbia origine una specifica *read*, e quindi per quale di esse conteggiarla. Esistono software come Alexa-seq [38] che stimano l'espressione conteggiando, per ogni isoforma, solamente le read che vi mappano unicamente e quindi quelle che allineano a livello di un esone associabile ad un'unica isoforma. Altri codici come Cufflinks, Miso [39] e RSEM [40] affrontano il problema costruendo delle funzioni di verosimiglianza che modellano il processo di RNA-seq. Le stime di espressione finali sono quelle che massimizzano tali funzioni, ovvero le più compatibili con le *read* ottenute da sequenziamento secondo il modello usato.
- analisi differenziale dell'espressione: dopo aver ottenuto le stime per l'espressione dei trascritti, il passo finale consiste nel valutare le differenze tra i diversi campioni. In realtà, data la variabilità che comunque esiste nei risultati dell'RNA-seq e la variabilità intrinseca delle grandezze biologiche, sarebbe importante avere a disposizione un certo numero di repliche per ogni campione per poter procedere ad una corretta stima delle differenze. Qualora si disponga di repliche biologiche, si possono ottenere stime empiriche della variabilità con software del tipo di Myrna [41]. Tuttavia, è più frequente che non sia presente un numero sufficiente di repliche. Per ovviare a questa mancanza, sono stati sviluppati diversi modelli (ad es. in EdgeR [42], DESeq [43], Cuffdiff [36]) che permettono di ottenere livelli di significatività per l'espressione differenziale, anche in assenza di un gran numero di repliche.

Analisi ChIP-seq. Le sequenze derivate da questa analisi dovrebbero risultare particolarmente arricchite in *read* provenienti da quelle porzioni di DNA che l'immunoprecipitazione ha selezionato rispetto al resto dei frammenti del DNA (*background*). Quindi, dopo l'allineamento, tipicamente si producono dei picchi a livello dei siti di legame tra fattore di trascrizione o istone e DNA che devono essere rilevati; quest'operazione risulta molto delicata per la relativa facilità con cui si possono avere falsi positivi e falsi negativi. Durante l'analisi occorre quindi trovare il giusto bilanciamento tra sensibilità (capacità di rilevare un picco) e specificità (capacità di minimizzare i falsi positivi), a seconda degli scopi. Esistono diverse decine di programmi disponibili (tra cui CisGenome [44], Erange [45], MACS [46]) che si differenziano principalmente per la tecnica di *peak-detection*, per il modo col quale si stabilisce successivamente la significatività di un picco e per la gestione del segnale di *background*. Dopo aver definito quali zone del riferimento corrispondono ai picchi di allineamento, l'analisi può procedere diversamente a seconda degli scopi:

- utilizzando *software* come MEME [47], Scope [48], CisGenome per estrarre i motivi maggiormente ricorrenti dalle regioni corrispondenti ai picchi. Tali software sono utili per individuare i motivi di legame di uno specifico TF (ognuno infatti ha una sequenza preferenziale di nucleotidi alla quale si lega);

- utilizzando un programma di *motif search* come MAST [49] per individuare in quali e quanti picchi si ha un riscontro per il motivo cercato. Questo programma permette di individuare se e dove si è legato al genoma un TF in base al motivo di legame noto.

Analisi delle variazioni geniche. In questo caso lo scopo è determinare le differenze statisticamente rilevanti tra uno o più campioni e una sequenza di riferimento (ad es. il genoma). Una *pipeline* piuttosto consolidata per la rilevazione delle varianti si basa su SAMtools [50] e prevede, dopo l'allineamento, i seguenti *step*:

- *pileup* delle sequenze: per ogni posizione sul riferimento vengono considerate le read che la ricoprono e se queste confermano o meno la base azotata prevista; successivamente viene calcolata la verosimiglianza dei dati per ogni possibile genotipo;
- chiamata delle varianti significative: tramite inferenza bayesiana[4] vengono rilevate le varianti significative. Questa fase richiede un'opportuna scelta dei parametri; ad esempio, le soglie per la rilevazione di varianti germinali (congenite e quindi presenti in tutte le cellule dell'organismo) possono essere piuttosto severe, mentre quelle per la rilevazione di varianti somatiche (non congenite, e quindi che possono essere solo in alcune cellule del campione) devono essere più basse.

Conclusioni

Le tecniche di NGS hanno permeato le attività di biologia molecolare in vari settori, dalle biotecnologie vegetali alla genetica medica, favorendo la comprensione di numerosi meccanismi alla base della fisiologia e della patologia delle specie viventi. Accanto agli straordinari progressi della biologia molecolare, in grado di realizzare nuove metodiche e protocolli di analisi *"customizzati"* su apposite piattaforme tecnologiche (composte da specifici strumenti e dal software di acquisizione dati), l'informatica e le tecniche di analisi dati stanno contribuendo in maniera significativa a decretare il successo delle nuove piattaforme e l'acquisizione di importanti progressi scientifici in questi settori.

ENEA, come altre istituzioni di ricerca in Italia e nel mondo, sta attivamente collaborando e partecipando a questo sforzo della comunità scientifica internazionale intervenendo, con le proprie competenze e le proprie strumentazioni, in numerosi programmi internazionali sulle specie vegetali di maggiore importanza

FIGURA 1 Il personale del laboratorio Genechron dello spin-off ENEA Ylichron S.r.l.
Fonte: ENEA

agroalimentare (grano, patata, pomodoro). Molte competenze tecnico-scientifiche del settore della genomica e NGS sono state trasferite ad un proprio spin-off (Ylichron s.r.l.) che le sta portando sul mercato negli ambiti della diagnostica medica, attivando nuove collaborazioni con importanti centri di ricerca medica per la realizzazione di nuovi servizi basati sull'utilizzo delle nuove tecniche di biologia molecolare.

Note

1. Per dare un'idea degli ordini di grandezza in gioco si pensi che il genoma del virus HIV è di 3,5 kB, del batterio E. coli 4,6 MB, quello umano 3,2 GB.
2. Il numero di volte per cui una data regione di DNA è sequenziata.
3. Operazione che consiste nel suddividere la sequenza di riferimento in sottostringhe s di breve lunghezza (10-20 basi) che vengono associate a degli indici (ad es interi binari crescenti) e ad un vettore che contiene tutte le posizioni del riferimento in cui ha inizio quella sottostringa s.
4. Procedimento per la stima di parametri che, partendo da un'idea della distribuzione di probabilità per i parametri (distribuzione a priori), determina, utilizzando anche i risultati osservati, la distribuzione detta "a posteriori", la cui massimizzazione fornisce la stima a cui si è interessati.

Bibliografia

[1] Potato Genome Sequencing Consortium. *Genome sequence and analysis of the tuber crop potato.* Nature 2011; 475(7355): 189-95.

[2] Tomato Genome Consortium. *The tomato genome sequence provides insights into fleshy fruit evolution.* Nature 2012; 485(7400): 635-41.

[3] International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome.* Nature 2001; 409, 860-921.

[4] Massa AN et al. *The transcriptome of the reference potato genome Solanum tuberosum Group Phureja clone DM1-3 516R44.* PLoS One 2011; 6(10): e26801.

[5] Rioux JD et al. *Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis.* Nat Genet 2007; 39(5): 596-604.

[6] Sanders SJ et al. *De novo mutations revealed by whole-exome sequencing are strongly associated with autism.* Nature 2012; 485(7397): 237-241.

[7] Turnbaugh PJ et al. *A core gut microbiome in obese and lean twins.* Nature 2009; 457(7228): 480-4.

[8] Jaenicke S et al. *Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing.* PLoS One 2011; 6(1): e14519.

[9] Schloss PD and Handelsman J. *Biotechnological prospects from metagenomics.* Curr Opin Biotechnol 2003; 14(3): 303-10.

[10] Parameswaran P et al. *Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity.* J Virol. 2012; 86(16): 8546-58.

[11] Jiang C and Pugh BF. *Nucleosome positioning and gene regulation: advances through genomics.* Nat Rev Genet 2009; 10(3): 161-72.

[12] Henikoff S. *Nucleosome destabilization in the epigenetic regulation of gene expression.* Nat Rev Genet 2008; 9(1): 15-26.

[13] Ku CS et al. *Studying the epigenome using next generation sequencing.* J Med Genet 2011; 48(11): 721-30.

[14] http://products.invitrogen.com/ivgn/product/3730XL

[15] http://454.com/products/gs-flx-system/index.asp

[16] http://www.illumina.com/systems/hiseq_2000_1000.ilmn

[17] https://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Next-Generation-Sequencing/Next-Generation-Sequencing-Systems-Accessories.html

[18] https://www.invitrogen.com/site/us/en/home/Products-and-Services/Applications/Sequencing/Semiconductor-Sequencing/pgm.html

[19] Margulies M et al. *Genome sequencing in microfabricated high-density picolitre reactors.* Nature 2005; 437(7057): 376-80.

[20] Myers EW et al. *A whole-genome assembly of Drosophila.* Science 2000; 287 (5461): 2196-2204.

[21] Batzoglou S et al. *ARACHNE: a whole-genome shotgun assembler.* Genome Res 2002; 12: 177-189.

[22] Pavel A et al. *An Eulerian path approach to DNA fragment assembly.* PNAS 2001; 98(17): 9748-9753.

[23] Zerbino DR and Birney E. *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res 2008; 18: 821-829.

[24] Simpson JT et al. *ABySS: A parallel assembler for short read sequence data,* Genome Res 2009; 19: 1117-1123.

[25] Li R et al. *SOAP2: an improved ultrafast tool for short read alignment.* Bioinformatics 2009; 25(15): 1966-7.

[26] Ning Z et al. *SSAHA: a fast search method for large DNA databases.* Genome Research 2001; 11(10): 1725-9.

[27] Rizk G and Lavenier *D. GASSST: global alignment short sequence search tool.* Bioinformatics 2010; 26(20): 2534-40.

[28] Rumble SM et al. *SHRiMP: accurate mapping of short color-space read.* PLoS Comput Biol 2009; 5(5): e1000386.

[29] Li H and Durbin R. *Fast and accurate short read alignment with Burrows-Wheeler Transform*. Bioinformatics 2009; 25: 1754-60.

[30] Langmead B et al. *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol 2009; 10: R25.

[31] Wang K et al. *MapSplice: accurate mapping of RNA-seq read for splice junction discovery.* Nucleic Acids Res 2010; 38: e178.

[32] Trapnell C et al. *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics 2009; 25: 1105-1111.

[33] Wu TD and Nacu S. *Fast and SNP-tolerant detection of complex variants and splicing in short read.* Bioinformatics 2010; 26: 873-881.

[34] De Bona F et al. *Optimal spliced alignments of short sequence read.* Bioinformatics 2008; 24: i174-i180.

Sp

Bibliografia

[35] Guttman M et al. *Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.* Nat Biotechnol 2010; 28: 503-510.

[36] Trapnell C et al. *Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol 2010; 28: 511-515.

[37] Birol I et al. *De novo transcriptome assembly with ABySS.* Bioinformatics 2009; 25 (21): 2872-2877.

[38] Griffith M et al. *Alternative expression analysis by RNA sequencing.* Nature Methods 2010; 7(10): 843-847.

[39] Katz Y et al. *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nature Methods 2010; 7: 1009-1015.

[40] Wang X et al. *Isoform abundance inference provides a more accurate estimation of gene expression levels in RNA-seq.* J Bioinform Comput Biol 2010; 8 (Suppl.1): 177–192.

[41] Langmead B et al. *Cloud-scale RNA-sequencing differential expression analysis with Myrna.* Genome Biol 2010; 11: R83.

[42] Robinson MD et al. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics 26, 139–140 (2010).

[43] Anders S and Huber W: *Differential expression analysis for sequence count data.* Genome Biology 2010; 11: R106.

[44] Ji H et al. *An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nat Biotechnol 2008; 26: 1293-1300.

[45] Mortazavi A et al. *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature Methods 2008; 5: 621-628.

[46] Zhang Y et al. *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology 2008; 9: R137.

[47] Bailey TL and Elkan C. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, 1994; pp. 28-36, AAAI Press, Menlo Park, California.

[48] Chakravarty A et al. *A novel ensemble learning method for de novo computational identification of DNA binding sites.* BMC Bioinformatics 2007; 8: 249.

[49] Timothy L et al. *Combining evidence using p-values: application to sequence homology searches.* Bioinformatics 1998; 14(1): 48-54.

[50] Li H et al. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009; 25: 2078-9.