



**UNIVERSITÀ
DEGLI STUDI
DI PADOVA**



**DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE**

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea in

INGEGNERIA INFORMATICA

**Stima della dimensione del genoma tramite k-mers:
confronto tra metodi computazionali**

Relatore:
Prof. Matteo Comin

Laureando:
Mattia Tamiazzo

Data di laurea 23/09/2022

Anno Accademico 2021-2022

Sommario

La dimensione del genoma è la quantità totale di DNA nucleare aploide presente nelle cellule di un organismo. La determinazione della dimensione del genoma costituisce un argomento di interesse, perché non esistono valori di riferimento assoluti che permettano di stabilire quale approccio sia più efficace, e perché i metodi sperimentali per la sua misurazione sono attualmente costosi dal punto di vista temporale ed economico. Una soluzione alla stima della dimensione del genoma con metodi computazionali è l'utilizzo di k-mers, sottostringhe di DNA di lunghezza k. Questa trattazione si pone l'obiettivo di analizzare e comparare vari approcci algoritmici pubblicati in letteratura per la stima della dimensione del genoma, che tuttavia spesso forniscono risultati contrastanti e di difficile valutazione assoluta. Grazie ai limitati requisiti richiesti rispetto ai metodi attuali, comunque, la stima basata sui k-mer risulta essere un approccio promettente, anche per lo sviluppo futuro di metodi più precisi ed efficienti.

Indice

1	Introduzione	1
1.1	Metodi di sequenziamento	1
1.2	Stima della dimensione del genoma	3
1.3	Sequenze di lunghezza k: i k-mer	3
2	Metodi analizzati	7
2.1	ALLPATHS-LG	7
2.2	GCE	10
2.3	GenomeScope	13
2.4	findGSE	15
2.5	MGSE	16
3	Valutazione dei metodi	19
3.1	Complessità e performance	19
3.2	Confronti tra i metodi	21
	Conclusioni	31
	Glossario	33
	Bibliografia	35

Capitolo 1

Introduzione

Il sequenziamento del DNA costituisce una tecnica fondamentale per lo studio del genoma di una specie, perché permette di determinare l'ordine dei nucleotidi che costituiscono il DNA. Tale processo trova applicazione in molti studi biologici che riguardano vari ambiti, come ad esempio la medicina riproduttiva, l'oncologia o l'infettivologia, attraverso indagini tra cellule diverse dello stesso individuo o lo studio delle mutazioni genetiche tra individui di una stessa specie [1].

1.1 Metodi di sequenziamento

1.1.1 Metodi di prima generazione

Le prime tecniche di sequenziamento del genoma furono sviluppate nella seconda metà del Novecento. Nel 1977 infatti vennero pubblicati due metodi di sequenziamento: il metodo Sanger, nel quale la sequenza di nucleotidi viene frammentata grazie a un terminatore di catena [2, 3], e il metodo di Maxam e Gilbert, in cui vengono utilizzati reagenti chimici per tagliare il DNA in frammenti in corrispondenza di basi specifiche [4]. Entrambi i metodi prevedono la misurazione dei frammenti creati tramite elettroforesi su gel di poliacrilammide con una corsia per base, in modo che siano separati in ordine di lunghezza e sia possibile dedurre l'ordine delle basi della sequenza in esame [5]. Mentre il metodo Maxam-Gilbert è stato progressivamente accantonato per la difficoltà tecnica e l'uso di sostanze tossiche che lo caratterizzano, il metodo Sanger è stato affinato con l'utilizzo di marcatori fluorescenti diversi per ogni base, che un lettore laser può distinguere restituendo la sequenza di basi di ogni frammento.

1.1.2 Shotgun assembly

Il sequenziamento del genoma può essere fatto su sequenze più lunghe, come un intero cromosoma, tramite *shotgun assembly*, metodo suggerito già nel 1979 [6]. Il genoma iniziale viene duplicato in modo da produrne più copie identiche, le quali vengono tagliate

1.2 Stima della dimensione del genoma

Il problema della misurazione della dimensione del genoma costituisce un argomento di interesse, perché oltre a fornire informazioni sulla sua evoluzione [10], permette di approssimare la quantità di dati che verranno prodotti nel sequenziamento e di valutare la complessità delle sequenze assemblate [11].

Di seguito sono elencate le due principali tipologie di metodi per la stima della dimensione del genoma, basate rispettivamente su approcci sperimentali o computazionali.

1.2.1 Metodi sperimentali

Inizialmente la ricerca scientifica ha cercato di stimare la dimensione del genoma con approcci biochimici, come ad esempio i metodi *Feulgen photometry* o *flow cytometry*. Tali metodologie però, oltre ad essere costose dal punto di vista economico e poco efficienti, devono basarsi su genomi specifici di riferimento [11, 12].

1.2.2 Metodi computazionali

Grazie all'aumento delle capacità computazionali, sono stati sviluppati metodi che possono calcolare la lunghezza del genoma utilizzando i dati ricavati dall'assemblaggio di shotgun. Dato che i dati assemblati sono di solito incompleti, è più conveniente cercarne di stimare la dimensione, utilizzando ad esempio i *k-mer*.

In questa trattazione verranno analizzati e confrontati vari approcci algoritmici che utilizzando i *k-mer* compiono la stima della dimensione del genoma.

1.3 Sequenze di lunghezza *k*: i *k-mer*

I *k-mer* sono tutte le sottostringhe di lunghezza *k* presenti nella sequenza del genoma [13]. Si prenda ad esempio la sequenza descritta in 1.1.

$$AGATTTCGC \tag{1.1}$$

I *k-mer* di lunghezza $k = 1$ saranno le quattro basi che formano la sequenza: *G, T, A, C*. Ponendo invece $k = 2$, i *k-mer* trovati saranno tutte le coppie formate da due basi: *AG, GA, AT, TT, TC, CG, GC*.

Allo stesso modo, per $k = 3$ le sequenze sono: *AGA, GAT, ATT, TTC, TCG, CGC*.

Il listing 1.1 nella pagina seguente mostra una semplice implementazione in pseudocodice per determinare i *k-mer* di lunghezza *k*, iterando la sequenza di input `seq` e dando in output tutte le sottostringhe di lunghezza *k* presenti in essa.

```

1  procedure k-mers(seq, k)
2    lunghezza = length(seq)
3    arr = array di L - k + 1 stringhe vuote
4
5    // La sequenza iniziale viene iterata,
6    // salvando il k-mer n-esimo nell'array di output
7    for n = 0 to L - k + 1 escluso do
8      arr[n] = sottostringa di seq da seq[n] a seq[n+k] escluso
9
10   return arr

```

Listing 1.1: Algoritmo in pseudocodice per la costruzione dei k-mer.

Dato che il numero di k-mer aumenta esponenzialmente all'aumentare del parametro k , sono necessari algoritmi più complessi per il calcolo dei k-mer, come ad esempio *Jellyfish* [13] o *KMC2* [14].

1.3.1 K-mer profile

Il *k-mer profile*, detto anche *k-mer spectrum*, rappresenta un indicatore della complessità del genoma preso in esame. Date in input le letture shotgun del genoma, esso mostra il numero di volte che ogni k-mer viene trovato rispetto la quantità di k-mer distinti presenti, ovvero la molteplicità di ciascun k-mer nella sequenza rispetto il numero di k-mer con quella molteplicità [15]. Un esempio di k-mer profile è mostrato dalla figura 1.2 tratta da [16], in cui si può notare come la natura del genoma influenzi direttamente il grafico in ogni sua componente.

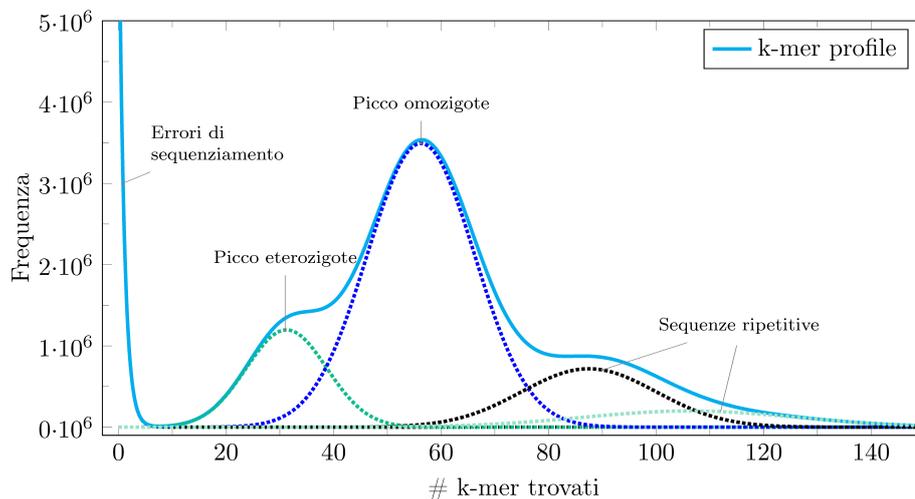


Figura 1.2: Composizione di un k-mer profile.

Ipotizzando che il genoma sia ideale, omozigote e senza ripetizioni, e che le letture siano state fatte senza errori con una certa copertura, il grafico del k-mer profile sarà una [distribuzione di Poisson](#) centrata sulla copertura media disponibile [17].

In casi reali invece, il genoma sarà eterozigote con una certa percentuale di eterozigosi e saranno presenti errori di sequenziamento; il k-mer profile presenterà tre picchi

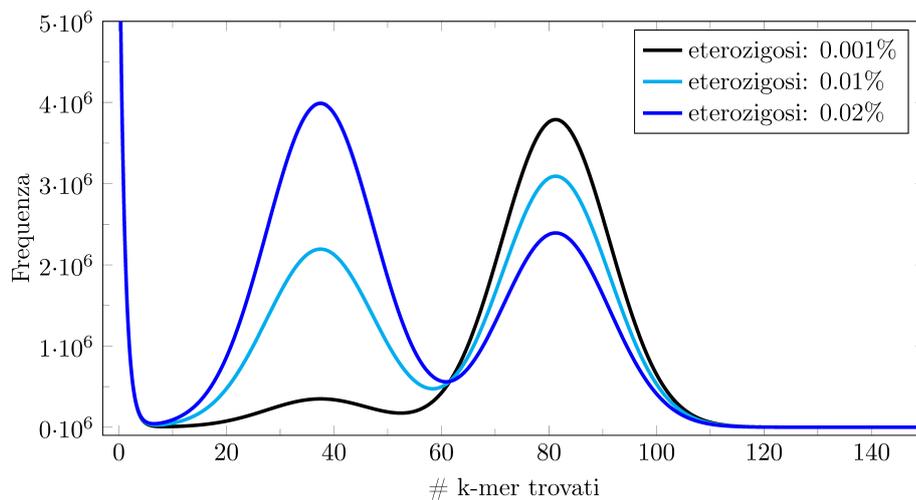


Figura 1.3: Variazione del grafico del k-mer profile al variare dell'eterozigosi.

principali [11]. Il primo picco del grafico corrisponde ai k-mer derivati da errori di sequenziamento, che accadono spesso ma che hanno bassa frequenza perché presentano poche occorrenze nelle letture di input; il secondo invece, rappresenta i k-mer eterozigoti e il terzo quelli omozigoti, presenti quindi su uno o entrambi gli alleli del set di cromosomi. I k-mer eterozigoti devono essere trattati più attentamente, perché possono risultare simili a quelli del primo picco, derivanti da errori di sequenziamento [16].

La lunga coda della distribuzione rappresenta invece le sequenze ripetitive, che occorrono con alta frequenza e sono presenti in un elevato numero di **locus**. Eventuali ripetizioni aggiungono al grafico ulteriori picchi, mentre errori nelle letture aumentano la varianza e producono distorsioni nel grafico.

La figura 1.3 tratta da [18] mostra come all'aumentare del **rapporto di eterozigosi** la quantità di k-mer eterozigoti del secondo picco diventi dominante rispetto ai k-mer omozigoti del terzo picco, che invece diminuiscono.

Capitolo 2

Metodi analizzati

In questo capitolo vengono presentati, in ordine cronologico di pubblicazione, i metodi presi in esame. Ciascun programma viene descritto svolgendo un'analisi dell'algoritmo che lo caratterizza, ed elencando le funzionalità previste per la gestione di eventi particolari, come ad esempio la presenza di errori di sequenziamento.

2.1 ALLPATHS-LG

ALLPATHS-LG è un programma che permette di eseguire il sequenziamento di un genoma tramite de novo assembly di letture shotgun, e che calcola implicitamente la dimensione totale del genoma. Esso si basa sul programma *ALLPATHS* [19, 20] sviluppato precedentemente e, rispetto al suo predecessore, permette l'assembly di genomi di dimensioni maggiori e con copertura minore, di gestire sequenze ripetitive, di correggere errori di lettura e di utilizzare in modo più efficiente le risorse disponibili durante il sequenziamento [21].

2.1.1 Algoritmo

L'algoritmo del programma si basa sul precedente software *ALLPATHS* [19]. Dato un numero minimo k di basi che si sovrappongono nelle letture shotgun, si definisce *branch* una sequenza di k basi (k-mer) che compare in due o più letture diverse, e la cui base successiva o precedente è diversa in ogni lettura. Spezzando il genoma in corrispondenza di ciascun branch, esso viene scomposto in un insieme di sequenze, dette *unipath*. Tali sequenze vengono create a partire dalle letture shotgun di input non allineate.

Formazione dei k-mer path Inizialmente negli shotgun viene corretto il maggior numero di errori di lettura, e vengono poi riconosciuti tutti i k-mer di lunghezza k . In ogni sequenza, ciascun k-mer viene numerato con un numero intero unico; a k-mer già trovati viene assegnato lo stesso valore. In questo modo ciascuna lettura potrà essere espressa come una sequenza di numeri interi, ognuno dei quali rappresenta un k-mer.

GCTGTCGGATCCTCTGTCCGGA		
GCTGTC	100	}
CTGTCC	101	
TGTCCG	102	
GTCCGA	103	
TCCGAT	104	
CCGATC	105	
CGATCC	106	
GATCCT	107	
ATCCTC	108	
TCCTCT	109	
CCTCTG	110	
CTCTGT	111	
TCTGTC	112	
CTGTCC	101	
TGTCCG	102	
GTCCGG	113	}
TCCGGA	114	

Figura 2.1: Esempio di numerazione dei k-mer e traduzione della sequenza in intervalli. Posto $k = 6$, vengono individuati tutti i k-mer presenti, ognuno dei quali viene numerato con un numero unico, riutilizzandolo nel caso di k-mer ripetuti. Usando gli intervalli, la sequenza iniziale viene quindi tradotta in $([100, 112], [101, 102], [113, 114])$.

I numeri della sequenza vengono poi raggruppati in intervalli, in modo da formare i *k-mer path*; essi associano ciascun intervallo al k-mer path a cui esso appartiene, permettendo una facile ricerca di tutte le sequenze che condividono un certo k-mer e semplificando quindi l'assembly delle letture shotgun. Si veda la figura 2.1 per un esempio di come avviene la numerazione dei k-mer e la creazione dei k-mer path. Le letture così tradotte costituiscono il database in cui è possibile effettuare ricerche per la costruzione degli unipath.

Formazione degli unipath Tutti i numeri degli intervalli trovati nei k-mer path vengono scanditi. L'obiettivo per ogni numero è trovare il più lungo intervallo senza interruzioni che lo contenga. Si cercano nel database tutti gli intervalli che contengono quel numero, e si sceglie l'intervallo continuo più lungo, che diventa un *unipath interval*. Il processo è ripetuto per tutti i k-mer che non sono stati ancora inclusi in un unipath interval.

Per ogni unipath interval viene preso il primo numero di k-mer; esso viene cercato nel database e, a partire da quest'ultimo, si determina, se presente, un suo possibile predecessore. Se ne esiste esattamente uno, l'unipath interval del primo numero viene collegato all'intervallo alla sua sinistra. Procedendo iterativamente, viene formato un unipath, che consiste quindi in un gruppo di k-mer contigui.

Isolamento dei seed unipath Tra gli unipath creati, è possibile isolare dei *seed unipath* attorno ai quali poi eseguire l'assembly. Preso l'insieme di tutti gli unipath, iterativamente si rimuovono alcuni di essi; quindi, preso un certo unipath u , si isolano quelli adiacenti a destra e a sinistra rispetto a u . Tra i due vicini di u viene dedotta la distanza; se essa è minore di una certa soglia, u può essere rimosso dall'insieme. Si procede in questo modo

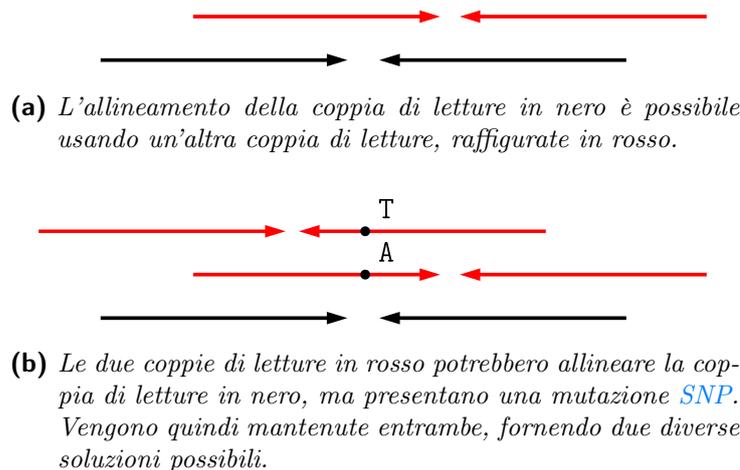


Figura 2.2: Esempi grafici del processo di gap filling.

per tutti gli unipath presenti finché continua a essere possibile la rimozione. Gli unipath rimasti costituiranno un seed unipath.

Assembly locale e globale Attorno ai seed unipath viene quindi assemblato il vicinato (*neighborhood*), cioè le regioni di 10 kb che precedono e seguono il seme. Per farlo, vengono creati due gruppi di sequenze, il primo (*primary read cloud*) contenente letture la cui posizione reale è vicina al seme, mentre il secondo (*secondary read cloud*) contiene brevi frammenti la cui sequenza può essere assemblata con le letture del primo gruppo. Il vicinato dei semi viene quindi assemblato utilizzando i due gruppi di letture, formando un *sequence graph* dell'assembly locale, cioè attorno al seed unipath corrispondente.

I vari assembly locali vengono fatti in parallelo e sono poi uniti per formare un sequence graph unico, relativo cioè all'assembly globale.

2.1.2 Gestione di genomi di grandi dimensioni

I predecessori di ALLPATHS-LG ottengono risultati promettenti solo per genomi di piccole dimensioni. Per gestire genomi di dimensioni più grandi, come quelli dei mammiferi, sono state fatte delle modifiche notevoli [21].

Il programma cerca di comprimere le ripetizioni in modo da favorirne l'allineamento. Se una sequenza ripetitiva è presente in due letture separate, il programma utilizza un'altra coppia di shotgun per poter allineare le prime due (*gap filling*). Le letture vengono unite se un'altra coppia fornisce una sovrapposizione, e viceversa. Il metodo può essere utilizzato anche se sono presenti **polimorfismi a singolo nucleotide (SNP)**, che forniscono più soluzioni di assemblamento. La figura 2.2 tratta da [21], mostra come avviene la gestione delle sequenze ripetitive nei due diversi casi.

La nuova versione del programma, inoltre, migliora la correzione degli errori di lettura: dato un k-mer, vengono analizzate tutte le letture che lo contengono; nel caso in cui una

singola lettura differisca dalla maggior parte delle altre, essa viene corretta se non ci sono voti in conflitto, altrimenti non viene modificata.

Ulteriori miglioramenti sono stati applicati anche per la gestione di sequenze a bassa copertura: dato che in questi casi la sovrapposizione tra letture può essere breve, in tali zone è preferibile utilizzare k-mer di dimensioni minori cioè scegliere un valore di k più basso. Il programma permette di utilizzare un valore di $k > 15$, ma solo nelle regioni che verrebbero assegnate a uno spazio vuoto tra altre due sequenze.

2.1.3 Stima della dimensione del genoma

Il programma, pur avendo come scopo primario la costruzione dell'assembly a partire dalle letture del genoma, produce in output un valore stimato della sua dimensione. Per farlo, dato il grafico del k-mer profile, esso identifica l'ascissa del punto di flesso f_v che si trova tra il primo e il secondo picco, cioè tra quello dei k-mer derivanti da errori di sequenziamento e quello dei k-mer omozigoti, e l'ascissa f_p corrispondente al picco eterozigote [11].

La dimensione del genoma G viene quindi stimata considerando gli N k-mer con copertura C compresi tra f_v e $3f_p/2$, scartando quindi dalle letture i k-mer a frequenze basse o molto alte, tramite la formula $G = N/C$.

2.2 GCE

Il programma *GCE* permette di modellare la distribuzione dei k-mer presenti nei dati sequenziati e di stimare la dimensione del genoma, le sequenze ripetitive e il [rapporto di eterozigosi](#) [22]. Il programma è open-source, scritto in C/C++ e si basa sulla statistica bayesiana.

2.2.1 Algoritmo

Conteggio dei k-mer Il programma fa uso del k-mer profile delle letture disponibili per stimare le caratteristiche del genoma. Inizialmente deve essere determinato il valore ottimale di k , che deve risultare grande abbastanza da far apparire ogni k-mer quasi unico all'interno del genoma, ma allo stesso tempo il più piccolo possibile per evitare un uso eccessivo di memoria durante l'estrazione dei k-mer. Per questo di solito si preferisce un valore k tale che $4^k > 5 \times G$, con G dimensione del genoma. In base al valore k scelto, può essere calcolato il k-mer profile. Il programma utilizza, oltre alle letture del genoma da sequenziare, un genoma di riferimento (*reference genome*).

Modello ideale Inizialmente, ipotizzando che entrambi i genomi siano ideali, cioè che il reference genome sia una sequenza casuale, senza eterozigosi o ripetizioni, e che le letture del genoma da sequenziare siano di uguale lunghezza e senza errori di sequenziamento,

la distribuzione delle frequenze dei k-mer seguirà una [distribuzione di Poisson](#) [17]. Considerando i valori n_{base} e n_{k-mer} il numero totale di basi e k-mer presenti nelle letture, e rispettivamente $c_{base} = n_{base}/G$ e $c_{k-mer} = n_{k-mer}/G$ le coperture per le basi e i k-mer nelle letture da sequenziare, una lettura di L basi genera $L - k + 1$ k-mer, quindi $n_{k-mer}/n_{base} = (L - k + 1)/L$. Si ricavano di conseguenza l'equazione 2.1 per il calcolo della copertura delle basi, e l'equazione 2.2 per la lunghezza del genoma.

$$c_{base} = c_{k-mer} \times L / (L - k + 1); \quad (2.1)$$

$$G = n_{k-mer} / c_{k-mer} = n_{base} / c_{base}. \quad (2.2)$$

Utilizzando i simboli $n = n_{k-mer}$ e $c = c_{k-mer}$, l'algoritmo deve determinare entrambi i parametri, per poi calcolare la lunghezza del genoma tramite l'equazione 2.2.

Il valore n può essere dedotto facilmente dal k-mer profile, ricavato il precedenza, come numero totale dei k-mer trovati. Per il valore della copertura c invece, vengono utilizzate le formule delle [distribuzioni di Poisson 2.3 e 2.4](#), che descrivono rispettivamente il k-mer profile, il quale rappresenta il numero di specie di k-mer ($P_{Kspecies}$), e il prodotto tra i valori del numero di specie di k-mer e la copertura, che rappresenta il numero di individui di k-mer ($P_{Kindividuals}$); x si riferisce alla copertura del k-mer profile.

$$P_{Kspecies}(x) = \frac{c^x}{x!} e^{-c}; \quad (2.3)$$

$$P_{Kindividuals}(x) = x P_{Kspecies}(x) / c = P_{Kspecies}(x - 1). \quad (2.4)$$

Dalle curve delle distribuzioni, è possibile stimare il valore di c tramite una delle due equazioni 2.5 o 2.6.

$$c = \frac{P_{Kspecies}(x + 1)}{P_{Kspecies}(x)} (x + 1); \quad (2.5)$$

$$c = \frac{P_{Kindividuals}(x + 1)}{P_{Kindividuals}(x)} x. \quad (2.6)$$

Gestione di sequenze ripetitive Nella maggior parte dei casi un genoma reale contiene sequenze ripetitive, le quali aumentano la difficoltà del sequenziamento. Viene quindi utilizzato come riferimento un genoma già sequenziato, nel quale i k-mer vengono classificati in base alla frequenza genomica i (*genomic frequency*), che rappresenta la frequenza dei k-mer nel genoma di riferimento. Per ciascuna classe i vengono calcolati il rapporto delle specie di k-mer ($a_i = n_{i,genomic,Kspecies} / n_{genomic,Kspecies}$) e il rapporto di k-mer individuali ($b_i = n_{i,genomic,Kindividuals} / n_{genomic,Kindividuals}$). Le ripetizioni causano più picchi nel k-mer profile, che possono essere rappresentati come composizione di [distribuzioni di Poisson](#). Le formule delle curve diventano quindi somma di [distribuzioni di Poisson](#),

descritte dalle formule 2.7 e 2.8.

$$P_{K_{species}}(x) = \sum_{i=1}^m a_i \times P_{K_{species},i}(x); \quad (2.7)$$

$$P_{K_{individuals}}(x) = \sum_{i=1}^m b_i \times P_{K_{individuals},i}(x). \quad (2.8)$$

La stima del parametro a_i viene fatta iterativamente basandosi su un modello bayesiano, utilizzando la formula 2.9 in cui il termine v_j rappresenta la probabilità della specie di k-mer con copertura j , mentre l'espressione $P(j|i)$ rappresenta la probabilità che una specie di k-mer con frequenza i sia trovata j volte nelle letture da sequenziare. Per il parametro c invece è stata ricavata la formula 2.10, che permette di stimare la dimensione del genoma tramite la formula $G = n_{k-mer}/c$, in cui il termine n_{k-mer} rappresenta il numero totale di k-mer.

$$a_{i,t+1} = \sum_{j=0}^w \frac{a_{i,t} P(j|i)}{\sum_{i=1}^m a_{i,t} P(j|i)} \times v_j; \quad (2.9)$$

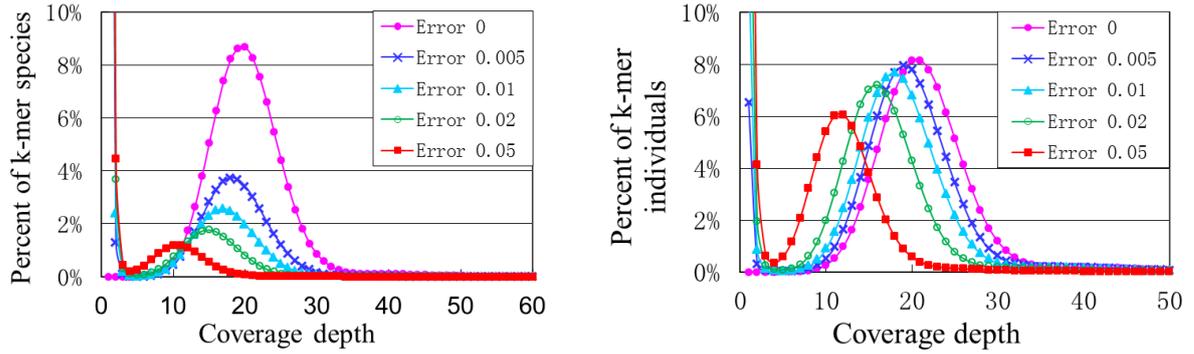
$$c_{t+1} = \frac{P_{K_{species},1,t+1}(x+1)}{P_{K_{species},1,t+1}(x)} \times (x+1). \quad (2.10)$$

Sequenziamento di genomi eterozigoti Sequenziando genomi diploidi reali, risulta utile stimare il [rapporto di eterozigosi](#), dovuto a mutazioni [SNP](#). Posto che i k-mer eterozigoti si distribuiscano in modo sparso nel genoma, saranno presenti $2 \times k$ k-mer eterozigoti attorno a ciascun sito SNP, che daranno origine a un nuovo picco nella curva di distribuzione dei k-mer.

2.2.2 Errori di sequenziamento e coverage bias

La gestione degli errori di sequenziamento viene fatta considerando i valori stimati di c e a_i . Il grafico della distribuzione dei k-mer in questi casi, infatti, presenta un grande picco iniziale, che riduce la dimensione degli altri picchi e li sposta verso sinistra, come mostrato dalla figura 2.3 a fronte tratta da [22]. Tramite i suddetti valori che rappresentano rispettivamente la copertura e il rapporto delle specie di k-mer, è possibile ricostruire la curva della distribuzione e confrontarla con la curva reale, in modo da determinare il numero di k-mer corretti e rimuovere gli errori di sequenziamento.

I *coverage bias*, che corrispondono a deviazioni rispetto alla distribuzione uniforme delle letture [23], sono invece più difficili da gestire, poiché appiattiscono le curve rendendo più difficile l'individuazione dei picchi e la stima di c . Anche se in presenza di tali errori i k-mer non vengono campionati in maniera equiprobabile, lo spettro di campionamento risulta continuo. Il nuovo modo di rappresentare le distribuzioni di specie e di individui di k-mer diventa quindi un modello continuo, che viene approssimato con un modello discreto



(a) Grafici delle specie di k -mer, al variare della percentuale di errore. (b) Grafici degli individui di k -mer, al variare della percentuale di errore.

Figura 2.3: Effetto degli errori di sequenziamento sui grafici delle specie e degli individui di k -mer. Il picco dei k -mer omozigoti viene appiattito e spostato a sinistra all'aumentare della percentuale di errore.

a livelli non uniformi a_k , nel quale la copertura c scelta corrisponde al maggior valore a_k , mentre il parametro a_i viene stimato come somma dei parametri a_k corrispondenti a ciascun picco. Una volta trovati i valori c e a_i , viene stimata la dimensione del genoma tramite la formula 2.2 a pagina 11.

2.3 GenomeScope

Il progetto open source *GenomeScope* cerca sia di stimare le caratteristiche del genoma completo, come ad esempio la sua lunghezza o il [rapporto di eterozigosi](#), sia di determinare le proprietà delle letture di DNA che prende in input, come la copertura (*read coverage*) o l'error rate [18]. Il programma per determinare tali caratteristiche utilizza il k -mer profile del genoma preso in esame, descritto nella sezione 1.3.1 a pagina 4.

2.3.1 Algoritmo

Il programma effettua una regressione non lineare dei dati iniziali, generando un profilo che cerca di approssimare il k -mer profile reale. Prendendo in input le letture del genoma che si vuole studiare, esso crea un modello che approssima il più possibile il k -mer profile. La funzione $f(X)$ scelta per l'interpolazione delle frequenze dei k -mer trovati è la somma di quattro [distribuzioni binomiali negative](#) $Y \sim \mathcal{NB}(X; p, n)$, per rappresentare rispettivamente i k -mer eterozigoti trovati nel genoma diploide una volta (unici) o tre volte (duplicati), e i k -mer omozigoti di cui si trovano due occorrenze (unici) o trovati quattro volte (duplicati). La funzione $f(X)$ è descritta dall'equazione 2.11 nella pagina [successiva](#), in cui G rappresenta un coefficiente di scala legato alla dimensione del genoma,

λ e ρ sono rispettivamente la media e la varianza della distribuzione.

$$f(X) = G \times (\alpha \mathcal{NB}(X; \lambda, \lambda/\rho) + \beta \mathcal{NB}(X; 2\lambda, 2\lambda/\rho) + \gamma \mathcal{NB}(X; 3\lambda, 3\lambda/\rho) + \delta \mathcal{NB}(X; 4\lambda, 4\lambda/\rho)). \quad (2.11)$$

I coefficienti α, β, γ e δ dipendono dai parametri r e d , che rappresentano rispettivamente il [rapporto di eterozigosi](#), cioè la percentuale di basi che sono specifiche a uno o due cromosomi omologhi, e la percentuale del genoma che è presente in due copie.

Lo scopo del programma è quindi determinare i coefficienti r, d, λ e ρ , oltre alla dimensione totale del genoma G . La funzione scelta $f(X)$, tramite cui poi può essere calcolata la dimensione del genoma, è quella che restituisce la minore somma dei quadrati degli errori residui (*Residual Sum of Square Error - RSSE*), che cioè minimizzi la somma tra i quadrati degli errori tra i valori osservati e quelli stimati, come descritto dall'equazione 2.12. Per dedurre i valori dei coefficienti, viene utilizzata la funzione `nls` del linguaggio di programmazione *R*, che compie il [fitting](#) dei dati alla funzione obiettivo.

$$RSSE = \sum_{x=E}^{+\infty} (kmer_{obs}[x] - kmer_{pred}[x])^2. \quad (2.12)$$

Al termine, il programma mostra all'utente i dati relativi al genoma trovati, come il [rapporto di eterozigosi](#), la media e la varianza della distribuzione, l'indice RSSE, che rappresenta la percentuale di k-mer non considerati dal modello, e la dimensione stimata del genoma.

La figura 2.4 [nella pagina successiva](#) mostra un confronto tra il k-mer profile reale e il modello costruito dal programma.

2.3.2 Gestione degli errori di sequenziamento

Eventuali errori di sequenziamento, ad esempio dovuti a duplicazioni con PCR o a sequenze contaminate, sono determinati solo empiricamente: dopo varie iterazioni del software in cui viene abbassata la soglia di copertura richiesta, i k-mer che non riescono ad essere rappresentati dal modello vengono identificati come errori di sequenziamento.

La stima dei k-mer sequenziati in modo errato è importante perché viene utilizzata per determinare la percentuale di basi errate nelle letture: una singola base inesatta infatti può dar luogo fino a k k-mer errati, aumentando notevolmente il numero di errori. GenomeScope permette una percentuale e di basi errate in ciascun k-mer. Tale valore è calcolato con un [fitting](#) dei k-mer errati a una [distribuzione binomiale](#) tramite la funzione `uniroot` del linguaggio di programmazione *R*. Questo metodo permette al programma di non dover assumere che la distribuzione degli errori di sequenziamento abbia una particolare forma, né di utilizzare un valore di soglia.

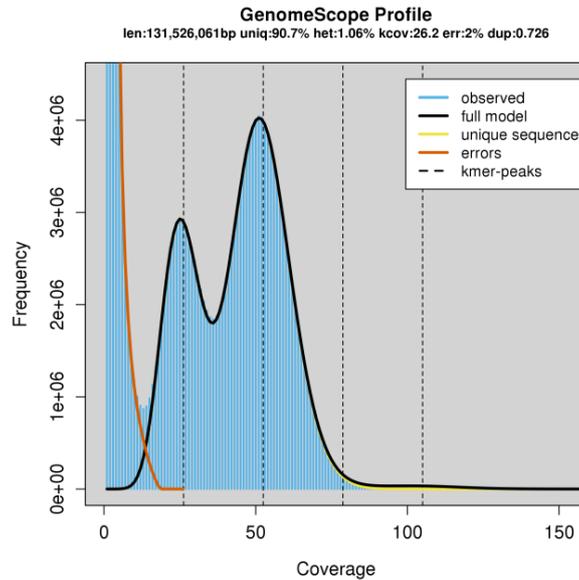


Figura 2.4: Modello del k-mer profile generato dal programma. L'istogramma in azzurro rappresenta i dati del k-mer profile reale, la curva nera il modello stimato e quella color arancione gli errori di sequenziamento stimati.

2.4 findGSE

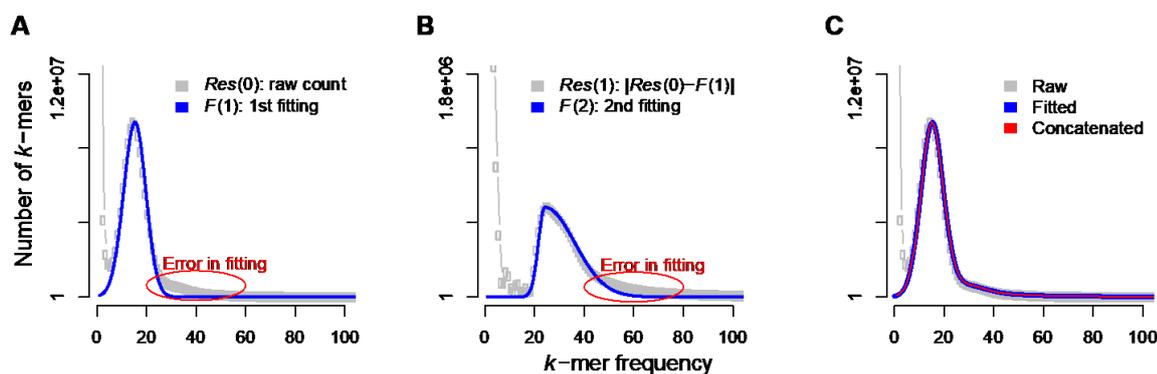
Il programma *findGSE* [11] ha come obiettivo principale la stima della lunghezza del genoma. Utilizzando le frequenze dei k-mer trovati nelle letture a disposizione, il programma compie una regressione non lineare dei dati utilizzando come funzione una [distribuzione normale asimmetrica](#) (*skew normal distribution* [24, 25]).

2.4.1 Algoritmo

Nel programma viene assunto che le frequenze dei k-mer possano essere approssimate da una [distribuzione normale asimmetrica](#) $Y \sim \mathcal{SN}(\xi, \omega^2, \alpha)$. Presa in input la distribuzione delle frequenze dei k-mer (k-mer profile), l'algoritmo effettua la regressione determinando i quattro parametri che descrivono una [distribuzione normale asimmetrica](#), la media ξ , la deviazione standard ω , l'asimmetria α e un fattore di scala s . Ad ogni iterazione, il programma cerca di minimizzare l'errore tra i dati di input e la funzione stimata, in modo da approssimare il più possibile il k-mer profile reale. La figura 2.5 nella [pagina seguente](#) mostra come viene effettuato iterativamente il [fitting](#) del k-mer profile utilizzando come funzione la suddetta distribuzione.

Dato un genoma aploide con G basi, il numero di k-mer possibili sarà $G - k + 1$. Ponendo C la copertura media dei k-mer, in modo che in media ogni k-mer sia trovato in C letture diverse, e N il numero di k-mer trovati nelle letture, la quantità di k-mer presenti nel genoma è descritta dall'equazione 2.13.

$$N = C \times (G - K + 1). \quad (2.13)$$



(a) Primo *fitting* dei dati dalle letture di input. (b) Secondo *fitting*, utilizzando anche i k-mer residui dalla precedente operazione. (c) *fitting* finale.

Figura 2.5: Esempio di *fitting* iterativo del k-mer profile.

Posta la dimensione del genoma molto maggiore del numero di basi utilizzate $G \gg k$, l'equazione 2.14 approssima la dimensione totale del genoma in analisi.

$$G \approx N/C. \quad (2.14)$$

A partire sia dal profilo reale che dal modello stimato, il programma calcola quindi il numero totale di k-mer trovati N e la copertura media dei k-mer C , per poi calcolare la dimensione del genoma attraverso l'equazione 2.14.

2.4.2 Analisi di genomi reali

Utilizzando letture reali, i dati devono essere precedentemente processati per minimizzare gli errori di stima del genoma. Per questo è consigliabile ridurre prima le letture con il programma *Skewer* [26], ed eliminare quelle con lunghezza minore di 33 bp. Per ottenere risultati migliori, inoltre, è consigliabile rimuovere le letture doppie dovute alla duplicazione PCR tramite il programma *FastUniq* [27], e le sequenze di simili a DNA dei mitocondri o dei cloroplasti con i programmi *BWA* [28] e *SAMtools* [29].

2.5 MGSE

Il programma *Mapping-based Genome Size Estimation (MGSE)* stima la dimensione del genoma attraverso il calcolo della copertura media a partire dall'assembly ad alta contiguità delle letture [12]. Lo script è open-source e scritto in Python, e processa le informazioni sulla copertura delle letture di input restituendo la dimensione stimata del genoma.

2.5.1 Algoritmo

Posto che le letture di input siano distribuite equamente sull'intera sequenza del genoma, il programma ne stima la dimensione calcolandone la copertura media. Se infatti sono noti il numero L di basi sequenziate e la copertura C in una certa posizione, la lunghezza totale N del genoma sarà pari a $N = L/C$.

Dato che le letture non forniscono la stessa copertura in tutte le regioni, ad esempio a causa di sequenze ripetitive, per una stima reale della dimensione del genoma è importante un calcolo corretto della copertura media. Per questo, il programma permette all'utente di inserire una lista di regioni di riferimento usate per il calcolo della media e della mediana della copertura. A questo scopo può essere utile il programma *Benchmarking Universal Single Copy Orthologs (BUSCO)* [30], che identifica un gruppo di sequenze *bona fidae*, cioè che sono a singola copia e che hanno solo una sola corrispondenza all'interno del genoma [31], le quali possono risultare adatte al calcolo della copertura media.

Il programma per poter effettuare la stima della dimensione del genoma riceve in input un file contenente la versione binaria compressa di sequenze allineate (file *bam* [29]), oppure un file contenente informazioni sulla copertura per ogni porzione della sequenza da sequenziare (file *cov* [32]). Dopo aver calcolato la media e la mediana della copertura, il programma produce in output le dimensioni del genoma stimate per entrambi i valori.

Capitolo 3

Valutazione dei metodi

In questo capitolo viene eseguita l'analisi di ciascun metodo, a partire dalla propria complessità e dalla precisione dimostrata. Vengono poi presentate due serie di confronti tra alcuni dei programmi analizzati, per cercare di determinare quale sia il migliore.

3.1 Complessità e performance

In questo paragrafo vengono analizzate per ciascun metodo la complessità, i requisiti richiesti per l'esecuzione e l'accuratezza dimostrata nei test eseguiti da ciascun gruppo di ricerca. Si noti che la misurazione fatta del numero di righe per ciascun programma considera solo il codice sorgente, escludendo gli spazi bianchi e i commenti.

3.1.1 ALLPATHS-LG

Dato che il programma ha come obiettivo principale l'assembly delle letture di input, esso presenta una complessità elevata (984 file, 209677 righe di codice). Inoltre, per far sì che il programma possa gestire genomi di grandi dimensioni, si rende necessario l'utilizzo di dispositivi server commerciali, in modo da svolgere l'elaborazione mantenendo in memoria grandi quantità di dati. Per esempio, l'assembly del genoma di un mammifero può essere fatto in alcune settimane utilizzando un server *Dell R815* con 48 processori e 512 GB di memoria RAM disponibile [21]. Altri metodi di sequenziamento tramite de novo assembly, come ad esempio il programma *SOAPdenovo* [33], riescono a completare l'assembly in un tempo minore a parità di dati in input e di risorse computazionali disponibili. Nonostante ciò, utilizzando letture di dimensioni minori ma a maggiore copertura, ALLPATHS-LG produce assembly di qualità più elevata rispetto agli altri programmi.

Il metodo, essendo molto più complesso e richiedendo maggiori capacità computazionali rispetto agli altri programmi analizzati, può risultare sovradimensionato per la sola stima della dimensione del genoma. Il valore stimato dal software, tuttavia, può risultare utile nel confronto con gli altri metodi.

3.1.2 GCE

Il programma presenta una complessità media (1150 righe di codice) e permette di stimare la dimensione del genoma tramite uno dei modelli descritti nel paragrafo 2.2 a pagina 10 in base al genoma che si deve analizzare, come ad esempio i modelli omozigote o eterozigote, continuo o discreto.

Considerando come genomi di riferimento altri progetti di sequenziamento tramite *de novo assembly*, come quello dell’*E. coli*-O104:H4 [34], della formica tagliafoglie [35], della patata [36] o del panda [37], le dimensioni stimate dal programma utilizzando dati simulati dimostrano un’elevata accuratezza.

Nella stima di genomi reali invece, dovendo eseguire un filtraggio e una correzione degli errori sui dati originali per diminuire la complessità della computazione e, quindi, il consumo di memoria, si registra una diminuzione della precisione [22]. Il valore dedotto dal programma, comunque, può risultare utile per determinare la strategia di sequenziamento di un genoma, e per guidare lo sviluppo di algoritmi per l’*assembly*.

3.1.3 GenomeScope

GenomeScope risulta un programma veloce, dal momento che processa i dati in meno di un minuto con consumi non eccessivi di memoria RAM, e mostrando una complessità media (501 righe di codice). Il programma, inoltre, restituisce in output un file testuale contenente le proprietà del genoma analizzato, oltre a immagini in alta qualità del modello costruito [18].

Nei test eseguiti dal gruppo di ricerca, il programma ha dimostrato un’alta affidabilità: sia nella valutazione di dati simulati di vari organismi variando il [rapporto di eterozigosi](#), il rapporto di duplicazione, la copertura e il numero di errori di sequenziamento, sia sulla stima di genomi reali combinati sinteticamente, il metodo riscontra una buona approssimazione delle caratteristiche delle sequenze. Nell’analisi di dati reali invece, si riscontra una precisione del 99,7% rispetto alle stime fatte con altri metodi, come genomi di riferimento o *flow cytometry*.

Per la maggior parte dei casi viene consigliato l’utilizzo di *k-mer* con valore $k = 21$, che può essere aumentato per genomi con dimensioni molto elevate (dimensione aploide $\gg 10$ Gb) o che presentano un elevato numero di sequenze ripetitive. Una scarsa copertura (in genere, minore di $25\times$) o un elevato rapporto di errore di sequenziamento, inoltre, potrebbero impedire al modello di convergere con il *k-mer* profile reale. La flessibilità del programma nel numero di distribuzioni utilizzate per il [fitting](#), infine, potrebbe risultare adeguata anche per lo studio di genomi poliploidi, che al momento risultano difficilmente trattabili [11].

3.1.4 findGSE

Il programma findGSE, che esegue il [fitting](#) di una [distribuzione normale asimmetrica](#) al k-mer profile, presenta una complessità media (969 righe di codice) e permette la stima della dimensione del genoma utilizzando risorse modeste.

Nell'analisi di genomi simulati la stima avviene quasi perfettamente, anche applicando variazioni al rapporto di errore di sequenziamento, alla copertura o al livello di eterozigosi. Il valore della stima della dimensione del genoma di *Arabidopsis thaliana* generata dal programma si dimostra leggermente minore al valore trovato tramite il metodo *flow cytometry*, con il quale comunque condivide una forte correlazione [11].

Il programma è stato utilizzato anche per la stima della dimensione di genoma umano. Prendendo come riferimento la versione GRCh38.p9 sviluppata da *The Genome Reference Consortium*, il programma ha riscontrato una differenza di 41 Mb tra i genomi maschile e femminile, similmente alla differenza di 49 Mb tra i due sessi del genoma di riferimento.

3.1.5 MGSE

MGSE presenta una complessità minore rispetto agli altri programmi presentati (323 righe di codice), e raggiunge le prestazioni migliori con assembly ad alta contiguità [12]. Rispetto alla maggior parte degli altri metodi basati sui k-mer, esso richiede un genoma di riferimento per il calcolo della copertura media, come ad esempio le sequenze individuate dal software *BUSCO*.

Il programma, se l'assembly include tutte le regioni a singola copia e almeno un'occorrenza di quelle ripetitive, si dimostra affidabile. Esso permette inoltre di gestire le regioni contenenti DNA affetto da contaminazioni batteriche o fungine, nel caso in cui esse non siano incluse nelle sequenze di riferimento, procedendo con l'esclusione delle letture corrispondenti.

Invece, nel caso in cui solo alcune sequenze di input siano state duplicate tramite PCR, risulta impossibile distinguere tali copie dalle sequenze presenti nel genoma reale in più di una regione. Il programma comunque prevede una certa tolleranza se la duplicazione è stata fatta equamente su tutte le letture di input.

3.2 Confronti tra i metodi

In questo paragrafo vengono mostrate due serie di confronti tra alcuni dei metodi presentati. In entrambi i casi a ciascun metodo vengono forniti gli stessi dati di input che, indipendentemente l'uno dall'altro, vengono processati per stimare la dimensione del genoma sotto specifiche condizioni.

3.2.1 Metodi findGSE, ALLPATHS-LG, GCE e GenomeScope

I dati e le figure della seguente analisi sono state ricavate da [11], che esegue un confronto tra i metodi findGSE, ALLPATHS-LG, GCE e GenomeScope. Si noti che nelle figure sono inclusi due metodi non trattati in questo documento (*naive* ed *egs*), i cui risultati non saranno pertanto analizzati. Il genoma di riferimento utilizzato per l'analisi di letture simulate di *A. thaliana* è una sequenza di circa 119 Mb.

Variazione della copertura Inizialmente vengono utilizzate letture simulate di un genoma omozigote con copertura minima $50\times$ e con l'1% di basi errate nel sequenziamento. Tutti i programmi si dimostrano precisi e stabili, stimando la dimensione del genoma tra 118 e 123 Mb a fronte di un valore reale di 119 Mb, come mostrato dalla figura 3.1. Riducendo gradualmente la copertura da $50\times$ a $10\times$, tutti i metodi riportano una stima corretta fino a $30\times$. Nel caso di una bassa copertura ($10\times$), i metodi ALLPATHS-LG e GenomeScope falliscono nel dare una stima della dimensione del genoma.

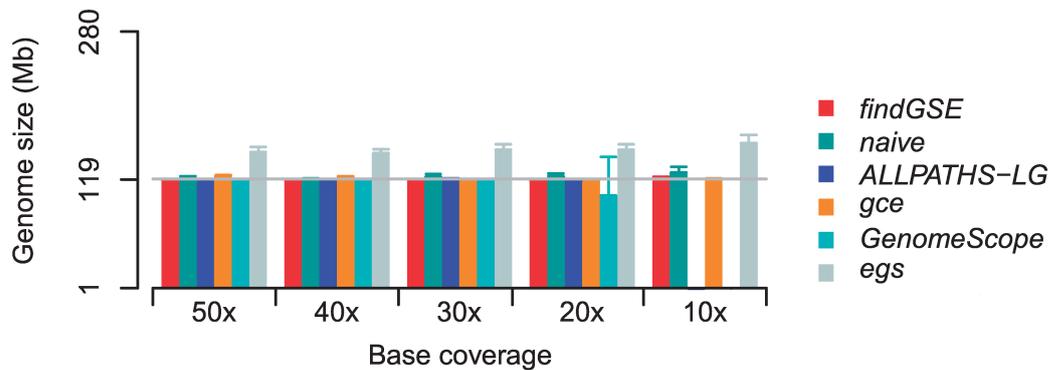


Figura 3.1: Stima della dimensione del genoma, con valore di copertura minima variabile.

Errori di sequenziamento Modificando invece il rapporto di errore di sequenziamento, i programmi generano risultati quasi sempre compatibili con la grandezza reale, come mostrato dalla figura 3.2 a fronte. Si noti che con un rapporto di errore elevato e non realistico (5%), mentre gli altri metodi mantengono la stessa precisione ALLPATHS-LG restituisce un valore instabile.

Influenza dell'eterozigosi Per la valutazione dell'impatto dell'eterozigosi, è stata considerata una sequenza con copertura $30\times$, un rapporto di errore di sequenziamento dell'1% e un rapporto di mutazioni *SNP* da 0,1 a 5,0%. Per livelli bassi di eterozigosi (da 0,1 a 1,0%) tutti i metodi risultano precisi e concordi, come mostrato dalla figura 3.3.

Aumentando il rapporto di eterozigosi (da 2,0 a 5,0%), la discrepanza tra i valori predetti e quello reale aumenta considerevolmente: su un totale di 280 stime per metodo infatti, i programmi ALLPATHS-LG e GenomeScope hanno generato valori doppi rispetto alla grandezza reale rispettivamente in 150 e 29 previsioni. Le stime erranee dei due

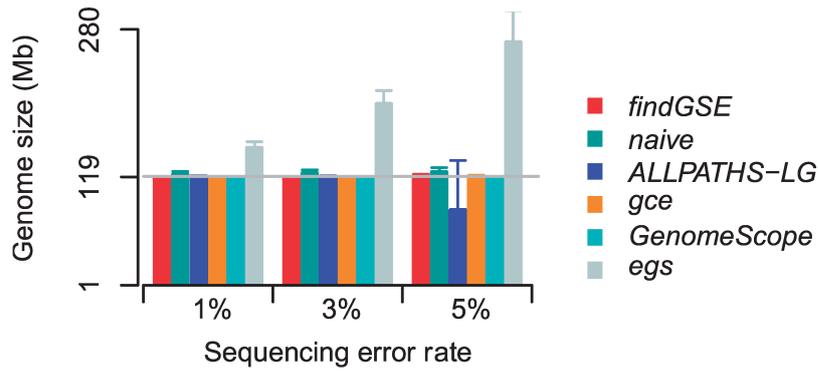


Figura 3.2: Stima della dimensione del genoma con tasso di errore di sequenziamento variabile.

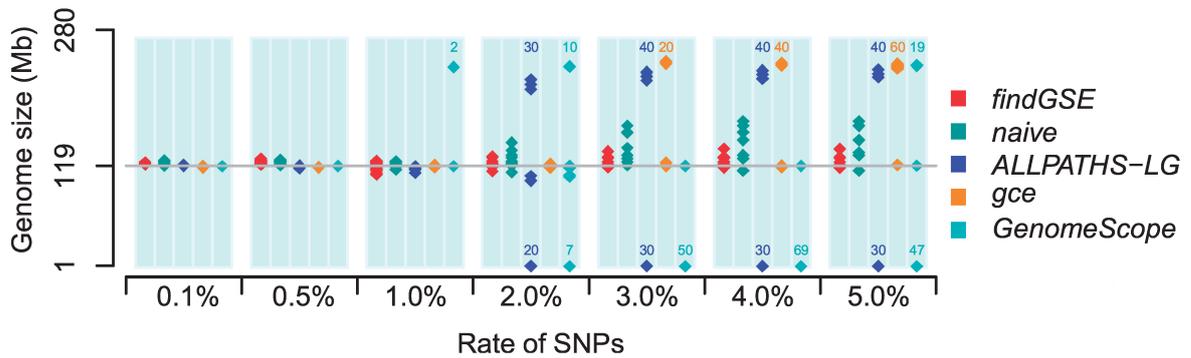


Figura 3.3: Stima della dimensione del genoma con livello di eterozigosi variabile e copertura base 30×. Si noti che i valori rappresentati indicano il numero di punti sovrapposti in una certa regione.

programmi sono dovute all’erronea selezione del picco eterozigote come se fosse quello omozigote, che porta a un calcolo non corretto della copertura e quindi al raddoppio della dimensione totale stimata. Gli stessi metodi hanno riportato valori nulli o estremamente bassi rispettivamente in 110 e 173 casi, generando quindi in totale valori erronei con probabilità del 92,9% per ALLPATHS-LG e del 72,1% per GenomeScope. Anche nel caso del metodo GCE, un alto tasso di eterozigosi fa generare al programma 120 valori errati su 280 totali, cioè nel 42,9% delle predizioni. Il metodo findGSE si dimostra poco influenzato dall’aumento del [rapporto di eterozigosi](#), che nel peggiore dei casi fa sovrastimare di poco la dimensione del genoma (125 ± 7 Mb).

Mantenendo elevato il tasso di errore di sequenziamento, l’aumento della copertura minima a 100× (figura 3.4) diminuisce il numero di stime non corrette per i metodi ALLPATHS-LG e GenomeScope, generando rispettivamente 60 (21,4%) e 110 (39,3%) valori errati. Il programma GCE invece, anche all’aumentare della copertura produce molti risultati errati, 210 su 280 totali (75%). Il metodo findGSE infine si dimostra molto accurato anche ad un alto [rapporto di eterozigosi](#), stimando una dimensione media di 121 ± 1 Mb. Il programma GenomeScope, tralasciando il problema di duplicazione della dimensione, si dimostra quello con i risultati più stabili tra i metodi analizzati.

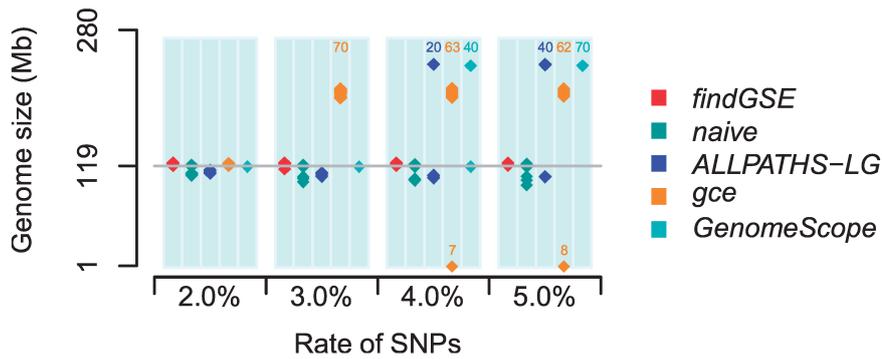


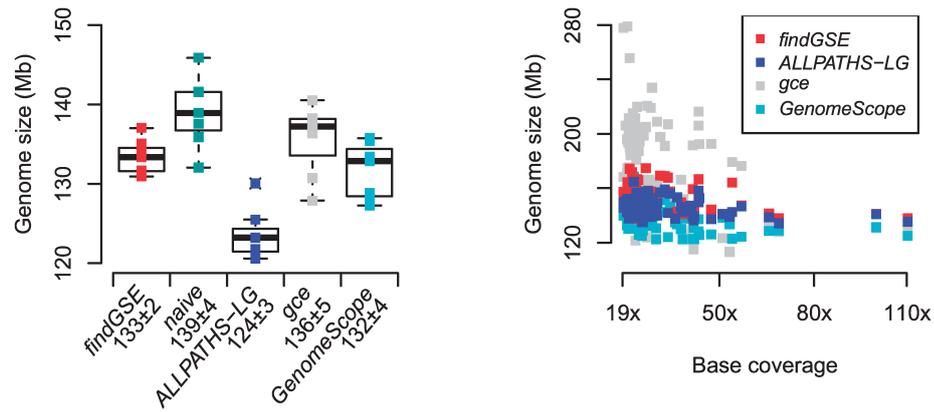
Figura 3.4: Stima della dimensione del genoma con livello di eterozigosi variabile e copertura base $100\times$. Si noti che i valori rappresentati indicano il numero di punti sovrapposti in una certa regione.

Stima di sequenze reali La riproducibilità dei metodi analizzati è stata verificata con la stima della dimensione di sette genomi reali. Le sequenze scelte per la verifica sono sequenziamenti di *Arabidopsis thaliana*, le cui differenze si assume non vadano a modificare considerevolmente la dimensione del genoma. La deviazione standard delle stime per ciascun metodo si presenta bassa, variando tra 2 Mb (findGSE) e 5 Mb (GCE), come mostra la figura 3.5a nella pagina successiva. Ciò può dimostrare la robustezza dei programmi rispetto ad eventuali variazioni nel sequenziamento del genoma, e l'alta riproducibilità che li caratterizza.

Per verificare l'indipendenza tra il valore stimato e la copertura di letture reali, i metodi sono stati utilizzati per stimare la dimensione del genoma di *A. thaliana* a partire da 89 sequenze con copertura base minima maggiore di $19\times$. Come mostra la figura 3.5b a fronte, le stime di ciascun programma risultano indipendenti da tale valore.

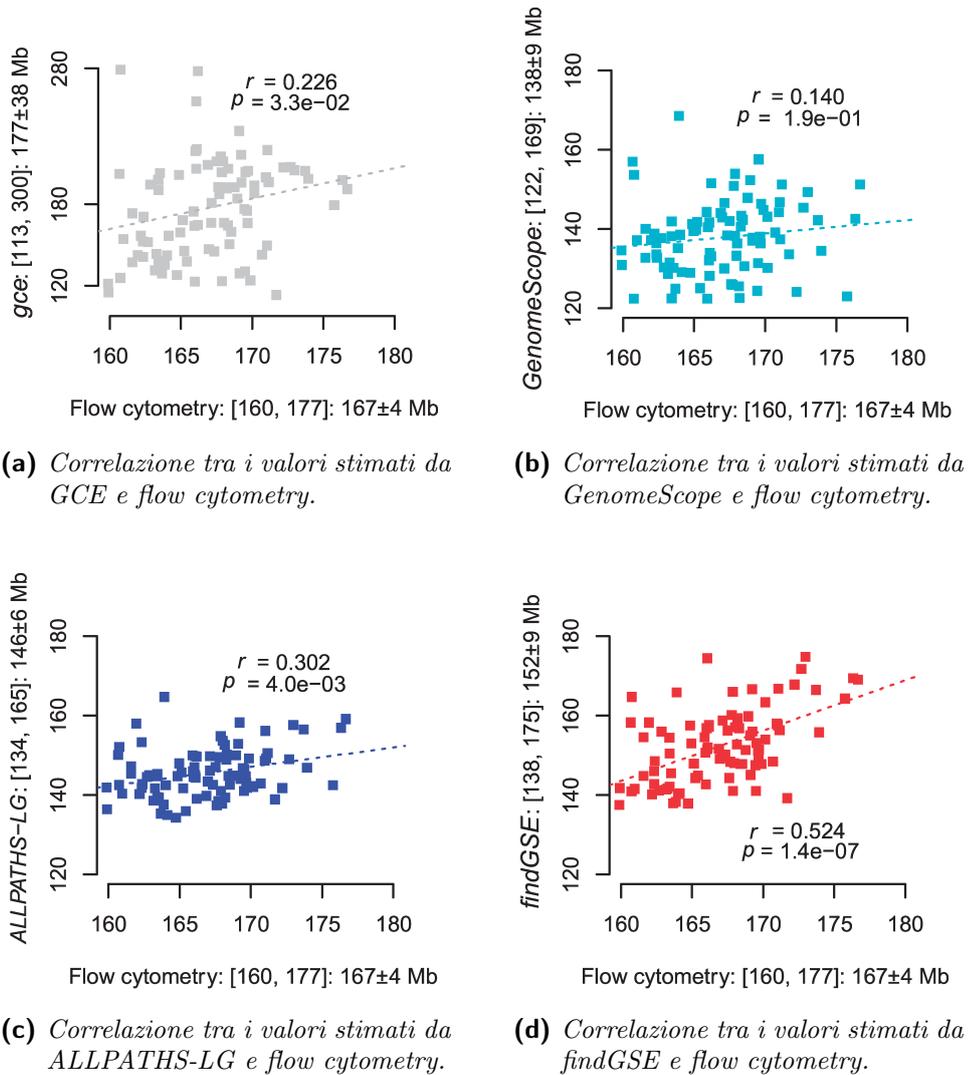
Confronto con flow cytometry Sebbene il metodo flow cytometry non possa restituire una dimensione certa del genoma, è considerato un metodo sperimentale che la possa almeno approssimare. Analizzando la dimensione di 89 sequenze di *A. thaliana*, la deviazione standard della stima fatta dai programmi varia tra 6 Mb e 9 Mb. Il metodo flow cytometry invece mostra una deviazione standard di 4 Mb, che suggerisce una maggior stabilità della predizione. Per ciascun programma è stato quindi misurato il [coefficiente di correlazione di Pearson](#) con il metodo flow cytometry (figura 3.6 nella pagina successiva), assumendo che un coefficiente più alto indichi una maggior capacità di rilevare variazioni reali della dimensione del genoma. La correlazione tra findGSE e il suddetto metodo risulta la più alta ($r = 0,524$).

I metodi findGSE, ALLPATHS-LG e GenomeScope stimano in media simili valori di dimensione (rispettivamente 152, 146 e 138 Mb), mentre GCE produce un risultato inespiegabilmente grande (177 Mb). Il valore medio stimato tramite flow cytometry invece è di 167 Mb, anche se tale metodo tende a sovrastimarne la dimensione.



(a) Valori stimati, media e deviazione standard di ciascun metodo. (b) Valori stimati da ciascun metodo in relazione alla copertura base.

Figura 3.5: Confronto tra i metodi analizzati nella stima della dimensione di *A. thaliana*.



(a) Correlazione tra i valori stimati da GCE e flow cytometry. (b) Correlazione tra i valori stimati da GenomeScope e flow cytometry. (c) Correlazione tra i valori stimati da ALLPATHS-LG e flow cytometry. (d) Correlazione tra i valori stimati da findGSE e flow cytometry.

Figura 3.6: Correlazione tra i metodi analizzati e flow cytometry. In ciascuna figura sono raffigurate anche la regressione lineare e il coefficiente di correlazione di Pearson r .

3.2.2 Metodi MGSE, GenomeScope, GCE e findGSE

I dati e le figure della seguente analisi sono state ricavate da [12], che esegue un confronto tra i metodi MGSE, GenomeScope, GCE e findGSE.

Dopo una comparazione iniziale processando due diversi ecotipi di *Arabidopsis thaliana*, i programmi vengono utilizzati per la stima della dimensione di genomi con grandezze e complessità maggiori, come le specie *Beta vulgaris*, *Brachypodium distachyon*, *Solanum lycopersicum*, *Zea mays* e *Vitis vinifera*.

Stima della dimensione di *A. thaliana* Inizialmente i programmi in esame vengono utilizzati per il calcolo della dimensione delle varianti Col-0 e Nd-1 di *Arabidopsis thaliana*. Per i metodi GenomeScope, GCE e findGSE la stima viene fatta utilizzando k-mer di taglia $k = 19, 21, 23$ e 25 , mentre per il programma MGSE vengono selezionate la media e la mediana di varie porzioni del genoma (selezione manuale dei geni a singola copia, sottosequenze contenenti i geni codificanti proteine, **esoni** dei gruppi precedenti, sequenze restituite dal software *BUSCO*).

La dimensione dell'assembly dell'ecotipo Col-0 di *A. thaliana* analizzato è di 120 Mb [38]. Molte delle stime fatte dai programmi generano risultati minori di tale valore, come mostra la figura 3.7. I metodi GenomeScope e MGSE (utilizzando le sequenze restituite da BUSCO) generano valori compatibili con la dimensione stimata; i metodi GCE e findGSE, invece, restituiscono rispettivamente valori in media più ridotti o elevati.

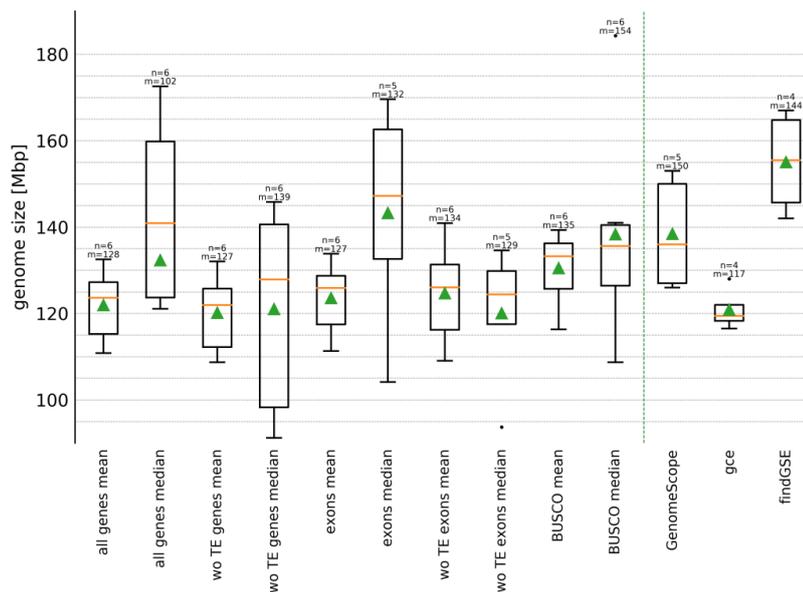


Figura 3.7: Dimensioni stimate dai metodi rispetto n campioni di ecotipo Col-0 di *A. thaliana*. Per il metodo MGSE vengono scelte la media o la mediana di porzioni diverse di genoma. La figura mostra per ciascun metodo il valor medio (in verde) e la mediana m (in giallo).

Nell'analisi dell'ecotipo Nd-1 mostrata dalla figura 3.8, il metodo MGSE (sequenze BUSCO) mostra una differenza evidente nella stima della dimensione utilizzando la me-

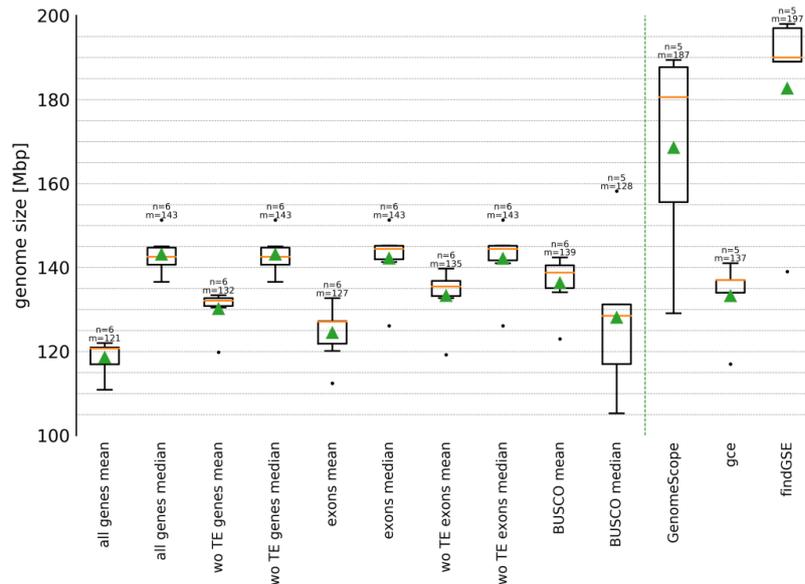


Figura 3.8: Dimensioni stimate dai metodi rispetto n campioni di *ecotipo* Nd-1 di *A. thaliana*. Per il metodo MGSE vengono scelte la media o la mediana di porzioni diverse di genoma. La figura mostra per ciascun metodo il valor medio (in verde) e la mediana m (in giallo).

dia o la mediana della copertura. Il valore stimato da MGSE attraverso la media delle sequenze BUSCO risulta il più affidabile, con una dimensione reale stimata di 138-140 Mb [39]. Mentre GCE restituisce un valore più plausibile, i metodi GenomeScope e findGSE generano una dimensione la cui correttezza risulta poco probabile.

Il confronto prosegue quindi con la stima della dimensione del genoma a partire da 1028 sequenze di *A. thaliana*. I programmi GCE, GenomeScope e findGSE nella maggior parte dei casi stimano valori compresi tra 120 e 200 Mb, mentre MGSE tra 120 e 160 Mb, come mostrato nella figura 3.9 nella pagina successiva. Si riporta comunque che tutti i metodi hanno generato alcuni valori errati, estremamente elevati o ridotti. In media, findGSE tende a stimare valori maggiori rispetto agli altri metodi basati sui k-mer.

Stima della dimensione di *Beta vulgaris* Per verificare l'accuratezza nella stima di genomi con maggiori dimensioni e complessità, i programmi vengono utilizzati per determinare la dimensione di *Beta vulgaris*. Dato che vengono utilizzate sequenze diverse come dati iniziali, sono attese differenze minime tra le letture di input, che potrebbero far variare i valori stimati dai metodi. Comunque, dato che al momento del test l'assembly più sviluppato del genoma di tale specie comprende 567 Mb della sequenza completa, i valori minori di tale soglia possono essere classificati come erranei.

Come mostrato dalla figura 3.10 nella pagina seguente, i metodi GenomeScope e GCE tendono a sottodimensionare la stima, restituendo valori in media di 500 Mb. Il programma findGSE, invece, genera risultati ad alta variabilità. Il metodo MGSE, utilizzando la media della copertura dei geni a singola copia o delle regioni BUSCO, sembra avere prestazioni migliori.

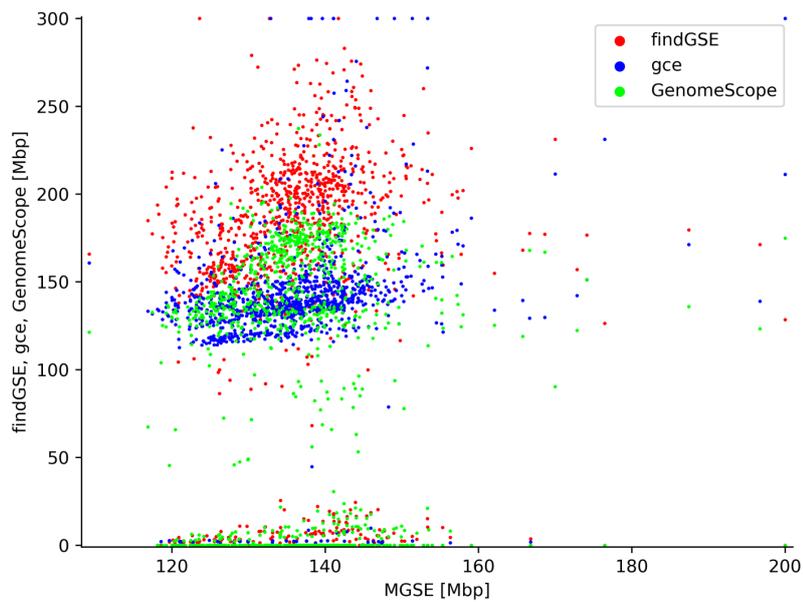


Figura 3.9: Stima della dimensione di 1028 sequenze di *A. thaliana*. I risultati dei metodi findGSE, GCE e GenomeScope vengono mostrati in funzione delle stime del metodo MGSE. Eventuali outlier vengono rappresentati sul bordo del grafico, permettendo una maggiore risoluzione alla nuvola centrale.

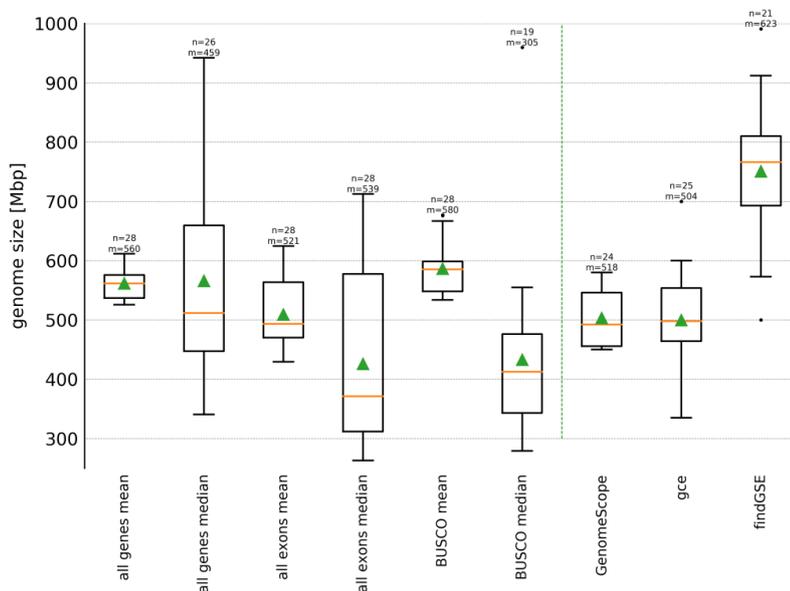


Figura 3.10: Dimensioni stimate dai metodi rispetto n campioni di *Beta vulgaris*. Per il metodo MGSE vengono scelte la media o la mediana di porzioni diverse di genoma. La figura mostra per ciascun metodo il valor medio (in verde) e la mediana m (in giallo).

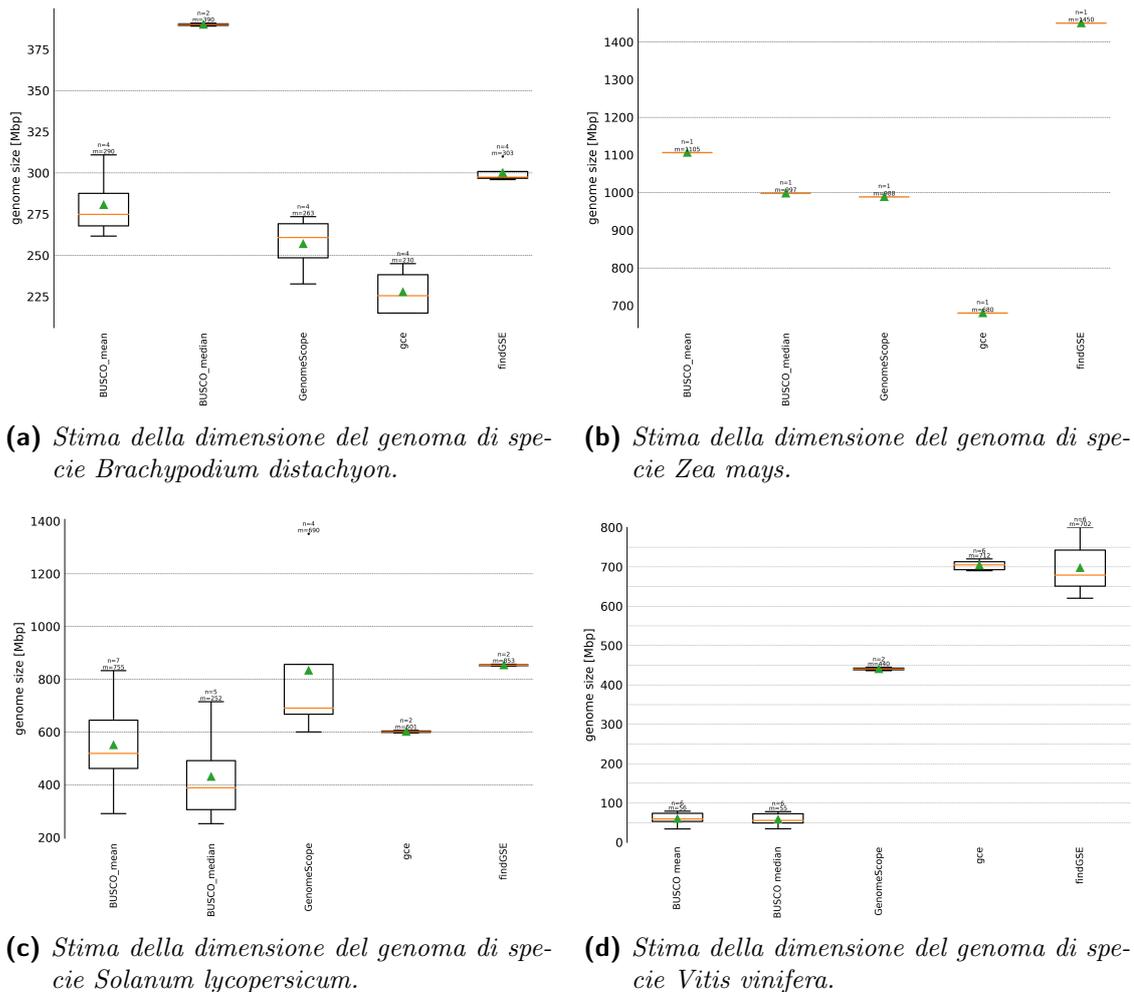


Figura 3.11: Dimensioni stimate dai metodi rispetto n campioni di genomi di varie specie. Per il metodo MGSE vengono scelte la media o la mediana delle regioni BUSCO. La figura mostra per ogni stima il valor medio (in verde) e la mediana m (in giallo).

Dimensione di genomi di altre specie L'analisi continua con la stima delle dimensioni di genomi di varie specie, ciascuna con caratteristiche diverse: il *Brachypodium distachyon* come modello di pianta erbacea, il pomodoro (*Solanum lycopersicum*) per il clade delle Asteridi, il mais (*Zea mays*) è stato incluso come specie monocotiledone ad alto contenuto di **trasposoni**, e la vite comune (*Vitis vinifera*) per la sua elevata eterozigosi.

I programmi, come mostrato dalla figura 3.11, restituiscono generalmente risultati appartenenti al medesimo intervallo. Nella stima della dimensione di *B. distachyon*, mentre MGSE, GenomeScope e GCE tendono a generare valori non corretti, findGSE stima una previsione ragionevole di 303 Mb, a fronte di una lunghezza stimata di circa 355 Mb [40]. La dimensione del genoma di *Z. mays* è ampiamente sottostimata da tutti i programmi, dato che la lunghezza stimata risulta di 2.4 Gb [41]. A differenza di MGSE e GCE, i programmi findGSE e GenomeScope stimano correttamente la lunghezza attesa di *S. lycopersicum* (950 Mb [42]). A causa dell'alta eterozigosi, MGSE restituisce un valore di soli 50 Mb per il genoma di *V. vinifera*, mentre il programma GenomeScope restituisce un valore plausibile, avendo una dimensione stimata di 475 Mb [43].

Conclusioni

Il presente elaborato si è posto l'obiettivo di analizzare e confrontare i principali metodi di stima della dimensione del genoma tramite k-mer.

Dopo una breve panoramica sui principali metodi di sequenziamento e di misurazione della dimensione del genoma attualmente disponibili, sono stati introdotti i k-mer e le loro principali proprietà. L'esposizione è quindi proseguita con l'analisi di ogni metodo, a partire dal punto di vista algoritmico e di gestione di casi particolari, fino alla valutazione della complessità e delle performance di ciascuno. Infine, l'analisi sperimentale dei risultati generati dai programmi ha permesso un confronto vero e proprio tra i metodi.

Dall'analisi degli approcci basati sui k-mer emerge che il metodo ALLPATHS-LG risulta sovradimensionato per la sola stima della dimensione e poco preciso in situazioni particolari, come nel caso di bassa copertura o alto tasso di errore di sequenziamento; il programma GCE, inoltre, tende solitamente a sottostimare la dimensione del genoma. GenomeScope, pur riscontrando problemi nel caso di una copertura scarsa nelle letture di input, sembra essere un metodo di stima discreto. Infine, findGSE sembra possedere l'approccio più promettente, avendo anche un'alta correlazione con il metodo flow cytometry.

Per quanto riguarda il programma MGSE, che presenta una complessità minore ma richiede informazioni più strutturate rispetto ai concorrenti, risulta un metodo efficiente nel caso di genomi di dimensioni ridotte. Infatti, la stima tramite la copertura ricavata dalle regioni BUSCO si dimostra precisa anche nei casi in cui altri metodi falliscono. Il programma, tuttavia, risulta poco accurato nell'analisi di genomi più particolari, come ad esempio un alto tasso di eterozigosi o un gran numero di [trasposoni](#).

Non esistendo attualmente un metodo affidabile che restituisca un valore di dimensione preciso per un genoma, risulta difficile determinare in modo assoluto il migliore tra i programmi trattati. In generale, si evince che alcuni metodi abbiano prestazioni migliori per certe tipologie di sequenze rispetto ad altri, ma che tutti riscontrino difficoltà nella trattazione di genomi di grandi dimensioni.

La stima della grandezza del genoma basata sui k-mer, assieme al metodo MGSE basato sull'individuazione di sequenze presenti in singola copia, comunque, anche grazie ai limitati requisiti richiesti, risultano essere approcci promettenti per lo sviluppo futuro di metodi più precisi ed efficienti.

Glossario

Clade Raggruppamento tassonomico costituito solo da un antenato comune e da tutti i suoi discendenti. [29](#)

Coefficiente di correlazione di Pearson Indice che esprime un'eventuale relazione di linearità tra due insiemi di dati, che consiste nel rapporto tra la covarianza di due variabili e il prodotto delle loro deviazioni standard. [24](#), [25](#)

Distribuzione normale asimmetrica Distribuzione continua di probabilità che generalizza la distribuzione normale, permettendo un indice di asimmetria diverso da zero. [15](#), [21](#)

Distribuzione binomiale Distribuzione discreta di probabilità di parametri n e p , che descrive il numero di successi in una sequenza di n esperimenti indipendenti, ciascuno con probabilità di successo p . [14](#)

Distribuzione binomiale negativa Distribuzione discreta di probabilità di parametri r e p , che descrive il numero di fallimenti in una sequenza di esperimenti indipendenti, ciascuno con probabilità di successo p , prima dell'occorrenza di un numero r di successi. [13](#)

Distribuzione di Poisson Distribuzione discreta di probabilità che esprime la probabilità di un numero dato di eventi che si verificano successivamente e in modo indipendente, fissato un numero medio di occorrenze. [4](#), [11](#)

Ecotipo Popolazione di organismi di una specie che si adatta geneticamente alle caratteristiche dell'habitat in cui vive. [26](#), [27](#)

Esoni Sequenza di DNA che viene copiata in filamenti di RNA messaggero, e che potrebbe codificare degli amminoacidi. Nel DNA gli esoni sono separati da introni, che invece non vengono copiati. [26](#)

Fitting Costruzione di una curva o di una funzione matematica che abbia la migliore corrispondenza ad una serie di punti. Essa può implicare l'interpolazione, nel caso si richieda un'esatta corrispondenza tra i punti e la funzione da ricercare. [14](#), [15](#), [16](#), [20](#), [21](#)

Locus Posizione specifica e fissata all'interno di un cromosoma in cui è posizionato un particolare gene o un'altra sequenza significativa. [5](#)

Rapporto di eterozigosi Rapporto tra il numero di siti eterozigoti e quello di siti omozigoti nel genoma di un individuo. [5](#), [10](#), [12](#), [13](#), [14](#), [20](#), [22](#), [23](#)

SNP - Polimorfismo a singolo nucleotide Variazione della sequenza di DNA a carico di un singolo nucleotide, presente almeno nell'1% degli individui di una popolazione. [9](#), [12](#), [22](#)

Trasposone Sequenze di DNA in grado di spostarsi da una posizione a un'altra del genoma. Il loro spostamento in genere può creare mutazioni, alterare l'identità genetica di una cellula o modificare la dimensione del genoma. [29](#), [31](#)

Bibliografia

- [1] J. Shendure e E. L. Aiden. «The expanding scope of DNA sequencing». In: *Nature Biotechnology* 30.11 (nov. 2012), pp. 1084–1094. ISSN: 1546-1696. DOI: [10.1038/nbt.2421](https://doi.org/10.1038/nbt.2421) (cit. a p. 1).
- [2] F. Sanger, S. Nicklen e A. R. Coulson. «DNA sequencing with chain-terminating inhibitors». In: *Proceedings of the National Academy of Sciences* 74.12 (gen. 1977), pp. 5463–5467. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463) (cit. a p. 1).
- [3] F. Sanger et al. «Nucleotide sequence of bacteriophage ϕ X174 DNA». In: *Nature* 265.5596 (feb. 1977), pp. 687–695. ISSN: 1476-4687. DOI: [10.1038/265687a0](https://doi.org/10.1038/265687a0) (cit. a p. 1).
- [4] A. M. Maxam e W. Gilbert. «A new method for sequencing DNA.» In: *Proceedings of the National Academy of Sciences* 74.2 (feb. 1977), pp. 560–564. DOI: [10.1073/pnas.74.2.560](https://doi.org/10.1073/pnas.74.2.560) (cit. a p. 1).
- [5] J. Shendure et al. «DNA sequencing at 40: past, present and future». In: *Nature* 550.7676 (ott. 2017), pp. 345–353. DOI: [10.1038/nature24286](https://doi.org/10.1038/nature24286) (cit. alle pp. 1, 2).
- [6] R. Staden. «A strategy of DNA sequencing employing computer programs». In: *Nucleic Acids Res* 6.7 (giu. 1979), pp. 2601–2610. DOI: [10.1093/nar/6.7.2601](https://doi.org/10.1093/nar/6.7.2601) (cit. a p. 1).
- [7] B. Wajid e E. Serpedin. «Do it yourself guide to genome assembly». In: *Briefings in Functional Genomics* 15.1 (nov. 2014), pp. 1–9. ISSN: 2041-2649. DOI: [10.1093/bfgp/elu042](https://doi.org/10.1093/bfgp/elu042) (cit. a p. 2).
- [8] C. Noune. «Dynamics, diversity and evolution of Baculoviruses». Tesi di dott. Queensland University of Technology, 2017. DOI: [10.5204/thesis.eprints.113154](https://doi.org/10.5204/thesis.eprints.113154) (cit. a p. 2).
- [9] J. M. Heather e B. Chain. «The sequence of sequencers: The history of sequencing DNA». In: *Genomics* 107.1 (gen. 2016), pp. 1–8. DOI: [10.1016/j.ygeno.2015.11.003](https://doi.org/10.1016/j.ygeno.2015.11.003) (cit. a p. 2).
- [10] T. R. Gregory. «Synergy between sequence and size in Large-scale genomics». In: *Nature Reviews Genetics* 6.9 (set. 2005), pp. 699–708. ISSN: 1471-0064. DOI: [10.1038/nrg1674](https://doi.org/10.1038/nrg1674) (cit. a p. 3).

- [11] H. Sun et al. «findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies». In: *Bioinformatics* 34.4 (ott. 2017), pp. 550–557. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btx637](https://doi.org/10.1093/bioinformatics/btx637) (cit. alle pp. 3, 5, 10, 15, 20–22).
- [12] B. Pucker. «Mapping-based genome size estimation». In: *bioRxiv* (apr. 2019). DOI: [10.1101/607390](https://doi.org/10.1101/607390) (cit. alle pp. 3, 16, 21, 26).
- [13] G. Marçais e C. Kingsford. «A fast, lock-free approach for efficient parallel counting of occurrences of k-mers». In: *Bioinformatics* 27.6 (gen. 2011), pp. 764–770. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btr011](https://doi.org/10.1093/bioinformatics/btr011) (cit. alle pp. 3, 4).
- [14] S. Deorowicz et al. «KMC 2: fast and resource-frugal k-mer counting». In: *Bioinformatics* 31.10 (gen. 2015), pp. 1569–1576. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btv022](https://doi.org/10.1093/bioinformatics/btv022) (cit. a p. 4).
- [15] D. Mapleson et al. «KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies». In: *Bioinformatics* 33.4 (feb. 2017), pp. 574–576. DOI: [10.1093/bioinformatics/btw663](https://doi.org/10.1093/bioinformatics/btw663) (cit. a p. 4).
- [16] J. Sohn e J. Nam. «The present and future of de novo whole-genome assembly». In: *Briefings in Bioinformatics* 19.1 (ott. 2016), pp. 23–40. ISSN: 1477-4054. DOI: [10.1093/bib/bbw096](https://doi.org/10.1093/bib/bbw096) (cit. alle pp. 4, 5).
- [17] X. Li e M. S. Waterman. «Estimating the repeat structure and length of DNA sequences using L-tuples». In: *Genome Res* 13.8 (ago. 2003), pp. 1916–1922. DOI: [10.1101/gr.1251803](https://doi.org/10.1101/gr.1251803) (cit. alle pp. 4, 11).
- [18] G. W. Vurture et al. «GenomeScope: fast reference-free genome profiling from short reads». In: *Bioinformatics* 33.14 (lug. 2017), pp. 2202–2204. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btx153](https://doi.org/10.1093/bioinformatics/btx153) (cit. alle pp. 5, 13, 20).
- [19] J. Butler et al. «ALLPATHS: de novo assembly of whole-genome shotgun micro-reads». In: *Genome Res* 18.5 (mag. 2008), pp. 810–820. DOI: [10.1101/gr.7337908](https://doi.org/10.1101/gr.7337908) (cit. a p. 7).
- [20] I. Maccallum et al. «ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads». In: *Genome Biol* 10.10 (ott. 2009), R103. DOI: [10.1186/gb-2009-10-10-r103](https://doi.org/10.1186/gb-2009-10-10-r103) (cit. a p. 7).
- [21] S. Gnerre et al. «High-quality draft assemblies of mammalian genomes from massively parallel sequence data». In: *Proceedings of the National Academy of Sciences* 108.4 (gen. 2011), pp. 1513–1518. DOI: [10.1073/pnas.1017351108](https://doi.org/10.1073/pnas.1017351108) (cit. alle pp. 7, 9, 19).
- [22] B. Liu et al. «Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects». In: *arXiv e-prints* (ago. 2013), arXiv:1308.2012. DOI: [10.48550/ARXIV.1308.2012](https://doi.org/10.48550/ARXIV.1308.2012) (cit. alle pp. 10, 12, 20).

- [23] M. G. Ross et al. «Characterizing and measuring bias in sequence data». In: *Genome Biology* 14.5 (mag. 2013), R51. ISSN: 1474-760X. DOI: [10.1186/gb-2013-14-5-r51](https://doi.org/10.1186/gb-2013-14-5-r51) (cit. a p. 12).
- [24] A. Azzalini. «A Class of Distributions Which Includes the Normal Ones». In: *Scandinavian Journal of Statistics* 12.2 (1985), pp. 171–178. ISSN: 03036898, 14679469. URL: <http://www.jstor.org/stable/4615982> (visitato il 05/08/2022) (cit. a p. 15).
- [25] A. Azzalini. «The Skew-normal Distribution and Related Multivariate Families». In: *Scandinavian Journal of Statistics* 32.2 (mag. 2005), pp. 159–188. DOI: [10.1111/j.1467-9469.2005.00426.x](https://doi.org/10.1111/j.1467-9469.2005.00426.x) (cit. a p. 15).
- [26] H. Jiang et al. «Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads». In: *BMC Bioinformatics* 15.1 (giu. 2014), p. 182. ISSN: 1471-2105. DOI: [10.1186/1471-2105-15-182](https://doi.org/10.1186/1471-2105-15-182) (cit. a p. 16).
- [27] H. Xu et al. «FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads». In: *PLOS ONE* 7.12 (dic. 2012), pp. 1–6. DOI: [10.1371/journal.pone.0052249](https://doi.org/10.1371/journal.pone.0052249) (cit. a p. 16).
- [28] H. Li e R. Durbin. «Fast and accurate short read alignment with Burrows–Wheeler transform». In: *Bioinformatics* 25.14 (mag. 2009), pp. 1754–1760. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324) (cit. a p. 16).
- [29] H. Li et al. «The Sequence Alignment/Map format and SAMtools». In: *Bioinformatics* 25.16 (giu. 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352) (cit. alle pp. 16, 17).
- [30] F. A. Simão et al. «BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs». In: *Bioinformatics* 31.19 (giu. 2015), pp. 3210–3212. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btv351](https://doi.org/10.1093/bioinformatics/btv351) (cit. a p. 17).
- [31] W. Li et al. «TruMatch—a BLAST post-processor that identifies bona fide sequence matches to genome assemblies». In: *Bioinformatics* 21.9 (gen. 2005), pp. 2097–2098. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bti257](https://doi.org/10.1093/bioinformatics/bti257) (cit. a p. 17).
- [32] B. Pucker e S. F. Brockington. «Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes». In: *BMC Genomics* 19.1 (dic. 2018), p. 980. ISSN: 1471-2164. DOI: [10.1186/s12864-018-5360-z](https://doi.org/10.1186/s12864-018-5360-z) (cit. a p. 17).
- [33] R. Li et al. «De novo assembly of human genomes with massively parallel short read sequencing». In: *Genome Research* 20.2 (2010), pp. 265–272. DOI: [10.1101/gr.097261.109](https://doi.org/10.1101/gr.097261.109) (cit. a p. 19).
- [34] D. Li et al. «Genomic data from Escherichia coli O104: H4 isolate TY-2482». In: *BGI Shenzhen* (2011) (cit. a p. 20).

- [35] S. Nygaard et al. «The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming». In: *Genome Research* 21.8 (2011), pp. 1339–1348. DOI: [10.1101/gr.121392.111](https://doi.org/10.1101/gr.121392.111) (cit. a p. 20).
- [36] X. Xu et al. «Genome sequence and analysis of the tuber crop potato». In: *Nature* 475.7355 (lug. 2011), pp. 189–195. ISSN: 1476-4687. DOI: [10.1038/nature10158](https://doi.org/10.1038/nature10158) (cit. a p. 20).
- [37] R. Li et al. «The sequence and de novo assembly of the giant panda genome». In: *Nature* 463.7279 (gen. 2010), pp. 311–317. ISSN: 1476-4687. DOI: [10.1038/nature08696](https://doi.org/10.1038/nature08696) (cit. a p. 20).
- [38] The Arabidopsis Genome Initiative. «Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*». In: *Nature* 408.6814 (dic. 2000), pp. 796–815. ISSN: 1476-4687. DOI: [10.1038/35048692](https://doi.org/10.1038/35048692) (cit. a p. 26).
- [39] B. Pucker et al. «A De Novo Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny». In: *PLOS ONE* 11.10 (ott. 2016), pp. 1–23. DOI: [10.1371/journal.pone.0164321](https://doi.org/10.1371/journal.pone.0164321) (cit. a p. 27).
- [40] B. S. Ozdemir et al. «*Brachypodium* Genomics». In: *International Journal of Plant Genomics* 2008 (feb. 2008), p. 536104. ISSN: 1687-5370. DOI: [10.1155/2008/536104](https://doi.org/10.1155/2008/536104) (cit. a p. 29).
- [41] G. Haberer et al. «Structure and architecture of the maize genome». In: *Plant physiology* 139.4 (dic. 2005), pp. 1612–1624. ISSN: 0032-0889. DOI: [10.1104/pp.105.068718](https://doi.org/10.1104/pp.105.068718) (cit. a p. 29).
- [42] A. Barone et al. «Structural and functional genomics of tomato». In: *International journal of plant genomics* 2008 (2008), pp. 820274–820274. ISSN: 1687-5370. DOI: [10.1155/2008/820274](https://doi.org/10.1155/2008/820274) (cit. a p. 29).
- [43] S. Myles et al. «Rapid Genomic Characterization of the Genus *Vitis*». In: *PLOS ONE* 5.1 (gen. 2010), pp. 1–9. DOI: [10.1371/journal.pone.0008219](https://doi.org/10.1371/journal.pone.0008219) (cit. a p. 29).