Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Strumenti Statistici per elaborazione dati su Sequenziamenti di Genoma Umano

Relatore:
Prof. Gastone Castellani

Presentata da:
Simone Giannini

Anno Accademico 2015/2016

Sommario

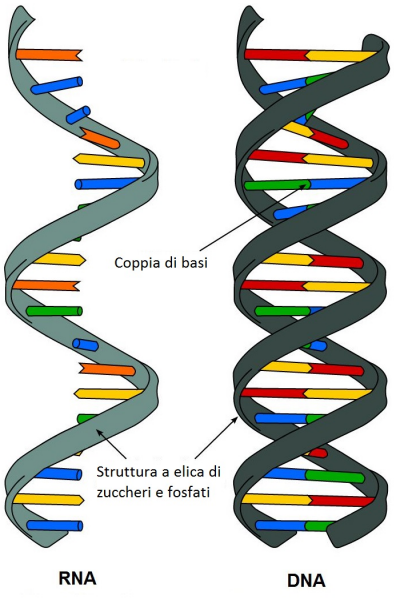
L'analisi del DNA è una delle chiavi per la comprensione della vita e dei suoi funzionamenti. Le tecniche di sequenziamento di nuova generazione NGS permettono una analisi parallela di molte sequenze che hanno reso possibili i sequenziamenti di genomi interi e l'impiego di questi dati in una vasta gamma di studi. In questa tesi verranno descritte le principali tecniche di sequenziamento NGS. Per quanto riguarda il genoma umano si tratteranno alcune tematiche di studio di varianti affrontate dal gruppo 1000Genomes. Nella fase conclusiva si introdurranno definizioni di statistica utili nell'affrontare l'elaborazione dei dati. Inoltre vengono descritti alcuni strumenti che permettono di svolgere questo tipo di analisi.

Indice

Capitolo 1

Introduzione al DNA

Il DNA (*acido desossiribonucleico*) è un polimero organico contenuto nelle cellule degli essere viventi e contiene tutte le informazioni genetiche necessarie alla sintesi cellulare. Questo polimero è formato da una doppia sequenza di monomeri detti nucleotidi, ognuno formato da un gruppo fosfato, uno zucchero pentoso (deossiribosio) e una base azotata. Due sequenze di queste unità formano la tipica struttura a doppia elica scoperta nel 1953 da James Watson e Francis Crick.
La catena è larga 2.2-2.6nm. Nel DNA sono presenti quattro tipi di basi azotate: adenina (A), timina (T), citosina (C) e guanina (G). Vedi Figura 1.1.



Le basi si legano al carbonio 1' dello zucchero. Le due catene vengono tenute insieme da legami idrogeno che si formano nelle coppie G-C e A-T, cioè una base purinica(G,A) forma legami con una base pirimidinica(C,T). Questi legami sono deboli e possono essere spezzati da alte temperatura o da azioni meccaniche come durante la replicazione del DNA. La stabilità delle coppie è differente poiché la coppia GC forma tre legami idrogeno ed è quindi sequenze contenenti molte di queste coppie sono maggiormente stabili.
Si sviluppa in lunghezza formando legami fosfodiesterici, cioè legando il carbonio 5' di un nucleotide al carbonio 3' del successivo tramite un gruppo fosfato. Ognuno è lungo 0.33nm.

L'RNA (*acido ribonucleico*) invece è formato da un singolo filamento ha come caratteristica la presenza della base uracile (U) al posto della timina. Quello messaggero

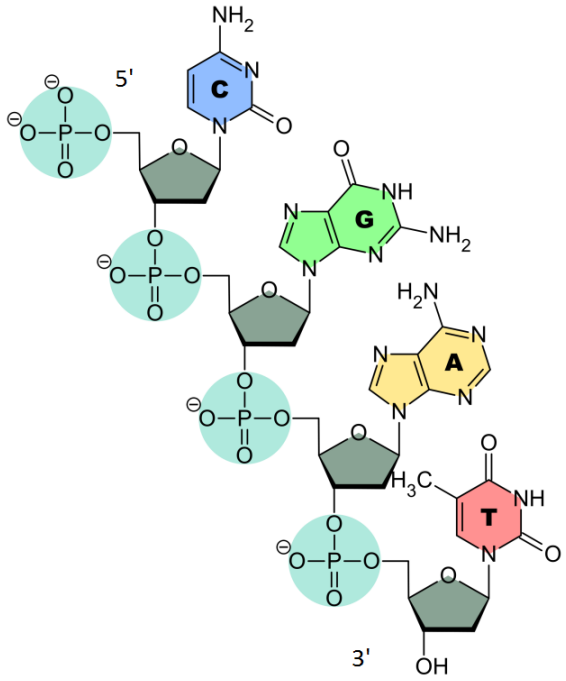


Figura 1.1: *Struttura chimica del filamento singolo di DNA.*

si forma nel processo di trascrizione del DNA da parte dell'enzima polimerasi. Viene utilizzato nella sintesi proteica per la formazione delle sequenze di aminoacidi.
Sia il DNA che l'RNA hanno una direzione, per il primo viene definita *senso* se la sue sequenza è la stessa del relativo mRNA. Quella sul filamento opposto viene detta *anti-senso*. Solo le sequenze di senso codificano per le proteine durante la polimerizzazione.
Il codice genetico è organizzato in *codoni*, ossia sequenze di tre nucleotidi, ai quali viene associato un aminoacido. Con questo vocabolario ci sono $4^3 = 64$ combinazioni alle quali vengono associati i 20 aminoacidi. Si ha perciò una ridondanza poiché ci saranno più codoni che identificano lo stesso aminoacido. Ne viene introdotto anche un altro detto codone di *stop* o di *nonsenso*, esso fa terminare la sequenza di aminoacidi durante la formazione delle proteine.

1.1 Il sequenziamento

Il sequenziamento del DNA è la determinazione delle sequenze di nucleotidi che formano il corredo genetico di un organismo.

Dentro il genoma di un organismo sono scritte tutte le sequenze di geni, codificanti proteine. Contiene anche informazioni per regolare l'espressione genica, ossia controllare le proprie funzioni sia esterne che interne. Lo studio delle sequenze è utile nella ricerca in biologia, in medicina nell'identificare malattie o agenti patogeni e conseguente sviluppo di cure appropriate. Il sequenziamento al giorno d'oggi si applica bene anche allo studio su larga scala del genoma umano.

Capitolo 2

Tecniche di sequenziamento Next-Generation

Le tecniche di sequenziamento di nuova generazione (NGS, *Next Generation Sequencing*) , hanno rivoluzionato la ricerca permettendo una acquisizione massiccia, parallela e ad alta definizione di dati dal DNA. Questi metodi risultano essere versatili e i dati raccolti si possono applicare, ad esempio, nell'identificazione su larga scala di polimorfismi a singolo nucleotide (SNP), nell'analisi di metilazione del DNA , nell'espressione del mRNA, nel sequenziamento completo del DNA etc.

2.1 Standard precedente, accenni sul Metodo di Sanger

Lo standard del sequenziamento è stato rappresentato dal metodo di Sanger (Premio Nobel per la Chimica nel 1980), detto anche metodo enzimatico o "chain termination method". E' necessario individuare il frammento di DNA che si vuole sequenziare e ottenerne un templato a singolo filamento. Vengono inseriti alcuni primer identici necessari alla polimerizzazione da parte della DNA-polimerasi. Nella polimerizzazione del DNA in questione non vengono utilizzati solo deossinucleotidi (dNTP) ma anche dideossinucleotidi (ddNTP).
Una molecola di DNA è formata da una catena di nucleotidi dNTP che si legano mediante legami fosfodiesterici. In questi legami un gruppo fosfato già legato al carbonio 5' dello zucchero del suo nucleotide, si lega al carbonio 3' dello zucchero di un altro nucleotide. Nel caso del ddNTP questo secondo legame non può avvenire a causa della mancanza di un gruppo idrossilico sul carbonio 3', fermando così la polimerizzazione. Poiché all'interno della soluzione sono presenti anche dNTP si ottengono molteplici pezzi di DNA di lunghezze diverse ognuna terminante con un certo ddNTP.
Si inseriscono separatamente i 4 tipi di nucleotidi, con una concentrazione adatta alla

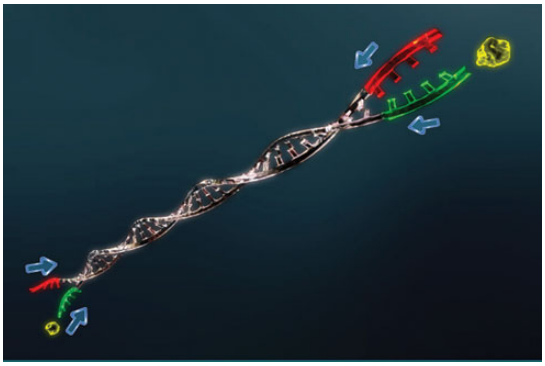
lunghezza del frammento da sequenziare, all'interno di quattro soluzioni identiche. I dideossinucleotidi, oppure i primer, vengono marcati radioattivamente.
Le soluzioni contenenti i nuovi filamenti di DNA con lunghezze diverse, polimerizzati in base ai nucleotidi disponibili, vengono passati in elettroforesi. L'elettroforesi è una tecnica che permette la separazione di particelle mediante un campo elettrico all'interno di un gel. I nucleotidi non sono neutri e quindi i frammenti di DNA vengono divisi in base alla loro lunghezza con una risoluzione di un nucleotide. Confrontando i risultati delle 4 soluzioni mediante una lastra autoradiografica o lampada UV si riesce a identificare la sequenza.

2.2 Roche 454

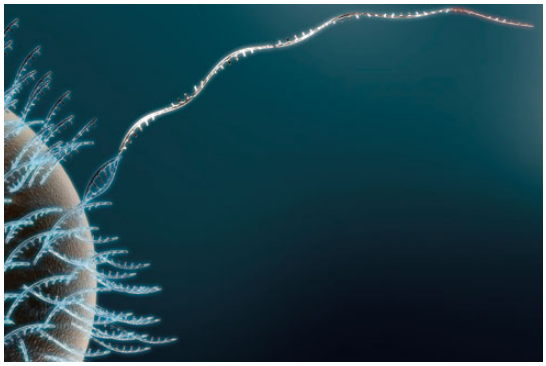
Roche 454 è una tecnica di sequenziamento per sintesi, ovvero si acquisiscono i dati durante la polimerizzazione delle catene di DNA. È stata introdotta sul mercato nel 2005 Si usa una tecnica di polimerizzazione a catena in una emulsione di micro-sfere di circa 28μm (em-PCR). Il sequenziamento vero e proprio avviene tramite Pirosequenziamento.

Tecnologia

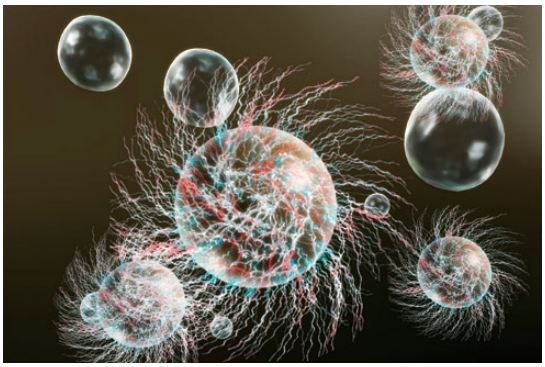
Per la preparazione dei campioni si frammenta il DNA per nebulizzazione in sequenze lunghe qualche centinaia di basi. [1] A questi frammenti biallelici si aggiungono, in entrambe le estremità, delle sequenze note. Questi poi vengono denaturati e uniti a una soluzione contenenti le micro-sfere, cresciute in modo da avere sulla loro superficie il frammento di DNA che combacia con le estremità precedentemente aggiunte ai frammenti.
Una emulsione di acqua e olio le isola durante la reazione di amplificazione em-PCR. In generale una PCR (Polymerase Chain Reaction) è una tecnica di amplificazione che permette di moltiplicare frammenti di acidi nucleici, conoscendone la sequenza iniziale e finale. Inserendo le sequenze note e il materiale necessario alla polimerizzazione, si ottiene una reazione a catena che va a completare le sequenze incomplete.
Ogni sfera sulla sua superficie conterà milioni di copie dello stesso frammento di DNA. Queste vengono poste su una superficie contenti celle molto piccole, adatte ad ospitare solo una micro-sfera e gli enzimi necessari.
Il Pirosequenziamento avviene dentro ognuna di queste celle simultaneamente tramite DNA polimerasi, ATP solforilasi, luciferasi, aspirasi, adenosinsolfofosfato (ASP) e luciferina. Inserito uno dei quattro dNTP, nel caso sia compatibile con il filamento di tempalto, la DNA polimerasi lo lega alla sequenza.
Il processo rilascia pirofosfato inorganico (PPi) che si trasforma in ATP dalla solforilasi, usando l'ASP come substrato. Grazie all'ATP e alla luciferasi si ottiene la conversione



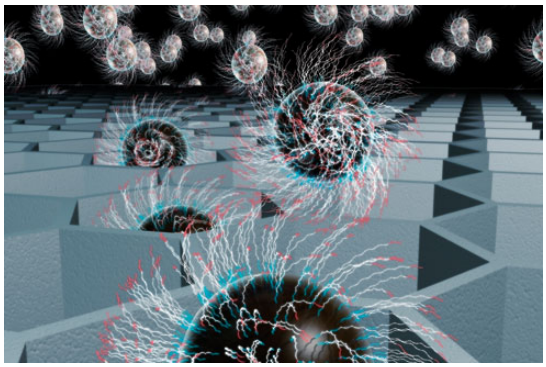
(a) *Preparazione del frammento.*



(b) *Accoppiamento delle sequenze con le microsfere.*



(c) *Amplificazione PCR.*



(d) *Posizionamento su superficie.*

Figura 2.1: *Immagini della sequenza di preparazione della Roche 454.*

della luciferina in ossiluciferina con la produzione di un segnale luminoso. Nel caso in cui venissero incorporati due o più nucleotidi uno di seguito all'altro, varierebbe solo l'intensità del segnale acquisito dal sensore CCD.
All'aggiunta dell'enzima aspirasi si ha la degradazione del dNTP non incorporato e dell'ATP prodotta. In questo modo è possibile inserire un altro dNTP e iterare il processo variando ogni volta il tipo di nucleotide. L'utilizzo dell'ATP come dNTP è da escludere poiché non si discriminerebbe più se la produzione di luce è riferita a una sequenza inglobata dalla polimerasi o semplicemente alla luciferasi, viene quindi sostituita con un suo analogo (adenosina-tio-trifosfato) che viene riconosciuta dal primo ma non dal secondo enzima. Vedi Figura 2.1.

Con questo metodo è possibile leggere migliaia di sequenze della lunghezza di 400-500 basi a ogni sequenziamento. La nuova versione del macchinario (GS FLX Titanium Sequencing Kit XL+) legge sequenze con lunghezze di 1kb [2].
Uno dei vantaggi che offre è che non soffre dell'alto contenuto di G-C e non esclude segmenti non clonabili. È invece meno efficace nella lettura di sequenze ripetitive ed omopolimeriche.
Il costo di sequenziamento è alto ma ha un basso costo se pesato sulla quantità di basi sequenziate, è quindi conveniente non usarlo per sequenziare frammenti genetici corti.

2.3 SOLiD

SOLiD è l'acronimo di "*Sequencing by Oligonucletide Ligation and Detection*" ed è una tecnica di sequenziamento sviluppato da Applied Biosystems e disponibile su mercato dal 2006. Viene anche chiamata "codifica a due basi" poiché la lettura avviene su una coppia di basi. Utilizza una amplificazione em-PCR simile al Roche ma su micro-sfere ancora più piccole. Non usa un enzima polimerasi ma DNA ligasi.

Tecnologia

La libreria da analizzare viene preparata aggiungendo opportune sequenze all'inizio e alla fine del materiale genetico frammentato.
Con l'utilizzo dell'amplificazione em-PCR, descritta precedentemente, si arricchiscono delle perline della dimensione di 1μm.
Con questa strategia solo il 30% di queste unità contiene il materiale genetico di interesse; si procede all'aggiunta di sfere in poliestere che si legano alla sequenza aggiunta ai frammenti da sequenziare. Il DNA così arricchito viene separato dal resto, il poliestere sciolto, ottenendo una percentuale di materiale genetico correttamente amplificato del 80%.
Viene poi trasferito e attaccato con legami covalenti su una superficie in vetro.
Il sequenziamento è composto da 5 parti, si introduce il primer apposito e vengono introdotte le sonde che verranno inserite a seguito attraverso la ligasi. Le sonde sono composte da 8 basi, partendo dal 3' abbiamo una coppia di basi che andranno combaciare con il filamento che si sta sequenziando e 5 basi universali. Al carbonio 5' è presente una tintura che mi codifica parzialmente la coppia di basi. Parzialmente perché i 4 colori che vengono usati sono distribuiti su 16 combinazioni di coppie, in modo da poterle discriminare conoscendo la base precedente. La sonda emana luce del relativo colore quando eccitata tramite un impulso laser, così facendo si staccano anche altre 3 basi rimanendo quindi con 5. La luce così prodotta viene rilevato A sequenza completata si lava il filamento prodotto e si inserisce un primer analogo che contiene un nucleotide in meno per poter sfasare le coppie. Si itera il procedimento in modo da avere due informazioni su ogni nucleotide da sequenziare. Il primer è conosciuto e quindi si ha l'informazione iniziale per decodificare il codice dei colori in basi. Vedi Figura 2.2.

La presenza di una doppia informazione su ogni base rende questo metodo molto accurato e ha un costo basso per base. Le sequenze di DNA sono abbastanza brevi, tipicamente 35 basi fino a un massimo di 150. Effettua oltre 500 milioni di letture per strumento per sequenziamento.[4]

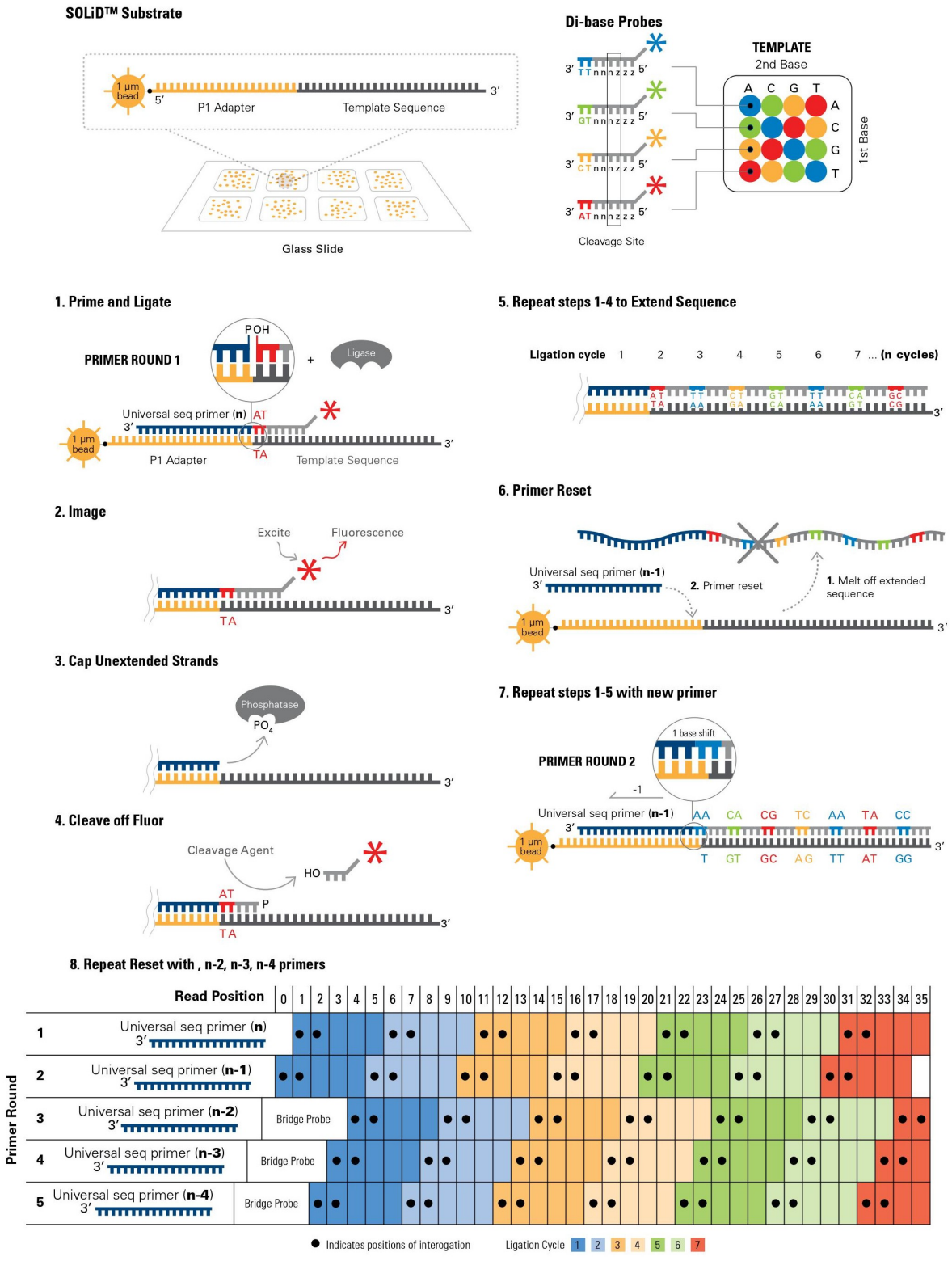


Figura 2.2: *Sequenza completa della tecnica SOLiD.*

2.4 ILLUMINA

Illumina è stato commercializzato inizialmente nel 2006 dalla Solexa, acquisita poi nel 2007. Il metodo è caratterizzato dall'uso di una particolare polimerasi che permette il sequenziamento tramite sintesi grazie a nucleotidi con marcatori fluorescenti di 4 colori.

Tecnologia

La preparazione della libreria avviene frammentando il DNA in segmenti con una lunghezza massima di 300bp[5]. Vengono aggiunti due sequenze adattatrici utili nelle prossime fasi.
Dopo essere stato denaturato, il materiale genetico viene immobilizzato su una superficie di vetro mediante le sequenze gemelle posizionate alle estremità del frammento. Si procede alla tecnica di amplificazione chiamata ponte PCR dove si formano sulla lastra delle isole contenenti le copie del DNA a seguito dei supporti complementari all'estremità libera. Utilizzando i nucleotidi e gli enzimi della PCR già descritta in precedenza, si ottengono delle isole compatte che rappresenteranno degli ottimi punti fissi e luminosi per l'acquisizione dati.
Dopo aver lavato gli elementi della PCR si inseriscono dei primer per iniziare la polimerizzazione dei nucleotidi che presentano diferenti marcatori fluorescenti. La DNA polimerasi lega queste basi una alla volta poiché non è possibile unire una nuova base fino a quando non si ha la fluorescenza, stimolata tramite laser.
Il segnale viene catturato da un sensore. È inoltre possibile ripetere l'acquisizione leggendo la sequenza opposta rispetto all'amplificazione effettuata, in questo modo si aumenta la qualità dei dati ottenuti.

I macchinari più moderni offrono 6 bilioni di letture a ciclo producendo un output da 1800 Gb in meno di tre giorni.[5]

2.5 Helicos tSMS

I sequenziatori della Helicos Biosciences utilizzano la tecnologia "true Single Molecule Sequencing" ed appartengono alla terza generazione di tecnologia di sequenziamento. Vennero introdotti sul mercato nel 2007
Si tratta di un sequenziamento per sintesi, la caratteristica principale è l'assenza di PCR. In questo modo si incorrono in molti errori dovuti all'amplificazione e alla perdita dell'informazione sulla quantità di materiale genetico.

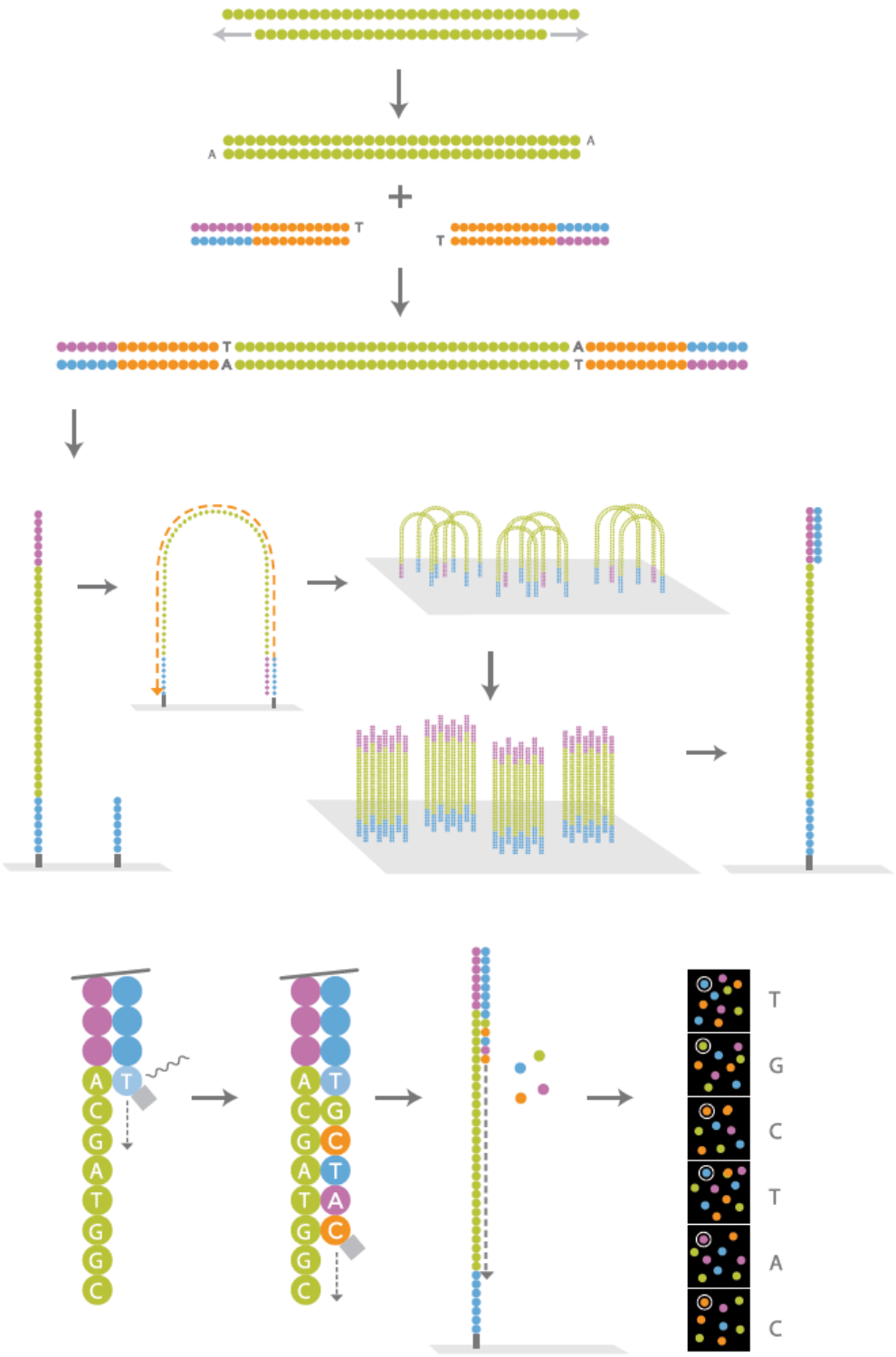


Figura 2.3: *Sequenza completa della tecnica ILLUMINA.*

Tecnologia

Il DNA viene denaturato e frammentato in sequenze lunghe qualche centinaio di basi.[7] Viene inserito un primer fluorescente e il tutto posizionato su di un substrato che li fissa con una densità pari a cento milioni di frammenti per centimetro quadro. La fluorescenza identifica i punti in cui è presente un segmento da sequenziare. Il sequenziamento avviene introducendo un tipo di nucleotide fluorescente alla volta, questo si lega tramite DNA polimerasi. Dopo aver sciacquato la superficie, si stimolato la fluorescenza mediante un laser e si acquisisce il segnale luminoso grazie a un sensore ad alta frequenza. Eliminato il marcatore fluorescente si ripete il processo per gli altri nucleotidi fino al completamento della sequenziamento. I cicli sono analoghi al sequenziamento per sintesi di Illumina.

Questa tecnica permette di analizzare molti milioni di frammenti di DNA simultaneamente, con un volume di sequenziamento dell'ordine del Gb.

2.6 Accenni a sequenziamenti di terza generazione

Negli ultimi anni si sono sviluppate tecniche che permettono di leggere sequenze di singolo DNA più lunghe senza alcuna PCR. Un esempio sono PACBio SMRT (single molecule real time sequencing) o l'utilizzo di nanopori i quali non verranno approfonditi.

2.7 Errori di sequenziamento

Nei sequenziamenti per sintesi sono presenti cluster di migliaia di copie dello stesso frammento; nel caso di Illumina le basi vengono eccitate da un laser per emettere fluorescenza. Tipicamente questo segnale è forte poiché generato simultaneamente da tutto il cluster e viene quindi acquisito da un dispositivo CCD.
L'instabilità, data dalla chimica stessa, causa errori di tipo stocastico. Possono essere di tre tipologie:

1. Errori di fase e prefase. Si hanno quando il segnale proveniente ad un cluster non è più univoco a causa di alcune sequenze che stanno sequenziando altre basi. Può essere causato da una errata incorporazione nel ciclo o nei precedenti.

2. Attenuazione del segnale. Tipicamente è causato da una perdita di materiale genetico sul campione stesso.

3. Errore di cross-talk. Causa una errata interpretazione del segnale.

Illumina per affrontare questi tipi di errori ha sviluppato "*Bustard*" che li gestisce separatamente. L'errore di cross-talk si tratta trasformando l'intensità del segnale in concentrazioni, definendo una matrice di interazione ed eliminando le sovrapposizioni mediante la matrice inversa.
Per l'attenuazione del segnale si rinormalizza le concentrazioni dividendo per il valore medio.
L'errore di fase viene affrontato tramite un modello markoviano.

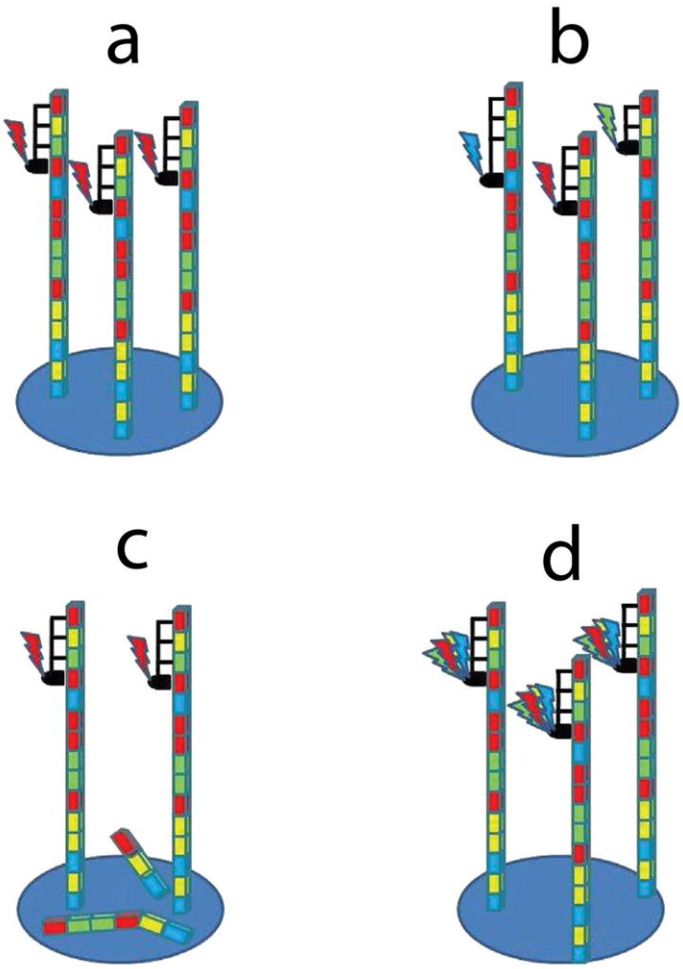


Figura 2.4: *Esempi di errori presenti in sequenziamento Illumina.* (**a**)Esempio di sequenziamento privo di errori stocastici. (**b**)Errore di fase. (**c**)Attenuazione del segnale. (**d**)Interpretazione errata del segnale

Capitolo 3

Sequenziamento del Genoma Umano

Lo studio del patrimonio genetico di un essere vivente rappresenta una firma unica dell'individuo e non solo della sua specie. Lo studio approfondito del genoma permette di comprendere i meccanismi cellulari e le strutture di certe proteine contenute nei geni, infatti il gene è la sequenza di DNA che codifica per una di queste. Il DNA degli organismi più complessi è composto da milioni di geni, infrapposti a sequenze non codificanti dei quali ancora non si conosce completamente il ruolo. Lo studio del genoma è esploso dopo l'introduzione delle tecniche di sequenziamento di nuova generazione che permettono l'acquisizione di grandi quantità di dati lavorando in modo parallelo.

3.1 Il Genoma Umano

Il DNA umano si trova nel nucleo della cellula in forma diploide, ossia sono contenuti due stessi geni codificanti la stessa proteina, ossia due alleli. Il numero di basi azotate del quale il genoma umano è composto si aggira sui 3.2 miliardi. Durante la fase di mitosi il DNA si sdoppia e si raggruppa formando i cromosomi. Il DNA nucleare umano forma 23 coppie di cromosomi omologhi. Per identificare la posizione dei geni viene introdotto il cosiddetto *locus genetico*.
Studiare il genoma umano, nello specifico è importante per la caratterizzazione e comprensione ci certe malattie, ma è utile anche per fare studi statistici sulle popolazioni per discriminarne la variazione.
Il primo progetto che prevedeva il sequenziamento del genoma umano (quindi una sequenza aploide del DNA) si chiama "*Human Genome Project (HGP)*"[8], è iniziato nel 1990 da una collaborazione internazionale e si è concluso nel 2003. Gli obbiettivi raggiunti era quello di comprendere le funzioni appartenenti al genere umano, identificare e mappare i geni contenuti in esso. Si è scoperto che il genoma contiene tra i 20000 e i 250000 geni. Il materiale codificante rappresenta solamente 1.5%. Il 36% è formato da introni(sequenze non codificanti la proteina, posizionati attorno all'esone), Pseudogeni,

regioni non traducibili (UTR) o frammenti genici. Il restante è DNA intragenico formato da sequenze ripetute. Mediante il confronto con altri organismi si è visto che non c'è correlazione tra complessità e numero di geni, infatti piante o altri animali non hanno una gran differenza in numero di geni rispetto all'essere umano.

Studi più recenti portati avanti dal gruppo *"The 1000 Genome Project Consortium"* hanno come scopo quello di catalogare le variazioni del genoma umano. Caratterizzando geograficamente e funzionalmente lo spettro delle variazioni sul genoma umano, si ha come scopo quello di costruire una risorsa che aiuti a capire il ruolo del DNA in certe patologie.

3.2 Varianti Genetiche

Uno studio coinvolgente 1092 individui di 14 diverse popolazioni, pubblicato nel 2012 dallo stesso gruppo, ha come risultato l'aver creato una mappa aplotipica di 38 milioni di polimorfismi a singolo nucleotide (SNP, *Single Nucleotide Polymorphism*), 1.4 milioni di piccoli inserimenti o mancanze (Indel, *in*sertion-*del*etion) e un numero di delezioni maggiori pari a 14000.

Nella fase pilota del progetto sono stati identificati il 95% degli SNP più comuni, ossia che si riscontrano con una frequenza maggiore del 5%. Le meno frequenti ed in particolare quelle fuori dall'esoma rimangono poco caratterizzate.
Le mutazioni puntuali meno frequenti hanno origini più recenti e sono importanti poiché potenzialmente funzionali e coinvolte, ad esempio, in varianti di sequenze proteiche soggette a selezione purificante debole, cioè che permangono per più tempo all'interno della popolazione prima di essere eliminate. Le difficoltà nell'identificarle è maggiore perché servono molti campioni e il potere di trovare associazioni è basso, inoltre molte sono specifiche di certe popolazioni.[12]
I 1092 individui sono stati scelti da 14 popolazioni diverse, suddivisibili in 4 aree geografiche: Europa, Asia dell'est, Africa sub-sahariana e americhe. I soggetti presi in considerazione con tecniche di sequenziamento differenti sono singole persone, coppie genitore figlio o tris genitori figli.
I dati raccolti sono frutto di analisi differenti: a bassa copertura sull'intero genoma (low-coverage whole genome sequence), sequenziamento profondo dell'esoma (targeted deep exome sequence) e informazioni provenienti da microarray. Questo scelta effettuata nella fase pilota è un compromesso utile per poter identificare il maggior numero di SNP e piccole inserzioni, ad eccezione dei più rari, contenendo i costi.

Sono state impiegate diverse tecnologie per convalidare la frequenza di accoppiamento dei campioni e valutare il tasso di falsa scoperta. Nelle regioni a bassa complessità si

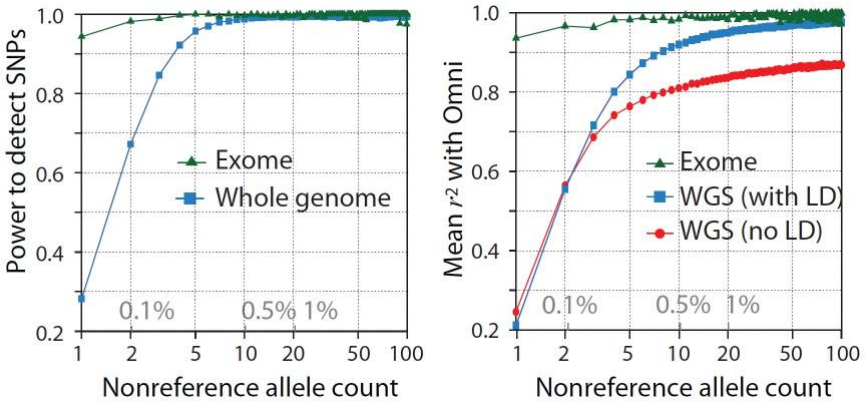


Figura 3.1: Potere di identificazione SNP come funzione del numero di varianti. Accuratezza del genotipo in funzione della frequenza di varianti

erano riscontrate delle ambiguità nella fase pilota, soprattutto per quanto riguarda gli indel.
Si definisce *"Genoma Accessibile"* la frazione di genoma per il quale dati di lettura brevi possono portare alla scoperta di varianti in modo affidabile. Nella fase pilota l'84% era accessibile, nella prima fase aumentando la lunghezza delle letture si è raggiunto il 94%, rendendo invalidi 1,7 milioni di SNP a bassa qualità.
Dal confronto tra dati esterni relativi agli SNP e il sequenziamento ad alta profondità si è stimato un potere di identificare la frazione di mutazioni puntuali, con una frequenza dell'1%, del 99.3% nel genoma e del 99.8% nell'esoma. Inoltre il potere di identificare quelle con una frequenza dello 0.1% si aggira sul 90% per l'esoma e sul 70% per il restante genoma. Nei siti in eterozigosi si ha una precisione oltre il 99% per gli SNP comuni e il 95% per quelli a frequenza di 0.5%. Se ai sequenziamenti completi del genoma a bassa intensità si aggiungono le informazioni di *LD* (linkage disequilibrium) la precisione nell'identificazione degli SNP con frequenza >1% è la medesima dei sequenziamenti ad alta intensità dell'esoma. Per quelli molto rari (<0.1%) non c'è differenza nell'inclusione del *LD* e si avrà una precisione bassa. Questa è naturalmente vicolata in primis dalla profondità del sequenziamento, ma anche dalla piattaforma da cui i dati provengono e da caratteristiche intrinseche della popolazione. L'accuratezza degli aplotipi prodotti è stata dedotta mediante l'analisi dei dati appartenenti alle terne genitori-figlio che sono stati sequenziati mediante "high coverage". Si deduce da questo che si compie un errore di fase ogni 300-400kb.

Per quanto riguarda le popolazioni l'obbiettivo del progetto *1000 Genome* era quello di identificare più del 95% degli SNP con frequenza dell'1% in una certa popolazione. L'obbiettivo è stato superato poiché si sono identificate nello studio "the Welcome Trust-funded UK10K project" circa 2500 genomi identificando $\sim$ 50%, 98%, 99.7% degli SNP

con frequenze rispettivamente di $\sim 0.1\%$, 1.0%, 5.0%. Con altre popolazioni, come per esempio nello studio *"the SardiNIA study"*, pur avendo circa 2000 genomi sequenziati, non si è raggiunto l'obbiettivo. Infatti se le popolazioni sono meno correlate l'obbiettivo è più difficile da raggiungere.

3.2.1 Costruzione di una mappa integrata di varianti

La risoluzione in aplotipi dei 1092 genomi è stato possibile mediante l'integrazione di dati provenienti da diverse tecnologie (Fase 1).
Gli individui vengono scelti in modo che non siano presenti parentele e si formano dei gruppi geografici o sulla base di caratteristiche ancestrali comuni con un minimo di 100 membri. La prima generazione di dati generati per ogni campione consistevano in sequenziamento completo a bassa copertura (con una media di 5x), sequenziamento ad alta densità su un target di 24Mb che comprendono più di 15000 geni e informazioni su array SNP ad alta densità.
Mediante differenti algoritmi si creano diverse varianti nell'ordinare i dati della lettura. Per poter ottenere i dati migliori, per ogni variante si identificano le qualità delle metriche come per esempio informazioni sull'unicità delle sequenze, la qualità delle prove a supporto delle varianti, la distribuzioni delle chiamate alle basi. Le informazioni multiple ottenute permettono di ordinare le varianti in modo da abbassare il coefficiente *FDR* (*False Discovery Rate*).
La verosimiglianza del genotipo viene identificata con il numero di copie presenti per sito e campione, nelle zone a espressione biallelica. Poiché l'evidenza di un singolo genotipo è debole nelle zone a basso sequenziamento (low coverage) e può essere molto variabile nella zona dell'esoma, vengono usati metodi statistici per inserire informazioni riguardanti zone in linkage disequilibrium che aiutano nell'identificazione degli aplotipi.

3.2.2 Varianti genetiche tra e dentro le popolazioni

Nell'analisi sono stati formati 5 macro gruppi in base alla predominanza di componenti ancestrali comuni: Europa(CEU, TSI, GRB, FIN, IBS), Africa (YRI, LWK, ASW), Est Asia(CHB, JPT, CHS), America(MXL, CLM, PUR). Le varianti con una frequenza $>10\%$ sono stati identificati su ogni popolazione. Dall'altro lato il 17% delle varianti a bassa frequenza nel range 0.5-5% sono osservabili in un solo gruppo ancestrale e il 53% delle varianti allo 0.5% sono solo all'interno di una singola popolazione.
Nei gruppi ancestrali le varianti comuni sono debolmente differenziate (secondo i valori di F_{st}, vedi Appendice **??** per la statistica F), anche se al di sotto dello 0.5% le frequenze delle varianti sono più del doppio. I gradi di variabilità varia tra le popolazioni scelte. Infatti nelle popolazioni IBS e FIN ci sono un eccesso di varianti rare che hanno avuto origine da alcune dinamiche di popolazione come per esempio selezione a collo di

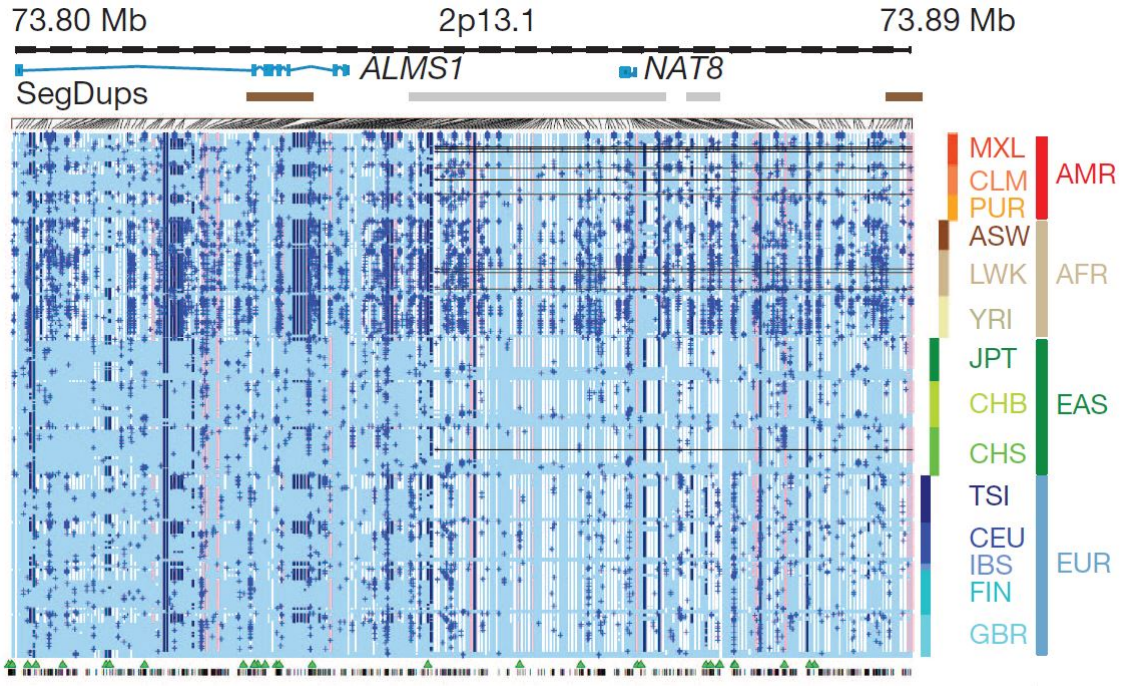


Figura 3.2: ***Esempio di distribuzione di varianti rare e comuni***[10]. Ogni riga rappresenta un aplotipo stimato per quella determinata popolazione. Gli alleli di riferimento sono indicati con il blu sullo sfondo. Le varianti con una frequenza oltre lo 0.5% sono indicate in rosa se ricavate da array SNP ad alta densità, in bianco quelle già note, in blu scuro le altre. Le varianti a frequenza <0.5% sono indicate con delle croci blu. Gli indel sono rappresentati da triangoli verdi e le nuove varianti da trattini. In alcuni aplotipi di certe popolazioni sono predenti dei tratti neri che indicano grossi tagli.

bottiglia. Alcune varianti comuni mostrano una forte differenziazione tra le popolazioni all'interno dei gruppi ancestrali e molte sembrano essere favorite da fattori esterni come adattamenti al luogo o comunque evolutivi.
La distribuzione delle frequenze alleliche derivate mostra sostanziali divergenze tra popolazioni sotto a una frequenza del 40%. Gli individui del gruppo ancestrale Africano contengono una densità tripla di varianti a bassa frequenza(0.5-5%) rispetto alle altre popolazioni, questo si pensa sia dovuto a un fenomeno evolutivo ancestrale a collo di bottiglia nelle popolazioni non africane. Tutte presentano un arricchimento di varianti rare (<0.5%) dovute alla crescita della popolazione e da adattamenti con differenti ambienti. Infatti le varianti rare hanno tipicamente origine più recente.

Le variazioni presenti almeno due volte all'interno della popolazione, vengono chiamate varianti f_2, tipicamente se recenti si trovano presenti nella medesima popolazione. Questo tipo di dato viene usato per eliminare errori di sequenziamento e discriminarli quindi da mutazioni a frequenza bassa, inoltre per analizzare la relazione di certe varianti nelle popolazioni. Tipicamente il 53% di queste vengono trovate all'interno della popolazione. Le restanti possono essere presenti in altre popolazioni a causa delle connessioni che ci sono tra le varie popolazioni sia recenti che passate. Si riesce con questi valori a identificare la probabilità di appartenenza a una certa popolazione. Per esempio se ho due individui, uno spagnolo(IBS) e un altro non spagnolo (IBS-X), con una certa variante f_2, il secondo è più probabile che appartenga alla popolazione americana piuttosto che a quella di un'altra nazione europea. Nell'est asia le popolazioni CHS e CHB mostrano un'alta condivisione di varianti f_2 rispetto alla JPT, anche se la popolazione JPT è più simile alla CHB piuttosto che alla CHS.
Evidenze indipendenti sull'età delle variazioni sono date dalla lunghezza della parte di aplotipo condivisa da cui appartengono. C'è una correlazione negativa tra la frequenza della variante e la lunghezza dell'aplotipo condiviso. Per le varianti a frequenza dell'1% gli aplotipi comuni sono lunghi 100-150kb. Errori di sincronizzazione ed errate chiamate alle basi si pensa che limitino di un fattore 2-3 la capacità di trovare sequenze comuni lunghe. In ogni caso la lunghezza degli aplotipi condivisi è utile per stimare l'età degli alleli. Dentro le popolazioni e tra quelle appartenenti allo stesso gruppo ancestrale si nota che le varianti f_2 sono in aplotipi condivisi più lunghi. Negli altri casi le varianti sono in sequenze molto corte.
È possibili stabilire la storia e determinare l'origine ancestrale di porzioni di DNA in una certa popolazione con caratteristiche miste. Il numero di varianti cambia tra le diverse popolazioni, varia anche la frequenza di varianti *sinonime* e *non-sinonime*, così come la proporzione di quelle nuove. Una variante si dice *sinonima* se la sua mutazione porta alla costruzione del medesimo peptide.

3.2.3 Considerazioni sui dati in ambito medico

I dati genomici raccolti sono largamente utilizzati nell'individuazione di malattie genetiche e nello studio dei tumori. La maggior parte degli SNP, rari o comuni, presenti nelle popolazioni non sono direttamente collegabili a conseguenze funzionali.
Gli individui tipicamente hanno più di 2500 varianti non-sinonime a posizione conservata, 20-40 identificate come pericolose e attorno a 150 varianti che comportano perdite di funzionalità(LOF *Loss-of-function*). La maggior parte di queste però sono comuni(>5%) o a bassa frequenza(0.5-5%) e quindi il numero delle rare è molto più basso; 130-140 varianti non-sinonime (per individuo), 10-20 LOF, 2-5 mutazioni dannose, 1-2 varianti identificate dal sequenziamento di alcuni tumori. Paragonando i dati con quelli il numero di varianti sinonimi si ha un eccesso di quelle rare, queste mutazioni possono essere sufficientemente gravi da non permettere una proliferazione nella popolazione ed avere quindi una frequenza bassa.

3.2.4 Studi Recenti-

Uno studio più recente [13] del gruppo 1000 Genomes Project ha ricostruito il genoma di 2,504 individui di 26 popolazioni diverse mediante una combinazione di tecniche low-coverage, sequenziamento profondo dell'esoma e l'uso di microarray. Le varianti trovate sono 88 milioni tra cui SNP e indel, identificando più del 99% delle varianti SNP con frequenza >1% per la maggior parte delle popolazioni ancestrali. Grazie a questi dati è stato possibile contribuire e convalidare 80 milioni di SNP sui 100 milioni totali contenuti ora nel catalogo dbSNP.

3.3 Varianti Strutturali-

Per varianti strutturali si intende una modificazione di sequenze intere di geni o un loro riarrangiamento. Alcuni esempi sono delezioni, inserzioni, duplicazioni e inversioni. Anche questo tipo di mutazioni sono coinvolte in numerose patologie. Uno studio del 2015[14] discute sulle varianti strutturali delle quali non è stato possibile trattare precedentemente se non per alcuni casi. Gli apoltipi di 26 popolazioni sono stati generati mediate ricostruzione da letture brevi del DNA e metodi statistici.
Rimangono ancora delle difficoltà nell'identificarle in regioni ad DNA ad alta complessità poiché presente su vari livelli. Con le normali tecniche di amplificazione PCR si perdono informazioni sulle quantità di materiale genetico che invece è una informazione cruciale per quanto riguarda l'identificazione di copie di sequenze.
Lo stesso studio è stato capace di individuare in totale 68818 SV di individui non correlati. Suddivisi in:

- 42279 delezioni bialleliche,
- 6025 duplicazioni bialleliche,
- 2929 mCNV (*multi allelci copy-number variants*), cioè variazioni del numero di alleli in un individuo,
- 786 inversioni,
- 168 inserzioni mitocondriali nucleari (NUMT)
- 16631 inserzioni mobili (MEI)

Il 60% di questi SV sono nuovi rispetto al database delle varianti genomiche DGV.

Questo risultato è stato possibile grazie ai dati della fase 3 di *1000 Genome Project*. I dati provengono dal sequenziamento completo del DNA (WGA,*whole-genome sequencing*) da tecnologie multiple, incluse letture lunghe e sequenziamenti a singolo filamento. I dati raccolti con Illumina WGS (lunghezza di lettura 100bp, 7.4x in media) sono stati mappati tramite due algoritmi indipendente, BWA e mrsFAST. Per l'identificazione delle varianti sono stati utilizzati 9 diversi algoritmi. L'indice di false scoperte FDR è stato calcolato mediante i dati raccolti con altre metodiche.
La sensibilità nell'identificazione di questa tipologia di dati è stata calcolata con l'aiuto di tecniche di sequenziamento differenti, quali *PacBio* e *deep-coverage*.

3.3.1 Proprietà di popolazione

Per l'analisi delle popolazioni sono stati formati 5 gruppi continentali (AFR, AMR, EAS, EUR, SAS). L'indice VAF (*Variant Allele Frequency*) definisce la frequenza con la quale una certa variante è presente nei sequenziamenti effettuati. Il 65% degli SV sono presenti con una bassa frequenza: $VAF < 0.2\%$. Quelle con $VAF > 2\%$ sono condivise in più continenti.
È stato osservato che il 72% degli SV con $VAF > 1\%$ e il 68% di quelli con $VAF > 0.1\%$ sono in linkage disequilibrium (LD) con un SNP ($r^2 > 0.6$), anche se varia molto considerando le diverse classi di varianti strutturali.
Dal catalogo di aplotipi basati su SV si nota un aumento di varianti per quanto riguarda la popolazione africana, compatibilmente con l'aumento degli SNP. È stato ricercato un segno di selezione adattiva studiando a stratificazione di alcuni SV nelle popolazioni i funzione dell'indice di frequenza. Sono state realizzate delle statistiche (V_{ST}) che possono essere applicate per indagare le stratificazioni di SV riferiti a uno o più alleli.

Capitolo 4

Analisi Statistica dei Dati

In qualsiasi esperimento ci sono dei dati che vanno interpretati, per farlo sono necessari metodi statistici. Questi sono utili anche nell'analisi dell'errore associati ai dati e a una loro eventuale elaborazione. La statistica in generale è alla base del metodo scientifico, non solo per una analisi descrittiva ma anche una inferenziale. È possibile dedurre informazioni da un campione di dati presi casualmente.
Come tutte le discipline anche la statistica ha una sua terminologia propria.
Un *evento* casule è un possibile risultato di un esperimento governato da una certo processo stocastico.
Una *popolazione* è un set di tutti i possibili eventi, cioè di tutti i potenzialmente osservabili.
La *probabilità* è un concetto base che esprime in un certo modo il grado di affidabilità. Si può pensare come la frequenza di un evento riferito a una serie infinita.

4.1 Distribuzioni

4.1.1 Definizione di distribuzione di Probabilità

Definiamo una funzione distribuzione F(t), tale che specifichi la probabilità P di trovare un valore di x minore di t:

$$F(t) = P\{x < t\}, \qquad con -\infty < t < \infty \tag{4.1}$$

Secondo gli assiomi della probabilità devono essere valide anche le seguenti proprietà per la funzione di distribuzione:

1. F(t) è non decrescente in funzione di t,
2. $F(-\infty) = 0$,

3. $F(\infty) = 1$.

Queste possono essere continue o discrete.

4.1.2 Distribuzioni Discrete

Una distribuzione discreta assegna la probabilità, dati un numero N di differenti eventi ordinati, a una variabile reale x_i, con i=1,...,N, con N finita o infinita. La probabilità $p(x_i)$ di osservare un valore x_i soddisfa la condizione di normalizzazione:

$$\sum_{i=1}^{N} x_i = 1 \tag{4.2}$$

Questa è definita da:

$$p(x_i) = P\{x = x_i\} = F(x_i + \epsilon) - F(x_i - \epsilon), \tag{4.3}$$

con ϵ numero positivo e inferiore alla distanza tra due valori adiacenti.

4.1.3 Distribuzioni Continue

Nel caso delle distribuzioni di continue introduciamo, al posto della probabilità discreta, la *densità di probabilità* $f(x)$, definita come:

$$f(x) = \frac{dF(x)}{dx}. \tag{4.4}$$

Questa è definita nell'intero raggio $-\infty < x < \infty$. Possiede le seguenti proprietà:

1. $f(-\infty) = f(+\infty) = 0$,
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

La probabilità $P\{x_1 \leq x \leq x_2\}$ di trovare una variabile random x nell'intervallo $[x_1, x_2]$ è data da:

$$P\{x_1 \leq x \leq x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\,dx. \tag{4.5}$$

4.1.4 Valore d'aspettazione

Il valore di aspettazione $E(u)$ di una quantità $u(x)$, che dipende dalla variabile casuale x, può essere ottenuto collezionando un numero infinito di valori x_i dalla distribuzione $f(x)$, e infine mediando su questi valori. Le definizioni sono le seguenti:

$$E(u(x)) = \sum_{i=1}^{\infty} u(x_i)p(x_i) \quad (\textit{distribuzione discreta}), \tag{4.6}$$

$$E(u(x)) = \int_{-\infty}^{\infty} u(x) f(x)\, dx \quad (\textit{distribuzione continua}). \tag{4.7}$$

Si assume che l'esistenza della serie e dell'integrale, restringendo così le condizioni sulle funzioni u, p, f. Definite c costante, u e v funzioni di x, seguono le seguenti relazioni che denotano la linearità di E:

$$E(c) = c, \tag{4.8}$$
$$E(E(u)) = E(u), \tag{4.9}$$
$$E(u + v) = E(u) + E(v), \tag{4.10}$$
$$E(cu) = cE(u). \tag{4.11}$$

Se x ed y sono variabili indipendenti vale anche:

$$E(u(x)v(y)) = E(u)E(v). \tag{4.12}$$

Viene anche indicato con la notazione seguente:

$$E(u) \equiv \langle u \rangle.$$

4.1.5 Valore medio

Il valore di aspettazione della variabile x è detto *valore medio*. Il valore medio di una variabile aleatoria rappresenta la previsione teorica del valore che mediamente tale variabile assumerà nell'ipotesi di eseguire un numero elevato di prove. Viene indicato con μ. Definizione:

$$E(x) \equiv \langle x \rangle = \mu = \sum_{i=1}^{\infty} x_i p(x_i) \qquad (\textit{distribuzione discreta}), \tag{4.13}$$
$$E(x) \equiv \langle x \rangle = \mu = \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad (\textit{distribuzione continua}). \tag{4.14}$$

Questo valore si distingue dalla *media dei valori*1, definita su un numero finito N di variabili, $x_1, \ldots, x_N$, definite dal simbolo $\overline{x}$:

$$\overline{x} = \frac{1}{N} \sum_i x_i. \tag{4.15}$$

4.1.6 Varianza

La *varianza* si indica come σ^2 e misura la larghezza della distribuzione. È definita come la deviazione quadratica media della variabile dal suo valore medio.

La radice quadrata della varianza è detta *deviazione standard* e si usa come incertezza per una certa variabile stocastica. La definizione è la seguente:

$$var(x) = \sigma^2 = E\big[(x-\mu)^2\big]. \tag{4.16}$$

Dalla definizione segue che

$$var(cx) = c^2 var(x), \tag{4.17}$$

e che σ/μ è indipendente dalla scala di x.
Dalle proprietà del valore di aspettazione $E()$ otteniamo

$$\begin{aligned} \sigma^2 &= E(x^2 - 2x\mu + \mu^2) \\ &= E(x^2) - 2\mu^2 + \mu^2 \\ &= E(x^2) - \mu^2, \end{aligned} \tag{4.18}$$

dalla quale segue che

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 = \langle x^2 \rangle - \mu^2. \tag{4.19}$$

Inoltre la varianza è invariante rispetto alla traslazione della distribuzione di un parametro a:

$$x \to x + a,\ \mu \to \mu + a \ \Rightarrow\ \sigma^2 \to \sigma^2.$$

Nel caso di due variabili indipendenti x_1 e x_2, che seguono distribuzioni differenti con valori medi μ_1 e μ_2, varianze σ_1^2 e σ_2^2, la somma $x = x_1 + x_2$, avrà come varianza:

$$\begin{aligned} \sigma^2 &= \langle (x - \langle x \rangle)^2 \rangle \\ &= \langle ((x_1\mu_1) + (x_2\mu_2))^2 \rangle \\ &= \langle (x_1\mu_1)^2 + (x_2\mu_2)^2 + 2(x_1\mu_1)(x_2\mu_2) \rangle \\ &= \langle (x_1\mu_1)^2 \rangle \langle + (x_2\mu_2)^2 \rangle + 2\langle (x_1\mu_1) \rangle \langle (x_2\mu_2) \rangle \\ &= \sigma_1^2 + \sigma_2^2. \end{aligned} \tag{4.20}$$

La deviazione standard di una certa variabile, formata dalla somma di altre due aleatorie, è dato dalla somma in quadratura delle singole deviazioni standard. Si può generalizzare alla somma $x = \sum x_i$ di N varianti:

$$\sigma^2 = \sum_{i=1}^{N} \sigma_i^2.$$

4.1.7 Distribuzione Gaussiana

La distribuzione gaussiana, chiamata anche *normale*, è una distribuzione di probabilità continua:

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)}. \tag{4.21}$$

Rispetta le condizioni di distribuzione di probabilità descritte precedentemente nelle sezioni 4.1.1, 4.1.3, come per esempio la norma,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)} = 1. \tag{4.22}$$

La somma di quantità distribuite normalmente è ancora normalmente distribuita:

$$\mu = \sum \mu_i, \quad \sigma^2 = \sum \sigma_i^2.$$

A essa convergono le distribuzioni discrete come la binomiale o quella di Poisson, ma anche la χ^2, nel limite di molti numeri, valore di aspettazione alto, molti gradi di libertà. Si adatta bene a molti fenomeni osservabili in natura. Infatti il *teorema del limite centrale* ci dice che il valore medio di un numero grande N di variabili indipendenti casuali, che seguono la stessa distribuzione con varianza σ_0^2, è soggetto a una distribuzione normale con $\sigma^2 = \sigma_0^2/N$.

4.1.8 t di Student

Una valutazione non approssimativa del livello di confidenza ricavato da un campione finito di N misure può essere intrapresa solo se si conosce la forma della distribuzione. Se la distribuzione è normale e si conosce la media è possibile verificarne la compatibilità di un set di N misure. La distribuzione di student descrive l'andamento di una variabile data da:

$$t = \frac{\overline{x} - \mu}{s} \quad con\ \overline{x} = \sum x_i/N. \tag{4.23}$$

Il numeratore è la differenza tra la media dei valori e quella della distribuzione gaussiana. La distribuzione è centrata in zero e ha come deviazione standard s.

$$s^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2. \tag{4.24}$$

La somma di destra se divisa per la varianza σ^2 della gaussiana, segue la distribuzione del χ^2 con $f = N - 1$ gradi di libertà. Dividendo t per la deviazione standard $\sigma/\sqrt{N}$.

La forma analitica della distribuzione di probabilità è:

$$h(t|f) = \frac{\Gamma((f+1)/2)}{\Gamma(f/2)\sqrt{\pi f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{f+1}{2}}. \tag{4.25}$$

4.1.9 Distribuzione Gamma

La *distribuzione Gamma* è una distribuzione di probabilità continua e comprende le distribuzioni esponenziale e chi-quadrato:

$$G(x|\nu,\lambda) = \frac{\lambda^\nu}{\Gamma(\nu)}\, x^{\nu-1}\, e^{-\lambda x}\,, \quad x > 0. \tag{4.26}$$

$\lambda > 0$ è un parametro di scala, $\nu > 0$ determina la forma della distribuzione. In generale la funzione $\Gamma(z)$ è la *Gamma di Eulero*, così definita:

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}\, e^{-t}\, dt. \tag{4.27}$$

Impone $\nu = 1$ si ottiene la distribuzione esponenziale:

$$f(x) = \lambda e^{-\lambda x}. \tag{4.28}$$

Per ottenere invece la distribuzione χ^2 con f gradi di libertà, è necessario imporre $\nu = f/2$ e $\lambda = 1/2$:

$$g_f(x) = \frac{1}{\Gamma(f/2)\, 2^{\,f/2}}\, x^{f/2-1}\, e^{-x/2}\,. \tag{4.29}$$

Nella statistica bayesiana è comune sia come distribuzione a priori che come distribuzione a posteriori.

Il valore atteso è

$$E(x) = \mu = \nu/\lambda\,, \tag{4.30}$$

mentre la varianza risulta

$$\sigma^2 = \nu/\lambda^2\,. \tag{4.31}$$

Se x segue la distribuzione Gamma $G(x|\nu,\lambda)$ allora anche αx segue la distribuzione Gamma $G(\alpha x|\nu,\alpha\lambda)$.

Data una successione di $x_1,\ldots,x$ di variabili aleatorie indipendenti, ognuna con distribuzione $G(x_i|\nu_i,\lambda)$, la somma $X = x_1 + \cdots + x_n$ segue la distribuzione $G(X|\nu_1 + \cdots + \nu_n,\lambda)$.

4.2 Introduzione alla statistica non-parametrica-

I metodi parametrici utilizzati per la soluzione di problemi hanno, come limitazione, la necessità di dover ricorrere all'introduzione di ipotesi molto restrittive, difficilmente giustificabili e interpretabili, irrealistiche, non sempre chiare, formulate ad hoc.
Le assunzioni che rendono valida l'applicazione di tali metodi sono di norma raramente soddisfatte e i risultati sono spesso ottenuti tramite approssimazioni. Questi metodi sono applicati in svariati campi e sono utili quando:

- non è nota la distribuzione,
- non esiste una normalizzazione,
- se l'inferenza riguarda variabili di tipo qualitativo,
- nel caso in cui il numero di campioni è inferiore al numero di variabili.

4.2.1 Mediana e rango

Nell'ambito non parametrico il valore maggiormente indicativo della distribuzione è la mediana e rappresenta il valore centrale. Per una distribuzione continua risulterà il valore Me tale per cui:

$$P(X \geq Me) = P(X \leq Me) = \frac{1}{2}, \tag{4.32}$$

Nel caso discreto, il rango di un valore centrale risulta essere la sua posizione ottenuta dopo aver ordinato la variabile.

4.3 Strumenti per l'elaborazione dei dati

Diversi studi hanno identificato una mutazione somatica in campioni tumorali utilizzando le moderne tecniche di sequenziamento. Queste informazioni vengono usate per l'identificazione della tipologia di carcinoma e identificare quindi una terapia mirata. Le tecniche di sequenziamento inerenti l'esoma, si stanno diffondendo poiché permettono di identificare oltre 25000 varianti SNP con costi contenuti.
Le metodologie utilizzate sono diverse e ognuna ha una propria peculiarità. Le difficoltà che bisogna superare sono: determinare frequenze alleliche basse a causa della vasta varietà di tumori, differenziare mutazioni da errori di sequenziamento o allineamento, classificare le mutazioni somatiche da quelle germinali, analizzare campioni misti di cellule sane e malate.
Le fasi dell'elaborazione dei dati sono principalmente due:

1. l'allineamento dai sequenziamenti e generazione dei primi parametri di validità del sequenziato,
2. confronto dei dati ottenuti (come per esempio la determinazione delle varianti).

4.3.1 Phred

Phred è un algoritmo che, mediante l'analisi dei picchi cromatografici, è in grado di risalire al alla sequenza che li ha generati ed assegnare ad ogni base un certo punteggio

(detto appunto punteggio *phred*) che definisce la qualità della scelta. Per fare questo prende in considerazione informazioni come l'ampiezza e la forma dei picchi, calcolando la qualità di ogni base. La qualità (Q) è legata alla probabilità di errore (P) mediante una relazione logaritmica:

$$Q = -10 \log_{10}(P). \tag{4.33}$$

Dato un valore Q la formula è semplicemente l'inversa della precedente

$$P = 10^{-Q/10}. \tag{4.34}$$

A un valore $Q = 10$ è associato un tasso di errore del 10%, per $Q = 20$ un errore dell'1%, per $Q = 30$ lo 0.1% e così via. Va considerato che avendo milioni, o addirittura miliardi di basi sequenziate, anche un valore come l'1% può portare a un gran numero di errori.

4.4 Allineamento dei dati

I software che effettuano questo tipo di elaborazione utilizzano un genoma di riferimento come per esempio *GRCh37/hg19* presente nelle banche dati. 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 formati tipicamente usati dagli strumenti di sequenziamento NGS sono il SAM(Sequence Alignment Map) e BAM(Binary Alignment Map).

-MOSAIK- Si adatta alle principali tecnologie NGS ed è l'unico allineatore a creare mappe in modo coerente rispetto a una molteplicità di dati. Utilizza l'algoritmo di Smith-Waterman che confronta segmenti di tutte le possibili lunghezze invece che guardare alla sequenza completa.

-BWA- Il programma si basa sulla trasformata di Burrows-Wheeler, ossia un algoritmo di compressione reversibile che permuta l'ordine dei caratteri, senza cambiarne il valore. Questo algoritmo tiene conto anche dei possibili gap.

4.5 Chiamata alle Varianti

Generalmente gli strumenti analizzano le mutazioni somatiche in modo o indipendente o simultaneo rispetto alle cellule mutate e sane.
Nel cercare queste mutazioni si incorre in troppi falsi positivi se si cerca di inglobare tutti gli eventuali positivi reali, oppure si perdono troppe mutazioni reali per ridurre il numero di quelle false. Così facendo nel primo caso si utilizza molto tempo per cercare di

discriminare la veridicità dei risultati, nel secondo caso invece si scartano delle mutazioni che potrebbero essere patogene.

4.5.1 GATK-LOD$_N$

GATK-LOD$_N$[17] è un metodo di analisi dei dati genomici che sfrutta le potenzialità di due strumenti standard: *MuTect* e *GATK*.

MuTect: è un software che permette un'identificazione delle mutazioni somatiche puntuali, è caratterizzato da una analisi simultanea di campioni misti. Pur avendo un tasso di identificazione inferiore ad altri metodi, ha il più alto tasso di convalida per mutazione. Il suo funzionamento può essere riassunto in tre punti[18].

- Nella prima fase si allineano le letture di tutto il campione, sia tumore che tessuto sano. In questo passaggio si ignorano sequenze contenenti molte letture o punteggi di qualità bassi che rappresentano più una fonte di rumore che di informazione.
- Una analisi statistica individua i siti che contengono mutazioni somatiche con alta confidenza. L'analisi statistica le identifica tramite l'uso di due parametri bayesiani. Il primo ha lo scopo di identificare nel tessuto tumorale quando si ha la mutazione rispetto alla sequenza di riferimento. La seconda si assicura che nel tessuto sano non siano presenti varianti. La classificazione viene effettuata calcolando il punteggio *LOD* (log odds) e comparandolo con il valore di soglia del logaritmo della probabilità a priori dell'evento considerato. Nel tumore:

$$LOD_T = log_{10}\left(\frac{P(dati\ osservati\ nei\ tumori|sito\ mutato)}{P(dati\, osservati\ nei\ tumori|sito\ non\ mutato)}\right). \tag{4.35}$$

  Nel tessuto non tumorale:

$$LOD_N = log_{10}\left(\frac{P(dati\ osservati\ nel\ tessuto\ normale|sito\ non\ mutato)}{P(dati\ osservati\ nel\ tessuto\ normale|sito\ mutato)}\right). \tag{4.36}$$

- Il processo successivo elimina gli artefatti creati dal sequenziamento stesso, dall'allineamento di sequenze brevi ed altre letture errate. Ad esempio, ci sono situazioni che portano a una lettura errata solo se si sequenzia in un verto verso, perciò è utile verificare se è presente su entrambe le direzioni verificando l'altro allele.

GATK (Genome Analysis Toolkit) è un insieme di strumenti efficaci nell'analisi dei dati genetici. I suoi metodo di chiamata alle varianti si differenzia da *MuTect* nel maggior spettro di individuazioni, ma anche per la minor validità delle stesse. L'analisi sulle cellule sane e tumorali avviene indipendentemente.

Papeline GATK-LOD$_N$

La papeline è molto simile a quelle utilizzate con gli strumenti GATK introducendo il punteggio LOD_N di MuTect.

Le sequenze lette sono sottoposte a controllo di qualità da parte dello script *fastq_ quality_ filter.pl* e *fastq_ quality_ trimmer.pl* contenuti in FASTX-Toolkit. Si è scelto un valore *phred* di soglia > 20, se non raggiunto le basi vengono segnalate. Se la lettura di una sequenza contiene più dell'80% di basi con basso phred, viene segnalata come anomala fino a quando non si raggiunge un valore di affidabilità maggiore, oppure viene scartata.

Per l'allineamento di queste sequenze si fa riferimento al genoma umano *GRCh37/hg19* usando BWA-MEN (con parametri di default). Un altro tool chiamato *Picard* si occupa delle procedure post allineamento. I dati vengono riordinati localmente attorno agli *Indel* e si ricalibrano i punteggi della qualità delle chiamate alle basi(BQSR, *base quality score recalibration*) contenuto in GATK3.0.

Per la chiamata alle varianti a singolo nucleotide si filtrano quelle trovate da GATK con lo strumento di ricalibrazione del punteggio di qualità (*Variant Quality Score Recalibration*).

Alle varianti SNV trovate si è adattato MuTect per trovare i valori del classificatore bayesiano LOD_N. Se il valore supera una certa soglia allora la mutazione è somatica. In questo filtraggio si selezionano solo letture con punteggi ≥ 8 nei campioni normali e ≥ 14 nei campioni tumorali. I risultati finali saranno una unione dei dati di MuTect e GATK-LOD$_n$.

Le varianti sono analizzate da *ANNOVAR* rispetto a database di commenti sul genoma umano. Si cercano corrispondenze, mediante i dati di altri studi come 1000 Genome, con varianti non-sinonimi e con perdita o guadagno del codone di stop. Le varianti presenti negli studi di riferimento con minor frequenza allelica (MAF)> 0.05 vengono rimosse.

Bibliografia

[1] Susmita Datta, Somnath Datta, Seongho Kim, Sutirtha Chakraborty, and Ryan S. Gill. *Statistical Analyses of Next Generation Sequence Data: A Partial Overview*. NIH-PA Author Manuscript, Departments of Bioinformatics Biostatistics Mathematics, University of Louisville USA, 2010

[2] Pirosequenziamento 454, Settembre 2016, URL: `http://454.com/applications/whole-genome-sequencing/index.asp`

[3] Ronaghi, Mostafa and Uhlén, Mathias and Nyrén, Pål. *A Sequencing Method Based on Real-Time Pyrophosphate*. SCIENCE 17 JUL 1998 : 363-365

[4] Heinz Breu, *A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction*, White Paper SOLiD™ System, USA, 07/2010

[5] Sequenziamento Illumina, Settembre 2016, URL: `http://www.illumina.com`

[6] Wilhelm J. Ansorge, *Next-generation DNA sequencing techniques*, New Biotechnology, Volume 25, Issue 4, April 2009, Pages 195-203, ISSN 1871-6784

[7] Sequenziamento tSMS, Settembre 2016, URL: `http://seqll.com/technology-information`

[8] Progetto Genoma Umano, Settembre 2016, URL: `http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml`

[9] International Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome.*, Nature, Ottobre 2004;431(7011):931-945, PMID 15496913.

[10] 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA.
An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632.

[11] The 1000 Genomes Project Consortium. *An integrated map of genetic variation from 1,092 human genomes: Supplementary Material.* doi:10.1038/nature11632

[12] Tennessen JA, Bigham AW, O'Connor TD, et al. *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.* Science (New York, NY). 2012;337(6090):64-69. doi:10.1126/science.1219240.

[13] The 1000 Genomes Project Consortium.
A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi:10.1038/nature15393.

[14] Sudmant PH, Rausch T, Gardner EJ, et al.,
An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81. doi: 10.1038/nature15394.

[15] Gerhard Bohm, Günter Zech. *Introduction to Statistics and Data Analysis for Physicists.* DESY, 2010. ISBN:978-3-935702-88-1

[16] John R. Taylor. *Introduzione all'analisi statistica degli errori, lo studio delle incertezze nelle misure fisiche.* Zanichelli, seconda edizione, 2012. ISBN:978-88-08-17656-1

[17] Ítalo Faria do Valle, Enrico Giampieri, Giorgia Simonetti, Antonella Padella, Marco Manfrini, Anna Ferrari, Cristina Papayannidis, Isabella Zironi, Marianna Garonzi, Simona Bernardi, Massimo Delledonne, Giovanni Martinelli, Daniel Remondini, Gastone Castellani.
Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. In fase di pubblicazione (Ottobre 2016).

[18] MuTect, Ottobre 2016,
`http://archive.broadinstitute.org/cancer/cga/mutect`.
Cibulskis, K. et al. *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnology (2013).doi:10.1038/nbt.2514

[19] Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, et al. (2014).
MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLoS ONE 9(3): e90581. doi: 10.1371/journal.pone.0090581