Alma Mater Studiorum · Università di Bologna

DIPARTIMENTO DI FISICA ED ASTRONOMIA

Corso di Laurea Magistrale in Fisica

Valutazione statistica della ripetibilità di esperimenti di sequenziamento del DNA in specie batteriche

Relatore:
Prof. Gastone Castellani

Presentata da:
Federico Mondaini

Correlatore:

Prof. Daniel Remondini

Italo Faria Do Valle

Anno Accademico 2015/2016

Abstract:

I progressi della biologia molecolare assieme alle nuove tecnologie di sequenziamento applicate su scala genomica alla genetica molecolare, hanno notevolmente elevato la conoscenza sulle componenti di base della biologia e delle patologie umane. All'interno di questo contesto, prende piede lo studio delle sequenze genetiche dei batteri, consentendo dunque, una migliore comprensione di ciò che si nasconde dietro le malattie legate all'uomo.

Il seguente lavoro di tesi si propone come obiettivo l'analisi del DNA del batterio ***Listeria monocytogenes****, un microrganismo presente nel suolo e in grado di contaminare l'acqua e gli alimenti. Lo scopo principale è quello di confrontare la variabilità tecnica e biologica, al fine di capire quali siano gli SNPs reali (Single Nucleotide Polymorphism) e quali artefatti tecnici.*

La prima parte, quindi, comprende una descrizione del processo di individuazione degli SNPs presenti nel DNA dei campioni in esame, in particolare di tre isolati diversi e tre copie.

Nella seconda parte, invece, sono effettuate delle indagini statistiche sui parametri relativi agli SNPs individuati, ad esempio il coverage o il punteggio di qualità assegnato alle basi. Il fine ultimo è quello di andare a verificare se sussistano particolari differenze tra gli SNPs dei vari isolati batterici.

Keywords: *DNA | SNP | Next Generation Sequencing | Coverage | Quality by Depth | Strand Odds Ratio | ANOVA | Multiple Comparisons |*

Indice

Introduzione

Il genoma, cioè il complesso dei geni che definiscono un individuo, è un documento immenso che viene trasmesso dalla cellula madre alla cellula figlia. Studiare il genoma di un organismo implica lo studio e la codifica del suo stesso DNA.
Tale codificazione è, dunque, importante in quanto le informazioni riguardanti un essere vivente vengono immagazzinate lungo la sequenza genetica. In questi ultimi anni il rapido sequenziamento del DNA ha radicalmente cambiato la biologia, tant'è che si è scoperto che solo l'1.5 % dei geni che costituiscono il patrimonio genetico umano è codificante, cioè attivo, mentre il 98.5 % è non codificante e viene denominato "junk-DNA".
Al contrario, nei batteri, quasi tutto il DNA risulta codificante, in quanto si è visto essere ricco di sequenze, e quindi informazioni, non ridondanti, cosa che invece è caratteristica del DNA non codificante. Quindi in termini di stringhe di nucleotidi, c'è proprio l'idea che dentro queste sequenze siano contenute delle informazioni importanti.

L'idea del seguente lavoro di tesi è stata sviluppata proprio all'interno di questo contesto. Lo scopo è stato quello di prendere l'intera sequenza genomica di un batterio, *Listeria monocytogenes*, e farne uno studio dal punto di vista informativo. In particolare, sono state effettuate delle analisi statistiche sulle variazioni della sequenza genetica del batterio considerato, ovvero sono stati studiati i cosiddetti SNP *(single nucleotide polymorphism)*, al fine di confrontare ciò che differenzia la variabilità biologica da quella tecnica.
Su tali variazioni, o mutazioni, si sono accumulate sufficienti evidenze sperimentali, così da considerare il loro ruolo importante nell'indurre la suscettibilità delle risposte dell'organismo nei confronti degli stimoli interni (endogeni) ed esterni (esogeni). Ad esempio, alcuni di questi SNP possono rendere ragione della resistenza del batterio stesso a un determinato agente decontaminante.

Nel primo capitolo si descrive la struttura del DNA, le sue caratteristiche e il concetto che sta alla base degli SNPs. Lo scopo è quello di fornire al lettore gli strumenti necessari per comprendere i dati su cui verranno effettuate le analisi.

Nel secondo capitolo si presenta il metodo di lavoro che ha portato all'identificazione degli SNPs. Vengono, quindi, mostrati nel dettaglio tutti i passaggi di una determinata *pipeline,* alla fine della quale si giunge all'individuazione di tutte le varianti presenti nei campioni dei batteri studiati.

Nel terzo capitolo vengono presentati i metodi statistici utilizzati per esaminare gli SNPs ottenuti, in particolare si descrive in che cosa consiste il Test di Bartlett, l'Analisi di Varianza (*ANOVA*), il Test dei confronti Multipli e infine che cosa sono i Box plots.

Nel quarto capitolo vengono mostrati i risultati derivanti dall' applicazione dei metodi statistici descritti, traendone infine le dovute conclusioni.

Capitolo 1

Il DNA

In questo capitolo si vuole dare una breve descrizione della struttura del DNA e del processo di sintesi delle proteine. Necessario a questo scopo è una esemplificata introduzione alla struttura della cellula e all'ambiente in cui sono situati e si svolgono i processi sui quali si vuole porre attenzione [1][2].

1.1 Organismi e cellule

La cellula è l'unità di struttura e di funzione di cui sono costituiti tutti gli organismi viventi: ogni loro attività dipende dal funzionamento delle cellule. Queste possono essere assai diverse tra loro, ma tutte hanno in comune le seguenti caratteristiche:

- ogni cellula è circondata da una membrana plasmatica, una sorta di sottilissima pellicola che separa la cellula dalle altre o dall'ambiente circostante, regolando anche l'ingresso e l'uscita dei materiali.
- la membrana plasmatica racchiude il citoplasma, che occupa la maggior parte dello spazio interno e nel quale avviene gran parte delle funzioni cellulari. Queste funzioni, nelle cellule meglio organizzate ed evolute, sono affidate a speciali categorie di organelli[1].
- tutte le cellule presentano un **nucleo** (che in quelle meno evolute è sostituito dal cosiddetto nucleoide) che contiene il DNA.

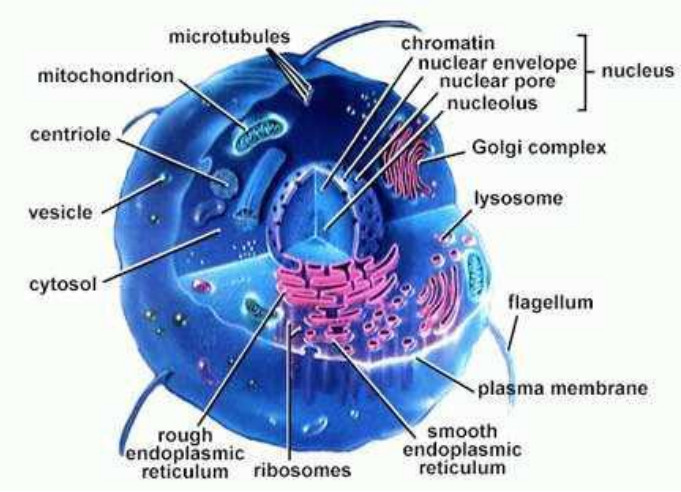


***Figura 1**: Un modello di cellula eucariote*

[1] Gli organelli costituiscono compartimenti delimitati da un tipo di membrana caratteristica.

Esistono due tipologie fondamentali di cellule: *cellule eucariote* (quelle che formano il corpo delle piante, degli animali e dell'uomo) e *cellule procariote* (per esempio i batteri).

Le **cellule procariote**, oltre a essere normalmente assai più piccole di quelle eucariote, sono anche molto più semplici dal punto di vista strutturale: in esse vi è il cosiddetto nucleoide sede del DNA e mancano del tutto organelli cellulari distinti.

Le **cellule eucariote**, al contrario, si caratterizzano proprio per la presenza di organelli distinti e provvisti di una loro membrana di separazione dal citoplasma; contengono un nucleo, che è separato dal resto con una membrana. Il nucleo contiene cromosomi che sono i portatori del materiale genetico**.**

Racchiusi nella membrana esterna ci sono organelli come centrioli, lisosomi, cloroplasti (che producono zucchero) e mitocondri che producono energia sotto forma di ATP, ecc.. .

Una singola cellula per produrre tessuti e organi deve:

- crescere
- dividersi
- differenziarsi

Queste tre fasi vanno a costituire il CICLO CELLULARE.

1.2 Struttura del DNA

Il DNA è la principale molecola portatrice di informazione in una cellula. Il DNA è un polinucleotide cioè una catena di piccole molecole chiamate nucleotidi. I nucleotidi sono composti da uno zucchero (che nelle molecole di DNA è il desossiribosio, uno zucchero a cinque atomi di carbonio) al quale si attaccano le basi azotate (adenina, timina, citosina, guanina) e un gruppo fosfato.

Esistono 4 differenti tipi di nucleotidi raggruppati in due generi chimici:

• ***purine***

ADENINA [**A**]

GUANINA [**G**]

• ***pirimidine***

CITOSINA [**C**]

TIMINA [**T**]

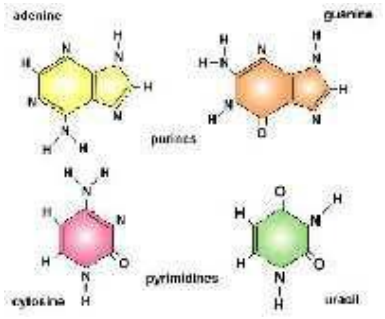


***Figura 2**: Purine e Pirimidine*

Differenti nucleotidi si legano assieme in un qualche ordine a formare un polinucleotide, ad esempio[2]:

A – G – T – C – C – A – A – G – C – T – T

Il polinucleotide può avere una qualsiasi lunghezza e qualsiasi sequenza; la sequenza ha una direzione, come questa:

A → G → T → C → C → A → A → G → C → T → T

L'orientazione del polinucleotide è data marcando le estremità con 5' a sinistra e 3' a destra (le due estremità della sequenza sono chimicamente differenti).
Due stringhe sono dette complementari se una può essere ottenuta dall'altra cambiando A con T e C con G, e cambiando la direzione della molecola in senso opposto. Ad esempio:

← T ← C ← A ← G ← G ← T ← T ← C ← G ← A ← A

questa è la complementare della sequenza superiore.

Specifiche coppie di nucleotidi possono formare legami deboli fra loro. A si lega con T, C si lega con G (per essere più precisi, due legami idrogeno possono venire a formarsi fra le coppie A-T, e tre legami idrogeno fra C-G).

Sebbene tali interazioni siano legami deboli, quando due lunghi polinucleotidi complementari si incontrano, essi tendono ad attaccarsi, come in figura 3.

```
C - G - A - T - T - G - C - A - A - C - G
|   |   |   |   |   |   |   |   |   |   |
G - C - T - A - A - C - G - T - T - G - C
```

***Figura 3**: unione di polinucleotidi complementari.*

Le linee verticali fra i due fili di DNA rappresentano le forze fra essi. Le coppie A-T e G-C sono chiamate *coppie-base* (bp). La lunghezza del DNA è misurata in coppie base di nucleotidi (nt).

[2] Per semplicità rappresentiamo l'intero nucleotide con la sua base azotata

Due catene di polinucleotidi complementari formano una struttura stabile, che assomiglia a un'elica ed è conosciuta come il DNA a doppia elica (circa 10 bp in questa struttura formano un intero giro che è lungo circa 3.4nm).

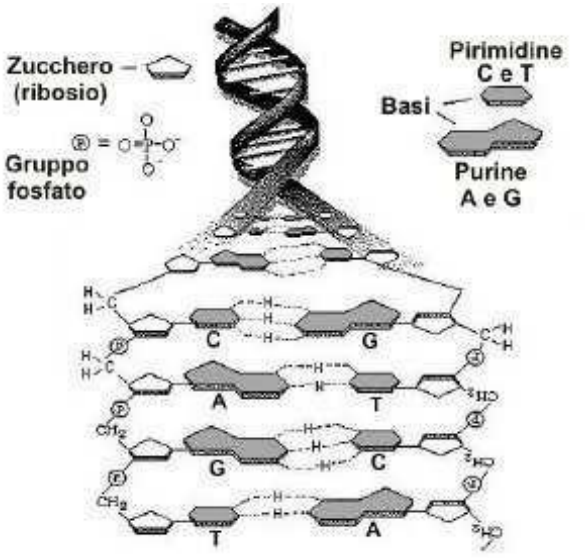


***Figura 4**: Struttura del DNA*

Notiamo che, dal momento che le stringhe sono complementari, ciascuna di esse determina completamente l'altra e quindi per avere l'informazione su tutto il DNA basterebbe prendere in considerazione una sola stringa della molecola del genoma. La massima quantità di informazione che può essere codificata in una tale molecola è quindi 2 bits volte la lunghezza della sequenza.

Durante il processo di sintesi del DNA (la replicazione del DNA) i due filamenti di DNA di ciascun cromosoma vengono srotolati e ciascun filamento dirige la sintesi del filamento di DNA ad esso complementare, generando così due doppie eliche di DNA, ciascuna delle quali è identica alla molecola parentale. Questo avviene durante la divisione cellulare e, dal momento che ciascuna doppia elica di DNA contiene un filamento che apparteneva alla molecola parentale e un filamento di DNA neosintetizzato, si dice che il processo replicativo è **semi-conservativo** [3].

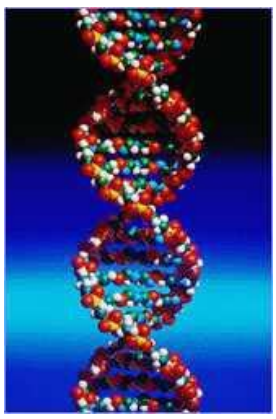


***Figura 5**: Il DNA*

1.3 SNP - Single Nucleotide Polymorphism

Le recenti scoperte inerenti al genoma umano hanno permesso di comprendere le basi che determinano le variazioni individuali nel funzionamento dell'organismo. E' infatti emerso chiaramente che tutti gli individui sono uguali per il 99,9% del loro patrimonio genetico, mentre lo 0,1% di variabilità determina le differenze fra gli individui. La variabilità genetica fa sì che ogni individuo risponda in maniera diversa agli stimoli ambientali e, quindi, sia in grado di adattarsi o meno a particolari condizioni. Le differenze genetiche tra individui sono spesso rappresentate da varianti genetiche puntiformi dette *SNP*.

Un **polimorfismo a singolo nucleotide** (in inglese ***Single Nucleotide Polymorphism*** o **SNP**, pronunciato *snip*) è un polimorfismo, cioè una variazione, del materiale genico a carico di un unico nucleotide. La figura 6 mostra in dettaglio che cos'è uno SNP.

Ad esempio, se le sequenze individuate in due pazienti sono AAGC**C**TA e AAGC**T**TA, è presente uno SNP che differenzia i due alleli C e T.

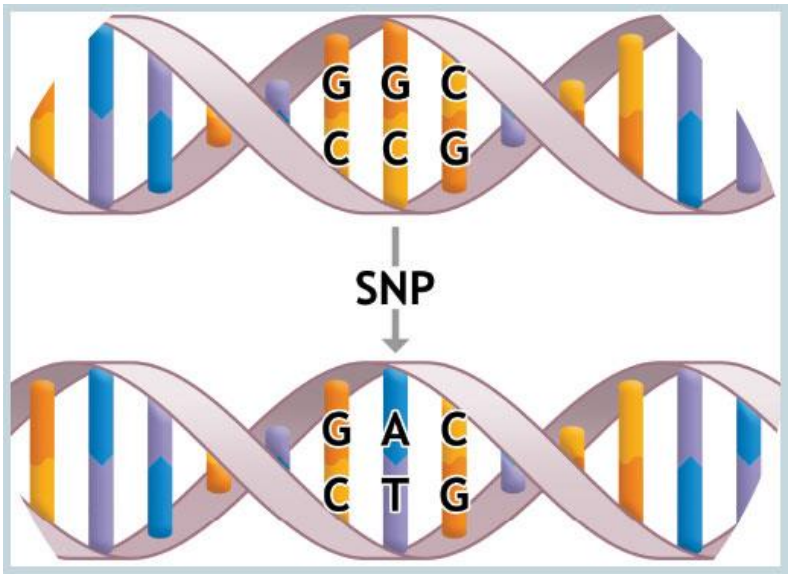


Figura 6: *Rappresentazione di uno SNP.*

All'interno di una popolazione, è possibile determinare una minor frequenza allelica, il rapporto tra la frequenza della variante più rara e quella più comune di un determinato SNP. Solitamente ci si guarda con maggiore attenzione da SNPs aventi un valore di minor frequenza allelica minore all'1%, trascurando nelle analisi la maggior parte degli SNPs che, anche per il loro elevatissimo numero, risultano poco maneggevoli. È importante notare che possono esistere variazioni notevoli tra popolazioni umane. Uno SNP molto comune in un determinato gruppo etnico può, dunque, essere molto raro in un'altra popolazione.

Lo studio degli SNPs è molto utile poiché variazioni anche di singoli nucleotidi possono influenzare lo sviluppo delle patologie o la risposta ai patogeni, agli agenti chimici, ai farmaci. Per tale motivo gli SNPs possono avere una grande importanza nello sviluppo di nuovi farmaci

e nella diagnostica, in quanto consentono di conoscere l'effetto che può avere un farmaco su un individuo ancor prima della somministrazione, attraverso uno screening degli SNPs presenti nel gene responsabile della metabolizzazione del farmaco stesso; queste sono le basi della farmacogenomica. Dal momento che gli SNPs sono perlopiù ereditati di generazione in generazione, essi vengono utilizzati in alcuni studi genetici.

1.4 Cromosomi, geni e RNA

In questo paragrafo si vuole fare chiarezza su termini che si useranno d'ora in poi per descrivere il processo di sintesi delle proteine.

In una cellula tipica ci sono uno o parecchi DNA a doppia elica arrotolati e organizzati come **cromosomi**. Nelle cellule eucariote i cromosomi hanno una struttura complessa in cui il DNA è arrotolato intorno a proteine strutturali chiamate istoni. Nell'uomo, ad esempio, ci sono 23 coppie di cromosomi, grandi abbastanza da essere visti al microscopio. La lunghezza totale del DNA umano, se potessimo srotolarlo, dovrebbe essere più di un metro. Anche i mitocondri contengono DNA ma in quantità assai minore rispetto ai cromosomi. IL DNA dei cromosomi e dei mitocondri forma il *GENOMA* dell'organismo. Tutte le cellule in un organismo contengono identici genomi, con alcune eccezioni, come risultati della replicazione del DNA a ciascuna divisione cellulare.

Il gene può essere visto come una sorta di messaggio in codice dal quale si parte per arrivare alla sintesi delle proteine. Una definizione di gene appropriata non è ancora stata concordata; si propone la seguente, con la consapevolezza della sua imprecisione:

> Un gene è un tratto continuo del DNA dal quale un complesso macchinario molecolare può leggere una informazione (codificata come una stringa di A, T, G e C) e produrre un particolare tipo di proteina o più differenti proteine.

In pratica, il messaggio contenuto nel DNA del nucleo viene trascritto in un altro acido nucleico, il cosiddetto RNA o acido ribonucleico. Similmente al DNA, l'RNA è costituito anch'esso da nucleotidi ma al posto della timina (T) ha l'uracile (U), al posto del desossiribosio è presente il ribosio e, inoltre, non ha una struttura ad elica ma è un'unica stringa. Comunque l'RNA può avere strutture spaziali complesse dovute a legami complementari fra le parti della stessa stringa.

L'RNA può codificare l'informazione genetica, è replicabile, forma complesse strutture 3D, e può anche agire come catalizzatore per certe reazioni chimiche relative allo splicing (una fase della sintesi proteica che verrà esposta più avanti).

In una rappresentazione lineare l'RNA si presenta come un insieme di geni uniti da sequenze dette regioni intergeniche la cui funzione è quella di essere riconosciute da molecole regolatrici, solitamente proteine. Sono chiamate anche regioni non codificanti in quanto non codificano per le proteine; qui la funzione del DNA è determinata direttamente dalla sua sequenza, non per mezzo di un qualche codice intermediario.

1.5 La sintesi proteica

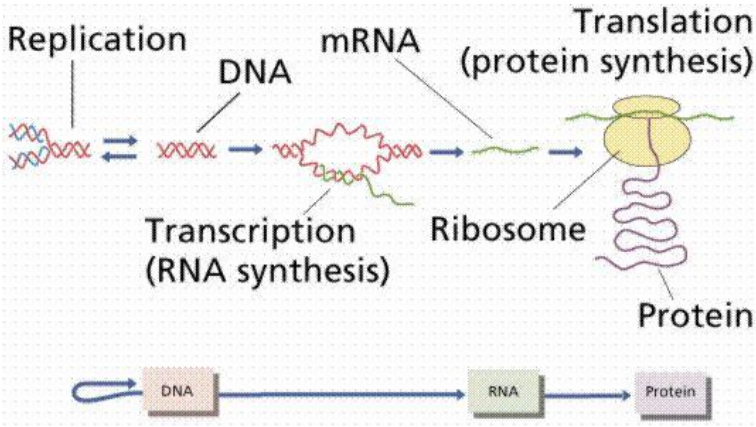


***Figura 7**: La sintesi proteica.*

Analizziamo ora il processo che porta alla creazione di una proteina. Possiamo schematizzare la sintesi delle proteine in tre fasi (trascrizione, splicing, traduzione) per le cellule eucariote; in due (trascrizione, traduzione) per le cellule procariote:

1. **Trascrizione**: Il messaggio genetico, che è codificato nella molecola di DNA, non viene utilizzato direttamente per la sintesi delle proteine che, tra l'altro, avviene nel citoplasma. Nelle cellule eucariote esso viene infatti trascritto in un complementare pre-mRNA dalla complessa proteina RNA polimerasi (per cellule procariote l'RNA è direttamente l'mRNA). Siccome il linguaggio del DNA e quello dell'RNA sono basati su sequenze di nucleotidi sostanzialmente simili, non vi è, nei due acidi nucleici, una sostanziale differenza di linguaggio molecolare e per questo il processo è chiamato trascrizione, cioè sostanzialmente "copiatura". Questo processo è simile alla replicazione del DNA, con la sola differenza che, ad essere copiato, è uno solo dei due filamenti del DNA e che la copiatura, o trascrizione, produce una molecola di mRNA complementare al tratto copiato, dove le basi complementari sono G − C e A − U (invece di A − T). Dei due filamenti di DNA, dunque, solo uno viene copiato, quello cioè che codifica la proteina e che è riconosciuto dall'enzima della trascrizione, la RNA polimerasi, per la presenza di una sequenza, chiamata promotore, collocata all'inizio del gene. La trascrizione termina in corrispondenza di un'altra sequenza specifica che dà il segnale di terminazione.

Nei batteri, dove non vi è una membrana nucleare, l'mRNA così prodotto viene immesso direttamente nel citoplasma ed è pronto per avviare la sintesi proteica. Negli eucarioti, invece, prima di essere inviato nel citoplasma, il pre-mRNA viene processato nella cosiddetta fase di splicing

2. **Splicing**: In questo meccanismo di "taglia e cuci", presente solo nelle cellule eucariote, vengono rimossi alcuni tratti del pre-mRNA chiamati introni e sono collegate insieme le sezioni rimanenti, chiamate esoni. Gli esoni sono la parte del gene che serve per la codifica delle proteine. Il meccanismo dello splicing dell'RNA dipende in modo cruciale da quella che è stata chiamata la regola *GT −AG*: quasi sempre gli introni cominciano con GT (o GU se ci si riferisce all'RNA) e terminano con AG.

   Grazie allo splicing, gli esoni vengono ad essere congiunti nello stesso ordine in cui si trovavano nel DNA. In questo modo si mantiene la collinearità fra gene e proteina, relativamente ai singoli esoni e alle parti corrispondenti della catena proteica, ma le distanze nei geni non corrispondono alle distanze nella proteina: la lunghezza del gene è definita dalla lunghezza dell'iniziale pre-mRNA invece che dalla lunghezza dell'mRNA. Notiamo che i geni eucarioti non sono necessariamente interrotti. Alcuni corrispondono direttamente alla proteina prodotta, come nei geni procarioti. Nel *Saccharomyces cerevisiae*, per esempio, molti geni non sono interrotti. In eucarioti evoluti, molti geni sono interrotti e gli **introni** sono molto spesso più lunghi degli esoni, creando geni che sono molto più grandi delle loro regioni codificanti. Tanto più il genoma diventa grande, tanto più gli introni sono estesi mentre gli esoni sono sequenze più brevi.

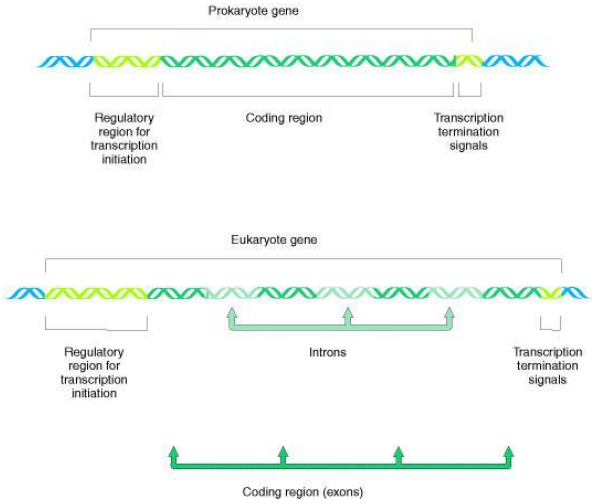


Figura 8: *regione codificante in cellule eucariote e cellule procariote.*

Il risultato della fase di splicing è l'**mRNA**.

3. **Traduzione**: Raggiunto il citoplasma l'mRNA viene tradotto nello specifico linguaggio delle proteine. Questo processo richiede una complessa serie di operazioni dovendosi trasferire l'informazione scritta nella sequenza dei quattro nucleotidi dell'mRNA, nella specifica sequenza dei venti possibili amminoacidi[3] delle proteine. Per ovviare a questa necessità il codice dovrà essere basato su triplette di nucleotidi (ovvero, tre nucleotidi adiacenti). Ciascuna tripletta è chiamata codone e codifica per un amminoacido. Dal momento che ci son $4^3 = 64$ codoni e solamente 20 amminoacidi, il codice è ridondante in quanto singoli amminoacidi possono corrispondere anche a due o più triplette diverse. Va, tuttavia, osservato che un codice così fatto non è mai ambiguo, in quanto la stessa tripletta codifica sempre per un unico amminoacido.

   Nel citoplasma l'mRNA forma un complesso con il ribosoma (una struttura composta da proteine ed RNA). Poiché i trinucleotidi e gli amminoacidi sono elementi strutturalmente incongruenti, sorge il problema di come ogni particolare codone possa trovare corrispondenza in un determinato amminoacido. Questo compito è affidato alla molecola **tRNA** (RNA di trasferimento), che traduce il linguaggio dei singoli nucleotidi in quello degli amminoacidi grazie a due sue proprietà fondamentali:

   1. rappresenta un singolo amminoacido al quale si lega in modo covalente;
   2. contiene una sequenza di tre nucleotidi, l'anticodone, complementare al codone che rappresenta il suo amminoacido. L'anticodone permette al tRNA di riconoscere il codone. È importante osservare il fatto che esiste almeno un tRNA per ogni amminoacido.

   Per la sua struttura il tRNA può portare ciascun amminoacido al ribosoma e riconoscere un codone nel mRNA; l'amminoacido portato dal tRNA è aggiunto alla nascente proteina.

   Da notare che, benché esistano 64 codoni, il numero di molecole di tRNA caratterizzate da codoni differenti è inferiore: esistono infatti solo 30 tipi di tRNA citoplasmatici e 22 tipi di tRNA mitocondriali. L'uso di tutti e 64 i codoni sui ribosomi citoplasmatici e mitocondriali è tuttavia possibile perché le regole di appaiamento tra basi sono meno

[3] sono i "mattoni" delle proteine. Vengono classificati in base alla natura dei loro gruppi laterali. Il nome di ogni amminoacido può essere indicato da abbreviazioni a tre o a una sola lettera.

rigide quando si tratta di riconoscimento tra codone e anticodone. L'ipotesi dell'imprecisione dell'accoppiamento (*wobble hypothesis*) afferma che l'appaiamento tra codone e anticodone segue le normali regole A − U, G − C per le basi nelle prime due posizioni di un codone, ma che nella terza posizione si potrebbe verificare una "incertezza" e che quindi è ammesso anche l'appaiamento G − U. La traduzione continua finché si incontra un codone di terminazione e cioè UAA, UAG, UGA nel caso degli mRNA codificati nel nucleo; UAA, UAG, UGA o AGG nel caso degli mRNA codificanti nei mitocondri (si ricorda che il codone d'inizio è quasi sempre AUG).

Il termine della traduzione è la parte finale dell'espressione del gene e il prodotto è una proteina, la sequenza della quale corrisponde alla sequenza codificata dell'mRNA.

Capitolo 2

Nel seguente capitolo verrà fornita, in primis, una breve descrizione del batterio *Listeria monocytogenes*, successivamente si passerà ad analizzare nel dettaglio la descrizione della *pipeline* che ha portato all'identificazione degli SNPs per il batterio considerato.

2.1 Listeria monocytogenes

Listeria monocytogenes è la specie di batteri patogeni che causa l'infezione della listeriosi. È un batterio facoltativamente anaerobico, capace quindi di sopravvivere sia in presenza che in assenza di ossigeno. È in grado di crescere e riprodursi all'interno di cellule ospiti ed è uno dei patogeni di origine alimentare più virulenti, responsabile negli USA di all'incirca 1600 infezioni e 260 morti all'anno.

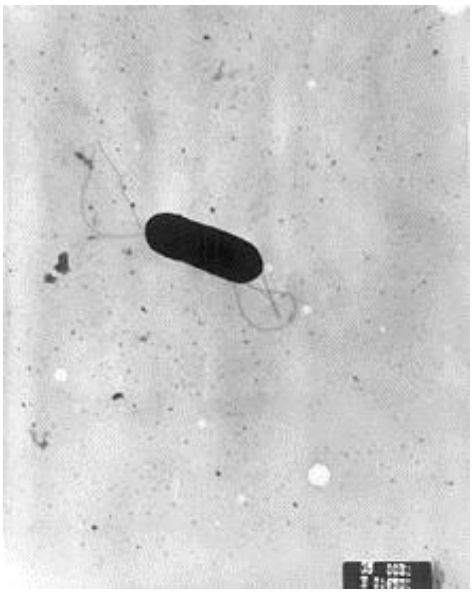


***Figura 9**: Listeria Monocytogenes visto tramite microscopio elettronico.*

Listeria monocytogenes è un batterio **Gram positivo** (rimane colorato di blu o viola dopo aver subito la colorazione di Gram), **asporigeno** (cioè che non genera spore, ovvero cellule avvolte da una membrana che le protegge dagli agenti atmosferici i quali potrebbero danneggiarne il contenuto), **aerobio-anaerobio facoltativo**. Il microrganismo cresce in un range di temperatura molto largo (tra i +3 °C e i 45 °C) con un optimum tra i 30 °C e i 38 °C. Presenta buona resistenza a varie condizioni di pH (tra 4,4 e 9,6) e temperatura, caratteristiche che lo rendono un potenziale contaminante di alimenti, anche se conservati in frigorifero.

Per il seguente lavoro di tesi, si aveva a disposizione 3+3=6 campioni del batterio (in particolare 3 isolati differenti con le relative copie) sui quali era già stato eseguito il sequenziamento tramite il metodo Illumina. Per una migliore comprensione è stata riportata la tabella 1 dove si elencano i nomi dei batteri con le relative copie; ognuno di quegli elementi è stato completamente analizzato secondo i metodi descritti nei paragrafi successivi [4].

Isolato	Copia
LIST1	LIST63
LIST2	LIST64
LIST3	LIST65

***Tabella 1**: Elenco dei batteri analizzati con le relative copie.*

2.2 Analisi computazionale su sequenze di DNA

Prima di addentrarsi nello studio della *pipeline* dell'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 (o pipeline) 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.
La figura 10, rappresentata in maniera poco dettagliata (in quanto le operazioni sono molto variabili), è una pipeline tipica per l'analisi delle sequenze esoniche.
In verde sono mostrati i processi, mentre in blu 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 successivo.
All'interno di questa tesi verrà approfondito ogni stadio dell'analisi, senza entrare troppo in dettagli specifici dipendenti dalla piattaforma o da particolari necessità.
Per ora ci si limita semplicemente a dire che la figura 10 mostra una pipeline per il sequenziamento del DNA utilizzata per identificare alcune varianti somatiche.
La prima fase, che prende il nome di ***Base calling*** è altamente specifica e collegata alla procedura di sequenziamento usata per generare sequenze di reads.

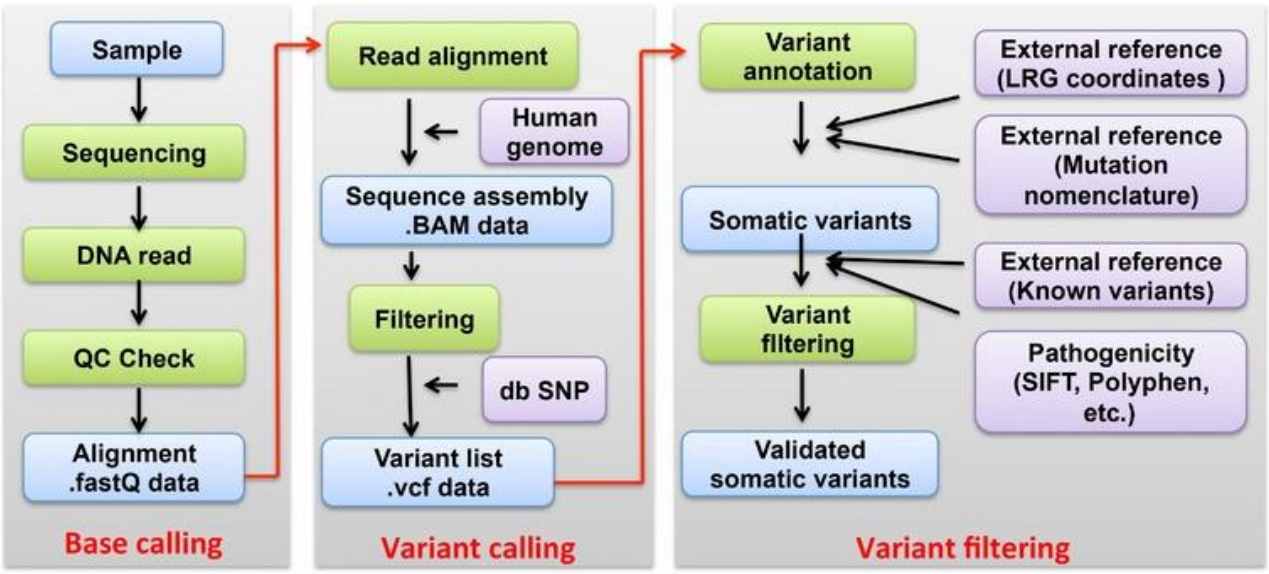


***Figura 10**: Pipeline utilizzata per l'identificazione degli SNPs.*

La seconda fase, quella di ***Variant Calling***, allinea le sequenze di reads al genoma di riferimento (in questo caso genoma umano), per ricostruire un completo assemblaggio delle sequenze e dedurre tutte le variazioni dovute al filtraggio.
Infine, l'ultimo step (***Variant Filtering***), salva queste variazioni utilizzando più fonti esterne come riferimenti [5].

La seguente pipeline è stata riportata come esempio generico per il sequenziamento del genoma umano, tuttavia in questa tesi, si prenderà in considerazione (come già accennato precedentemente) lo studio e l'analisi del genoma del batterio **Listeria monocytogenes.**

2.3 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 delle reads, svantaggio a cui si rimedia con letture ripetute e appositi processi computazionali nelle successive fasi di analisi.

2.4 Base calling & FastQC

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 alla mancanza di accuratezza delle letture e a valori non soddisfacenti di probabilità. Il formato dei dati di output più diffuso tra le piattaforme NGS è il FASTQ, un formato testuale di cui esistono diverse versioni.

1.1 FASTQ

FASTQ è diventato negli anni un formato molto comune per la condivisione di dati genetici di sequenziamento, poiché combina 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, in quanto aggiunge specifiche informazioni sull'attendibilità della lettura, rappresentando così la sequenza a disposizione con un livello di dettaglio maggiore, ma senza pesare sulla dimensione dei dati.
Grazie all'estrema semplicità del formato, il FASTQ è ampiamente utilizzato per l'interscambio di dati; tuttavia il FASTQ, anche se nato come evoluzione del FASTA, continua a soffrire dell'assenza di una definizione chiara e non ambigua, mancanza che ha portato all'esistenza di molte varianti incompatibili tra loro.

Proprio a causa di questa imprecisione, risulta necessario compiere alcuni controlli di qualità che confermino che i frammenti siano stati effettivamente letti correttamente.
Esistono numerosi programmi utilizzati per questo tipo di analisi, e uno dei più completi è proprio **FastQC**. Questo software permette una visualizzazione delle problematiche introdotte dal sequenziamento, con dettaglio sufficiente per spiegare anche le possibili cause della scarsa affidabilità dei dati [6].
Il programma riceve in ingresso un file nel formato *.fastq* in cui è stato raccolto l'elenco delle sequenze.
Questo programma valuta differenti aspetti per il controllo della qualità complessiva dei frammenti, e passa in rassegna: 1) la qualità della lettura delle singole basi, calcolata con opportuni algoritmi; 2) la verosimiglianza dell'assegnamento di ciascuna base della sequenza al nucleotide scelto; 3) le duplicazioni delle sequenze e alcune considerazioni sulle loro lunghezze.
Per ogni voce, poi, il programma, secondo opportuni standards, assegna automaticamente un simbolo che indichi il livello di qualità associato:


ottima qualità del dataset rispetto al parametro in esame;


qualità scarsa che necessita di verifiche;


qualità del tutto insufficiente: è necessario prestare attenzione.

Per ogni insieme di frammenti, pertanto, si propone una valutazione complessiva della qualità, con le indicazioni della bontà di lettura secondo i differenti parametri. Le analisi vengono compiute o sulla qualità complessiva delle diverse sequenze, o sulla valutazione della qualità media tra tutte le sequenze nella lettura delle diverse basi.
A causa del processo di lettura, infatti, la qualità varia a seconda di quante basi del frammento siano già state lette (in particolare, le prime basi lette vengono identificate univocamente, mentre più il processo va avanti, più la qualità si abbassa).

La Figura 11 propone, a titolo di esempio, la valutazione complessiva del dataset relativo al batterio LIST1 (Listeria) utilizzato per le successive analisi.

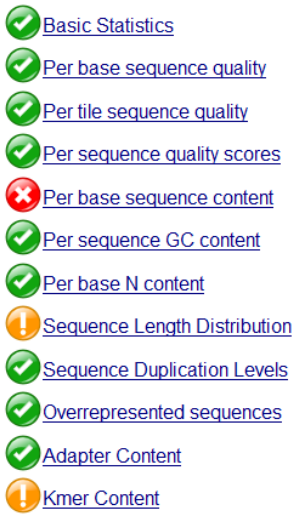


***Figura 11**: Valutazione della qualità complessiva del dataset relativo al batterio LIST1.*

Per ragioni di brevità, si è scelto di non riportare completamente l'analisi dettagliata della qualità relativa a tutti i campioni del batterio Listeria, come da schema riassuntivo proposto in Figura 11.
Poiché però si è parlato di qualità delle sequenze, risulta perlomeno utile prendere in considerazione il diagramma di figura 12, dove viene mostrata un'overview dell'intervallo dei valori di qualità per tutte le basi in ogni posizione nel file FastQ.

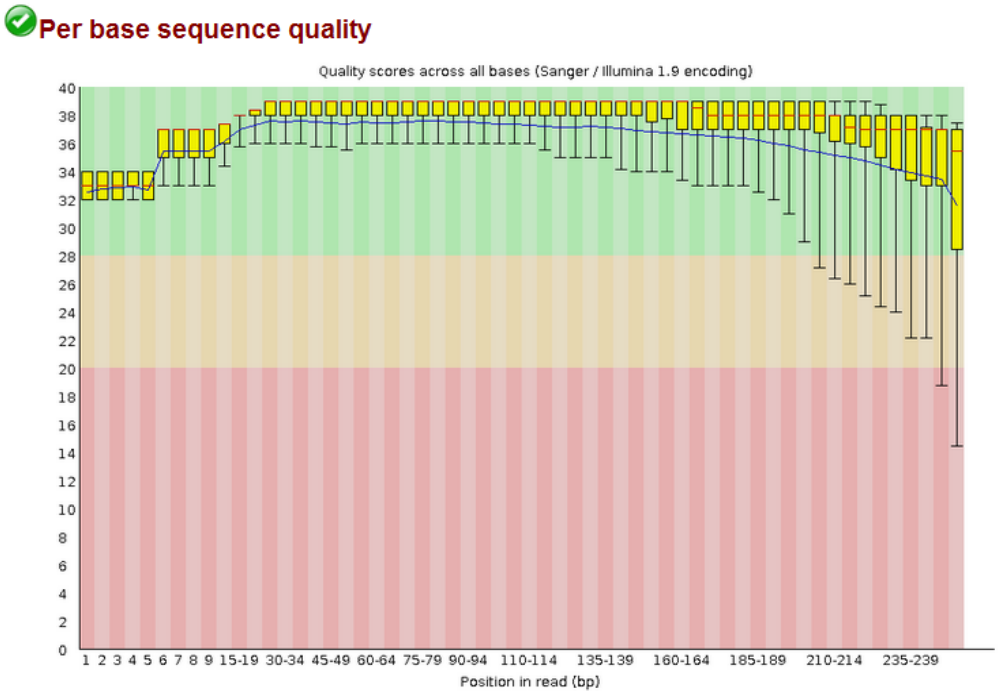


***Figura 12**: FastQC: qualità per ogni base della sequenza*

Per ogni posizione nella read è riportato un box plot che ha come elementi:

- La linea centrale rossa che rappresenta la mediana.
- I box gialli mostrano il range interquartile.
- La linea blu rappresenta la qualità media.

Sull'asse *y* del grafico è presente il *quality score*: più alto è questo valore e migliore è la chiamata della base. Sullo sfondo si possono distinguere tre colori che corrispondono a una chiamata molto buona della base (in verde), una chiamata di qualità rivedibile (arancione) e una di scarsa qualità (rosso). Nella maggior parte delle piattaforme che eseguono il sequencing, la qualità della chiamata si degrada con l'andare del processo: di conseguenza, è abbastanza comune vedere che verso la fine della read, le chiamate cadano nella zona arancione/rossa.

Tendendo in considerazione quanto detto poco sopra, questi risultati portano a concludere che il dataset inerente al batterio preso come esempio (LIST1), possiede complessivamente una buona qualità: come si può notare, tutte le reads cadono nella zona verde, la quale implica una qualità elevata. Tale procedimento è stato ripetuto per tutti gli altri campioni, nei quali si sono ottenuti risultati che hanno permesso di procedere verso lo step successivo.

2.5 Controllo di qualità - Trimmomatic

Il passo successivo della sequenza prende in considerazione il controllo di qualità delle reads.
In base ai risultati ottenuti con FastQC, è possibile decidere di ripulire il dataset delle reads, per facilitare e rendere più accurate le analisi successive. Le possibili operazioni sono:

- Scartare le estremità di bassa qualità.
- Scartare le reads con bassa qualità media.
- Scartare le reads che dopo le operazioni di trimming rimangono troppo corte.
- Rimuovere gli adattatori (adapters).

Al fine di eseguire una o più delle procedure elencate sopra, risulta necessario applicare un processo che prende il nome di "***trimming*** delle reads", che consiste fondamentalmente nel tagliare i frammenti. Esistono due metodologie di *trimming* che si possono utilizzare:

1) Trimming statico: si tagliano tutte le reads nello stesso punto;

2) Trimming dinamico (flexible): le reads vengono tagliate sia dall'estremità 5' che dall'estremità 3', finché la qualità resta sotto un valore di soglia definito. Le reads, quindi, non avranno più tutte la stessa lunghezza. Il vantaggio di questa seconda tecnica è quello di ottenere reads più corte, ma di maggiore qualità media.

Nel seguente lavoro di tesi si è scelto di utilizzare il software **Trimmomatic** [7], uno strumento capace non solo di leggere e tagliare i dati FASTQ prodotti da Illumina, ma anche di rimuovere gli adattatori.
Infatti, quando i dati sono sequenziati tramite Illumina, a questi vengono aggiunti degli adattatori per i frammenti, in modo tale che si aggancino alle "beads". Se questi adattatori non vengono rimossi, possono causare un assemblaggio non corretto o altri problemi. Inoltre, la qualità delle sequenze dipende anche dalla lunghezza delle reads, e le zone che possiedono un basso *quality score*, possono essere tagliate utilizzando Trimmomatic.

Le immagini che seguono mostrano un esempio di quello che accade alle reads una volta subìto il processo di trimming. In particolare, risulta evidente come e quanto la lunghezza delle reads nell'estremità destra sia stata notevolmente ridotta, dando luogo a una distribuzione più uniforme del quality score.

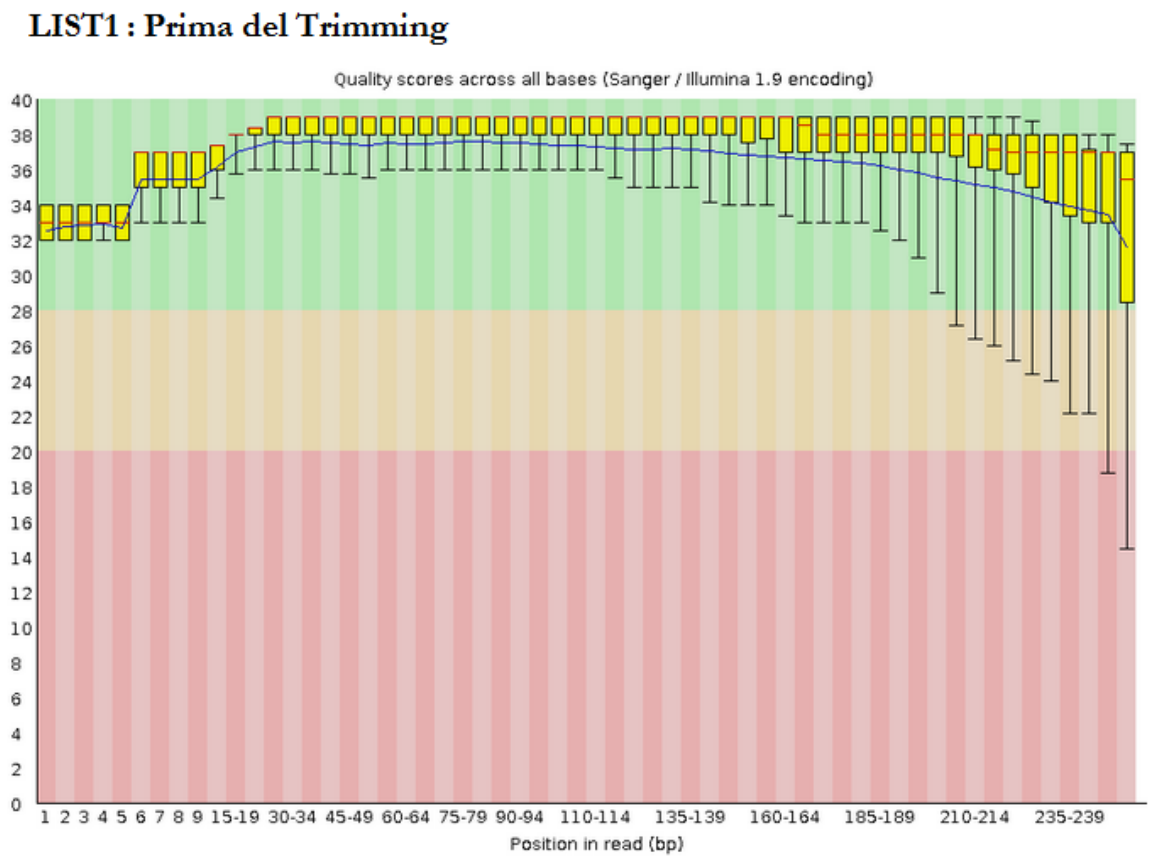


***Figura 13**: Quality score delle reads prima del trimming*

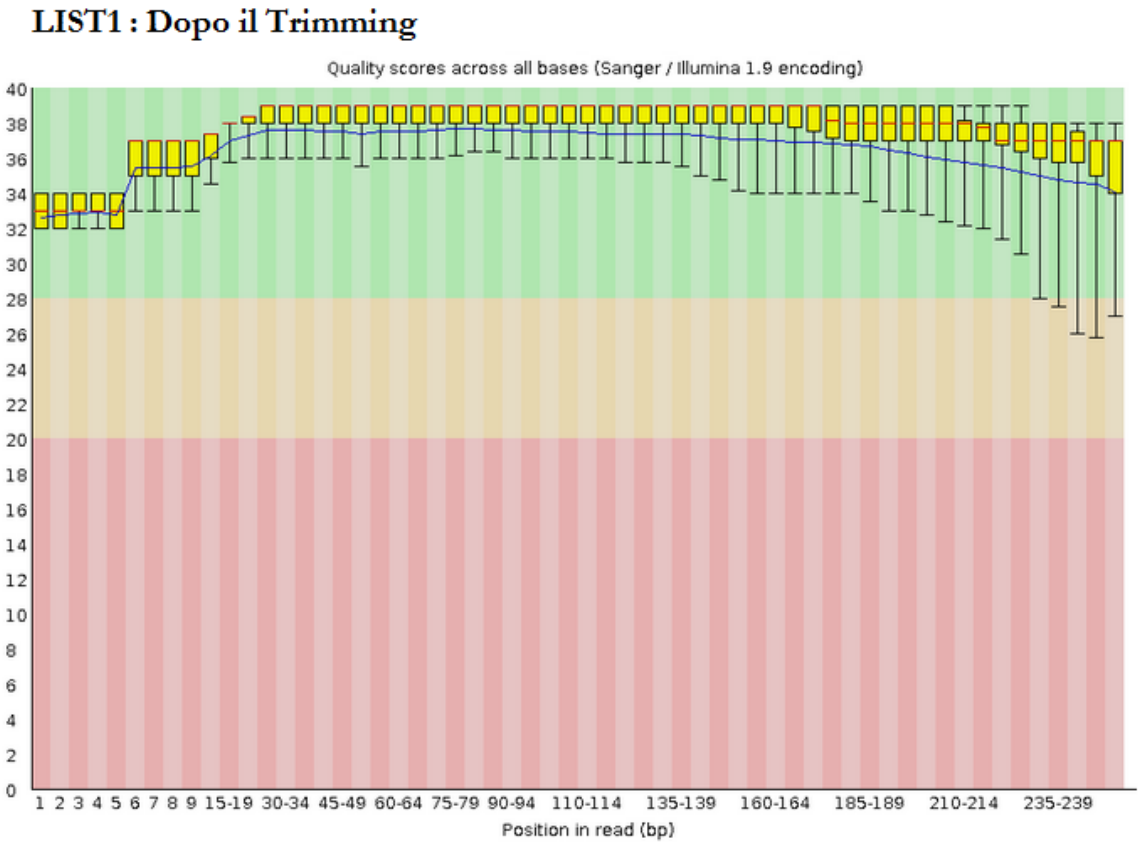


***Figura 14**: Quality score delle reads dopo il trimming.*

2.6 Allineamento al genoma di riferimento – BWA

Dopo aver raccolto, sequenziato i frammenti di DNA e verificato la qualità della lettura, si procede con l'allineamento delle reads al genoma di riferimento.

L'esigenza di un allineamento locale efficace, ha portato allo sviluppo di numerose tecniche di allineamento migliori della semplice ricerca di stringhe, con la valutazione in sequenza di tutto il genoma. Il numero di operazioni per compiere questa ricerca, infatti, è molto elevato e dell'ordine di $O(L_gL_rN_r)$ con $\mathbf{L_g}$ lunghezza del genoma, $\mathbf{L_r}$ lunghezza dei frammenti e $\mathbf{N_r}$ numero di frammenti da analizzare. Tale valore, nel caso del genoma umano, risulta molto alto contando all'incirca 3.3B di basi.
A partire dagli anni 2000 sono stati sviluppati numerosi algoritmi di allineamento locale. Alcuni di questi sono basati sulla scansione di tutto il genoma (come MAQ); altri, invece, sulla definizione della trasformata di **Burrows-Wheeler** (come SOAP2, Bowtie o BWA).
La trasformata di Burrows-Wheeler, oltre ad essere molto adatta alla compressione di stringhe e pertanto utilizzata in programmi come **bzip2**, è assai efficace per l'allineamento di brevi sequenze al genoma, soprattutto perché consente di memorizzare sinteticamente molte caratteristiche delle stringhe da analizzare, che altrimenti occuperebbero un'eccessiva quantità di memoria.

In questa sezione si prende in considerazione la tecnica del Burrows-Wheeler Alignemnt (**BWA**) [8], sviluppata da Heng Li e Richard Dublin che permette di allineare brevi frammenti alla sequenza del genoma umano, consentendo *mismatches* e introduzione di spazi. Questo algoritmo permette di allineare efficacemente i frammenti al genoma; dalla simulazione condotta dagli stessi H. Li e R. Durbin e riportata nel dettaglio in [9] sull'allineamento di frammenti lunghi 30 basi, si ottiene, per l'allineamento single-end, un'accuratezza dell'80.6%, risultato migliore degli altri algoritmi basati sulla trasformata di Burrows-Wheeler. Fondamentalmente il pacchetto software BWA è formato da tre algoritmi:

- BWA-backtrack
- BWA-SW
- BWA-MEM

Il primo di questi è stato creato per delle sequenze di reads in grado di contenere fino a 100bp, mentre gli altri due analizzano sequenze che variano da 70bp fino a 1Mbp.
BWA-MEM e BWA-SW condividono infatti caratteristiche simili, ma BWA-MEM, l'ultimo creato, è generalmente consigliato per queries specifiche per il fatto che risulta più veloce e accurato, motivo per cui è stato scelto nel seguente lavoro di tesi.

Una volta lanciato BWA-MEM, questo genera in output un file di tipo *.sam* (*sequence alignment map*), che contiene tutte le informazioni sul procedimento di allineamento e sul risultato.
In figura 15 si presenta un esempio di file *.sam* in cui si evidenzia la struttura che lo

compone. Esiste, infatti, una prima parte di intestazione (header) e una seconda in cui sono presentati nel dettaglio i risultati della procedura (alignment).

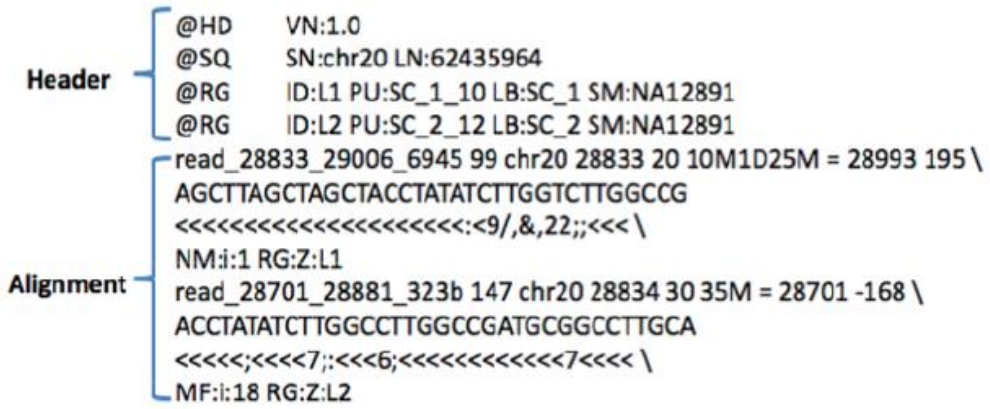


***Figura 15**: Esempio di file .sam*

La sezione di intestazione comprende generalmente:

- la definizione dell'indice di riferimento dell'allineamento (@HD);
- alcune informazioni sulla sequenza di riferimento, come la localizzazione sul genoma e la sua lunghezza (@SQ);
- alcune caratterizzazioni dei frammenti allineati, come l'identificativo del centro che li ha prodotti e la piattaforma con cui sono stati generati (@RG);
- le caratteristiche del tool che è stato utilizzato per l'allineamento, come il nome del software e la versione (@PG).

La sezione di allineamento contiene, invece, undici campi obbligatori che includono non solo delle informazioni generali sull'alignment, ma anche sulle caratteristiche dettagliate dei frammenti e della sequenza.
L'ultimo campo, infine, è un codice di qualità identificato con la codifica ASCII, che valuta nel complesso l'esito del processo di allineamento. Data la complessità del formato, esistono numerosi tool (samtools) che permettono l'estrazione di informazioni rilevanti dai file .sam e li convertono poi in formati più leggibili: ad esempio, il formato *.bed* che contiene semplicemente l'elenco dei frammenti allineati e la loro localizzazione sul genoma di riferimento.

2.6.1 Trasformata di Burrows – Wheeler (BWT)

La trasformata di Burrows-Wheeler[10] (abbreviata con BWT) è un algoritmo utilizzato in innumerevoli applicazioni per la compressione dei dati.

La trasformata, che costituisce gran parte dell'algoritmo, fu sviluppata da Wheeler nel 1983 e viene attualmente utilizzata come primo stadio di trasformazione della stringa da comprimere in molti programmi di compressione commerciali, primo tra tutti **bzip2**.

La trasformata BWT, prima di iniziare la fase di decodifica, deve necessariamente avere a disposizione l'intera sequenza codificata e, se applicata ad una stringa di caratteri, non effettua una vera e propria compressione, ma permuta in maniera reversibile l'ordine dei caratteri stessi.

Data quindi una sequenza $x_1\, x_2 \ldots x_N$ di lunghezza *N*, se ne creano *N - 1* permutazioni, ciascuna delle quali è uno shift ciclico della sequenza originale che prende il nome di *rotazione*. In pratica, si ottiene la *i*-esima rotazione concatenando la sottostringa degli ultimi caratteri con la sottostringa dei primi *N - i*, ottenendo le seguenti permutazioni della stringa di partenza:

$$x_{N-i+1} x_N x_1 x_2 \ldots x_{N-i} \qquad \forall i = 1,2,\ldots,N-1$$

Si riordinano poi le *N* sequenze in ordine lessicografico. Il codificatore invia la sequenza di lunghezza *N,* che si compone prendendo l'ultima lettera di ogni rotazione. La ricostruzione della stringa può essere assicurata in due diversi modi. Nel primo modo, il codificatore invia la posizione occupata dalla sequenza originale nella lista ordinata di rotazioni. Nel secondo modo, invece, il codificatore aggiunge un carattere speciale alla fine della stringa prima di effettuare la trasformazione di Burrows-Wheeler.
La sequenza così trasformata è strutturata in modo che si ottiene una compressione particolarmente efficiente. Viene presentato, ora, un esempio di trasformata di Burrows-Wheeler, per renderne più facile la comprensione.

Si prenda in considerazione la stringa campione "*BANANA"*. Si aggiunga ad essa un carattere finale facendo in modo che quest'ultimo non compaia all'interno del testo, al fine di distinguere l'ultimo carattere, quando si ricostruisce la stringa. Otteniamo una nuova stringa *BANANA@*, che costituirà la stringa di input.
Si costruisce, ora, una tabella con tutte le possibili rotazioni della stringa, ed infine si riorganizzano in ordine alfabetico le sue righe, come mostrato in figura 16.

B	A	N	A	N	A	@		A	N	A	N	A	@	**B**
A	N	A	N	A	@	B		A	N	A	@	B	A	**N**
N	A	N	A	@	B	A		A	@	B	A	N	A	**N**
A	N	A	@	B	A	N	⟹	B	A	N	A	N	A	**@**
N	A	@	B	A	N	A		N	A	N	A	@	B	**A**
A	@	B	A	N	A	N		N	A	@	B	A	N	**A**
@	B	A	N	A	N	A		@	B	A	N	A	N	**A**

Figura 16*: Esempio di trasformata di Burrows-Wheeler: Passo (1): elenco di tutte le rotazioni possibili sulla stringa "BANANA@"; Passo (2): riordinamento delle righe in ordine alfabetico.*

La stringa in output sarà l'ultima colonna della tabella riordinata alfabeticamente, cioè "*BNN@AAA*", stringa che contiene l'ultimo carattere di ognuna delle rotazioni effettuate. Viene descritto ora l'algoritmo inverso.

Si consideri una tabella vuota con righe e colonne pari al numero dei caratteri della stringa permutata. Conoscendo soltanto l'informazione della stringa permutata, è possibile ricostruire facilmente la stringa originale. L'ultima colonna mostra, infatti, quali siano i caratteri del file originale. Basterà soltanto riordinarli alfabeticamente, per ottenere quella che nella precedente figura 16 riordinata, costituiva la prima colonna. A questo punto, affiancando l'ultima colonna alla prima, si ottengono le coppie di caratteri successivi del file originale. Tali coppie, vengono ordinate alfabeticamente: in questo modo si ottiene la prima e la seconda colonna della figura 17. Affianchiamo a queste, di nuovo, l'ultima colonna ottenendo le triple di caratteri successivi del file originale. Continuando in questo modo, possiamo ricostruire l'intera stringa. A questo punto, la colonna contenente il carattere che indica la fine del testo come carattere iniziale rappresenta la stringa ordinata.

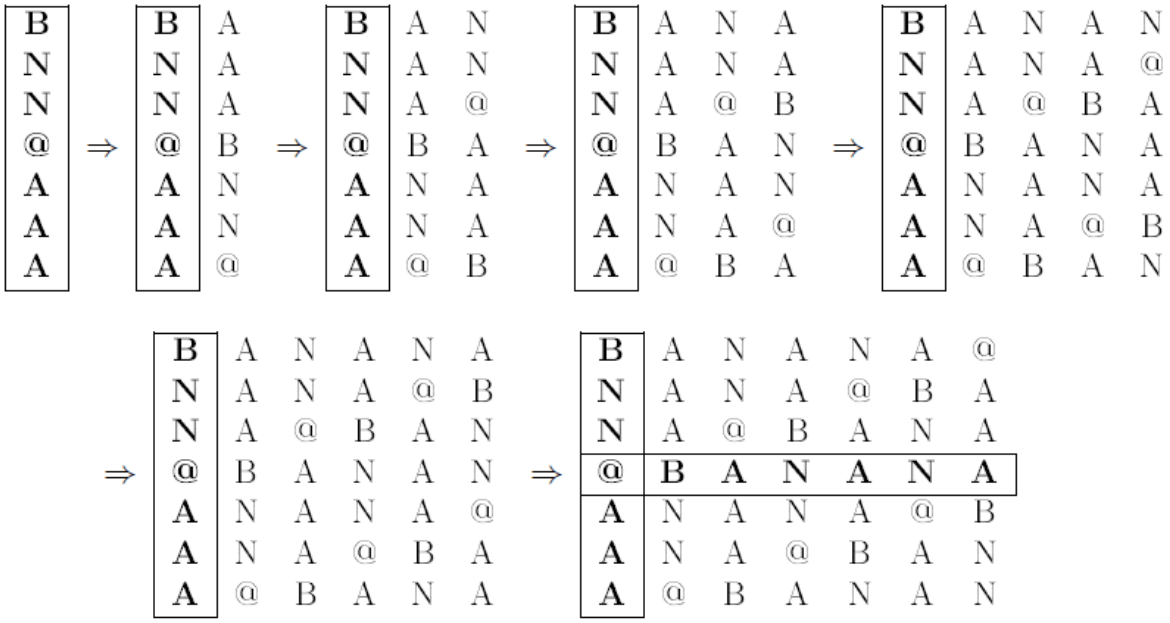


Figura 17: *Ricostruzione della stringa originale dalla stringa "BNN@AAA" trasformata attraverso il metodo BWT. Ad ogni passo si crea una nuova tabella riordinando le righe della tabella corrente ed aggiungendo ad esse la stringa "BNN@AAA" come prima colonna.*

2.7 Alignment Processing - PICARD

Come è stato descritto poco sopra, il file di tipo .sam, prodotto da BWA, deve essere manipolato in modo tale da poterne estrarre delle informazioni specifiche.
A tal fine, viene utilizzato il software PICARD-tools [11], che consiste in un insieme di linee di comando scritte in codice Java, adatto per la manipolazione di dati HTS (***High-Throughput Sequencing***) e formati come SAM/BAM/CRAM e VCF.

Vengono elencati di seguito i tools utilizzati per questa fase di alignment processing:

- PICARD - SortSam
- PICARD – MarkDuplicate
- PICARD – BuildBamIndex

Proprio come suggerisce il nome, il modulo *SortSam* ordina e indicizza un file di tipo SAM, producendo in output un file di indici in formato *.bam* (la versione binaria e compressa del SAM).
A questo punto, l'applicazione "*Mark Duplicate Reads*", è stata utilizzata per contrassegnare e rimuovere le reads duplicate (duplicati ottici e artefatti di PCR) presenti nel file BAM.
Eseguendo, infine, l'ultimo tool (*Build Bam Index*), questo genera un file ".bai" che è formato sostanzialmente da un indice del file BAM. Tale procedimento permette una ricerca più veloce dei dati nel file BAM stesso e svolge, in poche parole, la stessa funzione dell'indice in un database.

2.8 Variant Calling – GATK

La chiamata delle varianti, altrimenti detta **variant calling,** è la fase dell'analisi bioinformatica che segue quella dell'alignment. Infatti, una volta che le sequenze dei geni sono state ricostruite tramite l'allineamento delle reads ottenute nel sequenziamento, occorre individuare tutti i punti nei quali i geni differiscono dalle sequenze del genoma di riferimento (reference sequences). Queste varianti saranno naturalmente numerosissime e in massima parte irrilevanti, trattandosi, per lo più, di semplici polimorfismi alla base delle differenze interindividuali.

Il variant calling viene eseguito in automatico grazie all'ausilio di un software (nel seguente lavoro di tesi è stato utilizzato GATK) e, una volta terminata l'operazione, il risultato viene generalmente salvato in un file VCF (*Variant Call Format*).
Uno dei problemi maggiori di un'operazione di variant calling è dato dalla difficoltà nel riuscire a distinguere le varianti vere da quelle irreali, dovute ad artefatti del sequenziamento o ad errori nella fase di alignment. È poi dalla qualità del risultato di tale operazione che dipende la probabilità di identificare, o meno, la mutazione-malattia; oppure, nel caso degli studi di popolazione, la possibilità di determinare in modo affidabile la frequenza allelica dei polimorfismi.

Tuttavia, non è così semplice identificare queste varianti: infatti, volendo scendere più nel dettaglio, ci sono tre fattori che complicano maggiormente tale processo:

1. La **presenza di indels (inserzioni/delezioni)**, che possono essere erroneamente scambiati per varianti di singolo nucleotide (SNV).
2. **Errori di sequenziamento** dovuti ad artefatti della PCR (Polymerase Chain Reaction), i quali sono meno frequenti laddove si usino sistemi NGS basati sulle paired-end reads.
3. Variabilità della qualità del sequenziamento in corrispondenza delle **estremità delle reads.**

Alcuni tra i software di Variant Calling più utilizzati sono ATLAS 2, SOAPsnp, VarScan e GATK.

Il tool utilizzato nel seguente lavoro di tesi, che oltretutto risulta anche essere quello più sfruttato per l'identificazione e l'analisi delle varianti, è fornito dal software **GATK** (***Genome***

Analysis Toolkit) [12], lo stesso utilizzato nel progetto 1000 Genome Project e in The Cancer Genome Browser.
Proprio come PICARD e BWA, il software GATK è formato da numerosi tools che hanno come obiettivo primario quello di identificare le varianti. I tools presi in considerazione sono:

- GATK – HaplotypeCaller
- GATK - SelectVariants
- GATK – VariantFiltration

Il primo di questi, HaplotypeCaller, prende in input un file di tipo *.bam* restituendo in uscita un file VCF (Variant Call Format), ed è in grado di chiamare SNPs e indels contemporaneamente, tramite l'assemblaggio locale *de novo* di aplotipi (combinazioni di varianti alleliche) in una regione attiva. In altre parole, ogni volta che il programma incontra una regione che mostra segni di variazione, scarta le informazioni di mappatura esistenti e ricompone completamente le reads in quella regione.
Poiché spesso un file VCF contiene molti campioni e/o varianti, risulterà necessario creare un sottoinsieme di questi/e, in modo tale da facilitare i processi di analisi che seguiranno.

Il tool SelectVariant può essere utilizzato per questo proposito. Questo sottoinsieme di varianti, viene selezionato in base a:

- Criteri specifici che garantiscono l'inclusione (nel sottoinsieme) imponendo delle soglie relative a certi valori di annotazione, ad esempio: "DP > 1000" (profondità del coverage maggiore di 1000x) "AF < 0.25" (siti con frequenza allelica minore di 0.25).
- Tracce di concordanza o discordanza al fine di includere o escludere le varianti che sono presenti anche in altri callsets.
- Criteri come il loro tipo (ad esempio solo indels), lo stato di filtraggio, allelicità e così via…

L'ultimo tool preso in esame, VariantFilatration, è stato progettato per un filtraggio "pesante" delle varianti chiamate, ed è basato su determinati criteri. Questa operazione viene fatta andando a modificare alcuni campi all'interno della stringa che lancia il programma.
Il file restituito è sempre un formato VCF, ma questa volta si tratterà di un VCF "hard-filtered".

2.8.1 File VCF

Il file VCF [13] è un formato standardizzato generico, utilizzato per la memorizzazione della maggior parte delle varianti genetiche esistenti, tra cui SNPs, indels e varianti strutturali, associate ad annotazioni libere.

Tale tipo di file si divide principalmente in due parti:

- Una sezione di header
- Una sezione di dati.

L'intestazione (*header*) fornisce dei cosiddetti "meta-dati" i quali descrivono il corpo del file stesso. Ogni linea di meta-dati comincia con il delimitatore ##, mentre la linea di definizione della struttura comincia con il carattere #.
Le meta-informazioni 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 stesso.
Nel codice di esempio mostrato in figura sottostante, è illustrato un file VCF con varie meta informazioni e quattro 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.

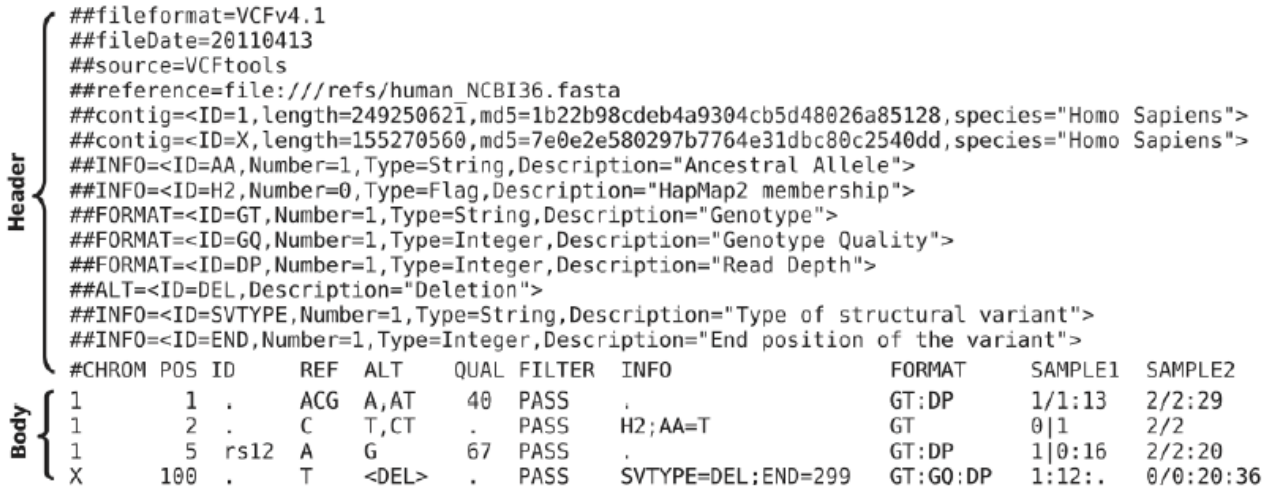


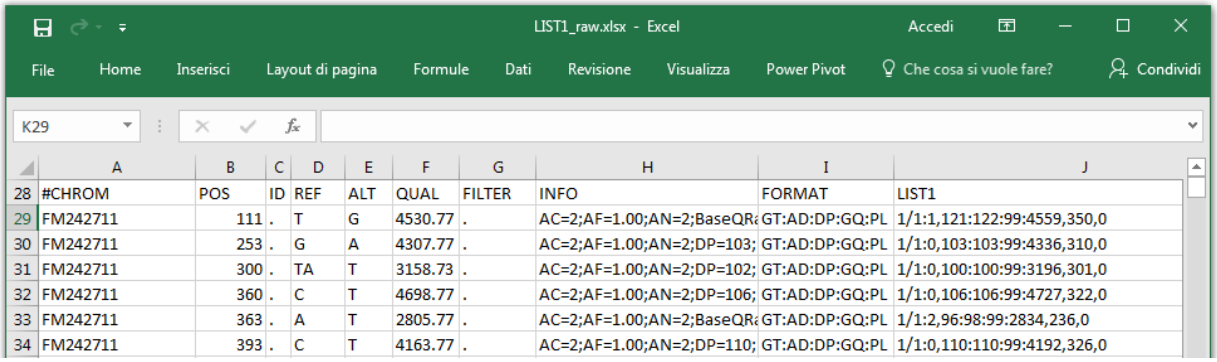
***Figura 18**: Esempio di un formato valido VCF*

Il corpo (*body*) del VCF segue l'intestazione, ed è sostanzialmente suddiviso in otto colonne.
Nella tabella sottostante vengono elencati i nomi delle colonne contenute nel corpo del file con la relativa descrizione.

	Nome	Descrizione
1	**CHROM**	Nome della sequenza (tipicamente un cromosoma) nella quale è stata chiamata la variazione. Questa sequenza è di solito conosciuta come 'sequenza di riferimento ', ovvero la sequenza nella quale si possono vedere le variazioni del nostro campione.
2	**POS**	Posizione della variazione nella sequenza data.
3	**ID**	Identificatore della variazione.
4	**REF**	La base di riferimento (o basi nel caso di un indel) in una data posizione in una sequenza di riferimento.
5	**ALT**	La lista degli alleli alternativi in quella posizione
6	**QUAL**	Rappresenta il "Quality Score" associato all'inferenza di un dato allele.
7	**FILTER**	È un flag che indica quale insieme di filtri la variazione ha superato.
8	**INFO**	Lista di parole chiave (campi) che descrivono la variazione. Campi multipli sono separati sa un ";" con valori opzionali.
9	**FORMAT**	Lista (opzionale) di campi che descrivono il campione.
+	**SAMPLES**	Per ogni campione descritto nel file, sono forniti i valori per i campi elencati in FORMAT.

***Tabella 2**: descrizione di tutti i campi del formato VCF.*

Una volta ottenuto il file VCF, questo viene caricato su un file di tipo *.xlsx* (Excel) per permettere una migliore visualizzazione dei dati contenuti.
La figura 19 mostra un esempio di estrazione dei dati *raw* per il campione LIST1. Come si può notare, tutti i campi presenti rispecchiano la struttura tipica del file VCF.



	A	B	C	D	E	F	G	H	I	J
28	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	LIST1
29	FM242711	111	.	T	G	4530.77	.	AC=2;AF=1.00;AN=2;BaseQR	GT:AD:DP:GQ:PL	1/1:1,121:122:99:4559,350,0
30	FM242711	253	.	G	A	4307.77	.	AC=2;AF=1.00;AN=2;DP=103;	GT:AD:DP:GQ:PL	1/1:0,103:103:99:4336,310,0
31	FM242711	300	.	TA	T	3158.73	.	AC=2;AF=1.00;AN=2;DP=102;	GT:AD:DP:GQ:PL	1/1:0,100:100:99:3196,301,0
32	FM242711	360	.	C	T	4698.77	.	AC=2;AF=1.00;AN=2;DP=106;	GT:AD:DP:GQ:PL	1/1:0,106:106:99:4727,322,0
33	FM242711	363	.	A	T	2805.77	.	AC=2;AF=1.00;AN=2;BaseQR	GT:AD:DP:GQ:PL	1/1:2,96:98:99:2834,236,0
34	FM242711	393	.	C	T	4163.77	.	AC=2;AF=1.00;AN=2;DP=110;	GT:AD:DP:GQ:PL	1/1:0,110:110:99:4192,326,0

***Figura 19** : esempio di un file VCF trattato con Excel.*

La parte di analisi dati che viene presentata nel capitolo 4, verte principalmente sullo studio del campo INFO del file VCF. Tale campo contiene al suo interno fino a 14 parametri utilizzati da

GATK per la chiamata delle varianti. Nonostante non siano visibili tutti (per ragioni di spazio), i parametri contenuti vengono indicati come segue:
AC, AF, AN, DP, FS, MLEAC, MLEAF, MQ, QD, SOR, BaseQRankSum, ClippingRankSum, ReadPosRankSum, MQRankSum.
La maggior parte di questi, i primi 10 solitamente, è comune a tutti gli SNPs, mentre gli ultimi 4 elencati, non possono essere calcolati per i siti omozigoti; cioè tali parametri compaiono soltanto nei siti eterozigoti, ovvero dove esistono diversi alleli alternativi allineati sul riferimento.
Al fine di favorire una migliore comprensione di ciò che si andrà ad analizzare, di seguito vengono elencati i parametri [14] e una loro breve descrizione:

- **Allele Count (AC)** – è il numero di alleli chiamati in un campione.
- **Allele Frequency (AF)** - è la misura della frequenza relativa di un allele in un locus genico nella popolazione.
- **Allele Number (AN)** – è il numero totale di alleli nel genotipo chiamato.
- **Depth of Coverage (DP)** - numero di volte che un nucleotide è stato "letto" durante il sequenziamento. È una misura di ridondanza. All'aumentare del coverage aumenta il grado di confidenza nel risultato del sequenziamento.
- **Fisher Strand (FS)**- L'annotazione "Fisher Strand" si riferisce a uno dei numerosi metodi che permettono di valutare se c'è un bias di strand nei dati. Questo metodo utilizza il Test esatto di Fisher per determinare se c'è un bias di strand nel filamento positivo o negativo del DNA di riferimento.
- **Maximum Likelihood Expectation for the Allele Counts (MLEAC)** per ogni allele alternativo, nello stesso ordine di come è elencato.
- **Maximum Likelihood Expectation for the Allele Frequency (MLEAF)**, per ogni allele alternativo, nello stesso ordine di come è elencato.
- **RMSMapping Quality (MQ)** – Questo parametro fornisce una stima della qualità della mappatura delle reads che supportano una chiamata variante.
  Vengono prodotti sia i dati grezzi (somma di quadrati e numero delle reads totali) che il valore quadratico medio (RMS). I dati grezzi vengono utilizzati per calcolare con precisione l'errore quadratico medio quando si prende in considerazione più di un campione.
- **Quality by Depth (QD) -** QD è il punteggio di qualità (QUAL) normalizzato per l'allele depth (AD) per una data variante. Per un singolo campione, *HaplotypeCaller* calcola il QD utilizzando il rapporto $\frac{\mathrm{QUAL}}{\mathrm{AD}}$.

- **Strand Odds Ratio (SOR):** Tale parametro rappresenta uno dei numerosi metodi che permettono di valutare se esiste un bias di strand nei dati.
  Il bias di strand è un tipo di errore nel sequenziamento in cui un filamento di DNA è favorito rispetto a un altro, e può provocare una non corretta valutazione della quantità di prove osservate per un allele rispetto all'altro.

Questo parametro, quindi, è una forma aggiornata del Fisher Strand Test, e tiene conto della grande quantità di dati nelle situazioni di alto coverage.

- **BaseQRankSum:** La colonna *BaseQRankSum* contiene una valutazione dei punteggi di qualità nelle reads che hanno una chiamata variante, la quale viene paragonata con il punteggio di qualità dell'allele di riferimento. Le varianti per le quali non è stato chiamato un corrispondente allele di riferimento, non possiedono un valore di BaseQRankSum. Allo stesso modo, non viene calcolato alcun valore per gli alleli di riferimento.
- **ClippingRankSum** - Questo parametro testa se i dati che supportano l'allele di riferimento mostrano più o meno un *base clipping* rispetto a quelli che supportano l'allele alternativo. Il risultato ideale è un valore prossimo allo zero, il che indica che c'è poca o nessuna differenza. Un valore negativo indica che le reads che supportano l'allele alternativo possiedono più basi *hard-clippled*, rispetto a quelle che supportano l'allele di riferimento. Al contrario, un valore positivo indica che le reads che supportano l'allele alternativo hanno un minor numero di basi *hard-clipped* rispetto a quelle che supportano l'allele di riferimento. Trovare una differenza statisticamente significativa in entrambi i casi, suggerisce che il processo di sequenziamento e/o la mappatura, possono essere stati affetti da un errore o da un artefatto.

- **ReadPosRankSum** – Esegue un test per vedere se c'è una evidenza di bias nella posizione degli alleli (all'interno delle reads che li supportano), tra gli alleli di riferimento e quelli alternativi.
  Vedere un allele solo vicino alle estremità delle reads è indicativo di un errore, perché è proprio lì che i sequenziatori tendono a fare il maggior numero di errori. Tuttavia alcune varianti, situate in prossimità dei bordi della regione sequenziata, saranno necessariamente coperte dagli estremi della reads, in modo tale che non si possa semplicemente fissare una soglia assoluta di "distanza minima dalla fine delle reads". Questo è il motivo per cui viene usato un *rank sum test*, cioè per valutare se c'è una differenza nel modo in cui l'allele di riferimento e l'allele alternativo sono supportati.
  Il risultato ideale è un valore prossimo allo zero, il che indica che c'è poca o nessuna differenza nella posizione in cui si trovano gli alleli rispetto alle estremità delle reads.
  Un valore negativo indica che l'allele alternativo si trova alle estremità delle reads più frequentemente rispetto all'allele di riferimento. Viceversa, un valore positivo indica che l'allele di riferimento si trova alle estremità delle reads più spesso rispetto all'allele alternativo.

- **MQRankSum** – Questa annotazione confronta la qualità di mappatura delle reads che supportano l'allele di riferimento con quelle che supportano l'allele alternativo. Anche in questo caso, il risultato ideale è un valore prossimo allo zero, il che indica che c'è poca o nessuna differenza. Un valore negativo indica che le reads che supportano l'allele alternativo hanno un *quality score* di mappatura più basso rispetto a quelli che supportano l'allele di riferimento. Al contrario, un valore positivo indica che le reads che supportano

l'allele alternativo hanno un *quality score* di mapping più alto rispetto a quelle che supportano l'allele di riferimento.

Capitolo 3

Nel seguente capitolo vengono descritti e spiegati nel dettaglio quali sono i test statistici utilizzati per analizzare i dati ottenuti. In particolare, ciò che si andrà a fare, sarà analizzare dal punto di vista statistico, alcuni parametri degli SNPs trovati (ad esempio il coverage), i cui risultati saranno esposti nel capitolo 4.

3.1 Analisi della Varianza (ANOVA)

L'analisi della varianza è un metodo sviluppato da Fisher, che è fondamentale per l'interpretazione statistica di molti dati biologici ed è alla base di molti disegni sperimentali. L'analisi della varianza [15] (in inglese: Analysis of Variance, abbreviata con l'acronimo ANOVA) è utilizzata per testare le differenze tra medie campionarie e per fare questo si prendono in considerazione le rispettive varianze.

Il principio alla base di questo test è quello di stabilire se due o più medie campionarie possono derivare da popolazioni che hanno la stessa media parametrica. Quando le medie sono solamente due è indifferente usare questo test o il ***t***-test, mentre si deve necessariamente utilizzare l'ANOVA quando le medie sono più di due, o quando si vuole suddividere la variabile di raggruppamento in più variabili, per eliminare eventuali fonti di variazione oltre a quella prodotta dal fattore di cui si vuole valutarne l'effetto.
Grazie allo "Statistics and Machine Learning Toolbox™" fornito da MatLab, è possibile eseguire l'analisi della varianza (ANOVA) di tipo 1, di tipo 2 e di tipo N; l'analisi multivariata della varianza (MANOVA); e l'analisi di covarianza (ANCOVA). Tuttavia nel seguente lavoro di tesi si è scelto di prendere in considerazione solo l'ANOVA di tipo 1.

3.1.1 One-way ANOVA

Tramite il comando `anova1` di MatLab, è possibile lanciare la funzione che permette di eseguire l'analisi di varianza di tipo 1. Lo scopo di tale analisi è quello di determinare se i dati provenienti da diversi gruppi hanno una media comune. Cioè, l'ANOVA di tipo 1, permette di scoprire se i diversi gruppi di una variabile indipendente, hanno effetti diversi sulla variabile di risposta y.

Questo tipo di analisi (ANOVA-1) è un caso speciale del modello lineare, ed è esprimibile mediante la seguente equazione:

$$y_{ij} = \alpha_j + \varepsilon_{ij}$$

dove y_{ij} è un'osservazione, nel quale *i* rappresenta il numero di osservazioni e *j* un diverso gruppo, α_j è la media della popolazione, ε_{ij} è il "residuo" o errore sperimentale con media zero e varianza costante.

L'analisi della varianza esegue un test per verificare le differenze delle medie tra i gruppi, scomponendo la varianza totale in due componenti:

- Varianza interna ai gruppi (anche detta *Varianza Within*) definita come: $y_{ij} - \bar{y}$
- Varianza tra i gruppi (*Varianza Between*) uguale a: $\bar{y}_j - \bar{y}$; dove $\bar{y}_j$ è il campione medio del gruppo *j* e $\bar{y}$ è il campione medio totale.

In altre parole, ANOVA separa la somma totale dei quadrati (SST) in una somma di quadrati dovuta all'effetto tra gruppi (SSR) e una somma di errori quadratici (SSE), per cui:

$$\underbrace{\sum_i \sum_j (y_{ij} - \bar{y})^2}_{SST} = \underbrace{\sum_j n_j (\bar{y}_j - \bar{y})^2}_{SSR} + \underbrace{\sum_i \sum_j (y_{ij} - \bar{y}_j)^2}_{SSE}$$

Dove n_j è la grandezza del campione per il *j*-esimo gruppo.

A questo punto l'ANOVA confronta la varianza tra gruppi con la varianza interna ai gruppi. Se il rapporto della varianza interna ai gruppi con la varianza tra gruppi risulta essere significativamente alto, allora si può concludere che le medie dei gruppi sono significativamente differenti tra loro. Questo rapporto si può misurare usando un test statistico che ha una distribuzione F con (k-1, N-1) gradi di libertà.

$$F = \frac{SSR/_{k-1}}{SSE/_{N-k}} = \frac{MSR}{MSE} \sim F_{k-1,N-k}$$

Dove MSR e MSE sono quadrati medi della regressione e dell'errore, *k* è il numero di gruppi e *N* è il numero totale di osservazioni. Se il *p-value* per la statistica F è minore del livello di significatività, il test rifiuta l'ipotesi nulla che le medie di tutti i gruppi sono uguali e conclude che almeno una media di un gruppo è diversa dalle altre. I livelli di significatività più comuni sono 0.05 e 0.01.

Affinché l'ANOVA possa essere eseguita è necessario che:

- Gli elementi che costituiscono i vari gruppi non siano oggetto di una particolare selezione, ma siano stati assegnati a caso (random).
- **I campioni devono essere tra loro indipendenti**, ovvero i dati osservati in un campione non devono essere influenzati da quelli osservati in un altro campione.

Quindi, prima di eseguire tale test, occorre **verificare che le varianze dei vari gruppi siano omogenee**.

3.2 Test di Bartlett

Prima di eseguire l'analisi della varianza, occorre verificare che le varianze dei vari gruppi siano omogenee (in statistica si parla di: *omoschedasticità* delle varianze).

Tra i vari test di omogeneità, è possibile utilizzare il cosiddetto *Test di Bartlett* [15], nel quale i dati presi in considerazioni sono **non bilanciati**, il che significa che i gruppi scelti hanno numerosità diverse.
Tale tipo di test, in particolare, viene utilizzato per verificare se più campioni di dati provengono da una distribuzione normale e possiedono la stessa varianza (ipotesi nulla), andando in contrasto con l'ipotesi alternativa che, almeno due dei campioni di dati, non hanno varianza uguale.

La statistica del test è esprimibile mediante la seguente formula:

$$T = \frac{(N-k)\ln s_p^2 - \sum_{i=1}^{k}(N_i - 1) * ln\, s_i^2}{1 + (\frac{1}{3(k+1)})\left[(\sum_{i=1}^{k}\frac{1}{(N_i-1)}) - \frac{1}{N-k}\right]}$$

dove:

- s_i^2 è la varianza dell' *i*-esimo gruppo
- N è il numero totale dei campioni
- N_i è la grandezza del campione dell'i-esimo gruppo
- k è il numero di gruppi
- s_p^2 è la varianza combinata

La varianza combinata viene definita come segue:

$$s_p^2 = \frac{\sum_{i=1}^{k}(N_i - 1){s_i}^2}{(N-k)}$$

La statistica segue la distribuzione di χ^2 con *p-1* gradi di libertà.
Il test di Bartlett è potente, ma seriamente sensibile alla non normalità della distribuzione e i risultati del test vanno considerati con una certa cautela.
D'altro canto non ci sono procedure particolarmente adatte nel caso di distribuzioni che si discostano sensibilmente dalla normalità.

3.3 Test di Kruskal - Wallis

In statistica, il test di Kruskal-Wallis[16] è un metodo non parametrico per verificare l'uguaglianza delle mediane di diversi gruppi; Prende il nome dai suoi autori William Kruskal e W. Allen Wallis.

Tale test richiede che le osservazioni siano trasformate in ranghi e può essere applicato nel caso di un esperimento completamente randomizzato.
Come per tutti i test non parametrici, l'ipotesi nulla non comprende relazioni riguardanti parametri delle popolazioni e non vengono utilizzate statistiche campionarie per la verifica delle ipotesi stesse.

Una volta riordinati i dati in ranghi, per ogni campione si calcola la somma dei ranghi ad esso relativi (R_i). Può essere calcolato il valore della statistica H :

$$H = \frac{12}{n_T(n_T+1)}\sum_i \frac{R_i^2}{n_i} - 3(n_T+1)$$

Ove n_T è il numero totale delle osservazioni ed n_i quelle appartenenti all' *i*-esimo gruppo.
La statistica H segue la distribuzione di χ^2 con *p-1* gradi di libertà, purché il numero di ripetizioni sia almeno 5. Se l'adattamento alla distribuzione del χ^2 non è valido, è possibile ricorrere ad apposite tavole di valori critici di H.

Nel caso in cui l'ipotesi nulla venga respinta, può essere interessante procedere a confronti multipli tra gruppi.

3.4 Confronti multipli (Multiple Comparisons)

Le tecniche di analisi di varianza testano se un set di gruppi ha media uguale o differente. Il rifiuto dell'ipotesi nulla, porta a concludere che non tutti i gruppi hanno la stessa media; tuttavia tale risultato, non fornisce alcuna informazione su quali di questi gruppi abbiano medie differenti e non è raccomandato eseguire una serie di *t*-test per determinare quali coppie di medie siano significativamente differenti.
Quando infatti si compiono dei *t*-test multipli, la probabilità che la differenza tra medie risulti significativa potrebbe essere dovuta al grande numero di test eseguiti.
Questi *t*-test utilizzano i dati provenienti dallo stesso campione, quindi non sono indipendenti, e tutto ciò rende più difficile quantificare il livello di significatività ottenuto per i test multipli.

Infatti, ogni volta che si rifiuta l'ipotesi nulla perché il *p-value* assume un valore minore del valore critico, è possibile che si stia sbagliando.
D'altronde, l'ipotesi nulla potrebbe tranquillamente essere vera, e quindi si potrebbe dire che i risultati significativi siano dovuti al caso.

Un *p-value* di 0.05 significa che c'è un 5% di possibilità di ottenere i risultati osservati, ***se*** l'ipotesi nulla è verificata. Questo però non significa che esiste un 5% di probabilità che l'ipotesi nulla sia vera.

Ad esempio, se si eseguono una serie di 100 test statistici, e per tutti questi l'ipotesi nulla risulta essere effettivamente vera, ci si aspetterebbe che circa 5 di questi test siano significativi con un livello di $P < 0.05$, solamente dovuto al caso. In questa situazione, si avrebbero all'incirca 5 risultati statisticamente significativi, dei quali tutti erano falsi positivi.
Il costo, in tempo, sforzi e forse soldi, potrebbe essere abbastanza alto, se si sono basate conclusioni importanti su questi falsi positivi.

Ecco che, quindi, che si sfrutta il metodo dei **confronti multipli** [17]**.** L'approccio classico per questo problema è quello di controllare il tasso di errore per il rifiuto dell'ipotesi nulla (errore di tipo 1). Invece di impostare un livello critico di P a 0.05 per la significatività del test, si può utilizzare un valore critico più basso. Se l'ipotesi nulla è vera per tutti i test, allora la probabilità di ottenere **un solo** risultato che sia significativo in questo nuovo valore critico, è 0.05. In altre parole, se tutte le ipotesi nulle sono vere, la probabilità che la famiglia di test comprenda uno o più falsi positivi dovuti al caso, è 0.05.

3.5 Box Plots

I Box Plots [18] sono una rappresentazione grafica utilizzata per descrivere la distribuzione di un campione tramite semplici indici di dispersione e di posizione. Tenendo sott'occhio l'immagine di figura 20, si possono elencare le seguenti caratteristiche:

- Il lato superiore e inferiore del rettangolo rappresentano sempre il primo e il terzo quartile del campione.
- La linea posizionata nel mezzo di ogni box non è altro che la **mediana**. Se la mediana non è centrata, allora viene evidenziata l'asimmetria del campione stesso.
- I segmenti che si estendono sia sopra che sotto il box, altrimenti chiamati *whiskers* ("baffi"), sono delimitati dal valore minimo e massimo.
- Le osservazioni che sono situate al di là del limite superiore e inferiore, rappresentano gli *outliers*. Di default, un outlier è un valore che è 1.5 volte più grande del range interquartile.
- Le tacche (***notches***) mostrano l'intervallo di confidenza della mediana tra i campioni. La larghezza di una tacca è calcolata in modo tale che i box plots di cui le tacche non si sovrappongono abbiamo una mediana differente al 5% di significatività. Tale livello di significatività si basa sull'assunzione di una distribuzione normale. In poche parole, paragonare le mediane dei box-plot, è come fare un test visuale delle ipotesi, analogo al *t*-test usato nelle medie.

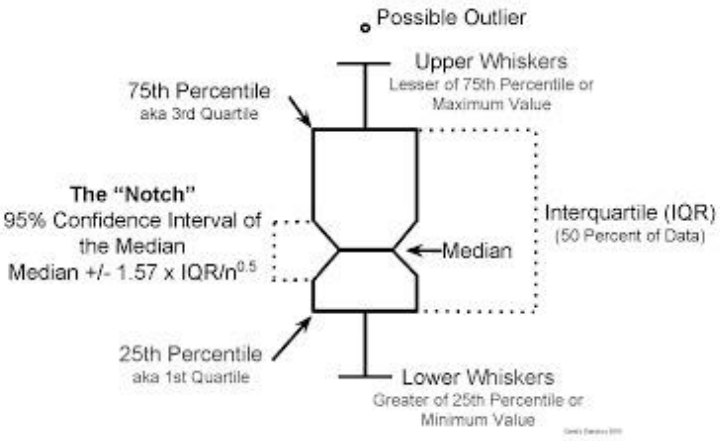


***Figura 20** : esempio di un box plot.*

L'intervallo dei *notches*, è basato sulla locazione dei quantili con l'estremo superiore e inferiore del notch definito da:

$$Notch = M \pm 1.57 \times \frac{IQR}{\sqrt{n}}$$

Dove:

- *M* è la mediana della distribuzione
- *IQR* (InterQuartile Range) è l'intervallo tra il primo e il terzo quartile
- *n* è il numero di osservazioni

Se gli intervalli di confidenza delle mediane di due box non si sovrappongono, allora c'è una forte evidenza (95% di confidenza), che le loro mediane differiscono.

Per concludere, si riporta l'immagine di figura 21 che mostra bene quale sia il legame che unisce un box plot con una *pdf* di una distribuzione normale.

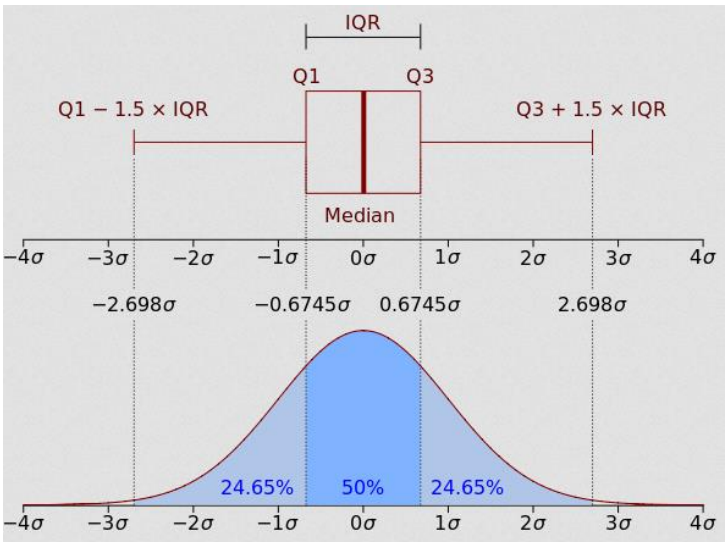


***Figura 21**: Relazione tra un box plot e la pdf di una distribuzione normale.*

Capitolo 4

Una volta portata a termine la pipeline del capitolo 2, si è ottenuto un file VCF per ogni batterio studiato (e la relativa copia). Si ricorda che tali files contengono all'interno tutte le varianti alleliche rilevate (SNPs).
I parametri del campo INFO, comuni a tutti gli SNPs e già descritti precedentemente, sono oggetto di studio di questo capitolo; chiaramente, non tutti sono stati analizzati, in quanto alcuni presentano dei valori costanti e poco significativi ai fini di un'analisi statistica.

Tutto ciò è visibile in Appendice, dove sono mostrati i box plots di quei parametri che si è scelto di non studiare. Dal momento che da tali box plots è possibile dedurre l'andamento delle distribuzioni, si può notare facilmente che molti di questi parametri assumono valori costanti (ad esempio AC=2, AN=2) oppure vicini allo zero, in quanto non comuni a tutti gli SNPs (BaseQRankSum, ClippingRankSum, ReadPosRankSum, MQRankSum); dunque, uno studio su queste variabili sarebbe parecchio marginale.

Nel presente capitolo si è scelto di approfondire lo studio dei parametri **DP**, **QD** e **SOR**, per ognuno dei quali si andrà a fornire una breve descrizione teorica e, successivamente, si mostreranno i risultati ottenuti dall'applicazione dei test statistici ai parametri stessi.

Prima di procedere con l'analisi, è stata condotta una breve indagine statistica per determinare non solo quanti SNPs hanno in comune un isolato batterico e la sua copia, ma anche per quanti SNPs differiscano l'uno dall'altro (vedi tabella 3.1)
La tabella 3.2, invece, è costruita sullo stampo di quella precedente, mettendo a confronto un isolato e l'altro.

Batterio A	Batterio B	# totale di SNP di A	# totale di SNP di B	# di SNP in comune	# di SNP di A non presenti in B	# di SNP di B non presenti in A
LIST1	**LIST63**	117574	116239	**113519**	4054	2719
LIST2	**LIST64**	118918	114605	**112937**	5944	1631
LIST3	**LIST65**	115063	116391	**112479**	2583	3911

***Tabella 3.1** : Calcolo del numero di SNP in cumune per ogni isolato e la sua copia.*

Batterio A	Batterio B	# totale di SNP di A	# totale di SNP di B	# di SNP in comune	# di SNP di A non presenti in B	# di SNP di B non presenti in A
LIST1	**LIST2**	117574	118918	**115161**	2413	3757
LIST2	**LIST3**	118918	115063	**113419**	5499	1644
LIST3	**LIST1**	115063	117574	**113182**	1881	4392

***Tabella 4.2** : Calcolo del numero di SNPs in comune tra un isolato e l'altro.*

4.1 Depth of Coverage (DP)

Come già descritto precedentemente, con il termine ***coverage*** [19] (copertura) vengono indicate quante basi in più sono state sequenziate, in relazione alla lunghezza totale di un genoma. Ad esempio, una copertura di 1X sta ad indicare che il numero di basi sequenziate è uguale alla lunghezza totale del genoma incognito, ma questo non vuol dire che tutto il genoma viene sequenziato.

Si supponga, ad esempio, di avere una sequenza target con le relative reads (ottenute dalla frammentazione del DNA di una copia).

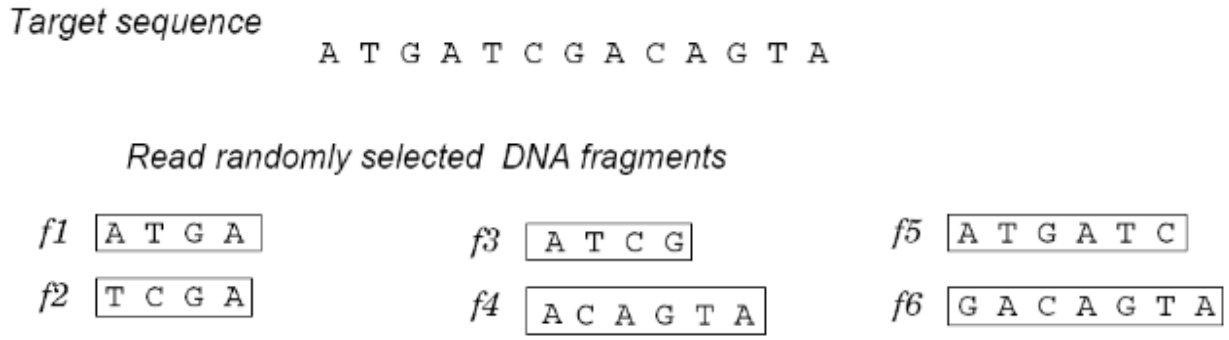


Figura 22: *esempi di reads ricavate dalla sequenza target.*

La sequenza target viene ricostruita andando a identificare le sovrapposizioni fra i frammenti, ottenendo quindi il risultato mostrato in figura 23.

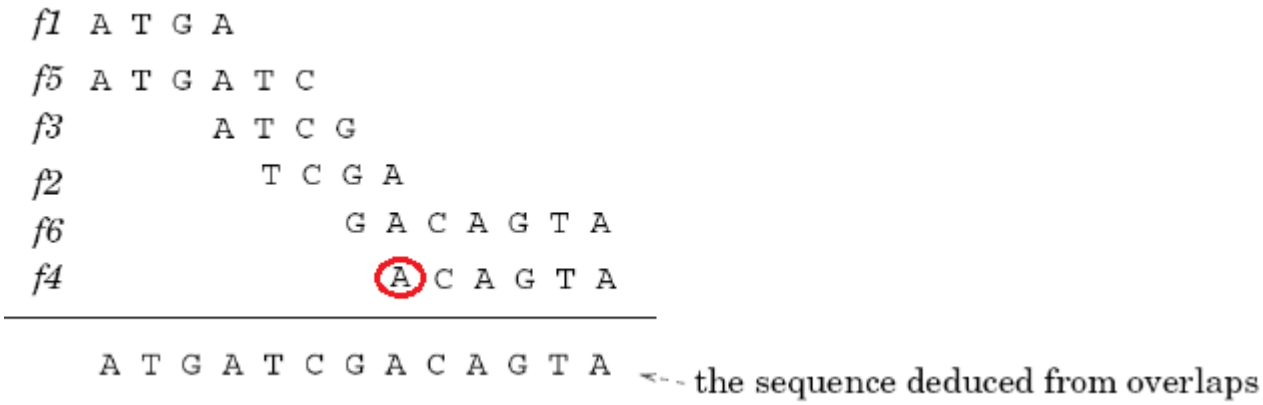


Figura 23: *sequenza ricavata dalla sovrapposizione delle reads; il nucleotide A ha un valore di coverage pari a 3.*

Si può ora determinare il coverage di un determinato nucleotide, andando semplicemente a contare il numero di volte che quello stesso nucleotide compare all'interno di una read.
In questo caso, il coverage del nucleotide cerchiato in rosso, avrà un valore pari a 3.

È possibile, inoltre, definire quello che è il ***coverage medio*** per l'intero genoma, il quale può essere calcolato prendendo in considerazione:

- Lunghezza del genoma originale (***G***),
- Numero di reads (***N***),

- Lunghezza media di una read (***L).***

Il coverage medio sarà dunque dato da:

$$\boldsymbol{Coverage\ medio} = \frac{N * L}{G}$$

Per esempio, un ipotetico genoma composto da 2000 bp, ricostruito da 8 reads con una lunghezza media di 500 nucleotidi, avrà una ridondanza 2×; il che significa che un preciso nucleotide compare all'incirca due volte, nel caso siano uniformemente distribuiti.

Nella tabella sottostante vengono invece riportati i risultati rispettivamente di media, varianza e deviazione standard del coverage per ciascun batterio preso in considerazione.

Depth of Coverage (DP)			
Chrom	**Media**	**Varianza**	**Std Dev**
LIST1	87,84 reads	380,05 reads	19,494 reads
LIST63	135,29 reads	835,32 reads	28,90 reads
LIST2	43,36 reads	120,15 reads	10,96 reads
LIST64	166,98 reads	1314,58 reads	36,26 reads
LIST3	120,10 reads	725,88 reads	26,94 reads
LIST65	127,20 reads	761,24 reads	27,59 reads

Tabella 5 : *Principali parametri statistici del parametro DP per batteri studiati.*

I grafici rappresentano, invece, gli scatter-plot relativi alle distribuzioni una contro l'altra del replicato (ad es. LIST1) e la sua copia (ad es. LIST63). Tali distribuzioni mostrano la "*depth of coverage*" associata a ognuna delle varianti alleliche presenti nelle sequenze genetiche dei batteri.

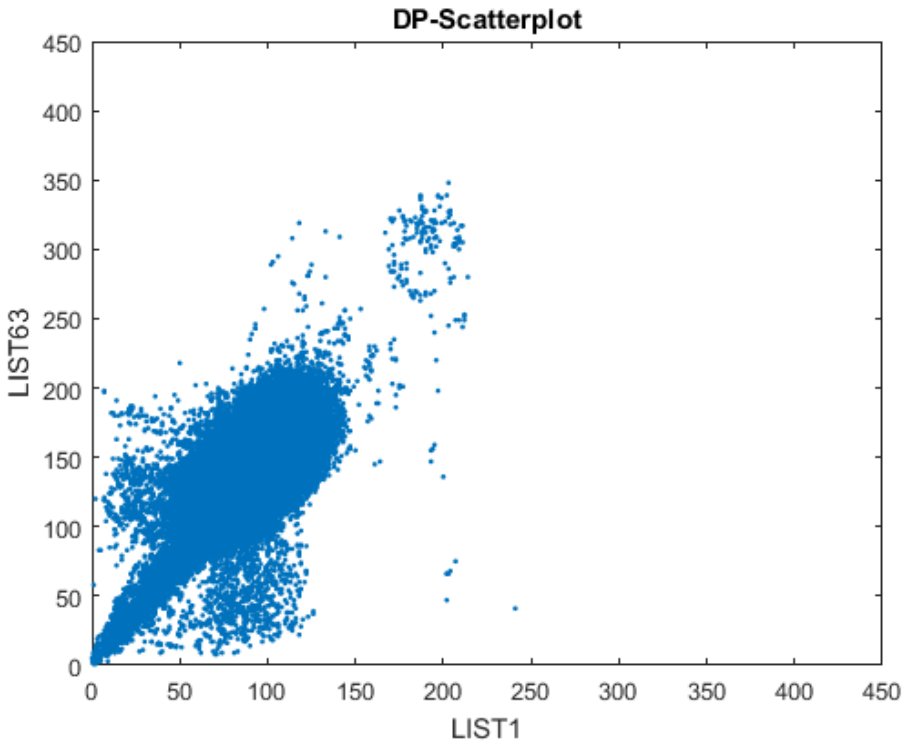


***Figura 24**: Scatter plot delle distribuzioni del coverage una contro l'altra dell'isolato (LIST1) e la sua copia (LIST63).*

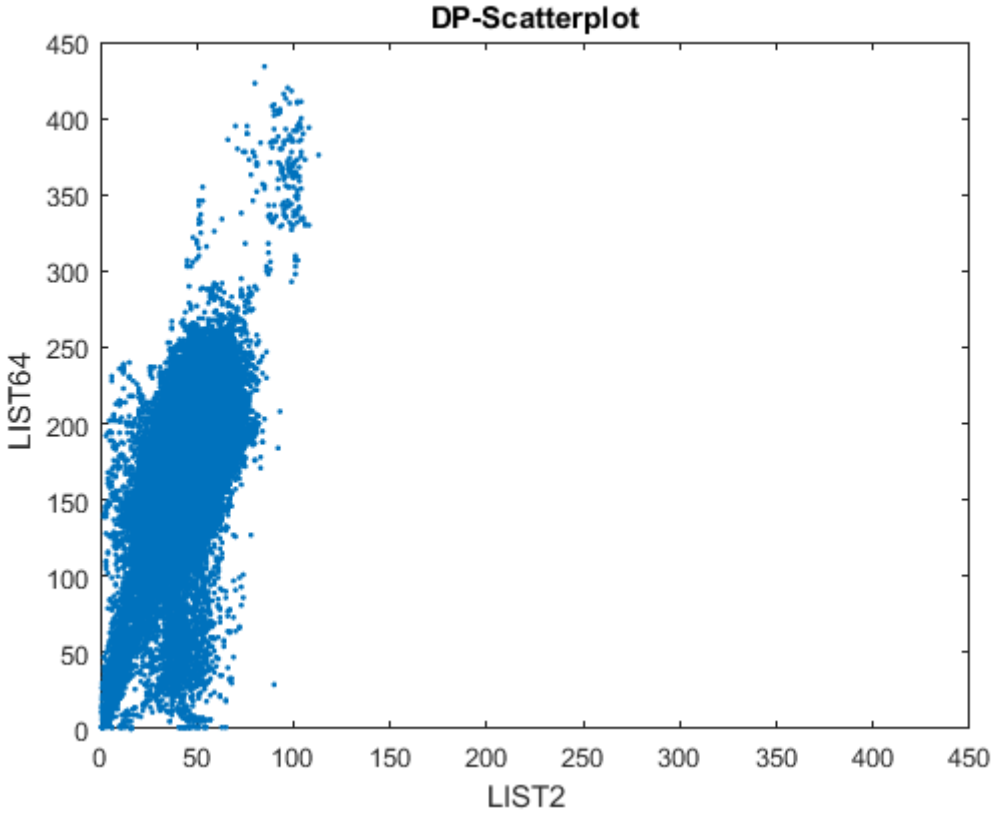


***Figura 25**: Scatter plot delle distribuzioni del coverage una contro l'altra dell'isolato (LIST2) e la sua copia (LIST64).*

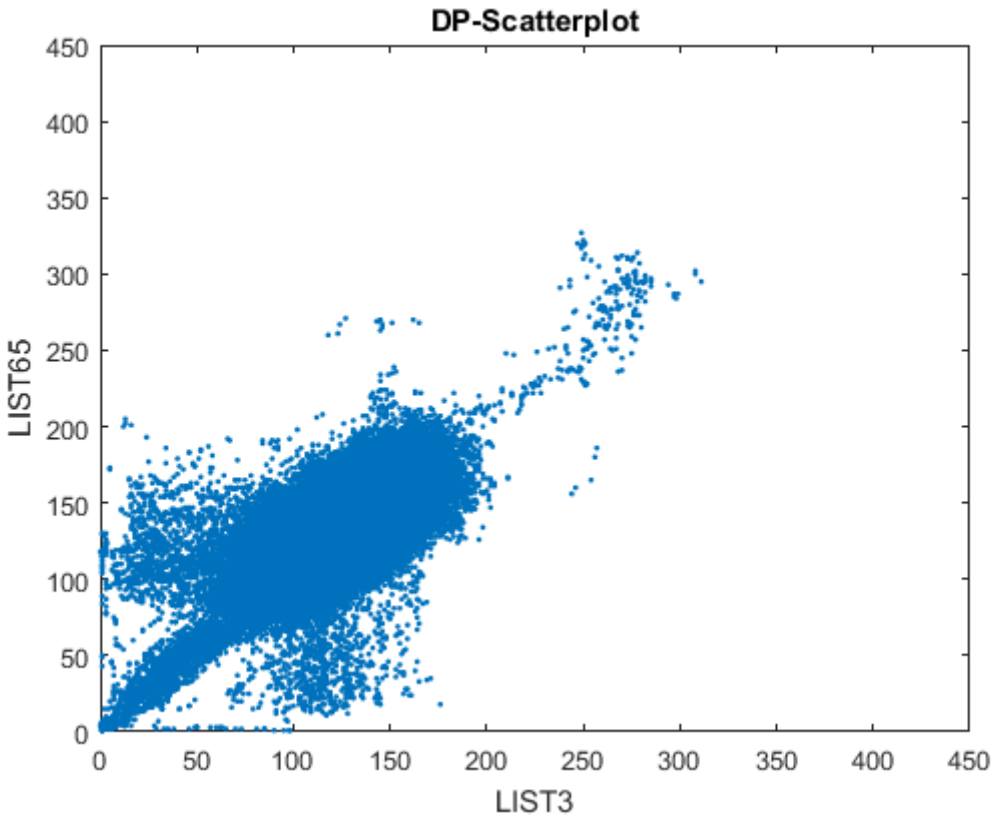


Figura 26 *: Scatter plot delle distribuzioni del coverage una contro l'altra dell'isolato (LIST3) e la sua copia (LIST65).*

Il confronto tra i vari scatter plots, avviene attraverso la valutazione dei coefficienti di correlazione associati a ogni coppia di variabili. Tali coefficienti vengono calcolati mediante la funzione di MatLab "`corrcoef`", la quale utilizza la seguente formula:

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2 \sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

in cui $\bar{x}$ è la media della prima serie di misure, ed $\bar{y}$ è la media della seconda serie di misure.
Il coefficiente di correlazione, dunque, assume un valore compreso tra -1 e 1, dove:

- Valori vicini a 1 indicano che esiste una relazione lineare positiva tra le colonne di dati.
- Valori vicini a -1 indicano che una colonna di dati ha una relazione lineare negativa con un'altra colonna di dati (in questo caso di parla di *anti-correlazione*).
- Valori vicini o uguali a 0 suggeriscono che non esiste una relazione lineare tra le colonne di dati.

Nella tabella sottostante si riportano i valori di tali coefficienti per ogni coppia considerata:

Campioni	Coefficiente di correlazione (DP)
LIST1 – LIST63	0.6944
LIST2 – LIST64	0.6058
LIST3 – LIST65	0.6961

Tabella 6: *coefficienti di correlazione per ogni coppia di batteri relativi al parametro DP.*

Facendo riferimento, quindi, alla tabella 5 e, sapendo che gli scatter plots mostrano la relazione che esiste tra le distribuzioni del coverage degli SNPs, si può dedurre che entrambi i tre grafici di figura 24, 25 e 26 hanno un buon grado di correlazione (si parla di ***correlazione positiva***), con valori compresi tra 0.6 e 0.7.

4.1.1 DP: BedTools Coverage

Una volta effettuate le misure del coverage relativo ai singoli SNPs per ognuno dei batteri elencati sopra, si è poi proceduto a misurare comunque la copertura, questa volta su tutto il genoma batterico.

Per fare ciò è stato preso in considerazione il programma **BedTools** [20]. Tale programma comprende una serie di utilities specifiche adatte per analizzare un'ampia gamma di dati genomici. Ad esempio, permette di intersecare, unire, contare, e smistare intervalli genomici da più file in formati ampiamente utilizzati, come BAM, BED, GFF/GTF e VCF.
Il tool preso in considerazione è appunto **BedTools – Coverage** che permette di calcolare sia la profondità (*depth*) che l'ampiezza (*breadth*) del coverage per le reads presenti in un file A, sulle caratteristiche di un file B.
Se ad esempio il file A rappresenta il genoma, mentre il file B rappresenta un determinato numero di intervalli, BedTools-Coverage suddivide il file A in un numero di intervalli determinati dal file B.
Poiché nel caso in questione si dispone di un file *.bam* lungo 2.912.690 basi (che rappresenta il file A) si è scelto di creare 2 file con estensione *.bed*, che rappresentano la suddivisione di quelle basi in intervalli di 55 bp e 110 bp.

Sono quindi stati prelevati i file *.bam* prodotti da PICARD per ognuno dei batteri citati, e proprio di questi è stato calcolato il coverage per ogni determinato intervallo.
Andando a dividere tutte le 2.912.690 basi con lunghezze da 55 bp e 110 bp, si ottiene un totale di 52.958 e 26.479 intervalli, in ognuno dei quali è stata calcolata la copertura.

I grafici risultanti sono mostrati nelle figure 27 e 28, dove non si notano particolari differenze nelle distribuzioni del coverage, nonostante nella prima immagine ci sia un numero doppio di intervalli rispetto alla seconda. Inoltre, la tabella 6 riporta i valori dei coefficienti di correlazione per ogni coppia di batteri esaminata.

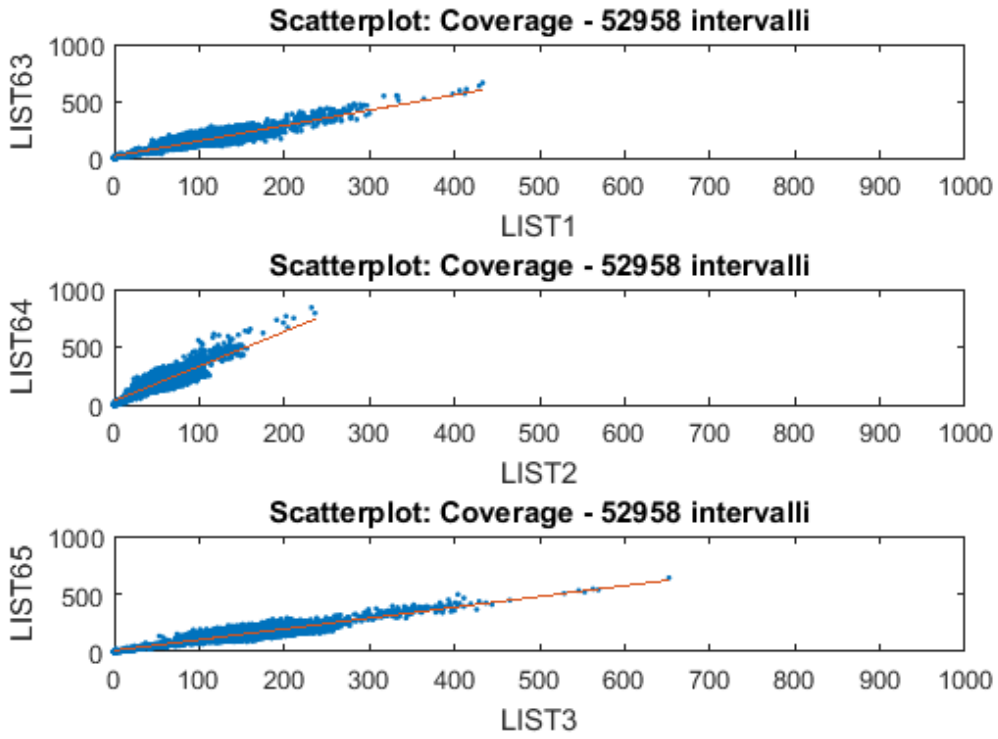


***Figura 27**: Scatter plot delle distribuzioni del coverage su tutto il genoma batterico ottenute andando a selezionare intervalli lunghi 55 bp.*

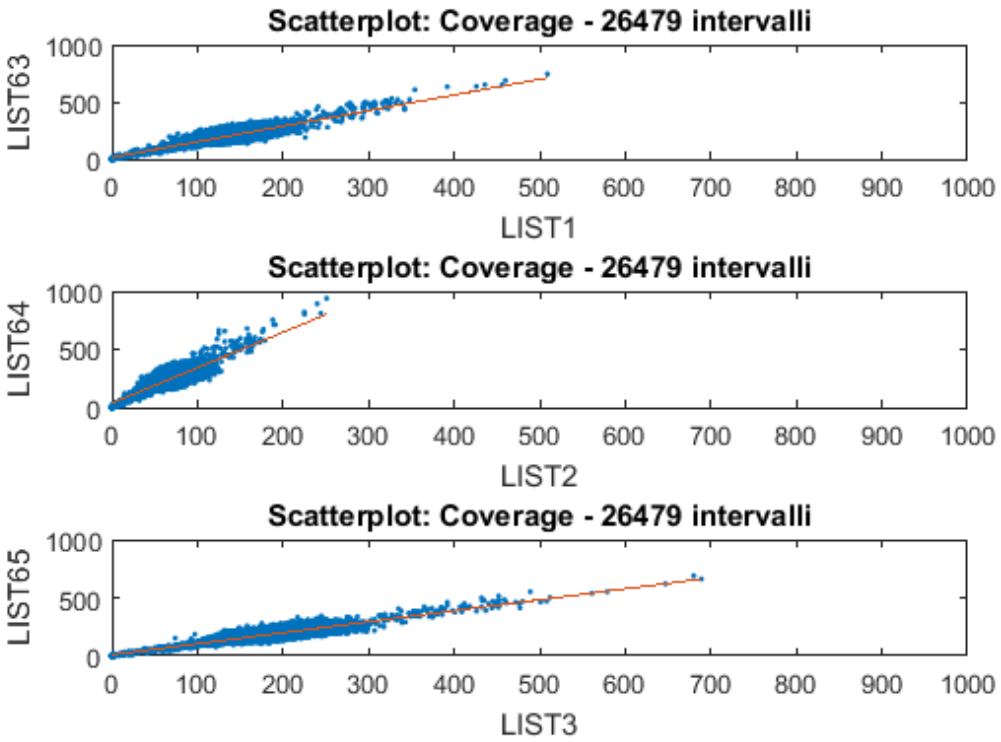


***Figura 28**: Scatter plot delle distribuzioni del coverage su tutto il genoma batterico ottenute andando a selezionare intervalli lunghi 110 bp.*

Campioni	Coeff. Corr. DP - 55 intervalli	Coeff. Corr. DP - 110 intervalli
LIST1 – LIST63	0.9148	0.9224
LIST2 – LIST64	0.8818	0.8925
LIST3 – LIST65	0.9249	0.9325

***Tabella 7**: coefficienti di correlazione per le coppie di batteri nel caso di una suddivisione in intervalli lunghi 55 e 110 bp.*

Come si può notare, i risultati ottenuti indicano un forte grado di correlazione tra ogni coppia di batteri selezionata. La tabella 6 evidenzia quanto detto, poiché i valori dei coefficienti si aggirano attorno a 0.9, indicando, quindi, che le variabili sono *correlate positivamente*.
Infine, a conferma di tutto, la linea di *best fit* presente negli scatter plots (in rosso) mostra come queste distribuzioni seguano un andamento lineare.

4.1.2 DP: Test di Bartlett

L'indagine statistica effettuata sul parametro DP, comincia con la costruzione di una matrice *M*, composta da 6 colonne, ognuna delle quali contiene tutti i valori del *coverage (DP)* dei batteri studiati.
A questo punto, si va ad eseguire il Test di Bartlett per verificare che i dati acquisiti provengano da una distribuzione normale e possiedano la stessa varianza (ipotesi nulla).
Per fare ciò, si utilizza la funzione di MatLab `vartestn(M)`. Tale funzione prende in ingresso la matrice *M*, e restituisce una tabella riassuntiva delle statistiche relative alle colonne della matrice *M*.
I risultati ottenuti sono mostrati in figura 29.

Group Summary Table

Group	Count	Mean	Std Dev
1	113519	87.841	19.4948
2	113519	135.287	28.9019
3	113519	43.148	11.339
4	113519	166.179	37.97
5	113519	119.004	29.1578
6	113519	126.037	30.0191
Pooled	681114	112.916	27.4984
Bartlett's statistic	161448.7		
Degrees of freedom	5		
p-value	0		

***Figura 29:** risultati del test di Bartlett per il parametro DP.*

Bisogna determinare ora se la statistica del test è significativa; per fare ciò si può procedere in due maniere. Un primo metodo è quello di andare a guardare il *p-value*.
Il *p-value* ottenuto è $p = 0$, ciò indica che `vartestn` rifiuta l'ipotesi nulla che le varianze sono

uguali su tutte le 6 colonne, a favore dell'ipotesi alternativa che, almeno una colonna, possiede una varianza diversa.

Il secondo metodo, invece, consiste nel confrontare il valore ottenuto con la statistica di Bartlett (visibile in figura 29) con il corrispettivo valore critico di χ^2, per 6-1 gradi di libertà che in questo caso risulta essere 11.07. La figura sottostante, infatti, riporta la tavola di distribuzione del χ^2, dove è stato cerchiato in rosso il valore corrispondente al caso in questione.

Tavola distribuzione CHI-QUADRATO

Gradi di libertà	***Livello di Probabilita' a***									
	1.00	***0.99***	***0.95***	***0.90***	***0.25***	***0.10***	***0.05***	***0.025***	***0.01***	***0.005***
1				0.02	1.32	2.71	3.84	5.02	6.64	7.88
2	0.01	0.02	0.10	0.21	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.12	0.35	0.58	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.71	1.06	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	1.15	1.61	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.64	2.20	7.84	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	2.17	2.83	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.73	3.49	10.22	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	3.33	4.17	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.94	4.87	12.55	15.99	18.31	20.48	23.21	25.19

***Figura 30**: Tavola di distribuzione χ^2*

Dato che il valore ottenuto con il test di Bartlett è molto superiore a quello critico, possiamo scartare l'ipotesi nulla e dire che le varianze non sono omogenee tra loro, per cui non è soddisfatta l'assunzione dell'omogeneità delle varianze necessaria per eseguire l'ANOVA. Tutto ciò, deriva dal fatto che il test di Bartlett è sensibile agli scostamenti dalla normalità, cioè ne risente se i campioni provengono da distribuzioni non normali.
E in effetti, andando a confrontare le diverse distribuzioni del parametro DP (coverage) riportate in figura 31, si può notare immediatamente che gli istogrammi evidenziano un aspetto asimmetrico, il che suggerisce una distribuzione non gaussiana dei dati.

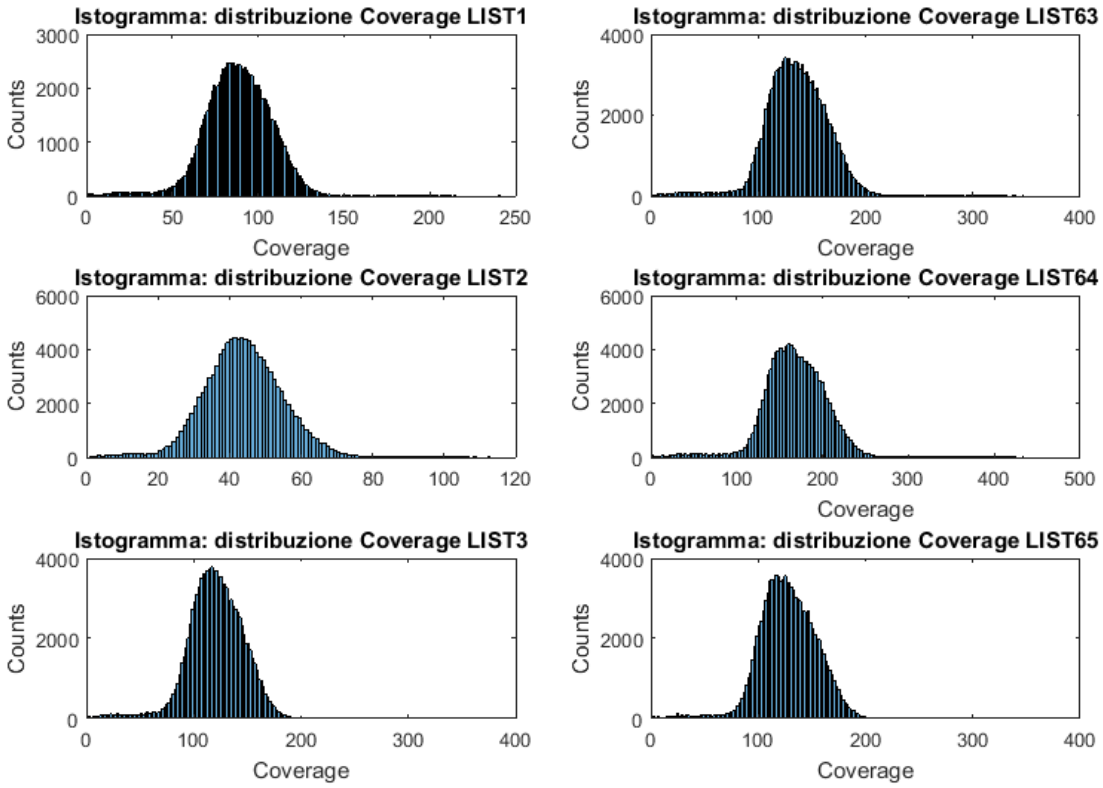


***Figura 31**: Istogrammi delle misure del coverage per gli SNPs dei batteri in esame.*

4.1.3 DP: ANOVA

Si prosegue quindi lanciando il comando `anova1(M)`, dove *M* è sempre la matrice costruita inizialmente. I risultati ottenuti sono mostrati nella figura 32.

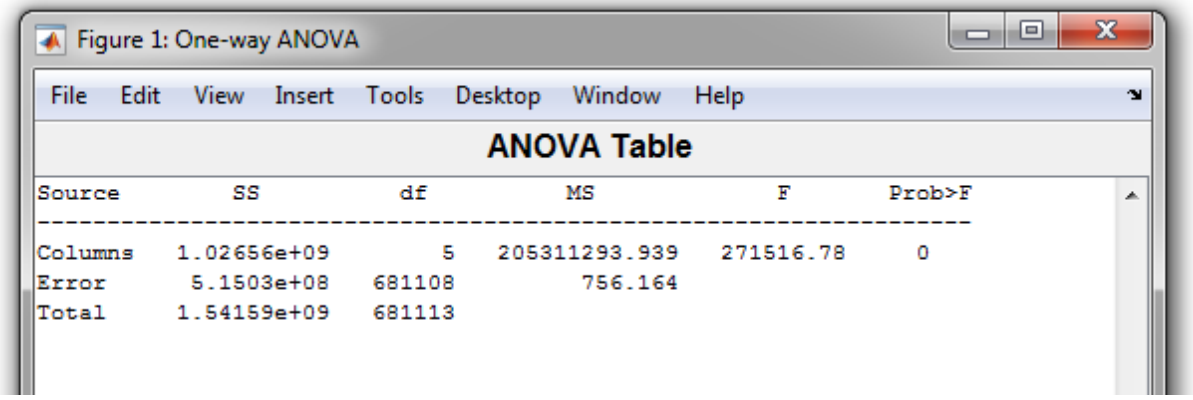


***Figura 32**: ANOVA: tabella riassuntiva.*

La tabella relativa all'ANOVA mostra la varianza tra i gruppi (`Columns`) e la varianza interna ai gruppi (`Error`). `SS` rappresenta la somma dei quadrati, mentre `df` è il grado di libertà.
Il grado di libertà totale è dato dal numero complessivo di osservazioni meno 1, nel caso in questione sarebbe 681113. Il grado di libertà tra i gruppi è, invece, dato dal numero di gruppi meno uno, quindi 6-1= 5. Di conseguenza, il grado di libertà interno ai gruppi è rappresentato dal numero totale dei gradi di libertà meno i gradi di libertà interni ai gruppi, ovvero 681108.
`MS` rappresenta l'errore quadratico medio dato da `SS/df` per ogni fonte di variazione.

La statistica del test, indicata con *F*, non è altro che il rapporto degli errori quadratici medi, quindi:

$$F = \frac{205311293.939}{756.164}$$

Il *p*-value, infine, rappresenta la probabilità che *F* assuma un valore maggiore o uguale del test statistico. Quindi che, P (F > 271516.78).

Seguendo questa definizione, si può concludere che il valore del *p-value* ottenuto, ovvero 0, suggerisce che la differenza tra le medie delle colonne della matrice *M* è significativa, mettendo in evidenza, di conseguenza, il fatto che i coverage messi a confronto non hanno la stessa media.

4.1.4 DP: Kruskal - Wallis

Dal test precedentemente condotto, si era concluso che tutte le colonne della matrice *M* non avevano una media uguale.

Il test che ora si prende in considerazione, ovvero il test di *Kruskal – Wallis*, ripete la stessa analisi applicando una procedura **non** parametrica, che è una versione non parametrica dell'ANOVA di tipo 1. L'assunzione che c'è dietro a questo tipo di test è che le misure provengano da una distribuzione continua, ma non necessariamente normale. Il test è basato sull'analisi di varianza utilizzando non il valore dei dati, ma il rango assegnato a ciascun dato.

Eseguendo, dunque, il comando di MatLab `kruskalwallis(M)`, viene restituita in output una tabella come quella di figura 33 in cui il basso valore del *p-value* conferma che il test di Kruskal-Wallis è in accordo con il risultato ottenuto dall'analisi della varianza.

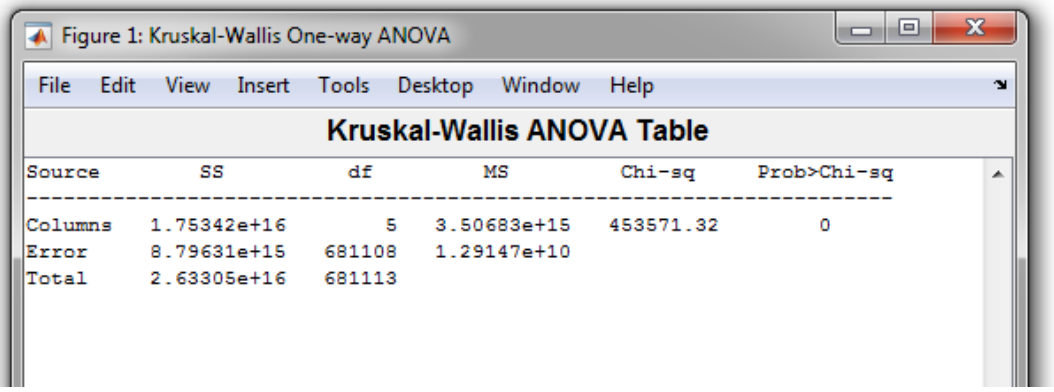


***Figura 33**: risultati del test di Kruskal-Wallis.*

Viene dunque rifiutata l'ipotesi nulla che tutti i 6 campioni dei dati provengano da una distribuzione continua con un livello di significatività dell'1%.

4.1.5 DP: Multiple Comparisons

Proprio come descritto nel capitolo 3, nel caso in cui l'ipotesi nulla del test di *K-W* venga respinta, può essere interessante procedere ai confronti multipli tra gruppi.

A tal proposito, si esegue la funzione di MatLab `multcompare(stats)`, per determinare quelle coppie dei gruppi che hanno una media significativamente differente. Tale funzione restituisce in output una matrice con i risultati del confronto a coppie, utilizzando le informazioni contenute in `stats` (prodotte da `anova1`). Per informazioni dettagliate sul meccanismo di tale funzione, si può consultare la documentazione online riportata in [21].

I risultati ottenuti sono mostrati in figura 34.

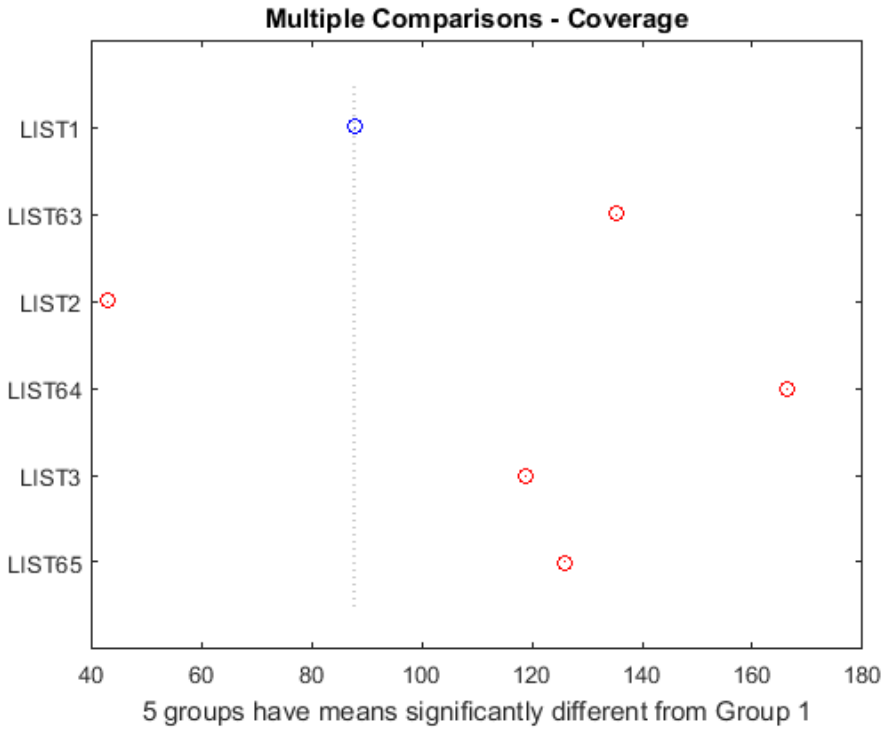


***Figura 34**: diagramma dei confronti multipli tra i campioni.*

Nel grafico soprastante, la media di ogni gruppo è rappresentata da un cerchio, mentre l'intervallo è descritto da una linea che si estende al di fuori del simbolo.
In questo grafico, due gruppi hanno medie significativamente differenti se, e solo se, i loro intervalli sono disgiunti (proprio come accade nel caso in questione); al contrario, se gli intervalli risultano sovrapposti, non sussistono particolari differenze
Poiché il grafico ha una scala abbastanza ampia, è stata riportata un'immagine ingrandita del cerchio blu (figura 35), che rappresenta il gruppo LIST1, al fine di fornire una migliore comprensione.

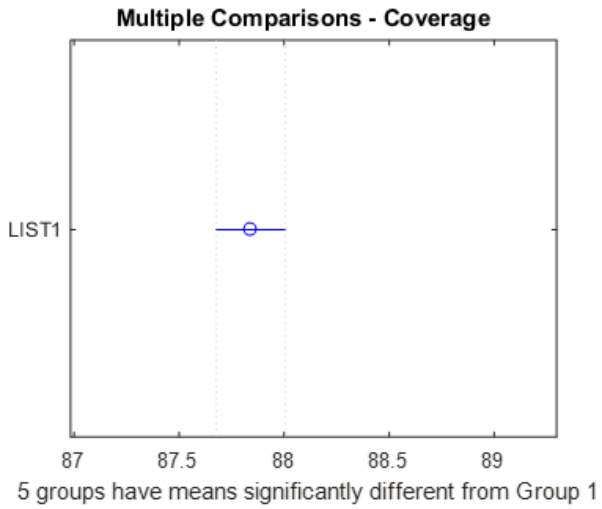


***Figura 35**: ingrandimento dell'immagine di figura 34 dove è visibile l'intervallo per il campione LIST1*

Dall'analisi condotta si può notare che, tra i 6 gruppi analizzati, 5 di questi hanno una media significativamente differente dal gruppo 1; stessa cosa vale per il gruppo 2 e così via per gli altri, in quanto **non** è presente una sovrapposizione di intervalli.
Inoltre, la tabella sottostante che mostra i risultati del confronto a coppie, conferma quanto detto. Infatti, è evidente che tutti i *p-value* assumono il valore zero; ciò indica che la media del coverage per gli SNP di ogni gruppo è diversa per tutti i sei campioni presi in considerazione.

Campione A	Campione B	Lim. Inferiore	Differenza medie	Lim. Superiore	*p-value*
LIST1	LIST63	-47,78	-47,45	-47,12	0,00
LIST1	LIST2	44,36	44,69	45,02	0,00
LIST1	LIST64	-78,67	-78,34	-78,01	0,00
LIST1	LIST3	-31,49	-31,16	-30,83	0,00
LIST1	LIST65	-38,52	-38,20	-37,87	0,00
LIST63	LIST2	91,81	92,14	92,47	0,00
LIST63	LIST64	-31,22	-30,89	-30,56	0,00
LIST63	LIST3	15,95	16,28	16,61	0,00
LIST63	LIST65	8,92	9,25	9,58	0,00
LIST2	LIST64	-123,36	-123,03	-122,70	0,00
LIST2	LIST3	-76,19	-75,86	-75,53	0,00
LIST2	LIST65	-83,22	-82,89	-82,56	0,00
LIST64	LIST3	46,85	47,17	47,50	0,00
LIST64	LIST65	39,81	40,14	40,47	0,00
LIST3	LIST65	-7,36	-7,03	-6,70	0,00

Tabella 8: *tabella riassuntiva contenente i risultati del test dei confronti multipli.*

In tabella:

- Le colonne 1 e 2 mostrano i gruppi confrontati.
- La colonna 4 rappresenta la differenza tra le medie di quei gruppi.
- Le colonne 3 e 5 determinano il limite superiore e inferiore di un 95% di intervallo di confidenza per la media effettiva.
- La colonna 6 contiene il *p-value* per un test d'ipotesi in cui la corrispondente differenza di medie è uguale a zero.

4.1.6 DP: Box Plots

L'immagine di figura 36 mostra, invece, i box plots del coverage per gli SNP dei batteri studiati. Osservandola, si può affermare che le mediane dei primi quattro gruppi sono significativamente differenti con un intervallo di confidenza del 95%, in quanto le tacche (*notches*) dei boxplot non si sovrappongono.

Guardando, invece, la coppia LIST3-LIST65, si potrebbe constatare che le mediane di entrambi i box si aggirano attorno allo stesso valore. Nella tabella 8 è stata appositamente cerchiata in rosso la similarità tra le mediane di questi ultimi due gruppi; tuttavia, tale somiglianza apparente, entra in contrasto con il *p-value* trovato (0), visibile in tabella 7 alla riga 15.
Le croci rosse, che si estendono sia sopra che sotto ogni box plots, rappresentano tutti i possibili *outliers*, che, come si può notare sono molto numerosi. Infine, si può dire che linea rossa interna appare abbastanza centrata per ogni gruppo, il che indica che le distribuzioni dei campioni sono quanto meno simmetriche.

Mediane dei campioni					
LIST1	LIST63	LIST2	LIST64	LIST3	LIST65
88 reads	135 reads	43 reads	166 reads	**120 reads**	**126 reads**

***Tabella 9**: valori delle mediane dei campioni, in rosso sono stati cerchiati i valori simili per la coppia LIST3-LIST65*

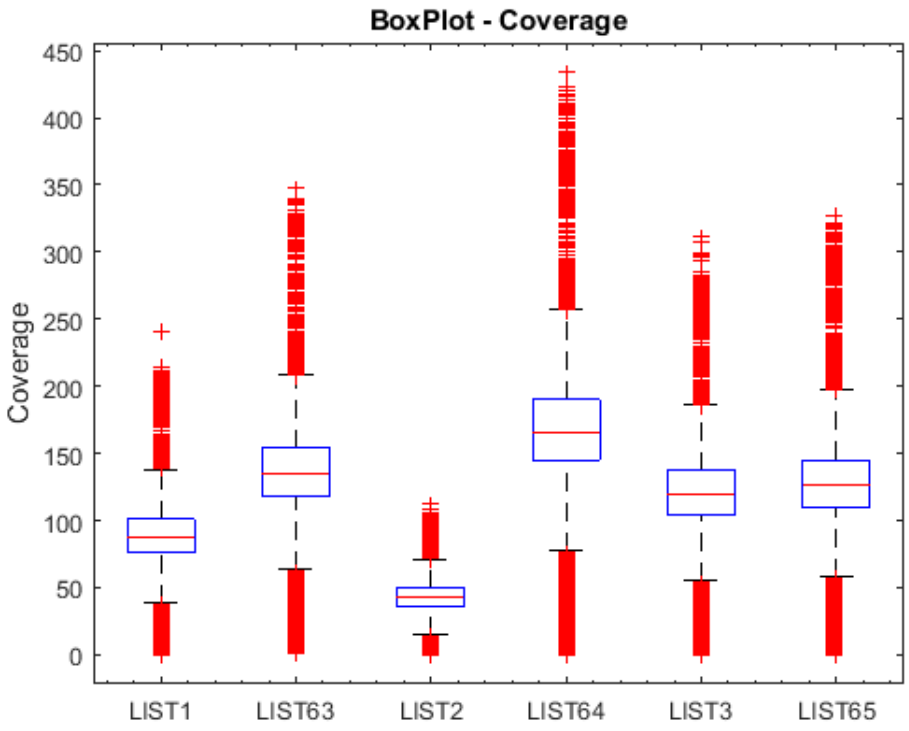


***Figura 36**: Box plots del coverage per gli SNP dei batteri.*

Ora, pertanto, si analizzerà l'intervallo di confidenza delle mediane, i cosiddetti **notches**, dei box plots relativi a LIST3 e LIST65, per verificare se effettivamente tale somiglianza sia significativa.

Si sono quindi prese le distribuzioni LIST3 e LIST65 e ne è stato fatto nuovamente un test parametrico (*Kruskal - Wallis*) al fine di verificare l'uguaglianza tra le mediane. I risultati ottenuti e i relativi box plots sono illustrati in figura 37 e 38:

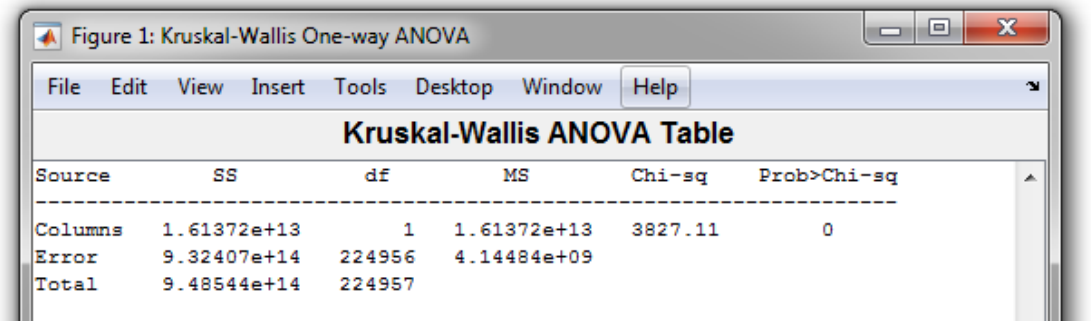


***Figura 37**: risultati del test di K-W eseguito per i campioni LIST3 e LIST65.*

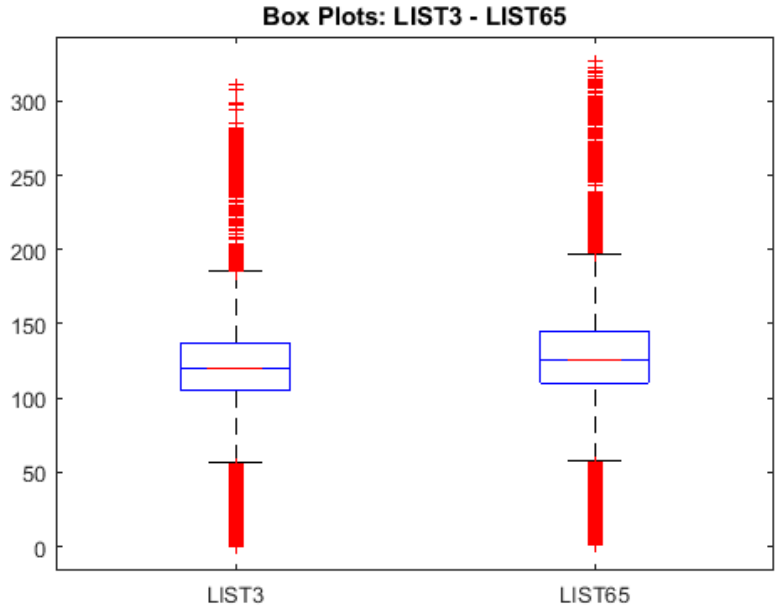


***Figura 38**: Box plots per LIST3 e LIST65.*

Come si può notare, gli intervalli di confidenza delle mediane sono talmente piccoli che non è possibile distinguerli. Questo fatto è dovuto all'elevato numero *n* di osservazioni, ricordando che le tacche (notches) si calcolano mediante la seguente formula:

$$Notch = M \pm 1.57 \times \frac{IQR}{\sqrt{n}}$$

È evidente come il test scarti l'ipotesi di uguaglianza tra le mediane, in quanto il *p-value* ottenuto assume valore zero, smentendo quanto evidenziato prima sull'apparente similarità dei risultati di tabella 8.

Dall'analisi complessiva, si può concludere che i 6 gruppi rappresentanti il coverage per i batteri studiati, non presentano particolari somiglianze, poiché ogni gruppo mostra dei risultati a sé stanti, sia per quanto riguarda la continuità delle distribuzioni, sia per quanto riguarda la similarità tra le mediane.

4.2 Quality by Depth (QD)

Il punteggio di qualità di una base (Q), detto anche **phred quality score** [22], è un valore intero che rappresenta la stima di probabilità di commettere un errore, ad esempio che la base chiamata non sia corretta. Detta *P* la probabilità di commettere un errore, si avrà che:

$$P = 10^{-Q/10}$$

e

$$Q = -10\, log_{10}(P)$$

È da notare che un quality score pari a 3, implica P=0.5, ovvero esiste una probabilità del 50% di commettere un errore su quella chiamata; e valori più bassi di Q rappresentano un alto valore di P. La figura 39 è una rappresentazione grafica del Phred Quality Score che mostra la relazione esistente tra il grado di accuratezza e la probabilità di errore.

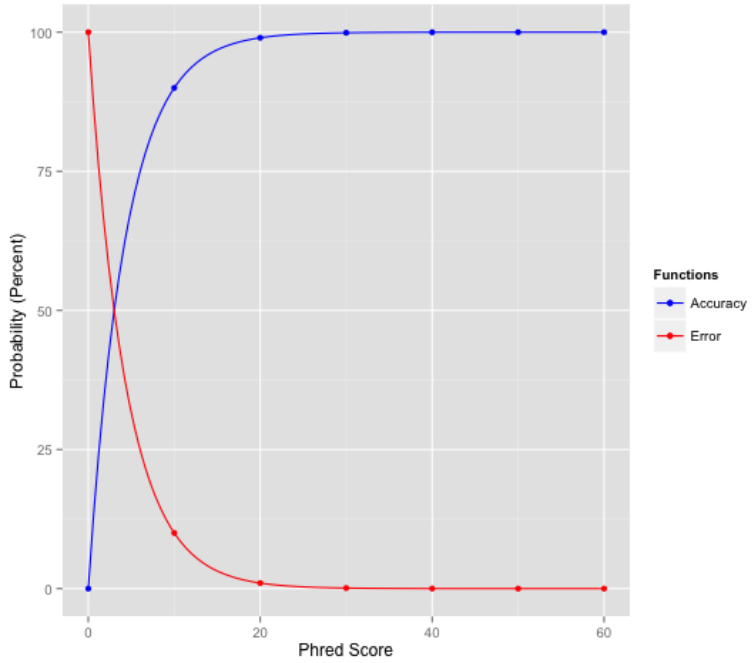


Figura 39 [23]*: rappresentazione grafica del phred quality score.*

La linea rossa mostra l'errore, mentre la linea blu mostra l'accuratezza. Si può notare come al decrescere dell'errore, aumenta simmetricamente il grado di accuratezza.

Il parametro QD studiato nel seguente lavoro di tesi, è definito dalla relazione:

$$QD = \frac{QUAL}{Allele\ Depth}$$

e non è altro che il punteggio di qualità, normalizzato per la quantità di coverage disponibile Tale normalizzazione viene eseguita per il seguente motivo:

Si supponga, ad esempio, di avere due varianti (A e T), dove a entrambe è stato assegnato un punteggio di qualità pari a 10. Tuttavia, la variante A è supportata da un coverage di 50 reads, mentre la variante T da 100 reads. Poiché entrambe possiedono lo stesso punteggio Q, verrebbe naturale pensare che l'errore associato a tali chiamate sia pari a P= 0.1;

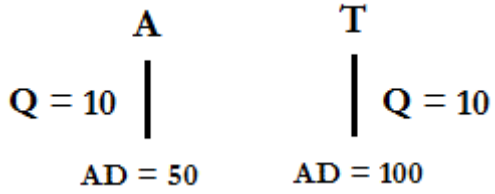


***Figura 40**: esempio di due varianti con stesso quality score ma con un valore di coverage differente.*

Secondo tale valore, quindi, esiste una probabilità del 10% di commettere un errore su entrambe le chiamate.

Tuttavia una chiamata è supportata da 50 reads, mentre l'altra da 100 reads, per cui, applicando la formula scritta sopra per il calcolo del QD, si avrà che:

$$QD_A = \frac{10}{50} = 0.2$$

$$QD_T = \frac{10}{100} = 0.1$$

Si può concludere, pertanto, che in generale, è meglio avere un valore alto di QD: questo perché la variante supportata da più reads dovrebbe avere, in proporzione, un punteggio di qualità nettamente più alto.

Nella tabella sottostante vengono riportati i risultati rispettivamente di media, varianza e deviazione standard del parametro in considerazione.
Invece, i grafici collocati subito dopo la tabella rappresentano gli scatter-plot relativi alle distribuzioni una contro l'altra del replicato (ad es. LIST1) e la sua copia (ad es. LIST63). Tali immagini mostrano la distribuzione del parametro QD associata a ognuna delle varianti alleliche presenti nelle sequenze geniche dei batteri trattati.

Quality by Depth (QD)			
Chrom	**Media**	**Varianza**	**Std Dev**
LIST1	29,97	9,65	3,10
LIST63	29,95	9,58	3,09
LIST2	30,01	9,45	3,07
LIST64	29,96	9,50	3,08
LIST3	29,97	9,55	3,09
LIST65	29,97	9,39	3,06

***Tabella 10**: Media, Varianza e StdDev del parametro QD per i batteri studiati.*

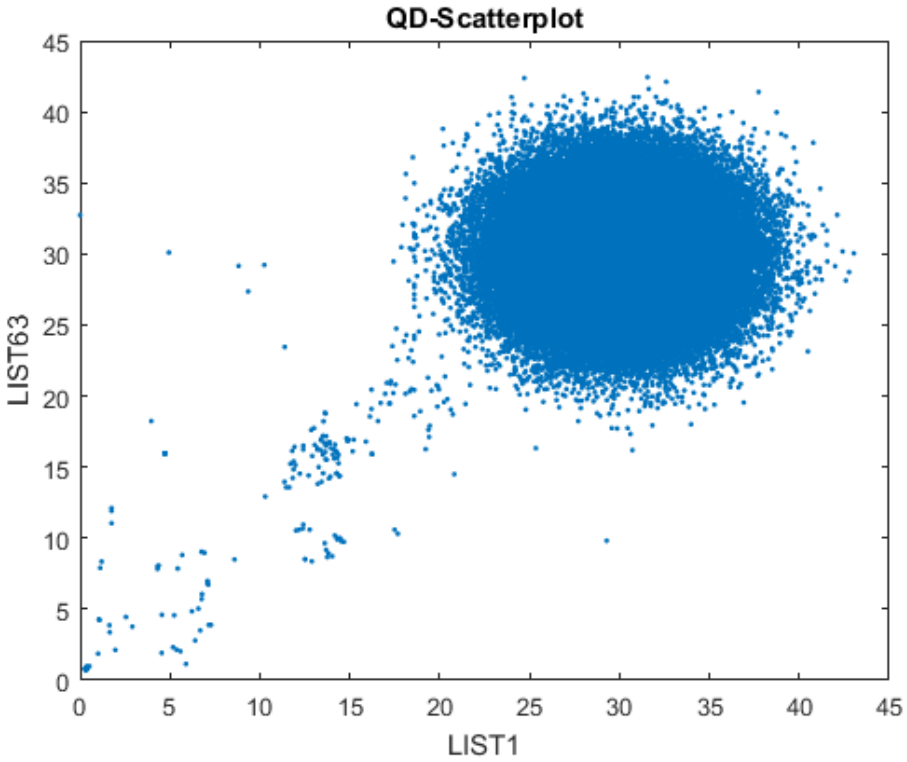


***Figura 41**: Scatter plot delle distribuzioni del parametro QD una contro l'altra dell'isolato (LIST1) e la sua copia (LIST63).*

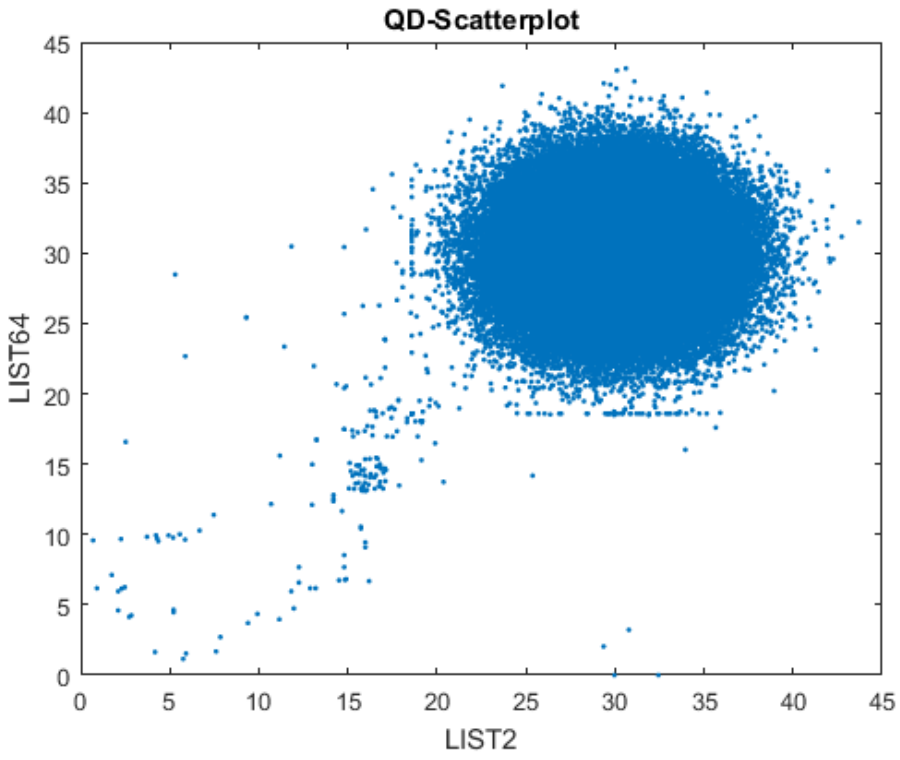


***Figura 42:** Scatter plot delle distribuzioni del parametro QD una contro l'altra dell'isolato (LIST2) e la sua copia (LIST64).*

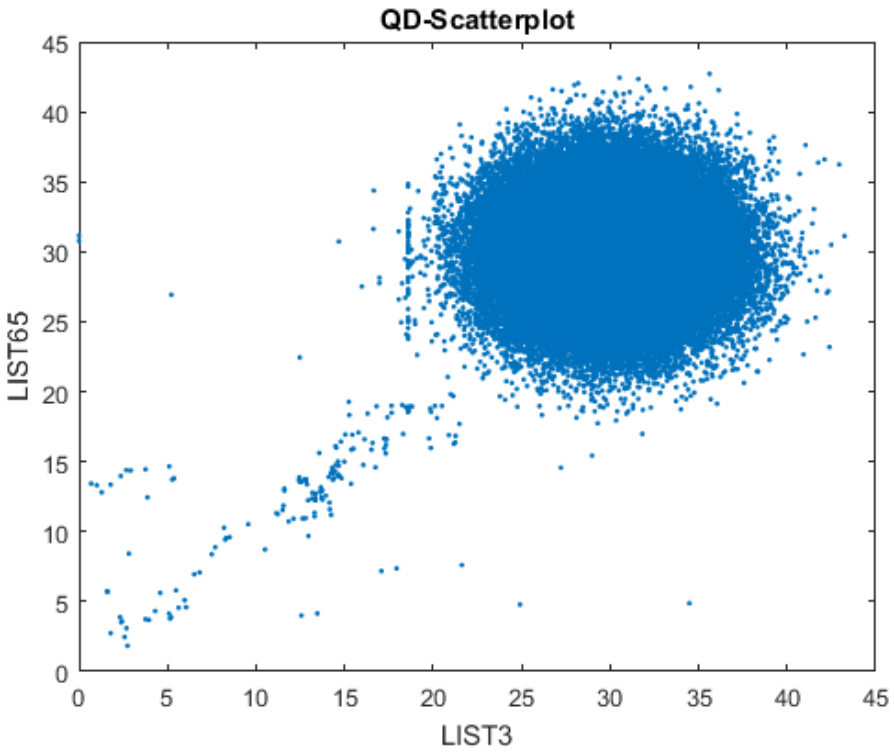


Figura 43: *Scatter plot delle distribuzioni del parametro QD una contro l'altra dell'isolato (LIST3) e la sua copia (LIST65).*

Anche nel seguente caso, per confrontare le distribuzioni, si è deciso di procedere al calcolo del coefficiente di correlazione.

I risultati ottenuti sono seguenti:

Campioni	Coefficiente di correlazione (QD)
LIST1 – LIST63	0.0611
LIST2 – LIST64	0.0414
LIST3 – LIST65	0.0482

Tabella 11*: coefficienti di correlazione per ogni coppia di batteri relativi al parametro DP.*

Gli scatter plots considerati mostrano la relazione che esiste tra le distribuzioni del parametro QD degli SNPs. Osservando i valori riportati in tabella 10, si può notare che entrambi i tre grafici di figura 41, 42 e 43 hanno un grado di correlazione vicinissimo allo 0.
Nel seguente caso, le variabili si dicono *incorrelate;* non esiste quindi alcun legame che unisce tali parametri.

4.2.1 QD: Test di Bartlett

Proprio come per il coverage, prima di eseguire l'ANOVA, risulta necessaria una verifica della normalità delle distribuzioni. Si è, dunque, costruita, anche in questo caso, una matrice *M* composta da 6 colonne, contenente però i valori del parametro QD relativi a ogni batterio.

L'esecuzione del Test di Bartlett, fornisce un *p-value* pari a zero come visibile nell'immagine di figura 44.
Tale valore indica che l'ipotesi nulla viene scartata e che le distribuzioni considerate non sono normali.

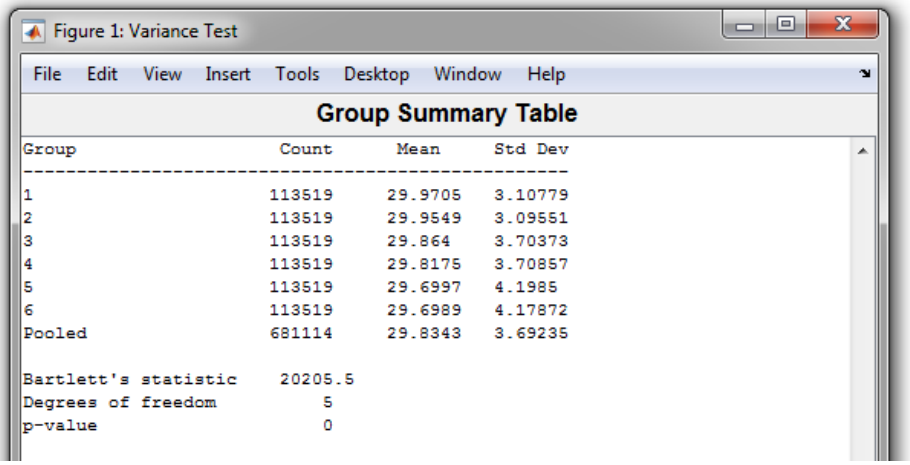


Figura 44: *test di Bartlett per il parametro QD*

Gli istogrammi sottostanti confermano la non normalità di tali distribuzioni. A un primo sguardo si potrebbe notare che l'andamento assume la forma di una curva di gauss, tuttavia questo fatto viene smentito dalle lunghe code (seppure poco visibili) presenti in quasi tutti i grafici, confermando appunto quanto detto.

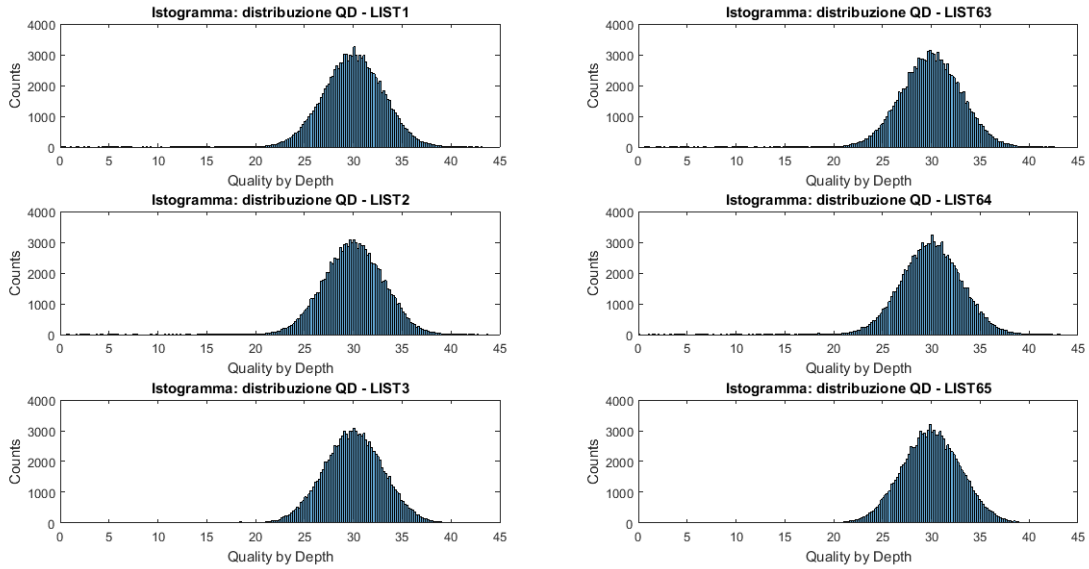


***Figura 45**: Istogrammi delle misure di QD per gli SNPs dei batteri in esame.*

4.2.2 QD: ANOVA & K-W

Vengono riportati, ora, i risultati ottenuti dopo aver eseguito i test di analisi di varianza (ANOVA) e la sua versione non parametrica (Test di K-W).

Le figure 46 e 47 mostrano quanto ottenuto:

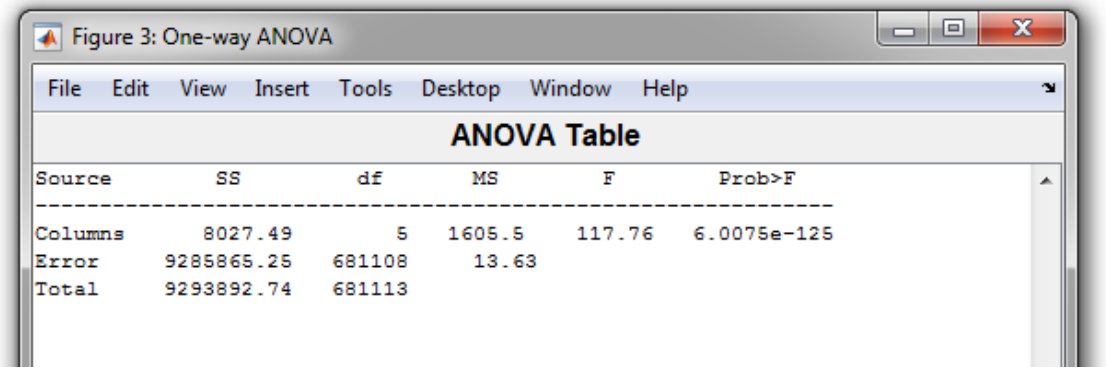


***Figura 46:** Risultati del test di Analisi di Varianza per il parametro QD.*

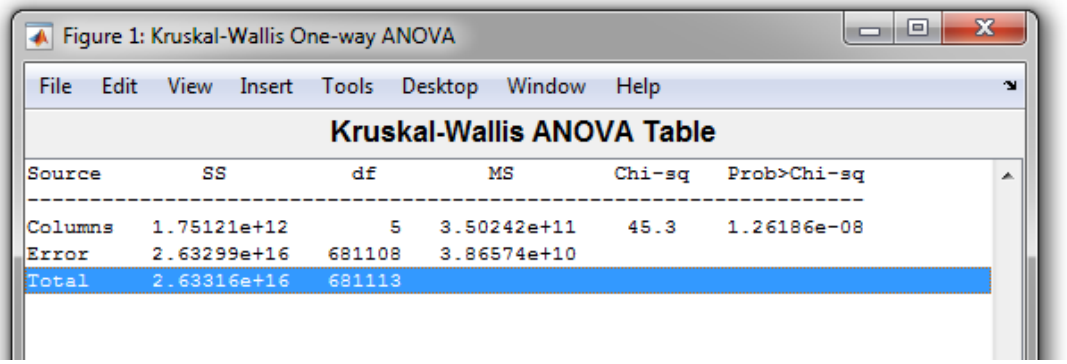


***Figura 47**: risultati del test di Kruskal-Wallis per il parametro QD.*

Il valore del *p-value* (prossimo allo 0) riscontrato sia per l'ANOVA che per il test di Kruskal-Wallis, porta a rifiutare l'ipotesi nulla, confermando il fatto che le distribuzioni considerate hanno una media significativamente differente.

4.2.3 QD: Multiple Comparisons

Anche in questo caso, poiché il test di *K-W* rifiuta l'ipotesi nulla, risulta opportuno effettuare l'operazione dei confronti multipli per vedere quali gruppi hanno una media significativamente differente dagli altri.

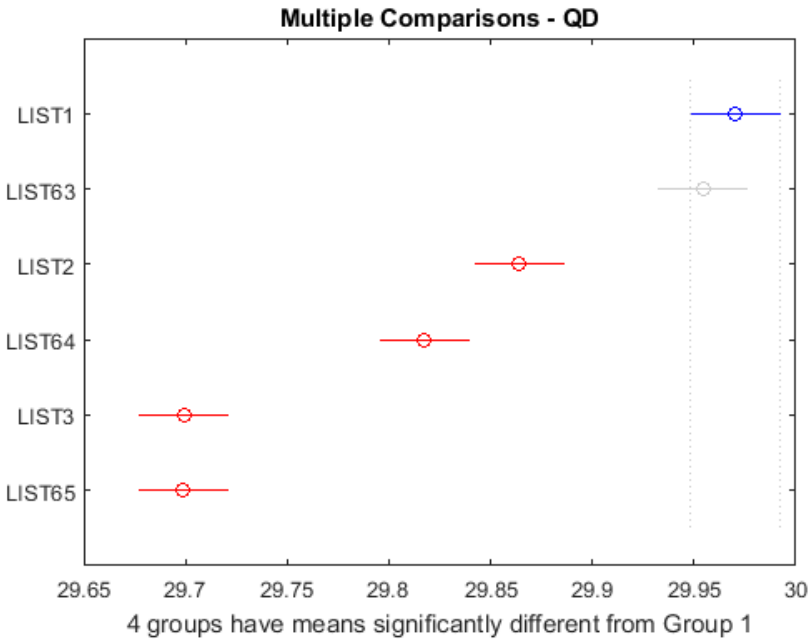


***Figura 48**: diagramma dei confronti multipli tra i campioni per il parametro QD.*

Come si può notare dal grafico di figura 48, le coppie LIST1-LIST63 e LIST3-LIST65 presentano intervalli che si sovrappongono (quasi completamente nel secondo), il che indica che le medie di questi gruppi **non** sono particolarmente differenti e, di conseguenza, **non** viene scartata l'ipotesi nulla. Per quanto riguarda invece la coppia LIST2-LIST64, sembra che non ci

sia una particolare sovrapposizione di intervalli, rivelando quindi una significativa differenza tra le medie.

La tabella 11, conferma quanto detto: infatti i valori del *p-value* più alti si trovano in corrispondenza delle coppie citate sopra, dove per una migliore visualizzazione sono stati cerchiati in rosso. Tali valori indicano l'accettazione dell'ipotesi nulla H_0.

Campione A	Campione B	Lim. Inferiore	Differenza medie	Lim. Superiore	*p-value*
LIST1	**LIST63**	**-0,03**	**0,02**	**0,06**	**0,92**
LIST1	LIST2	0,06	0,11	0,15	0,00
LIST1	LIST64	0,11	0,15	0,20	0,00
LIST1	LIST3	0,23	0,27	0,31	0,00
LIST1	LIST65	0,23	0,27	0,32	0,00
LIST63	LIST2	0,05	0,09	0,14	0,00
LIST63	LIST64	0,09	0,14	0,18	0,00
LIST63	LIST3	0,21	0,26	0,30	0,00
LIST63	LIST65	0,21	0,26	0,30	0,00
LIST2	LIST64	0,00	0,05	0,09	0,03
LIST2	LIST3	0,12	0,16	0,21	0,00
LIST2	LIST65	0,12	0,17	0,21	0,00
LIST64	LIST3	0,07	0,12	0,16	0,00
LIST64	LIST65	0,07	0,12	0,16	0,00
LIST3	**LIST65**	**-0,04**	**0,00**	**0,04**	**1,00**

***Tabella 12**: tabella riassuntiva contenente i risultati del test dei confronti multipli, in rosso sono cerchiati i valori per cui si è ottenuto un valore alto del p-value.*

4.2.4 QD: Box Plots

La figura 49, rappresenta i box plots per le distribuzioni del parametro QD degli SNP dei batteri studiati.
Dall'immagine, risulta evidente che la mediana di tutti i gruppi, rappresentata dalla linea rossa interna ai box, si aggira attorno allo stesso valore. Si può, quindi, concludere, con un 95% di confidenza, che le mediane stesse non differiscono particolarmente l'una dall'altra. A conferma di ciò, la tabella 12 evidenzia la forte somiglianza tra le mediane dei campioni.

Le croci rosse che si estendono sia sopra che sotto ogni box plots, rappresentano tutti i possibili *outliers*, che, come si può notare, sono molto numerosi.

Infine, poiché la linea rossa interna appare abbastanza centrata per ogni gruppo, si può concludere che le distribuzioni dei campioni sono abbastanza simmetriche.

Mediane dei campioni					
LIST1	**LIST63**	**LIST2**	**LIST64**	**LIST3**	**LIST65**
30,01	29,98	30,02	29,97	29,97	29,95

***Tabella 13**: valori delle mediane dei campioni.*

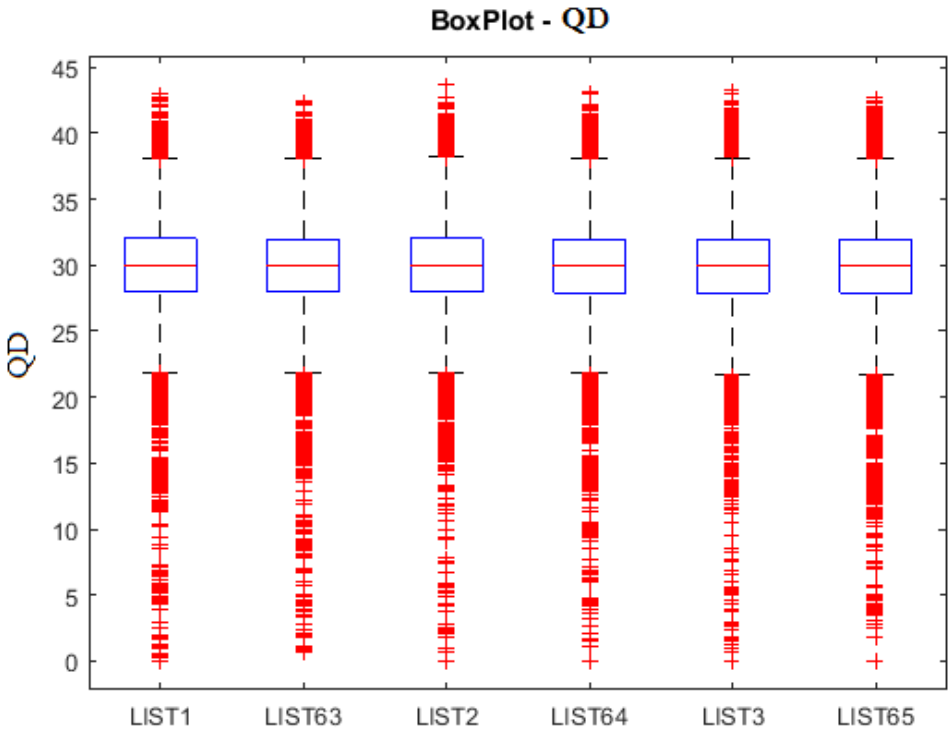


***Figura 49**: Box plots per le distribuzioni di QD relative agli SNP dei campioni.*

4.3 Strand Odds Ratio (SOR)

Una molecola di DNA possiede una struttura a doppia elica ed è quindi costituita da due filamenti, in cui uno viene chiamato filamento positivo (**sense strand**) e l'altro filamento negativo (**anti-sense strand**) come visibile in figura 50.

Il parametro SOR [24] è un metodo che permette di valutare se esiste un bias di strand nei dati; in particolare viene utilizzato per determinare se c'è un bias di strand tra il filamento positivo e quello negativo, per l'allele di riferimento o quello alternativo. Il valore calcolato da questo parametro è riportato in scala *ln*.

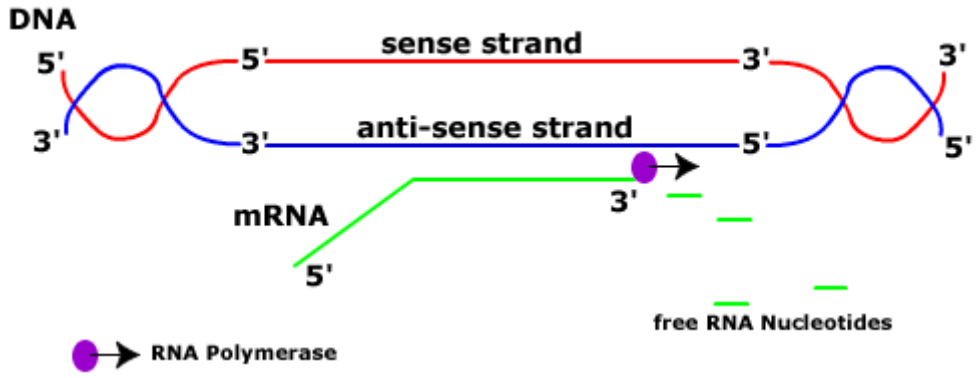


***Figura 50**: rappresentazione della lettura del DNA (in questo caso da parte dell'RNA polimerasi). Il filamento che viene letto da 5' a 3' (rosso) è il sense strand, mentre quello che viene letto da 3' a 5' (blu) è l'anti-sense strand.*

Di seguito viene riportata la tabella contenente i principali parametri statistici relativi al parametro SOR e gli scatter-plot delle distribuzioni una contro l'altra per il battere isolato con la sua copia.

Strand Odds Ratio (SOR)			
Chrom	**Media**	**Varianza**	**Std Dev**
LIST1	0,96	0,13	0,36
LIST63	0,88	0,05	0,23
LIST2	1,04	0,15	0,39
LIST64	0,86	0,05	0,21
LIST3	0,92	0,11	0,33
LIST65	0,88	0,05	0,21

***Tabella 14:** Media, Varianza e StdDev per il parametro SOR.*

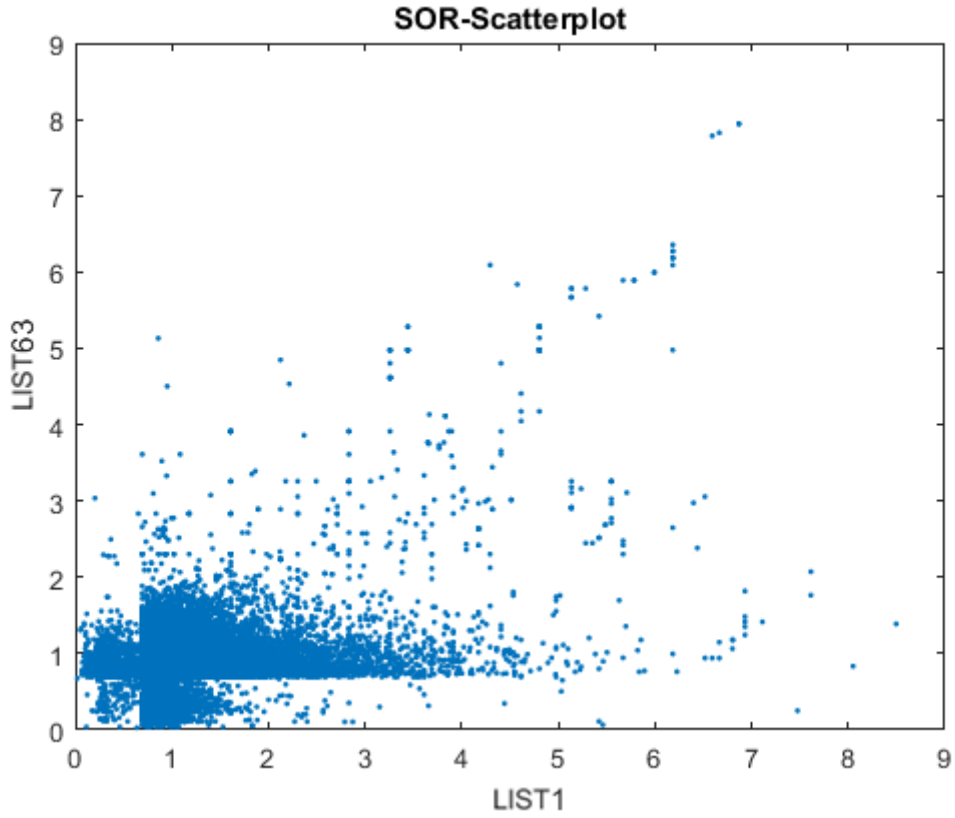


Figura 51: *Scatter plot delle distribuzioni del parametro SOR una contro l'altra dell'isolato (LIST1) e la sua copia (LIST63).*

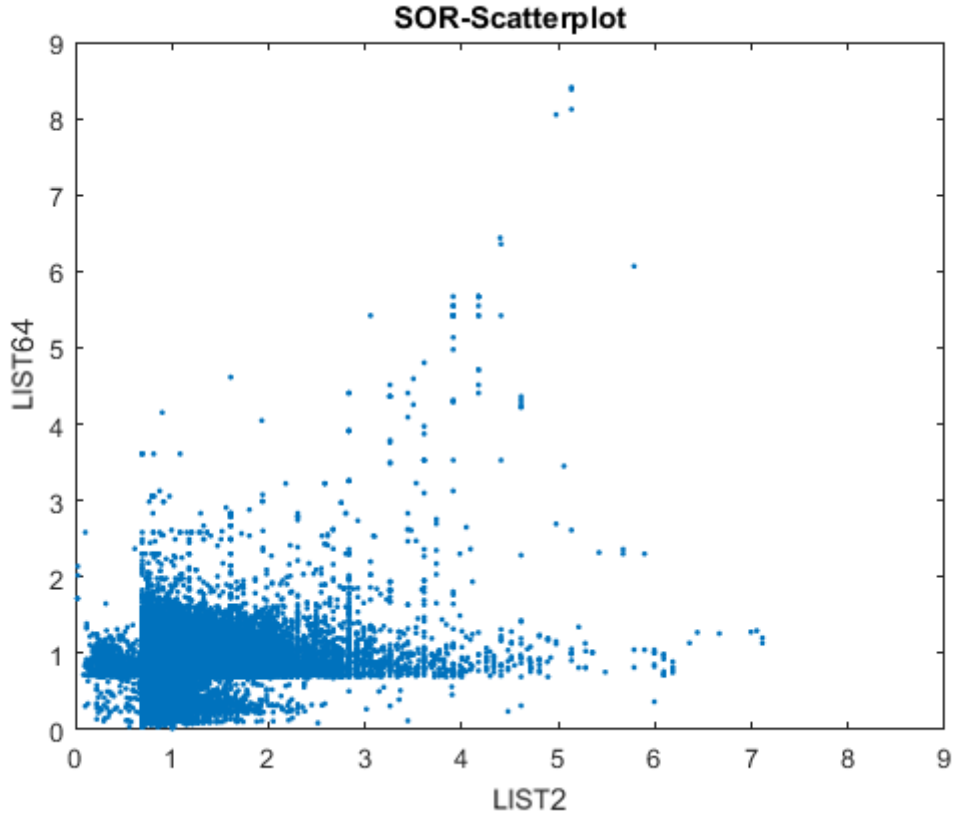


Figura 52*: Scatter plot delle distribuzioni del parametro SOR una contro l'altra dell'isolato (LIST2) e la sua copia (LIST64).*

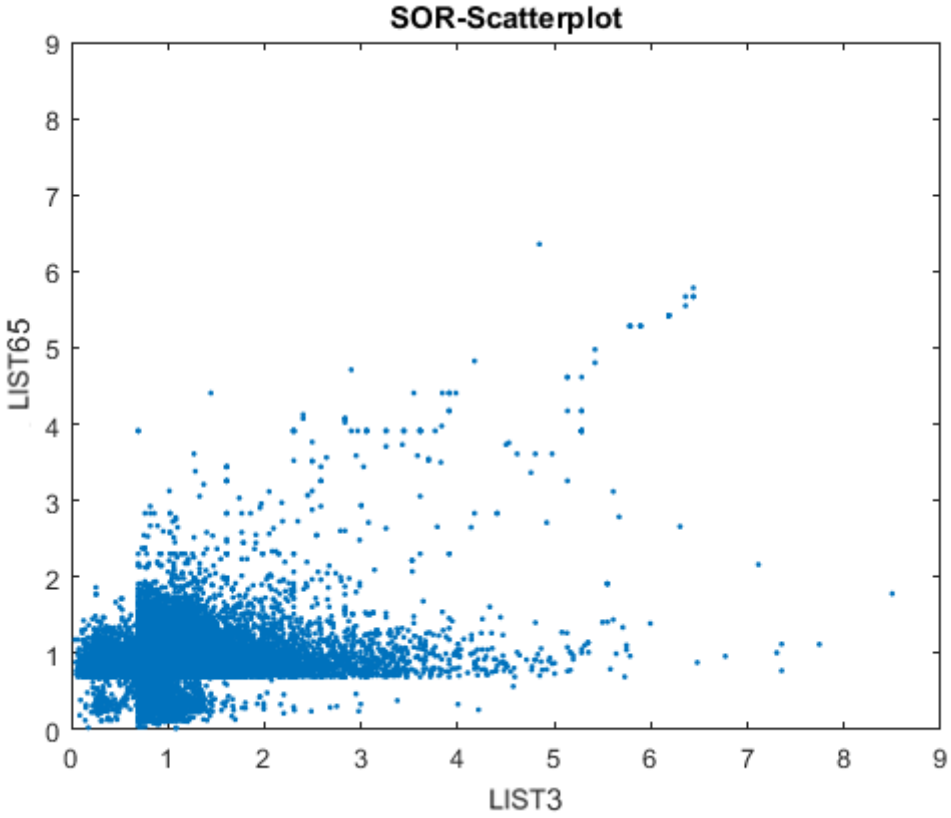


***Figura 53**: Scatter plot delle distribuzioni del parametro SOR una contro l'altra dell'isolato (LIST3) e la sua copia (LIST65).*

Il successivo calcolo del coefficiente di correlazione, servirà per confrontare i grafici appena mostrati.

I risultati ottenuti sono i seguenti:

Campioni	Coefficiente di correlazione (SOR)
LIST1 – LIST63	0.2944
LIST2 – LIST64	0.1842
LIST3 – LIST65	0.2477

***Tabella 15**: coefficienti di correlazione per ogni coppia di batteri relativi al parametro SOR.*

Gli scatter plots considerati, mostrano la relazione che esiste tra le distribuzioni del parametro SOR degli SNPs, e la tabella 14 mette in evidenza come le tre coppie dei campioni hanno un grado di correlazione debole con valori compresi tra 0.18 e 0.30. Anche in questo caso si parla di *correlazione positiva.*

4.3.1 SOR: Test di Bartlett

Proprio come è stato fatto per il coverage e per il parametro QD, anche in questo caso è stato eseguito un test di Bartlett per la verifica della normalità delle distribuzioni. Lanciando il comando `vartestn` si ottiene quanto riportato in figura 54:

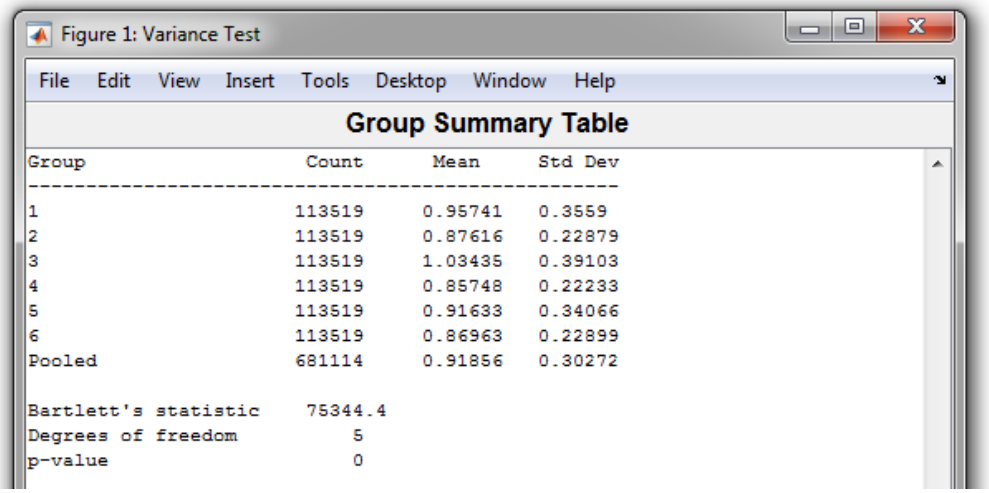


***Figura 54**: Risulti ottenuti per il test di Bartlett.*

Prendendo in considerazione il *p-value (p = 0)* e il risultato della statistica dal test di Bartlett, molto maggiore rispetto al valore critico del Chi^2 per 6-1 gradi di libertà, si può concludere che viene rifiutata l'ipotesi nulla. Il test, dunque, suggerisce che le varianze non sono uguali per tutti 6 i gruppi, per cui non è soddisfatta l'assunzione dell'omogeneità necessaria per eseguire l'ANOVA.

A conferma di questo fatto, sono stati riportati gli istogrammi contenenti l'andamento del parametro studiato. Guardando l'immagine di figura 55, ci si accorge subito che le distribuzioni non sono normali, ma presentano un forte grado di asimmetria, con valori per lo più compresi tra 0 e 2.

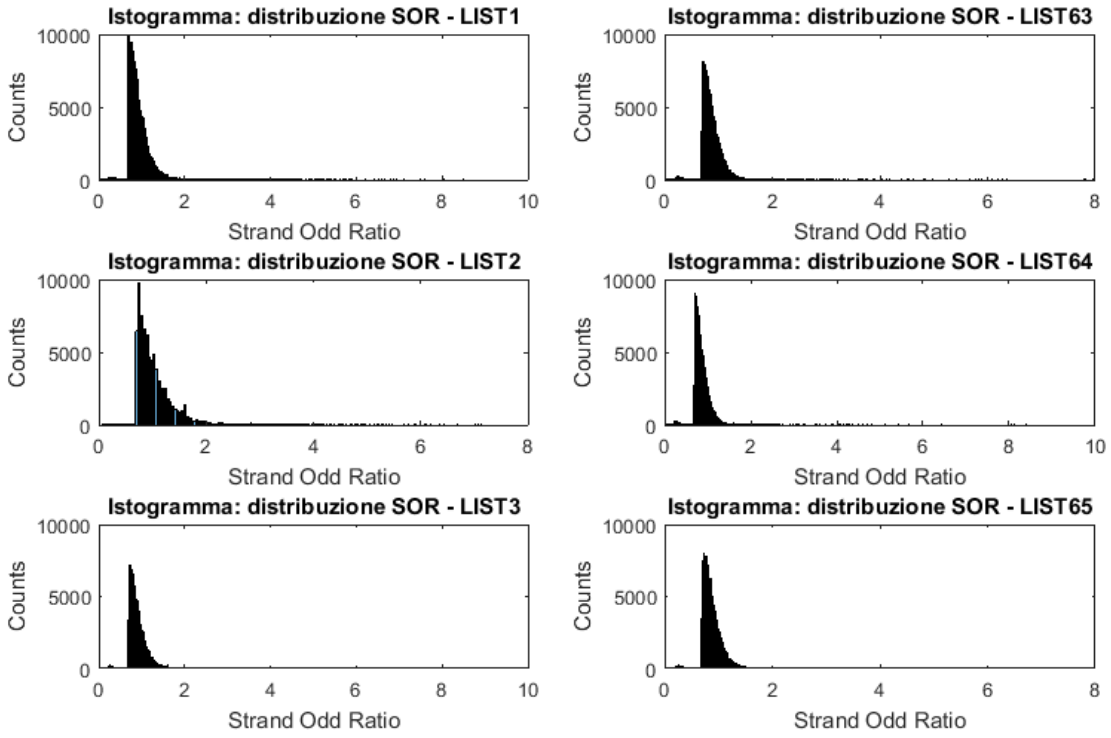


Figura 55: *Istogrammi delle misure di SOR per gli SNPs dei batteri in esame. Si può notare il forte grado di asimmetria per ogni distribuzione.*

4.3.2 SOR: ANOVA & K-W

Anche in questo caso, nonostante il test di Bartlett indichi la non normalità delle distribuzioni, è stato eseguito il test per l'analisi di varianza (ANOVA), assieme alla sua versione non parametrica (K-W). I risultati di questi test sono riportati rispettivamente nelle figure 56 e 57.

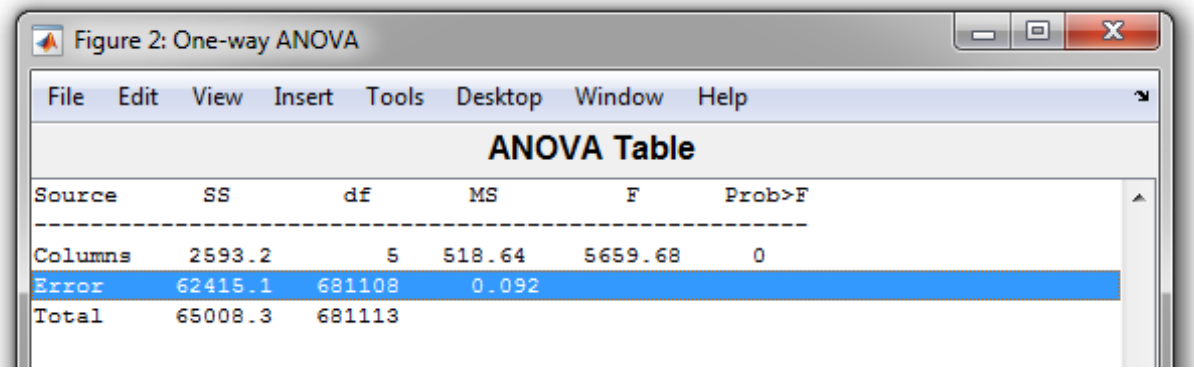


Figura 56*: risultati del test di Analisi di Varianza per il parametro SOR.*

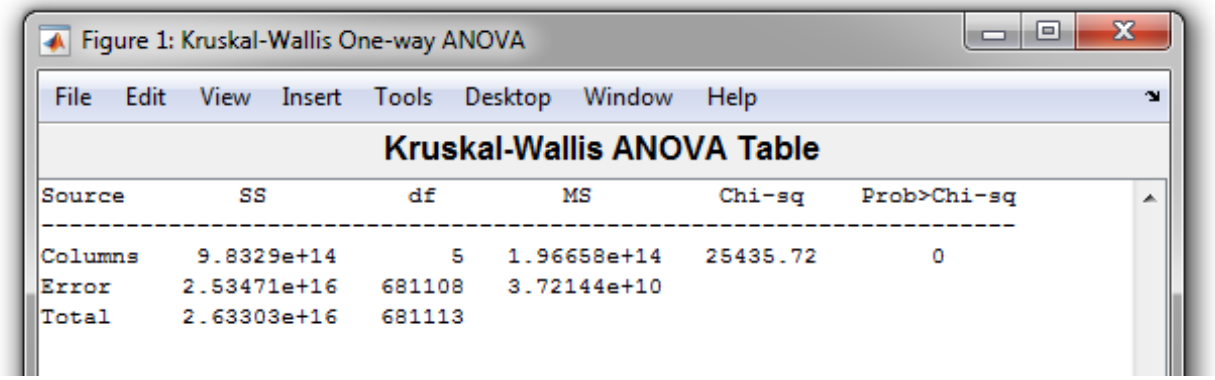


***Figura 57**: risultati del test di Kruskal-Wallis per il parametro SOR.*

In entrambi i casi il valore del *p-value* di 0, indica che la differenza tra le medie dei sei gruppi è significativa, risultando quindi in accordo con il test di Bartlett.

4.3.3 SOR: Multiple Comparisons

Viene eseguito, in ultima analisi, il test dei confronti multipli per il parametro SOR. Tale test è utilizzato per verificare quali gruppi hanno una media significativamente differente dagli altri. I risultati, illustrati in figura 58, portano a concludere che nessun gruppo presenta una media significativamente uguale ad un'altra, in quanto gli intervalli indicati non mostrano alcuna sovrapposizione. In aggiunta, guardando i valori del *p-value* della tabella 15, si può notare che hanno tutti un valore pari a 0, rifiutando quindi l'ipotesi nulla che i 6 gruppi abbiano la stessa media.

***Figura 58**: Diagramma dei confronti multipli tra gruppi. Non si verifica nessuna sovrapposizione di intervalli.*

Campione A	Campione B	Lim. Inferiore	Differenza medie	Lim. Superiore	*p-value*
LIST1	LIST63	0,08	0,08	0,08	0,00
LIST1	LIST2	-0,08	-0,08	-0,07	0,00
LIST1	LIST64	0,10	0,10	0,10	0,00
LIST1	LIST3	0,04	0,04	0,04	0,00
LIST1	LIST65	0,08	0,09	0,09	0,00
LIST63	LIST2	-0,16	-0,16	-0,15	0,00
LIST63	LIST64	0,02	0,02	0,02	0,00
LIST63	LIST3	-0,04	-0,04	-0,04	0,00
LIST63	LIST65	0,00	0,01	0,01	0,00
LIST2	LIST64	0,17	0,18	0,18	0,00
LIST2	LIST3	0,11	0,12	0,12	0,00
LIST2	LIST65	0,16	0,16	0,17	0,00
LIST64	LIST3	-0,06	-0,06	-0,06	0,00
LIST64	LIST65	-0,02	-0,01	-0,01	0,00
LIST3	LIST65	0,04	0,05	0,05	0,00

***Tabella 16**: tabella riassuntiva contenente i risultati del test dei confronti multipli.*

4.3.4 SOR: Box Plots

La figura 59, rappresenta i box plots per le distribuzioni del parametro SOR degli SNP dei batteri studiati.

Ancora una volta, dall'immagine si potrebbe dedurre che le mediane dei box plots non differiscano particolarmente l'una dall'altra, in quanto le tacche (notches) appaiono sovrapposte. Tuttavia, tale impressione viene smentita dai *p-value*, i cui valori ottenuti assumono il valore 0. Quindi, sebbene i valori delle mediane presenti nella tabella 16, si aggirino tutti attorno alla stessa cifra, eccezione fatta per il battere LIST2 che possiede un valore più alto rispetto agli altri gruppi, tale vicinanza non garantisce al 95% la similarità tra le 6 mediane.
Questo fatto è dovuto alla non sovrapposizione dei notches (intervallo di confidenza delle mediane) permettendo di giungere alla conclusione che le mediane differiscono significativamente.

La linea rossa interna a tutti i box, non appare troppo centrata in nessun gruppo, il che indica che le distribuzioni dei campioni possiedono un certo grado di asimmetria, in accordo con la figura 55 vista sopra.
Infine, le numerosissime croci rosse che si estendono sia sopra che sotto ogni box, evidenziano gli *outliers*.

Mediane dei campioni					
LIST1	LIST63	LIST2	LIST64	LIST3	LIST65
0,878	0,838	**0,941**	0,826	0,853	0,838

***Tabella 17**: valori delle mediane dei campioni.*

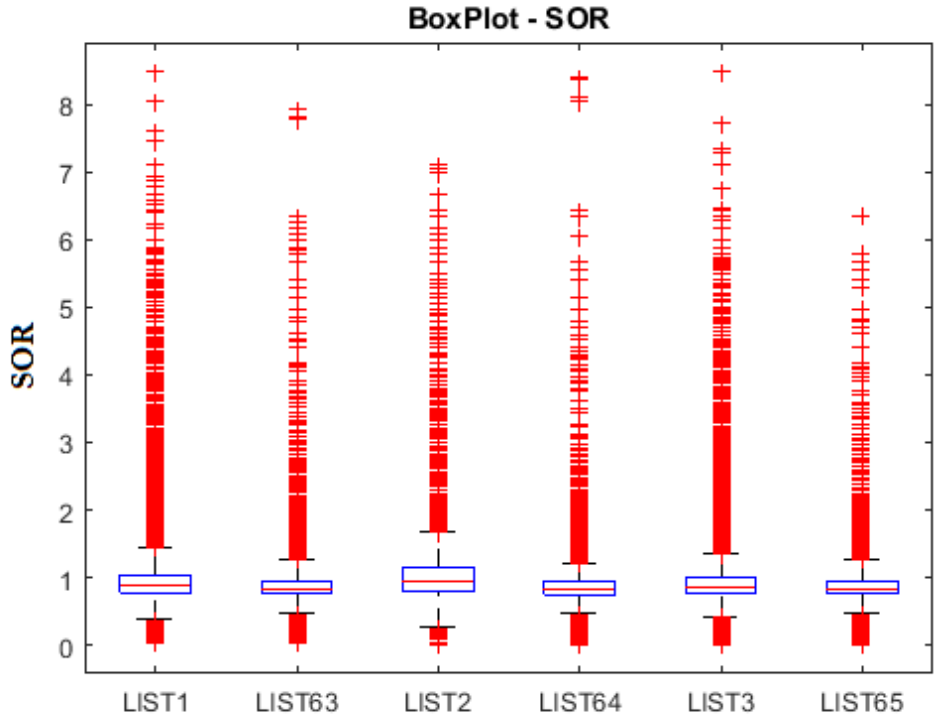


***Figura 59**: Box plots per le distribuzioni di SOR relative agli SNPs dei campioni,*

Conclusioni

Il presente lavoro di tesi è basato sullo studio degli SNPs presenti nel DNA di sei campioni del batterio ***Listeria monocytogenes***; in particolare sono stati analizzati tre isolati diversi con le tre relative repliche.

Il metodo con cui tali SNPs sono stati individuati, è risultato essere molto solido ed efficace, nonostante presenti alcuni svantaggi legati ai tempi di calcolo. Infatti, la pipeline seguita, comprende l'utilizzo di programmi che richiedono periodi di tempo parecchio lunghi per l'esecuzione (ad esempio BWA) e per i quali si sta cercando di ottimizzare la procedura.
Nonostante ciò, la parte di analisi statistica, seppure ancora preliminare, fornisce dei risultati quantomeno incoraggianti i quali indicano che la ripetibilità del sequenziamento in tale specie batterica, è abbastanza buona.
I risultati ottenuti mettono in evidenza non solo l'elevato numero di SNPs in comune individuati tra i vari isolati batterici e le proprie copie, ma anche il numero di SNPs in comune tra un isolato e l'altro.

Per quanto riguarda, invece, i parametri presi in considerazione, relativi a tali varianti, come ad esempio il "Coverage" e il "Quality by Depth", questi forniscono dei risultati incoraggianti. Infatti, nel primo parametro si è riscontrato un buon grado di correlazione tra un isolato batterico e la sua copia, soprattutto quando si è andati ad analizzare il grado di copertura su tutto il genoma. Per quanto riguarda il secondo, invece, si può notare che, facendo riferimento ai box plots, sussiste un buon grado di similarità tra le distribuzioni.

Chiaramente, tale studio statistico, seppure promettente, è ancora incompleto in quanto necessita di un'analisi più dettagliata, che richiede l'utilizzo non solo di più campioni a disposizione, ma anche di tempi ben più lunghi.

Ringraziamenti

Desidero ricordare tutti coloro che mi hanno aiutato nella stesura della tesi con suggerimenti, critiche ed osservazioni: a loro va la mia gratitudine, anche se a me spetta la responsabilità per ogni errore contenuto in questo lavoro.
Ringrazio *in primis* il professor Gastone Castellani, che mi ha permesso di intraprendere questo lavoro e senza il quale questa tesi non esisterebbe. Successivamente ringrazio il professor Daniel Remondini per la disponibilità e il supporto tecnico.
Un ringraziamento speciale va inoltre a Italo Faria Do Valle, che mi ha accompagnato e guidato con grande esperienza durante tutto il periodo lavorativo.

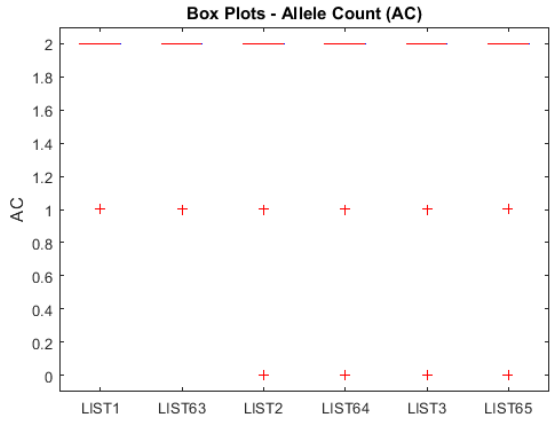
Infine, un mega grazie va a tutti quelli che mi sono stati vicino e che mi hanno sopportato con immensa pazienza.

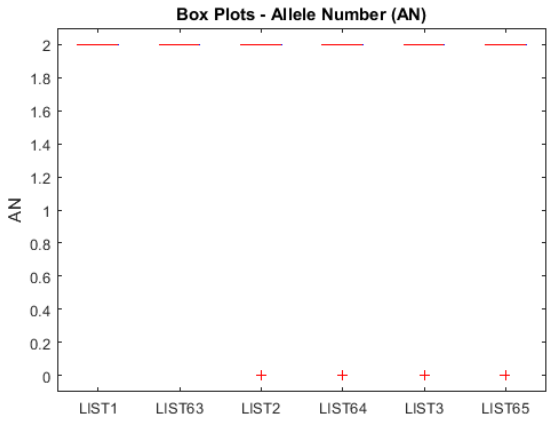
Grazie di tutto.

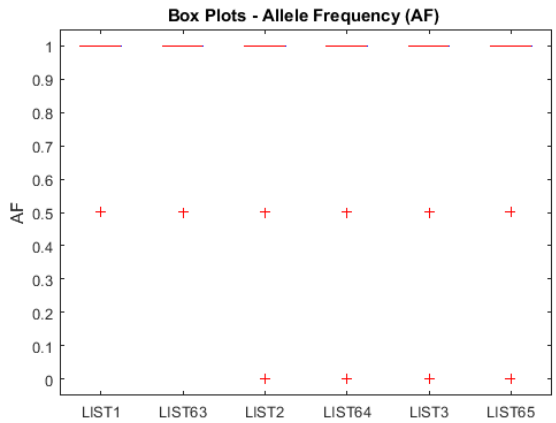
Appendice

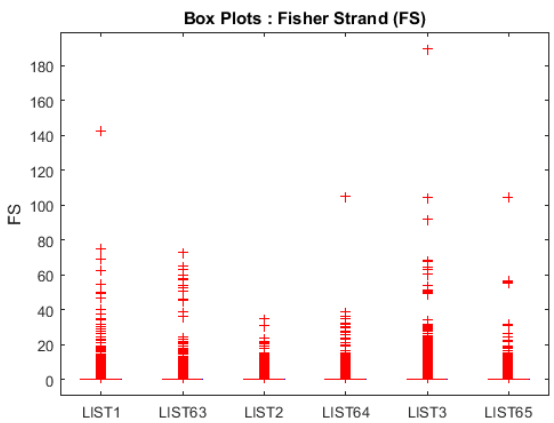
Di seguito, sono riportati i box plots dei parametri che si è scelto di non studiare poiché poco significativi ai fini di indagini statistiche.

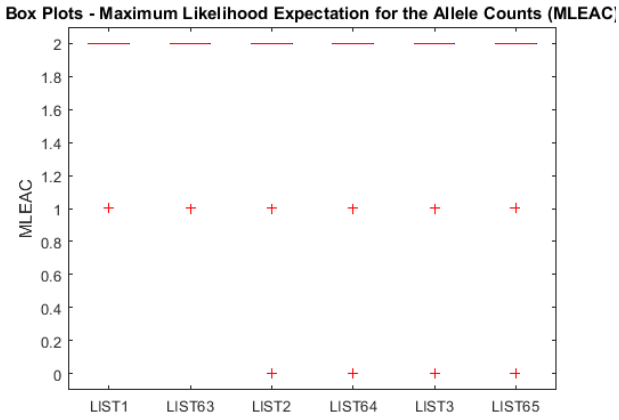
Si può notare come AC, AN, AF, MLEAC e MLEAF assumano valori fissi come 1 o 2. I restanti invece (FS, BaseQRankSum, ClippingRankSum, ReadPosRankSum, MQRankSum), hanno la mediana dei valori centrata sullo zero, questo perché tali parametri non sono comuni a tutti gli SNPs; di conseguenza, gli *outliers* rappresentano tutti quegli SNPs che possiedono un valore di quel determinato parametro.

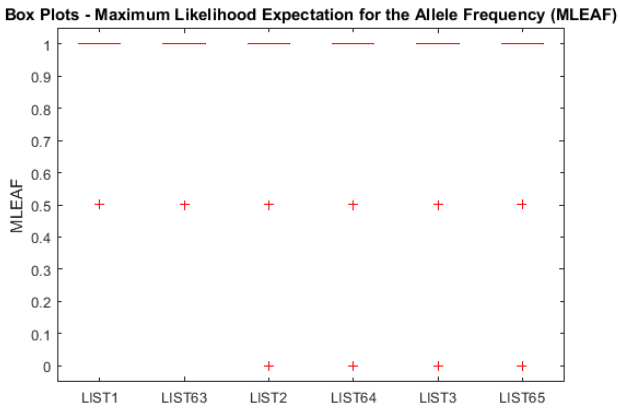


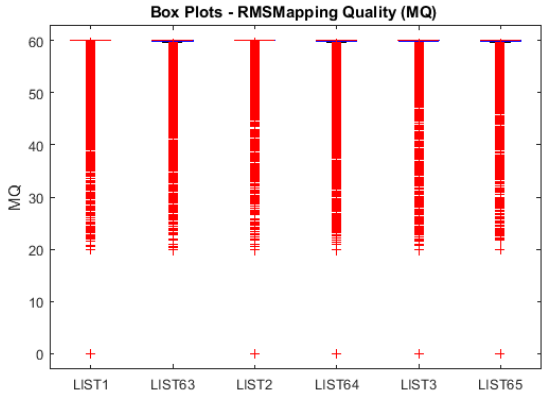


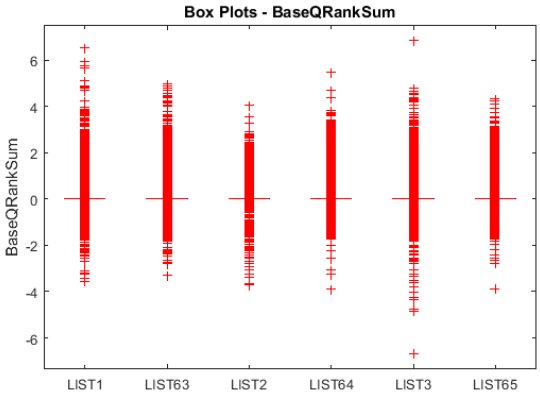


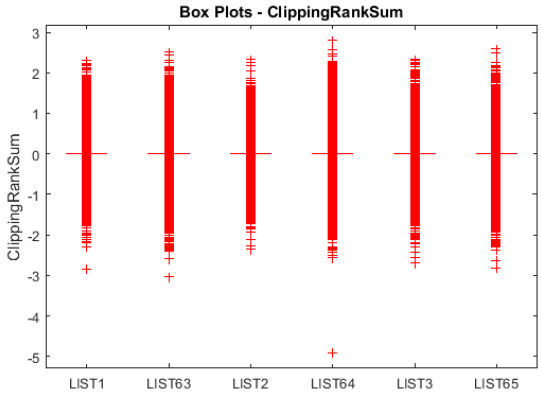


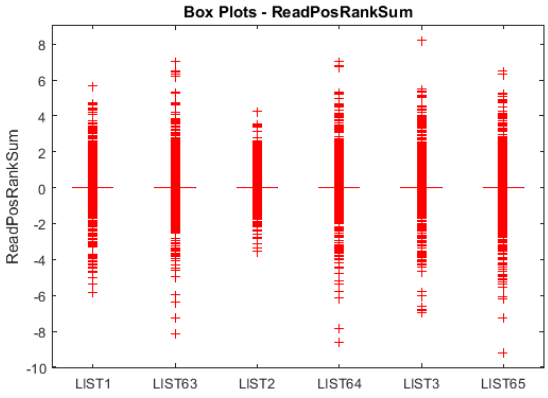

Box Plots - Maximum Likelihood Expectation for the Allele Counts (MLEAC)
MLEAC
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65

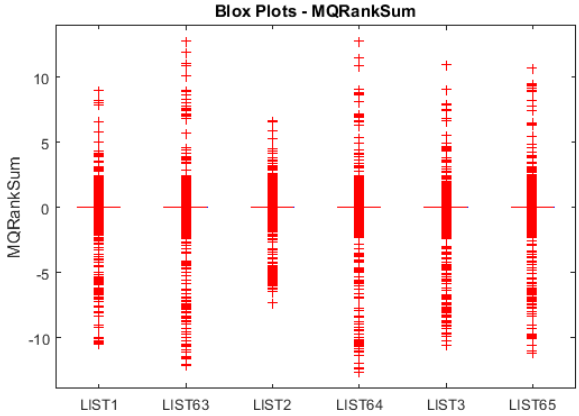

Box Plots - Maximum Likelihood Expectation for the Allele Frequency (MLEAF)
MLEAF
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65


Box Plots - RMSMapping Quality (MQ)
MQ
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65


Box Plots - BaseQRankSum
BaseQRankSum
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65


Box Plots - ClippingRankSum
ClippingRankSum
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65


Box Plots - ReadPosRankSum
ReadPosRankSum
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65


Blox Plots - MQRankSum
MQRankSum
LIST1
LIST63
LIST2
LIST64
LIST3
LIST65

Bibliografia

1. Geoffrey M. Cooper. *La cellula un approccio molecolare*. Zanichelli, 1998.
2. Benjamin Lewin. *Il Gene VI*. Zanichelli, 1999.
3. Tom Strachan and Andrew P. Read. *Human molecular genetics 2*. Bios, 1999.
4. https://en.wikipedia.org/wiki/*Listeria_monocytogenes*
5. Berglind H, Pawitan Y, Kato S, Ishioka C, Soussi T. *Analysis of Tp53 mutation status in human cancer cell lines: a paradigm for cell line cross-contamination*. Cancer Biol Ther, 2008
6. https://biof-edu.colorado.edu/videos/dowell-short-read-class/day-4/fastqc-manual
7. http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic Manual_V0.32.pdf.
8. https://insidedna.me/tool_page_assets/pdf_manual/bwa.pdf.
9. H. Li, R. Durbin: *Fast and accurate short read alignment with BurrowsDWheeler transform*, Bioinformatics, 2009, vol. 25 n. 14.
10. https://en.wikipedia.org/wiki/Burrows%E2%80%93Wheeler_transform
11. https://broadinstitute.github.io/picard/
12. https://software.broadinstitute.org/gatk/
13. 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.
14. https://software.broadinstitute.org/gatk/gatkdocs/3.7-0/AnnotationModules
15. A. Camussi, F. Möller, E. Ottaviano, M. Sari Gorla - "*Metodi statistici per la sperimentazione biologica*", Zanichelli.
16. https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_oneway_analysis_of_variance
17. http://www.biostathandbook.com/multiplecomparisons.html
18. https://en.wikipedia.org/wiki/Box_plot
19. Cossu, Carla (2013) *Nuovi approcci molecolari per lo studio di malattie monogeniche rare: utilizzo dell'exome sequencing per la ricerca di geni malattia,* Ph.D Thesis.
20. http://bedtools.readthedocs.io/en/latest/content/tools/coverage.html
21. https://it.mathworks.com/help/stats/multcompare.html
22. http://www.phrap.com/phred/
23. http://gatkforums.broadinstitute.org/gatk/discussion/4260/phred-scaled-quality-scores
24. https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_gatk_tools_walkers_annotator_StrandOddsRatio.php