Alma Mater Studiorum · Università di Bologna

**SCUOLA DI SCIENZE**
**Corso di Laurea in Informatica Magistrale**

# UNA BASE DATI PER IL KNOWLEDGE DISCOVERY IN GENETICA MEDICA

**Tesi di Laurea in Complementi di Basi di Dati**

**Relatore:**
**Chiar.mo Prof.**
**DANILO MONTESI**

**Presentata da:**
**STEFANO GIOVANNI RIZZO**

**Sessione II**
**Anno Accademico 2012/2013**

# Indice

# 1 | Introduzione

Le tecnologie di sequenziamento NGS (Next Generation Sequencing) hanno permesso il sequenziamento del DNA a velocità senza precedenti e a costi sempre più bassi. Come conseguenza, negli 8 anni dalla nascita di queste tecnologie ad oggi, il numero di dati relativi a sequenze biologiche è aumentato considerevolmente, e ci si aspetta che questa quantità cresca sempre più rapidamente. I grandi progetti genomici come *HapMap*, che cataloga le varianti genetiche comuni negli esseri umani, e *1000 Genomes*, che è giunto a descrivere i genomi di 1,092 individui provenienti da 14 popolazioni [Abecasis et al., 2012], hanno contribuito alla crescita esponenziale di questa quantità di dati (vedere figura 1.1) e allo sviluppo di tecnologie sempre più efficienti.

Insieme al sequenziamento su larga scala di campioni di DNA è cresciuto l'interesse da parte della comunità scientifica biologica, medica, bioinformatica. In un arco temporale relativamente breve sono stati sviluppati e prodotti strumenti hardware, software di analisi, algoritmi, banche dati pubbliche, database privati e infrastrutture a supporto della genomica.

Una prima conseguenza di questa esplosione nell'interesse scientifico e commerciale verso la genomica è la mancanza di standard di riferimento per i dati biologici. Anzitutto la competizione tra le case produttrici di macchine NGS ha portato alla produzione di tecnologie diverse tra loro e in continua evoluzione, imponendo formati differenti per i dati in output. Lo sviluppo di numerosi strumenti software per l'elaborazione dei dati NGS e per l'analisi statistica delle informazioni genetiche ha contribuito ulteriormente alla diffusione di formati diversi e incompatibili tra loro.

Un risultato diretto della crescente quantità di letture del DNA è la dimensione dei dati generati, che continua a rappresentare una sfida per le infrastrutture che mantengono questi dati e per i software e gli algoritmi di analisi statistica delle sequenze. Basti pensare che fino al 2012 sono state sequenziate più di $13 \times 10^{15}$ basi nucleotidiche e che per un genoma umano, contenente circa 3 miliardi di basi nucleotidiche, sono necessari circa 100 gigabyte di dati.

In figura 1.1 è mostrato in giallo l'andamento del costo del sequenziamento pre-NGS in basi nucleotidiche per dollaro, e in rosso il costo del sequenziamento con tecnologia NGS. Come è evidente la quantità di basi sequenziabili con un dollaro ha subito una crescita esponenziale dalla nascita delle tecnologie NGS. Il progetto 1000 Genomes ad esempio ha prodotto nei primi 6 mesi una quantità doppia di dati rispetto a quanto fosse stato sequenziato nei 30 anni precedenti [Stein et al., 2010]. Ancora più interessante è rapportare questa crescita al costo dello storage in megabyte per dollaro, il cui andamento (in blue nella figura 1.1) segue la legge di Moore. Come molti sanno la legge di Moore, coniata dal cofondatore di Intel Gordon Moore, descrive la crescita esponenziale nel tempo della complessità nei circuiti integrati, affermando che il numero di transistor in un circuito integrato raddoppia approssimativamente ogni 18 mesi. I costi di sequenziamento e di conseguenza la quantità di dati prodotti hanno un andamento molto più veloce, dimezzando approssimativamente ogni 5 mesi. I dati biologici sono quindi diventati più costosi da memorizzare che da generare, e la loro entità è tale da essere ormai considerati Big Data.

Lo scopo principale nell'analisi di queste grandi quantità di informazioni genetiche umane è lo studio delle relazioni tra varianti genetiche e malattie, predisposizioni a malattie e caratteristiche fisiche in genere. La descrizione delle varianti di un individuo rappresenta il suo genotipo, mentre le sue caratteristiche fisiche, comprese possibili malattie, il suo fenotipo. Gli studi di associazione sul genoma si pongono come obiettivo la scoperta di queste relazioni mediante analisi statistiche e algoritmi di machine learning.

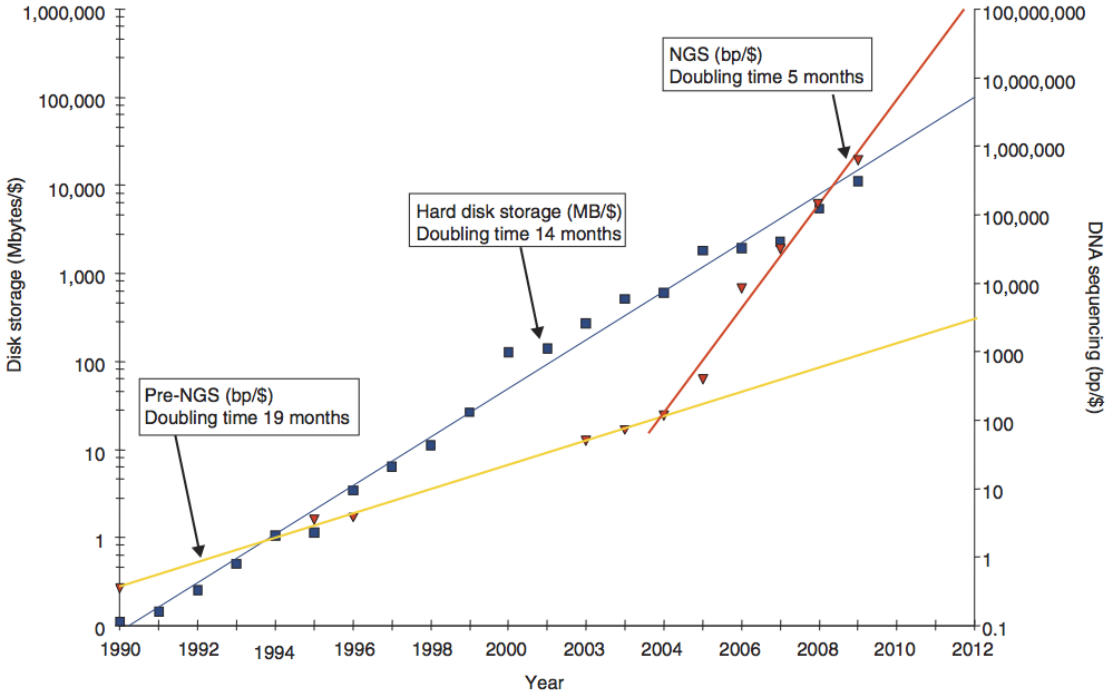


Figura 1.1: ( Rappresentazione grafica dei costi di sequenziamento e di storage dal 1990 al 2012 - immagine da Stein Genome Biology 2010 **11**:207, su licenza Creative Commons, doi:10.1186/gb-2010-11-5-207)

Al fine di condurre questi studi è necessario integrare informazioni sul fenotipo, che quindi riguardano le caratteristiche fisiche dell'individuo, ai dati genetici del soggetto. L'integrazione di questi dati non soltanto ha permesso, su larga scala, di individuare le responsabilità di specifiche varianti genetiche, ma ha aperto la strada verso la cosiddetta Medicina Personalizzata, nella quale i farmaci e le terapie applicate siano scelti su misura per ogni individuo. La mancata digitalizzazione delle cartelle cliniche o, laddove questo avvenga, l'assenza di standard per le descrizioni cliniche, rende il processo d'integrazione impossibile nella pratica.

Riassumendo, le difficoltà che ostacolano gli studi genomici *in silico*[1] sono:

[1]La locuzione latina, letteralmente nel silicio, si è diffusa di recente per indicare simulazioni matematiche al computer di entità o fenomeni chimici e biologici, in contrapposizione a *in vitro* e *in vivo*.

- Assenza di standard nella rappresentazione dei dati biologici, memorizzati in file con formati incompatibili tra loro.
- Quantità dei dati genomici in crescita esponenziale, per cui strumenti software e infrastrutture hardware tradizionali non sono sufficienti.
- Mancata di digitalizzazione e problemi di ambiguità nei termini per le descrizioni cliniche.

Questa problematica ha guidato i principi di progettazione e lo studio dei requisiti del sistema proposto.

Lo scopo dello studio svolto sui dati genetici, sugli standard nei quali vengono memorizzati e sui modelli di dati finora realizzati per rendere questi dati fruibili e pronti per l'analisi, è quello di definire i requisiti per la progettazione di un modello di dati ad hoc che possa supportare dati e metadati finora prodotti.

Questi requisiti dipendono sia dagli standard dei dati attualmente esistenti, in quanto rappresentano il modello per l'origine dei dati, sia dalle necessità successive di analisi. Quest'ultimo punto è di fondamentale importanza, perché da questo dipenderà l'efficienza degli algoritmi di analisi e l'utilizzo appropriato dello spazio di storage.

In questa prospettiva si prenderanno in considerazione le tecnologie Big Data, sia per le soluzioni in termini di spazio e scalabilità, sia per l'analisi distribuita su più nodi.

Nel predisporre l'integrazione dei dati clinici sarà indispensabile considerare ontologie di dominio medico e biologico, come base per la modellazione di un sistema per cartelle cliniche elettroniche.

### Organizzazione di questa tesi

Nel capitolo 2 viene descritto passo per passo il processo che va dal sequenziamento di un campione di DNA all'associazione tra varianti genetiche e malattie.

Nel capitolo 3 si analizzano nel dettaglio i formati standard e non standard dei file utilizzati in bioinformatica nei vari passi del processo, con lo scopo di astrarne un modello di dati unico.

Nel capitolo 4 si passano in rassegna i database genetici più noti, derivandone il modello relazionale da interfacce di sottomissione, documentazioni e fonti varie.

Nel capitolo 5 viene proposto un modello matematico dei dati biologici, insieme alle funzioni alla base dei processi di analisi, per dare una visione d'insieme formale dei temi trattati.

Nel capitolo 6 vengono descritti in dettaglio gli studi di associazione tra varianti genetiche e caratteristiche fenotipiche[2], ponendo le basi di progettazione per l'integrazione di dati genetici con informazioni cliniche.

Nel capitolo 7 viene progettato un database e un'infrastruttura a partire dai requisiti che emergono dai precedenti capitoli.

Nel capitolo 8 si pone l'accento sui requisiti non funzionali del sistema software appena proposto, basato su un dataset che può essere considerato Big Data, e in quanto tale richiede tecnologie e approcci di programmazione mirati alla scalabilità alle capacità computazionali.

[2]Le caratteristiche fisiche osservabili di un individuo, tra cui rientrano patologie e disfunzioni.

# 2 | Analisi computazionale su sequenze

Prima di addentrarsi nello studio dei workflow[1] di analisi genetica è bene dire che il processo di elaborazione digitale, che porta dalla lettura grezza delle sequenze di DNA all'identificazione di varianti genetiche, è un processo molto complesso e variabile. La complessità del workflow è dovuta principalmente all'eventualità di possibili errori nelle fasi di lettura e di allineamento, che avvengono con una probabilità maggiore nei nuovi strumenti NGS rispetto ai sequenziatori pre-NGS.

La continua evoluzione dei metodi e delle tecnologie hardware e software del settore, inoltre, impone frequenti cambiamenti nei formati di dati standard e negli strumenti software che filtrano e analizzano le sequenze. Lo sviluppo di un workflow ben definito e automatizzato per l'analisi dei dati genetici è diventato di fondamentale importanza negli ultimi anni [Koboldt et al., 2010].

In figura 2.1 è illustrata, in un modo generico e senza entrare nei dettagli, in quanto sono molto variabili, una pipeline tipica per l'analisi delle sequenze esoniche. In verde sono mostrati i processi e in beige i dati di output/input. Se si considera il diagramma come uno modello layer a strati, i dati di output generati da uno strato rappresentano i dati di input per lo strato sottostante.

Approfondirò di seguito ogni stadio dell'analisi, senza entrare troppo in dettagli specifici dipendenti dalla piattaforma o da particolari necessità di analisi.

[1] Per workflow o pipeline intendiamo una sequenza ordinata di attività, in cui l'input di un'attività è rappresentato dall'output dell'attività precedente

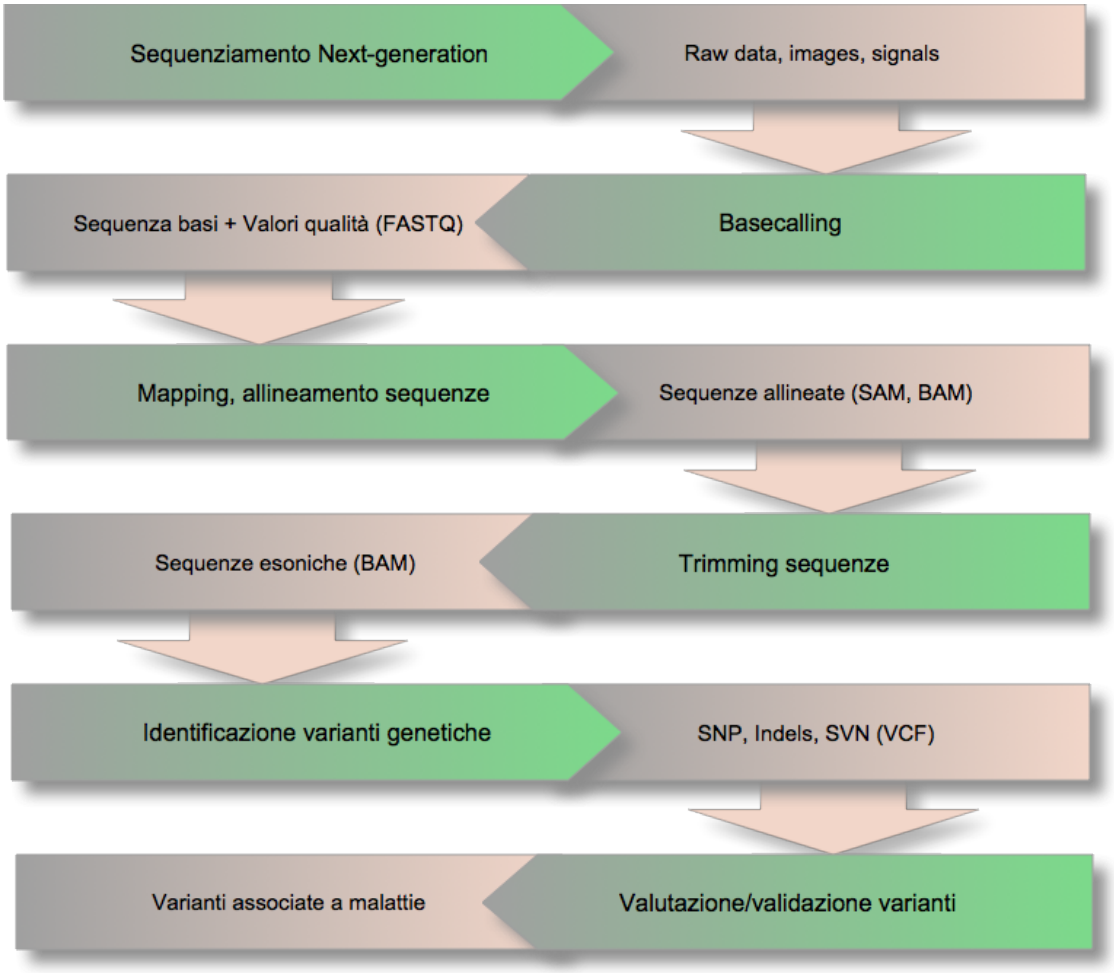


Figura 2.1: Workflow del sequenziamento esomi.

## 2.1 Sequenziamento Next-generation

Il NGS (Next-generation sequencing) è un nome scelto per indicare tutte quelle piattaforme di sequenziamento, e le relative tecnologie, nate dopo il 2005, che hanno rivoluzionato il processo di sequenziamento permettendo la parallelizzazione dello stesso, a beneficio di costi e prestazioni. La crescita esponenziale nella velocità di sequenziamento delle macchine NGS, capaci di generare molti milioni di sequenze lette per ogni esecuzione, ha spostato il collo di bottiglia dalla generazione delle sequenze alla gestione ed analisi dei dati.

La velocità di esecuzione del sequenziamento per queste tecnologie va però a discapito dell'accuratezza nel basecalling e nell'allineamento, svantaggio a cui si rimedia con letture ripetute e appositi processi computazionali nelle successive fasi di analisi.

### 2.1.1 Sequenziamento dell´esoma

Con la possibilità di sequenziare il patrimonio genetico completo di un individuo in pochi giorni, approcci WGS (*Whole Genome Sequencing*, sequenziamento dell'intero genoma) vengono progettati per scoprire variazioni genetiche che contribuiscono a malattie rare o comuni. Nonostante la diminuzione nei costi di sequenziamento questi approcci restano molto dispendiosi nella gestione e nell'analisi di un numero grande di campioni [Hedges et al., 2009]. Per questo motivo sono nati metodi alternativi, che si concentrano solo su frazioni dell'interno genoma, e rappresentano approcci convenienti per identificare le varianti genetiche potenzialmente associate alle patologie.

L'esoma è un termine derivato da Genoma, e sta ad indicare l'insieme di tutti gli esoni presenti nel genoma, ovvero di tutte le sottosequenze del DNA che possono codificare proteine. L'approccio WES (*Whole Exome Sequencing*, sequenziamento dell'intero esoma) è ritenuto una valida alternativa al WGS per diversi motivi:

- Gli esoni codificanti concorrono alla maggior parte delle variazioni funzionali [Botstein and Risch, 2003, Ng et al., 2008]
- L'esoma costituisce solo l'1% del genoma umano, richiedendo quindi il sequenziamento di sole 30 milioni di basi nucletodiche (Mbp) [Ng et al., 2009]
- La quantità totale di lavoro per il sequenziamento, l'analisi e la gestione dell'intero esoma, rispetto all'approccio WGS, è in rapporto 1/20 [Nielsen et al., 2010]
- I polimorfismi a singolo nucleotide (SNP) che appaiono nelle regioni codificanti sono la causa più comune per le malattie Mendeliane (le malattie dipendenti da un solo gene). [Horner et al., 2010]

## 2.2 Basecalling

Il base calling è un processo ad opera dello strumento di sequenziamento NGS, che associa ad ogni nucleotide letto un valore di probabilità per ogni base azotata. Spesso la stessa sequenza viene letta più volte per ovviare all'inaccuratezza delle letture e a valori di probabilità non soddisfacenti. Il formato dei dati di output più diffuso tra le piattaforme NGS è il FASTQ, un formato testuale di cui esistono diverse versioni. La versione Sanger FASTQ è lo standard de facto, ed è il formato accettato dall'NCBI Sequence Read Archive per l'invio dei dati raccolti.

## 2.3 Allineamento sequenze

Il primo processo bioinformatico nell'analisi delle sequenze di DNA è quello di allineamento.

L'allineamento è il processo di mappatura tra le sequenze lette e la sequenza di un genoma di riferimento. Molti dei software di allineamento disponibili oggi sono basati su due algoritmi principali: il metodo hashed-based e il metodo Burrows-Wheeler Transform [Li and Durbin, 2009].

Gli algoritmi hash-based usano una tabella di hash, costruita a partire dal genoma di riferimento o dalle sequenze lette, per mappare l'insieme delle sequenze lette nelle relative posizioni del genoma.

Algoritmi più recenti si basano invece sullo string matching usando la trasformata di Burrows-Wheeler (BWT). Gli algoritmi BWT riordinano la sequenza del genoma di riferimento raggruppando in una struttura dati le sequenze che appaiono più volte, viene poi creato un indice di riferimento e usato per un rapido piazzamento delle sequenze lette sul genoma di riferimento. Il vantaggio principale degli algoritmi BWT sta nella loro velocità, risultano infatti molto più veloci degli algoritmi hash-based.

La fase successiva è quella di assemblaggio, effettuata quasi sempre dagli stessi tool di allineamento. In questa fase le sequenze lette e allineate vengono composte tra loro a formare la sequenza genetica originale del campione in input. Anche

per questa fase viene utilizzata la sequenza di riferimento del genoma umano, tuttavia mediante appositi algoritmi di overlapping sarebbe possibile ricomporre la sequenza originale senza nessuna sequenza di riferimento. E' quello che avviene ad esempio nel sequenziamento di DNA per organismi di cui non si possiede il genoma.

La sequenza allineata viene memorizzata in formato SAM (Sequence Alignment/Map, vedere sezione 3.6) . Al termine della pipeline di allineamento il SAM viene convertito in BAM (SAM binario), un formato molto più compresso.

## 2.4 Trimming sequenze

La fase successiva a quella di allineamento è la fase di raffinamento della sequenza ottenuta. Esistono infatti diverse problematiche scaturite da un sequenziamento con macchina NGS, principalmente quelle dovute all'allineamento di sequenze molto corte. Per esempio, poichè ogni sequenza corta è allineata indipendentemente, sequenze di Indel (mutazioni o ricombinazioni che risultano nell'inserimento di un codone in più) possono non essere allineate con la sequenza originale. Prima di procedere a successive analisi, quindi, vengono effettuati dei processi di controllo e miglioramento della qualità delle letture.

In questa fase inoltre, nel caso in esame, si procede all'estrazione delle sole sequenze esoniche, che rappresentano una piccola frazione dell'intero genoma, ma che si crede codifichino la maggior parte delle variazioni funzionali ([Botstein and Risch, 2003], [Ng et al., 2010]). Poichè il sequenziamento può catturare anche frammenti di DNA originati da regioni non codificanti, viene applicato un filtro utilizzando una lista di posizioni conosciute per le sequenze esoniche, escludendo ogni lettura che non si sovrappone a queste posizioni.

## 2.5 Identificazione e valutazione di varianti genetiche

Per facilitare la ricerca di cause recessive o dominanti, le varianti vengono divise tra mutazioni omozigote e mutazioni eterozigote. Nell'identificazione delle varianti entrano in gioco tutte le variabili finora ottenute - valori di probabilità dei base call ricalibrati, mappatura delle varianti, letture esoniche sovrapposte - al fine di evitare possibili falsi positivi.

Il tool più utilizzato al momento per l'identificazione e l'analisi delle varianti è fornito nel software GATK (Genome Analysis Toolkit) [GATK, 2013, 2013], utilizzato anche nel progetto *1000 Genome Project* e nel *The Cancer Genome Browser* [Broad Institute, 2010, ]. Oltre a un insieme di tool, il GATK è in realtà anche una struttura di API Java, progettata per lo sviluppo di tool di analisi delle sequenze già allineate. Questo software è attualmente utilizzato anche in grandi progetti di sequenziamento come 1000 Genomes Project e The Cancer Genome Atlas.

Nella valutazione delle varianti identificate è di fondamentale importanza basarsi sulle varianti pre-filtrate, rese accessibili dai progetti HAPMap Project e dbSNP, come training data per il clustering dei dati da analizzare.

Il formato di file per la memorizzazione delle varianti è il variant call format (VCF), un formato testuale che può memorizzare SNP, DIP e varianti strutturali più lunghe. Il formato prevede anche uno standard per l'annotazione di informazioni come genotipo, allele ancestrale, profondità di lettura, qualità della mappatura, etc.

## 2.6 Identificare i geni associati alle malattie

L'approccio bioinformatico alle malattie è un approccio riduzionista: si tenta di individuare i geni e i loro prodotti che causano una malattia. Una volta identificati questi geni, la sfida resta quella di scoprire in che modo le mutazioni individuate causano la malattia, cercando quindi di collegare il genotipo al fenotipo.

### 2.6.1 Classificazione per causa

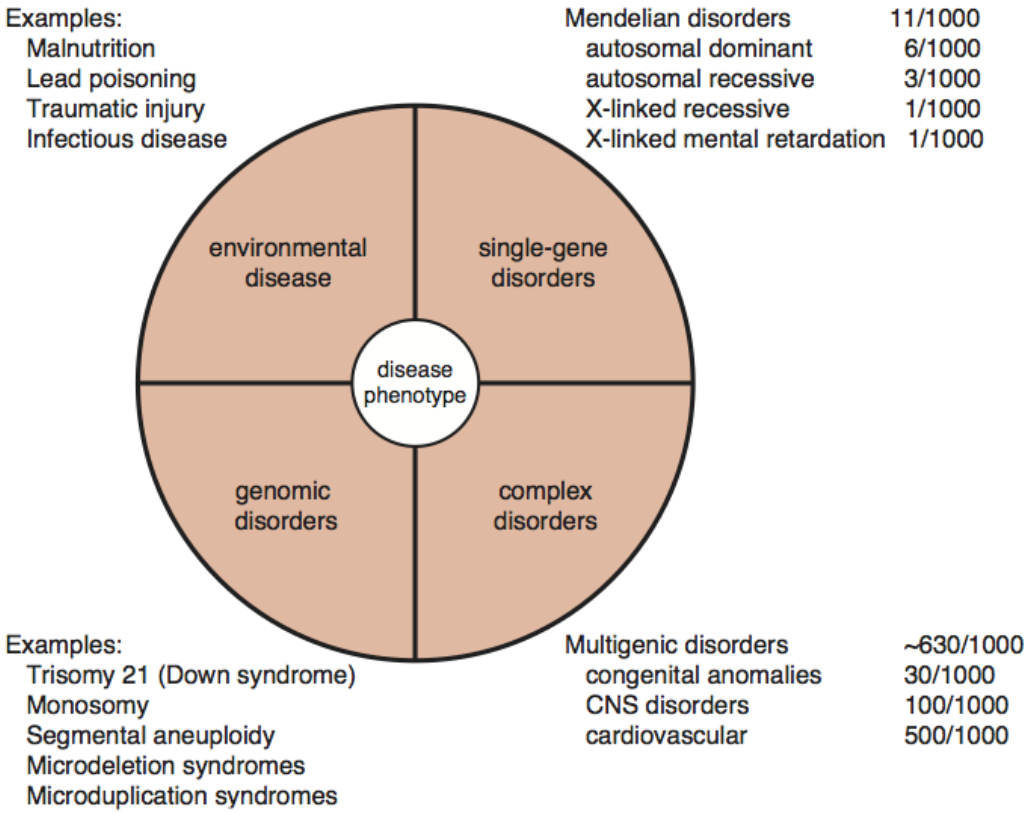


Figura 2.2: Cause genetiche e incidenza delle malattie.

Le malattie umane possono essere suddivise in base alla causa: disordini a singolo gene (mutazioni in un singolo gene); disordini complessi (derivati dalla mutazione in due o più geni, come il cancro e la schizofrenia); disordini genomici (quelli derivanti da anormalità cromosomiche); e malattie ambientali (come le malattie infettive). In figura 2.2 sono mostrati i valori di incidenza di queste classi di disordini. Nel complesso, i disordini multigenici sono molto più comuni dei disordini a gene singolo.

Per ovvi motivi è più semplice individuare le cause genetiche che portano alle malattie a singolo gene, infatti allo stato attuale le malattie monogeniche conosciute sono già state mappate nel genoma. I disordini genomici sono relativamente

facili da individuare e sono abbastanza comuni. Solitamente comportano delle anormalità di larga scala per le quali è presente un cromosoma in più (trisomia) o in meno (monosomia), come nella sindrome di Down (trisomia 21). In tutte le categorie, comunque, le cause e lo sviluppo di una malattia derivano dall'influenza di fattori sia genetici che ambientali.

Negli ultimi anni l'interesse della ricerca è indirizzato ai disordini complessi, come l'asma o la schizofrenia, dove i polimorfismi singoli hanno una debole influenza nel loro sviluppo. A causa della complessità del processo che porta a queste malattie, si utilizzano metodi stocastici di inferenza probabilistica per individuare la posizione del geni coinvolti in una malattia.

### 2.6.2 Analisi dei collegamenti

Il Linkage Analysis è una tecnica usata principalmente per la localizzazione di malattie ereditarie a singolo gene. Il collegamento genetico (genetic linkage) è la tendenza dei geni vicini ad essere ereditati insieme durante la meiosi, evitando la separazione su cromatidi diversi durante la ricombinazione cromosomica. Negli studi di linkage si analizzano le regioni di cromosoma ereditate all'interno delle famiglie. Seguendo il pattern di trasmissione in un albero genealogico, l'analisi dei collegamenti può essere usata per localizzare il gene di una malattia. La malattia di Huntington, un disordine neurodegenerativo, è stato il primo disordine ad essere mappato usando l'analisi dei collegamenti.

### 2.6.3 Studi di associazione sull'intero genoma

Mentre le basi genetiche di migliaia di malattie a singolo gene sono state già trovate, è molto più difficile identificare le cause genetiche di malattie umane comuni che coinvolgono più geni. Gli studi di associazione GWAS (*Genome Wide Association Studies*, studi di associazione sull'intero genoma) forniscono un approccio importante per raggiungere questo scopo. Con questo approccio vengono analizzati i markers (SNP, DIP) in individui affetti dalla malattia e individui sani, per

identificare le differenze nella frequenza di variazioni. I risultati saranno tanto più attendibili quanto più sarà grande il campione analizzato.

## 2.7 Diagnosticare malattie a partire dal genoma

Come già accennato, per molte malattie sono stati identificati i geni responsabili e il loro locus all'interno del genoma. I dati raccolti durante gli anni sono stati organizzati in diversi database, fruibili gratuitamente online, tra questi i più conosciuti sono OMIM [Hamosh et al., 2005], GeneCards [Safran et al., 2003], Cardiff, DMDM [Peterson et al., 2010].

Il progetto di più grande importanza in questo campo è OMIM (Online Mendelian Inheritance in Man). OMIM è una fonte autorevole e molto vasta che contiene informazioni riguardanti tutte le malattie mendeliane, ovvero ereditarie, collegate a più di 12000 geni. Nato negli anni '60 come catalogo delle malattie ereditarie, OMIM è su web dal 1995 a cura del NCBI ed è aggiornato regolarmente ogni giorno. Ogni voce contiene molti link a risorse genetiche esterne, come i database genomici o proteomici.

GeneCards è un database incentrato sui geni, che collega questi ai loro prodotti (proteine) e al loro coinvolgimento nelle malattie.

Cardiff Human Gene Mutation Database è un altro database che mette in relazione collezioni di mutazioni genetiche a malattie umane ereditarie.

Il database DMDM (Domain Mapping of Disease Mutations) [Peterson et al., 2010] fornisce, per ogni dominio di proteine, le mutazioni e gli SNP che codificano proteine coinvolte nella nascita o nello sviluppo di malattie. Il database fa uso di dati raccolti da altri progetti già citati, come OMIM, Swiss-Prot, RefSeq, e i modelli di dominio NCBI CDD.

# 3 | Formati dei File Genomici Standard

*"Orsù, scendiamo laggiù e confondiamo la loro lingua, affinchè l'uno non comprenda più il parlare dell'altro*

— Genesi 11,7, (La Torre di Babele)

La gran parte dei formati storici per le sequenze nucleotidiche sono formati flat file. Un flat file è un file contenente dei record privi di relazioni strutturali. Per interpretare un flat file è necessario conoscere le proprietà di formattazione del file.

I flat file per i dati genetici sono flat file ASCII delimitati, ovvero possono contenere record di ampiezza variabile, delimitati da apposite stringhe (codici). Ogni codice indica la tipologia di informazione contenuta dopo il codice. Ad esempio nel formato flat file per sequenze nucleotidiche dell'EBML l'inizio della sequenza vera e propria è delimitato dal codice SQ.

Utilizzando un file per ogni sequenza, il flat file contiene tutte le informazioni di riferimento di quella sequenza, come descrizione, parole chiave, referenze bibliografiche relative al lavoro dal quale sono stati estratti i dati riportati, posizione della sequenza nel genoma, ecc.

Uno dei maggiori problemi in bioinformatica riguarda il dover trattare una profusione di formati di file, spesso con standard scarsamente definiti o ambigui. Alcuni di questi formati flat file, costruiti ad hoc per le esigenze del momento, sono diventati poi col tempo standard *de facto*.

Nelle sezioni che seguono si analizzeranno in dettaglio i formati più comuni per i dati di sequenziamento, con l'intenzione ultima di progettare un modello di dati comune che possa supportare le informazioni finora memorizzate in file.

## 3.1 FASTA

Il formato FASTA è un formato per sequenze di DNA e di amminoacidi. E' stato originariamente inventato da Bill Pearson come formato di input per i tool della suite FASTA, di sua produzione [Pearson and Lipman, 1988].

Nella figura 3.1 si mostra un esempio di file FASTA contenente una sequenza di amminoacidi:

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATL
PAEKMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEE
KYNLTSVLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPE
SEQFRADHPFLFLIKHNPTNTIVYFGRYWSPMKILELPFASGDLSMLVLLMKILELPFASGDLSMLV
```

Figura 3.1: Standard FASTA per una sequenza di amminoacidi.

Un file in formato FASTA comincia con una riga singola di descrizione della sequenza. Le righe di descrizione (*defline*) si distinguono dalle righe di sequenza per il simbolo di maggiore (>) all'inizio della riga. Nella definizione dello standard si raccomanda di non utilizzare più di 80 caratteri per *defline*.

La riga di descrizione solitamente contiene il nome o un identificatore univoco per la sequenza, insieme a svariate altre informazioni. La struttura di questa intestazione e la tipologia di informazioni in essa contenute non sono standardizzate, ma i tanti database di sequenze hanno imposto ognuno il suo standard di intestazione FASTA. Una convenzione comune a questi standard di intestazione è che la riga di descrizione debba identificare univocamente la sequenza successiva, mediante una serie di attributi, come numeri progressivi e database di provenienza, separati dal carattere . Nella tabella 3.1 si riassume la sintassi delle linee di descrizione [Madden, 2003] per i principali database pubblici.

| Database pubblico | Sintassi identificatore |
|---|---|
| GenBank | gb\|accession\|locus |
| EMBL Data Library | emb\|accession\|locus |
| DDBJ, DNA Database of Japan | dbj\|accession\|locus |
| NBRF PIR | pir\|\|entry |
| Protein Research Foundation | prf\|\|name |
| SWISS-PROT | sp\|accession\|entry name |
| Brookhaven Protein Data Bank | pdb\|entry\|chain |
| Patents | pat\|country\|number |
| GenInfo Backbone Id | bbs\|number |
| General database identifier | gnl\|database\|identifier |
| NCBI Reference Sequence | ref\|accession\|locus |

Tabella 3.1: Tabella riassuntiva della sintassi per gli identificatori nel formato FASTA

## 3.2 FASTQ

FASTQ è diventato negli anni un formato molto comune per la condivisione di dati genetici di sequenziamento, combinando sia la sequenza di basi che un *quality score* associato ad ogni base nucleica, ovvero un punteggio di attendibilità sulla lettura di quella base all'interno della sequenza. Il formato FASTQ nasce come un'estensione del formato FASTA, aggiungendo le informazioni sull'attendibilità della lettura, rappresentando così la sequenza di sequenziamento con un livello di dettaglio maggiore, ma senza pesare sulla dimensione dei dati, come invece accadrebbe considerando lo spazio dei colori o l'immagine di cattura per le letture NGS.

Grazie all'estrema semplicità del formato, il FASTQ è ampiamente utilizzato per l'interscambio di dati, tuttavia, anche se nato come evoluzione del FASTA, continua anch'esso a soffrire dell'assenza di una definizione chiara e non ambigua, mancanza che ha portato all'esistenza di molte varianti incompatibili tra loro. Nello stato attuale, infatti, ogni macchina di sequenziamento NGS produce in

output un formato FASTQ differente.

### 3.2.1 FASTQ SANGER

La variante originale del formato FASTQ è quella Sanger. Il formato FASTQ è stato inventato nel 2000 nel Wellcome Trust Sanger Institute da Jim Mullikin, poi gradualmente diffuso negli anni successivi, ma senza ma essere formalmente documentato. Ciò che più si assomiglia ad una descrizione ufficiale del formato da parte di Sanger si può trovare sul sito web di MAQ/BWA [Li et al., 2008, Li and Durbin, 2009], ma perfino quest'ultima definizione risulta incompleta. In figura 3.2 si mostra una lettura in formato FASTQ dal database NCBI SRA.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGGGTTTTG
AATTTCAAACCCTTTTCGGTTTCCAACCTTCCAAAGCAATGCCAAT
+SRR014849.1 EIXKN4201CFU84 length=93
+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA1EA5'9B
:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@ /=<?7=9<2A8==
```

Figura 3.2: Il formato FASTQ Sanger in un file di esempio, contenente sequenze e *quality score*.

La struttura di FASTQ può essere scomposta in 4 sezioni di testo:

1. Una linea di titolo per la sequenza, che funge da identificatore, delimitata dal simbolo @. Questa linea non ha un formato definito e non ha un limite di lunghezza, permettendo annotazioni arbitrarie e commenti.

2. La sequenza vera e propria, senza simbolo delimitatore iniziale oltre alla nuova linea. La sequenza di DNA è nel formato FASTA, e come nel formato FASTA non pone limiti sui caratteri utilizzabili, tuttavia ci si attiene alla nomenclatura IUPAC per l'espressione delle basi nucleotidiche e degli amminoacidi. Gli spazi vuoti come lo spazio o il tab non sono permessi.

3. Il carattere delimitatore + indica la fine della sequenza e l'inizio della linea per i quality score. Originariamente questa linea includeva anche la prima linea di descrizione, come mostrato nell'esempio NCBI precedente, ma più di frequente questa linea contiene soltanto il carattere di delimitazione.

4. L'ultima linea di testo è quella dei punteggi di qualità. Si utilizza un simbolo, corrispondente ad un punteggio di qualità, per ogni base nucleotidica. Di conseguenza la linea dei *quality score* dovrà essere di lunghezza uguale alla linea della sequenza di basi.

   Come si nota dall'esempio i valori di qualità sono caratteri ASCII stampabili (solitamente i caratteri ASCII compresi tra 33-126). Il valore numerico corrispondente si ottiene con un semplice mapping, considerando come lo zero il primo carattere dell'intervallo ASCII in considerazione (ad esempio l'ASCII 33). Il valore di errore vero e proprio, su scala logaritmica, è stimato invece in modo diverso per ogni variante FASTQ.

### 3.2.2 I quality score in FASTQ

Mentre per quanto riguarda la sequenza in sé, tutti i formati seguono lo standard Sanger, questo non vale per la rappresentazione dei valori di qualità. La stima del valore di qualità e la sua rappresentazione dipendono fortemente dalle scelte di ciascuna casa di produzione NGS.

| Nome versione | Sequenza ASCII | | Quality Score | |
|---|---|---|---|---|
| | Intervallo | Offset | Tipologia | Intervallo |
| Sanger standard | 33-126 | 33 | PHRED | da 0 a 93 |
| Solexa, Illumina <1.3 | 59-126 | 64 | Solexa | da -5 a 62 |
| Illumina 1.3+ | 64-126 | 64 | PHRED | da 0 a 62 |

Tabella 3.2: Le tre varianti FASTQ descritte, con le relative caratteristiche dei valori di qualità. Gli intervalli sono da considerarsi inclusivi.

### PHRED

Il software PHRED legge i file di lettura del DNA grezzi, contenenti il segnale o l'immagine rilevata, interpretando di volta in volta il dato catturato come una delle 4 basi [Ewing et al., 1998, Ewing and Green, 1998]. Facendo questo, assegna una probabilità a questa interpretazione, definita in termini della stimata probabilità di errore $P_e$:

$$Q_{\text{PHRED}} = -10 \times log_{10}(P_e) \tag{3.1}$$

Insieme a questo standard di valutazione per la qualità delle interpretazioni, PHRED ha introdotto un nuovo formato di file, conosciuto come il formato QUAL, per mantenere questi dati separati dalla sequenza vera e propria. Ad esempio, utilizzando sempre record NCBI SRA mostrato in precedenza, un file QUAL corrispondente può essere come il file in figura 3.3

```
>SRR014849.1 EIXKN4201CFU84 length=93
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 22 37 31 22 16 11
6 1 26 34 30 11 33 26 30 21 33 26 25 36 32 16 36 32
16 36 32 20 6 24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 23 32 26 30 24
33 26 35 31 14 28 27 30 22 28 24 27 17 32 23 28 28
```

Figura 3.3: File di esempio per il formato QUAL.

I valori di qualità PHRED rappresentano ormai uno standard *de facto* per questo genere di dati. Per esempio, la macchina di sequenziamento Roche 454 permette la conversione dell'output in file FASTA e QUAL. I valori PHRED sono utilizzati anche direttamente all'interno dei file, come nei file SAM, Staden Experiment, ACE e FASTQ.

### FASTQ SOLEXA

Nel 2004 la società NGS Solexa ha introdotto la sua specifica versione di FASTQ, incompatibile e indistinguibile dalla versione Sanger [Bennett, 2004]. Nonostante

il formato FASTQ originale memorizzi un valore di qualità per ogni base della sequenza, nella versione di Solexa vengono memorizzati i valori di probabilità di tutte le quattro basi nucleotidiche, per ogni base in sequenza. Inoltre per rappresentare con più precisione i valori bassi di qualità, è stata ridefinito il valore logaritmico di qualità [Bentley et al., 2008] in questo modo:

$$Q_{\text{Solexa}} = -10 \times log_{10}(\frac{P_e}{1 - P_e}) \tag{3.2}$$

Nonostante questa differenza nella stima dell'errore tra valori PHRED e valori Solexa, la conversione da un valore all'altro risulta facilmente calcolabile:

$$Q_{\text{PHRED}} = 10 \times log_{10}(10^{\text{Q}_{\text{Solexa}}/10} + 1) \tag{3.3}$$

$$Q_{\text{Solexa}} = 10 \times log_{10}(10^{\text{Q}_{\text{PHRED}}/10} - 1) \tag{3.4}$$

### 3.2.3 FASTQ ILLUMINA

Nonostante inizialmente le macchine NGS Illumina utilizzassero la variante FASTQ di Solexa, dalla versione 1.3 in poi si è optato per la rappresentazione con valori PHRED. Tuttavia, al posto di supportare a pieno la variante Solexa, Illumina ha introdotto un'altra versione di FASTQ, incompatibile con i valori PHRED originali ma intercambiabile con le prime versioni dei file Solexa FASTQ.

La versione FASTQ di Illumina 1.3+ codifica i valori PHRED in un intervallo di 64 valori, ovvero i valori PHRED da 0 a 62 inclusi (ASCII 64-126). Bisogna però considerare che i valori di qualità Illumina prodotti dalla macchina di sequenziamento sono in realtà compresi tra 0 e 40.

## 3.3 CSFASTA e CSFASTQ

Le macchine di sequenziamento ABI SOLiD lavorano sullo spazio dei colori e non sullo spazio delle sequenze [Pandey et al., 2008], portando ABI a introdurre una particolare versione del formato FASTA, il Color Space FASTA (CSFASTA), con un file associato QUAL contenente i valori di qualità. Un'altra versione, che include i valori di qualità nello stesso file, è il formato Color Space FASTQ (CSFASTQ).

## 3.4 GenBank

Il formato GenBank è lo standard di sottomissione per i record del database GenBank, il database NCBI per sequenze di DNA e proteine. In un record GenBank sono presenti, oltre alla sequenza vera e propria, informazioni a essa correlate: riferimenti bibliografici, informazioni sulla funzione della sequenza, posizione delle regioni codificanti, posizioni delle mutazioni. Queste informazioni sono organizzate in campi del flat file GenBank, ognuno dei quali comincia con l'identificatore del campo, in maiuscolo, seguito da un carattere di tabulazione e dal dato in questione. L'identificatore può essere espresso con il nome intero o con un'abbreviazione di due o tre lettere (ad esempio si può usare REFERENCE o RF). Nella figura 3.4 si illustra il formato per il file GenBank:

```
LOCUS       nome del locus, lunghezza e tipo della sequenza,
            classificazione dell'organismo, data di inserimento
DEFINITION  descrizione del record
ACCESSION   numero di ammissione al database originale
KEYWORDS    parole chiave per i riferimenti incrociati
SOURCE      organismo sorgente del DNA
ORGANISM    descrizione dell'organismo
REFERENCE   riferimenti bibliografici
COMMENT     funzione biologica o informazioni sul database
FEATURES    informazioni su particolari parti della sequenza,
            con indicazione della posizione o dell'intervallo in sequenza
    source      (intervallo di sequenza) organismo sorgente
    misc_signal (intervallo di sequenza) tipo di funzione o segnale
    mRNA        (intervallo di sequenza) mRNA
    CDS         (intervallo di sequenza) regione codificante proteine
    intron      (intervallo di sequenza) posizione di un introne
    mutation    (posizione nella sequenza) variazione nella sequenza per una mutazione
BASE COUNT  conteggio del numero di A,C,G,T e altri simboli all'interno della sequenza
ORIGIN      questo delimitatore indica l'inizio della sequenza vera e propria
    1 gattacagatt acagattaca gattacagatt acagattaca gattacagatt
    51 acagattaca gattacagatt acagattaca gattacagatt acagattaca
    101 gattacagatt acagattaca gattacagatt acagattaca gattacagatt
//          questo delimitatore indica la fine della sequenza
```

Figura 3.4: Formato dello standard GenBank per le sequenze nucleotidiche.

## 3.5 EMBL

I record e i flat file di sottomissione dei database *NCBI GenBank*, *EMBL EBI* e *DDBJ*, ovvero i database facenti parte del progetto sincronizzato *International Nucleotide Sequence Database Collaboration*, sono molto simili tra loro. I record DDBJ sono quasi completamente identici ai corrispettivi di GenBank. Il flat file di EMBL invece si distingue leggermente dal formato GenBank, come illustrato nelle specifiche dello standard in figura 3.5

```
ID  codice di identificazione per la sequenza all'interno del database
AC  numero di ammissione nel database sorgente
DT  data di inserimento e modifica
KW  parole chiave per i riferimenti incrociati
OS, OC  organismo di origine
RN, RP, RX, RA, RT, RL riferimenti bibliografici
DR  codice identificativo in altri database
CC  descrizione della funzione biologica
FH, FT  informazioni su posizioni o intervalli specifici della sequenza
        source
        misc_signal
        mRNA
        CDS
        intron
        mutation
SQ      conteggio del numero di A,C,G,T e altri simboli
gattacagatt acagattaca gattacagatt acagattaca gattacagatt acagattaca 60
gattacagatt acagattaca gattacagatt acagattaca gattacagatt acagattaca 120
.
.
.
```

Figura 3.5: Formato dello standard EMBL per le sequenze nucleotidiche.

## 3.6 SAM e BAM

Il formato **SAM** (Sequence Alignment/Map format) è il formato testuale standard per gli output intermedi e finali dei software di allineamento. L'utilizzo più comune che si fa di questi software è quello di allineare il segmento di DNA in input (solitamente un file FASTQ) su un genoma di riferimento. Il formato è un flat file delimitato da tabulazioni e dal carattere @ per le linee di intestazione. Se

presente, l'intestazione deve precedere i dati di allineamento. I dati di allineamento sono organizzati con un allineamento per ogni linea di testo. Per ogni linea di allineamento saranno definiti 11 campi obbligatori, in un ordine prestabilito e delimitati dal carattere di tabulazione. Se un particolare dato non è disponibile il campo non può essere lasciato vuoto, al suo posto viene inserito il carattere 0 oppure * (dipende dal campo).

La definizione della sintassi per i campi degli allineamenti è descritta in 3.3. La descrizione completa e il significato di ogni campo necessiterebbe un lungo approfondimento e, dato che questo esula dalla scopo di questa tesi, si rimanda alle specifiche ufficiali [Sam and Specification, 2011].

Il formato **BAM** è un formato di codifica per i file SAM, compresso nel formato BGZF (*Blocked GNU Zip Format*). BGZF è un formato di compressione a blocchi implementato sullo standard gzip. L'obiettivo di BGZF è quello di fornire una buona compressione insieme alla possibilità di accedere al file BAM in modo non sequenziale per eseguire interrogazioni indicizzate. Il formato BGZF è compatibile con gunzip, il che rende possibile l'estrazione di un file BGZF utilizzando un tool gzip. Per una descrizione completa del formato BAM si rimanda alle specifiche del formato [Sam and Specification, 2011].

| Ordine | Campo | Tipo | Regexp/Intervallo | Descrizione |
|---|---|---|---|---|
| 1 | QNAME | String | [!-?A-~]1,255 | Nome della lettura |
| 2 | FLAG | Int | [0,216 -1] | FLAG bitwise |
| 3 | RNAME | String | [!-()+-<>-~][!-~]* | Nome di rif. sequenza |
| 4 | POS | Int | [0,229 -1] | Posizione della sequenza |
| 5 | MAPQ | Int | [0,28 -1] | Qualità mapping (PHRED) |
| 6 | CIGAR | String | ([0-9]+[MIDNSHPX=])+ | stringa CIGAR estesa |
| 7 | RNEXT | String | [!-()+-<>-~][!-~]* | Nome di rif. prossima seq. |
| 8 | PNEXT | Int | [0,229 -1] | Posizione prossima sequenza |
| 9 | TLEN | Int | [-229 +1,229 -1] | Lunghezza sequenza inserita |
| 10 | SEQ | String | [A-Za-z=.]+ | Sequenza di basi |
| 11 | QUAL | String | [!-~]+ | ASCII della qualità PHRED |

Tabella 3.3: Specifiche per le linee di allineamento nel formato SAM, con: ordine del campo, nome, tipo di dato, definizione sintassi del campo mediante regular expression e intervallo di valori accettabili, descrizione del campo.

## 3.7 SCF

La genomica moderna produce una quantità di dati scientifici mai vista prima, arrivando a generare, con una sola macchina NGS e in pochi giorni, la quantità di basi generate dal progetto genoma umano in 10 anni. Come si nota dalle specifiche dei formati *trace file* finora visti, ovvero i file contenenti le letture sequenziali di DNA, i valori di qualità, e a volte i dati grezzi di cattura, poca attenzione viene prestata per le dimensioni finali dei file prodotti. Il problema della dimensione dei trace file è invece uno dei maggiori interessi per la comunità bioinformatica. Il primo reale tentativo nato da queste motivazioni è il formato SCF [Dear and Staden, 1992].

Il formato SCF contiene i dati per una lettura singola e include:

- La traccia dei punti di campionamento
- La sequenza di basi dedotta dai segnali campionati
- Le posizioni delle basi in sequenza nella traccia di campionamento

| Tipo di informazione | Percentuale occupata |
|---|---|
| Basi dedotte | 1% |
| Stima accuratezza base dedotta | 4% |
| Traccia delle ampiezze per ognuna delle 4 basi | 88% |
| Offset della base rispetto alla traccia | 4% |
| Commenti testuali vari (identificatori, date etc.) | 0.2% |

Tabella 3.4: La tabella mostra le percentuali sulla dimensione non compressa di un file SCF.

| Formato | Dimensione Originale | Dimensione Compressa | Rapporto |
|---|---|---|---|
| ABI | 18 158 424 | 8 427 773 | 0.464 |
| SCFv2 | 7 887 845 | 3881 662 | 0.492 |
| SCFv2 | 7 887 845 | 2 396 562 | 0.304 |

Tabella 3.5: Confronto tra dimensioni originali e compresse dei formati ABI, SCFv2 ed SCFv3

- La stima dell'accuratezza per la deduzione (calling) delle basi

Le specifiche di formato includono la codifica di questi dati in tipi e strutture dati C. Il risultato finale in termini di dimensioni del file è mostrato in percentuale nella tabella 3.4. Nella tabella 3.5 sono mostrate invece le differenze di dimensione tra i formati ABI, SCFv2 e SCFv3 per la stessa quantità di informazione genetica. Come si può notare dalla tabella versioni successive di SCF migliorano lo spazio utilizzato su disco con un approccio loss-less. E' stato determinato infatti che soltanto cambiando la disposizione del tipo di informazione all'interno del file si otteneva con gzip una compressione maggiore.

## 3.8 ZTR e SFR

Sebbene successivamente all'SCFv3 sono stati proposti formati più efficienti, soprattutto per la predisposizione alla compressione come nel caso del CTF[1], lo stesso gruppo di ricerca che sviluppò SCF ha proposto 10 anni dopo un nuovo formato, il ZTR.

La progettazione di ZTR prende in parte spunto dalle tecniche di compressione PNG, e si basa sui seguenti principi chiave:

- Estensibilità: non potendo prevedere facilmente quali dati sarà necessario memorizzare in un trace file in futuro, è necessario un meccanismo per incorporare nuove informazioni mantenendo valido il formato
- Dimensione ridotta: ridurre al minimo la dimensione dei dati è un vantaggio sia per lo spazio di storage sia per le infrastrutture di rete.
- Velocità:la velocità di codifica e decodifica non deve essere peggiore dello standard SCF.
- Specifiche pubbliche: le specifiche e il codice sorgente per la codifica devono essere aperti e resi pubblici.

Il formato SRF (Short Read Format) è uno standard basato su ZTR, strutturato come un container per sequenze ZTR e progettato per garantire una maggiore flessibilità sui dati contenuti. Nella struttura infatti è presente anche un blocco XML dove l'utente può memorizzare informazioni addizionali.

## 3.9 VCF e file di varianti

Il format VCF [Danecek et al., 2011] è un formato standardizzato generico per memorizzare la maggior parte delle varianti genetiche esistenti, tra cui SNP, indel

[1]CTF è un formato sviluppato da Jean Thierry-Mieg per il NCBI, che specifica il suo algoritmo di compressione proprietario, ottenendo risultati superiori a SCF. Non esistono pubblicazioni in merito.

e varianti strutturali, associate ad annotazioni libere. Il file VCF è formato da una sezione di header e da una sezione di dati. L'intestazione contiene un numero arbitrario di meta-dati, organizzati su più linee, e una linea di definizione per la struttura della sezione di dati. Ogni linea di meta-dati comincia con il delimitatore ##, mentre la linea di definizione della struttura comincia con il carattere #, e i campi definiti al suo interno sono delimitati da tabulazioni.

Le meta informazioni nell'intestazione possono essere usate per descrivere il mezzo con cui è stato creato il file, la data di creazione, la versione della sequenza di riferimento, i software usati e tutte le informazioni rilevanti sulla storia del file.

Nel codice di esempio 3.6 è illustrato un file VCF con varie meta informazioni nell'intestazione e 4 diverse varianti. Gli header obbligatori sono il ##fileformat e la linea di definizione dei campi #CHROM, le altre linee sono informazioni sul file e sulla sequenza da cui sono state estrapolate le varianti. Nel *body* del file è possibile osservare uno SNP alla terza riga: nel campo REF si indica la sequenza di riferimento che può variare, in questo caso un singolo nucleotide, nel campo ALT si mostrano invece le alternative nella sequenza analizzata.

```
Header  ##fileformat=VCFv4.1
        ##fileDate=20110413
        ##source=VCFtools
        ##reference=file:///refs/human_NCBI36.fasta
        ##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
        ##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
        ##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
        ##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
        ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
        ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
        ##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
        ##ALT=<ID=DEL,Description="Deletion">
        ##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
        ##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
        #CHROM POS ID     REF  ALT    QUAL FILTER  INFO                FORMAT     SAMPLE1  SAMPLE2
Body    1        1  .     ACG  A,AT    40  PASS    .                   GT:DP      1/1:13   2/2:29
        1        2  .     C    T,CT    .   PASS    H2;AA=T             GT         0|1      2/2
        1        5  rs12  A    G       67  PASS    .                   GT:DP      1|0:16   2/2:20
        X      100  .     T    <DEL>   .   PASS    SVTYPE=DEL;END=299  GT:GQ:DP   1:12:.   0/0:20:36
```

Figura 3.6: Esempio di un formato valido VCF [Danecek et al., 2011]

.

# 4 | I grandi Database biologici

*"Se ho visto più lontano, è perché stavo sulle spalle di giganti.*
— Isaac Newton, (da una lettera a Robert Hooke)

Nell'ultimo decennio, progetti di larga scala hanno generato un'enorme quantità di dati biologici a livello molecolare. Questi dati sono stati conservati e organizzati in diversi MBDB (*Molecular Biology Databases*, Database Biologici Molecolari). Questi database includono informazioni genomiche, proteomiche, trascrittomiche, metabolomiche e negli ultimi casi di ricerca integrata, informazioni cliniche. Esistono al momento più di mille MBDB disponibili online al pubblico, alcuni ad accesso libero altri commerciali. Questi database sono divenuti col tempo il fulcro della ricerca scientifica di settore, permettendo la correlazione di dati eterogenei e assistendo i ricercatori nei loro studi. Questi database hanno un'incredibile varietà di dati, e permettono agli utenti di trovare facilmente correlazioni complesse, ad esempio quale gene codifica per un enzima che svolge una particolare funzione nell'organismo specifico, o quali sono gli enzimi che partecipano in un pathway metabolico e qual'è la loro struttura tridimensionale.

## 4.1 I grandi portali bioinformatici

L'**NCBI** (*National Center for Biotechnology Information*, Centro Nazionale per le Informazioni Biotecnologiche), mantiene 31 database differenti. Al centro di

questi database si trova Entrez, un database retrieval system che permette la ricerca testuale e le query booleane.

Entrez permette in un attimo di cercare tra tutti i database dell'NCBI, ritornando il numero di record trovati per ogni database, incluse le sequenze di DNA e di proteine, rispettivamente dei database GenBank e Proteins, tassonomie, genomi, data set di popolazioni, struttura delle proteine, letteratura biomedica PubMed, e molte altre tipologie di risultato.

La feature My NBCI di Entrez permette inoltre di memorizzare le ricerche e le preferenze di un utente, e può inviare automaticamente un email all'utente con l'aggiornamento delle ricerche salvate.

L'**EMBL-EBI** (*European Molecular Biology Laboratory-European Bioinformatics Institute*) ospita più di 50 database e 50 strumenti bioinformatici di analisi. I database EMBL-EBI includono sequenze di DNA e proteine (rispettivamente EMBL e UniProt), strutture delle proteine, geni, interazioni molecolari, tools per l'allineamento, letteratura e molto altro. Il server EMBL-EBI offre agli utenti diversi browser, EB-eye permette ricerche in tutti i database, i risultati potranno poi essere estratti con EBI Dbfetch. Anche EMBL-EBI mette a disposizione numerosi tool di analisi per sequenze, strutture proteiche e geni.

Esistono numerosi database pubblici oltre a quelli appena descritti, commerciali e non, messi a disposizione da altri istituti di ricerca. In questo capitolo verranno analizzati soltanto i database che si vuole integrare nel sistema in progettazione, o per i quali si vuole garantire interoperabilità. Un elenco esaustivo di database genomici è stato compilato in tabella 4.1

## 4.2 Database di sequenze di nucleotidi

**NCBI GenBank**, **EMBL Nucleotide Sequence Database** (EMBL) e il **DNA Data Bank of Japan** (DDBJ) sono database, i più grandi al mondo per la loro tipologia, che contengono sequenze nucleotidiche disponibili liberamente al pubblico. Le tre organizzazioni sincronizzano i loro dati su una base giornaliera per assicurare copertura mondiale, per perseguire un grande progetto di collaborazione,

| Nome | Tipologia | Link |
|---|---|---|
| Nucleic Acid Research Journal | Database letteratura | http://nar.oxfordjournals.org/ |
| PubMed | Database letteratura | http://www.ncbi.nlm.nih.gov |
| ISI Web of KnowledgeSM | Database letteratura | http://www.ncbi.nlm.nih.gov/ |
| NCBI | Portale per banche dati e strumenti | http://www.ncbi.nlm.nih.gov/ |
| EMBL-EBI | Portale per banche dati e strumenti | http://www.ebi.ac.uk/ |
| Entrez Nucleotide | Database di sequenze nucleotidiche | http://www.ncbi.nlm.nih.gov |
| EMBL Nucleotide Sequence Database | Database di sequenze nucleotidiche | http://www.ebi.ac.uk |
| DNA Data Bank of Japan | Database di sequenze nucleotidiche | http://www.ddbj.nig.ac.jp/ |
| Entrez Protein | Database sequenze proteiche | http://www.ncbi.nlm.nih.gov |
| UniProt | Database sequenze proteiche | http://beta.uniprot.org |
| SwissProt | Database sequenze proteiche | http://ca.expasy.org |
| KEGG | Database metabolomici | http://www.genome.jp |
| EcoCyc | Database metabolomici | http://ecocyc.org/ |
| GeneOntology | Ontologia | http://www.geneontology.org/ |
| HPO | Ontologia | www.human-phenotype-ontology.org/ |
| PATO | Ontologia | obofoundry.org/wiki/index.php/PATO:Main_Page |
| FMA | Ontologia | http://sig.biostr.washington.edu/projects/fm/AboutFM.html |
| OMIM | Database di associazioni | www.omim.org/ |
| GeneCards | Database di associazioni | www.genecards.org/ |
| DMDM | Database di associazioni | bioinf.umbc.edu/dmdm/ |

Tabella 4.1: Tabella riassuntiva dei database biologici con descrizione e URL

l'International Nucleotide Sequence Database Collaboration. La collaborazione tra queste tre istituzioni, l'americana NCBI, l'europea EMBL e la giapponese DDBJ, ha portato a grossi benefici nella comunità scientifica.

Le sequenze contenute non sono solo umane ma per diversi organismi, ogni record di sequenza contiene una descrizione, il nome scientifico dell'organismo sorgente e riferimenti bibliografici. I record non possono essere aggiornati o corretti senza il consenso del submitter.

I record di GenBank, DDBJ ed EMBL includono le sequenze per ogni gene, i WGS (Whole Genome Shotgun), RNA, sequenze sintetiche (sequenze di DNA create artificialmente) e sequenze ambientali (sequenze di DNA raccolta dall'ambiente il cui organismo di appartenenza è ancora sconosciuto). Per la sua completezza e per il suo ruolo primario e centrale, l'insieme GenBank/EMBL/DDBJ è il fulcro della maggior parte degli MBDB.

### 4.2.1 Schema Relazionale GenBank

Una definizione relazionale del database GenBank non esiste. Per ricostruire una versione relazionale della banca dati GenBank, è stato necessario dedurne la strut-

tura dal formato di sottomissione e esportazione dei record. In figura 4.1 è illustrato il modello relazionale per GenBank. Per assegnare i nomi a campi ed entità è stata utilizzata la stessa terminologia delle descrizioni del file di sottomissione GenBank, si rimanda quindi alla sezione 3.4 per una descrizione dettagliata dei campi.

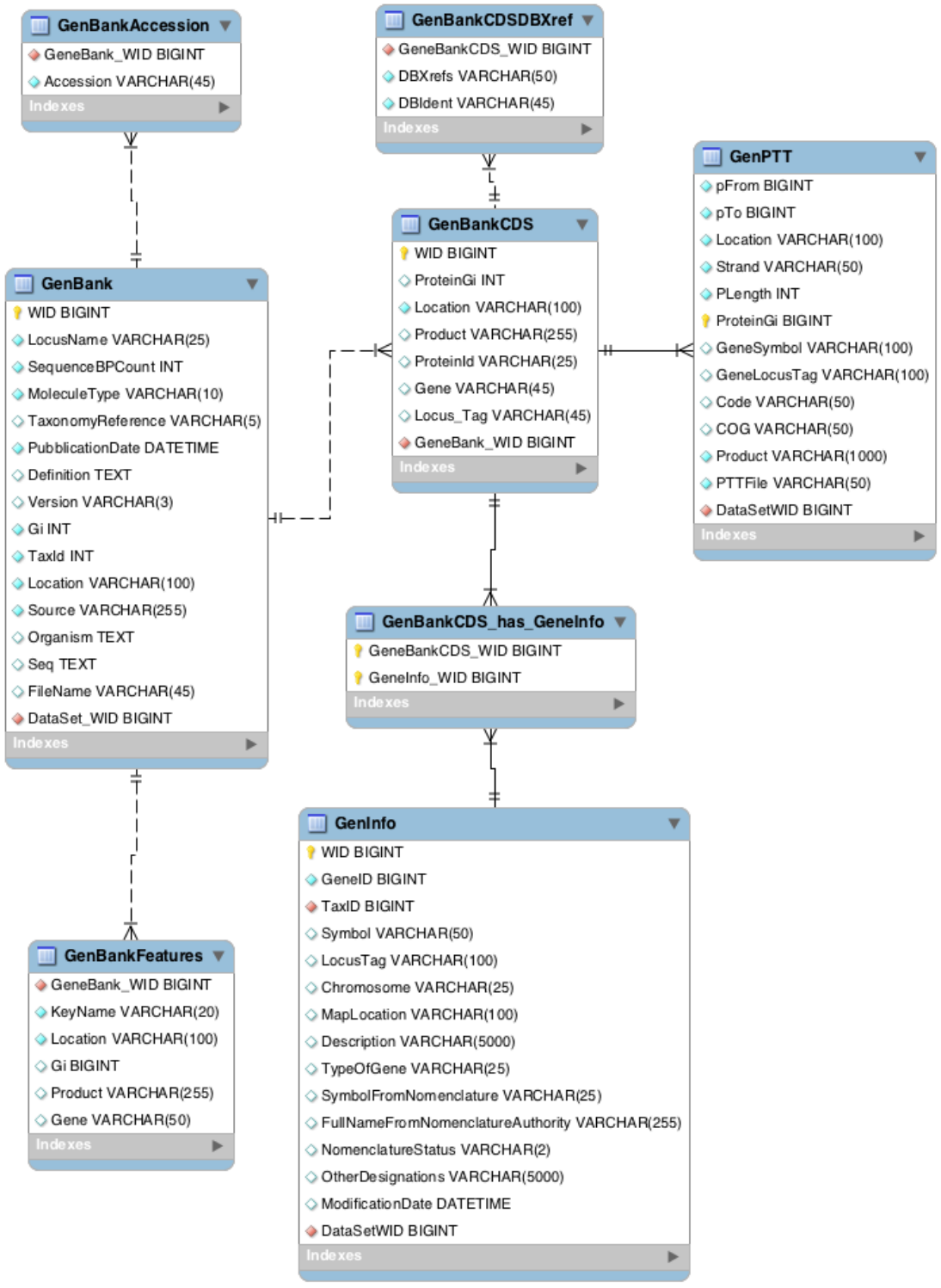


Figura 4.1: Schema relazionale del database di sequenze nucleotidiche GenBank

## 4.3 Database di ontologie

I database ontologici medici sono fondamentali per integrare le caratteristiche fenotipiche ai dati genetici in modo non ambiguo. L'ontologia studiata e integrata nel sistema in progettazione è lo Human Phenotype Ontology, un progetto inizialmente sviluppato utilizzando le informazioni del database OMIM (*Online Mendelian Inheritance in Man*). La scelta di questa ontologia, oltre ad evitare ambiguità nell'uso dei termini per la descrizione fenotipica, permette l'interoperabilità come molti database biologici, poichè considerata l'ontologia fenotipica di riferimento.

Il database è disponibile per il download sul sito del progetto [HPO, 2013, ] in formato OWL (Web Ontology Language). OWL è una famiglia di linguaggi per la rappresentazione di ontologie e basi di conoscenza. Attraverso alcune semplici trasformazioni è possibile riportare la struttura di un'ontologia OWL su uno schema SQL [Irina Astrova, 2007].

### 4.3.1 Schema Relazionale HPO

In figura 4.2 è illustrato il modello relazionale per l'ontologia HPO. L'unica entità è quella della classe fenotipica, che contiene:

- L'identificatore univoco su HPO.
- Il termine univoco che descrive la classe fenotipica.
- Una descrizione più dettagliata del fenotipo.
- I possibili sinonimi per il termine della classe.
- Il codice completo della classe in HPO, composto dalla stringa HP: concatenata all'identificativo numerico.
- Chiavi esterne per riferimenti a record di altre ontologie, quali FMA (*Foundational Model of Anatomy Ontology*), un'ontologia per l'anatomia umana, e PATO (*Phenotypic Quality Ontology*). Queste ontologie fanno tutte parte della fondazione OBO (*Open Biomedical Ontologies*)

Le relazioni sono tutte relazioni *n:n* tra la classe del fenotipo e se stessa, ovvero mettono in relazione le classi fenotipiche tra di loro. Una classe fenotipica può infatti avere più superclassi, più classi alternativi, più sintomi, dove per sintomi si intendono altre classi fenotipiche.

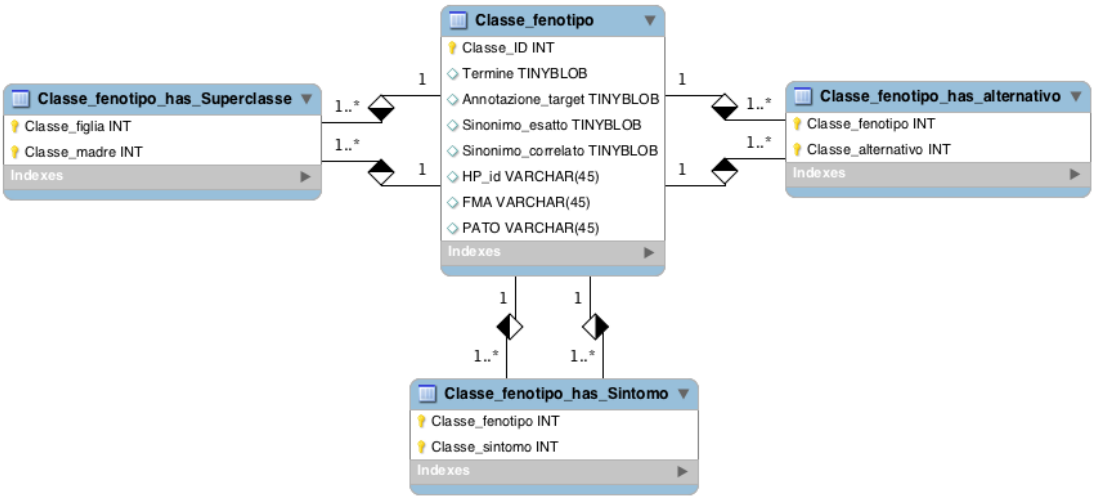


Figura 4.2: Schema relazionale dell'ontologia HPO, ottenuto dall'astrazione dell'ontologia in formato OWL.

## 4.4 Modelli Relazionali

BioSQL è un modello relazionale generico per database biologici, che supporta la gestione e la correlazione di features [1], annotazioni, tassonomie di riferimento, ontologie di termini specifici, sequenze biologiche sia proteomiche che genomiche. BioSQL fa parte di un progetto che include strumenti e framework in diversi linguaggi per la gestione e l'analisi di dati biologici: BioJava, BioPerl, BioRuby, BioPython.

Lo schema relazionale di BioSQL è stato estrapolato mediante un reverse engineering dello script di creazione delle tabella. La figura 4.3 mostra soltanto una

[1] Per features di una entità biologica, come una sottosequenza o un gene, si intende in bioinformatica una informazione pertinente a quella entità o una caratteristica peculiare di quella entità

parte dello schema relazionale di BioSQL, in cui si è cercato di raggruppare entità fondamentali a BioSQL o importanti per la progettazione del database in oggetto.

Come si può notare, un'entità importante per grado di connessioni nel diagramma è la tabella *term*. Per tutte le informazioni di tipo testuale, infatti, si utilizza un vocabolario controllato, i cui termini afferiscono a una particolare ontologia.

Ogni record in database è considerato un *bioentry*, con diversi codici identificativi assegnati al momento dell'inserimento, come avviene per tutti i database pubblici studiati. Le entità biologiche bioentry appartengono a una tassonomia, e sono descritte da annotazioni e da features.

Le features possono essere in relazione tra loro, e possono avere una posizione all'interno della sequenza biologica.

Le sequenze sono un particolare tipo di *bioentry*. Per ogni sequenza, insieme al codice identificativo e alla versione, vengono specificati la lunghezza e l'alfabeto (ad esempio A,C,G,T nel caso di sequenze nucleotidiche). Qualora la sequenza provenga da un database esterno conosciuto, si può mettere in relazione con l'entità *biodatabase*.

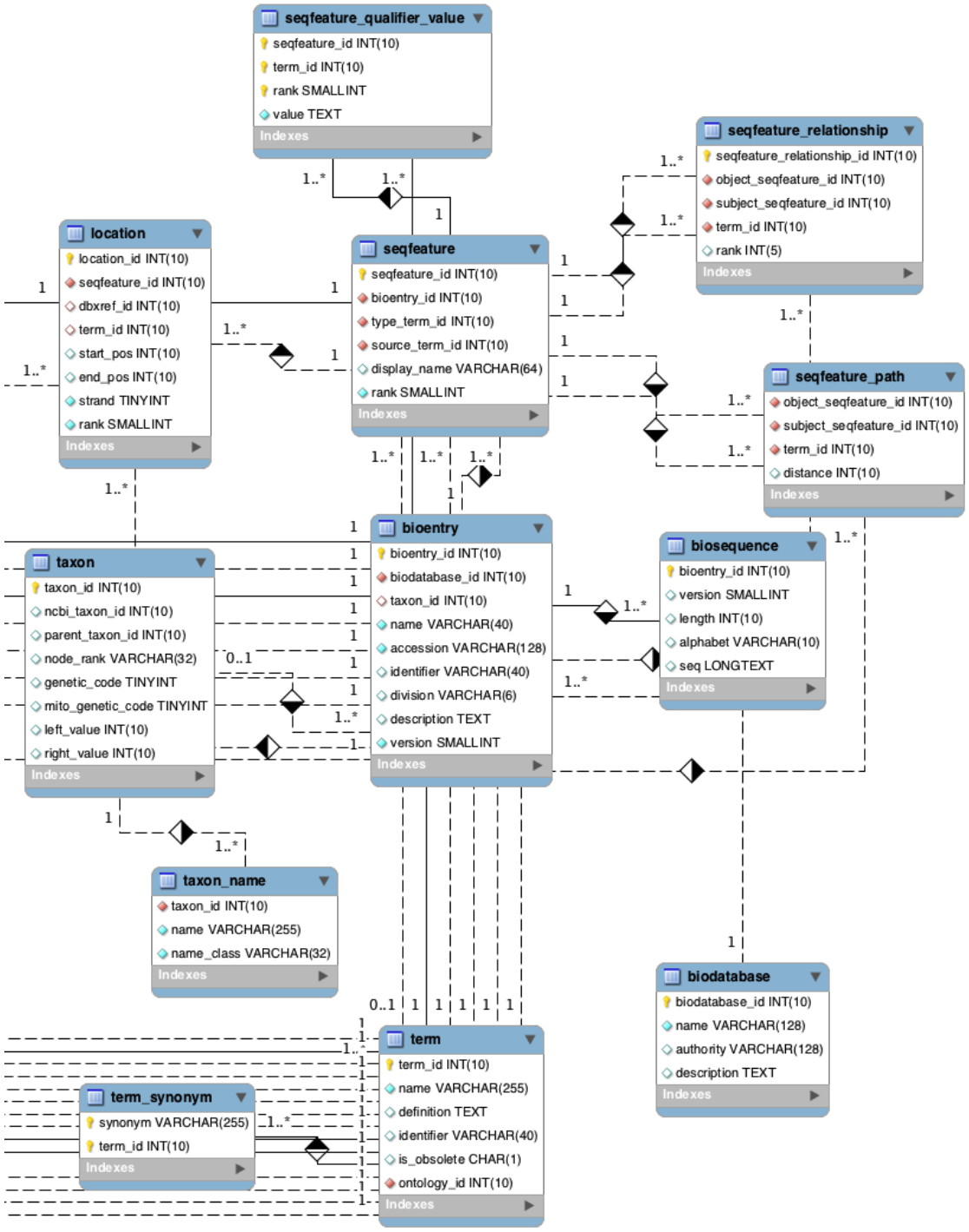


Figura 4.3: Sezione dello schema relazionale di BioSQL.

Ad esempio, una sequenza in formato GenBank come in figura 4.4 verrà memorizzata come il record *bioentry* mostrato in tabella 4.2

```
LOCUS       S63169S6                 22 bp    DNA     linear   PRI 25-AUG-1993
DEFINITION  NDP=Norrie disease {first three exons, microdeletion regions}
ACCESSION   S63178
VERSION     S63178.1  GI:386456
...
//
```

Figura 4.4: Sequenza di esempio in formato GenBank

| name | accession | identifier | division | description | version |
|---|---|---|---|---|---|
| S63169S6 | S63178 | 386456 | PRI | NDP=Norrie disease {first three exons, microdeletion regions} | 1 |

Tabella 4.2: Record bioentry di esempio.

Ogni *bioentry* può avere soltanto una *biosequence* ad essa associata e viceversa. Le sequenze possono avere il loro numero di versione, in modo indipendente dal record *bioentry* associato. La lunghezza della sequenza viene pre-calcolata e memorizzata per facilitare le possibili query su questo dato. Nel modello BioSQL risulta di estrema facilità ottenere informazioni su metadati come su dati biologici con delle semplici query SQL. In figura 4.5 è mostrata una query per selezionare la descrizione del *bioentry* associato alla sequenza più lunga in database. In figura 4.6 invece si cercano tutti i *bioentry* le cui sequenze contengono la stringa GATTACA.

```
SELECT   bioentry.description
 FROM     bioentry
          JOIN biodatabase USING (biodatabase_id)
          JOIN biosequence USING (bioentry_id)
 WHERE    biodatabase.name = 'genbank'
 ORDER BY biosequence.length DESC
 LIMIT 1
```

Figura 4.5: Query per cercare la sequenza più lunga in database.

```
SELECT  bioentry.*
 FROM    bioentry JOIN biosequence USING (bioentry_id)
 WHERE   biosequence.seq LIKE "%GATTACA%"
   AND   biosequence.alphabet = 'dna'
```

Figura 4.6: Query per cercare le sequenze contenenti la stringa GATTACA.

# 5 | Modellazione dei dati e delle funzioni sul DNA

> *"La scienza è conoscenza affermata mediante argomenti logici.*
>
> — Platone, Teeteto

Un modello matematico è un modello astratto che descrive, mediante il linguaggio matematico, le proprietà strutturali e il comportamento di un sistema. Eykhoff (1974) definisce il modello matematico come

> Una rappresentazione degli aspetti essenziali di un sistema esistente (o un sistema da costruire) che presenti in modo fruibile la conoscenza sul sistema.

Uno dei grandi vantaggi dei modelli matematici è la loro capacità di connettere i diversi componenti di un sistema complesso, come può esserlo un sistema di analisi delle sequenze genetiche. Grazie a un modello matematico è possibile osservare e misurare un sistema a un livello basso, convertire questi dati in parametri di un modello, combinando conoscenze matematiche e conoscenza sul tema specifico, e usare il modello per integrare questa conoscenza nella speranza di ottenere intuizioni su un livello più alto di funzionamento del sistema. In particolare si vuole tentare di ridurre fenomeni biochimici molto complessi a semplici interazioni tra sequenze definite, col fine ultimo di modellare un sistema GWAS (Genome Wide Association Studies) in tutti gli step di analisi, a partire dalla sequenza testuale di basi nucleotidiche.

## 5.1 La sequenza del DNA

Come ben si sa, il DNA è composto da due catene lineari di nucleotidi (dette strand) parallele che formano una doppia elica. Ci sono 4 differenti nucleotidi, caratterizzati dalla loro base azotata:

- Adenina (A)
- Guanina (G)
- Citosina (C)
- Timina (T)

Le basi A e G sono dette puriniche mentre le basi C e T sono dette pirimidiniche. L'alfabeto di una sequenza testuale del DNA è quindi formato dall'insieme

$$\Sigma = \{A, G, C, T\}$$

Definiamo quindi una sequenza

$$S = s_1 s_2 ... s_n$$

come una sequenza di DNA di $n$ basi nucleotidiche, dove $s_i \in \{A, G, C, T\}$.

Possiamo ora considerare il linguaggio L di tutte le possibili sequenze nucleotidiche come

$$L = \Sigma^*$$

Nel linguaggio appena definito si assume che le sequenze genetiche non contengano particolari pattern o strutture ripetitive. Nella realtà il genoma eucariotico, come è quello umano, ha proprietà strutturali ben definite sia sulle lunghe sequenze che nei segmenti corti. Nei prossimi paragrafi si analizzerà la struttura del genoma umano e si proporranno linguaggi formali atti a descriverla.

## 5.2 Il Genoma umano

Una sequenza genomica tipica umana ha più di 3 miliardi di basi nucleotidiche. Se prendiamo come riferimento il genoma HGR (*Human Genome Reference*, Genoma Umano di Riferimento), la sequenza di DNA del genoma è una sequenza

$$SG \in L \qquad |SG| = n$$

con $n = 3,324,592,091$. Questo numero fa riferimento alla lunghezza della sequenza del genoma di riferimento HGR, ma è molto variabile da persona a persona, a causa di varianti strutturali *indel* (insertion/deletion), per le quali ognuno di noi possiede sottosequenze in più o in meno nel nostro genoma totale. Sulla quantità e sull'intervallo di lunghezza di queste varianti strutturali non si hanno ancora informazioni certe. In media si considerano centinaia di migliaia di indel, della lunghezza tra 1 bp e 10000 bp (*base pair*).

Il genoma umano può essere suddiviso in diverse regioni in base allo scopo delle sequenze. La prima suddivisione è tra regioni che codificano proteine e regioni non codificanti. Nel genoma umano le regioni codificanti proteine sono rare, ammontando al 2,7% della sequenza totale. Oltre ai geni codificanti proteine, altre regioni codificano molecole di RNA strutturali, quali i componenti di RNA dei ribosomi e gli RNA Transfer.

Elementi ripetitivi di funzione sconosciuta danno conto di frazioni molto grandi dei nostri genomi. Le sequenze LINE (*long interspersed elements*, lunghi elementi interspersi) e le sequenze SINE (*short interspersed elements*, corti elementi interspersi), costituiscono rispettivamente il 21% e il 13% del genoma. Infine, sequenze ancora più altamente ripetute, DNA satellite, DNA minisatellite e DNA microsatellite, possono presentarsi in decine o persino centinaia di migliaia di copie, costituenti complessivamente fino al 15% del genoma.

Nella tabella 5.1 si schematizzano le proprietà di lunghezza e distribuzione delle sequenze appena descritte, che insieme costituiscono la struttura totale del genoma umano. L'unità di misura *bp* indica la coppia di basi (la base nucleotidica e la sua base complementare).

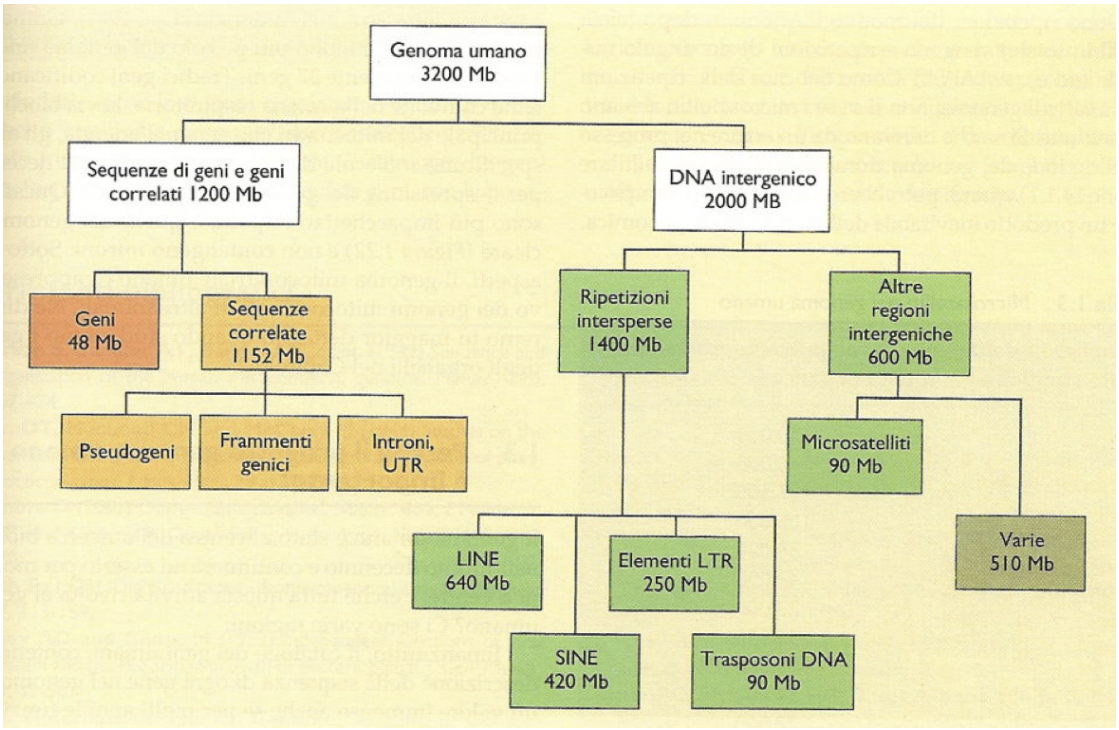


Figura 5.1: Macrostruttura gerarchica del genoma umano. L'unità di misura Mb indica 1 milione di basi nucleotidiche.

| Sequenza | Lunghezza | Classe |
|---|---|---|
| Esone codificante proteine | 122 bp in media | Sequenze di DNA uniche |
| Esone tRNA e rRNA | 150-300 bp | Moderatamente ripetitivo |
| Introne | Enorme variabilità | Moderatamente ripetitivo |
| LINE | >5 kb | Moderatamente ripetitivo |
| SINE | 200-300 bp | Moderatamente ripetitivo |
| Satellite | 5-220 bp | In tandem - Altamente ripetitivo |
| Minisatellite telomerico | 10-100 bp | In tandem - Altamente ripetitivo |
| Minisatellite ipervariabile | 6-50 bp | In tandem - Altamente ripetitivo |
| Microsatellite | 1-10bp | In tandem - Altamente ripetitivo |

Tabella 5.1: Le sequenze ripetute nel DNA. In bp (base pair) si indica la lunghezza media.

## 5.3 Un linguaggio formale per le sequenze genetiche

Trattare il genoma come un linguaggio può permettere di generalizzare le informazioni strutturali contenute nelle sequenze biologiche e investigarle utilizzando i metodi della teoria dei linguaggi formali. La teoria dei linguaggi formali può essere usata per modellare fenomeni biologici e in generale meccanismi genetici [Fernau, 2003].

Molti aspetti dei linguaggi formali sono simili ad alcuni processi biologici:

- Le *grammatiche pure* non fanno differenza tra simboli terminali e non terminali, così che tutte le parole generate dalle regole grammaticali sono inserite nel linguaggio generato. Questa nozione vale anche in campo biologico, dato che tutti i simboli in una sequenza di DNA hanno lo stesso ruolo.

- La *regola di cancellazione* $A \rightarrow \epsilon$ modella l'evento di cancellazione (*deletion*) nelle sequenza di DNA

- La *chain rule* $A \rightarrow B$ riflette i polimorfismi a singolo nucleotide nel DNA (SNP)

- La *repetition rule* $A \rightarrow AA$ modella le sequenze di DNA ad alto livello di ripetizioni come le ripetizioni *tandem*

- La *regola di produzione* $A \rightarrow BC$ modella la crescita nella sequenza del DNA.

- La *regola stocastica* modella le mutazioni casuali nelle sequenze di DNA.

I tentativi di ricondurre le sequenze di acidi nucleici a un linguaggio formale nascono fin dai primi anni dalla scoperta, da parte di Watson e Crick, della struttura del DNA. Infatti, proprio mentre venivano scoperte e descritte le caratteristiche del DNA e il funzionamento del codice genetico, il campo della linguistica veniva rivoluzionato dai lavori di Noam Chomsky.

Nel 1984 [Prof and Received, 1984] è stato proposto l'uso di linguaggi formali per descrivere sottoinsiemi di tutte le possibili parole (sottosequenze di DNA) che occorrono nel DNA e nell'RNA, dimostrando questo concetto con un linguaggio definito da un automa a stati finiti per l'RNA dei fagi di gruppo I. Sebbene non si conoscesse ancora struttura e contenuto del genoma umano, questo approccio permise di caratterizzare pattern grammaticali nelle informazioni genetiche.

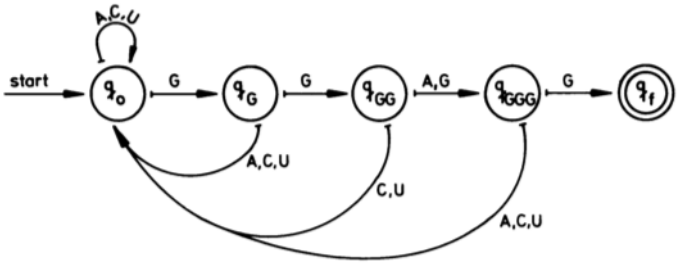


Figura 5.2: Il diagramma di un automa a stati finiti. Gli archi orientati indicano il simbolo di input della sequenza di RNA che induce la transizione. L'automa come si può notare accetta tutte le sequenze del linguaggio {A,G,C,U}* che terminano per GGGG o GGAG, ovvero il sito di legame al ribosoma che precede il cappuccio del gene.

Altre ricerche su questo tema non hanno prodotto direttamente grammatiche per il linguaggio genetico, ma piuttosto hanno usato formalismi grammaticali come strumenti per quelli che essenzialmente sono studi teorici sull'informazione [Jimenez-Montano, 1984, Ebeling and Jimenez-Montano, 1980] o hanno utilizzato analisi statistiche a livello di vocabolari, seguendo un approccio più tradizionale di linguistica comparativa [Pietrokovski et al., 1990, Pevzner et al., 1989, Brendel et al., 1986].

Solo successivamente è stata ripresa la mera formalizzazione linguistica, mediante grammatiche generative, di fenomeni biologici come la regolazione dei geni [Collado-Vides, 1989], la struttura e l'espressione dei geni[Searls, 1989], la ricombinazione e altre forme di mutazione e riarrangiamento[Searls, 1989], e sono state poste le basi per l'analisi computazionale delle sequenze di dati genetici [Searls, 1989, Searls and Noordewier, 1991].

Un approccio di linguistica formale applicato alle sequenze genetiche che tuttora viene utilizzato per gli studi sul tema è stato riproposto nel 1993 da Searls [Searls and Dong, 1993] e si basa sui *Lindenmayer systems* o *L-systems* [Lindenmayer, 1968]. Sebbene le ricerche appena descritte siano tutte basate su grammatiche di Chomsky, i sistemi di Lindenmayer, risalenti al 1968, sono stati originalmente sviluppati proprio come base per una teoria assiomatica dello sviluppo biologico.

### 5.3.1 L-Systems per il genoma umano

Recentemente è stato proposto un sistema [Damasevivcius, 2010] basato su L-grammar per risolvere uno dei più importanti problemi bioinformatici: individuare in una sequenza di DNA regioni con specifiche funzioni come i promotori (sequenza corte che precedono l'inizio dei geni) e i siti di splicing (punti di giunzione tra esoni e introni dove avviene lo splicing).

Gli L-Systems o L-grammars sono speciali classi di grammatiche parallele usate per modellare la crescita di organismi viventi, come lo sviluppo delle piante. Questa classe di grammatica può anche essere utilizzata per modellare la morfologia di vari organismi.

Negli L-Systems, le regole di produzione sono applicate in parallelo e possono sostituire tutte le lettere di una determinata parola contemporaneamente. Inoltre non c'è distinzione tra caratteri terminali e caratteri non terminali. La natura ricorsiva delle regole negli L-Systems porta ad ottenere forme ricorsive e frattali nei linguaggi generati, un'altra proprietà condivisa con le sequenze di DNA.

Per modellare le sequenze del DNA si utilizza una L-grammar stocastica libera dal contesto, ovvero una grammatica libera dal contesto probabilistica, definita dalla tupla

$$G = \{V, \omega, R, P\} \tag{5.1}$$

Dove:

- $V = \{A, C, G, T\}$ è un insieme di simboli (l'alfabeto) contenente gli elementi che possono essere rimpiazzati, nel nostro caso i 4 nucleotidi.

| | |
|---|---|
| Variabili: | A,C,G,T |
| Inizio: | CCCGAA |
| Regole: | 0.15:$(A \rightarrow CTGT)$, |
| | 0.95: $(C \rightarrow CGGTA)$, |
| | 1.00: $(G \rightarrow CTC)$, |
| | 0.98: $(T \rightarrow GCA)$ |

| | |
|---|---|
| Variabili: | A,C,G,T |
| Inizio: | CCTTT |
| Regole: | 0.13:$(A \rightarrow CGGGCA)$, |
| | 0.10: $(C \rightarrow CCCCG)$, |
| | 0.47: $(G \rightarrow ACGCC)$, |
| | 0.51: $(T \rightarrow AGACAT)$ |

Tabella 5.2: Regole nella grammatica di Lindenmayer per la generazione di giunzioni introne-esone ed esone-introne

- $\omega = V^k$ è la stringa di lunghezza K di simboli che definisce lo stato iniziale del sistema
- $R \subseteq V^1 \times V^L$ è un insieme finito di regole di produzione che definiscono il modo in cui uno specifico nucleotide può essere rimpiazzato dalla combinazione di altri nucleotidi. Una regola consiste di 2 stringhe–il predecessore e il successore.
- $P$ è un insieme di probabilità $p_j \in \lfloor 0, 1 \rfloor$ che una regola di produzione $r_j \in R$ sarà applicata.

Data la complessità delle sequenze nel genoma umano, per definire la grammatica di Lindenmayer si è utilizzato un processo di inferenza grammaticale, ovvero si è indotta una grammatica formale, sotto forma di regole di produzione, da un insieme di osservazioni, utilizzando tecniche di machine learning (ad esempio un classificatore SVM).

Nella tabella 5.2 si mostrano le regole grammaticali generate da un SVM per le giunzioni esone-introne e introne-esone.

## 5.4 DNA Walk

La visualizzazione dei dati è un tema di particolare interesse per la bioinformatica, in quanto si pone come obiettivo la ricerca di pattern all'interno delle sequenze

di DNA, con la speranza di poter comprendere completamente le funzioni biologiche sottostanti [Fitch and Sokhansanj, 2000]. In particolare negli ultimi anni le attenzioni si sono soffermate su funzioni di visualizzazione che mettessero in luce le caratteristiche ricorsive delle informazioni genetiche e dei fenomeni biologici, spesso sotto forma di frattali[Peng et al., 1992].

Per poter analizzare le informazioni genetiche nelle sequenze sia nei segmenti brevi che nella relazioni tra segmenti distanti, sono necessarie tecniche computazionalmente efficienti, che prendano in considerazione un sottoinsieme di queste informazioni ma mantengano il loro scopo ottenendo risultati significativi. La soluzione più ovvia è quella di considerare soltanto le sottosequenze codificanti, che come è stato accennato precedentemente rappresentano solo il 2%-3% di tutto il genoma.

Una di queste tecniche di visualizzazione è il DNA Walk [Peng et al., 1992]. Il DNA Walk è una rappresentazione vettoriale di una sequenza di DNA, trasformata in una traiettoria su un piano. Due paia di nucleotidi complementari (A-T, G-C) vengono usate come direzioni vettoriali, in modo tale che la sequenza di DNA si sposti verso l'alto con A, verso il basso con T, verso destra con G, verso sinistra con C, visualizzando una traiettoria (vedere figura 5.3).

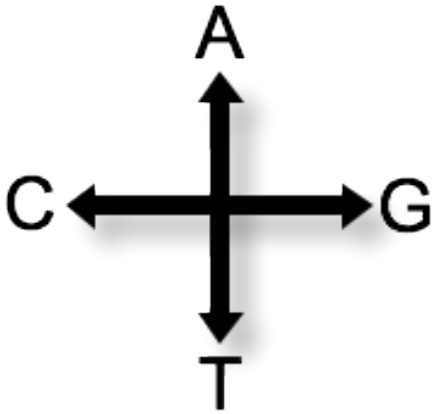


Figura 5.3: Il DNA Walk viene costruito muovendo ogni nucleotide un pixel nella direzione illustrata nel diagramma.

La funzione di DNA Walk è utile in quanto rende visibili diversi tipi di pattern all'interno delle sequenze genomiche. Grossi raggruppamenti di ripetizioni, palindromi, telomeri, inclinazioni GC (ovvero linee diagonali sul diagramma di visualizzazione che indicano una forte presenza di basi G e C ripetute) possono essere facilmente riconosciute semplicemente osservando il grafico del cammino.

## 5.5 Il codice genetico

Lo studio del codice genetico e delle analogie strutturali tra questo codice e le codifiche in informatica è un'attività a cui è stata dedicata molta ricerca, sia con lo scopo di creare un *DNA-computer* che per lo sviluppo della bioinformatica. Per studiare matematicamente il processo di codifica degli amminoacidi, gli insiemi di elementi biologici (acidi nucleici) possono essere rappresentati mediante matrici. Come per i precedenti modelli, prendiamo in considerazione un alfabeto di 4 simboli, le 4 basi nucleotidiche $\{A, G, C, T\}$. Durante la trascrizione, ovvero quando l'informazione viene trasmessa mediante mRNA, T viene sostituito con U (Uracile). Inoltre, la trascrizione fa sì che le basi associate siano quelle dell'altro strand, ovvero le basi opposte. Le basi A,G,C,T di DNA vengono associate rispettivamente alle basi U,C,G,A nell'RNA.

Nel codice genetico l'unità di base è una stringa di 3 basi nucleotidiche, detta codone o terzina. Avendo un alfabeto di 4 simboli, le possibili terzine sono

$$4^3 = 64$$

Tre di questi 64 codoni (UAA, UAG, UGA) sono codoni di stop–nessun amminoacido corrisponde al loro codice. I restanti 61 codoni rappresentano 20 differenti amminoacidi. Queste terzine di codice genetico nell'mRNA che codificano specifici acidi durante il processo di traduzione, hanno un organizzazione matematicamente logica. Possiamo rappresentare questa traduzione da un punto di vista matematico mediante una funzione che mappa i codoni negli amminoacidi

$$g : C \to A,$$

dove $\mathbf{C} = \{(x_1x_2x_3) : x_i \in \mathbf{R} = \{A, C, G, U\}\}$ è l'insieme di codoni, e $\mathbf{A} = \{Ala, Arg, Asp, ..., Val, UAA, UAG, UGA\}$ è l'insieme di amminoacidi e codoni di terminazione. Nella tabella 5.3 è illustrata in modo completo la mappatura tra i codoni e gli amminoacidi. L'intestazione verticale di sinistra nella tabella indica il primo nucleotide, l'intestazione in alto indica il secondo nucleotide, l'ultima colonna indica il terzo nucleotide.

Come si nota intuitivamente, la funzione g è una funzione suriettiva, in quanto ogni elemento del codominio (insieme degli amminoacidi) è immagine di almeno un elemento del dominio (insieme dei codoni). La funzione g non è però iniettiva: a più codoni dell'insieme $C$ corrisponde lo stesso amminoacido nell'insieme $A$.

**TODO: matrici stocastice Genetic code, attributive mappings and stochastic matrices**

## 5.6 Cromosomi e locus

Il termine locus è a volte erroneamente confuso con il concetto di gene, ma si riferisce in realtà a una posizione della mappa genomica. Una definizione più precisa è data dalle *Rules and Guidelines from the International Committee on Standardized Genetic Nomenclature for Mice* che afferma: "Un locus è un punto nel genoma, identificato da un marcatore, che può essere in qualche modo mappato. Non deve necessariamente corrispondere a un gene; può, per esempio, identificare un segmento anonimo. Un singolo gene può contenere più *loci* al suo interno (ognuno definito da un marcatore)[...]" [Epp, ] .

La nomenclatura per definire un locus prende in considerazione il cromosoma di cui fa parte, il braccio del cromosoma, la regione, la banda. Questa nomenclatura, definita e aggiornata nel corso dei decenni nell'ISCN (International System for Human Cytogenetic Nomenclature)[Shaffer et al., 2005] è lo standard internazionale per l'individuazione di specifiche posizioni all'interno delle sequenze genetiche.

Un esempio di locus può essere 6p21.3. Il primo numero indica il cromosoma, in questo caso il cromosoma 6. La lettera successiva indica il braccio del cromosoma, in cui distinguiamo braccio corto e braccio lungo (vedi figura 5.4), in questo caso p

| Matrice tridimensionale di codifica | | | | | |
|---|---|---|---|---|---|
| | U | C | A | G | |
| U | UUU Phe [F] | UCU Ser [S] | UAU Tyr [Y] | UGU Cys [C] | U |
| | UUC Phe [F] | UCC Ser [S] | UAC Tyr [Y] | UGC Cys [C] | C |
| | UUA Leu [L] | UCA Ser [S] | UAA Ter[end] | UGA Ter [end] | A |
| | UUG Leu [L] | UCG Ser [S] | UAG Ter [end] | UGG Trp [W] | G |
| C | CUU Leu [L] | CCU Pro [P] | CAU His [H] | CGU Arg [R] | U |
| | CUC Leu [L] | CCC Pro [P] | CAC His [H] | CGC Arg [R] | C |
| | CUA Leu [L] | CCA Pro [P] | CAA Gln [Q] | CGA Arg [R] | A |
| | CUG Leu [L] | CCG Pro [P] | CAG Gln [Q] | CGG Arg [R] | G |
| A | AUU Ile [I] | ACU Thr [T] | AAU Asn [N] | AGU Ser [S] | U |
| | AUC Ile [I] | ACC Thr [T] | AAC Asn [N] | AGC Ser [S] | C |
| | AUA Ile [I] | ACA Thr [T] | AAA Lys [K] | AGA Arg [R] | A |
| | AUG Met [M] | ACG Thr [T] | AAG Lys [K] | AGG Arg [R] | G |
| A | GUU Val [V] | GCU Ala [A] | GAU Asp [D] | GGU Gly [G] | U |
| | GUC Val [V] | GCC Ala [A] | GAC Asp [D] | GGC Gly [G] | C |
| | GUA Val [V] | GCA Ala [A] | GAA Glu [E] | GGA Gly [G] | A |
| | GUG Val [V] | GCG Ala [A] | GAG Glu [E] | GGG Gly [G] | G |

Tabella 5.3: Tabella di mappatura tra codoni e amminoacidi.

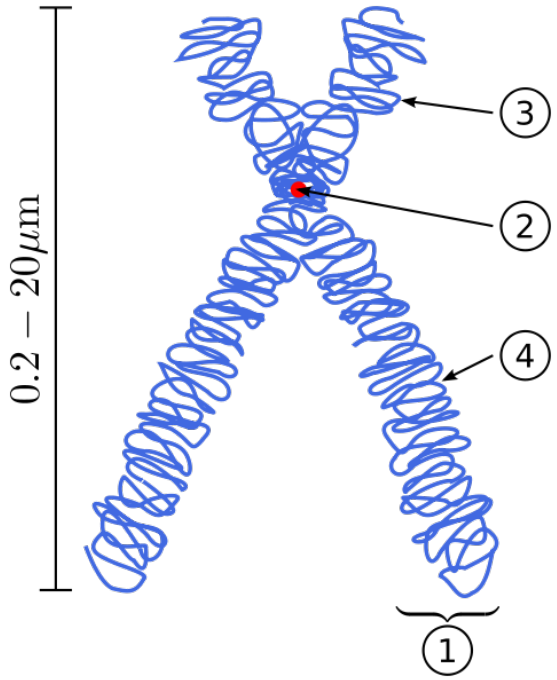


Figura 5.4: Componenti del cromosoma: (1) Cromatide (2) Centromero (3) Braccio corto (4) Braccio lungo - Sorgente: Wiki Commons, File:Chromosome.svg

indica il braccio corto del cromosoma (*p* sta per *petit* in francese). Le cifre prima del punto vanno lette come regione 2, banda 1, mentre la cifra dopo il punto indica la sottobanda. Guardare la tabella 5.4 per una descrizione più dettagliata della nomenclatura.

## 5.7 Le varianti genetiche

La genotipizzazione è il processo di analisi che, dato il dna di un individuo, ottiene l'insieme delle variazioni nella sequenza, ovvero di varianti nella sequenza che la rendono unica rispetto agli altri individui. Il genotipo è infatti la costituzione genetica di un singolo individuo.

La sequenza di DNA del genoma umano è per il 99,9% identica per tutti gli individui. Dell'interno genoma, ovvero 3,2 miliardi di nucleotidi, 3 milioni di questi sono nucleotidi varianti, che cambiano da individuo a individuo. La particolare

| Componente | Descrizione |
|---|---|
| 6 | Il numero 6 indica il cromosoma di riferimento. Nel genoma umano esistono 21 cromosomi numerati più i 2 cromosomi X e Y |
| p | Indica il braccio del cromosoma che contiene il locus. p (da *petit*) indica il braccio corto, mentre q (la lettera successiva alla p) indica il braccio lungo del cromosoma |
| 21.3 | Questo valore indica la posizione all'interno del braccio del cromosoma. |

Tabella 5.4: La nomenclatura ISCN 1995 per i locus nelle sequenze di DNA

forma di gene che ognuno di noi possiede in un determinato locus (posizione nel cromosoma) è detta allele. I ciascun locus si possiedono 2 alleli, uno ereditato dal padre, l'altro dalla madre.

Le **mutazioni cromosomiche** sono mutazioni grandi e per questo portano a gravi conseguenze. Le mutazioni cromosomiche possono avvenire a livello di struttura (delezione, duplicazione, inversione, traslocazione) o nel numero dei cromosomi (monosomie e trisomie).

Le **mutazioni geniche o puntiformi** sono dovute nella maggior parte dei casi per la sostituzione di una singola base nucleotidica con un'altra. Esistono poi altre variazioni dovute all'**inserimento** o alla perdita (**delezione**) di una base nel DNA (*indels*). Tra le mutazioni puntiformi possiamo distinguere **mutazioni silenti**, per le quali il codone risultante dalla mutazione continua ad essere associato allo stesso amminoacido, senza nessuna conseguenza nei successivi processi biologici, *mutazioni di senso*, ovvero tutte le normali caratteristiche che ci distinguono gli uni dagli altri, e *mutazioni non senso*, quando il codone risultante dalla mutazione codifica per il simbolo di stop, e la proteina prodotta sarà più corta di quella originale.

Tra le varianti nel DNA le più interessanti sono i **polimorfismi**, ovvero le variazioni normali tra gli individui. Un determinato sito nel DNA è definito polimorfismo se di esso si conoscono almeno due forme alleliche, la più rara delle quali

ha una frequenza di almeno l'1%.

Lo **SNP** (Single Nucleotide Polymorphism) è il polimorfismo più semplice, a un solo nucleotide. Questo polimorfismo è molto abbondante all'interno del genoma, si conta infatti la presenza di uno SNP per ogni 1000 basi nel genoma umano.

## 5.8 Trovare gli SNP nell'era NGS

Prima dell'avvento delle tecnologie di Next Generation Sequencing, i metodi di genotipizzazione, ovvero i processi per l'identificazione dei polimorfismi in un determinato genoma, erano per la maggiorparte attività biochimiche di laboratorio: metodi enzimatici (metodo Sanger, RFLP, FEN, Primer extension, OLA), metodi basati sulle proprietà fisiche del DNA (DGGE, SSCP, DHPLC), metodi basati sull'ibridazione (TaqMan assay, ASO, DASH, Molecular beacons). Con le macchine NGS, e quindi con la possibilità di sequenziare per intero un genoma, o un esoma, si è aperta la strada alla genotipizzazione in silico. Una volta ottenuta la sequenza completa adesso è possibile allinearla a un genoma di riferimento (HGR) e andare a cercare direttamente i marcatori in posizioni predeterminate nella nuova sequenza. Data quindi una sequenza già ottenuta, allineata con il genoma di riferimento HGR, e le posizioni conosciute degli SNP all'interno del genoma, la funzione di genotipizzazione in silico

$$gn : (L_a, P_s) \rightarrow (\Sigma \times P_s)$$

con $L_a$ insieme delle stringhe del linguaggio allineate all'HGR, $P_s$ insieme di posizioni degli SNP, ovvero di indici della stringa dove esiste un polimorfismo a singolo nucleotide e $\Sigma$ l'alfabeto dei 4 nucleotidi.

## 5.9 Funzioni di associazione geni-malattia

L'avanzamento delle tecnologie di sequenziamento ha rapidamente cambiato i metodi e la ricerca della medicina genetica. Gli Whole-Genome Sequencing (WGS, sequenziamento dell'intero genoma) e Whole-Exome Sequencing (WES, sequenzia-

mento dell'interno esoma) hanno provato essere dei metodi efficaci e fattibili per la scoperta delle cause genetiche di malattie rare e complesse.

Nonostante i costi siano scesi esponenzialmente negli ultimi anni, il Whole-Exome Sequencing resta un processo ancora costoso per poterne vedere l'utilità negli studi di associazione su larga scala. Inoltre l'enorme quantità dei dati, sempre crescente, continua a rappresentare una grossa sfida per le infrastrutture di storage e analisi di queste informazioni. Al contrario, una recente tendenza è quella di puntare sul sequenziamento dell'intero esoma (WES), mantenendo la completezza delle informazioni genetiche più importanti e riducendo i costi a una frazione rispetto a quelli del WGS.

Il WES è stato già utilizzato per l'identificazione di difetti molecolari nelle malattie a singolo gene, per capire meglio fattori genetici associati a malattie più complesse, e per supportare la diagnosi nei pazienti, migliorandone l'affidabilità.

Nell'analisi NGS, che è sinonimo di GWAS, si considerano spesso tutti gli step, già ampiamente discussi nei capitoli precedenti, che culminano nella ricerca degli SNP. Il passo ultimo è quello di ricercare i polimorfismi che con una certa probabilità possono essere coinvolti nello sviluppo di un dato fenotipo. In alcuni casi vengono prima estratti i polimorfismi (genotipo) di più individui e poi questi vengono confrontati, in altri l'individuazione dei polimorfismi e la relativa associazione con particolari fenotipi avviene nello stesso task di confronto tra sequenze di individui diversi.

Nel corso degli ultimi anni a questo scopo sono stati sviluppati più di un centinaio di tool [Pabinger et al., 2013], per la ricerca dei polimorfismi e il confronto tra genomi/esomi. Una funzione tipica di studio di associazione GWAS è la seguente:

$$f_{\text{GWAS}} : (\{S_1, S_2, S_3, ...\}, \{M_1, M_2, M_3, ...\}) \rightarrow (\Sigma \times P_s \times O_{\text{PROB}})$$

Dove le sequenze $S_i$ sono le sequenze di individui sani per la malattia in esame, le sequenze $M_i$ sono le sequenze degli individui malati, $\Sigma$ è l'insieme alfabeto delle basi, $P_s$ è l'insieme delle posizioni nei cromosomi, $O_{\text{PROB}}$ è l'insieme delle probabilità che l'SNP individuato sia associato alla malattia.

Come accennato molti tool e relativi algoritmi sono stati sviluppati per l'associazione gene-malattia. Tuttavia la maggior parte restituiscono risultati diversi sulle stesse sequenze. [Kraft and Cox, 2008]

Definizione delle funzioni di studio delle associazioni a partire da un numero grande di stringhe $\{S_1, S_2, ..., S_n\} \in L*$.

Per uno studio approfondito delle funzioni di associazione GWAS si rimanda al relativo capitolo.

# 6 | Studi di associazione geni-malattie

L'obiettivo degli studi di associazione GWAS (Genome Wide Association Studies) e EWAS (Exome Wide Association Studies) è quello di determinare le relazioni tra fenotipi e genotipi, così da individuare quei geni o quelle sottosequenze del DNA che comportano particolari caratteristiche fisiche, patologie congenite o predisposizione a disfunzioni. Attualmente gli studi GWA sono gli strumenti più potenti nell'identificare i geni e le varianti associati alle malattie [Hirschhorn and Daly, 2005].

Per fenotipo nello specifico si intende tutto l'insieme di caratteristiche umane osservabili, come le sue proprietà morfologiche, fisiologiche, di sviluppo, o di comportamento. Tra le caratteristiche fenotipiche che vanno a costituire il fenotipo rientrano anche le patologie, non solo quelle congenite, e in generale il quadro clinico di un individuo.

Il genotipo è rappresentato invece dal corredo genetico di un individuo, identificabile nella sua sequenza di DNA. Per ridurre la complessità nell'analisi delle relazioni tra genotipo e fenotipo, possiamo più semplicemente considerare come genotipo l'insieme di varianti genetiche che distinguono un individuo dagli altri. Come già visto nei capitolo precedenti esistono diverse tipologie di varianti: inserimenti e cancellazioni, polimorfismi a singolo nucleotide (SNP), varianti strutturali. Tipicamente, da pochi anni a questa parte, vengono presi in considerazione soltanto gli SNP[1]. Gli SNP sono presenti in grandi quantità in un genoma e hanno dimostrato essere una rappresentazione molto soddisfacente del genotipo umano.

Nell'affrontare questo problema, bisogna tener conto di un secondo fattore, che

[1] In uno studio di associazione le varianti che costituiscono il genotipo, come gli SNP, vengono anche chiamate *markers* o marcatori

è quello ambientale. Le caratteristiche ambientali sono tutte quelle condizioni non dipendenti dalla costituzione genetica ma scaturite dall'ambiente circostante (ad esempio radiazioni, inquinamento, altitudine) o da abitudini comportamentali (ad esempio l'alimentazione, l'attività fisica). Nel complesso possiamo schematizzare questa interazione nel modo seguente:

$$genotipo(G) + ambiente(A) \rightarrow fenotipo(P) \tag{6.1}$$

Gli studi di associazione sono quindi delle analisi statistiche su un insieme relativamente grande di dati, che contiene per ogni individuo preso in considerazione le sue varianti genetiche, le sue caratteristiche fenotipiche, e in alcuni casi informazioni di carattere ambientale.

## 6.1 Linkage Disequilibrium

Il Linkage Disequilibrium (LD) è un insieme di associazioni statistiche, a livello di popolazione, tra marcatori genetici (varianti genetiche) e caratteristiche fenotipiche. L'analisi del LD su più individui rappresenta un approccio spaziale allo studio delle cause genetiche di tratti quantitativi, ovvero quei caratteri fenotipici che variano in modo continuo e non discreto. Gli studi di associazione GWAS si basano su questo approccio.

Esistono diverse metodologie statistiche per determinare queste associazioni.

## 6.2 Scelta degli SNP

Come accennato, per caratterizzare il genotipo (genotipizzazione) di un individuo si utilizza l'insieme degli alleli presenti in punti predeterminati del genoma, ovvero laddove si conosce siano presenti polimorfismi a singolo nucleotide (SNP).

La scelta di questi SNP è di fondamentale importanza per l'affidabilità di uno studio GWAS, questa selezione infatti influisce direttamente sulla presenza o meno di falsi positivi. L'insieme degli SNP, ad esempio, varia significativamente in

base al gruppo etnico di discendenza di un individuo. Per questo motivo il progetto HapMap ha analizzato il genotipo di centinaia di individui dividendolo in tre categorie genealogiche: popolazioni di ascendenza africana, asiatica ed europea.

Nel caso del sequenziamento di un intero genoma (WGS) o di un intero esoma (WES) si utilizza una mappa di SNP per estrarre nelle esatte posizioni gli alleli, ovvero il valore (base nucleotidica) che assume la sequenza in quel punto.

Attualmente esiste anche la possibilità di utilizzare dei Chip SNP per sequenziare soltanto le basi relative agli SNP conosciuti, abbassando di molto i costi della genotipizzazione.

La selezione degli SNP per il processo di genotipizzazione in ambito WGS spesso viene effettuata basandosi su Chip SNP commerciali (es. Affymetrix SNP Array, Illumina HumanHap) e sul database pubblico HapMap [Barrett and Cardon, 2006, Li et al., 2008, Frazer et al., 2007]. HapMap contiene gli SNP per i quali ogni allele occorre minimo nell'1% della popolazione, e viene considerata come la fonte più autorevole per la scelta degli SNP, tanto che molti Chip SNP commerciali si basano direttamente sulle collezioni HapMap.

Nella scelta di un Chip SNP per sequenziare, o nella selezione di un insieme di SNP da considerare in uno studio GWA, si prendono in considerazione gli indici di copertura, ovvero una stima di quanto l'insieme degli SNP scelti L'indice di copertura locale di un Chip SNP si ottiene prendendo in considerazione regioni cromosomiche della dimensione di 1Mbp, adattando la formula di Barret e Cardon [Barrett and Cardon, 2006]:

1. R: il numero di SNP comuni in HapMap

2. T: il numero di SNP nel Chip che si sta valutando

3. L: il numero di SNP non presenti sul Chip ma a cui è attribuito un fattore di ricombinazione maggiore di una certa costante[2] con un altro SNP presente nel chip nel raggio di 250 mila basi.

Il valore di copertura locale è stimato con la formula

[2]Il fattore di ricombinazione deve essere $r^2 > 0.8$

$$[L/(R-T) \times (G-T) + T]/G \tag{6.2}$$

Nella piattaforma GWAS proposta, questo valore, di fondamentale importanza per la valutazione dei risultati, può essere automaticamente calcolato e associato ad ogni genotipo, così da essere preso in considerazione nelle successive analisi statistiche.

## 6.3 Analisi con PLINK

PLINK è un insieme di strumenti open-source per gli studi di associazione, che permettono lo studio di grandi dataset di genotipi e fenotipi [Purcell et al., 2007]. PLINK è stato inizialmente sviluppato nel 2007 per rispondere alla necessità di analizzare grandi volumi di dati provenienti dai microarray SNP. Il tool si presenta come un programma command-line, scritto in C/C++, ma esistono anche interfacce GUI basate su java, come gPLINK, e interfacce per l'integrazione con il software statistico R [PLINK, 2013, ].

PLINK non è soltanto un tool per l'analisi statistica di associazione tra genotipo e fenotipo, ma offre strumenti complementari per la gestione dei dati, statistiche generali, stima delle relazioni tra individui, registrazione, ordinamento, fusione e inversione degli *strand* di DNA, estrazione di sottoinsiemi di dati e molte altre funzioni versatili.

PLINK accetta dati di input in diversi formati. Il formato più comune è PED/MAP. PED/MAP suddivide l'informazione in due parti: i file MAP contengono la posizione di tutti i marker presi in considerazione, con un identificatore univoco per ciascuno; i file PED contengono il genotipo e il fenotipo degli individui presi in esame, insieme alle relazioni di parentela con gli altri individui e al sesso.

Molto utili al dimensionamento degli studi di associazione sono le funzionalità di testing di PLINK. E' possibile ad esempio simulare dati genotipici, scegliendo il numero di casi affetti e il numero di casi non affetti, e analizzando poi i risultati su

un Manhattan plot [3] per giudicare la capacità di rilevare l'associazione a partire dalla quantità di dati simulata.

### 6.3.1 Formati PED/MAP

Un file in formato **PED** è un flat file delimitato da spazi bianchi (carattere spazio o carattere tabulazione), che contiene il fenotipo e il genotipo degli individui soggetti allo studio di associazione. Una caratteristica importante dell'origine dati PED è che ogni file PED fa riferimento a una sola caratteristica fenotipica, e per ogni individuo viene espresso se il soggetto è affetto o non affetto dal fenotipo in questione. In 6.1 è illustrata la struttura del formato. Il formato non è rigido ma può essere personalizzato per alcuni aspetti. PLINK infatti prevede opzioni command-line in cui specificare proprietà di formato specifiche, ad esempio risulta molto utile la possibilità di utilizzare file trasposti (TPED/TFAM), in cui gli SNP sono contenuti in righe e ogni colonna rappresenta un individuo.

Un file **MAP**, come i file PED, è un flat file delimitato da spazi bianchi (spazio o tabulazione) per le colonne e un nuova linea per ogni record. Il file MAP contiene le posizioni nei cromosomi per ogni SNP che è stato *genotipizzato* e tipicamente è strutturato come in tabella 6.2.

Come per i file PED, esistono molte varianti disponibili per i file map. Per esempio, la distanza genetica[4] può essere specificata in centi-morgan con un'opzione da command line.

Ad ogni riga del file MAP corrispondono due colonne del file PED. Si assume che le colonne di PED [5] e le righe di MAP siano ordinate nello stesso modo, ovvero che il marcatore SNP descritto nella riga 1 del file MAP è riferito alle colonne 7 e 8 nel file PED.

[3] Un particolare grafico di dispersione, in cui le coordinate genomiche sono visualizzate sull'asse delle X e il logaritmo negativo dell'associazione sull'asse Y

[4] Il Morgan e il cM sono unità di misura della distanza genetica tra due loci, ovvero tra due posizioni nel DNA. Da Wikipedia: Due loci genici distano di un cM quando danno luogo ad una ricombinazione ogni cento meiosi, ovvero quando ogni meiosi dà come prodotto 0,01 geni (posti su loci distanti 1 cM) ricombinanti.

[5] a partire dalla settima colonna

| Numero colonna | Nome | Descrizione |
|---|---|---|
| Colonna 1 | FID | Identificatore alfanumerico univoco della famiglia di appartenenza |
| Colonna 2 | IID | Identificatore alfanumerico univoco dell'individuo |
| Colonna 3 | PID | Identificatore alfanumerico univoco del padre |
| Colonna 4 | MID | Identificatore alfanumerico univoco della madre |
| Colonna 5 | Sex | Sesso dell'individuo. E' codificato con le costanti: 1 = uomo; 2 = donna; qualsiasi altro numero = sconosciuto. Se il sesso di un individuo non è conosciuto, qualsiasi carattere che non sia uno o 2 può essere usato. |
| Colonna 6 | Phe | Fenotipo dell'individuo. Ogni file PED può contenere non più di un fenotipo, e se lo contiene deve essere indicato nella colonna 6. Il campo fenotipo può essere una variabile binaria o numerica. Nel caso sia binaria la costante indica lo stato di affezione in questo modo: 1 = non affetto; 2 = affetto; 0 = dato mancante. Nel caso la variabile sia quantitativa questo viene automaticamente rilevato, se il valore è diverso da 0, 1 o 2, si assume che il fenotipo sia un tratto quantitativo. Inoltre nel caso quantitativo la costante per indicare il dato mancante è il valore negativo di default -9. |
| Colonna 7...n | Gen | Genotipo, costituito dalla lista delle basi azotate nelle posizioni indicate dai markers (nel file MAP). A partire dalla colonna 7, ogni colonna contiene uno SNP dell'individuo nell'ordine delle posizioni del file MAP. E' importante tener presente che di default PLINK assume che i marker siano biallelici, ovvero per ogni SNP si hanno 2 basi azotate, una per il cromosoma ereditato dalla madre, uno per il cromosoma ereditato dal padre. Ad esempio le colonne 7 e 8 contengono la coppia di genotipi per l'SNP1, le colonne 9 e 10 per l'SNP2 e così via. |

Tabella 6.1: Formato di un file PED con la descrizione delle colonne.

| Numero colonna | Descrizione |
|---|---|
| Colonna 1 | Numero del cromosoma |
| Colonna 2 | Identificatore del marcatore |
| Colonna 3 | Distanza genetica (in Morgan) |
| Colonna 4 | Posizione fisica della base (in unità bp, *base pairs*) |

Tabella 6.2: Formato di un file MAP con la descrizione delle colonne.

## 6.4 Il progetto HapMap

Il progetto HapMap [Tanaka, 2003] ha come obiettivo quello di determinare la varianti comuni nelle sequenze della popolazione umana e rendere questi dati facilmente fruibili.

Ogni genoma umano si differenzia dagli altri per circa lo 0,1% della sequenza totale, e tra queste variazioni la più importante è la più studiata è quella a singolo nucleotide (SNP).

Nello specifico lo scopo di HapMap è quello di mappare tutti gli aplotipi comuni della popolazione umana mondiale. L'aplotipo è una sequenza di alleli appartenenti a SNP consecutivi su un particolare cromosoma: gli SNP vicini tra loro nella sequenza del DNA non sono soggetti a ricombinazione cromosomica.

Per chiarire meglio l'idea sul rapporto tra sequenza di DNA, SNP e aplotipi si osservi la figura 6.1: nella sezione **a** viene mostrata la stessa sottosequenza di DNA per quattro individui diversi. La maggiorparte della sequenza del DNA mostrata è identica, ma 3 basi differiscono laddove è presente una variazione. Ogni SNP ha due possibili alleli, ad esempio nel primo SNP l'allele può essere C oppure T. Nella sezione **b** si mostra la composizione degli aplotipi. Un aplotipi è fatto da una particolare combinazione di alleli negli SNP vicini. In questo caso sono mostrati i genotipi per 20 SNP, estesti su 6000 basi di DNA. Sono mostrate solo le basi variabile, comprese quelle individuate nella sezione **a** della figura. Nella sezione **c** si dimostra come la genotipizzazione dei 3 SNP mostrati basti a identificare tutti e 20 gli SNP adiacenti. Ad esempio se un particolare cromosoma ha il pattern A-T-C nei tre tag SNP, questo pattern determinerà l'aplotipo 1.

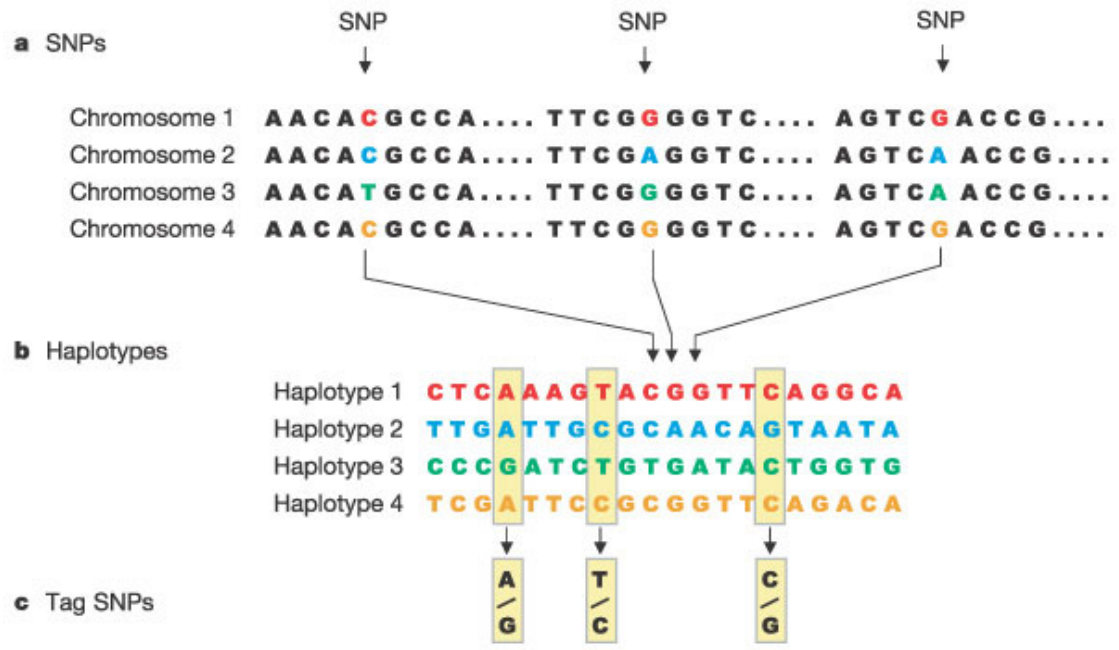


Figura 6.1: Illustrazione di SNP, aplotipi e annotazioni SNP su una sequenza di DNA.

## 6.5 dbSNP

Il database dbSNP è una banca dati mantenuta da NCBI per catalogare i polimorfismi a singolo nucleotide di riferimento, attualmente contenente più di 10 milioni di varianti SNP uniche [Smigielski et al., 2000]. Come per gli altri database NCBI, dbSNP è pubblico e permette la sottomissione di nuovi dati. Nel momento della sottomissione di una variante il sistema restituisce un identificatore unico per la sottomissione, detto *ss*. Se la variante sottomessa è unica, ovvero non era mai stata inserita in database prima di allora, dbSNP assegna alla variante un identificatore detto *rs*.

Il codice *rs* è stato scelto come riferimento nel sistema in progettazione per identificare univocamente le varianti SNP e garantire l'interoperabilità con il database dbSNP, e con tutti i database che utilizzano questo indice.

## 6.6 SNPedia

SNPedia (pronunciato SNIPìdia) è una risorsa di tipo wiki che documenta le scoperte degli studi GWAS descrivendo per ogni SNP il suo coinvolgimento in tratti fenotipici e in particolare in patologie umane, in una forma leggibile sia dagli utenti che dai software [Cariaso and Lennon, 2012].

Ogni voce di SNPedia è strutturata in modo tale da permettere l'associazione automatizzata del fenotipo descritto con un genotipo o con un insieme di genotipi (varianti genetiche). Ad ogni SNP è associato il relativo identificatore univoco del database NCBI dbSNP.

Oltre all'interfaccia wiki di SNPedia, il progetto mette a disposizione i dati sotto forma di servizio DAS (Distributed Annotation System) [Prlić et al., 2007] con il supporto dell'istituto europeo di bioinformatica EBI, e in formato GFF3, lo standard di input per il software GBrowser [Stein et al., 2002].

DAS è un protocollo di rete client-server basato su HTTP molto utilizzato nei database bioinformatici, in cui il client richiedere un URL e riceve una risposta XML. L'esistenza di questo servizio permette l'integrazione automatizzata di questi dati in una piattaforma GWAS.

GBrowser (Generic Genome Browser) è un'applicazione web-based per visualizzare sequenze genomiche, annotazioni e altre informazioni biologiche. Include tra le feature la possibilità di navigare nella sequenza, ingrandendo o scorrendo regioni arbitrarie di una sequenza, la possibilità di raggiungere una precisa posizione o ricercare un testo all'interno di tutte le annotazioni.

SNPedia possiede all'interno delle voci riferimenti ad altri database tra cui il già citato dbSNP, HapMap, Ensembl, PharmGKB.

# 7 | Modello integrato e interoperabile

## 7.1 Integrazione dei dati

Gli studi di associazione descritti nel capitolo 6 sono basati sul principio di integrazione tra dati genetici e dati fenotipici degli individui. L'integrazione di dati eterogenei biologici e medici, che possono andare anche oltre la sola descrizione fenotipica, porta inconfutabilmente benefici alla scoperta di nuova conoscenza in campo medico [Biesecker, 2010, Eronen and Toivonen, 2012]. Un esempio di questo approccio alla ricerca bioinformatica è il consorzio americano eMerge Network [McCarty et al., 2011], che ha già mostrato negli ultimi due anni importanti risultati scientifici [Gottesman et al., 2013].

### 7.1.1 Modello eMerge Network

Il progetto eMerge Network è nato nel 2007 con lo scopo di collegare e combinare tra loro sorgenti dati di diversa natura sparse nel territorio (vedi 7.2), tra cui banche dati genetiche e banche dati cliniche, con lo scopo di realizzare studi GWA con un alto potenziale di ricerca, nella prospettiva di realizzare per la prima volta una Medicina Personalizzata nella quale la genomica e sia incorporata nel tessuto sanitario. L'innovazione introdotta dal progetto nelle metodologie degli studi di associazione sta nel derivare le informazioni del fenotipo direttamente dalle EMR

(Electronical Medical Record), ovvero dalla digitalizzazione delle cartelle cliniche. In sintesi gli obiettivi principali del progetto erano:

- Utilizzare i dati clinici EMR per realizzare un sistema di fenotipizzazione robusto e automatizzato
- Condurre studi di associazione sull'intero genoma (GWAS) utilizzando i fenotipi derivati
- Esplorare le implicazioni etiche, legali e sociali associate al GWAS basato su EMR e alla condivisione dei dati su larga scala.

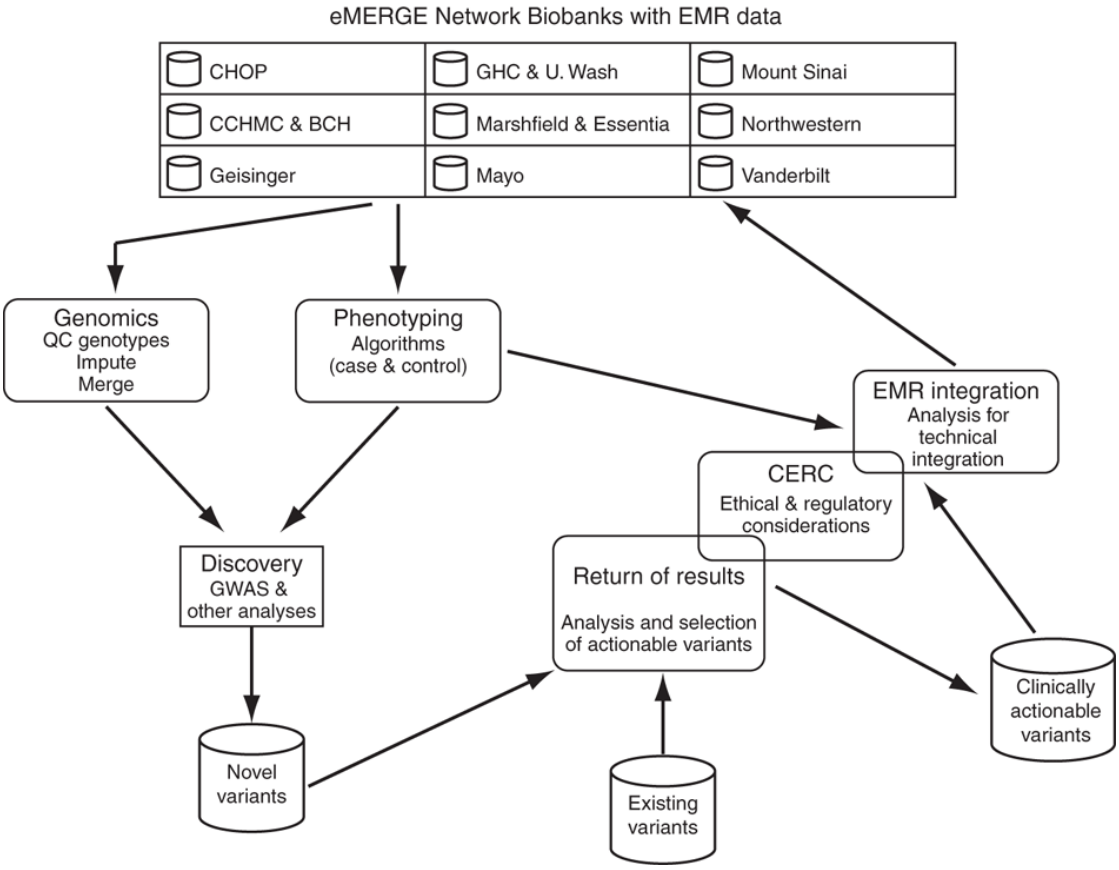


Figura 7.1: Modello della rete eMerge, con le attività svolte e le basi dati coinvolte (immagine da Genetics in Medicine (2013) 15, 761–771 doi:10.1038/gim.2013.72, su licenza Creative Commons)

## 7.2 Distribuzione geografica dell'infrastruttura

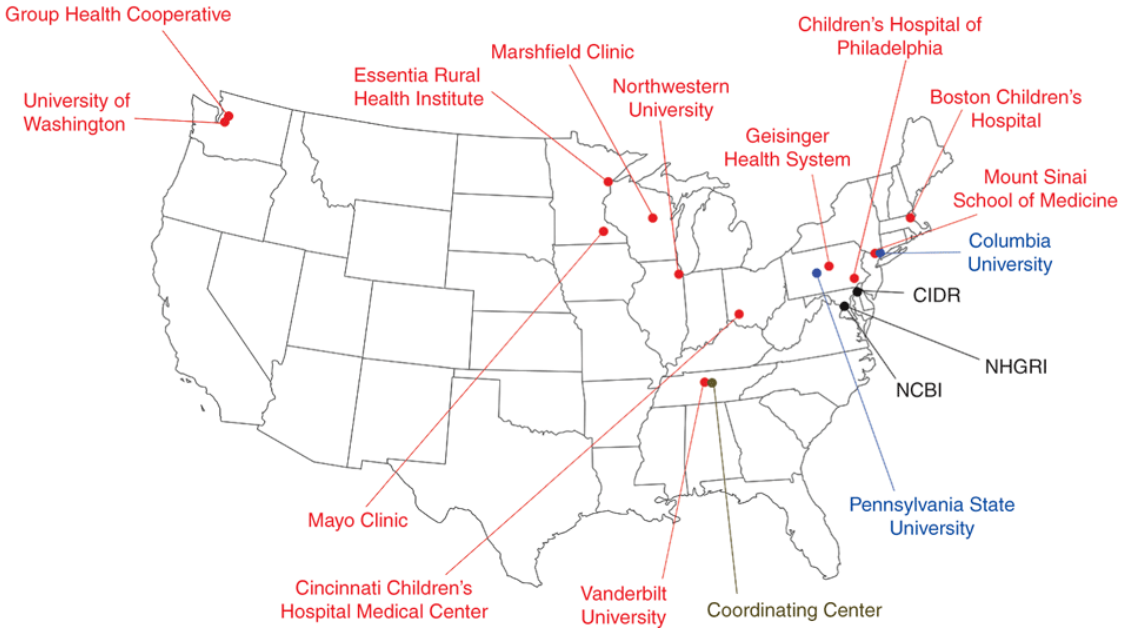


Figura 7.2: Localizzazione di infrastrutture, siti e servizi affiliati alla rete eMerge o a supporto di questa. In rosso sono indicati i 9 membri della rete eMerge-II, in grigio sono indicati i centri di coordinamento di eMerge, in blu i siti di affiliazione eMerge in nero i centri che forniscono servizi e supporto al progetto. (immagine da Genetics in Medicine (2013) 15, 761–771 doi:10.1038/gim.2013.72, su licenza Creative Commons)

## 7.3 Predisposizione per studi di associazione

L'analisi statistica di dati SNP per gli studi di associazione GWAS e EWAS tipicamente implica la gestione e l'integrazione di informazioni cliniche dei pazienti, compresi dati fenotipici, con gli SNP del genoma. I metodi attuali di memorizzazione di questi dati, su formati flat file, pongono diversi problemi nell'ambito GWAS, tra cui:

- Validazione dei dati dei pazienti e dei dati SNP.

- Problemi di performance nell'operare su grandi dataset.
- Necessità di aggiornare con precisione porzioni di dati che cambiano di frequente, come i dati clinici dei pazienti.
- Conversione dei dati per utilizzo in software di analisi

L'utilizzo di un database relazionale per gestire le sorgenti dati degli studi GWAS rappresenterebbe una soluzione per tutti e quattro i punti elencati. L'ultimo punto è il meno ovvio. Si consideri la semplice attività di convertire i dati GWAS, costituiti da informazioni genetiche e fenotipiche, in formati standard per strumenti software statistici, come PED/MAP o TPED/TFAM [1]. Un approccio comune per utilizzare questi dati sui diversi tool statistici è quello di estrarre i dati dei pazienti e gli SNP mediante uno script ad hoc, per produrre un secondo file compatibile con il software da utilizzare. Se si suppone ad esempio che alcuni di questi dati siano corrotti, ci si accorgerà del problema solo osservando i risultati del software statistico (ad esempio R), e questo sarà il primo momento in cui l'errore sarà rilevato. A questo livello sarà molto difficile risalire a quale sia il dato corrotto nel dato sorgente. Inoltre, data la varietà di macchine NGS esistenti e la varietà di formati compatibili di input per gli strumenti statistici, l'approccio ad hoc diventa molto complesso e molto fragile.

Un approccio migliore a questo problema è l'utilizzo di database relazionali. I database relazionali rappresenterebbero una soluzione diretta al problema della validazione dei dati mancanti o corrotti. Tuttavia, spesso questa soluzione viene evitata per questioni di performance. In un ambiente di ricerca GWAS si ha a che fare con aggregazioni di dati *Big Data*, e in questo contesto l'utilizzo di database relazionali pone molti limiti di prestazione sulle funzioni di analisi e sulle complesse manipolazioni di dati necessarie.

Nonostante ciò, è possibile progettare un modello dei dati *sharded* [Data et al., 2012], ovvero è possibile partizionare il database su più tabelle, facendo in modo ad esempio che record afferenti allo stesso fenotipo siano memorizzati nella stessa tabella.

[1] PED/MAP e TPED/TFAM sono formati per il software PLINK, uno strumento largamente utilizzato per l'analisi di dati genotipici. Vedere la sezione 6.3

Per aggirare le limitazioni di performance dei DBMS relazionali sono stati proposti approcci basati su sistemi NoSQL e tecnologie Map/Reduce (vedere capitolo 8.2).

Riguardo alla validazione dei dati, a differenza dell'approccio semplicistico basato su file, i DBMS validano i dati al momento del caricamento nella base dati. In questo modo, l'errore è identificato nei dati di origine prima che questi vengano elaborati.

Il problema della varietà nei numerosi formati di dati in origine (vedere capitolo 3) può essere facilmente risolto con la scrittura di una procedura di esportazione dei dati dal database. Lo stesso vale per i dati di origine delle sequenza, per i quali potranno essere messi a punto procedure ETL (Extract - Transform - Loading), mentre per quanto riguarda i dati clinici è preferibile utilizzare una piattaforma di data entry specifica.

## 7.4 Fenotipizzazione

L'analisi delle correlazioni fenotipiche delle mutazioni genetica è stata per lungo tempo un metodo essenziale per scoprire la funzione biologica dei geni. L'analisi fenotipica ha giocato un ruolo centrale nella mappatura di geni che sono causa di malattie.

## 7.5 Human Phenotype Ontology

Lo Human Phenotype Ontology (HPO) è uno strumento per permettere analisi computazionali del fenoma umano su larga scala. L'ontologia contiene attualmente oltre 9500 termini, ognuno dei quali descrive una particolare anomalia fenotipica. I termini sono organizzati come un grafo aciclico direzionato, e sono connessi tra loro mediante una relazione di "is-a", così che un termine rappresenti una caratterizzazione più specifica del termine padre. Esempio: "Anormalità dei piedi" is-a "Anormalità degli arti inferiori". Ogni termine nell'HPO descrive un'anormalità clinica. Questi termini possono essere generici, come "Anormalità del sistema muscoloscheletrico" o molto specifici come "Atrofia corioretinale".

Il modello dei dati per l'ontologia HPO è stato già descritto nel paragrafo 4.3. L'integrazione nel database progettato della banca dati HPO è la soluzione migliore per memorizzare e associare informazioni cliniche dei pazienti ai dati genomici.

## 7.6 Modello generico progettato

In figura 7.3 è illustrato lo schema del modello relazionale sviluppato a partire dai formati standard e dai database pubblici finora esaminati. Ai requisiti comuni di questi modelli conosciuti si sono aggiunti i requisiti studiati insieme ai ricercatori del Centro di Genetica Medica del Policlinico Sant'Orsola di Bologna.

Il modello relazionale supporta sia i dati genetici prodotti ad ogni passo della pipeline di analisi studiata nel capitolo 2, sia i dati clinici dei pazienti. Il modello dei dati dipende molto dal tipo di studio che si andrà fare, in questo caso è stato incentrato sugli studi di associazione genome-wide o exome-wide mediante genotipizzazione SNP. Tuttavia il modello supporta anche altre tipologie di varianti, mediante l'uso di una rappresentazione relazionale del formato VCF, mentre gli studi possibili restano limitati allo studio su sequenze nucleotidiche, escludendo altri tipi di sequenze biologiche (proteine, rna).

Bisogna tener presente che il modello presentato è limitato al solo scopo espositivo, restringendo di molto le potenzialità di un simile progetto. Si può considerare come il nucleo di un modello interoperabile, facilmente estendibile ad analisi ben più complesse. Basti considerare che la chiave primaria *rs* per le varianti SNP è un riferimento diretto al database SNPedia. SNPedia potrebbe essere facilmente importato, o interrogato via HTTP dal sistema. In questo modo sarà possibile risalire al codice del gene interessato dallo SNP, e da qui connettere le informazioni sulle malattie associate già conosciute (GeneOntology, OMIM, GeneCards), sul coinvolgimento nella produzione di proteine (UniProt, SwissProt), sui relativi effetti farmacologici (KEGG), sulle pubblicazioni in merito (PubMed). Anche con il solo modello qui illustrato, sarebbe possibile connettere in un grafo informazioni su più livelli di studio, similmente al progetto Biomine[Eronen and Toivonen, 2012],

con il vantaggio di avere a disposizione dati genetici e informazioni fenotipiche originali.

Il diagramma è suddiviso in due parti. Nel riquadro Genotipo con sfondo violetto sono rappresentate le entità che contengono informazioni genetiche su un individuo: sequenze di DNA, varianti, genotipizzazioni basate su SNP, dati di allineamento delle sequenze. Nella parte con sfondo verde etichettato con nome Fenotipo sono invece illustrate entità e relazioni per informazioni non genetiche sul paziente. L'entità *patient* rappresenta l'intersezione tra i due insiemi, e sarà utilizzata nelle interrogazioni per il join tra dati genetici e informazioni fenotipiche.

Ogni record genetico è un entità *genetic_ entry*, che può essere un insieme di varianti di diverso tipo (*VCF*), un allineamento di sequenze (*bam_ data*), una genotipizzazione sotto forma di alleli (*genotype_ SNP*) in posizioni definite (*variant_ SNP*), o semplicemente una sequenza di nucleotidi con gli eventuali valori di qualità.

L'entità *quality_ sec* è stata modellata in modo tale da garantire la flessibilità nell'espressione dei quality score. Come si è osservato dall'analisi dei diversi formati FASTQ (capitolo 3), i punteggi di qualità vengono rappresentati con codifiche, intervalli e calcoli di probabilità differenti. Per questo motivo si è cercato di rendere la rappresentazione su modello relazionale di questo dato compatibile con i formati studiati e flessibile per l'introduzione futura di nuovi standard.

Per quanto riguarda il fenotipo ogni descrizione medica è un insieme di termini predefiniti facenti parte di un vocabolario controllato di termini, l'entità *term*, estrapolati da un'ontologia come HPO (Human Phenotype Ontology). L'esito di un esame clinico (*medical_ exam*) può essere confermato o meno. Qualora un esame venga confermato viene associato al paziente una descrizione fenotipica *fenotype_ desc*.Ogni paziente (*patient*) può essere sottoposto a più esami clinici e accumulare nel tempo più descrizioni fenotipiche, ognuna delle quali riporterà la data della diagnosi.

La relazione *parent_ of* permette di rappresentare legami di parentela tra i soggetti sotto osservazione. Con delle semplici interrogazioni è possibile a partire dalla relazione genitore figlio estrapolare qualsiasi altra relazione di parentela tra

gli individui in database. Questa relazione è indispensabile nel condurre studi di associazione su malattie ereditarie e ancor di più nel produrre una diagnosi clinica, attraverso uno studio congiunto su dati genetici e clinici.

Di seguito sono mostrate delle interrogazioni di esempio. La prima interrogazione in figura 7.4 è la più semplice. Vengono selezionate le prime 100 sequenze di DNA ordinate per lunghezza e si restituiscono le informazioni del relativo record genetico con un solo join.

In figura 7.5 si mostra un'interrogazione per ottenere il numero di donne e il numero di uomini che hanno, tra le sequenze di DNA in database, una particolare sottosequenza, scelta casualmente in questo esempio. Da notare che tra l'entità record genetico *genetic_ entry* e i diversi record biologici, tra cui varianti, allineamenti e sequenze, esiste una relazione 1:1, e che ad ogni record biologico corrisponde un solo individuo. Per questo motivo nelle entità biologiche è inclusa la chiave esterna dell'individuo, così da ridurre il numero di join necessari per connettere i dati biologici alle informazioni sul fenotipo.

Un'interrogazione più complessa è descritta in figura 7.6. Per rendere l'esempio più realistico, è stata scelta una variante dalla collezione SNPedia. La variante catalogata come 'Rs1805007' è stato dimostrato essere coinvolta nel gene dei capelli rossi e nella sensibilità ad alcuni anestetici, utilizzati principalmente dai dentisti. L'allele di rischio è il 'T', mentre l'allele più comune e senza fattori di rischio è il 'C'. Nell'interrogazione si cerca la configurazione biallelica 'T;T', che nelle donne rappresenta il fattore di rischio più alto per la sensibilità agli anestetici. Nella query di esempio quindi si selezionano i pazienti donna con la configurazione biallelica 'T;T' per la variante in questione, e a solo scopo di esempio si ottengono le informazioni sul padre di queste pazienti.

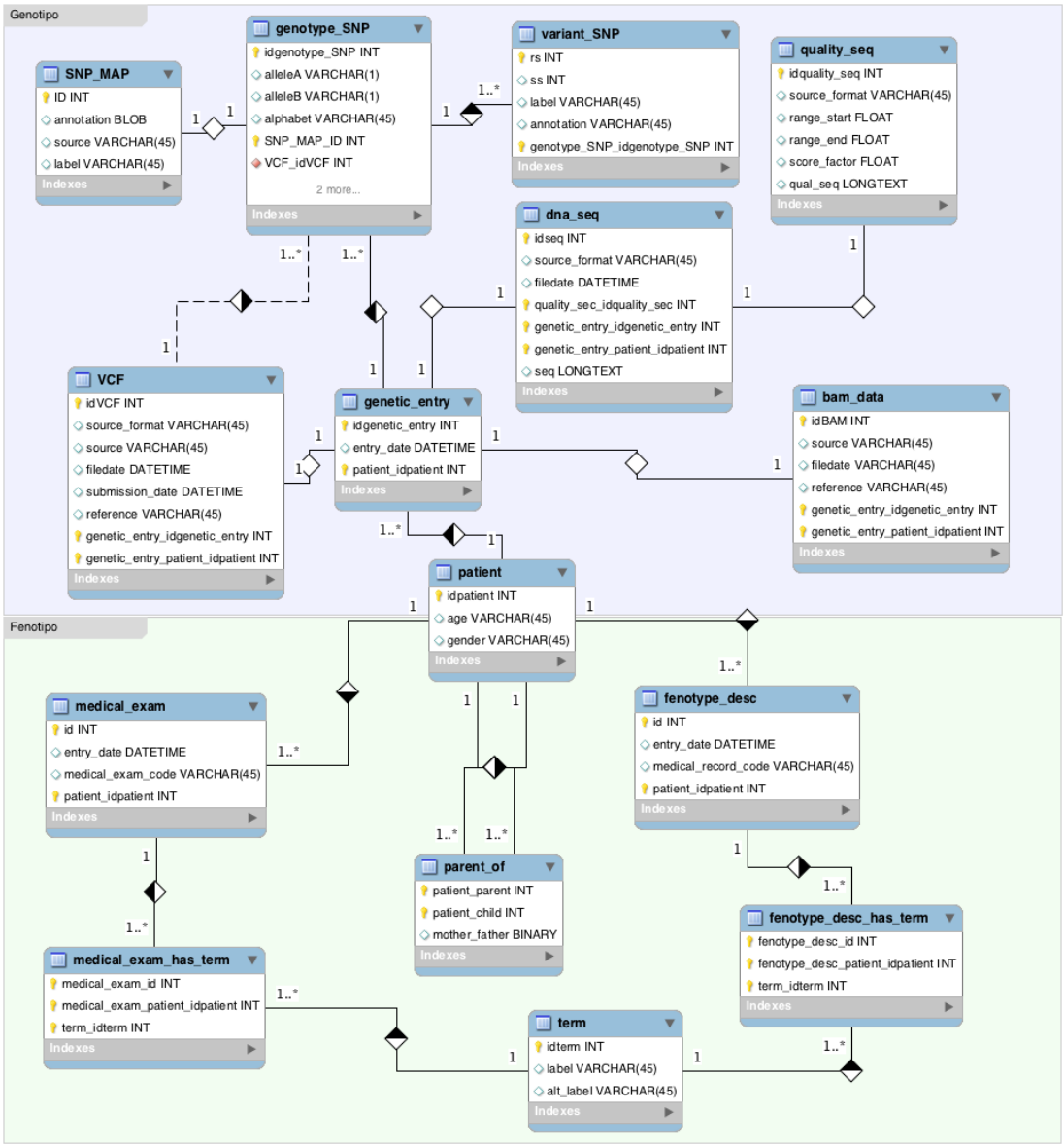


Figura 7.3: Schema astratto del modello relazionale per il database progettato. Per questioni di impaginazione e chiarezza sono state omesse alcune relazioni.

```
SELECT g.*
FROM genetic_entry g
JOIN dna_seq d ON g.id_genetic_entry = d.genetic_entry_idgenetic_entry
ORDER BY CHAR_LENGTH(d.seq)
LIMIT 100
```

Figura 7.4: Query di esempio sul modello relazionale proposto per ottenere informazioni sulle prime 100 sequenze più lunghe.

```
SELECT p.gender, COUNT(p.gender) as QUANTI
FROM patient p
JOIN dna_seq d ON p.idpatient = d.genetic_entry_patient_idpatient
WHERE d.seq LIKE "ATTCCGTAAAAAGATTACAAAAAAGATTACA"
GROUP BY gender
```

Figura 7.5: Query per contare quanti uomini e quante donne possiedono nel loro genoma la sottosequenza 'ATTCCGTAAAAAGATTACAAAAAAGATTACA'

```
SELECT p.*
FROM patient p
JOIN parent_of q ON p.idpatient = q.patient_parent
JOIN patient c ON c.idpatient = q.patient_child
JOIN genotype_SNP t ON c.idpatient = t.genetic_entry_patient_idpatient
JOIN variant_SNP v ON t.idgenotype_SNP = v.genotype_SNP_idgenotype_SNP
WHERE
variant_SNP.rs = 1805007 AND
genotype_SNP.alleleA = 'T'
genotype_SNP.alleleB = 'T'
c.gender = "FEMALE" AND
p.gender = "MALE"
```

Figura 7.6: Query più complessa per uno studio su una variante generica coinvolta nella sensibilità agli anestetici.

Il modello proposto permette di memorizzare dati eterogenei, eseguire interrogazioni più o meno complesse per ottenere informazioni di associazione tra genotipi, fenotipi ed ereditarietà delle varianti genetiche, e permette l'interoperabilità o l'integrazione, in base alle possibili implementazioni del sistema software di gestione e

analisi, con sorgenti dati esterne. Sebbene un sistema basato su database relazionale classico possa andar bene per un piccolo centro di ricerca genetica, condurre studi su larga scala diventerebbe impossibile in termini di tempo di esecuzione delle interrogazioni e capacità di storage. Per questo motivo, nel capitolo 8 si prenderanno in considerazione soluzioni Big Data al sistema in esame.

# 8 | I dati genetici come Big Data

Il termine **Big Data** è spesso associato alla gestione e all'analisi di dati prodotti da social network, dalla registrazione delle interazioni tra utenti e grandi portali web o da enormi quantità di testo. Tuttavia con l'avvento del Big Science, ovvero lo studio scientifico attraverso la registrazione di enormi quantità di dati e alla successiva analisi knowledge discovery, tecnologie e approcci BigData sono stati applicati a vari settori scientifici, a cominciare dagli studi sugli esperimenti LHC (*Large Hadron Collider*, grande collisore di adroni), per finire alle grandi collezioni di dati genomici. Alcuni framework per il mantenimento e l'analisi delle sequenze si basano già adesso su tecnologie MapReduce [McKenna et al., 2010, O'Connor et al., 2010], e l'attenzione verso il BigData nel mondo bioinformatico è in costante aumento, proporzionalmente alla crescita della quantità di dati biologici sequenziati. In tabella 8.1 sono elencati e descritti diversi progetti, relativamente recenti, che fanno uso di tecnologie BigData per la gestione, la manipolazione e l'analisi di dati genetici e per gli studi GWA.

Il concetto di Big Data associato ai database biologici e medici diventa ancora più realistico nell'idea di una sanità personalizzata. La sanità personalizzata, o medicina personalizzata, indica un approccio molto specifico alla prevenzione, diagnosi e terapia, attività che potranno essere studiate su misura per ogni individuo, mettendo in conto il suo patrimonio genetico (genoma), il suo fenotipo (fenoma), e informazioni sull'esposizione ambientale, i rapporti tra l'individuo e l'habitat (esposoma). Il mantenimento di una tale quantità di dati, rilevati in momenti diversi nella vita di un individuo e accumulati per lo studio in funzione del tempo, necessita di sistemi adatti alla gestione di Big Data e scalabili.

| Funzione | Algoritmo | Descrizione | Riferimento |
|---|---|---|---|
| Genomic sequence mapping | CloudAligner | Un'applicazione basata su MapReduce per mappare le letture corte generate dalle macchine NGS. | [Nguyen et al., 2011] |
| | CloudBurst | Un algoritmo parallelo per mappare sequenze NGS a genomi di riferimento. | [Schatz, 2009] |
| | SEAL | Un kit di applicazioni per l'allineamento, la manipolazione e l'analisi di sequenze corte di DNA. | [Pireddu et al., 2011] |
| | BlastReduce | Un algoritmo di mapping per l'allineamento parallelo ottimizzato per la scoperta di SNP, la genotipizzazione e la genomica personale. | [BlastReduce, ] |
| Genomic sequencing analysis | Crossbow | Una pipeline di software che combina gli algoritmi di Bowtie e SoapSNP per il risequenziamento dell'intero genoma. | [Langmead et al., 2009] |
| | Contrail | Un algoritmo per l'assemblaggio de novo, ovvero senza genoma di riferimento, di letture corte basta sui grafi di de Bruijin. | [Schatz et al., 2010] |
| | CloudBrush | Un assemblatore distribuito. | [Chang et al., 2012] |
| RNA sequence analysis | Myrna | Una pipeline in cloud per calcolare l'espressione dei geni in grandi dataset di sequenze RNA. | [Langmead et al., 2010] |
| | FX | Uno strumento per la stima dei livelli di espressione genetica e delle varianti genomiche. | [Hong et al., 2012] |
| | Eoulsan | Una soluzione flessibile e integrata per l'analisi di sequenze RNA. | [Jourdren et al., 2012] |
| Sequence file management | Hadoop-BAM | Una libreria scalabile per la manipolazione di sequenze NGS allineate. | [Niemenmaa et al., 2012] |
| | SeqWare | Un insieme di strumenti di analisi per sequenze NGS con una base dati HBase. | [O'Connor et al., 2010] |
| | GATK | A gene analysis tool-kit for next-generation resequencing data. | [McKenna et al., 2010] |
| GPU bioinformatics software | GPU-BLAST | Una versione accelerata dell'algoritmo BLAST, che utilizza le GPU in cloud per l'allineamento di sequenze. | [Vouzis and Sahinidis, 2011] |
| | SOAP3 | Un algoritmo di allineamento per sequenze corte che utilizza schede grafiche multi-processore. | [Liu et al., 2012] |
| Search engine implementation | Hydra | Un motore di ricerca MapReduce per database di sequenze proteomiche. | [Lewis et al., 2012] |
| | CloudBlast | Implementazione BLAST scalabile su cloud. | |
| Miscellaneous | BioDoop | Un insieme di strumenti per la gestione di sequenze FASTA e per la conversione di sequenze. | [Leo et al., 2009] |
| | BlueSNP | Un algoritmo su R per le analisi computazionalmente intensive su grandi dataset integrati di genotipi e fenotipi. | [Huang et al., 2013] |
| | Quake | Un'applicazione per il rilevamento e la correzione di errori nelle sequenze di DNA. | [Kelley et al., 2010] |
| | YunBe | Un algoritmo su cloud per l'identificazione di biomarcatori in un insieme di geni | [Zhang et al., 2012] |

Tabella 8.1: Rassegna completa delle implementazioni di software e database genetici su tecnologie BigData.

## 8.1 Hadoop

Hadoop è un insieme di framework open source che supporta applicazioni distribuite data-intensive. Hadoop è basato sui precedenti lavori di Google ovvero GFS (Google File System), MapReduce e BigTable.

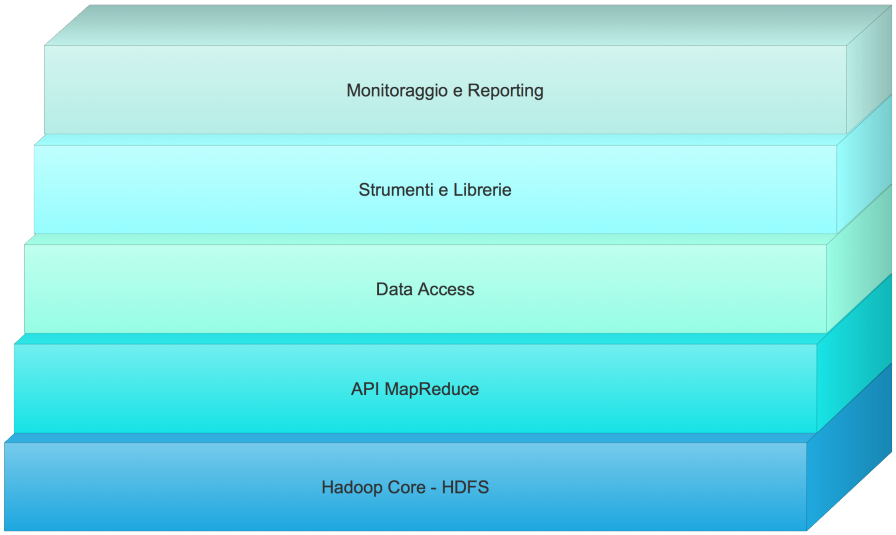


Figura 8.1: L'insieme di framework, API e strumenti di Hadoop può essere rappresentato in un architettura a layer.

Con Hadoop è possibile realizzare un cluster BigData, scalabile in modo arbitrario, senza preoccuparsi dell'affidabilità e dello spostamento dei dati all'interno del cloud, in quanto gestiti in modo trasparente dal framework. Il framework include un motore MapReduce (MapReduce oppure YARN) e un file system distribuito, HDFS (vedere fig. 8.1). Esistono diverse distribuzioni di Hadoop, alcune di queste open source:

- Apache Hadoop
- Cloudera
- Hortonworks
- MapR

- Amazon AWS
- Windows Azure HDInsight

In figura 8.2 è mostrata l'architettura del cluster Hadoop e le interazioni tra i suoi componenti. Un utente generico richiede l'esecuzione di un job MapReduce, che può essere una classe java che implementa le funzioni Map e Reduce oppure una query SQL in Hive. Il JobTracker, il componente Hadoop che gestisce in batch le operazioni sui dati richieste, suddivide il lavoro tra i nodi del cluster. In ogni nodo è presente un tracker locale, che divide il lavoro del nodo in più task, in base alla definizione della funzione di mapping. I componenti HDFS, in verde, gestiscono la divisione dei dati in blocchi, la sincronizzazione e l'affidabilità dei dati distribuiti, lo spostamento dei dati tra diversi nodi. La gestione è affidata a un nodo gestore virtuale, il NameNode, che può trovarsi in uno dei DataNode oppure su un nodo a sè.

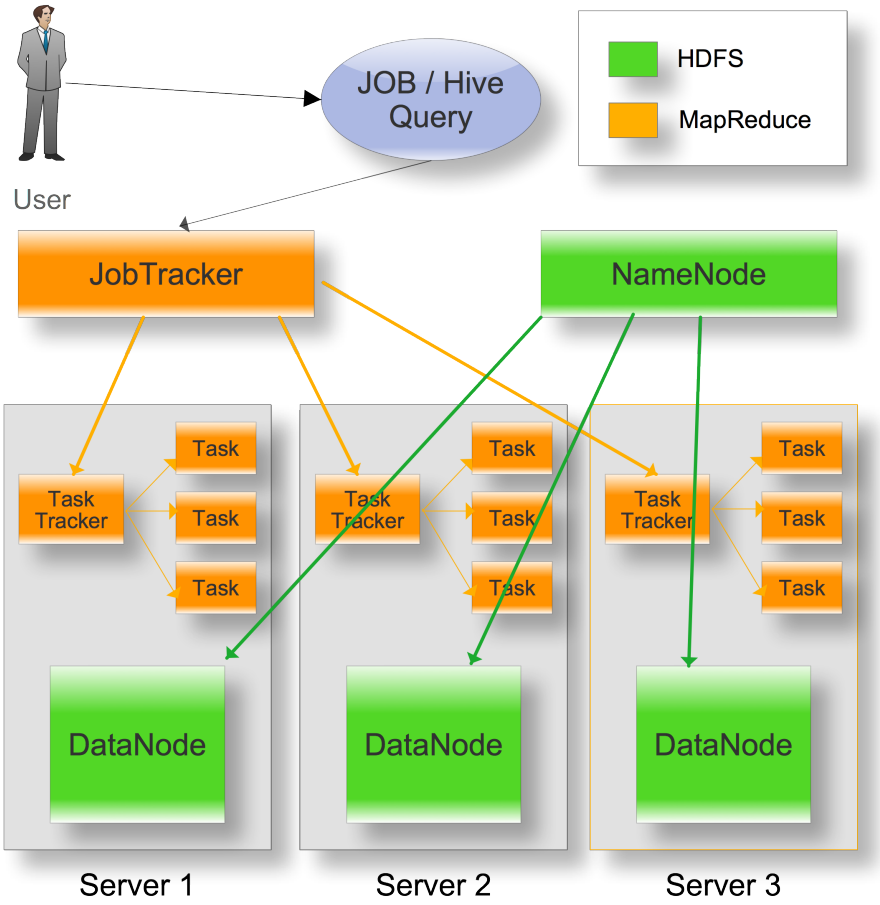


Figura 8.2: L'architettura dei nodi Hadoop, in arancione i componenti del framework MapReduce, in verde i componenti gestiti da HDFS.

## 8.2 MapReduce

MapReduce è un modello di programmazione per l'elaborazione di data set estremamente grandi, inizialmente sviluppato da Google agli inizi del 2000 per ottenere scalabilità nel sistema di ricerca [Dean and Ghemawat, 2008]. MapReduce si basa su principi di elaborazione parallela e distribuita senza dipendere da un database in particolare. La flessibilità di MapReduce sta nella sua capacità di processare elaborazioni distribuite su una grande quantità di dati in cluster. Le caratteristiche

principali di MapReduce sono:

- Parallelizzazione automatica
- Distribuzione dei dati automatica
- Tolleranza ai guasti
- Estensibilità
- Flessibilità nel linguaggio di programmazione
- Strumenti per monitorare lo stato del sistema

Come accennato, lo scopo di una funzione distribuita MapReduce è in genere quello di eseguire un'operazione, di complessità arbitraria, su una vasta quantità di dati, spesso misurabile in termini di PetaByte. Per questa ragione i task MapReduce non sono eseguiti istantaneamente, vengono invece pianificati temporalmente da Hadoop come *jobs* in *batch*. La coda di jobs può essere monitorata, così come lo stato di ogni job. Quando un job termina la sua esecuzione viene prodotto, oltre a una directory contenente i risultati prodotti, un report dettagliato della sua esecuzione, comprendente il numero di task Map e Reduce eseguiti, i tempi di esecuzione, il numero di dati prodotti e ricombinati ad ogni stadio intermedio.

I dati di input, uniformemente distribuiti tra i nodi del cluster Hadoop, vengono dapprima elaborati come task locali in ogni nodo, successivamente nella fase reduce i risultati intermedi vengono ricombinati e viene prodotto il risultato su un nodo master. L'architettura distribuita di MapReduce e il file system di base HDFS rendono possibile la scalabilità orizzontale, un requisito fondamentale per il sistema in esame, per il quale i dati sono sparsi in più nodi geograficamente distanti.

Il paradigma MapReduce, dividendo il lavoro totale in sotto-lavori più piccoli, uno per ogni nodo, rende possibile l'elaborazione distribuita e soprattutto riduce enormemente lo spostamento di dati tra i nodi dell'infrastruttura.

La tipologia di operazioni programmata nel modello MapReduce rassomiglia per molti versi a quella delle interrogazioni SQL. Per questo motivo sono stati prodotti diversi framework per l'interrogazione diretta di dati su Hadoop mediante del-

le query SQL-like[1]. La soluzione più utilizzata per interrogare i dati con un linguaggio simile a SQL è l'infrastruttura di data warehouse Hive [Thusoo et al., 2010]. Una query SQL eseguita su Hive viene automaticamente tradotta in un task Hadoop, e la sua esecuzione viene distribuita sui nodi del cluster Hadoop seguendo il paradigma MapReduce.

### 8.2.1 Modello di programmazione MapReduce

MapReduce è basato su un paradigma di programmazione funzionale largamente ispirato al modello Lisp. Tipicamente, il programmatore deve implementare due funzioni:

- Map (in_key, in_value) -> (out_key, intermediate_value) list
  - La funzione `Map` riceverà in input delle coppie di chiavi e valori, e successivamente ai cicli di elaborazione produrrà un insieme intermedio di coppie chiave-valore.
  - Funzioni di libreria saranno poi usate per raggruppare insieme tutti i valori associati ad una chiave intermedia K, per poi passarli alla funzione Reduce.
- Reduce (out_key, intermediate_value list) -> out_value list
  - La funzione `Reduce` accetta in input una chiave intermedia K e l'insieme di valori per la chiave.
  - La funzione unisce insieme questi valori per formare un insieme di valori possibilmente più ridotto.
  - Il riduttore restituisce uno o nessun valore per ogni invocazione.
  - I valori intermedi sono forniti alla funzione Reduce mediante un iteratore. La funzione iteratore permette di gestire una grande lista di valori che non potrebbero mai entrare in memoria in un solo passo.

[1]Il linguaggio delle interrogazioni in Hive non rispetta completamente le specifiche ANSI SQL

Generalizzando il funzionamento del paradigma MapReduce, è possibile rappresentarlo graficamente come in figura 8.3. I cerchi in azzurro sono i nodi del cluster Hadoop. Ogni rettangolo verticale indica un dato nella sotto forma di coppia *(key,value)*, in cui il colore indica la chiave del dato. Rettangoli con lo stesso colore rappresentano dati con la stessa chiave. Le fasi dell'operazione MapReduce sono indicate in alto. La funzione di shuffle non deve essere implementata dall'utente, è uno stadio intermedio in cui i dati provenienti dai diversi nodi vengono ricombinati insieme per chiave.

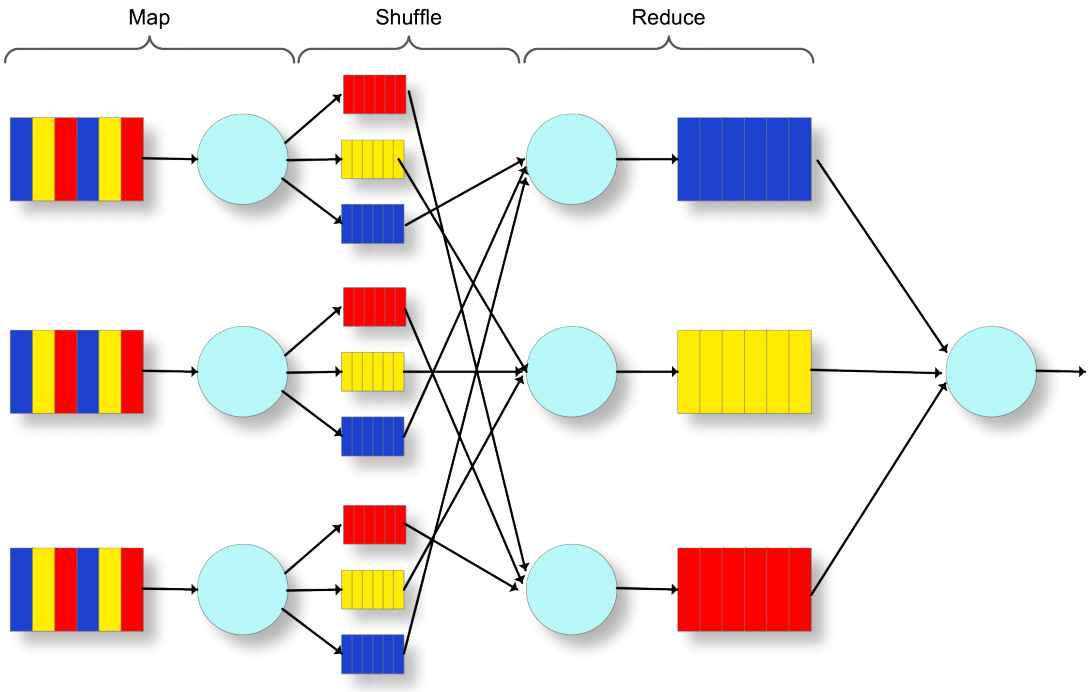


Figura 8.3: Astrazione del processo MapReduce.

### Esempio

Prima di applicare il paradigma di programmazione MapReduce alle funzioni GWAS è bene mostrarne il funzionamento mediante un esempio tipico. Si consideri il problema di contare il numero di occorrenze di ogni parola in una grande quantità di documenti. Questo genere di problema rappresenta per MapReduce l'equivalente di un Hello World, in quanto è l'applicazione più semplice del para-

digma, ma allo stesso tempo svolge una funzione utile nel concreto. Il problema inoltre può essere ricondotto con facilità ad alcune funzioni statistiche del GWAS, in cui i dati sono estratti da esperimenti diversi e ricombinati in base ad attributi comuni, come un particolare fenotipo o delle varianti SNP.

In pseudocodice MapReduce può essere espresso come segue:

```
map(String key, String value):
// key:  nome documento
// value:  contenuto documento
for each word w in value:
EmitIntermediate(w, "1");

reduce(String key, Iterator values):
// key:  a word
// values:  a list of counts
int result = 0;
for each v in values:
result += ParseInt(v);
Emit(AsString(result));
```

La funzione Map così definita emette ogni diversa parola trovata nel testo, insieme a un valore, in questo caso il valore costante '1'. Successivamente, nel processo di *shuffle* gestito direttamente dal framework MapReduce, queste coppie *(Chiave,Valore)* vengono ricombinate in modo da raggruppare insieme le coppie con la stessa chiave in una lista. Infine, la funzione Reduce prende le liste di valori e le somma, producendo delle coppie chiave valore in cui la chiave è la parola e il valore è la somma di tutte le occorrenze in tutti i testi.

Per chiarire meglio il concetto, si è illustrato in figura 8.4 il processo di MapReduce del codice discusso. In figura sono mostrati i dati di input e i risultati intermedi soltando per due nodi, il nodo 1 e il nodo 3. Nella fase, quella di mapping, ogni nodo esegue un task in cui produce per ogni parola trovata nel testo una coppia *(parola,1)*. Durante la ricombinazione locale, ogni nodo unisce insieme le coppie con la stessa chiave in una sola coppia, incrementando il valore a ogni

ricombinazione. Il risultato intermedio di questa fase è che ogni nodo ha risolto il suo sottoproblema, ovvero contare le occorrenze di ogni parola per i dati locali. Nella fase di shuffle questi risultati intermedi sono uniti tra loro in base alla chiave. Questa fase produce un insieme di coppie *(parola, [x,y,z])* in cui la chiave è una parola e il valore è una lista contenente il conteggio di occorrenze della parola per ogni nodo. Nella fase di riduzione gli elementi delle liste vengono sommati tra loro, producendo un insieme di coppie *(parola,conteggio)* come risultato finale.

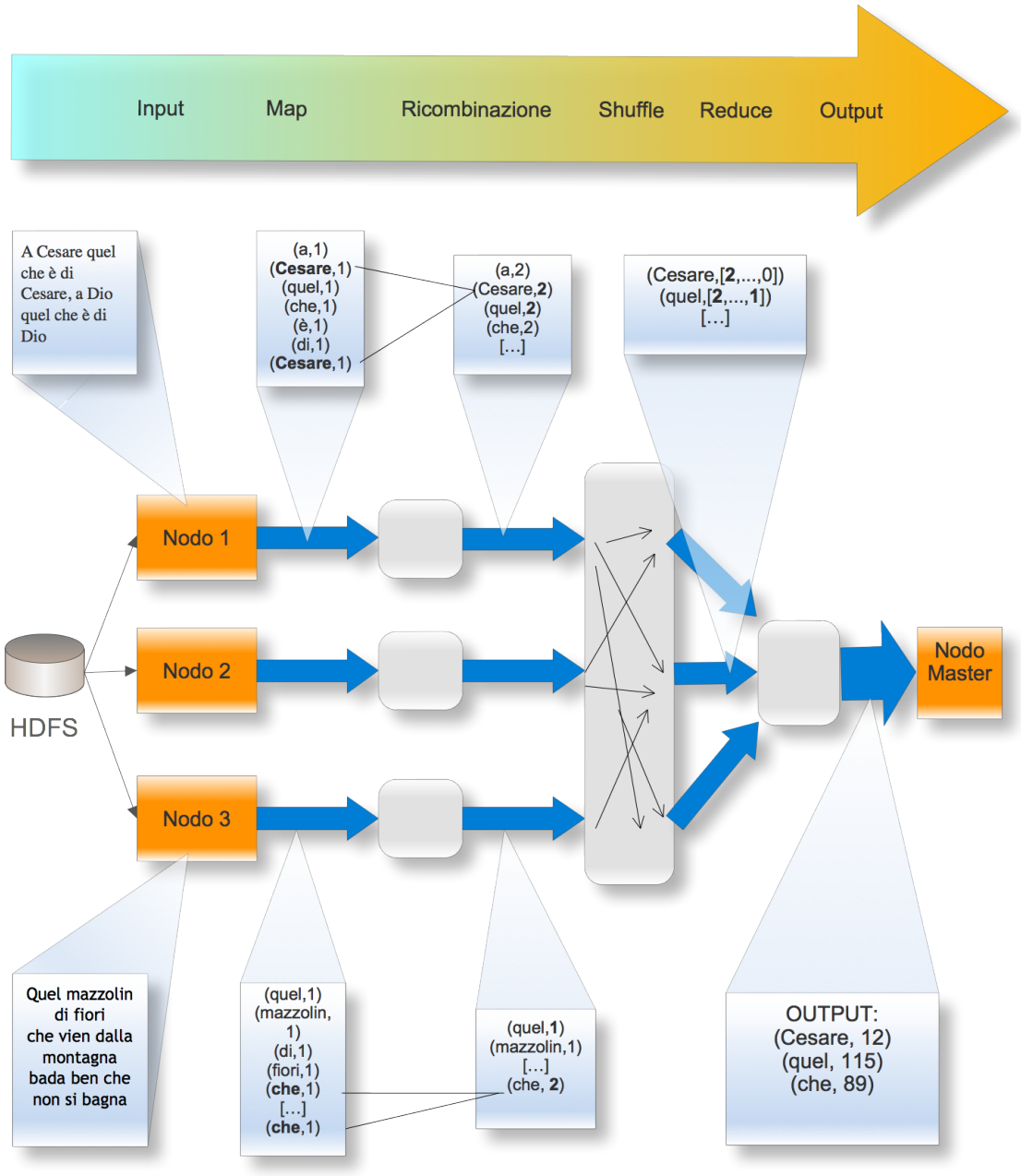


Figura 8.4: Le fasi di un job MapReduce per il codice di esempio di un algoritmo Word Count.

| Nodo 1 | | Nodo 2 | | Nodo 3 | |
|---|---|---|---|---|---|
| Chiave | Valore | Chiave | Valore | Chiave | Valore |
| SNP1 | A | SNP1 | A | SNP1 | A |
| SNP2 | T | SNP2 | A | SNP2 | A |
| SNP3 | C | SNP3 | C | SNP3 | C |
| SNP1 | G | SNP1 | G | SNP1 | A |
| SNP2 | T | SNP2 | T | SNP2 | T |
| SNP3 | C | SNP3 | C | SNP3 | C |

Tabella 8.2: Coppie chiave-valore di esempio per uno studio di associazione su MapReduce

## 8.3 Applicazione di MapReduce agli studi di associazione in esame

Le analisi statistiche degli studi di associazione tra varianti genetiche e caratteristiche fenotipiche hanno molto in comune con le operazioni tipicamente eseguite su un sistema MapReduce. Nel database che si è modellato nel capitolo 7 è stato associato il fenotipo di ogni individuo alle varianti della sua sequenza genetica. Supponiamo di voler analizzare l'insieme di SNP comuni ad un particolare fenotipo. Si assume di aver già a disposizione un sottoinsieme di individui che condividono lo stesso fenotipo, ad esempio ottenuto con una query in Hive sul database. Questo sottoinsieme può essere considerato come il dataset di record di un file PED (vedere capitolo 6), in cui si ha la lista degli alleli per tutti gli SNP del genoma (o dell'esoma). Possiamo rappresentare questo dataset come coppie chiave-valore come in tabella 8.2

In figura 8.5 è illustrato il processo MapReduce per risolvere il problema appena descritto. Al contrario dell'esempio precedente di conteggio delle parole, in questo caso il valore per ogni chiave non è l'unità numerica, ma una delle quattro costanti che compongono l'alfabeto del DNA {A,C,T,G}. Nella fase Map i dati genotipici in input vengono uniti insieme per SNP uguali, ottenendo delle liste di alleli per lo

stesso SNP. Successivamente avviene un altro mapping in cascata, non mostrato nella figura per chiarezza, in cui avviene il conteggio di occorrenze dello stesso allele per ogni SNP. Nella fase di Shuffle le coppie con lo stesso SNP (stessa chiave) vengono ricombinati e spostati sullo stesso nodo. Infine nella fase Reduce avviene il conteggio finale. In questo modo nell'output si potrà osservare qual'è l'aplotipo più comune per gli individui in esame, che ricordiamo condividere lo stesso fenotipo.

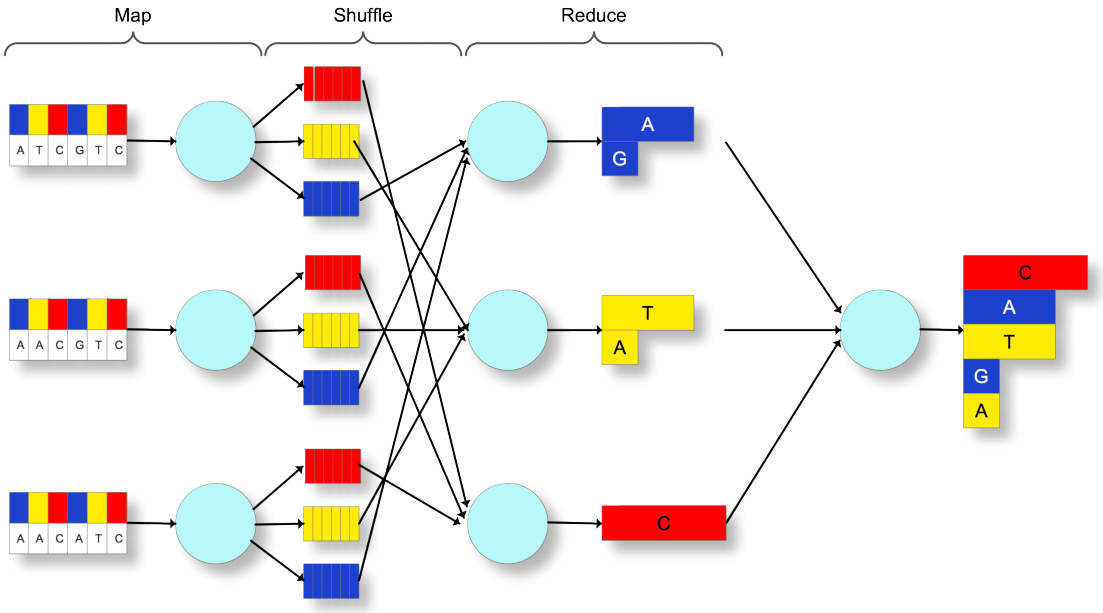


Figura 8.5: Esempio di applicazione del processo MapReduce ad una analisi statistica delle varianti genetiche.

Intuitivamente questa operazione non ha una valenza statistica in quanto i dati e le variabili da tenere in considerazione sono molteplici, tra cui l'insieme dei genotipi per i quali il fenotipo in esame non è presente. La falla più evidente è che gli alleli in comune risultanti da questa operazione potrebbero essere condivisi anche da individui non affetti dal fenotipo. Tuttavia questo task potrebbe far parte di una serie di operazioni in uno studio GWAS reale, e l'esecuzione di queste operazioni su framework MapReduce riducono significativamente i tempi necessari al loro completamento (vedi fig. 8.6).

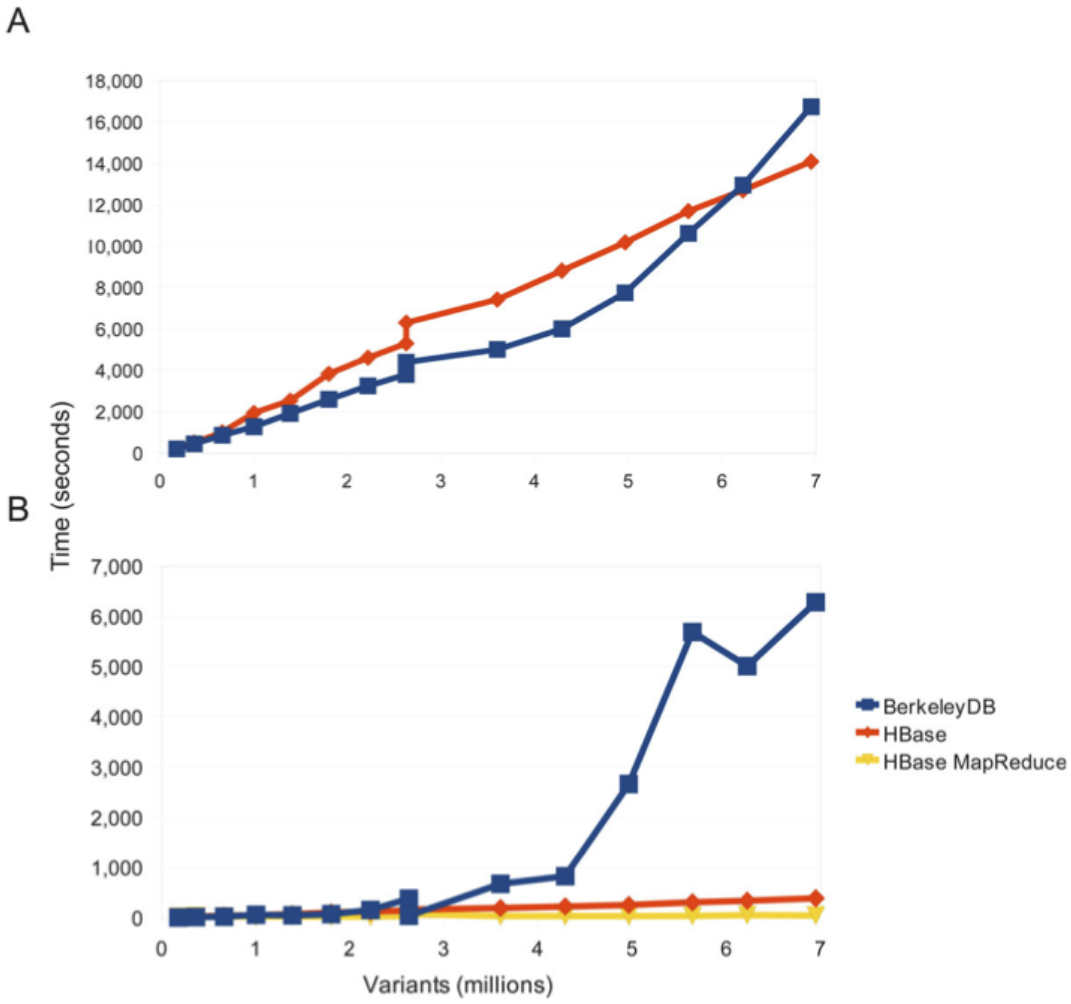


Figura 8.6: Confronto di prestazioni per le operazioni di caricamento (A) e interrogazione (B) per i database BerkleyDB, HBase e HBase Mapreduce

# 9 | Conclusioni

Partendo dallo studio dei formati dei file biologici per finire ai database biologici Big Data, la ricerca approfondita sui dati e sui database genetici ha portato a una chiara consapevolezza dei requisiti per un sistema innovativo e dei principi cardine che dovrebbero guidare la sua progettazione. L'interesse per la genomica, considerata tra i temi più caldi nella comunità scientifica, ha contribuito negli ultimi anni a creare un mondo caotico e in continua evoluzione. Questo aspetto, sebbene sia di per sé una complicazione, rappresenta un'opportunità, per l'industria e la ricerca informatica, di definire degli standard e di sviluppare dei sistemi sulla base di queste mancanze.

Per far fronte alle difficoltà create dall'esistenza di formati incompatibili e semistrutturati dei dati biologici, si è proposto un modello relazionale. E' importante, per garantire continuità nella ricerca e negli strumenti utilizzati, mantenere la compatibilità con formati divenuti nel corso degli anni degli standard de-facto. Un dettaglio da considerare, nella visione futura di una standardizzazione dei dati biologici, è che uno modello standard in database non può bastare, visto che a generare i dati biologici sono strumenti hardware che finora possono soltanto produrre in output dei file su disco.Un altro passo in questa direzione quindi potrà essere la definizione di uno standard di file universale (ad esempio uno schema XML) per realizzare questa compatibilità a partire dai livelli più bassi del flusso di analisi.

Al termine di questo studio approfondito sui GWAS risulta evidente la necessità di integrare i dati fenotipici mediante l'uso di ontologie. Com'è ovvio non basterà la predisposizione di un database ma dovrà essere sviluppato un sistema di data

entry, progettato in modo tale da permettere l'inserimento veloce dei dati clinici [Chiang et al., 2003, Cole et al., 2006]. Potrebbe inoltre essere necessario digitalizzare, con tecnologia OCR, i dati cartacei finora prodotti. Sebbene in questa tesi ci si sia soffermati esclusivamente su analisi di associazione basate sulla semplice intuizione di mettere in relazione, statisticamente, varianti genetiche comuni a malattie comuni, nel corso degli anni molti studi più complessi sono stati proposti, molti dei quali coinvolgono l'interazione tra geni nella produzione di proteine e altri processi biologici, spostandosi dalla genomica alla *metabolomica*. L'analisi GWAS così proposta resta la più diffusa e discussa per il semplice fatto che sia quasi totalmente automatizzabile, tuttavia dopo 6 anni di studi genomici sono in molti ad esprimere le loro perplessità sulla validità dell'ipotesi *varianti comuni = malattie comuni* [Visscher et al., 2012]. Come già accennato, un'estensione del modello di base proposto potrebbe considerare l'integrazione dei dati privati con i dataset pubblici, nell'intento di realizzare analisi ben più complesse. Queste analisi potranno basarsi sulla connessione automatica in un grafo di tutte le informazioni correlate, come nel progetto Biomine [Eronen and Toivonen, 2012], o saranno semplicemente un supporto alla ricerca manuale di un operatore, sia negli studi di associazione sia nella diagnosi. L'integrazione potrà essere locale, importando nel database locale gli schemi e i dati pubblici, come nel caso Biomine, o realizzata sfruttando l'interoperabilità dei repository pubblici, quasi tutti provvisti di interfaccia webservice, DAS o RDF, interrogabile in SPARQL.

Valutando il volume di dati, di dimensioni notevoli e in continua crescita, e l'interesse dimostrato dalla comunità bioinformatica nei confronti del BigData, si ritengono queste tecnologie opportune per il sistema proposto. L'approccio BigData, conveniente per i costi di storage e per l'efficienza delle analisi distribuite, potrebbe diventare indispensabile nei prossimi anni, vista la crescita esponenziale nella produzione di dati biologici. Se in futuro saranno sviluppati dei formati file standard biologici, in una prospettiva BigData questi potranno essere direttamente memorizzati senza la trasformazione e l'importazione in un modello relazionale, sfruttando le peculiarità del file system HDFS e dell'approccio non relazionale NoSQL.

# Bibliografia

[Epp, ] MGI-Guidelines for Nomenclature of Genes, Genetic Markers, Alleles, & Mutations in Mouse & Rat.

[Abecasis et al., 2012] Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[Barrett and Cardon, 2006] Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nature Genetics*, 38:659–662.

[Bennett, 2004] Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, 5:433–438.

[Bentley et al., 2008] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A.,

Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59.

[Biesecker, 2010] Biesecker, L. G. (2010). Exome sequencing makes medical genomics a reality. *Nature Genetics*, 42:13–4.

[BlastReduce, ] BlastReduce, . Blastreduce: high performance short read mapping

with mapreduce. .

[Botstein and Risch, 2003] Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33:228–237.

[Brendel et al., 1986] Brendel, V., Beckmann, J. S., and Trifonov, E. N. (1986). Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of biomolecular structure & dynamics*, 4(1):11–21.

[Broad Institute, 2010, ] Broad Institute, 2010. Integrative Genomics Viewer. `http://www.broadinstitute.org/igv`, note = Acceduto il: 9/09/2013.

[Cariaso and Lennon, 2012] Cariaso, M. and Lennon, G. (2012). SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Research*, 40:D1308–12.

[Chang et al., 2012] Chang, Y.-J., Chen, C.-C., Chen, C.-L., and Ho, J.-M. (2012). A de novo next generation genomic sequence assembler based on string graph and MapReduce cloud computing framework. *BMC genomics*, 13 Suppl 7:S28.

[Chiang et al., 2003] Chiang, M. F., Cao, H., Sharda, P., Hripcsak, G., and Starren, J. B. (2003). An Experimental System for Comparing Speed, Accuracy, and Completeness of Physician Data Entry using Electronic and Paper Methods. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, 2003:812.

[Cole et al., 2006] Cole, E., Pisano, E. D., Clary, G. J., Zeng, D., Koomen, M., Kuzmiak, C. M., Seo, B. K., Lee, Y., and Pavic, D. (2006). A comparative study of mobile electronic data entry systems for clinical trials data collection. *International Journal of Medical Informatics*, 75:722–729.

[Collado-Vides, 1989] Collado-Vides, J. (1989). A transformational-grammar approach to the study of the regulation of gene expression. *Journal of theoretical biology*, 136(4):403–25.

[Damasevivcius, 2010] Damasevivcius, R. (2010). Structural analysis of regulatory DNA sequences using grammar inference and Support Vector Machine. *Neurocomputing*, 73(4-6):633–638.

[Danecek et al., 2011] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158.

[Data et al., 2012] Data, A. L. L., Related, I. S., Appeal, T. H. E., Nosql, O. F., Relational, T. H. E., and Data, B. I. G. (2012). Scaling Big Data With Your Relational Database. *Database Trends Applications*, 26:p24–24.

[Dean and Ghemawat, 2008] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

[Dear and Staden, 1992] Dear, S. and Staden, R. (1992). A standard file format for data from DNA sequencing instruments. *DNA sequence : the journal of DNA sequencing and mapping*, 3(2):107–10.

[Ebeling and Jimenez-Montano, 1980] Ebeling, W. and Jimenez-Montano, M. A. (1980). On grammars, complexity, and information measures of biological macromolecules. *Mathematical Biosciences*, 52(1):53–71.

[Eronen and Toivonen, 2012] Eronen, L. M. and Toivonen, H. T. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, 13:119.

[Ewing and Green, 1998] Ewing, B. and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research*, 8:186–194.

[Ewing et al., 1998] Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8:175–185.

[Fernau, 2003] Fernau, H. (2003). Parallel Grammars: A Phenomenology. *Grammars*, 6(1):25–87.

[Fitch and Sokhansanj, 2000] Fitch, J. and Sokhansanj, B. (2000). Genomic engineering: moving beyond DNA sequence to function. *Proceedings of the IEEE*, 88(12):1949–1971.

[Frazer et al., 2007] Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., De Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Altshuler, D., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N.,

Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Tsunoda, T., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Zeng, C., Zhao, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Gibbs, R. A., Belmont, J. W., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Wheeler, D. A., Yakub, I., Gabriel, S. B., Onofrio, R. C., Richter, D. J., Ziaugra, L., Birren, B. W., Daly, M. J., Altshuler, D., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861.

[GATK, 2013, 2013] GATK, 2013 (2013). The Genome Analysis Toolkit. `http://www.broadinstitute.org/gsa/wiki/index.php`. Acceduto il: 7/10/2013.

[Gottesman et al., 2013] Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. a., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. a., Brilliant, M., Carey, D. J., Chisholm, R. L., Chute, C. G., Connolly, J. J., Crosslin, D., Denny, J. C., Gallego, C. J., Haines, J. L., Hakonarson, H., Harley, J., Jarvik, G. P., Kohane, I., Kullo, I. J., Larson, E. B., McCarty, C., Ritchie, M. D., Roden, D. M., Smith, M. E., Böttinger, E. P., and Williams, M. S. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(10):761–771.

[Hamosh et al., 2005] Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33:D514–D517.

[Hedges et al., 2009] Hedges, D. J., Burges, D., Powell, E., Almonte, C., Huang, J., Young, S., Boese, B., Schmidt, M., Pericak-Vance, M. A., Martin, E., Zhang, X., Harkins, T. T., and Züchner, S. (2009). Exome Sequencing of a Multigenerational Human Pedigree. *PLoS ONE*, 4:8.

[Hirschhorn and Daly, 2005] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature reviews. Genetics*, 6(2):95–108.

[Hong et al., 2012] Hong, D., Rhie, A., Park, S.-S., Lee, J., Ju, Y. S., Kim, S., Yu, S.-B., Bleazard, T., Park, H.-S., Rhee, H., Chong, H., Yang, K.-S., Lee, Y.-S., Kim, I.-H., Lee, J. S., Kim, J.-I., and Seo, J.-S. (2012). FX: an RNA-Seq analysis tool on the cloud. *Bioinformatics (Oxford, England)*, 28:721–3.

[Horner et al., 2010] Horner, D. S., Pavesi, G., Castrignanò, T., De Meo, P. D., Liuni, S., Sammeth, M., Picardi, E., and Pesole, G. (2010). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11:181–97.

[HPO, 2013, ] HPO, 2013. HPO hpo.

[Huang et al., 2013] Huang, H., Tata, S., and Prill, R. J. (2013). BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters. *Bioinformatics (Oxford, England)*, 29:135–6.

[Irina Astrova, 2007] Irina Astrova, Nahum Korda, A. K. (2007). Storing OWL Ontologies in SQL Relational Databases.

[Jimenez-Montano, 1984] Jimenez-Montano, M. A. (1984). On the syntactic structure of protein sequences and the concept of grammar complexity. *Bulletin of Mathematical Biology*, 46(4):641–659.

[Jourdren et al., 2012] Jourdren, L., Bernard, M., Dillies, M.-A., and Le Crom, S. (2012). Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. *Bioinformatics (Oxford, England)*, 28:1542–3.

[Kelley et al., 2010] Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11:R116.

[Koboldt et al., 2010] Koboldt, D. C., Ding, L., Mardis, E. R., and Wilson, R. K. (2010). Challenges of sequencing human genomes. *Briefings in Bioinformatics*, 11:484–498.

[Kraft and Cox, 2008] Kraft, P. and Cox, D. G. (2008). Study Designs for Genome-Wide Association Studies. *Advances in Genetics*, 60:465–504.

[Langmead et al., 2010] Langmead, B., Hansen, K. D., and Leek, J. T. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome biology*, 11:R83.

[Langmead et al., 2009] Langmead, B., Schatz, M. C., Lin, J., Pop, M., and Salzberg, S. L. (2009). Searching for SNPs with cloud computing. *Genome biology*, 10:R134.

[Leo et al., 2009] Leo, S., Santoni, F., and Zanetti, G. (2009). Biodoop: Bioinformatics on Hadoop. *2009 International Conference on Parallel Processing Workshops*.

[Lewis et al., 2012] Lewis, S., Csordas, A., Killcoyne, S., Hermjakob, H., Hoopmann, M. R., Moritz, R. L., Deutsch, E. W., and Boyle, J. (2012). Hydra: a scalable proteomic search engine which utilizes the Hadoop distributed computing framework. *BMC bioinformatics*, 13:324.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25:1754–1760.

[Li et al., 2008] Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18:1851–1858.

[Lindenmayer, 1968] Lindenmayer, A. (1968). Mathematical models for cellular interactions in development I. Filaments with one-sided inputs. *Journal of Theoretical Biology*, 18(3):280–299.

[Liu et al., 2012] Liu, C.-M., Wong, T., Wu, E., Luo, R., Yiu, S.-M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., Li, R., and Lam, T.-W. (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics (Oxford, England)*, 28:878–9.

[Madden, 2003] Madden, T. (2003). The BLAST Sequence Analysis Tool.

[McCarty et al., 2011] McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I., Jarvik, G., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., Struewing, J. P., Wolf, W. A., and Team, T. E. (2011). The Electronic Medical Records & Genomics (eMERGE) Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20:1297–1303.

[Ng et al., 2008] Ng, P. C., Levy, S., Huang, J., Stockwell, T. B., Walenz, B. P., Li, K., Axelrod, N., Busam, D. A., Strausberg, R. L., and Venter, J. C. (2008). Genetic Variation in an Individual Human Exome. *PLoS Genetics*, 4:15.

[Ng et al., 2010] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42:30–35.

[Ng et al., 2009] Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461:272–276.

[Nguyen et al., 2011] Nguyen, T., Shi, W., and Ruden, D. (2011). CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. *BMC research notes*, 4:171.

[Nielsen et al., 2010] Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., and Wang, T. (2010). Visualizing genomes: techniques and challenges. *Nature methods*, 7(3 Suppl):S5–S15.

[Niemenmaa et al., 2012] Niemenmaa, M., Kallio, A., Schumacher, A., Klemelä, P., Korpelainen, E., and Heljanko, K. (2012). Hadoop-BAM: directly manipulating next generation sequencing data in the cloud. *Bioinformatics (Oxford, England)*, 28:876–7.

[O'Connor et al., 2010] O'Connor, B. D., Merriman, B., and Nelson, S. F. (2010). SeqWare Query Engine: storing and searching sequence data in the cloud. *BMC bioinformatics*, 11 Suppl 12:S2.

[Pabinger et al., 2013] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, pages bbs086–.

[Pandey et al., 2008] Pandey, V., Nutter, R. C., and Prediger, E. (2008). *Applied Biosystems SOLiD System: Ligation-Based Sequencing*. Wiley-VCH Verlag GmbH & Co. KGaA.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 85(8):2444–8.

[Peng et al., 1992] Peng, C.-K., Buldyrev, S., Goldberger, A., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. (1992). Fractal landscape analysis of DNA walks. *Physica A: Statistical Mechanics and its Applications*, 191(1):25–29.

[Peterson et al., 2010] Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A., and Kann, M. G. (2010). DMDM: domain mapping of disease mutations. *Bioinformatics (Oxford, England)*, 26:2458–2459.

[Pevzner et al., 1989] Pevzner, P. A., Borodovsky MYu, and Mironov, A. A. (1989). Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *Journal of biomolecular structure & dynamics*, 6(5):1013–26.

[Pietrokovski et al., 1990] Pietrokovski, S., Hirshon, J., and Trifonov, E. N. (1990). Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. *Journal of biomolecular structure & dynamics*, 7(6):1251–68.

[Pireddu et al., 2011] Pireddu, L., Leo, S., and Zanetti, G. (2011). SEAL: a distributed short read mapping and duplicate removal tool. *Bioinformatics (Oxford, England)*, 27:2159–2160.

[PLINK, 2013, ] PLINK, 2013. PLINK R plugins e librerie plink per r. `http://pngu.mgh.harvard.edu/purcell/plink/rfunc.shtml`. Acceduto il: 7/10/2013.

[Prlić et al., 2007] Prlić, A., Down, T. A., Kulesha, E., Finn, R. D., Kähäri, A., and Hubbard, T. J. (2007). Integrating sequence and structural biology with DAS. *BMC Bioinformatics*, 8:333.

[Prof and Received, 1984] Prof, L. and Received, F. R. G. (1984). Nucleic Acids Research. 12(5):2561–2568.

[Purcell et al., 2007] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, Manuel and Bender, D. M. J., Sklar, P., de Bakker, P., Daly, M., and Sham,

P. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81:559–575.

[Safran et al., 2003] Safran, M., Chalifa-Casp, V., Shmueli, O., Rosen, N., Benjamin-Rodrig, H., Ophir, R., Yanai, I., Shmoish, M., and Lancet, D. (2003). The GeneCards™ family of databases: GeneCards, GeneLoc, GeneNote and GeneAnnot. *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*.

[Sam and Specification, 2011] Sam, T. and Specification, F. (2011). The SAM Format Specification (v1.4-r985). *Read*, pages 1–11.

[Schatz, 2009] Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics (Oxford, England)*, 25:1363–1369.

[Schatz et al., 2010] Schatz, M. C., Sommer, D., Kelley, D., and Pop, M. (2010). De novo assembly of large genomes using cloud computing. In *CSHL Biology of Genomes conference*.

[Searls and Dong, 1993] Searls, D. and Dong, S. (1993). A Syntactic Pattern Recognition System For Dna Sequences. *Proceedings of 2nd International conference on Bioinformatics, Supercomputing and Complex genome analysis*.

[Searls and Noordewier, 1991] Searls, D. and Noordewier, M. (1991). Pattern-matching search of DNA sequences using logic grammars. In *[1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application*, volume i, pages 3–9. IEEE Comput. Soc. Press.

[Searls, 1989] Searls, D. B. (1989). Investigating the Linguistics of DNA with Definite Clause Grammars. pages 189 – 208.

[Shaffer et al., 2005] Shaffer, L. G., Slovak, M. L., and Cambell, L. J. (2005). *ISCN 2005: an international system for human cytogenetic nomenclature (2005)*.

[Smigielski et al., 2000] Smigielski, E. M., Sirotkin, K., Ward, M., and Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research*, 28:352–355.

[Stein et al., 2010] Stein, L. D. et al. (2010). The case for cloud computing in genome informatics. *Genome Biol*, 11(5):207.

[Stein et al., 2002] Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., and Lewis, S. (2002). The Generic Genome Browser: A Building Block for a Model Organism System Database. *Genome Research*, 12:1599–1610.

[Tanaka, 2003] Tanaka, T. (2003). The International HapMap Project. *Nature*, 426:789–796.

[Thusoo et al., 2010] Thusoo, A., Sarma, J., Jain, N., Shao, Z. S. Z., Chakka, P., Zhang, N. Z. N., Antony, S., Liu, H. L. H., and Murthy, R. (2010). Hive - a petabyte scale data warehouse using Hadoop. *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*.

[Visscher et al., 2012] Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

[Vouzis and Sahinidis, 2011] Vouzis, P. D. and Sahinidis, N. V. (2011). GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, 27:182–188.

[Zhang et al., 2012] Zhang, L., Gu, S., Liu, Y., Wang, B., and Azuaje, F. (2012). Gene set analysis in the cloud. *Bioinformatics (Oxford, England)*, 28:294–5.