"Next-generation sequencing": dalla ricerca di base alla diagnostica*

Karl V. Voelkerding[1,2], Shale A. Dames[1], Jacob D. Durtschi[1]
[1]ARUP Institute for Experimental and Clinical Pathology, Salt Lake City, UT, USA
[2]Department of Pathology, University of Utah, Salt Lake City, UT, USA

Traduzione a cura di Maurizio Ferrari e Angela Brisci

ABSTRACT

Next-generation sequencing: from basic research to diagnostics. For the past 30 years, the Sanger method has been the dominant approach and gold standard for DNA sequencing. The commercial launch of the first massively parallel pyrosequencing platform in 2005 ushered in the new era of high-throughput genomic analysis now referred to as next-generation sequencing (NGS). This review describes fundamental principles of commercially available NGS platforms. Although the platforms differ in their engineering configurations and sequencing chemistries, they share a technical paradigm in that sequencing of spatially separated, clonally amplified DNA templates or single DNA molecules is performed in a flow cell in a massively parallel manner. Through iterative cycles of polymerase-mediated nucleotide extensions or, in one approach, through successive oligonucleotide ligations, sequence outputs in the range of hundreds of megabases to gigabases are now obtained routinely. Highlighted in this review are the impact of NGS on basic research, bioinformatics considerations, and translation of this technology into clinical diagnostics. Also presented is a view into future technologies, including real-time single-molecule DNA sequencing and nanopore-based sequencing. In the relatively short time frame since 2005, NGS has fundamentally altered genomics research and allowed investigators to conduct experiments that were previously not technically feasible or affordable. The various technologies that constitute this new paradigm continue to evolve, and further improvements in technology robustness and process streamlining will pave the path for translation into clinical diagnostics.

INTRODUZIONE

Nel 1977 sono stati pubblicati 2 lavori fondamentali relativi ai metodi di sequenziamento del DNA. Allan Maxam e Walter Gilbert hanno descritto un metodo in cui frammenti di DNA marcati in posizione terminale venivano sottoposti a scissione chimica base-specifica e i prodotti della reazione successivamente separati mediante elettroforesi su gel (1). Con un approccio alternativo, Frederick Sanger et al. hanno descritto l'uso di analoghi sintetici di deossinucleotidi terminatori, che incorporati nella catena nascente di DNA ne impedivano l'ulteriore allungamento in maniera base-specifica (2). Il perfezionamento e la commercializzazione di questo ultimo approccio hanno portato alla sua diffusione nella comunità scientifica e, di conseguenza, nella diagnostica clinica. In una applicazione industriale su larga scala, il metodo Sanger è stato impiegato nell'ambito del "Progetto Genoma Umano" per il sequenziamento completo del primo genoma umano; obiettivo che è stato raggiunto nel 2003, dopo 13 anni di lavoro e con un costo stimato di 2,7 miliardi di dollari. In confronto, nel 2008 il sequenziamento del genoma umano ha richiesto 5 mesi di lavoro con un costo di circa 1,5 milioni di dollari (3). Quest'ultimo risultato evidenzia la rapida evoluzione nel campo delle tecnologie di "next generation sequencing" (NGS), che sono emerse nel corso degli ultimi 5 anni. Attualmente, sono disponibili in commercio 5 piattaforme NGS e altre sono in fase di sviluppo. Inoltre, nell'agosto 2008 lo "US National Human Genome Research Institute" (NHGRI) ha annunciato il finanziamento di una serie di progetti nell'ambito del programma "Revolutionary genome sequencing technologies", che ha come obiettivo il sequenziamento del genoma umano con un costo pari o inferiore a 1000 dollari (http://www.genome.gov/ 27527585). Questo articolo descrive le tecnologie NGS, prende in esame il loro impatto nel campo della ricerca di base e ne esplora la potenziale applicabilità alla diagnostica molecolare.

PIATTAFORME NGS

Tutte le piattaforme NGS disponibili presentano una caratteristica tecnologica comune: il sequenziamento

**Questo articolo è stato tradotto con il permesso dell'American Association for Clinical Chemistry (AACC). AACC non è responsabile della correttezza della traduzione. Le opinioni presentate sono esclusivamente quelle degli Autori e non necessariamente quelle dell'AACC o di Clinical Chemistry. Tradotto da Clin Chem 2009;55:641-58 su permesso dell'Editore.*

parallelo massivo di molecole di DNA amplificate in modo clonale o di singole molecole di DNA separate spazialmente in una cella a flusso ("flow cell"). Questa strategia rappresenta un cambiamento radicale rispetto al metodo di sequenziamento descritto da Sanger, che si basa sulla separazione elettroforetica di frammenti di lunghezza diversa ottenuti mediante singole reazioni di sequenziamento. Nelle tecnologie NGS, invece, il sequenziamento viene effettuato mediante cicli ripetuti di estensioni nucleotidiche ad opera di una DNA-polimerasi o, in alternativa, mediante cicli iterativi di ligazione di oligonucleotidi. Poiché la procedura è parallela e massiva, tali piattaforme consentono di sequenziare da centinaia di Mb (milioni di paia di basi) fino a Gb (miliardi di paia di basi) di DNA in un'unica seduta analitica, a seconda del tipo di tecnologia NGS utilizzata. Le piattaforme disponibili sono illustrate di seguito.

Roche/454 Life Sciences

La tecnologia 454 (http://www.454.com) deriva dalla convergenza di due metodiche: il pirosequenziamento e la "polymerase chain reaction" (PCR) in emulsione. Nel 1993, Nyren et al. hanno descritto un approccio di sequenziamento basato sulla rilevazione della chemiluminescenza derivante dal rilascio del pirofosfato a seguito dell'incorporazione di deossinucleosidi trifosfato (dNTP) ad opera di una DNA-polimerasi (4). Successivamente il perfezionamento di tale sistema, da parte di Ronaghi et al., ha consentito lo sviluppo commerciale della metodica di pirosequenziamento (5, 6). Parallelamente, Tawfik e Griffiths hanno descritto l'amplificazione di singole molecole di DNA in micro-compartimenti, costituiti da emulsioni di acqua-olio (7). Nel 2000, Jonathan Rothberg ha fondato la società 454 Life Sciences, che ha portato allo sviluppo della prima piattaforma NGS disponibile commercialmente (GS 20), commercializzata nel 2005. Grazie alla combinazione tra PCR di singole molecole in emulsione e pirosequenziamento, Margulies et al. alla 454 Life Sciences hanno effettuato il sequenziamento dopo frammentazione casuale ("shotgun sequencing") dell'intero genoma di Mycoplasma genitalia (580.069 paia di basi) in un'unica seduta analitica, utilizzando la piattaforma GS 20 e ottenendo una copertura o livello di ridondanza ("coverage") del 96% e un'accuratezza del 99,96% (8). Nel 2007, 454 Life Sciences è stata acquisita da Roche Applied Science, che ha introdotto una seconda versione della piattaforma 454, la GS FLX. Tale sistema, che si basa sulla stessa tecnologia di GS 20, prevede l'impiego di una "slide" a fibre ottiche definita piastra "picotiter". Nel suo formato più recente, sulla superficie della "slide" sono incisi circa 3,4 x 10^6 pozzetti in ciascuno dei quali avviene la reazione di sequenziamento (con volume nell'ordine dei picolitri), le cui pareti presentano un rivestimento in metallo per aumentare la discriminazione tra segnale e rumore di fondo. Per il sequenziamento (Figura 1) viene generata una libreria di frammenti di DNA-stampo mediante nebulizzazione o sonicazione. Successivamente alle estremità di tali frammenti di DNA a doppio filamento (lunghi diverse centinaia di basi) vengono legati degli oligonucleotidi adattatori. La libreria viene quindi diluita a una concentrazione tale da ottenere singole molecole di DNA, che sono denaturate e ibridate a singole biglie rivestite da sequenze complementari a quelle degli oligonucleotidi adattatori. Le biglie sono catturate nelle goccioline di un'emulsione acqua-olio, in cui avviene la reazione di amplificazione clonale (mediante PCR in emulsione) di ogni singola molecola di DNA legata a ciascuna biglia. Dopo l'amplificazione, le biglie che portano immobilizzati i prodotti di amplificazione clonale vengono recuperate dall'emulsione, separate mediante un'ulteriore diluizione, depositate nei singoli pozzeti di una piastra "picotiter" e combinate con gli enzimi necessari per la reazione di sequenziamento. La piastra viene introdotta nel sequenziatore GS FLX e funge da cella a flusso, in cui sono effettuati cicli iterativi di pirosequenziamento mediante aggiunte successive di dNTP. All'interno di ogni pozzetto contenente i prodotti di

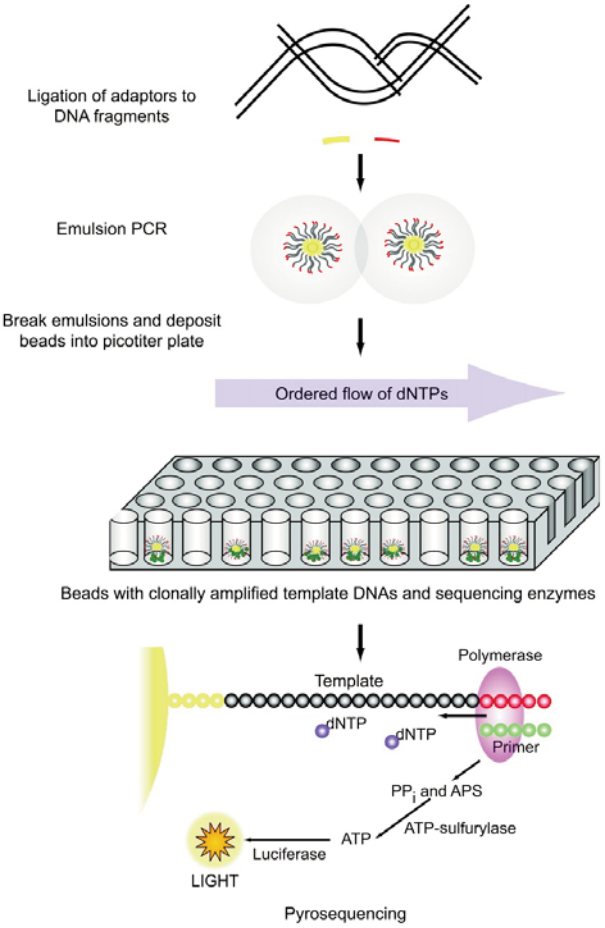

Figura 1
Sequenziamento mediante piattaforma Roche 454 GS FLX. Il DNA stampo è frammentato, legato a oligonucleotidi adattatori e amplificato clonalmente mediante PCR in emulsione. Dopo amplificazione, le biglie sono depositate in pozzetti di una "slide" a fibre ottiche (piastra "picotiter") insieme agli enzimi necessari per la reazione di sequenziamento. La piastra "picotiter" agisce da cella a flusso in cui sono effettuati cicli iterativi di pirosequenziamento. In ogni pozzetto, ad ogni incorporazione nucleotidica, si ha la produzione di pirofosfato (PPi) e la generazione di un segnale luminescente. APS, adenosina 5'-fosfosolfato.

amplificazione clonale, ad ogni incorporazione nucleotidica si ha la produzione di pirofosfato e la conseguente generazione di un segnale luminescente, che viene trasmesso attraverso le fibre ottiche della piastra e registrato da una camera "charge-coupled device" (CCD). Dopo l'aggiunta di ciascun dNTP, il segnale luminoso derivante da ciascun pozzetto viene acquisito, analizzato per valutare il rapporto tra segnale e rumore di fondo, filtrato in base a criteri di qualità e successivamente tradotto mediante algoritmi in una sequenza lineare. Con l'introduzione di una nuova chimica, denominata "Titanium", una singola seduta analitica mediante la piattaforma GS FLX genera circa 1 x 10^6 risultati di reazioni di sequenza ("reads"), corrispondenti a sequenze di lunghezza ≥400 bp (paia di basi), per un totale di circa 500 milioni di paia di basi (Mb) sequenziate.

Un punto di forza della tecnologia 454 è la lunghezza delle "reads", che facilita l'assemblaggio *de novo ("*de novo assembly*")* di interi genomi (9). Uno dei problemi maggiori è rappresentato dalla accuratezza della determinazione di omopolimeri di lunghezza >3-4 bp. Un omopolimero di 6 basi dovrebbe teoricamente generare un segnale di luminescenza pari al doppio del segnale generato da un omopolimero di 3 basi. Operativamente, si ottiene un segnale di luminescenza variabile e, all'aumentare della lunghezza dell'omopolimero, le stime della sua lunghezza diventano meno accurate (8, 10). Il rivestimento in metallo delle pareti dei pozzetti della piastra "picotiter" dovrebbe migliorare l'accuratezza nella determinazione della lunghezza dell'omopolimero. Il numero di volte per cui una data regione di DNA è sequenziata ("coverage depth") e l'accuratezza dei risultati ottenuti mediante la tecnologia 454 sono discussi nella sezione "Analisi dei dati".

Illumina/Solexa

Nel 1997 Balasubramanian e Klenerman hanno ideato un approccio per il sequenziamento di singole molecole di DNA legate a microsfere. L'anno successivo hanno fondato Solexa, ma l'obiettivo di sequenziare singole molecole di DNA non è stato raggiunto, portando invece allo sviluppo di una metodica basata sull'amplificazione clonale del DNA. A partire dal 2006, la prima piattaforma per il sequenziamento di frammenti corti ("short reads") di DNA (Solexa Genome Analyzer) è stata introdotta sul mercato e, in seguito, acquisita da Illumina (http://www.Illumina.com). Lo strumento utilizza una cella a flusso consistente in una "slide" otticamente trasparente costituita, sulla superficie, da 8 comparti a cui sono legati degli oligonucleotidi di ancoraggio (Figura 2). Il DNA stampo è frammentato in segmenti lunghi alcune centinaia di basi e modificato per generare estremità tronche 5'-fosforilate. L'attività polimerasica del frammento di Klenow è utilizzata per aggiungere una singola base di adenina all'estremità 3' del frammento di DNA stampo fosforilato. Questa aggiunta prepara i frammenti di DNA per il processo di ligazione ad oligonucleotidi adattatori, che presentano al 3' una sporgenza di una singola base di timina per aumentare l'efficienza di ligazione. Gli oligonucleotidi adattatori sono complementari agli oligonucleotidi ancorati alla cella a flusso. In condizioni di diluizione limite, il DNA stampo a singolo filamento legato agli adattatori è aggiunto alla cella a flusso e immobilizzato mediante ibridazione agli oligonucleotidi di ancoraggio. A differenza della PCR in emulsione, i frammenti di DNA sono amplificati nella cella a flusso mediante una amplificazione "a ponte" ("bridge amplification"), che si basa sulla cattura dei filamenti di DNA ripiegati ad arco che si ibridano a un oligonucleotide di ancoraggio adiacente. Cicli multipli di amplificazione convertono la singola molecola di DNA stampo in un "cluster" di frammenti ripiegati, amplificati clonalmente, ciascuno composto approssimativamente da 1000 ampliconi clonali. In una singola cella a flusso possono essere generarti circa 50 x 10^6 "cluster" separati. Per il sequenziamento, i "cluster" sono denaturati e una successiva reazione di scissione chimica e lavaggio consente di ottenere solo i filamenti "senso"

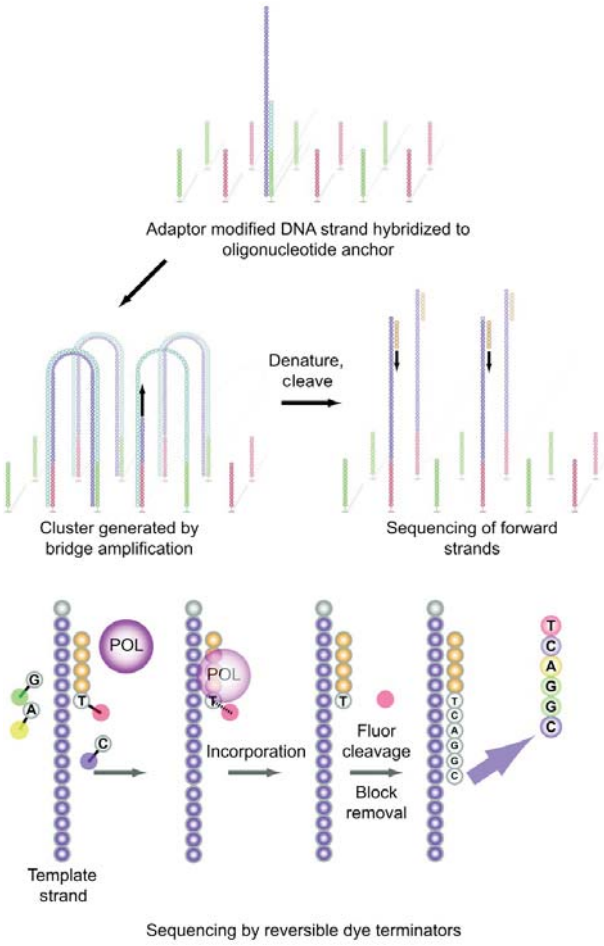

Figura 2
Sequenziamento mediante piattaforma Illumina Genome Analyzer. Il DNA a singolo filamento modificato con oligonucleotidi adattatori è aggiunto alla cella a flusso e immobilizzato mediante ibridazione. L'amplificazione "a ponte" genera "cluster" di frammenti amplificati clonalmente. I "cluster" sono denaturati e recuperati; il sequenziamento ha inizio con l'aggiunta di "primer", DNA polimerasi e 4 nucleotidi terminatori reversibili marcati con composti fluorescenti. Dopo l'incorporazione delle basi, la fluorescenti è rilevata e registrata. La fluorescenza e il blocco sono rimossi prima del successivo ciclo di sintesi.

("forward"). Il sequenziamento dei filamenti "senso" avviene mediante l'ibridazione di un "primer" complementare alla sequenza dell'oligonucleotide adattatore; successivamente, si ha l'aggiunta di una DNA polimerasi e di una miscela di 4 nucleotidi terminatori "reversibili" marcati con fluorofori differenti. I nucleotidi terminatori sono incorporati in base alla complementarietà della sequenza di ciascun filamento, all'interno di ogni "cluster" clonale. Dopo l'incorporazione, i reagenti in eccesso vengono rimossi con un lavaggio e la fluorescenza relativa ad ogni "cluster" viene rilevata e registrata. Mediante successivi passaggi, il gruppo di blocco dei nucleotidi terminatori "reversibili" viene rimosso e la marcatura fluorescente viene allontanata tramite un lavaggio, consentendo di effettuare il ciclo successivo di sequenziamento. Questo processo iterativo di "sequenziamento per sintesi" richiede circa 2,5 giorni per generare "reads" di 36 bp di lunghezza. Con 50 x 10^6 "cluster" per ogni cella a flusso vengono prodotte più di un miliardo di basi (Gb) per ogni seduta analitica (11).

La nuova piattaforma Genome Analyzer II presenta delle modifiche di tipo ottico che consentono l'analisi di "cluster" a densità più elevata. Inoltre, il miglioramento nelle chimiche di sequenziamento unito alla possibilità di leggere sequenze più lunghe di 50 basi dovrebbe consentire un ulteriore miglioramento nelle prestazioni della piattaforma. Illumina e altre tecnologie NGS hanno escogitato strategie per sequenziare entrambe le estremità delle molecole di DNA stampo. Tale possibilità fornisce informazioni che facilitano l'allineamento e l'"assembly" specialmente di "short reads" (12, 13). Con la piattaforma Illumina l'accuratezza nell'identificazione delle basi ("base calling") diminuisce con l'aumentare della lunghezza della "read" (14). Questo fenomeno è dovuto principalmente all'aumento di segnali di interferenza ("dephasing noise"), costituendo un problema tecnico per il processo di sequenziamento. Durante un dato ciclo di sequenziamento, i nucleotidi possono essere incorporati in eccesso o in difetto o la rimozione del blocco può fallire. Con i cicli successivi, questi segnali di errore si accumulano producendo una popolazione eterogenea di filamenti di varia lunghezza all'interno di un "cluster". Questa eterogeneità riduce la specificità del segnale e la precisione nel "base calling", specialmente all'estremità 3' delle "reads". Attualmente, si stanno sviluppando modificazioni nella chimica di sequenziamento e negli algoritmi per l'analisi e l'interpretazione dei dati per attenuare questo fenomeno di "dephasing" (15). Ricercatori del "Wellcome Trust Sanger Institute", che hanno una grande esperienza con la piattaforma Illumina, hanno pubblicato una serie di miglioramenti tecnici per la preparazione delle libreria di frammenti di DNA, inclusi metodi per aumentare la riproducibilità nella frammentazione mediante sonicazione con onde acustiche focalizzate, per aumentare l'efficienza della ligazione degli oligonucleotidi adattatori tramite un processo di ligazione alternativo e per ridurre, tramite un protocollo modificato di estrazione da gel, gli errori nell'assegnazione delle coppie G+C che sono stati osservati nelle "reads" ottenute mediante la piattaforma

Applied Biosystems

La piattaforma "Supported Oligonucleotide Ligation and Detection" (SOLiD) System 2.0, distribuita da Applied Biosystems (http://www.solid.appliedbiosystems.com), è una tecnologia per il sequenziamento di "short reads" di DNA basata su reazioni di ligazione. Questo approccio è stato sviluppato nel laboratorio di George Church e applicato nel 2005 al risequenziamento del genoma di Escherichia coli (17). Nel 2007 Applied Biosystems ha migliorato tale tecnologia consentendo la distribuzione della strumentazione SOLiD. La preparazione dei campioni è simile quella effettuata per la tecnologia 454, in cui i frammenti di DNA sono legati a oligonucleotidi adattatori, immobilizzati su biglie e amplificati clonalmente mediante PCR in emulsione. Le biglie recanti i prodotti di amplificazione clonale vengono fissate sulla superficie di vetro funzionalizzata di una cella a flusso, sulla quale ha luogo la reazione di sequenziamento che viene innescata tramite l'ibridazione di un "primer" complementare all'adattatore a livello della giunzione adattatore-DNA stampo (Figura 3). Invece di fornire un gruppo ossidrile al 3' per l'estensione mediata dalla DNA polimerasi, il "primer" è orientato in modo da fornire un gruppo fosfato al 5' per consentire la ligazione a sonde durante il primo passaggio di sequenziamento mediante ligazione ("ligation sequencing"). Ogni sonda consiste in un ottamero costituito (in direzione 3'→ 5') da 2 basi specifiche seguite da 6 basi degenerate con uno dei 4 marcatori fluorescenti legato all'estremità 5'. Le 2 basi specifiche corrispondono a una delle 16 possibili combinazioni di doppiette di basi (per es. TT, GT e così via). Nel primo passaggio di "ligation sequencing" sono presenti una ligasi termostabile e le sonde rappresentanti le 16 possibili combinazioni, che competono tra loro per l'ibridazione alla sequenza stampo immediatamente adiacente al "primer". Dopo il legame delle sonde, viene effettuata una reazione di ligazione a cui segue una fase di lavaggio per eliminare le sonde non ibridate. I segnali di fluorescenza vengono rilevati e la porzione marcata delle sonde ligate viene tagliata e allontanata tramite un lavaggio in modo da rigenerare un gruppo fosfato all'estremità 5'. Nei cicli successivi di sequenziamento, le sonde vengono ligate al gruppo fosfato all'estremità 5' del pentamero precedente. Vengono effettuati 7 cicli di ligazione successivi, definiti come un "round", per estendere il primo "primer". Successivamente, il filamento sintetizzato viene denaturato e viene ibridato un nuovo "primer" di sequenziamento, sfasato di una base (n-1) rispetto al "primer" usato nel passaggio precedente. In totale sono realizzati 5 "round", utilizzando ogni volta nuovi "primer" sfasati di basi successive (n-2, n-3 e così via), e questo tipo di approccio permette che ogni nucleotide del DNA stampo venga sequenziato due volte. Mediante una seduta analitica della durata di 6 giorni sono generate "reads" di lunghezza pari a 35 bp, la cui composizione nucleotidica viene interpretata in base ai risultati ottenuti dalla ligazione delle diverse sonde con le 16 combinazioni possibili di doppiette di basi. Grazie all'utilizzo di "primer" sfasati, vengono sequenziate diverse basi dell'oligonucleotide adattatore. Questa informazione fornisce una sequenza di riferimento

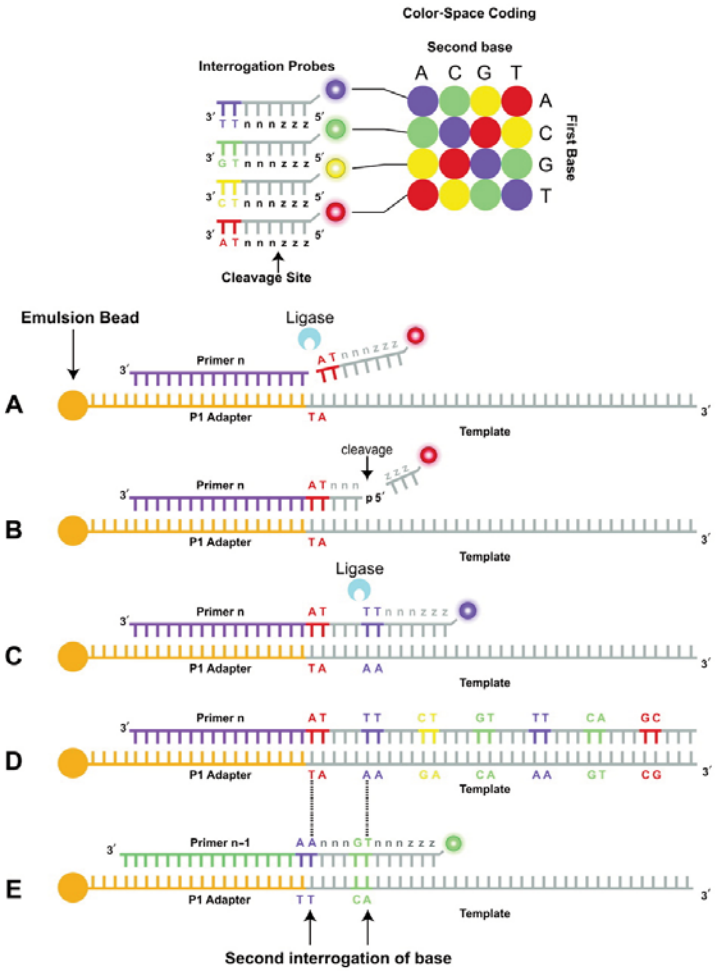

Figura 3

Sequenziamento mediante piattaforma Applied Biosystems SOLiD.

In alto: sistema di codificazione a 2 basi della piattaforma. Ogni sonda consiste in un ottametro costituito (in direzione 3'→5') da 2 basi specifiche seguite da 6 basi degenerate (nnnzzz) con uno dei 4 marcatori fluorescenti legato all'estremità 5'. Le 2 basi specifiche corrispondono a una delle 16 combinazioni possibili di doppiette di basi. In basso: (A) il DNA stampo legato all'oligonucleotide adattatore P1, con il "primer" (n) ibridato, è interrogato dalle sonde che rappresentano le 16 combinazioni possibili di doppiette di basi. In questo esempio, le 2 basi specifiche, complementari al DNA stampo, sono AT; (B) dopo l'ibridazione e la ligazione della sonda il segnale di fluorescenza viene rilevato, la porzione della sonda corrispondente alle ultime 3 basi degenerate viene tagliata e prima del secondo passaggio di sequenziamento, l'estremità 5' della sonda viene fosforilata (non mostrato); (C) ibridazione e ligazione della sonda successiva; (D) estensione completa del "primer" (n) durante il primo "round" costituito da 7 cicli di ligazione; (E) il prodotto ottenuto a seguito dell'estensione del "primer" (n) è denaturato dall'oligonucleotide adattatore legato al DNA stampo e viene effettuato un secondo "round" di sequenziamento con il "primer" (n-1). Grazie all'utilizzo progressivo di "primer" sfasati rispetto al "primer" iniziale (in questo esempio i "primer" sono sfasati di una base: n-1), vengono sequenziate le basi dell'oligonucleotide adattatore, che funge da sequenza di riferimento di partenza, che in combinazione con il sistema di codificazione a 2 basi (in cui ogni coppia di basi adiacenti è correlata a un fluoroforo specifico) permette di decifrare algoritmicamente la sequenza stampo. Mediante tale piattaforma, ogni nucleotide del DNA stampo viene sequenziato due volte.

di partenza che viene utilizzata in combinazione con un sistema di codificazione a 2 basi (in cui ogni coppia di basi adiacenti è correlata a un fluoroforo specifico) per decifrare algoritmicamente la sequenza stampo (Figura 4). Introducendo nello strumento 2 celle a flusso per ogni seduta analitica vengono generati fino a oltre 4 Gb di dati di sequenza. Prima della reazione di ligazione, i filamenti di DNA stampo non estesi vengono bloccati chimicamente ("capping") per diminuire segnali di interferenza aspecifici ed evitare il fenomeno di "dephasing". Il "capping" dei filamenti non estesi associato all'uso di una ligasi ad alta fedeltà e alla doppia interrogazione di ogni base nucleotidica del DNA stampo durante cicli indipendenti di ligazione consentono, secondo il produttore, di ottenere una sequenza consenso ("consensus") con un'accuratezza pari al 99,9% per un frammento noto di DNA stampo sequenziato in "reads" di lunghezza pari a 25 nucleotidi con un "coverage" di 15 volte.

In maniera indipendente, il laboratorio Church ha collaborato con Motion and Dover Systems per sviluppare e introdurre una piattaforma alternativa di "ligation sequencing", il Polonator G.007 (http://www.polonator.org).

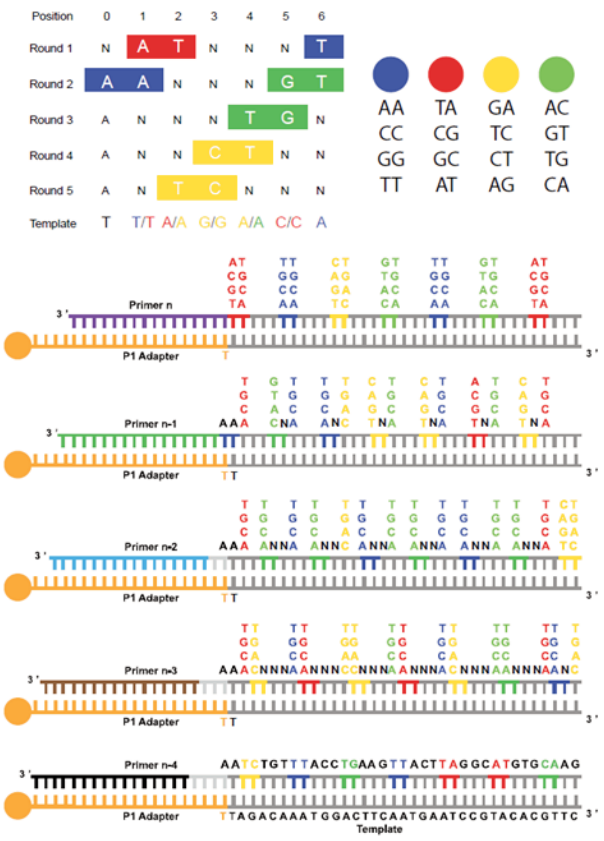

Figura 4

Sistema di codificazione a 2 basi (in cui ogni coppia di basi adiacenti è correlata a un fluoroforo specifico) e analisi algoritmica della sequenza di DNA stampo mediante la piattaforma SOLiD. In alto a sinistra: interrogazione di una sequenza stampo recante le basi TAGACA (in posizione 1-6, con base T della sequenza dell'oligonucleotide adattatore in posizione 0). A destra: sistema di codificazione a 2 basi. In basso: sequenziamento del DNA stampo mediante ligazione e analisi algoritmica. L'estensione del "primer" (n) interroga le basi del DNA stampo in posizione 1 e 2, 6 e 7, 11 e 12, e così via. Nel corso del primo "round" non viene fornita alcuna informazione relativa alla sequenza poiché ciascun fluoroforo rappresenta 4 combinazioni possibili di coppie di basi. L'estensione del "primer" (n-1) permette di decifrare la base della sequenza in posizione 1, sapendo che la base in posizione 0 della sequenza dell'oligonucleotide adattatore P1 è una T. Le estensioni dei "primer" (n-2) e (n-3) forniscono ulteriori informazioni prima della codificazione a 2 basi del DNA stampo effettuata al quinto "round" ("primer" n-4). Le basi indicate come "N" mostrano le posizioni, lungo il DNA stampo, che sono sequenziate due volte prima di essere decifrate.

Nella Tabella 1 sono riassunte le caratteristiche delle piattaforme GS FLX, Genome Analyzer e SOLiD.

Helicos BioSciences (sequenziamento a singola molecola)

Helicos BioSciences (http://www.helicosbio.com) ha ora reso disponibile la prima piattaforma per il sequenziamento a singola molecola (HeliScope), che è in grado, secondo il produttore, di fornire sequenze per un totale di 1 Gb/giorno. Questa tecnologia ha origine dal lavoro di Braslavsky et al. pubblicato nel 2003 (18). Eliminato il processo di amplificazione clonale, il metodo si basa sulla frammentazione del DNA stampo e sulla sua poliadenilazione all'estremità 3', in cui l'adenosina terminale risulta essere marcata con un fluoroforo. I filamenti di DNA poliadenilati vengono denaturati e ibridati a oligonucleotidi di poli(dT) immobilizzati sulla superficie di una cella a flusso, con una densità di cattura fino a un massimo di 100 x 10^6 filamenti di DNA stampo per cm^2. Le coordinate delle posizioni dei singoli filamenti di DNA catturati sono quindi registrate da una camera CCD e il fluoroforo viene rimosso ed eliminato tramite un lavaggio prima di effettuare la reazione di sequenziamento. Per il sequenziamento vengono aggiunti alla cella a flusso una DNA-polimerasi e uno dei 4 dNTP marcati con il fluoroforo Cy5, la cui fluorescenza viene acquisita per determinare l'incorporazione della base nei singoli filamenti di DNA. Dopo rimozione del fluroforo e lavaggio, tale processo viene ripetuto con il dNTP successivo marcato con Cy5. Ogni ciclo di sequenziamento, che consiste in aggiunte successive di DNA polimerasi e di ciascuno dei 4 dNTP marcati, è chiamato "quad". In genere, viene effettuato un numero di circa 25-30 "quad", che permette di ottenere "reads" di 45-50 bp di lunghezza. La piattaforma Helicos è stata utilizzata per sequenziare il genoma di 6407 bp del batteriofago M13 (19). Questo studio ha permesso di dimostrare sia il suo potenziale che le principali problematiche tecniche, comuni a tutti i metodi di sequenziamento a singola molecola basati sul sequenziamento-per-sintesi. In primo luogo, l'accuratezza della reazione di sequenziamento era notevolmente migliorata introducendo il sequenziamento in doppio di ogni molecola stampo ("2-pass sequencing"). In secondo luogo, era rilevata una minore accuratezza nel sequenziamento di tratti omopolimerici a causa dell'aggiunta ad ogni ciclo, da parte della DNA polimerasi, di un numero addizionale di basi dello stesso tipo rispetto a quelle contenute in tali regioni. Helicos ha quindi brevet-

Tabella 1

Confronto tra le diverse piattaforme NGS

	Roche 454 GS FLX	Illumina Genome Analyzer	Applied Biosystems SOLiD	Metodo Sanger
Metodo di sequenziamento	Pirosequenziamento	Nucleotidi terminatori "reversibili" fluorescenti	"Ligation sequencing"	Nucleotidi terminatori fluorescenti
Lunghezza "reads"	400 bp	36 bp	35 bp	800 bp
Tempo di analisi	10 ore	2,5 giorni	6 giorni	3 ore
Basi totali analizzate	500 Mb	1,5 Gb	4 Gb	800 bp

tato dei dNTP marcati definiti terminatori virtuali ("virtual terminators"), che sembra riducano la processività della DNA polimerasi in modo che vengano aggiunte solo singole basi, migliorando l'accuratezza del sequenziamento degli omopolimeri. E' interessante notare che la percentuale dei filamenti di DNA con lunghezza intorno ai 50 nucleotidi che può essere sequenziata è notevolmente inferiore rispetto a quella ottenuta sequenziando frammenti di lunghezza minore, intorno ai 25 nucleotidi, probabilmente a causa delle strutture secondarie (per es. ad ansa) assunte dalle molecole stampo.

IMPATTO DELLA TECNOLOGIA NGS SULLA RICERCA DI BASE

Nei 4 anni successivi alla messa in commercio della prima piattaforma, le tecnologie NGS hanno notevolmente accelerato la crescita di vari settori di ricerca genomica, consentendo di effettuare esperimenti che in precedenza non erano tecnicamente possibili o convenienti da un punto di vista economico. Di seguito vengono descritte le principali applicazioni delle piattaforme NGS e l'analisi dei dati prodotti.

Analisi genomica

La potenzialità delle tecnologie NGS di effettuare analisi ad alta processività è stata sfruttata per sequenziare interi genomi, da quelli di microorganismi a quelli umani (3, 8, 9, 11, 20-24), incluso il recente sequenziamento del genoma di cellule citogeneticamente normali, derivanti da un paziente con leucemia mieloide acuta, che ha portato all'identificazione di nuove mutazioni genetiche tumore-specifiche (25). La capacità della tecnologia 454 di sequenziare frammenti più lunghi rispetto a quelli che possono essere sequenziati con le tecnologie Illumina e SOLiD, adatte al sequenziamento di "short reads", facilita il "de novo assembly" in assenza di un genoma di riferimento. Per il risequenziamento, entrambe le tecnologie per l'analisi di frammenti lunghi o corti sono state utilizzate con successo. In uno studio comparativo, le piattaforme 454, Illumina e SOLiD hanno tutte accuratamente rilevato variazioni di singole basi con un "coverage depth" per allele ≥15 volte (20) (gli aspetti critici relativi al "coverage depth" sono discussi più avanti nella sezione "Analisi dei dati"). La lunghezza maggiore dei frammenti che possono essere sequenziati mediante la piattaforma 454 permette di ottenere informazioni relative ad aplotipi nucleotidici all'interno di regioni di alcune centinaia di basi e tale metodo è più adatto per identificare inserzioni e delezioni di media grandezza e per produrre allineamenti in zone contenenti sequenze ripetute. Ulteriori studi sono necessari per comparare le prestazioni tecnologiche delle varie piattaforme nel rilevare inserzioni e delezioni. Ogni piattaforma ha una strategia alternativa per sequenziare entrambe le estremità dei frammenti delle librerie di DNA ("paired-end sequencing"). Oltre a raddoppiare la produzione dei risultati, il fatto di sapere che le "reads" sono associate tra loro all'interno di un certo frammento aumenta la possibilità di allineamento e "assembly", specialmente per sequenze corte. Il "paired-end sequencing" è stato utilizzato per mappare variazioni genomiche strutturali, incluse delezioni, inserzioni e riarrangiamenti (12, 13, 26, 27). La capacità di sequenziare interi genomi umani mediante piattaforme NGS ad un costo sostanzialmente ridotto ha dato nuovo impulso allo sviluppo di progetti internazionali volti a sequenziare migliaia di genomi umani nei prossimi dieci anni (http://www.1000genomes.org), che consentiranno la caratterizzazione e la catalogazione delle variazioni genetiche umane a un livello senza precedenti.

Risequenziamento genomico di regioni di interesse

Il sequenziamento di sottoregioni genomiche e di gruppi di geni viene attualmente impiegato per identificare polimorfismi e mutazioni in geni implicati nei tumori e in regioni del genoma umano osservate avere un coinvolgimento in malattie genetiche, mediante studi di "linkage" e di associazione sull'intero genoma (28, 29). Soprattutto in quest'ultimo scenario le regioni di interesse possono raggiungere dimensioni da centinaia di Kb a diverse Mb. Per un miglior uso delle tecnologie NGS per il sequenziamento di tali regioni candidate, vengono introdotti nel disegno sperimentale diversi passaggi di arricchimento genomico, sia tradizionali che nuovi. La sovrapposizione di ampliconi (di circa 5-10 Kb) derivanti da reazioni di PCR a lungo raggio ("long-range PCR") può permettere di sequenziare fino a diverse centinaia di Kb, mentre per regioni genomiche più estese questo approccio non risulta molto pratico. Più recentemente, l'arricchimento è stato raggiunto grazie all'ibridazione di DNA genomico umano, frammentato e denaturato, a oligonucleotidi sonda di cattura complementari a regioni di interesse e alla successiva eluizione del DNA arricchito (30-33). Le sonde di cattura possono essere immobilizzate su una superficie solida (Roche NimbleGen, http://www.nimblegen.com; Agilent Technologies, http://www.agilent.com; e Febit, http://www.febit.com) o utilizzate in soluzione (Agilent). L'"array" NimbleGen nel suo formato attuale contiene 350.000 oligonucleotidi di 60-90 bp di lunghezza, che in genere sono distanziati tra loro da 5-20 nucleotidi, dai quali sono esclusi oligonucleotidi complementari alle regioni ripetute. Per l'arricchimento, 5–20 µg di DNA genomico vengono frammentati e legati a oligonucleotidi "linker" contenenti sequenze che fungono da "primer" per reazioni di PCR universali. Questo materiale è denaturato, ibridato a un "array" per 3 giorni, eluito e il DNA arricchito è amplificato mediante PCR prima della preparazione della libreria NGS. In studi riportati in letteratura, sull'"array" 350K sono state catturate fino a 5 Mb di sequenze, tra cui il 60-75% mappanti le regioni di interesse, mentre il rimanente mappanti regioni non di interesse derivanti da catture non specifiche. NimbleGen sta inoltre sviluppando un "array" di 2,1 x 10^6 posizioni per la cattura di regioni genomiche più ampie. La tecnologia di Agilent, basata su l'impiego di sonde di cattura in soluzione, prevede l'uso di oligonu-

cleotidi fino a 170 bp di lunghezza, che recano ad ogni estremità sequenze per l'innesco di una PCR universale e siti di riconoscimento per endonucleasi di restrizione. La libreria di oligonucleotidi viene amplificata mediante PCR, digerita tramite enzimi di restrizione e legata ad adattatori contenenti il sito promotore per la T7 polimerasi. In vitro, in presenza di UTP biotinilato viene effettuata la trascrizione in modo da generare sequenze di cattura di cRNA biotinilato a singolo filamento. Per la cattura, 3 µg di DNA genomico frammentato e denaturato sono ibridati in soluzione alle sequenze di cRNA per 24 ore. Dopo l'ibridazione, ibridi a doppio filamento di DNA e cRNA sono legati a biglie magnetiche rivestite da streptavidina; il cRNA viene poi digerito enzimaticamente, liberando il DNA a singolo filamento arricchito, che viene successivamente processato per le piattaforme NGS. Un approccio di arricchimento alternativo sviluppato da RainDance Technologies (http://www.raindancetechnologies.com) utilizza una nuova tecnologia microfluidica in cui singole coppie di "primer" per la reazione di PCR per regioni genomiche di interesse sono segregate in goccioline di un'emulsione acqua-olio e successivamente riunite per creare una libreria di "primer". Separatamente, vengono preparate delle gocce in emulsione contenenti il DNA genomico e i reagenti per la reazione di PCR. Sono generati due flussi separati, uno di goccioline con la libreria di "primer" e l'altro di goccioline contenenti DNA genomico e reagenti per la reazione di PCR. I due flussi vengono fatti convogliare e le goccioline di entrambi sono mescolate in un rapporto di 1:1. Poiché tale processo avviene attraverso un canale microfluidico, le goccioline sono sottoposte a un impulso elettrico che le porta a fondersi. Le goccioline fuse contenenti le singole coppie di "primer" e il DNA genomico/reagenti per la reazione di PCR sono depositate in una piastra da 96 pozzetti e amplificate. Dopo l'amplificazione, gli ampliconi sono recuperati dall'emulsione e mescolati per essere analizzati mediante le tecnologie NGS.

Metagenomica

Le tecnologie NGS hanno avuto un impatto enorme sullo studio della diversità microbica in campioni ambientali e clinici. Dal punto di vista operativo, il DNA genomico è estratto dal campione di interesse e convertito in una libreria per essere sequenziato mediante una piattaforma NGS. Le sequenze ottenute vengono allineate a sequenze di riferimento di microorganismi che si ipotizza possano essere presenti nel campione. Possono essere distinte specie strettamente correlate e possono essere desunte anche le specie più distanti dal punto di vista filogenetico. Inoltre, il "de novo assembly" di dati ottenuti può fornire informazioni a sostegno della presenza di specie note o potenzialmente nuove. Si ottengono informazioni genomiche di natura qualitativa e l'analisi delle proporzioni relative delle "reads" può essere utilizzata per ottenere informazioni quantitative sulle singole specie microbiche. Ad oggi, la maggior parte delle analisi di metagenomica mediante le tecnologie NGS è stata effettuata con la tecnologia 454 in quanto la lunghezza maggiore dei frammenti sequenziati facilita l'allineamento a genomi microbici di riferimento e il "de novo assembly" di genomi microbici non ancora caratterizzati. Esempi di studi di metagenomica includono analisi di popolazioni microbiche nell'oceano (34, 35) e nel suolo (36), l'identificazione di un nuovo arenavirus in pazienti trapiantati (37) e la caratterizzazione della microflora presente nella cavità orale umana (38) e nell'intestino di gemelli obesi e magri (39).

Sequenziamento del trascrittoma

Le tecnologie NGS hanno fornito un nuovo e potente approccio, denominato "RNA-Seq", per la mappatura e la quantificazione dei trascritti nei campioni biologici. Il RNA totale, privato della componente ribosomiale o il RNA con la coda di poli(A) (mRNA) è isolato e convertito in cDNA. Un protocollo tipico comporta la generazione di un primo filamento di cDNA attraverso la retrotrascrizione mediata da "primer" esamerici a composizione casuale e la generazione successiva del secondo filamento di cDNA ad opera di una RNasi H e di una DNA polimerasi. Il cDNA viene frammentato e legato ad adattatori per l'analisi mediante le tecnologie NGS. Per i piccoli RNA, come i microRNA e gli "short interfering" RNA, sono comunemente impiegati i seguenti approcci di isolamento, che possono anche essere usati in combinazione: l'isolamento preferenziale attraverso un metodo di arricchimento di piccoli RNA o la selezione in base alla taglia tramite separazione elettroforetica su gel. Per unire le sequenze degli adattatori alla molecola di RNA viene utilizzata una RNA ligasi; questo passaggio è spesso seguito da una reazione di PCR prima di procedere con le tecnologie NGS. Dopo il sequenziamento, le "reads" sono allineate a un genoma di riferimento, comparate a sequenze di trascritti noti o utilizzate per un "de novo assembly" per costruire una mappa di trascrizione. Sebbene la tecnologia di RNA-Seq sia ai primi stadi di sviluppo, essa ha già mostrato alcuni vantaggi rispetto agli "array" di espressione genica (40). In primo luogo, gli "array" permettono di analizzare solamente sequenze genomiche note, mentre l'approccio di RNA-Seq, non essendo vincolato da tale limitazione, consente la caratterizzazione dei trascritti senza una conoscenza a priori dei siti genomici di inizio trascrizione. RNA-Seq, rispetto agli "array", è inoltre un metodo che permette di ottenere una risoluzione a livello di singole basi e mostra una maggior capacità di distinguere isoforme di RNA, di determinare l'espressione allelica e di rilevare le variazioni di sequenza. I livelli di espressione sono dedotti dal numero totale di "reads" che mappano per gli esoni di un gene, normalizzate per la lunghezza degli esoni che possono essere mappati in modo univoco. I risultati ottenuti con questo tipo di approccio hanno mostrato una stretta correlazione con quelli ottenuti mediante esperimenti di PCR quantitativa e di "RNA-spiking". L'intervallo dinamico del RNA-Seq per determinare i livelli di espressione è di 3-4 ordini di grandezza, rispetto ai 2 ordini di grandezza degli "array" di espressione. In questo contesto, RNA-Seq ha mostrato un miglioramento delle prestazioni per l'identificazione quantitativa di trascritti prodotti sia ad alti

che a bassi livelli. RNA-Seq viene attualmente impiegato per confermare e aggiornare informazioni geniche, incluse le regioni alle estremità 5' e 3', e le regioni di giunzione tra esoni e introni; questa seconda informazione si ottiene mappando le "reads" rispetto alle giunzioni esoniche caratterizzate da siti "consensus" di "splicing" GT-AG. Possono essere desunte sia informazioni di tipo qualitativo che di tipo quantitativo relative a fenomeni di "splicing" alternativo. RNA-Seq è stato applicato per studi su una varietà di organismi, tra cui Saccharomyces cerevisiae, Arabidopsis thaliana, topi e cellule umane (40-51).

Mappatura delle proteine leganti il DNA e analisi della cromatina

La caratterizzazione delle proteine regolatrici associate al genoma è progredita enormemente grazie all'introduzione delle tecnologie di immunoprecipitazione della cromatina e all'ibridazione a "microarray" (ChIP-on-chip) (52). Con tale approccio, le proteine associate al DNA genomico sono ibridate chimicamente (tipicamente tramite trattamento blando con formaldeide) ai loro siti di legame e il DNA è frammentato mediante sonicazione o digestione con nucleasi micrococcica. Le proteine ibridate con il DNA sono immunoprecipitate con anticorpi specifici per le proteine di interesse. Il DNA nell'immunoprecipitato viene purificato e ibridato a un "array" di oligonucleotidi costituito da sequenze genomiche, consentendo l'identificazione dei siti di riconoscimento delle proteine leganti il DNA ("DNA-binding proteins"). Questo approccio è stato impiegato con successo per identificare i siti di legame dei fattori di trascrizione e delle proteine istoniche.

La tecnologia ChIP-on-chip è attualmente sostituita in una varietà di approcci sperimentali con ChIP-Seq, in cui il DNA recuperato dell'immunoprecipitato viene convertito in una libreria che viene analizzata mediante piattaforme NGS. Le "reads" ottenute sono mappate al genoma di interesse di riferimento per generare una mappa sull'intero genoma ("genome-wide") dei siti di riconoscimento delle "DNA-binding proteins" (53-55). Gli studi che fino ad oggi hanno esaminato i siti genomici di legame delle proteine di trascrizione umane "Neuron restrictive silencer factor" (NRSF) e "Signal transducer and activator of transcription 1" (STAT1) mostrano che la risoluzione ottenuta tramite la tecnologia ChIP-Seq è maggiore rispetto a quella ottenuta mediante ChIP-on-chip, come evidenziato dal fatto che oltre alla conferma di siti caratterizzati in precedenza ne sono stati identificati di nuovi (56, 57). Analogamente all'approccio di RNA-Seq, il ChIP-Seq presenta l'importante vantaggio di non richiedere la conoscenza a priori delle posizioni genomiche di legame proteico. Oltre allo studio relativo ai fattori di trascrizione, le piattaforme NGS sono anche utilizzate per mappare la metilazione a livello del genoma. Un approccio tradizionale, che è stato applicato per lo studio di sottoregioni genomiche o di interi genomi, prevede la conversione del DNA mediante bisolfito seguita da un'analisi con le tecnologie NGS (58, 59). Attualmente si sta cercando di sviluppare una variante della tecnologia ChIP-Seq, in cui il profilo di metilazione viene analizzato associando il processo di immunoprecipitazione all'utilizzo di un anticorpo monoclonale diretto contro le citosine metilate e l'analisi successiva su piattaforme NGS (60).

ANALISI DEI DATI

Gli esperimenti con piattaforme NGS generano una quantità di informazioni senza precedenti e questo rappresenta una sfida per la gestione, l'archiviazione e, soprattutto, l'analisi dei dati (61). Come primo tipo di informazione si ottiene un'enorme quantità di immagini dei segnali di luminescenza o di fluorescenza acquisiti dalla superficie della cella a flusso, registrate dopo ogni passaggio di sequenziamento iterativo (Figura 5). Questa massa di informazioni richiede un sistema interconnesso ("data-pipeline system") con capacità operativa molto elevata per l'archiviazione, la gestione e l'elaborazione dei dati. La dimensione delle informazioni generata durante una singola seduta analitica effettuata con le tecnologie 454 GS FLX, Illumina e SOLiD è rispettivamente di circa 15 GB, 1 TB e 15 TB. L'operazione di elaborazione principale del "data-pipeline system" consiste nella conversione delle immagini acquisite in "reads",

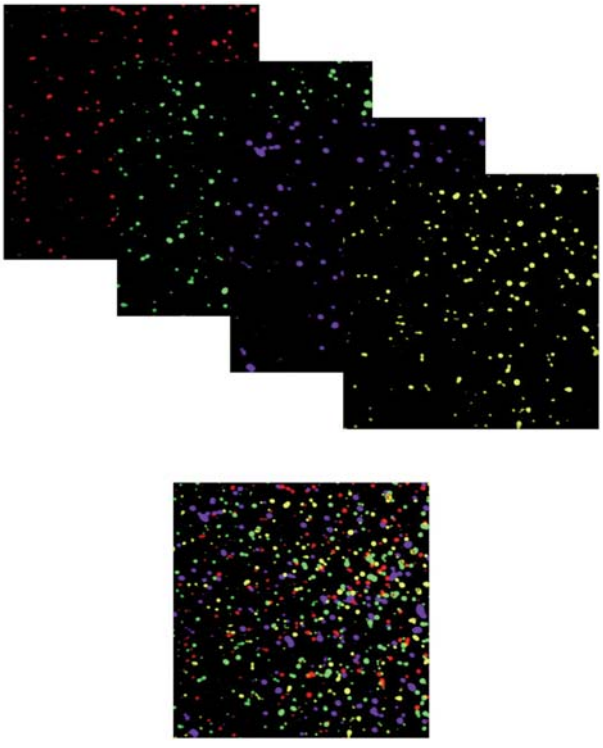

Figura 5
Immagine a colori della cella a flusso della piattaforma Illumina. Ogni segnale di fluorescenza deriva da un "cluster" di DNA stampo amplificato clonalmente. Il pannello superiore mostra i segnali di fluorescenza emessi alle 4 lunghezze d'onda dai diversi fluorofori. Le immagini vengono elaborate per identificare i singoli "cluster" e per eliminare il rumore di fondo o eventuali interferenze. Il pannello inferiore rappresenta un'immagine composita dei 4 canali di fluorescenza.

definita "base calling", che richiede un intenso lavoro computazionale. In primo luogo, le singole biglie o i relativi "cluster" vengono identificati e localizzati attraverso una serie di immagini. Mediante l'impiego di algoritmi specifici per ogni piattaforma vengono quindi valutati una serie di parametri di immagine quali intensità, rumore di fondo e presenza di eventuali segnali aspecifici per generare sequenze nucleotidiche e assegnare a ciascuna base punteggi di qualità ("quality scores"), correlati alla probabilità di errore. Sebbene molti ricercatori utilizzino i "software" "data-pipeline" specifici per ogni piattaforma per il processo di "base calling", sono stati sviluppati anche programmi alternativi che utilizzano "software" più avanzati e che prevedono anche la valutazione di parametri statistici. Le caratteristiche di questi programmi alternativi prevedono l'inclusione di basi ambigue nelle "reads" una migliore esclusione delle basi poste all'estremità delle "reads", che presentano un basso "quality score" (62), e l'impiego di una serie di dati per il training all'utilizzo del "software" di analisi (15). L'introduzione di queste caratteristiche ha permesso di ridurre gli errori di lettura e di migliorare l'allineamento, data in specifico la tendenza attuale di generare "reads" più lunghe. Questi vantaggi, tuttavia, devono essere valutati in base alle necessarie capacità operative del computer, richieste dall'ampio volume di dati di immagine analizzati.

I "quality scores", calcolati durante il processo di "base calling" delle piattaforme NGS, forniscono importanti informazioni per l'allineamento, l'"assembly" e l'analisi di variazioni. Sebbene la valutazione della qualità vari a seconda del tipo di piattaforma, i calcoli sono tutti effettuati in base allo storico "phred score", introdotto nel 1998 per i dati di sequenza di Sanger (63, 64). Il valore di qualità del "phred score" (q) utilizza una scala matematica per convertire a una scala logaritmica la probabilità stimata per l'identificazione non corretta di una base (e):

$$q = -10 \cdot \log_{10}(e).$$

Le probabilità di identificare una base in modo non corretto pari a 0,1 (10%), 0,01 (1%) e 0,001 (0,1%) producono, rispettivamente, un valore di "phred score" di 10, 20 e 30. I tassi di errore per le piattaforme NGS stimati tramite i "quality scores" dipendono da diversi fattori, inclusi i rapporti tra segnale e rumore di fondo, l'interazione tra biglie o "cluster" adiacenti e il fenomeno di "dephasing". E' stato compiuto un lavoro enorme per capire e migliorare l'accuratezza dei "quality scores" e dei fattori alla base degli errori (10, 14), comprese le inaccuratezze nelle lunghezze delle "reads" omopolimeriche ottenute mediante la tecnologia 454 e degli errori sistematici ("bias") di identificazione delle singole basi generati con il formato Illumina. Studi relativi a questi tipi di errore hanno portato allo sviluppo di esempi di "software" che non richiedono processi aggiuntivi di "base calling", ma migliorano l'accuratezza dei "quality scores" e quindi l'accuratezza del sequenziamento (65, 66). I "quality scores" sono uno strumento importante per escludere "reads" ed eliminare basi che mostrano una bassa qualità, per migliorare l'accuratezza dell'allineamento e per determinare una sequenza "consensus" e la presenza di basi varianti ("variant calls") (67).

L'allineamento e l'"assembly" risultano più difficoltosi sui dati ottenuti tramite le piattaforme NGS rispetto a quelli ottenuti tramite sequenziamento con il metodo Sanger, a causa della minor lunghezza delle "reads" che sono generate nel primo caso. Una limitazione nell'allineamento e nell'"assembly" di "short reads" è l'impossibilità di allineare in modo univoco ampie porzioni di una serie di "reads" quando la lunghezza delle stesse diventa troppo ridotta. Allo stesso modo, il numero di "reads" allineate in modo univoco si riduce quando si effettuano allineamenti a genomi o a sequenze di riferimento più grandi e complessi, che hanno una maggior probabilità di contenere sequenze ripetute. Un esempio è riportato da uno studio che indica come il 97% del genoma di Escherichia coli può essere allineato in modo univoco con "reads" di 18 bp di lunghezza, mentre soltanto il 90% del genoma umano può essere allineato in modo univoco con "reads" di 30 bp (68, 69). L'allineamento univoco o l'"assembly" è ridotto non solo dalla presenza di sequenze ripetute, ma anche dalle omologie condivise all'interno di famiglie geniche strettamente correlate e con pseudogeni. L'allineamento non univoco di "reads" viene risolto dal "software" assegnando le "reads" a posizioni multiple o lasciando lacune ("gaps") nell'allineamento stesso. Nel processo di "de novo assembly" queste reads saranno escluse, portando a insiemi contigui di "reads" ("contigs") assemblati, più corti e numerosi. Questi fattori sono rilevanti per la scelta della piattaforma di sequenziamento più appropriata e alla relativa lunghezza delle "reads" prodotte, specialmente per processi di "de novo assembly" (9).

I tassi di errore delle "reads" per le singole piattaforme NGS sono più elevati di quelli che si osservano per il sequenziamento col metodo Sanger. In quest'ultimo caso, la maggior accuratezza deriva non solo dal grado di avanzamento della chimica, ma anche dal fatto che i picchi dei ferogrammi corrispondono a reazioni di estensione con terminazioni multiple e altamente ridondanti. L'accuratezza per le piattaforme NGS è ottenuta mediante il sequenziamento ripetuto di una data regione di interesse mediante una procedura massiva e parallela in cui ogni sequenza contribuisce all'intensità di "coverage", consentendo di definire una sequenza "consensus". Le fasi di "assembly", allineamento e analisi dei dati derivanti dalle piattaforme NGS richiedono un numero adeguato di "reads" tra loro sovrapposte, definito "coverage". In pratica, il "coverage" per una data regione di sequenza è variabile e diversi fattori, oltre alla variabile casuale di Poisson della preparazione della libreria, possono contribuire a questa variabilità, incluse la ligazione differenziale degli oligonucleotidi adattatori alle sequenze di DNA stampo e l'amplificazione differenziale durante la generazione clonale del DNA stampo (11, 70). Al di là degli errori di sequenza, un "coverage" non adeguato può portare a un mancato rilevamento di un'effettiva variazione nucleotidica, dando origine a risultati falsi-negativi per campioni eterozigoti (3, 11). Per la piattaforma 454 diversi studi hanno dimostrato che "coverage" minori di 20-30 volte sono sufficienti a ridurre l'accura-

tezza nell'identificazione di un polimorfismo (65). Per il formato Illumina sono state utilizzate intensità di "coverage" 50-60 volte più elevate in modo da migliorare l'allineamento delle "short reads", l'"assembly" e l'accuratezza, sebbene un "coverage" di 20-30 volte possa essere sufficiente per alcune applicazioni di risequenziamento (14). Come osservato in precedenza, uno studio comparativo sul genoma di lievito ha mostrato come le tecnologie 454, Illumina e SOLiD abbiano tutte identificato variazioni di singole basi in maniera corretta, quando l'intensità di "coverage" per allele era ≥15 volte (20). Nel caso di sequenze che non vengono allineate a causa di una consistente diversità rispetto alla sequenza di riferimento si possono verificare delle lacune di copertura ("coverage gaps"). L'allineamento di sequenze ripetute a regioni di interesse contenenti tali sequenze può influire sul "coverage" apparente. In base al "software" di allineamento impiegato, le "reads" che si allineano ugualmente bene a più siti possono essere assegnate in modo casuale a tali siti o in alcuni casi possono essere scartate. Con "software" per il "de novo assembly", le "reads" che mostrano allineamenti ambigui di norma non sono considerate, generando gruppi multipli di "reads" allineate ("contigs"), senza alcuna informazione su come sono ordinati.

Sono stati sviluppati e resi disponibili per la comunità scientifica un grande numero di programmi per l'allineamento e l'"assembly", la maggior parte dei quali utilizza il sistema operativo Linux, mentre alcuni sono disponibili per Windows. Molti richiedono un sistema operativo a 64-bit e possono usare "random access memory" (RAM) ≥16 MB e un'unità centrale di elaborazione (CPU) "multicore". La quantità di dati, l'"hardware", i pacchetti "software" e le impostazioni portano i tempi di elaborazione da pochi minuti a diverse ore, sottolineando la necessità di una adeguata potenza di calcolo. Anche se sta crescendo il numero di algoritmi disponibili per l'allineamento e l'"assembly", occorre valutare se sia preferibile la velocità o l'accuratezza con cui sono valutati molti ma non tutti i possibili allineamenti e cercare di trovare un equilibrio tra allineamento ideale e efficienza computazionale.

Le caratteristiche dei "software" per le piattaforme NGS variano in base al tipo di applicazione e, in generale, possono includere allineamento, "de novo assembly", visualizzazione dell'allineamento e programmi per l'identificazione di variazioni. Inoltre sono in fase di sviluppo alcuni strumenti per l'analisi statistica dei dati ottenuti tramite le piattaforme NGS (come JMP Genomics, SAS Institute). Pacchetti "software" disponibili per l'allineamento e l'"assembly" a una sequenza di riferimento includono Zoom (71), MAQ (67), Mosaik (72), SOAP (73) e SHRiMP (http://compbio.cs.toronto.edu/shrimp/), che consente l'analisi per la piattaforma SOLiD. Software per il "de novo assembly" comprendono Edina (70) EULER-SR (74), SHARCGS (75), SSAKE (69), Velvet (76) e SOAPdenovo (http://soap.genomics.org.cn/). Un recente "software" commerciale per i processi di allineamento e il "de novo assembly" include pacchetti forniti da DNAStar (www.dnastar.com), SoftGenetics (www.softgenetics.com) e CLC bio (www.clcbio.com), che consentono la visualizzazione dei dati per la rappresentazione dell'allineamento delle "reads", dell'intensità di "coverage", delle annotazioni genomiche e delle analisi di variazioni. La Figura 6 mostra alcuni esempi di come i dati derivanti da piattaforme NGS possono essere visualizzati mediante l'impiego di 2 differenti "software".

L'analisi dei dati ottenuti mediante la tecnologia RNA-Seq rappresenta una sfida particolare e richiede l'allineamento di sequenza sia a livello delle regioni dei trascritti in cui è avvenuto lo "splicing", sia a livello delle code di poli(A). Tuttavia, il "software" attualmente in uso ha consentito analisi approfondite, quali il riconoscimento di sequenze "consensus" a livello delle giunzioni di "splicing" e di regioni di giunzione introne-esone con un basso "coverage" di allineamento (41). L'identificazione delle diverse isoforme dei trascritti prevede la mappatura delle "reads" a giunzioni di "splicing" note e putative e richiede che ogni isoforma sia supportata da "reads" contenenti giunzioni di "splicing" indipendenti e multiple con siti di inizio di trascrizione differenti (51). Il "software" Erange è stato utilizzato per l'analisi del trascrittoma di topo (43). Questo "software" mappa sia "reads" uniche ai loro siti di origine genomica che "reads" che si allineano a diversi siti, o "multireads", che si allineano al sito di origine più probabile. "Reads" che non mappano a esoni noti sono raggruppate in base all'omologia con interi esoni o parti di esoni candidati. La natura semiproporzionale dei dati relativi al trascrittoma derivanti da piattaforme NGS permette la quantificazione del RNA prodotto in base al "coverage" dei dati assemblati e allineati. Per quantificare i trascritti, Erange utilizza valori normalizzati di "reads" uniche, in cui si sono verificati fenomeni di "splicing", e "multireads". Per studi sui microRNA, incluse analisi di struttura secondaria per brevi RNA a forcina, allineamenti a banche dati di microRNA noti e ricerche tramite piattaforme NGS per filamenti di microRNA complementari, sono necessarie ulteriori considerazioni analitiche, come riportato in studi per lo sviluppo di chicchi di riso (77) e di embrioni di pollo (78).

L'applicazione della tecnologia ChIP-Seq ha portato a metodi di analisi e "software" che sfruttano i vantaggi del ChIP-in-chip e permettono di ottenere una quantità maggiore di dati. La risoluzione a livello della singola base consente una stima migliore della posizione del sito di legame mediante l'uso di programmi quali QuEST (79) e MACS (80). I dati allineati alle regioni di legame delle proteine presentano tipicamente due picchi caratteristici, ognuno dei quali è costituito o solo da "reads" senso ("forward") o antisenso ("reverse"). Questi picchi sono dei marcatori distintivi dei corti frammenti di DNA immunoprecipitati ottenuti con la tecnologia ChIP-Seq, che presentano un sito di legame vicino al centro e sono utilizzati dal "software" per stimare la posizione del sito di legame prossima alla posizione media del picco. Caratteristiche addizionali del programma includono innovazioni nelle analisi statistiche per minimizzare gli errori nell'identificazione dei siti di legame e migliorare la stima della probabilità di errore e l'analisi dei motivi.

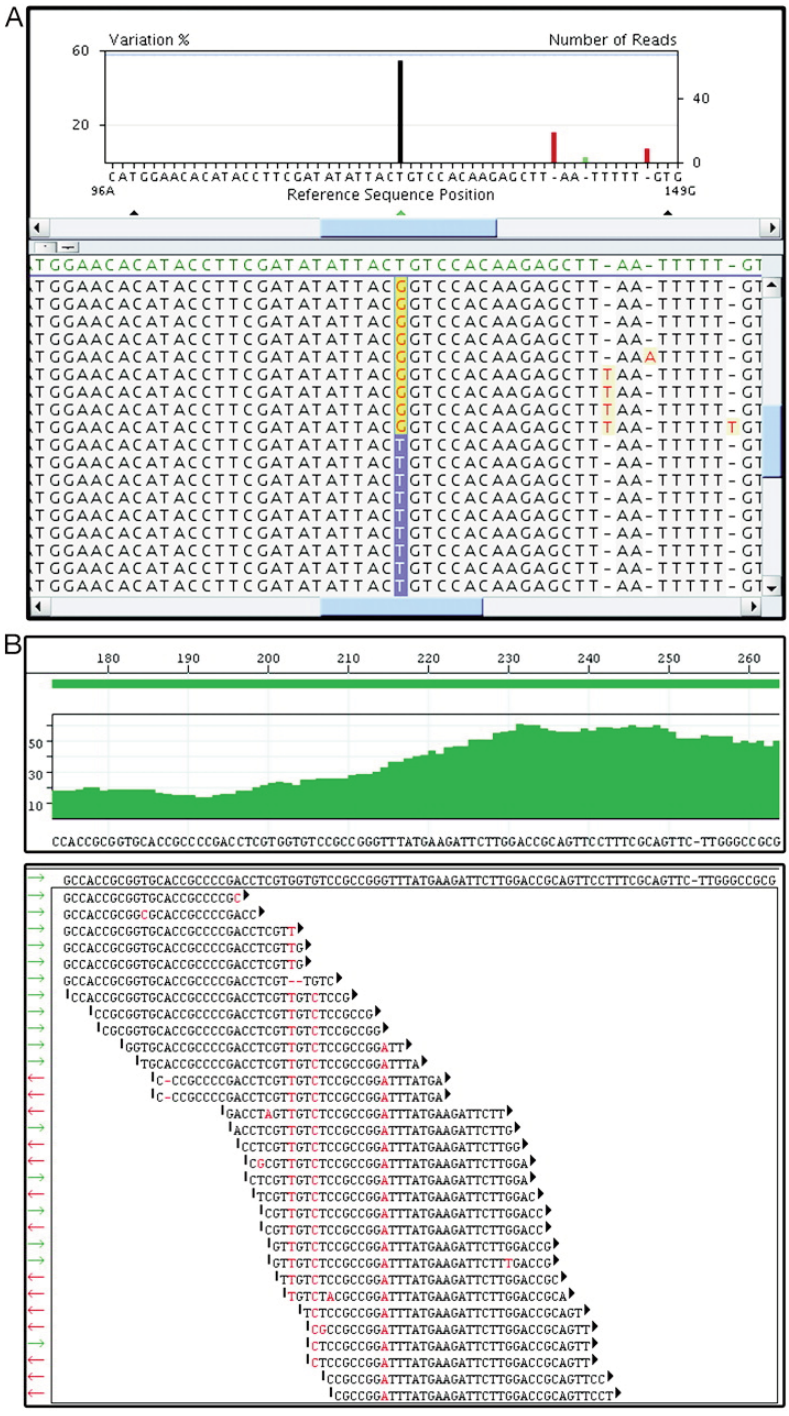

Figura 6

Esempi di visualizzazione dei dati derivanti da piattaforme NGS ottenuti mediante l'impiego di 2 differenti "software".
(A) Visualizzazione dei dati derivanti dalla piattaforma GS FLX, relativi al gene "cystic fibrosis transmembrane conductance regulator" (CFTR) (ATP-binding cassette sub-family C, member 7), mediante il software "Amplicon Variant Analyzer" (Roche). Il pannello in basso mostra l'allineamento di 18/68 "reads" alla sequenza di riferimento (indicata in alto). La colonna evidenziata mostra la presenza nelle "reads" di un polimorfismo (SNP) in eterozigosi. Inserzioni singole T/A, che non si allineano con la sequenza di riferimento, possono indicare degli errori. Il pannello in alto rappresenta la percentuale di variazione rispetto alla sequenza di riferimento (barre verticali) e il "coverage" ("number of reads"). (B) Visualizzazione dei dati derivanti dalla piattaforma Illumina, relativi al Mycobacterium massiliense, mediante il software DNAStar SeqMan Pro. Il pannello in basso mostra l'allineamento delle "reads" alla sequenza di riferimento, indicata in alto. Le frecce a sinistra indicano l'orientamento della sequenza; sono visualizzate le basi identificate come differenti dalla sequenza di riferimento (in rosso). Le tre colonne identiche (in rosso) visualizzano i presunti SNP. Le altre basi in rosso possono essere degli errori. Il pannello in alto rappresenta il "coverage" delle "reads" con le posizioni dell'allineamento relativo indicate sopra il grafico. Per le informazioni relative all'analisi del gene CFTR, effettuata su campioni residui di DNA non identificabili, vedere i "Ringraziamenti".

PROSPETTIVE CLINICHE FUTURE

Dall'impatto che le piattaforme NGS hanno avuto nel campo della ricerca di base è possibile prevedere quale potrà essere la loro applicabilità nella diagnostica molecolare. I temi chiave che dovranno essere affrontati in questa fase di transizione includono la complessità delle procedure tecniche, la robustezza, l'accuratezza e i costi. Sulla base di tutte queste valutazioni, le piattaforme NGS trarranno il beneficio dall'ottimizzazione di un processo in continua evoluzione, che include l'automazione, il perfezionamento della chimica, la riduzione dei costi e il miglioramento nella gestione dei dati. Attualmente il costo per le analisi effettuate mediante le piattaforme NGS risulta essere consistente in termini di investimento per l'attrezzatura (da circa 600.000 dollari per le piattaforme Roche 454 Life Sciences, Illumina e Applied Biosystems SOLiD a 1,35 milioni di dollari per il formato HeliScope) e nei reagenti per la reazione di sequenziamento (da circa 3500-4500 dollari per le piattaforme Illumina, Applied Biosystems e Roche/454 a 18.000 dollari per il formato HeliScope). Tuttavia, il costo per base è sostanzialmente inferiore rispetto a quello richiesto dal sequenziamento col metodo Sanger e questo dato, combinato all'enorme quantità di dati che sono prodotti, rende facile capire perché centri per lo studio del genoma, strutture di base e aziende con contratti di sequenziamento abbiano prontamente adottato questa nuova tecnologia. Considerazioni sul flusso di lavoro includono il fatto che la preparazione di una libreria di DNA stampo richiede passaggi multipli di biologia molecolare e, a seconda del tipo di piattaforma, 2-4 giorni per essere completata. In aggiunta alle competenze di biologia molecolare, l'analisi dei dati comporta anche la necessità di competenze di bioinformatica e la conoscenza del sistema operativo Linux. La possibilità per le tecnologie NGS di effettuare analisi ad alta processività può essere agevolata analizzando campioni multipli in compartimenti separati della cella a flusso. Inoltre, sequenze univoche di identificazione o "codici a barre" possono essere legati ai singoli campioni, che possono essere successivamente mescolati e sequenziati. Dopo sequenziamento, le sequenze dei singoli campioni sono dedotte sulla base della decodifica dei dati (81-83).

Nei laboratori più grandi l'applicazione delle tecnologie NGS nella diagnostica clinica è nelle prime fasi di sviluppo e viene impiegata per studi che richiedono una grande quantità di informazioni di sequenza, quantificazioni relative e rilevazioni ad alta sensibilità. Esempi che rispondono a questi criteri comprendono l'identificazione, precedentemente riportata, di mutazioni in cellule tumorali derivanti dal circolo sanguigno o da biopsie. Nell'ambito delle malattie mitocondriali, le tecnologie NGS possono essere utilizzate per sequenziare l'intero genoma mitocondriale di 16,5 Kb, determinare la percentuale di mutazioni eteroplasmiche e analizzare i geni nucleari i cui i prodotti proteici influenzano il metabolismo mitocondriale, il tutto in un'unica seduta analitica. Nel laboratorio degli Autori è in corso il sequenziamento dei genomi micobatterici come approccio per migliorare l'identificazione dei microorganismi e agevolare le indagini cliniche di tipo epidemiologico. E' stata riportata l'identificazione della quasi-specie HIV e la sua relativa quantificazione, incluso il fatto che questa possa essere utilizzata per monitorare l'insorgenza di una resistenza al farmaco (84). Per quanto riguarda la genetica umana, c'è una necessità crescente di analizzare geni multipli che, se mutati, portano alla manifestazione di fenotipi clinici che si sovrappongono. Per esempio, nella patogenesi della cardiomiopatia ipertrofica sono coinvolti 16 geni differenti (85, 86). In tali contesti, per una diagnosi approfondita, sarà necessario sequenziare più di 100.000-200.000 bp. Per questo tipo di sfida tecnica, l'associazione delle piattaforme NGS ai metodi di arricchimento genomico sopra descritti offrono un approccio promettente.

Recentemente, i gruppi di ricerca di Y.M. Dennis Lo e Stephen Quake hanno applicato le tecnologie NGS all'identificazione di aneuploidie cromosomiche fetali (87, 88). Studi precedenti avevano dimostrato che nel sangue materno durante la gravidanza, insieme ad acidi nucleici (DNA e RNA) liberi di derivazione materna, sono presenti acidi nucleici liberi di origine fetale. Sono stati sviluppati diversi approcci analitici che utilizzano acidi nucleici liberi fetali per determinare la presenza di aneuploidie nel feto, inclusa l'analisi del mRNA placentare derivante dai cromosomi di interesse (ad es. il cromosoma 21) e la determinazione del dosaggio cromosomico relativo di un gran numero di loci di interesse rispetto a loci di riferimento mediante analisi di "PCR digitale" (89-91). Definito il concetto di dosaggio cromosomico relativo, i gruppi di Lo e Quake hanno mostrato, in maniera indipendente, la fattibilità della trasformazione del DNA libero nel plasma materno in una libreria di DNA stampo per la piattaforma Illumina, successivamente sequenziata e mappata a una sequenza di riferimento di genoma umano. La valutazione del numero delle "reads" che mappano su ciascun cromosoma permette di definire il dosaggio relativo di ogni cromosoma. Se fosse presente un'aneuploidia nel feto, il numero delle "reads" che mappano sul cromosoma colpito dovrebbe essere statisticamente in eccesso nella serie dei dati analizzati. Questa previsione è stata confermata per le gravidanze con trisomia 21, con ulteriori elementi di prova ottenuti per le gravidanze con trisomia 18 e 13. Questi studi hanno aperto una nuova strada per l'identificazione delle aneuploidie fetali e forniscono un punto di partenza per l'analisi, in condizioni fisiologiche e fisiopatologiche, del DNA libero nel circolo mediante piattaforme NGS.

TECNOLOGIE ALL'ORIZZONTE

Le nuove tecnologie di sequenziamento a singola molecola, in fase di sviluppo, possono diminuire il tempo di sequenziamento, ridurre i costi e semplificare la preparazione del campione. Il sequenziamento-per-sintesi in tempo reale è stato sviluppato da VisiGen (http://www.visigenbio.com) e Pacific Biosciences (http://www.pacificbiosciences.com). L'approccio introdotto da VisiGen prevede l'utilizzo di una DNA polimera-

si modificata con una molecola fluorescente, che funge da "donatore". La polimerasi, immobilizzata su una superficie di una "slide" di vetro, catalizza l'estensione di un nuovo filamento di DNA a partire dal "primer" ibridato ai DNA stampo. I nucleotidi sono marcati con molecole fluorescenti che fungono da "accettori" e durante l'incorporazione delle basi viene utilizzata energia luminosa per eccitare la molecola "donatore" e dare origine al fenomeno di "trasferimento di energia per risonanza" ["fluorescence resonance energy transfer" (FRET)] tra la polimerasi e i gruppi funzionali fluorescenti (accettori) dei nucleotidi, che, essendo legati al gruppo fosfato in posizione γ, vengono rilasciati durante la formazione del legame fosfodiesterico. L'azienda prevede che la piattaforma sarà costituita da un "array" massivo e parallelo di molecole di DNA polimerasi immobilizzate che consentirà di generare 1 x 10^6 bp di sequenza al secondo.

Pacific Biosciences effettua un sequenziamento a singola molecola in tempo reale che utilizza dNTP marcati con molecole fluorescenti legate ai gruppi fosfato. Il sequenziamento del DNA viene eseguito all'interno di migliaia di pozzetti di reazione di 50-100 nm di diametro, che sono fabbricati con una sottile pellicola di rivestimento in metallo depositata su una struttura che indirizza le onde elettromagnetiche nello spettro del visibile ("optical waveguide"), costituita da un substrato solido trasparente di biossido di silicio. Ogni pozzetto di reazione rappresenta una camera nanofotonica in cui solo la terza parte inferiore è permeabile alla luce, producendo un volume di rilevazione di circa 20 zeptolitri (20 x 10^{-21} L). I complessi DNA polimerasi/DNA stampo sono immobilizzati sul fondo dei pozzetti e sono aggiunti i dNTP marcati con 4 fluorocromi differenti. Poiché la DNA polimerasi incorpora i nucleotidi complementari, ogni base resta all'interno del volume di rilevazione per 10 millisecondi, tempo maggiore rispetto a quello necessario affinché un nucleotide diffonda all'interno e all'esterno del volume di rilevazione. L'eccitazione laser permette che gli eventi di incorporazione che avvengono nei singoli pozzetti siano acquisiti attraverso l'"optical waveguide" e la lunghezza d'onda del segnale di fluorescenza rilevata corrisponde all'incorporazione di un dNTP specifico. Per il sequenziamento, Pacific Biosciences utilizza una DNA polimerasi modificata del batteriofago phi29 (DNA polimerasi phi29), da cui sono state potenziate le proprietà cinetiche per il sistema di incorporazione dei dNTP marcati con molecole fluorescenti legate a gruppi fosfato. Inoltre, la DNA polimerasi phi29 è un enzima altamente processivo e con attività di scalzamento del filamento di DNA ("strand-displacement"). Avvalendosi di queste proprietà, Pacific Biosciences ha dimostrato come, utilizzando come DNA stampo una molecola di DNA circolare a singolo filamento, sia possibile sequenziare "reads" superiori alle 4000 basi di lunghezza. In tale configurazione, la DNA polimerasi phi29 effettua fasi multiple di sintesi con "strand-displacement" di tutto il DNA stampo circolare. Il tasso medio di sintesi del DNA stampo è stato stimato di circa 4 basi/s. Gli errori osservati, comprese delezioni, inserzioni ed errate incorporazioni nucleotidiche, possono essere compensati tramite la generazione di una sequenza "consensus" derivata da cicli multipli di sequenziamento del DNA stampo. Sono in fase di sviluppo ulteriori miglioramenti nella chimica e nella strumentazione della piattaforma, che si prevede venga messa in commercio nel 2010 (92-94).

Un sistema ancora più avveniristico è rappresentato dal sequenziamento basato sul monitoraggio del passaggio di molecole di DNA attraverso dei nanopori di 2-5 nm o più di diametro. I nanopori vengono fabbricati all'interno di membrane inorganiche (nanopori allo stato solido), assemblando canali proteici in membrane lipidiche, o disposti in canali nanofluidici a base polimerica. In alcuni formati, la corrente è applicata attraverso le membrane dei nanopori per guidare la traslocazione di molecole di DNA cariche negativamente attraverso i pori stessi, mentre vengono monitorati i cambiamenti nella conduttanza elettrica della membrana, misurati nell'ordine dei picoampère. La NABsys (http://www.nabsys.com) si sta dedicando a un approccio che associa la tecnologia dei nanopori a quella del sequenziamento tramite ibridazione in cui singole molecole di DNA stampo a singolo filamento sono ibridate a una libreria di esameri di una sequenza nota. Il DNA ibridato è interrogato attraverso un nanoporo, in base ai differenti cambiamenti di corrente nelle regioni in cui gli esameri si sono appaiati. I risultati di ibridazione vengono utilizzati per mappare le regioni ibridate e determinarne le sequenze. Oxford Nanopore Technologies (http://www.nanoporetech.com) sta sviluppando un metodo di sequenziamento basato sulla tecnologia dei nanopori, che utilizza un canale proteico di α-emolisina in un doppio strato lipidico ricostituito. I nanopori sono situati in pozzetti indipendenti di un "array", in cui sono introdotte singole molecole di DNA che vengono progressivamente digerite tramite l'azione di una esonucleasi. Le singole basi nucleotidiche rilasciate entrano nel nanoporo e alterano la corrente elettrica, creando un cambio di corrente caratteristico per ciascuna base (95, 96). Anche se l'approccio è apparentemente futuristico, NHGRI sta stanziando notevoli finanziamenti per sviluppare varie tecnologie basate sull'utilizzo dei nanopori nell'ambito del programma che ha come obiettivo il sequenziamento del genoma umano con un costo pari a 1000 dollari. Per ulteriori informazioni relative alla tecnologia dei nanopori si consultino le rassegne recentemente pubblicate (97, 98).

CONCLUSIONI

Negli ultimi anni abbiamo assistito alla comparsa delle tecnologie NGS che condividono come caratteristica comune il sequenziamento massivo parallelo di molecole di DNA amplificate clonalmente. Nel 2008 è stata lanciata la prima piattaforma NGS basata sul sequenziamento a singola molecola. Come prospettive future sono in fase di studio tecnologie di sequenziamento a singola molecola in tempo reale e approcci basati sulla tecnologia dei nanopori. Le piattaforme NGS hanno avuto un impatto sostanziale sulla ricerca genomica di base in termini di analisi su vasta scala e fattibilità. Per i prossimi anni è previsto un passaggio di tali piattaforme ad appli-

cazioni di diagnostica clinica. Elementi essenziali per rendere questa transizione possibile sono l'esigenza di ottimizzare i processi, in particolare la fase di preparazione del campione, associata a miglioramenti nella robustezza della tecnologia e nella caratterizzazione dell'accuratezza attraverso studi di validazione. L'enorme quantità di dati di sequenza generati rappresenteranno una sfida bioinformatica per il laboratorio clinico. Oltre all'elaborazione dei dati, l'interpretazione dei risultati di sequenza richiederà una caratterizzazione ulteriore delle variazioni genomiche presenti nelle regioni analizzate. Sebbene il trasferimento delle piattaforme NGS alla diagnostica clinica richieda ancora un lavoro notevole, le applicazioni potenziali sono numerose ed entusiasmanti.

RINGRAZIAMENTI

L'analisi del gene CFTR, mostrata nella Figura 6, è stata effettuata su campioni residui di DNA non identificabili, secondo l'approvazione del Comitato Etico dell'Università dello Utah, protocollo per soggetti umani no. 7275.

BIBLIOGRAFIA

1. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA 1977;74:560-4.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977;74:5463-7.
3. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872-6.
4. Nyren P, Pettersson B, Uhlen M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. Anal Biochem 1993;208:171-5.
5. Ronaghi M, Karamohamed S, Pettersson B, et al. Real-time DNA sequencing using detection of pyrophosphate release. Anal Biochem 1996;242:84-9.
6. Ronaghi M, Uhlen M, Nyren P. A sequencing method based on real-time pyrophosphate. Science 1998;281:363-5.
7. Tawfik DS, Griffiths AD. Man-made cell-like compartments for molecular evolution. Nat Biotechnol 1998;16:652-6.
8. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
9. Pearson BM, Gaskin DJ, Segers RP, et al. The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828). J Bacteriol 2007;189:8402-3.
10. Huse SM, Huber JA, Morrison HG, et al. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007;8:R143.
11. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53-9.
12. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420-6.
13. Campbell PJ, Stephens PJ, Pleasance ED, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722-9.
14. Dohm JC, Lottaz C, Borodina T, et al. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;36:e105.
15. Erlich Y, Mitra PP, delaBastide M, et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 2008;5:679-82.
16. Quail MA, Kozarewa I, Smith F, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods 2008;5:1005-10.
17. Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 2005;309:1728-32.
18. Braslavsky I, Hebert B, Kartalov E, et al. Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 2003;100:3960-4.
19. Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. Science 2008;320:106-9.
20. Smith DR, Quinlan AR, Peckham HE, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18:1638-42.
21. Quinn NL, Levenkova N, Chow W, et al. Assessing the feasibility of GS FLX pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics 2008;9:404.
22. Satkoski JA, Malhi R, Kanthaswamy S, et al. Pyrosequencing as a method for SNP identification in the rhesus macaque (Macaca mulatta). BMC Genomics 2008;9:256.
23. Borneman AR, Forgan AH, Pretorius IS, et al. Comparative genome analysis of a Saccharomyces cerevisiae wine strain. FEMS Yeast Res 2008;8:1185-95.
24. Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. Nature 2008;456:60-5.
25. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66-72.
26. Kim PM, Lam HY, Urban AE, et al. Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history. Genome Res 2008;18:1865-74.
27. Chen J, Kim YC, Jung YC, et al. Scanning the human genome at kilobase resolution. Genome Res 2008;18:751-62.
28. Yeager M, Xiao N, Hayes RB, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008;124:161-70.
29. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008;455:1069-75.
30. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903-5.
31. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522-7.
32. Okou DT, Steinberg KM, Middle C, et al. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;4:907-9.
33. Porreca GJ, Zhang K, Li JB, et al. Multiplex amplification of large sets of human exons. Nat Methods 2007;4:931-6.
34. Huber JA, Mark Welch DB, Morrison HG, et al. Microbial population structures in the deep marine biosphere. Science 2007;318:97-100.
35. Sogin ML, Morrison HG, Huber JA, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci USA 2006;103:12115-20.
36. Urich T, Lanzen A, Qi J, et al. Simultaneous assessment of soil microbial community structure and function through

analysis of the meta-transcriptome. PLoS ONE 2008;3:e2527.
37. Palacios G, Druce J, Du L, et al. A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 2008;358:991-8.
38. Keijser BJ, Zaura E, Huse SM, et al. Pyrosequencing analysis of the oral microflora of healthy adults. J Dent Res 2008;87:1016-20.
39. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature 2008;457:480-4.
40. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.
41. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 2008;320:1344-9.
42. Wilhelm BT, Marguerat S, Watt S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 2008;453:1239-43.
43. Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621-8.
44. Lister R, O'Malley RC, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008;133:523-36.
45. Cloonan N, Forrest AR, Kolle G, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 2008;5:613-9.
46. Marioni JC, Mason CE, Mane SM, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18:1509-17.
47. Morin R, Bainbridge M, Fejes A, et al. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 2008;45:81-94.
48. Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 2008;18:610-21.
49. Emrich SJ, Barbazuk WB, Li L, et al. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 2007;17:69-73.
50. Pan Q, Shai O, Lee LJ, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008;40:1413-5.
51. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456:470-6.
52. Ren B, Robert F, Wyrick JJ, et al. Genome-wide location and function of DNA binding proteins. Science 2000;290:2306-9.
53. Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell 2007;129:823-37.
54. Schones DE, Zhao K. Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 2008;9:179-91.
55. Valouev A, Ichikawa J, Tonthat T, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 2008;18:1051-63.
56. Johnson DS, Mortazavi A, Myers RM, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497-502.
57. Robertson G, Hirst M, Bainbridge M, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 2007;4:651-7.
58. Korshunova Y, Maloney RK, Lakey N, et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 2008;18:19-29.
59. Cokus SJ, Feng S, Zhang X, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008;452:215-9.
60. Marguerat S, Wilhelm BT, Bahler J. Next-generation sequencing: applications beyond genomes. Biochem Soc Trans 2008;36:1091-6.
61. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008;24:142-9.
62. Rougemont J, Amzallag A, Iseli C, et al. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics 2008;9:431.
63. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186-94.
64. Ewing B, Hillier L, Wendl MC, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175-85.
65. Brockman W, Alvarez P, Young S, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 2008;18:763-70.
66. Smith AD, Xuan Z, Zhang MQ. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 2008;9:128.
67. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 2008;18:1851-8.
68. Whiteford N, Haslam N, Weber G, et al. An analysis of the feasibility of short read sequencing. Nucleic Acids Res 2005;33:e171.
69. Warren RL, Sutton GG, Jones SJ, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007;23:500-1.
70. Hernandez D, Francois P, Farinelli L, et al. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 2008;18:802-9.
71. Lin H, Zhang Z, Zhang MQ, et al. ZOOM! Zillions Of Oligos Mapped. Bioinformatics 2008;24:2431-7.
72. Smith DR, Quinlan AR, Peckham HE, et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. Genome Res 2008;18:1638-42.
73. Li R, Li Y, Kristiansen K, et al. SOAP: short oligonucleotide alignment program. Bioinformatics 2008;24:713-4.
74. Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. Genome Res 2008;18:324-30.
75. Dohm JC, Lottaz C, Borodina T, et al. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res 2007;17:1697-706.
76. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18:821-9.
77. Zhu QH, Spriggs A, Matthew L, et al. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. Genome Res 2008;18:1456-65.
78. Glazov EA, Cottee PA, Barris WC, et al. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res 2008;18:957-64.
79. Valouev A, Johnson DS, Sundquist A, et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 2008;5:829-34.
80. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
81. Binladen J, Gilbert MT, Bollback JP, et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2007;2:e197.

82. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. Nat Protoc 2008;3:267-78.
83. Meyer M, Stenzel U, Myles S, et al. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 2007;35:e97.
84. Wang C, Mitsuya Y, Gharizadeh B, et al. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 2007;17:1195-201.
85. Fokstuen S, Lyle R, Munoz A, et al. A DNA resequencing array for pathogenic mutation detection in hypertrophic cardiomyopathy. Hum Mutat 2008;29:879-85.
86. Morita H, Rehm HL, Menesses A, et al. Shared genetic causes of cardiac hypertrophy in children and adults. N Engl J Med 2008;358:1899-908.
87. Chiu RW, Chan KC, Gao Y, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci U S A 2008;105:20458-63.
88. Fan HC, Blumenfeld YJ, Chitkara U, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 2008;105:16266-71.
89. Fan HC, Quake SR. Detection of aneuploidy with digital polymerase chain reaction. Anal Chem 2007;79:7576-9.
90. Lo YM, Lun FM, Chan KC, et al. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc Natl Acad Sci USA 2007;104:13116-21.
91. Dennis Lo YM, Chiu RW. Prenatal diagnosis: progress through plasma nucleic acids. Nat Rev Genet 2007;8:71-7.
92. Korlach J, Marks PJ, Cicero RL, et al. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci USA 2008;105:1176-81.
93. Levene MJ, Korlach J, Turner SW, et al. Zero-mode waveguides for single-molecule analysis at high concentrations. Science 2003;299:682-6.
94. Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science 2008;323:133-8.
95. Astier Y, Braha O, Bayley H. Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 2006;128:1705-10.
96. Wu HC, Astier Y, Maglia G, et al. Protein nanopores with covalently attached molecular adapters. J Am Chem Soc 2007;129:16142-8.
97. Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol 2008;26:602-11.
98. Rhee M, Burns MA. Nanopore sequencing technology: research trends and applications. Trends Biotechnol 2006;24:580-6.