Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria dell'Informazione

Analisi Critica di Tecniche di Sequenziamento di Nuova Generazione

Laureando:

Giorgio Vitale

Relatore:

Dott.ssa Cinzia Pizzi

Anno accademico 2012/2013

Indice

Sommario

Le tecnologie di sequenziamento di nuova generazione, con la capacità di fornire, a prezzo contenuto, milioni di sequenze di DNA per *run*, hanno cambiato il modo di approcciarsi all'analisi dei genomi. L'abbattimento dei costi e dei tempi introdotti da questi strumenti hanno avuto ricadute in diversi ambiti disciplinari, come la ricerca contro il cancro, la metagenomica, la ricerca sugli animali e la selezione biologica. Queste discipline hanno tutte trovato, nelle tecniche NGS, degli strumenti potenti e validi per il loro sviluppo.
In questa tesi si compie un'analisi delle tecniche NGS, studiandone il funzionamento generale, i sistemi che le implementano, gli algoritmi grazie a cui questi strumenti funzionano, i problemi e le future possibili evoluzioni.
Come dimostrazione pratica del loro potenziale, viene anche analizzato nel dettaglio un algoritmo, SHRiMP, per l'analisi e il filtraggio delle *reads* prodotte da queste tecnologie.

1 Introduzione

Il sequenziamento del DNA è un ambito di ricerca in piena evoluzione. Per anni la maggior parte delle applicazioni si sono basate su variazioni ed evoluzioni dell'algoritmo Sanger, in grado di generare *reads* lunghe fino a 800bp.
Da una decina d'anni tuttavia le tecnologie di sequenziamento di nuova generazione(NGS) hanno dimostrato le loro potenzialità, con *reads* corte e *throughput* sempre più alti, arrivando presto a soppiantare le tecnologie precedenti. Questo, unito allo sviluppo di varie materie, dalla chimica alla bioinformatica, hanno portato alla possibilità di effettuare sequenziamenti di DNA, a prezzi contenuti, fino ad allora irrealizzabili.
La bioinformatica in questo processo di sviluppo è stata fondamentale. Grazie ad essa sono stati sviluppati molteplici algoritmi, basati su tecniche diverse, come tabelle hash, indici, *spaced-seed*, per ottimizzare analisi su moli di dati sempre più vaste. Uno dei motivi per cui le NGS sono al centro dell'attenzione delle tecnologie del settore è che, con le loro caratteristiche e le loro potenzialità, hanno consentito a diverse discipline di svilupparsi in modi che fino a pochi anni fa erano impensabili. Tuttora le NGS trovano applicazione in ambiti diversi, dalla ricerca contro il cancro, all'analisi del DNA umano, allo studio sugli animali.
La tesi è strutturata come segue. Nel capitolo 2 si descrive il metodo Sanger, passando poi alle tecnologie di nuova generazione, con una descrizione del metodo utilizzato e con un analisi degli strumenti in commercio basati su queste tecniche.
Nel capitolo 3 si fa una panoramica sugli algoritmi bioinformatici che le tecniche NGS implementano e i problemi che devono risolvere.
Il capitolo 4 si concentra sulle applicazioni, mostrando come le NGS vengano utilizzate in ambiti differenti. Le limitazioni e i problemi che sorgono nell'utilizzo di queste tecniche sono descritti nel capitolo 6.
Nel capitolo 5 viene proposto lo studio di un algoritmo, SHRiMP, basato su *spaced-seed* e ottimizzato per l'analisi di *reads* corte, il cui obiettivo è quello di selezionare i dati forniti dalle applicazioni NGS, filtrandoli secondo euristiche e correggendo eventuali errori. Oltre alla descrizione dell'algoritmo vengono riportati e commentati i risultati di alcuni esperimenti svolti dagli sviluppatori di questo algoritmo. Si conclude con alcuni cenni sulle prospettive future delle tecniche di sequenziamento e le loro applicazioni.

2 Sequenziamento di Genomi

2.1 Panoramica Storica

Il sequenziamento di DNA è composto di tre fasi: la preparazione del campione da analizzare, il sequenziamento fisico e il riassemblaggio.
La preparazione del campione consiste nello spezzare il genoma in più frammenti di lunghezza contenuta, a seconda del metodo seguito, questi frammenti possono essere amplificati in vari modi.
Nella fase di sequenziamento fisico, ogni base in ogni frammento viene identificata, creando quindi delle *reads*. Una *read* è una stringa di caratteri, appartenenti ad un alfabeto finito e ben definito. Grazie a quest'astrazione è possibile utilizzare tutti quegli algoritmi per l'analisi di stringhe, come ad esempio la ricostruzione della superstringa più lunga che contenga i frammenti, oppure gli algoritmi specializzati nel confronto tra stringhe con mismatch.
Il numero di *reads* contigue trovate si indica come *read length*.
Nel riassemblaggio, si sfruttano software bioinformatici per concatenare *reads* sovrapposte, allungando quindi la lunghezza dei frammenti. Maggiore la lunghezza della sequenza, migliori i risultati, avendo infatti dati più veritieri su cui lavorare. Molte applicazioni infatti necessitano di *reads* il più lungo possibili per fornire risultati precisi.

La prima generazione di sequenziatori genomici fu sviluppata da Sanger nel 1975 (*chain-termination method*) [15][16] e, parallelamente, da Maxam e Gilbert, nel 1977(*chemical sequencing methods*) [3].
Da questi due metodi, risultò poi che il metodo Sanger fosse meno complesso tecnicamente e più portato allo sviluppo, al contrario del metodo a sequenziamento chimico, che quindi riuscì a svilupparsi molto meno.
Dalla sua proposta, il metodo Sanger ha subito diverse evoluzioni. Al giorno d'oggi, durante la fase di preparazione, frammenti di DNA di differente lunghezza sono generati dalla stessa posizione(Figura 1A). Ogni frammento, terminante con una certa base, viene marcato. A questo punto tutti i vari frammenti vengono ricombinati insieme, completando il sequenziamento.
Tramite questo metodo si può arrivare a *reads* lunghe mediamente 800basi, con punte fino a 1000basi.

Inizialmente lo scopo di questi strumenti era arrivare a sequenziare l'intero genoma umano, purtroppo però si riscontrarono diverse limitazioni: il tempo necessario a sequenziare il DNA e l'alto costo di esecuzione rendevano proibitivo l'uso di questi strumenti su sequenze troppo lunghe(si pensi che, con le tecnologie del tempo, ci vollero 10 anni, e 3 miliardi di dollari, per sequenziare l'intero genoma umano). Al giorno d'oggi, grazie alle tecnologie di nuova generazione, questi limiti sono stati superati, arrivando a sequenziarlo in modo più economico e con tempi più rapidi.

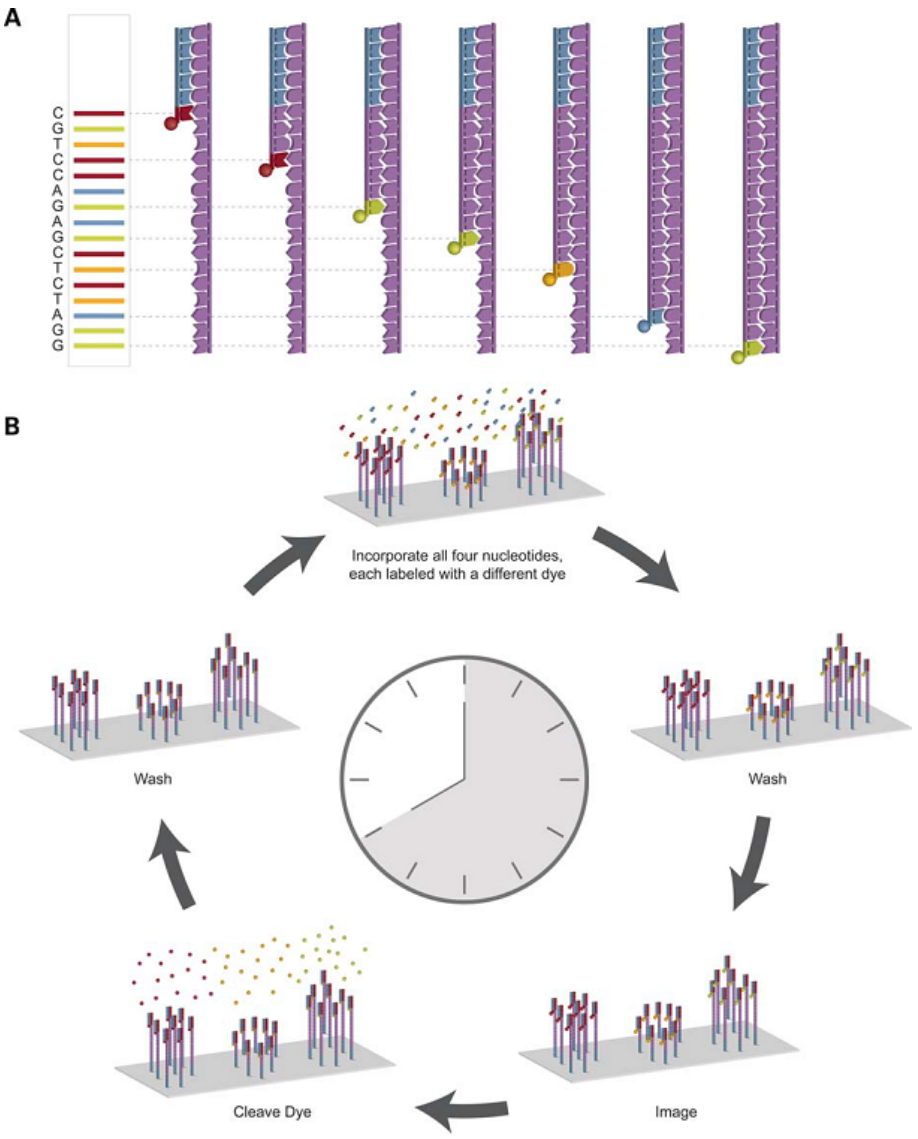

Figura 1: Metodo di funzionamento delle vecchie versioni di sequenziatori. (A) Implementazione moderna del metodo Sanger (B) Struttura del metodo Illumina.[13]

2.2 Next Generation Sequencing (NGS)

Di seguito si descrive, prima in termini generali, poi in termini specifici di ciascun sistema, il funzionamento delle tecnologie NGS.

2.2.1 Descrizione

Le tecnologie di nuova generazione per il sequenziamento del DNA consentono di ottenere velocità e throughput elevati, rendendo possibile il completamento in alcune settimane di analisi, che con il metodo Sanger avrebbero impiegato anni. Il vantaggio di queste tecnologie è infatti la possibilità di ottenere la sequenza di DNA amplificando il frammento, senza quindi doverlo clonare. Per quanto riguarda la quantità di errori di sequenziamento ottenuti, non si nota, invece, una sostanziale differenza col metodo Sanger. Tuttavia l'elevato *throughput* delle nuove tecnologie consente di sequenziare interi genomi in tempi estremamente rapidi.

Lo schema organizzativo delle NGS può essere diviso in due macro-blocchi, uno afferente alla biologia, uno alla bioinformatica. Nel primo, il genoma viene sequenziato, assegnando ai nucleotidi che compongono i frammenti la lettera iniziale del corrispondente acido nucleico (Adenina, Guanina, Citosina e Timina), ottenendo quindi delle stringhe di caratteri. Nel secondo, queste sequenze vengono analizzate, di modo da unirle, dove possibile, scartarle o correggerle in caso di errori, e quindi fornire all'utilizzatore del software dei risultati validi, con la minor quantità di informazioni superflue.

Le tecniche di sequenziamento di nuova generazione seguono tutte un particolare metodo di sviluppo, illustrato schematicamente in Figura 2.

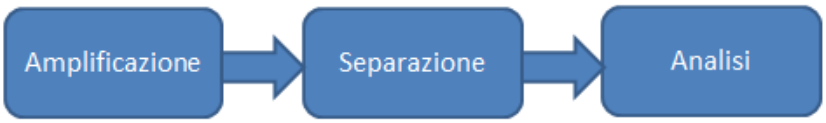

Figura 2: Passi fondamentali nell'analisi biologica delle tecnologie di nuova generazione

Amplificazione La tecnologia si basa sull'analisi della luce emessa da ogni nucleotide, che permette di identificarne il tipo. Purtroppo, la luce che ognuno di questi emette è troppo ridotta, deve quindi essere amplificata. Per amplificarla si utilizza di solito la PCR(reazione a catena della polimerasi), una tecnica

che consente la moltiplicazione, e quindi l'amplificazione, di frammenti di acidi nucleici dei quali si conoscano le sequenze nucleotidiche iniziali e finali. Si ottiene quindi un filamento amplificato, che per essere studiato deve essere separato.

Separazione Il filamento, amplificato al punto precedente, viene qui separato. La separazione può essere eseguita usando un picotiter plate(PTP), una specie di vetrino in grado di dividere i vari nucleotidi.

Analisi Una volta che i nucleotidi sono stati separati, è possibile analizzarli. L'analisi viene effettuata rilevando la luce che ogni nucleotide emette, poiché la luce emessa da ogni tipo di nucleotide è unica.

2.2.2 Sistemi commerciali per NGS

Vengono qui descritti i principali sistemi che realizzano le tecniche di sequenziamento di nuova generazione, e le cui evoluzioni fondamentali vengono riassunte in Figura 4.

Roche 454 System Di questa generazione di sequenziatori, è stato il primo a venire commercializzato, nel 2005. Questo sistema utilizza il pirosequenziamento e l'amplificazione PCR.
Inizialmente questo sistema arrivava a *reads* di 100-150bp, producendo circa 200000*reads*, con un throughput per run di 20Mb.
Nel 2008 ne è stata proposta un'evoluzione, il sequenziatore 454 GS FLX Titanium, che arrivava a *read* lunghe 700bp, con una precisione del 99.9% dopo il filtraggio, con un output di 0.7Gb per *run* in 24 ore.
Nel 2009, la combinazione del metodo GS Junior con sistema 454 GS portò l'output a 14Gb per *run*[39].
Ulteriori evoluzioni hanno portato al GS FLX+, in grado di sequenziare *read* lunghe fino a 1kb.
La grande velocità, unita alla lunghezza delle *reads* prodotte, sono i punti forti di questo sistema. Tuttavia il costo dei reagenti resta un problema da risolvere.

AB SOLiD System È stato commercializzato dalla Applied Biosystems nel 2006. Il sistema utilizza il metodo di sequenziamento a due basi, basato sul *ligation sequencing*, ossia effettua l'analisi del filamento in entrambe le direzioni.
Inizialmente la lunghezza delle *reads* era di sole 35 bp, e l'output arrivava a 3Gb per run. Grazie al sequenziamento a due basi, l'accuratezza dell'applicazione arrivava al 99.85% dopo il filtraggio.
Nel 2010 viene rilasciato il SOLiD 5500x1, con *reads length* di 85bp, precisione del 99.99% e output di 30Gb per *run*. Un *run* completo è possibile eseguirlo in 7 giorni.
Il problema principale di questo metodo è la lunghezza delle *reads*, che per molte applicazioni non è sufficiente.

Lo strumento viene utilizzato per sequenziare interi genomi, sequenziamenti mirati, epigenomica.

Illumina GA/HiSeq System Nel 2006 la Solexa rilascia l'analizzatore genomico(GA), nel 2007 l'azienda viene acquistata dall'Illumina. Il sistema sfrutta il sequenziamento per sintesi(SBS) e l'amplificazione *bridge*, un'alternativa al PCR.
All'inizio l'output dell'analizzatore era di 1Gb per *run*, portato poi a 50Gb per *run* nel 2009.
Nel 2010 viene rilasciato l'HiSeq 2000, che utilizza le stesse strategie del predecessore, arrivando però a 600Gb per *run*, ottenibile in 8 giorni. Nel prossimo futuro dovrebbe poter arrivare addirittura ad 1Tb per *run*.
Anche la lunghezza delle *reads* è stata migliorata, passando da 35bp a circa 200bp nelle ultime versioni.
Il costo per operazione è tra i più bassi, rispetto ai vari sequenziatori. Unica pecca è l'accuratezza è di circa il 98%.

Helicos single-molecule sequencing device, Heliscope È stato presentato per la prima volta nel 2007. Il metodo, a differenza dei precedenti, utilizza una tecnica che analizza le molecole singolarmente, in questo modo è possibile ottenere un'accuratezza ancora maggiore, non sporcando il genoma con reagenti chimici[20].
Anche per questo sistema si ha un throughput nell'ordine delle Gigabasi.
Lo svantaggio principale di questo metodo resta tuttavia la bassa capacità di gestire gli indels in modo corretto, con un conseguente aumento di errori.
Altro problema è la lunghezza delle *reads*, che non ha mai superato le 50bp.

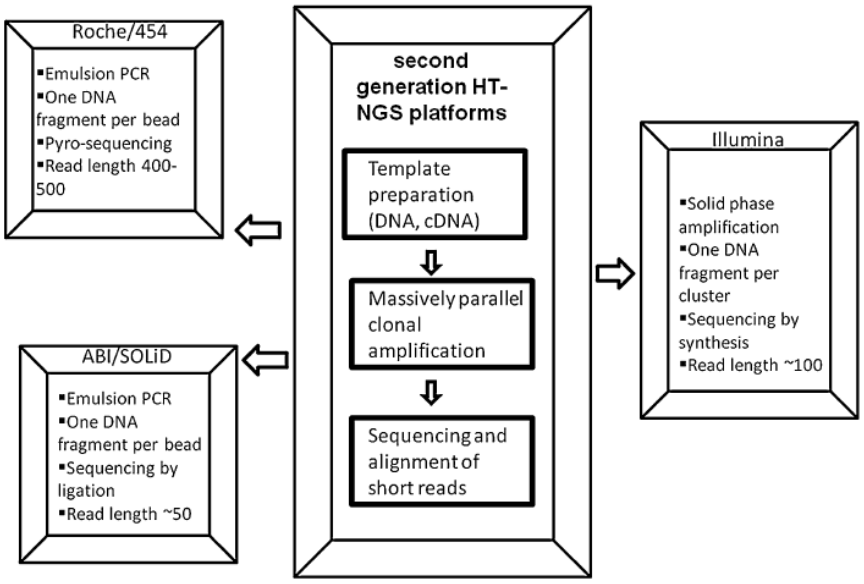

Figura 3: Schema di funzionamento delle principali piattaforme NGS. Immagine tratta da [8].

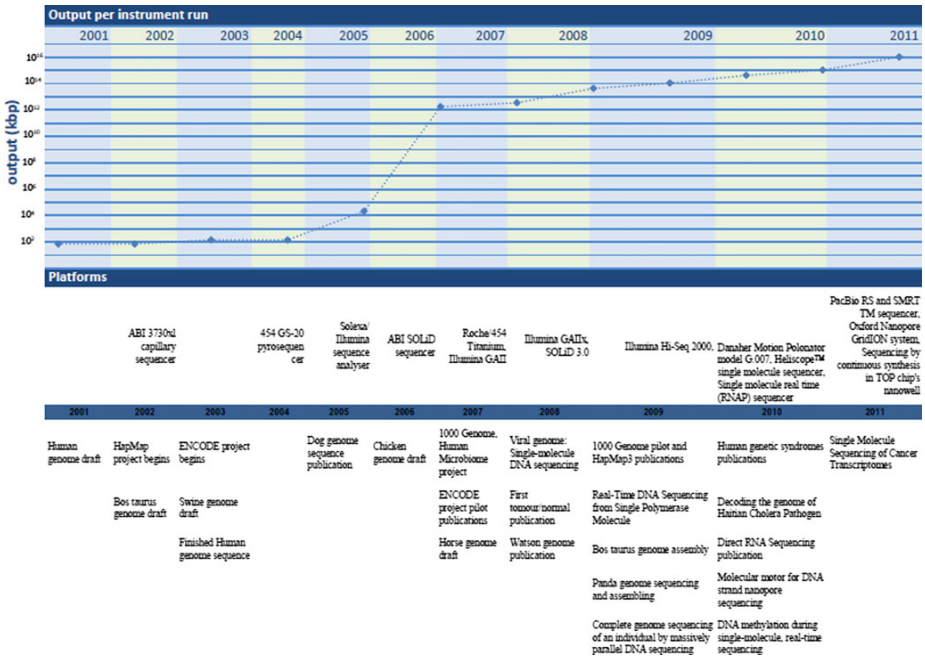

Figura 4: Evoluzioni nel sequenziamento genomico negli ultimi anni(aggiornato al 2011). In alto, scala della quantità di dati che i sistemi sono in grado di produrre per *run*. Al centro, timeline dei sequenziatori ad elevato parallelismo. In basso, timeline di progetti, pietre miliari e articoli rilevanti. Immagine tratta da [8].

3 Bioinformatica in NGS

Per quanto le varie tecnologie di sequenziamento seguano strade diverse, forniscono come output delle sequenze. Delle sottostringhe di questa sequenze, di lunghezza variabile da alcune decine fino a diverse centinaia di basi, vengono combinate insieme, di solito in un file FASTQ, un tipo di file utilizzato in biologia per memorizzare sequenze genetiche e i relativi *quality scores*, ossia il punteggio che l'algoritmo che effettua l'analisi assegna alla stringa, e che viene poi utilizzato per scegliere il match migliore contro il genoma di riferimento.
A questo punto, l'analisi bioinformatica si divide in tre passi: allineamento, che ricerca corrispondenze tra le *reads* e il genoma di riferimento, *variant calling*, che tenta di separare differenze dovute ad errori genetici dagli errori strumentali compiuti nell'analisi, filtraggio e annotazione[29](Figura 5), che tentano di allineare le *reads* al genoma di riferimento.

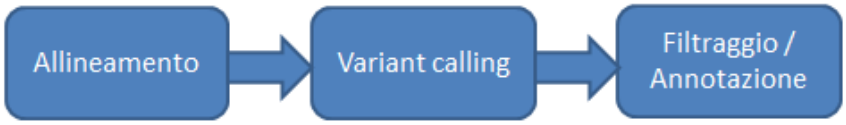

Figura 5: Passi fondamentali nell'analisi bioinformatica dell'output prodotto da tecnologie di sequenziamento di nuova generazione

3.1 Allineamento

L'allineamento è il processo attraverso cui si mappano *reads* corte ad un genoma di riferimento. Si tratta di un compito complesso, in quanto il software deve confrontare ogni *reads* con ogni posizione del DNA di riferimento. Si tratta di un passaggio computazionalmente impegnativo, e dispendioso in termini temporali.
I file SAM(Sequence Alignment Map) e BAM(Binary Alignment Map) sono gli standard di riferimento per il salvataggio dei dati ottenuti dall'allineamento per le tecnologie di nuova generazione. Esistono numerosi software in commercio, gratuiti o a pagamento, per svolgere questo compito. La maggior parte di questi usa un metodo basato su indici, che sono più veloci nella ricerca di posizioni di allineamento senza *gap* nel genoma di riferimento. Altri algoritmi consentono invece la ricerca di allineamenti con *gap*. I vari approcci al problema prevedono l'utilizzo di tabelle hash(e.g., MAQ, ELAND), algoritmi basati sulla trasformata di Burrows-Wheeler(e.g., BWA, Bowtie, SOAP2), un algoritmo di compressione reversibile che permuta l'ordine dei caratteri, senza cambiarne il valore, *genome-based hash*(e.g., Novoalign, SOAP), oppure un approccio *spaced-seeds*(e.g., SHRiMP), una variante del pattern matching che consente la presenza di un certo numero di errori in posizioni della stringa fissate.
Gli algoritmi possono fornire il risultato migliore basandosi su un'euristica, altri invece possono fornire tutti i risultati che danno un riscontro.
Si differenziano inoltre tra algoritmi che tengono conto dei gap(e.g., BWA, Bowtie2) e algoritmi che invece non ne tengono conto(e.g., MAQ, Bowtie).

3.2 Variant Calling

Dopo l'allineamento delle *reads*, il DNA in analisi può essere confrontato al genoma di riferimento, avendo quindi la possibilità di individuare le variazioni. Queste variazioni possono essere causa di malattie, oppure possono essere solo del rumore, senza alcun effetto dannoso.
La complessità di questo argomento risiede nel dover identificare le variazioni vere, ossia mutazioni del genoma, da variazioni dovute ad errori di sequenziamento. Il continuo sviluppo e perfezionamento delle tecnologie di nuova generazione porterà auspicabilmente vantaggi anche in questo ambito, migliorando l'input che l'analisi delle variazioni riceve.
La difficoltà nell'analisi delle variazioni è causata principalmente dalla presenza di indels, ovvero fenomeni di inserimenti(*insertion*) o cancellazioni(*deletion*). Questi infatti sono la causa della maggior parte dei falsi-positivi rilevati da questi algoritmi, numero che aumenta se non viene eseguito un allineamento con *gap*. Altra causa del fenomeno sono gli errori compiuti nella preparazione delle sequenze, dovuti a problemi nell'analisi PCR.
La soluzione a questo problema è aumentare la sensibilità degli strumenti nell'analisi iniziale, e migliorare i software utilizzati nell'allineamento, nonché avere un database di confronto di notevoli dimensioni, così da avere a disposizione un *set* per il confronto più ampio.

3.3 Filtraggio e annotazione

Con i passaggi precedenti viene generata una lista di migliaia di potenziali differenze tra il genoma in corso di studio e quello di riferimento. Il passaggio successivo è quindi quello di determinare quali di queste variazioni, non dovute ad errori di sequenziamento, contribuiscono al problema in corso di studio, quali cioè siano legate in qualche modo al fenomeno, eliminando tutte quelle che, secondo studi pregressi, oppure secondo euristiche, si stima non siano collegate all'oggetto in esame, riducendo quindi la quantità di materiale da analizzare.
Il metodo prevede di alterare i risultati ottenuti, rimuovendo variazioni che seguono modelli di altri fenomeni, e annotare, cercando informazioni di variazioni, identificando quelle che hanno un riscontro col processo in esame.
Il filtraggio può essere eseguito col confronto delle *reads* contro un albero genetico di riferimento, cercando quindi tutti quegli elementi di cui si conosce la funzione, oppure, nel caso di cancro ad esempio, analizzando il genoma sequenziato di tessuti malati e tessuti sani dell'individuo, escludendo quelle variazioni presenti in entrambi i tessuti.
Oltre al filtraggio, l'annotazione fornisce un altro strumento per selezionare e restringere il campione in esame, effettuando ricerche su studi precedenti, oppure applicando modelli degli effetti funzionali predetti.
L'efficacia e il basso costo di questi strumenti sta portando alla scoperta di un numero di variazioni genetiche sempre più ampio, dando la possibilità di identificare le cause delle varie malattie. Per esempio, è stato verificato che circa 1300 posizioni di un gene o di una sequenza all'interno di un cromosoma sono associate a 200 malattie, ma che solo una parte di queste sia causa della malattia.

3.4 Algoritmi per l'allineamento

Per velocizzare l'allineamento, gli algoritmi si appoggiano a strutture dati, che appartengono principalmente a due categorie: tabelle hash e algoritmi basati su indici di suffissi.

3.4.1 Algoritmi basati su tabelle hash

Tutti gli algoritmi basati su tabelle hash seguono lo stesso paradigma. Vengono salvate le posizioni di ogni k-mer in una tabella hash, usando la k-mer come chiave, confrontando poi la stringa col genoma di riferimento, cercando un match esatto, detto *seed*.
I *seed* vengono poi uniti senza *gap*, per migliorare successivamente l'allineamento usando l'algoritmo Smith-Waterman.
Questo algoritmo è stato perfezionato in modi diversi.

Spaced-seed Un'evoluzione del metodo precedente è consentire che, al momento del confronto della *read* col genoma di riferimento, in determinate posizioni della parola siano presenti degli errori. Questo metodo consente di

aumentare significativamente la precisione dell'analisi. *Seed* in cui si permette questo sono detti *Spaced-seed*. Il numero di match che si verificano sono il peso del *seed*. Algoritmi che usano questa strategia sono il SOAP, SHRiMP, MAQ.

Filtro q-gram e seed-hit multiplo Un problema che i metodi precedenti hanno è che non consentono la presenza di *gap* all'interno del *seed*, che vengono trovati solo nei passaggi successivi dell'analisi. Il filtro q-gram permette di costruire l'hash table tenendo già conto di queste informazioni. Si basa sull'idea che una parola, lunga w, con al più k differenze, tra mismatch e *gap*, rispetto alla stringa di riferimento, deve avere almeno $(w+1)-(k+1)q$, sottostringhe comuni lunghe q. Algoritmi che utilizzano questo algoritmo sono il SSAHA2 e il BLAT.

3.4.2 Algoritmi basati su suffix/prefix trie

Gli algoritmi in questa categoria essenzialmente riducono il problema dei match inesatti a quello dei match esatti, tutto questo in due passaggi: identificazione dei match esatti e costruzione di un allineamento per i match inesatti, utilizzabile poi con i match esatti.
Per trovare match esatti, gli algoritmi si basano su suffix/prefix trie, suffix tree, suffix array, FM-index. Il vantaggio dei trie/tree è che l'allineamento contro più copie identiche di una sottostringa viene fatto una sola volta, perché tutte queste stringe si trovano nella stessa posizione dell'albero. Questo concetto può essere traslato in termini di sottointervalli consecutivi nelle strutture dati basate su array.

Trie, prefix/suffix tree e FM-index Un suffix trie è una struttura dati che memorizza tutti i suffissi di una stringa, consentendo veloci *string matching*. In Figura 6A è mostrato un esempio di prefix trie.

Trovare tutti i match esatti di una stringa è equivalente a cercare un percorso dalla radice al nodo, in cui la stringa, costruita concatenando la *label* di ogni nodo, corrisponda alla stringa di partenza rovesciata. Se un tale percorso esiste, allora la stringa è sottostringa.
La complessità per verificare se una stringa è sottostringa è lineare, indipendentemente dalla stringa di riferimento. Un trie tuttavia occupa uno spazio $O\left(L^2\right)$, con L la lunghezza della stringa di riferimento.
Utilizzare i trie non è quindi praticabile, nemmeno per piccoli genomi. Per ovviare al problema, sono state proposte diverse strutture dati, tra cui il suffix tree(Figura 6C), che è sfruttato in diversi algoritmi, dato che occupa uno spazio lineare($17n$ bytes) e consente ricerche in tempi lineari.
Ulteriori migliorie a livello di memoria utilizzata sono state raggiunte con i FM-index, con cui è stato possibile salvare l'intero genoma umano con 8GB.
Algoritmi basati su queste strutture dati sono OASIS(suffix tree), Segemehl(suffix array), BWA, SOAP2(FM-index). Il FM-index è la struttura dati più utilizzata, soprattutto per la poca memoria richiesta.

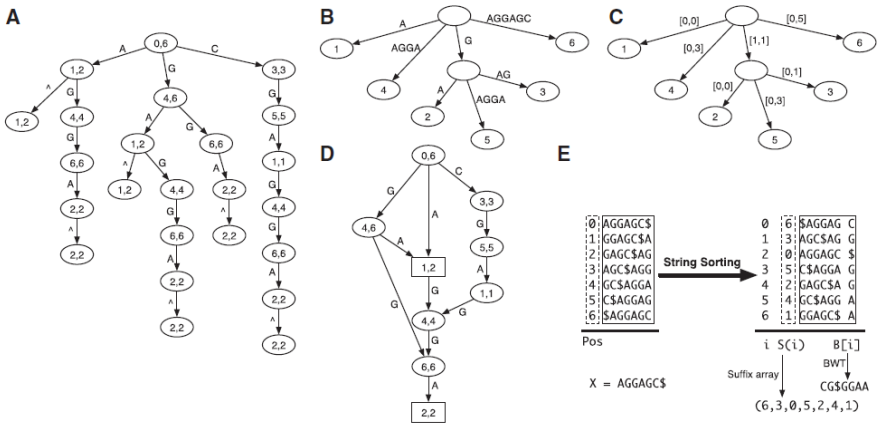

Figura 6: Strutture dati basate su prefix trie. (A) Prefix trie della stringa AGGAGC, dove il simbolo OE indica l'inizio della stringa. I numeri in ogni nodo indicano l'intervallo del suffix array della sottostringa rappresentata dal nodo, che è la concatenazione delle stringhe presenti sui rami dal nodo alla radice. (B) Prefix trie compresso, ottenuto raggruppando tutti i nodi con un solo figlio. (C) Prefix tree con, al posto delle stringhe, l'intervallo corrispondente nella stringa originale. (D) Grafo diretto dei prefissi(prefix DAWG), ottenuto facendo collassare i nodi del prefix trie sugli intervalli identici del suffix array. (E) Costruzione del suffix array e trasformata di Burrows-Wheeler della stringa. \$ indica la fine della stringa. L'intervallo del suffix array di una sottostringa W è l'intervallo massimale nel suffix array, con tutti i suffissi nell'intervallo aventi W come prefisso. Immagine tratta da [19].

I match inesatti sono trattati in modi diversi, tra i tanti si cita quello utilizzato da Bowtie e BWA, che prendono brevi sottostringhe dal genoma di riferimento e le confrontano con la stringa in analisi, limitando i mismatch a dei valori fissati. Queste sottostringhe vengono poi ricombinate, ottenendo un allineamento con *gap*.

3.5 Problematiche

La lunghezza attualmente ottenibile dalle NGS è di circa 1kbp per il Roche 454, mentre di 200bp per l'Illumina e di 100bp per l'ABI SOLiD. A confronto col metodo Sanger, che poteva arrivare a 700-900bp, l'unico sistema competitivo è il metodo Roche 454, gli altri risultano inutilizzabili per problemi in cui si abbia bisogno di sequenze più lunghe, come ad esempio il sequenziamento del genoma umano con un livello di dettaglio maggiore(*reads* più lunghe consentono confronti più accurati). Tuttavia, anche le 1000bp raggiungibili dal metodo Roche 454 sono ormai lunghezze inaccettabili per questo genere di problemi.
Avere *reads* corte dà problemi anche in altre applicazioni, come il *de novo assembly*, che si occupa partendo dalle *reads*, di ricomporre il genoma, senza utilizzarne uno di riferimento. Questo sistema risulta di difficile applicazione con le NGS per la brevità dei frammenti da unire, brevità che causa la valutazione errata dei frammenti, da cui si ha una ricostruzione errata del genoma.

Oltre alla lunghezza ridotta delle *reads*, un altro problema è sorto col progredire di questi sequenziatori. I limiti tecnologici dell'hardware sono di fatto diventati il collo di bottiglia per questi strumenti, sia come limiti di calcolo, sia come limiti di immagazzinamento dati. Alle elevate potenze di calcolo si è rimediato parallelizzando il lavoro su centinaia di macchine, soluzione dispendiosa ma efficiente, potendo studiare le *reads* in modo indipendente, ricombinandole in seguito. L'immagazzinamento dati al giorno d'oggi non è un problema così serio ma, nei prossimi anni, col volume sempre crescente di informazioni che questi algoritmi offrono, si dovrà trovare una soluzione ai costi degli strumenti e alle velocità di accesso ai dati.

4 Ambiti di applicazione

In questo capitolo si descriveranno alcuni tra i diversi ambiti applicativi in cui le NGS svolgono un ruolo fondamentale.

4.1 Epigenetica

L'epigenetica è una branca della genetica che descrive tutte quelle modifiche ereditabili che variano l'espressione genica pur non alterando la sequenza del DNA[44].

Recenti ricerche nel settore hanno gettato le fondamenta del Human Epigenome Project(HEP), che si propone di catalogare le metilazioni del DNA(il termine metilazione è usato in chimica per definire l'addizione o la sostituzione di un gruppo metile su vari substrati) su larga scala genomica.
Le tecnologie NGS, con il loro throughput elevato, possono abbattere significativamente i tempi per queste ricerche.
Attualmente le tecnologie NGS sono state applicate a diverse aree dell'epigenetica.

4.1.1 Classificazione di pattern di metilazione

Catalogare pattern di metilazioni del DNA su larga scala è una delle applicazioni più comuni dell'epigenetica, ed è l'obiettivo principale dell'HEP[26].
Esistono tre approcci per il rilevamento delle metilazioni su larga scala: la riduzione dello smaltimento degli enzimi endonucleasi attraverso tecnologia a microarray(un metodo di sequenziamento di terza generazione), sequenziamento mediante bisolfito e immunoprecipitazione. Con il sequenziatore Roche 454 è stato possibile migliorare il sequenziamento tramite bisolfito[22] arrivando, in alcuni esperimenti, ad ottenere 1600 sequenze, contro le 20 generate con i metodi precedenti. Anche per l'Illumina è stata proposta una soluzione simile.

4.1.2 Localizzazione siti di legame

Modifiche degli istoni, ossia proteine più abbondanti associate ai cromosomi, nel DNA si pensa controllino l'espressione genetica, regolando la forza dell'interazione tra DNA e istone[10].
Storicamente il metodo di riferimento per questi studi era il ChIP(chromatin immunoprecipitation), che lega proteine al DNA.
Il metodo è stato migliorato utilizzando prima il sistema Sanger, poi il Roche/454(STAGE).
Le tecniche NGS hanno permesso lo sviluppo di una nuova tecnologia, ChIP-Seq, per il rilevamento di modifiche di istoni su larga scala genomica, grazie alla quale sono stati effettuati studi più approfonditi sull'interazione tra proteine e DNA.
Su grandi estensioni di DNA, ChIP-Seq ha il potenziale per ottenere la massima risoluzione, poiché la sua precisione dipende solamente dalla dimensione dei

frammenti che riceve in ingresso.
Attualmente la tecnologia di riferimento per questo tipo di analisi è l'Illumina. Con le tecniche NGS, ChIP-Seq è in grado di fare analisi prima irrealizzabili, sia grazie all'elevato throughput, sia con l'amplificazione PCR, che permette di estrarre dati senza rovinarli, come invece accadeva con tecnologie precedenti.

4.1.3 Mappatura di nucleosomi

Le tecnologie di nuova generazione sono state utilizzate per la mappatura delle posizioni dei nucleotidi, nonché di altre informazioni rilevanti. I nucleosomi sono tra questi fattori rilevanti, modificando la regolazione dei geni, e vengono generalmente associati ad una minor capacità del DNA di sfruttare le proteine. Due recenti studi hanno utilizzato la digestione MNase(un metodo per legare acidi nucleici), seguita da un sequenziatore Roche/454 per mappare le posizioni di alcuni nucleosomi su larga scala genetica[40].
I dati forniti dal metodo ChIP-Seq possono essere utilizzati per risolvere anche questo problema, permettendo il rilevamento delle posizioni dei nucleosomi su larga scala genomica. Studi in questo senso sono stati compiuti con successo[6]. L'utilizzo dei dati ChIP-Seq porta tre vantaggi: possono essere mappati solo nucleotidi associati ad un certo istone, permettendo quindi di affinare le ricerche; il posizionamento dei nucleosomi è regolato da modifiche agli istoni, che possono ad esempio segnalare per la rimozione un nucleosoma; il metodo non è quantitativo, poiché la quantità di *reads* mappate in una certa regione è correlata con la presenza di modifiche ad istoni, che non sono per forza correlate alla quantità di nucleosomi.

4.2 Metagenomica

La metagenomica è una branca della genomica che si occupa di analizzare insiemi di DNA differenti, con l'obiettivo di stabilire quali e quanti organismi siano presenti. Le tecnologie di sequenziamento di nuova generazione, con la possibilità di sequenziare senza dover duplicare i DNA in esame, cosa che, in questo contesto, spesso non è possibile, e col loro elevato *throughput*, consentono analisi irrealizzabili fino a pochi anni fa.
La metagenomica trova applicazioni diverse nello studio dei batteri a contatto col corpo umano, in effetti in questo senso sono già stati fatti diversi studi, che cercato tutti di individuare la relazione esatta tra certe popolazioni di batteri e il nostro organismo.
Oltre a studi sul corpo umano e sui batteri, la metagenomica è applicata in settori molto diversi, come gli studi sul suolo[38] o sull'oceano[1].

4.3 Ricerca sul genoma umano

La ricerca sul genoma umano è stata sviluppata in special modo da due gruppi, lo Human Genome Project(HGP) e dalla Celera Genomics che hanno fatto la storia, sequenziando per la prima volta l'intero genoma umano. Entrambi hanno

seguito differenti strade, l'HGP ha sfruttato un algoritmo basato su mappe, mentre il gruppo Celera ha preferito usare il metodo shotgun (Figura 7). Questi

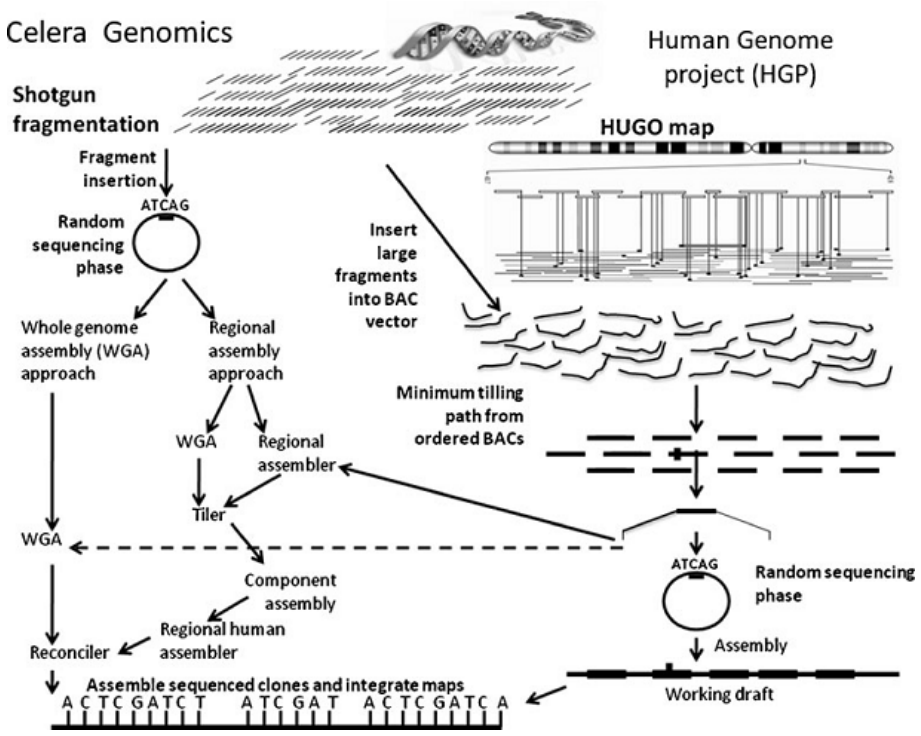

Figura 7: Metodi per generare la prima sequenza completa del genoma umano: sulla destra il metodo a mappe del HGP, sulla sinistra il metodo shotgun usato dal Celera Group. Immagine tratta da [8].

metodi dimostrarono presto i loro limiti, con costi computazionali eccessivi. In questo contesto le Next Generation Technologies trovarono presto applicazione.

4.3.1 Sequenziamento de novo e ri-sequenziamento

Il sequenziamento *de novo* è un metodo per creare una trascrizione del DNA senza utilizzare genomi di riferimento. Le tecnologie di nuova generazione hanno reso questo strumento, applicato a svariati organismi, tra cui gli esseri umani, un'operazione più veloce. Sono stati infatti sequenziati sottoinsiemi(dal 7 al 30%) di genomi di diversi individui, usando i metodi già visti appartenenti alle NGS, Roche 454, Illumina, ABI SOLiD.
Risultati simili non sono stati raggiunti solo con le tecnologie di seconda generazione, per esempio si era arrivati a sequenziare il 90% del genoma di una persona utilizzando l'Helicos[12]. Il vantaggio delle NGS tuttavia resta, col loro meccanismo consentono operazioni di selezione e controllo impensabili prima, si

veda ad esempio lo studio del DNA del feto, utilizzando tecniche HT-NGS sul sangue della madre[11].

4.3.2 Sequenziamento di RNA

Le tecnologie di nuova generazione hanno trovato applicazione anche nello studio di porzioni di RNA. Tra le tante, si cita lo studio effettuato sui miRNA(piccole molecole endogene di RNA non codificante, a singolo filamento, di 20-22 nucleotidi), grazie al quale sono stati effettuati progressi nello studio sul cancro[18].
Sono stati proposti diversi strumenti che sfruttano il sequenziatore Illumina, risolvendo anche i problemi tecnici dovuti alla mancanza di database esaurienti sui miRNA.
Oltre all'Illumina, anche il sequenziatore 454/Roche ha trovato applicazione nella ricerca contro il cancro, capire la trascrizione(processo mediante il quale le informazioni contenute nel DNA vengono trascritte enzimaticamente in una molecola complementare di RNA) è essenziale per trovare e descrivere quegli elementi che causano l'insorgere delle malattie.
Utilizzando le tecnologie NGS è quindi sempre più semplice ottenere trascrizioni, rendendo più efficacie la ricerca di tratti comuni tra le varie sequenze.
A questo va aggiunto che le *reads* corte prodotte con NGS, specialmente da SOLiD e Illumina, si adattano perfettamente al profiling dell'espressione genetica.

4.3.3 Variazioni genetiche e rilevamento delle mutazioni

Una delle aspettative sulle NGS è di facilitare lo studio delle variazioni genetiche sull'intero genoma umano, scoprendo tutte le variazioni, comuni e rare, nella popolazione umana. In questo settore, il progetto *1000 Genomes Project* sta dando uno dei maggiori contributi, con un lavoro decennale su cui ancor oggi vengono eseguite analisi. Scopo originale del progetto era quello di scoprire la funzione dei geni appartenenti al genoma umano, obiettivo non raggiunto nella sua interezza a causa di limiti tecnologici. Con lo sviluppo delle NGS, che consentono un sequenziamento più preciso, si spera di riuscire presto a costruire una mappa completa del genoma umano e quindi poterlo analizzare.
Altra applicazione delle NGS è il rilevamento di mutazioni del DNA a causa di malattie o farmaci[35].

4.3.4 Ricerca contro il cancro e biomarcatori

Gli sviluppi delle NGS hanno portato allo sviluppo di nuovi approcci terapeutici e diagnostici per il trattamento del cancro. Individuare le mutazioni del DNA che il cancro ha indotto è un'operazione sempre più rapida.
Altra applicazione di interesse sono i biomarcatori personalizzati.
Un biomarcatore è un frammento della sequenza di DNA causa di malattia o di una certa predisposizione patologica. I biomarcatori vengono modificati in modo da rilevare la presenza di riarrangiamenti specifici di un certo tumore nei campioni di sangue dei pazienti, così da generare poi un biomarcatore che contrasti la malattia.

4.4 Ricerca sul genoma animale

Sono disponibili genomi completi di diversi animali, tra cui cani, gatti, polli, cavalli. La prima iniziativa di sequenziamento del genoma animale è iniziata nel 2001, poco dopo il sequenziamento del genoma umano.
Inizialmente vennero organizzati dei workshop dall'Animal Genome Research(AGR) e dal National Academy of Sciences(NAS), con l'obiettivo di individuare gli scopi di ricerche nel settore, raccogliere fondi per la costruzione di database di genomi delle principali specie domestiche.
Con dei database sempre più forniti, c'è ora la possibilità di eseguire analisi sulla selezione genetica(genomics breeding for selection, GBS) dei capi di bestiame, così da migliorarne la qualità.
In questo momento la ricerca sugli animali non è un settore in grande sviluppo, non trovando i fondi necessari. È possibile che, con i miglioramenti della ricerca sull'uomo, anche questo contesto possa accrescersi.

5 Esempio pratico, SHRiMP

5.1 Descrizione

L'algoritmo SHRiMP(the SHort Read Mapping Package), proposto nel 2009[41] per il *mapping* e l'analisi di *reads* corte(25-70 basi), risulta essere molto efficiente in caso di geni polimorfici(si parla di polimorfismo genetico quando una variazione genetica ha una prevalenza maggiore dell'1%). È stato ottimizzato per l'utilizzo dell'output fornito dalla piattaforma SOLiD.
Il metodo sfrutta in un primo momento un veloce *k-mer hashing*(un hash basato su chiavi lunghe k) e, successivamente, un'implementazione dell'algoritmo Smith-Waterman[42]. In questo modo viene eseguito un allineamento completo di tutte le *reads* prodotte su ogni area del genoma di riferimento che potrebbe essere ad esse omologa. A questo passaggio segue un altro algoritmo, specializzato nel *mapping* di *color-space reads*, che consente di allineare, in modo non euristico, il genoma campione con quello di riferimento, raffinando ulteriormente i risultati del passaggio precedente, aumentando la qualità dell'output fornita dal sistema SOLiD. Per concludere, viene utilizzato un altro algoritmo per la verifica della qualità degli allineamenti prodotti.
Poiché una *reads* potrebbe avere un riscontro in più posizioni del genoma, è stato introdotto un metodo statistico per selezionare l'allineamento più probabile, riducendo il numero di falsi positivi. In Figura 8 è presente uno schema dei passi fondamentali dell'algoritmo.

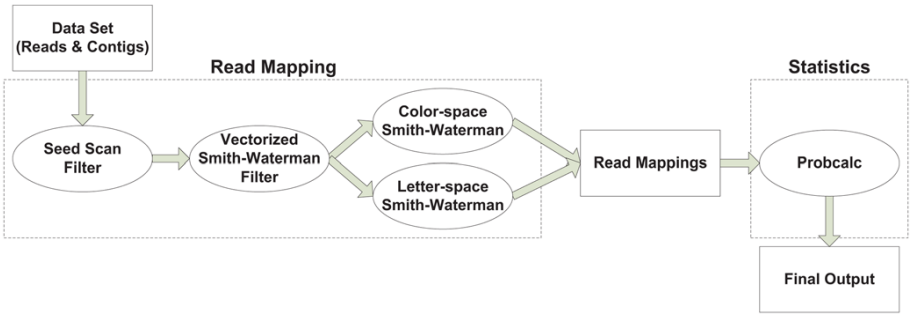

Figura 8: Workflow dell'algoritmo SHRiMP: vengono filtrati e mappati i dati, dopodiché, con la suite PROBCALC, si selezionano i risultati migliori. Immagine tratta da [41]

5.2 Algoritmo

5.2.1 Algoritmo di Read Mapping

L'algoritmo utilizza tre strumenti precedentemente sviluppati: l'approccio *q-gram filter*[23], gli *spaced seeds*[2] e i gli *specialized vector computing hardware*[27]. Questi metodi vengono sfruttati per filtrare in modo rapido tutte le va-

rie *reads*. Successivamente, si utilizza l'algoritmo basato sullo Smith-Waterman per valutare con più precisione i risultati.

Spaced seeds Il metodo si basa sull'idea che alcune posizioni della *read* possano non ottenere corrispondenza, ammettendo quindi errori in posizioni prestabilite. Il metodo è un evoluzione rispetto ai classici sistemi, in cui si richiedeva una corrispondenza perfetta, in quanto consente di tener conto di errori del genoma.

Q-gram filters Questo strumento utilizza più seed per determinare se si è verificata un'occorrenza. Nel metodo SHRiMP si richiede infatti che ci sia un numero predeterminato di seed che mappano in un'area del genoma.

Vectorized Smith-Waterman Se una *read* supera i passaggi precedenti, viene eseguito un rapido controllo tra due regioni, per verificare le somiglianze. Questo viene fatto con una versione modificata dell'algoritmo Smith-Waterman, sfruttando così le potenzialità delle CPU moderne. Per ogni *read* vengono salvate la posizione e la valutazione degli allineamenti migliori, il cui numero può essere impostato.

Allineamento finale Finiti i passaggi precedenti, si esegue un allineamento di tutte le *reads* rimaste, con tutte le locazioni potenziali(le migliori ottenute dal passaggio precedente). Il metodo con cui questo passaggio viene compiuto dipende dalle tecnologie con cui si interfaccia l'applicazione, esistono moduli per dati forniti dall'Illumina Solexa e dall'AB SOLiD.

5.2.2 Algoritmo per l'allineamento color-space

Col sequenziatore AB SOLiD è stata introdotta anche una nuova tecnica di sequenziamento, che legge coppie di basi sovrapposte e le assegna un colore(quattro possibili, in genere vengono numerati da 0 a 3). Ogni base viene analizzata due volte, una volta col predecessore, una col successore. Due schemi del metodo sono mostrati in Figura 9.

Uno dei vantaggi del sequenziatore AB SOLiD è la capacità di distinguere tra errori di sequenziamento ed errori dovuti a SNPs biologico, cioè ad una variazione genetica. Un errore SNPs scambia due *reads* adiacenti nel codice color-space, mentre un errore di sequenziamento non è probabile capiti a due *reads* consecutive.
Il metodo ha tuttavia delle limitazioni: un singolo errore di sequenziamento comporta un errore per tutte le *reads* successive, traducendole tutte in un colore diverso.

Per questi motivi, gli sviluppatori di SHRiMP hanno proposto una soluzione differente. Invece di tenere traccia di una singola codifica, si memorizzano tutte e quattro le possibilità. A questo punto, l'algoritmo, secondo un'euristica, può decidere, se stima che si sia verificato un errore, di cambiare la codifica corrente

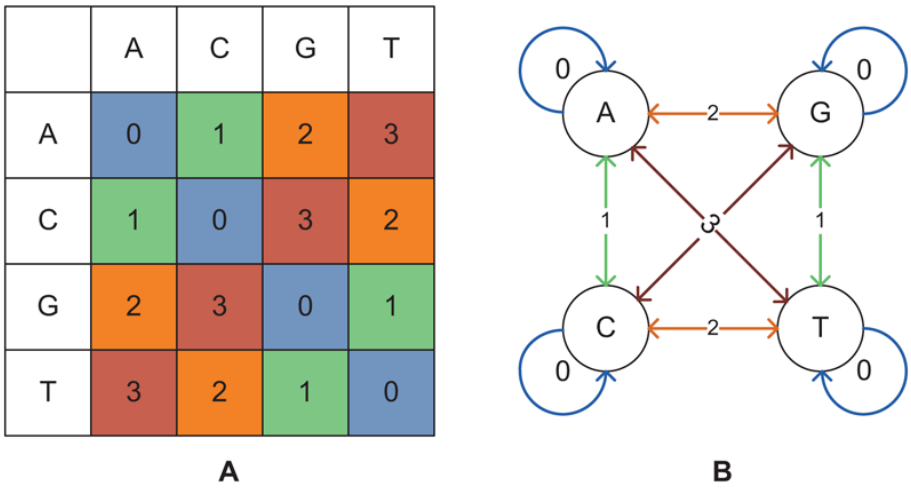

Figura 9: Rappresentazioni del metodo di codifica *color-space* utilizzato dall'AB SOLiD. (A) Rappresentazione standard delle relazioni tra le basi. (B) Rappresentazione con un automa stati finiti. Immagine tratta da [41]

con una delle tre rimanenti, correggendo l'errore.
Questo algoritmo, nell'applicazione, viene utilizzato come ultimo passaggio, preferendo la versione AB SOLiD per i primi passaggi di selezione visti al capitolo precedente.

5.2.3 Calcolo statistico per le reads e le coppie

Una volta mappate le *reads*, viene applicato un metodo statistico su ogni *reads* e su ogni coppia, in modo da stimare la probabilità che si sia verificato un errore.

Statistica per le reads Data la brevità delle *reads* è stato necessario sviluppare una teoria adeguata. Ne è stata quindi sviluppata una in grado di modellare *reads* brevi, tenendo conto anche di indels. Vengono analizzati i dati seguendo questa statistica, calcolata da PROBCALC:

- *pchance*: probabilità che si verifichi un hit per caso
- *pgenome*: probabilità che l'hit sia stato generato dal genoma

Un buon sequenziatore dobrebbe avere *pchance* bassa(circa 0) e *pgenome* alta(prossima all'1). *pchance* è la probabilità che una *read* si allinei con un buon punteggio ad un genoma della stessa lunghezza, ma con composizione nucleica casuale(ossia, che la *read* si allinei per fortuna). Per tenere conto di tutto questo, si contano tutte le *k-mers* con la stessa *pchance*, definendo questo numero Z.

Per calcolare *pchange* sull'intero genoma, si assume che le posizioni siano indipendenti tra loro, e si valuta la probabilità che non ci sia un match in nessuna di queste.

$$p_{chance} = 1 - \left(1 - cf\,(r) * \frac{Z}{4}\right)^{2g} \tag{1}$$

dove r è la lunghezza dell'allineamento, g la lunghezza del genoma, $cf\,(r)$ il fattore di correzione.

pgenome indica la probabilità che una hit si sia verificata nel genoma, e che quindi sia legittima(indels, errori genetici). Si stima per prima cosa il tasso di ogni evento, utilizzando il metodo bootstrap(un metodo statistico per l'estrapolazione dei dati), usando poi questi dati per costruire uno stimatore per ogni evento. Ad esempio, stimato il tasso di errore medio Γ_ϵ, si può stimare la probabilità che una *read* sia stata generata da questi errori:

$$p_\epsilon \approx \left(\frac{r}{n_\epsilon}\right) \Gamma_\epsilon^{n_\epsilon} \left(1 - \Gamma_\epsilon\right)^{r - n_\epsilon} \tag{2}$$

dove n_ϵ è il numero di errori osservati nella hit attuale, e r è la lunghezza dell'allineamento. Allo stesso modo si possono stimare p_{indel} e p_{subs} per indels e sostituzioni. A questo punto si può definire *pgenome* come:

$$p_{genome} = p_\epsilon p_{subs} p_{indel} \tag{3}$$

Per concludere, si definisce la misura di qualità dell'hit come la differenza normalizzata

$$normodds_{hit} = \frac{\frac{pgenome_{hit}}{pchance_{hit}}}{\sum \forall hits \frac{pgenome}{pchance}} \tag{4}$$

Questo valore indica la credibilità relativa di un'hit, rispetto ad altre. Una singola hit avrà normodds=1, due hit equiprobabili avranno normodds=0.5. Con invece una hit molto vicina al risultato e l'altra più lontana, la prima si avvicinerà ad 1, l'altra sarà più prossima allo 0.

Statistica per le coppie SHRiMP fornisce anche una stima dell'attendibilità delle coppie, combinando il valore ottenuto al punto precedente con distribuzioni empiriche contenute nella propria libreria. Più nello specifico, si calcola la distanza d tra tutte le coppie, la cui media viene detta μ, quindi si assegna ad ogni coppia un *pchanche* e *pgenome*.
Data $p_c\,(g)$ la probabilità *pchance* di una *read*, con g la lunghezza del genoma, si definisce la *pchance* della coppia di *reads*, read_1, read_2, come:

$$p_c = p_{c,read_1}(g) p_{c,read_2}(|\mu - d + 1|) \tag{5}$$

dove g è la lunghezza del genoma usata in PROBCALC, μ è ma distanza media tra le coppie e d è la distanza tra le due *reads*. In altre parole, questa formula indica la probabilità che una *read* con lo stesso livello della prima si allinei ovunque nel genoma, per caso, e che un'altra *read* si allinei al genoma alla

Tabella 1: Tempo di esecuzione per l'algoritmo SHRiMP per mappare 500'000 *reads* lunghe 35bp su un genoma di riferimento di 180Mb, utilizzando un computer single Core2 2.66GHz. Dati tratti da [41]

k-mer	(7,8)	(8,9)	(9,10)	(10,11)	(11,12)	(12,13)
% scansioni k-mer	10.1%	16.5%	18.9%	13.4%	9.8%	7.4%
% Filtro SW	88.8%	75.4%	49.8%	30.2%	20.1%	14.9%
% Allineamento SW completo	1.1%	8.0%	30.7%	55.5%	68.8%	76.2%
% Tempo	1d 21h 34m	6h 18m	1h 36m	50m 28s	37m 52s	32m 32s

stessa distanza d.
Assumento ora che p_g sia la *pgenome* di una *read*, si può allora calcolare la *pgenome* di ogni coppia con

$$p_g = p_{g,1} p_{g,2} T \tag{6}$$

con T la *tail probability*(probabilità che una variabile aleatoria devii di una certa quantità dal valore atteso) della distanza della coppia, secondo la distribuzione calcolata all'inizio in modo empirico.

5.3 Analisi sperimentale

Si riportano qui i risultati dei test effettuati da Rumble *et al*[41] a titolo d'esempio delle potenzialità dell'algoritmo.

Analisi dei tempi di esecuzione Nel primo esperimento si è effettuata l'analisi di 135 milioni di *reads* lunghe 35 bp estratte da un *Ciona savignyi* appartenente alla famiglia dei tunicati(un animale marino), rapportandole al genoma di riferimento[24].
L'analisi è stata fatta studiando 500'000 *reads* prese casualmente, mappandole quindi al genoma di riferimento. In Tabella 1 sono riportati i tempi d'esecuzione complessivi per l'esecuzione dell'algoritmo su un pc domestico(single Core2, 2.66GHz), indicando i tempi che le varie fasi hanno richiesto.
Si noti come, aumentando la dimensione delle k-mer, si riducano significativamente i tempi. Questa riduzione è dovuta principalmente alla miglior selettività del filtro che, per k-mer più lunghe, genera meno hit, riducendo quindi i costi per le operazioni successive.

Analisi della precisione È stato effettuato anche un altro esperimento, questa volta per verificare se l'algoritmo era in grado di fare analisi precise anche di *reads* con pattern complessi di mutazioni, inserimenti, cancellazioni. Per fare questa verifica, si è analizzato un *Ciona savignyi* risequenziato, scelto per via dell'alto tasso di polimorfismo che ha.

Tabella 2: Risultati del mapping di 135 milioni di bp di *Ciona savignyi* utilizzando SHRiMP e il mapper dell'AB SOLiD. Dati tratti da [41].

	SHRiMP	SOLiD mapper
Uniquely-Mapped reads	51,856,904 (38.5%)	15,268,771 (11.3%)
Non-Uniquely-Mapped reads	64,252,692 (47.7%)	12,602,387 (9.4%)
Reads non mappate	18,657,736 (13.8%)	106,896,174 (79.3%)
Copertura media	10.3	3.0
Copertura mediana	8	1
SNPs	2,119,720	383,099
Eliminazioni (1-5bp)	51,592	0
Inserzioni (1-5bp)	19,970	0

Utilizzando il sistema AB SOLiD sono state generate 135 milioni di *reads*, lunghe ciascuna 35bp. Queste sono poi state confrontate al genoma di riferimento[24] con SHRiMP. Il sequenziamento ha occupato un totale di 48 ore, utilizzando 250 core da 2.33GHz.
Per chiarezza, è stato effettuato lo stesso test utilizzando anche il mapper dell'AB SOLiD, impostato per mappare *reads* con al più 3 *mismatch*.

In tabella 2 è possibile osservare i risultati di questo esperimento. SHRiMP ha mappato un numero di *reads* 4.2 volte maggiore superiore a quello del mapper AB SOLiD, e ha trovato 5.5 volte in più di SNP. Il mapper AB SOLiD tuttavia è stato molto più veloce, richiedendo 255 ore CPU per effettuare l'analisi, un tempo 50 volte inferiore. La differenza tra i due metodi è tuttavia evidente nell'analisi degli indels, dove il mapper AB SOLiD è inutilizzabile, mentre SHRiMP funziona egregiamente.

Analisi di dati simulati Per verificare se effettivamente SHRiMP funzionasse bene con situazioni di forte polimorfismo(SNPs e indels) è stato organizzato un altro esperimento. Presi dei campioni casuali dal *Ciona savignyi*, sono stati introdotti diversi errori, sia indels che SNPs, in percentuali dal 2% al 7%, e quindi si sono mappati questi campioni modificati contro il genoma originale. Nell'analisi si sono considerate solo so le *reads* con un'unica top hit. Sono state poi valutate la precisione, cioè il rapporto tra la frazione di *reads* che hanno trovato un riscontro univoco, e il totale delle *reads*, e il *recall*, la frazione di *reads* con un hit univoco corretto. In tabella 3 si possono vedere i risultati di questo esperimento.

Conclusioni Secondo i dati sperimentali ottenuti, perché l'algoritmo sia utilizzabile in modo efficiente, necessita di k-mer lunghe almeno 10, in modo da ridurre nei primi passaggi la quantità di dati in analisi. Lo strumento si dimostra inoltre più performante del mapper SOLiD, anche se l'aspetto più interessante è la capacità di rilevare le *indels*. Nella rilevazione di *indels* e SNPs

Tabella 3: Precisione del mapping color-space. Ogni cella indica la precisione(prec.) e il recall(rec.) nel mapping delle *reads*, con diversi gradi di polimorfismo. Per una migliore analisi, i risultati sono stati divisi per in base al numero di SNPs e al numero di indels. Dati tratti da [41].

Numero di SNPs	0		1		2		3		4	
Massimo numero di indels	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
0	85.7	83.2	84.8	81.3	83.5	76.6	80.6	65.2	75.6	46.8
1	83.8	79.4	82.2	74.0	79.4	62.6	72.8	43.2	63.1	24.7
2	83.2	77.1	80.8	69.6	77.9	56.6	68.2	36.4	56.4	18.9
3	80.7	71.0	79.6	64.2	73.6	48.3	66.5	31.5	57.1	16.6
4	78.0	65.4	76.5	56.1	71.4	41.9	60.6	23.9	50.3	12.4
5	75.9	58.9	73.0	48.1	69.7	36.6	57.0	21.3	46.0	12.7

si trovano invece precisioni buone, ma non elevate: il valore massimo arriva a circa l'85%. Tuttavia ulteriori sviluppi del metodo potrebbero sopperire a questi limiti, rendendo SHRiMP uno strumento ancor più competitivo.

6 Future Evoluzioni

I continui studi sulle tecnologie, la chimica e i materiali, hanno ormai aperto la strada anche alle tecnologie di terza generazione, che stanno venendo sviluppate e migliorate in questi anni. Si descrivono di seguito alcune tra le più promettenti.

Single-molecule real-time sequencing(SMRT) È il primo sequenziatore di terza generazione ad analizzare direttamente le singole molecole. Poiché le molecole sono della dimensione dell'ordine dei 10nm di diametro, si ha necessità di riuscire a separare in modo efficiente le varie molecole, in modo che non causino errori nell'analisi.
Grazie a questo metodo di funzionamento, il sequenziatore SRMT può produrre *reads* lunghe mediamente 1'000bp, arrivando a punte di 10'000bp.
Altro vantaggio rispetto alle NGS sono i costi e i tempi di preparazione, che vengono abbattuti, non necessitando più di reagenti per le reazioni PCR.
Le potenzialità del SMRT sono ancora maggiori, consentendo l'analisi di informazioni cinetiche, come l'osservazione real-time di come il ribosoma trasforma l'mRNA.
Purtroppo restano dei problemi da risolvere, come nella tecnologia Helicos. Gli errori nella lettura dei dati, causati da indels, sono sopra il 5%. Altro problema è il throughput, che ancora non è ai livelli delle tecnologie di nuova generazione.

Fluorescence resonance energy transfer È un nuovo metodo in via di sviluppo, prevede di utilizzare una tecnologia simile all'Helicos, legando cioè un nucleotide ad un reagente che, una volta uniti, emette un segnale di risonanza(fluorescence resonance energy transfer, FRET). Questo metodo ha le potenzialità per milioni di basi al secondo.

Tunnelling e approccio basato su transmission-electron-microscopy Si tratta di una tecnica che utilizza la microscopia elettronica che, sfruttando il transmission electron microscopy(TEM) rileva gli atomi che identificano univocamente i diversi nucleotidi che compongono la struttura del DNA. Al momento, non è ancora stato realizzato un sistema che usi questo strumento ma, si dovesse riuscire a realizzarlo, si potrebbero avere *reads* molto lunghe, ad un prezzo estremamente contenuto.

Scanning tunneling microscope tips Reveo sta sviluppando questa tecnologia, in cui il DNA viene posizionato su una superficie conduttrice per rilevare le basi elettronicamente, utilizzando uno scanner(STM). La procedura per depositare la sequenza di DNA sulla superficie non è ancora stata trovata ma, come il precedente, si dovesse trovare un modo per applicare questa teoria, si potrebbero ottenere velocità e lunghezza delle *reads* notevoli e costi significativamente abbattuti.

Nanopore La maggior parte delle tecnologie di sequenziamento basate su nanopore studiano il transito di una molecola di DNA attraverso un buco, studiando la variazione del campo elettrico si possono individuare le basi. Poiché questa tecnologia utilizza DNA non modificato, c'è la possibilità che riescano a funzionare ad estrema velocità. Sono stati proposti diversi metodi basati sulle nanopore, tra questi si citano quello prodotto dalla Oxford Nanopore(sfrutta la combinazione di tre molecole), quello basato sulla MspA(una proteina), le differenze tra questi sono solamente il tipo di nanopora utilizzato per studiare il DNA.
Una possibile evoluzione di questo sistema è lo studio in parallelo di più nanopore contemporaneamente. Il primo metodo che mette in pratica questa tecnologia è stato sviluppato[4].

Tecnologia basata su transistor L'IBM sta sviluppando un apparecchio in grado di rilevare elettronicamente le singole basi in una singola molecola di DNA. Di nuovo, velocità, lunghezza delle *read*, basso costo, sono tutte qualità che potrebbero essere migliorate con questo sistema, con un limite teorico che si stima essere sulle 500.000.000 di *reads* per ogni transistor, per ogni secondo. Attualmente resta solo un problema da risolvere per poter utilizzare questa tecnologia, si deve dimostrare che il segnale di una singola base possa essere distinto dal segnale delle basi vicine.

7 Conclusioni

Il rapido sviluppo delle tecnologie di sequenziamento di nuova generazione evidenzia come uno dei futuri sviluppi necessari alla crescita del settore dovrà essere il miglioramento dei sistemi che utilizzano i dati forniti dalle NGS.
La disponibilità di algoritmi che consentano analisi efficienti di questi dati, infatti, potrebbero trasformare la medicina nel prossimo futuro, dando nuovi modi di diagnosticare malattie e produrre farmaci.
Si potrebbero inoltre sviluppare nuovi metodi di analisi e sequenziamento di DNA, RNA, assemblamento De Novo.
Inevitabilmente ci saranno ricadute anche in ambiti commerciali, come il miglioramento delle specie negli allevamenti e la medicina personalizzata, ossia la possibilità di personalizzare le cure per il paziente in base ad analisi genomiche.
Al giorno d'oggi le NGS con *reads* corte(25-50 basi) e medio-lunghe(500 basi) hanno trovato diverse applicazioni. Per il futuro la speranza è che si creino metodi in grado si sequenziare *reads* di lunghezze superiori alle 1000bp.
Con lo svilupparsi della terza generazione di sequenziatori anche questi nuovi obiettivi non sembrano più così lontani.

Riferimenti bibliografici

[1] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta e G. J. Herndl. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. U. S. A.*, 103:12115–12120, 2006.

[2] A. Califano, I. Rigoutsos. Flash: a fast look-up algorithm for string homology. *IEEE Computer Society Conference*, pages 353–359, 1993.

[3] A. M. Maxam, W. Gilbert. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA*, 74:560–564, 1977.

[4] B. McNally, A. Singer, Z. Yu, Y. Sun, Z. Weng e A. Meller. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Lett.*, 10:2237–2244, 2010.

[5] B. Wold, R. M. Myers. Sequence census methods for functional genomics. *Nature*, Methods 5:19–21, 2008.

[6] C. D. Schmid, P. Bucher. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell*, 131:831–832, 2007.

[7] C. Liu, T. Wong, E. Wu, R. Luo, S. Yiu, Y. Li, B. Wang, C. Yu, X. Chu, K. Zhao, R. Li e T. Lam. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, 28:878–879, 2012.

[8] C. S. Pareek, R. Smoczynski e A. Tretyn. Sequencing technologies, genome sequencing. *Journal of Applied Genetics*, 52:413–435, 2011.

[9] D. C. Schwartz, M. S. Waterman. New generations: Sequencing machines and their computational challenges. *Journal of Computer Science and Technology*, 25:3–9, 2010.

[10] D. E. Schones, K Zhao. Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9:171–191, 2008.

[11] D. J. Burgess. Human disease: Next-generation sequencing of the next generation. *Nature Rev Genet*, 12:78–79, 2011.

[12] D. Pushkarev, N. F. Neff e S. R. Quake. Single-molecule sequencing of an individual human genome. *Nature Biotechnology*, 27:847–850, 2009.

[13] E. E. Schadt, S. Turner e A. Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19:R227–R240, 2011.

[14] E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402, 2008.

[15] F. Sanger, A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94:441–448, 1975.

[16] F. Sanger, S. Nicklen e A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA*, 74:5463–5467, 1977.

[17] G. P. Patrinos, W. J. Ansorge. *Molecular Diagnostics(Second Edition) - Novel Next-Generation DNA Sequencing Techniques for Ultra High-Throughput Applications in Bio-Medicine.* Elsevier, 2010.

[18] G. Ramsingh, D. C. Koboldt, M. Trissal, K. B. Chiappinelli, T. Wylie, S. Koul, L. W. Chang, R. Nagarajan, T. A. Fehniger, P. Goodfellow, V. Magrini, R. K. Wilson, L. Ding, T. J. Ley, E. R. Mardis e D. C. Link. Complete characterization of the microRNAome in a patient with acute myeloid leukemia. *Blood*, 116:5316–5326, 2010.

[19] H. Li, N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11:473–483, 2010.

[20] J. F. Thompson, K. E. Steinmann. Single molecule sequencing with a HeliScope genetic analysis system. *Curr Protoc Mol Biol*, 2010 Ottobre, 2007.

[21] J. Shendure, H. Ji. Next-generation DNA sequencing. *Nature biotechnology*, 26, 2008.

[22] K. H. Taylor, R. S. Kramer, J. W. Davis, J. Guo, D. J. Duff, D. Xu, C. W. Caldwell e H. Shi. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, 67:8511–8518, 2007.

[23] K. Rasmussen, J. Stoye e F. W. Myers. Efficient q-gram filters for finding all ematches over a given length. *Journal of Computational Biology*, 13:296–308, 2006.

[24] K. S. Small, M. Brudno, M. M. Hill e A. Sidow. A haplome alignment and reference sequence of the highly polymorphic ciona savignyi genome. *Genome Biology*, 8, 2007.

[25] K. V. Voelkerding, S. A. Dames e J. D. Durtschi. Next-generation sequencing: From basic research to diagnostics. *Clinical Chemistry*, 55:641–658, 2009.

[26] M. Esteller. The necessity of a human epigenome project. *Carcinogenesis*, 27:1121–1125, 2006.

[27] M. Farrar. Striped Smith-Waterman speeds database searches six times over other simd implementations. *Bioinformatics*, 23:156–161, 2007.

[28] M. M. Haque, T. S. Ghosh, D. Komanduri e S. S. Mande. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics*, 25:1722–1730, 2009.

[29] M. P. Dolled-Filhart, M. Lee Jr., C. Ou-yang, R. R. Haraksingh e J. C. Lin. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *The Scientific World Journal*, 2013:10, 2013.

[30] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén e P. Nyren. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem*, 242:84–89, 1996.

[31] M. Ruffalo, T. LaFramboise e M. Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27:2790–2796, 2011.

[32] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genetics*, 24, 2008.

[33] O. Morozova, M. A. Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92:255—-264, 2008.

[34] P. Nyren, A. Lundin. Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Anal. Biochem*, 151:504–509, 1985.

[35] P. Senapathy, A. Bhasi, J. Mattox, P. S. Dhandapany e S. Sadayappan. Targeted genome-wide enrichment of functional regions. *PLoS One*, 5, 2010.

[36] R. Li, C. Yu, Y. Li, T. Lam, S. Yiu, K. Kristiansen e J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25:1966–1967, 2009.

[37] R. Li, Y. Li, K. Kristiansen e J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24:713–714, 2008.

[38] S. Leininger, T. Urich, M. Schloter, L. Schwark, J. Qi, G. W. Nicol, J. I. Prosser, S. C. Schuster e C. Schleper. Archaea predominate among ammoniaoxidizing prokaryotes in soils. *Nature*, 442:806–809, 2006.

[39] S. M. Huse, J. A. Huber, H. G. Morrison, M. L. Sogin e D. M. Welch. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, 8, 2007.

[40] S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan e A. Z. Fire. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.*, 16:1505–1516, 2006.

[41] S. M. Rumble, P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow e M. Brudno. SHRiMP: Accurate mapping of short color-space reads. *PLoS Computational Biology*, 5, 2009.

[42] T. F. Smith, M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.

[43] T. J. Treangen, S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13:36–46, 2012.

[44] V. E. A. Russo, R. A. Martienssen e A. D. Riggs. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

[45] W. J. Ansorge. Next-generation DNA sequencing techniques. *New Biotechnology*, 25:195–203, 2009.