Quiz show «algoritmi deterministici»

Quiz show «algoritmi deterministici»

- Simulazione dell’esame scritto
- **Una sola risposta esatta** (a volte alcune possono sembrare giuste ma sono errate solo per qualche dettaglio)
- **1 punto** per ogni risposta corretta ma **normalizzazione** se alcune domande risultano poco chiare e sbagliate in massa
- Domande soprattutto (ma non solo) sulla parte teorica. Quella pratica (soprattutto) sarà testata dal progetto

Quiz show «algoritmi deterministici»

- Simulazione dell'esame scritto
- **Una sola risposta esatta** (a volte alcune possono sembrare giuste ma sono errate solo per qualche dettaglio)
- **1 punto** per ogni risposta corretta ma **normalizzazione** se alcune domande risultano poco chiare e sbagliate in massa
- Domande soprattutto (ma non solo) sulla parte teorica. Quella pratica (soprattutto) sarà testata dal progetto
- Note sul Progetto: a Dicembre, dopo le 60 ore del corso "ufficiale", ci rendiamo disponibili per qualche tutoraggio pratico per aiutarvi con i progetti

Quiz show «algoritmi deterministici»

- **Se aveste solo Illumina short reads e voleste assemblare un genoma, cosa usereste?**
  - E’ impossibile computazionalmente effettuare un assemblaggio usando solo short reads
  - Un metodo basato su grafi di De Brujin
  - Un metodo basato su grafi delle sovrapposizioni (overlap-graph)
  - Un metodo basato sulla trasformata di Burrow-Wheeler
  - Un metodo che usi `bwa` per mappare ogni read contro tutte le altre

Quiz show «algoritmi deterministici»

- **Se aveste solo Illumina short reads e voleste assemblare un genoma, cosa usereste?**
  - E' impossibile computazionalmente effettuare un assemblaggio usando solo short reads
  - Un metodo basato su grafi di De Brujin
  - Un metodo basato su grafi delle sovrapposizioni (overlap-graph)
  - Un metodo basato sulla trasformata di Burrow-Wheeler
  - Un metodo che usi `bwa` per mappare ogni read contro tutte le altre

Quiz show «algoritmi deterministici»

- **Quale è il maggior collo di bottiglia dei grafi delle sovrapposizioni (overlap graph) nell'assemblaggio di genomi?**
  - Non si possono usare su short reads
  - Non si possono usare con long-reads
  - Sono computazionalmente inefficienti all'aumentare dei dati
  - Introducono molti errori e riarrangiamenti

Quiz show «algoritmi deterministici»

- **Quale è il maggior collo di bottiglia dei grafi delle sovrapposizioni (overlap graph) nell'assemblaggio di genomi?**
  - Non si possono usare su short reads
  - Non si possono usare con long-reads
  - Sono computazionalmente inefficienti all'aumentare dei dati
  - Introducono molti errori e riarrangiamenti

Quiz show «algoritmi deterministici»

- **Come aiutano i k-mer a far rispettare le condizioni di ricerca di un ciclo Euleriano? Selezionare la risposta errata.**
  - Rendono la coverage/read depth piu' uniforme
  - Permettono la costruzione di un grafo di De Brujin
  - Usando k-mer corti, la maggio parte di questi sono corretti anche per reads con errori
  - Rendono la composizione GC piu' uniforme

Quiz show «algoritmi deterministici»

- **Come aiutano i k-mer a far rispettare le condizioni di ricerca di un ciclo Euleriano? Selezionare la risposta errata.**
  - Rendono la coverage/read depth piu' uniforme
  - Permettono la costruzione di un grafo di De Brujin
  - Usando k-mer corti, la maggio parte di questi sono corretti anche per reads con errori
  - Rendono la composizione GC piu' uniforme

Quiz show «algoritmi deterministici»

- **Quale non è una proprietà dei minimizer**
  - Rappresentando una sequenza con L piu’ piccole, ci attendiamo meno minimizers
  - Sono robusti ad errori
  - Fungono da seeds/ancore durante la mappatura di long reads
  - Sono utilizzati in programmi di metagenomica

Quiz show «algoritmi deterministici»

- **Quale non è una proprietà dei minimizer**
  - Rappresentando una sequenza con L piu’ piccole, ci attendiamo meno minimizers
  - Sono robusti ad errori
  - Fungono da seeds/ancore durante la mappatura di long reads
  - Sono utilizzati in programmi di metagenomica

Quiz show «algoritmi deterministici»

- **Con il comando** `bwa index` **cosa effettuiamo?**
  - Creiamo la trasformata di Burrow-Wheeler per un set di sequenze da mappare
  - Creiamo degli array di suffissi (suffix array) parziali per un genoma su cui vogliamo mappare delle reads
  - Effettuiamo una mappatura di reads con bwa
  - Cerchiamo i seeds per il processo di mappatura con bwa

Quiz show «algoritmi deterministici»

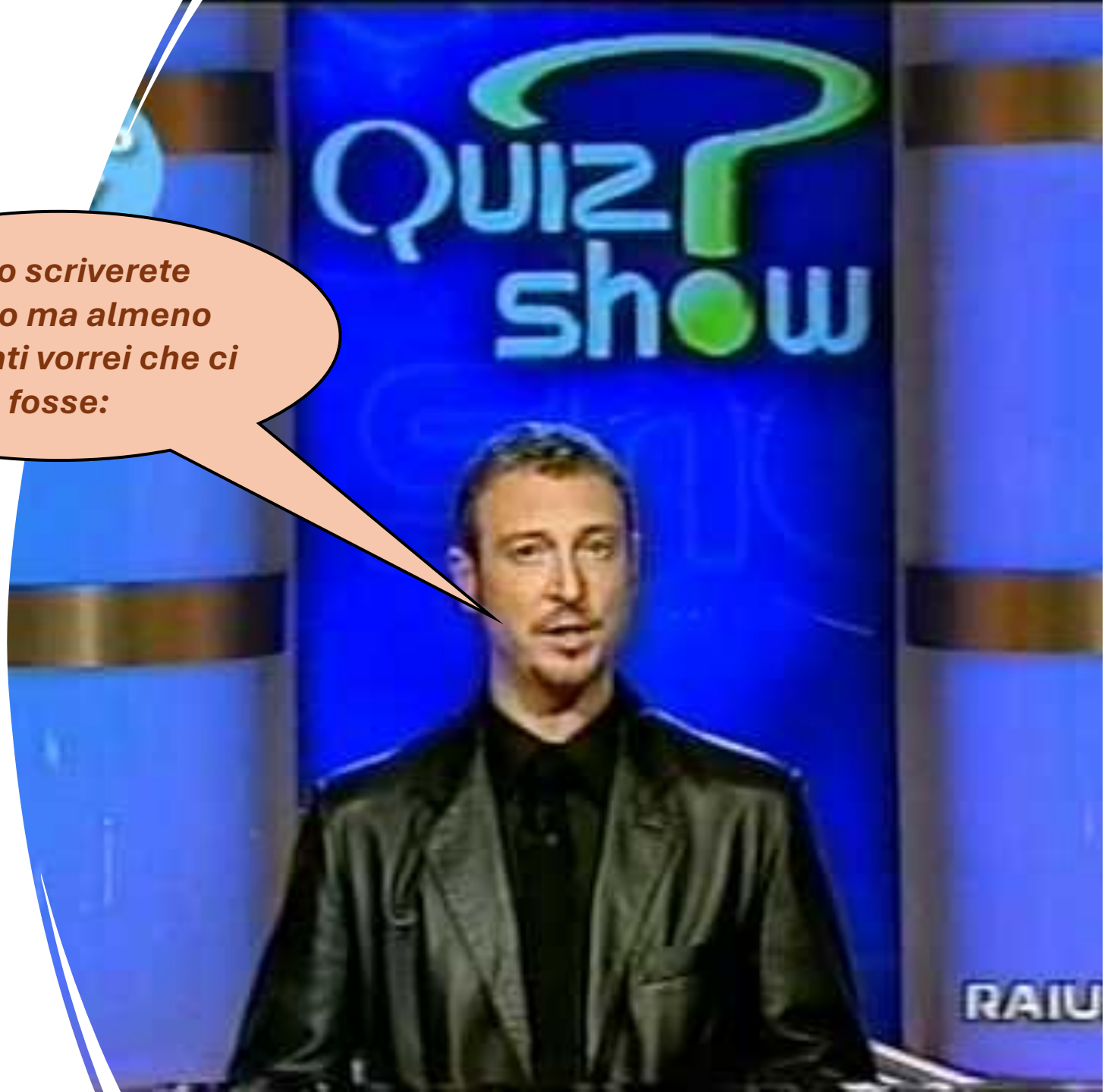
- **Con il comando** `bwa index` **cosa effettuiamo?**
  - Creiamo la trasformata di Burrow-Wheeler per un set di sequenze da mappare
  - Creiamo degli array di suffissi (suffix array) parziali per un genoma su cui vogliamo mappare delle reads
  - Effettuiamo una mappatura di reads con bwa
  - Cerchiamo i seeds per il processo di mappatura con bwa

Quiz show «algoritmi deterministici»

Esempio di domanda aperta:

- Descrivete vantaggi e buone pratiche dell’utilizzo della linea di comando per la genomica computazionale

Quiz show «algoritmi deterministici»



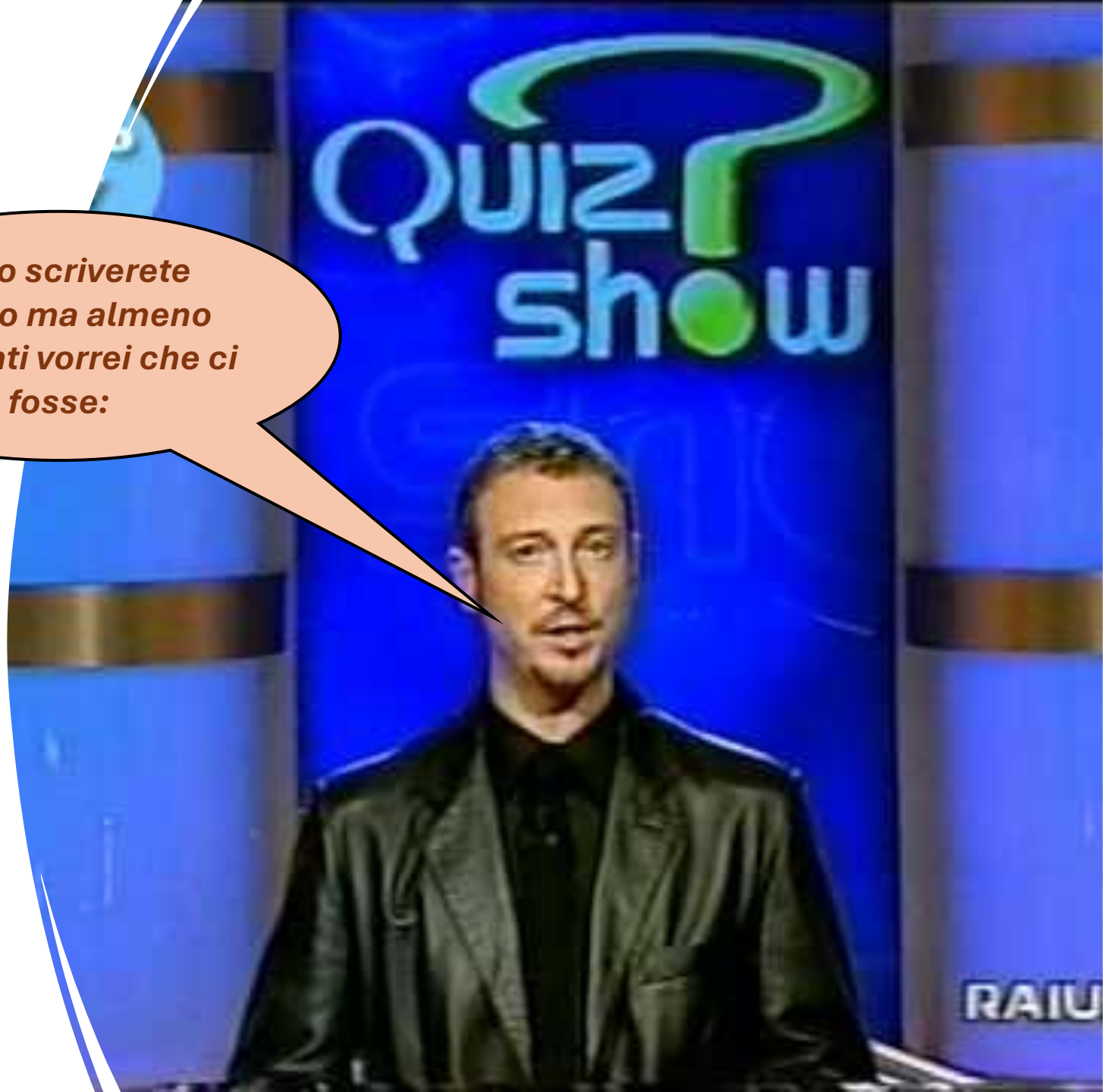
Esempio di domanda aperta:

- Descrivete vantaggi e buone pratiche dell'utilizzo della linea di comando per la genomica computazionale

Risposta

- Possibilità di lanciare programmi altrimenti non utilizzabili
- Riproducibilità (scientifica)
- Riproducibilità (per noi, facile rieseguire il codice)
- Possibilità di «customizzare» i comandi e le operazioni con argomenti ed analisi scritte da noi

Quiz show «algoritmi deterministici»



Esempio di domanda aperta:

- Descrivete vantaggi e buone pratiche dell'utilizzo della linea di comando per la genomica computazionale

Risposta

- Possibilità di lanciare programmi altrimenti non utilizzabili
- Riproducibilità (scientifica)
- Riproducibilità (per noi, facile rieseguire il codice)
- Possibilità di «customizzare» i comandi e le operazioni con argomenti ed analisi scritte da noi
- *…. molte altre, anche non dette a lezione….*
- *Efficienza di calcolo (no GUI) e possibilità di lavorare su server in remoto*

Quiz show «algoritmi deterministici»

Esempio di esercizio (qui come domanda aperta):

- Rappresentate la sequenza `CGGAACGAT` utilizzando minimizers con parametri L=6 (dimensione della finestra) e k=3 (dimensione del k-mer) e ordine alfabetico.

Quiz show «algoritmi deterministici»

Esempio di esercizio (qui come domanda aperta):

- Rappresentate la sequenza CGG**AAC**GAT utilizzando minimizers con parametri L=6 (dimensione della finestra) e k=3 (dimensione del k-mer) e ordine alfabetico.

Risposta

- **AAC**

Classi di algoritmi

Categoria	Definizione	Esempi	Pro e cons
Deterministici	Stesso input -> stesso output	Burrow-Wheeler Transform per la mappatura; De Brujin Graphs per l'assemblaggio	Esatti e prevedibili.
Probabilistici	Il modello include casualità (randomness).	Modelli binomiali per la genotipizzazione, MCMC, Hidden-Markov come CHROMHMM, PhastCons, etc.	C'è un po' di probabilità. Facilmente interpretabili. Danno stime dell'incertezza nella soluzione
Euristici	Scorciatoie per problemi difficili	Filtraggio di varianti	Approssimati. Semplici ma sporchi.
Machine-learning/AI	Data driven; eterogenei	DeepVariant; AlphaFold.	Potenti. Necessitano moltissimi dati. Poco interpretabili

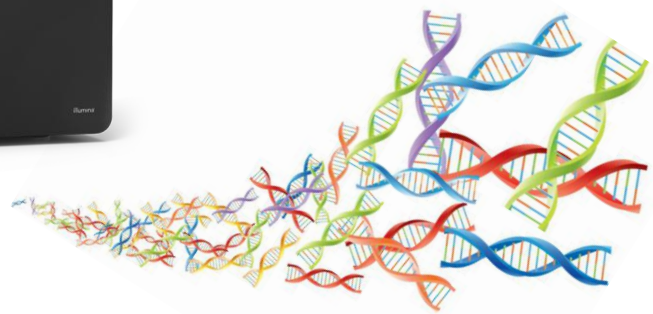
Classi di algoritmi

Categoria	Definizione	Esempi	Pro e cons
Deterministici	Stesso input -> stesso output	Burrow-Wheeler Transform per la mappatura; De Brujin Graphs per l'assemblaggio	Esatti e prevedibili.
Probabilistici 	Il modello include casualità (randomness).	Modelli binomiali per la genotipizzazione, MCMC, Hidden-Markov come CHROMHMM, PhastCons, etc.	C'è un po' di probabilità.  Facilmente interpretabili. Danno stime dell'incertezza nella soluzione
Euristici	Scorciatoie per problemi difficili	Filtraggio di varianti	Approssimati. Semplici ma sporchi.
Machine-learning/AI	Data driven; eterogenei	DeepVariant; AlphaFold.	Potenti. Necessitano moltissimi dati. Poco interpretabili

Vari classi di algoritmi sono utilizzati nella genotipizzazione

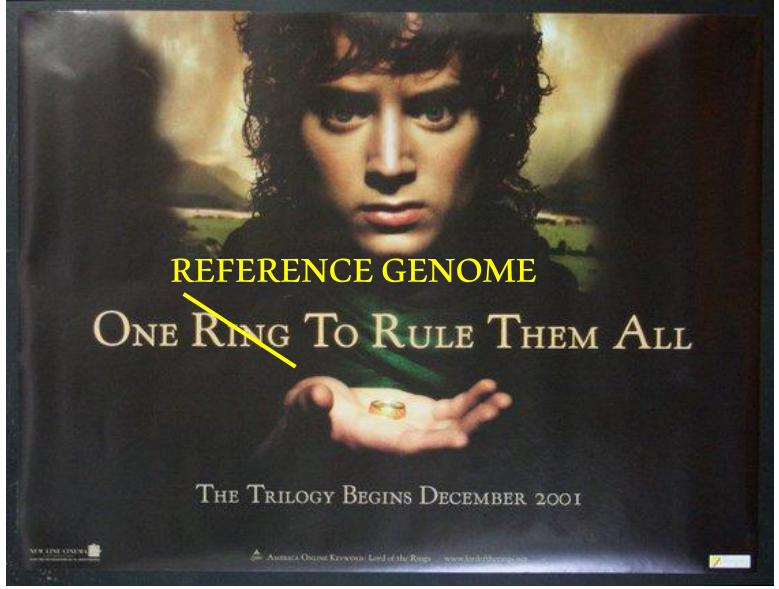
Categoria	Definizione	Esempi	Pro e cons
Deterministici	Stesso input -> stesso output	Burrow-Wheeler Transform per la mappatura; Input per la genotipizzazione	Esatti e prevedibili.
Probabilistici 	Il modello include casualità (randomness).	Da **semplici modelli binomiali** con errore uniforme fino a modelli complessi e **Hidden Markov Model** per l'imputazione	C'è un po' di probabilità.  Facilmente interpretabili. Danno stime dell'incertezza nella soluzione
Euristici	Scorciatoie per problemi difficili	Genotipizzazione e filtraggio di varianti	Approssimati. Semplici ma sporchi.
Machine-learning/AI	Data driven; eterogenei	**DeepVariant** per la genotipizzazioni	Potenti. Necessitano moltissimi dati. Poco interpretabili

Genoma (di referenza)

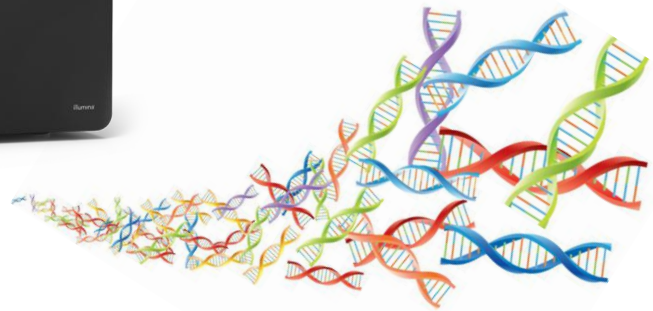


sequenze

assemblaggio



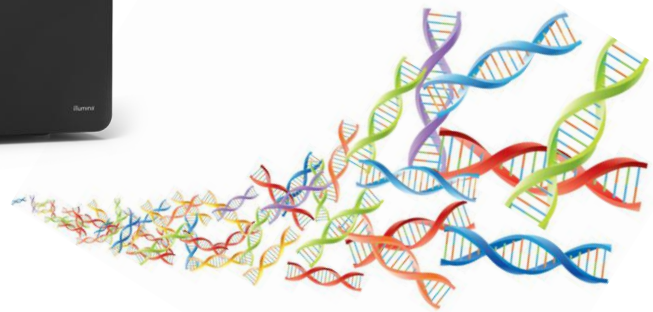
Genoma (di referenza)



sequenze

assemblaggio

Genoma (di referenza)



sequenze

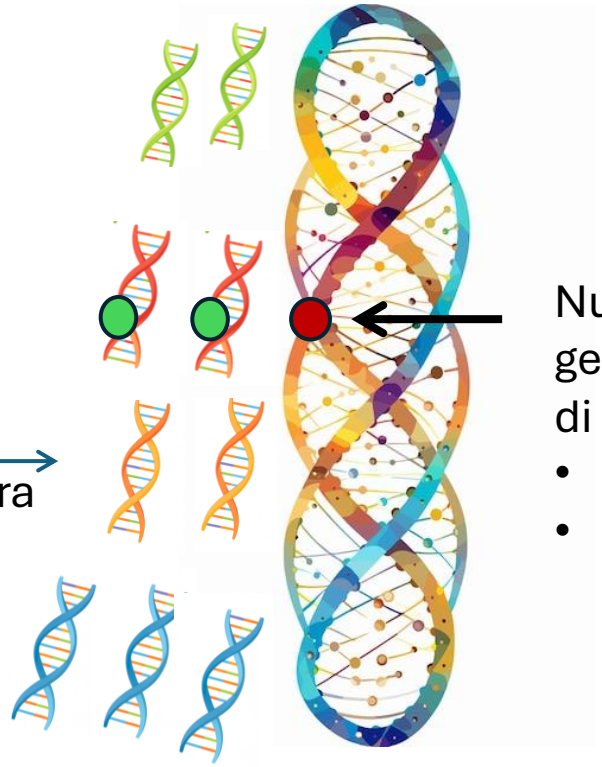
mappatura

La mappatura tollera alcuni nucleotidi differenti dalla referenza (errori e vere differenze)

sequenze

mappatura

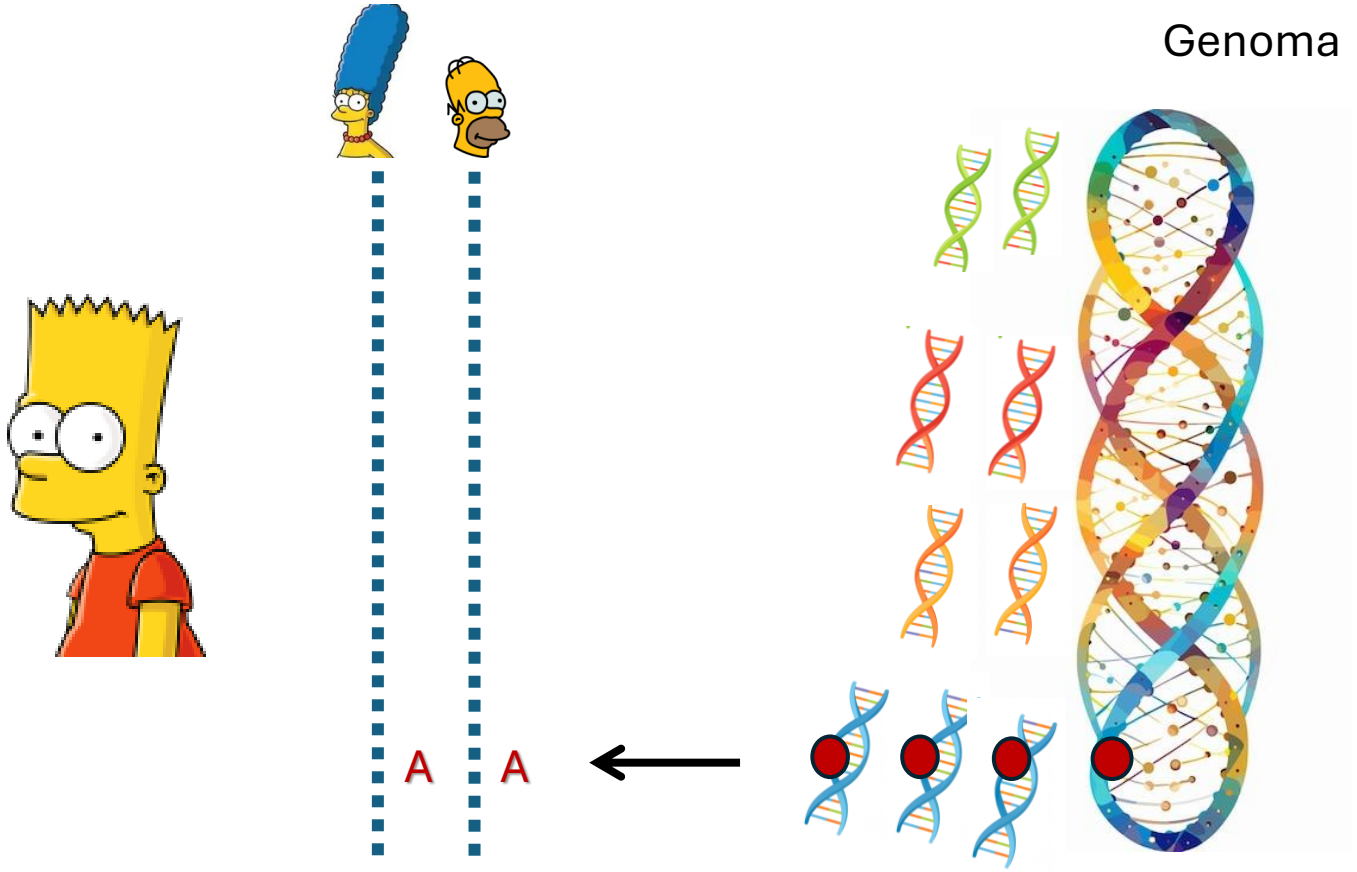
Genoma (di referenza)



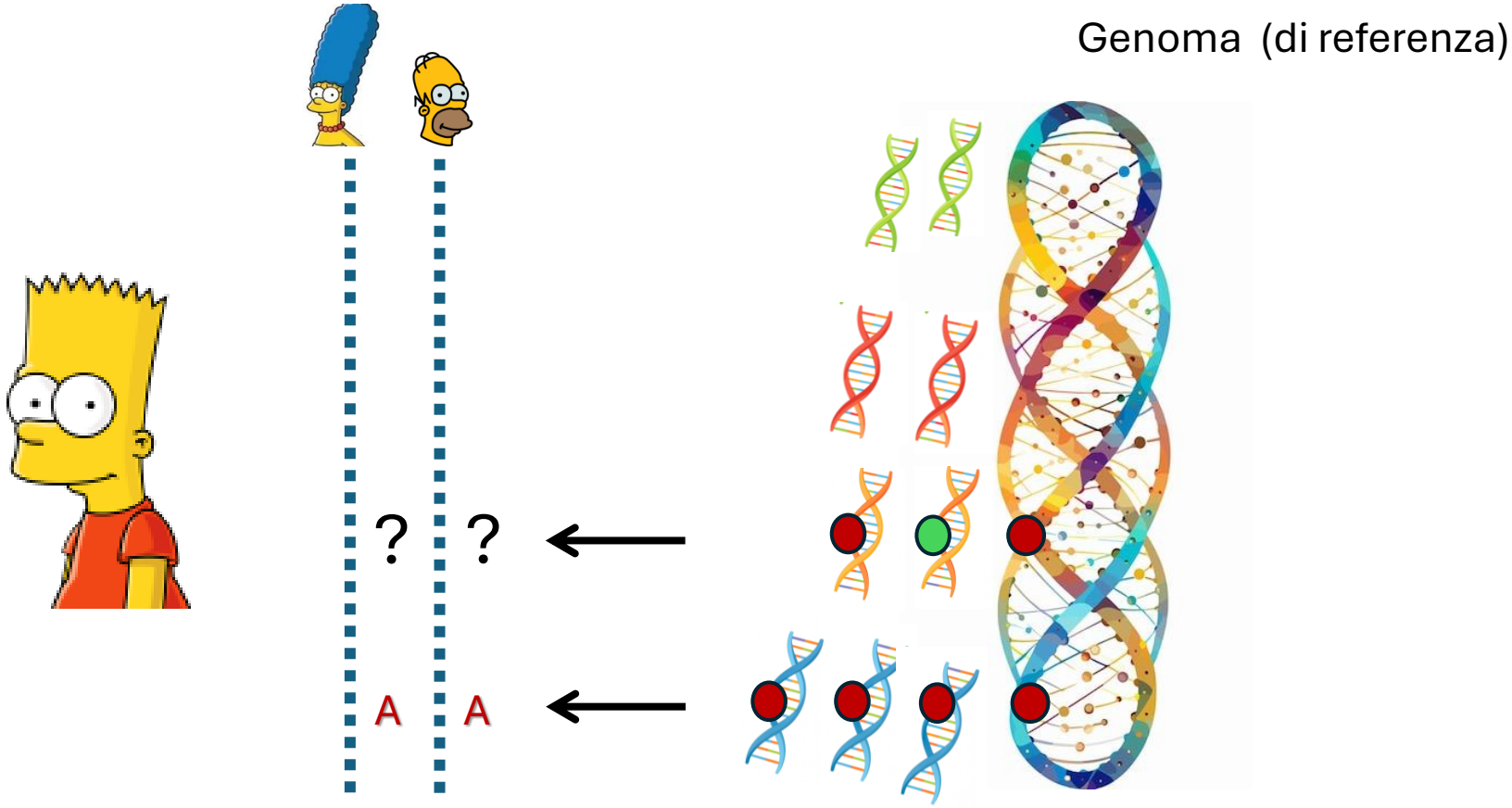
Nucleotidi differenti tra genoma
di referenza e sequenze:

- errori
- differenze tra individui

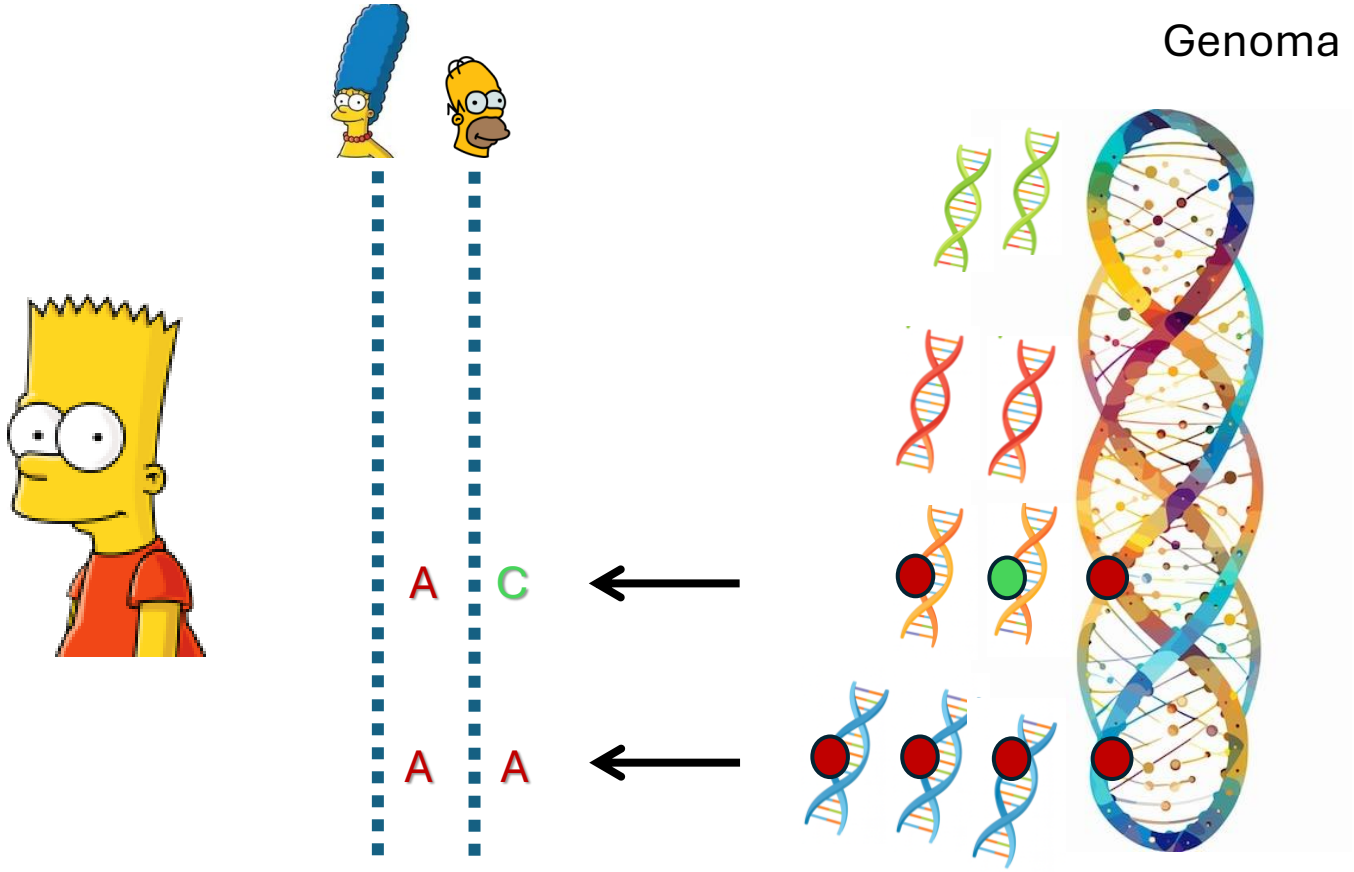
I genotipi (diploidi): un sito omozigote (uguale alla referenza)



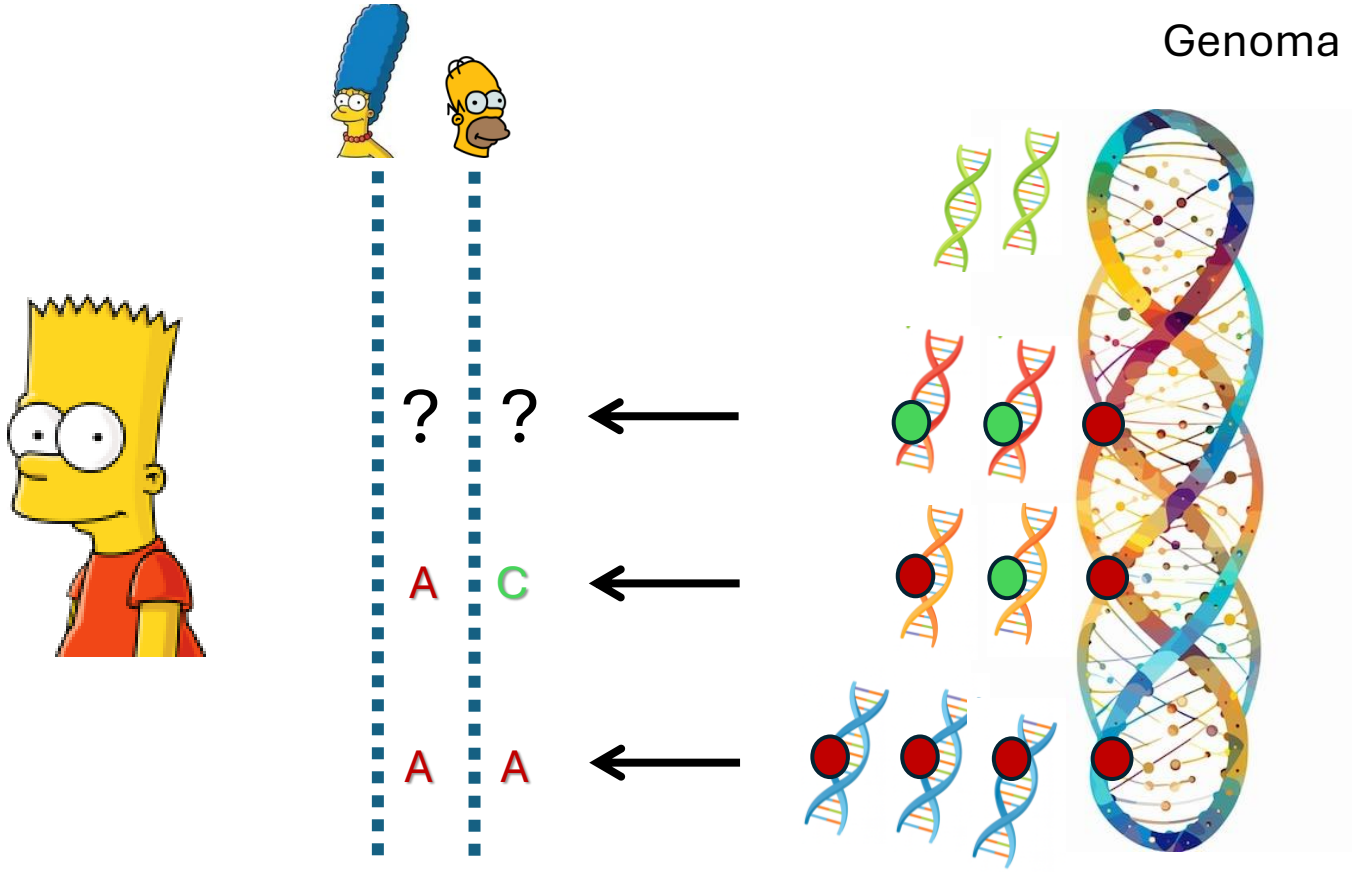
I genotipi (diploidi): un sito eterozigote



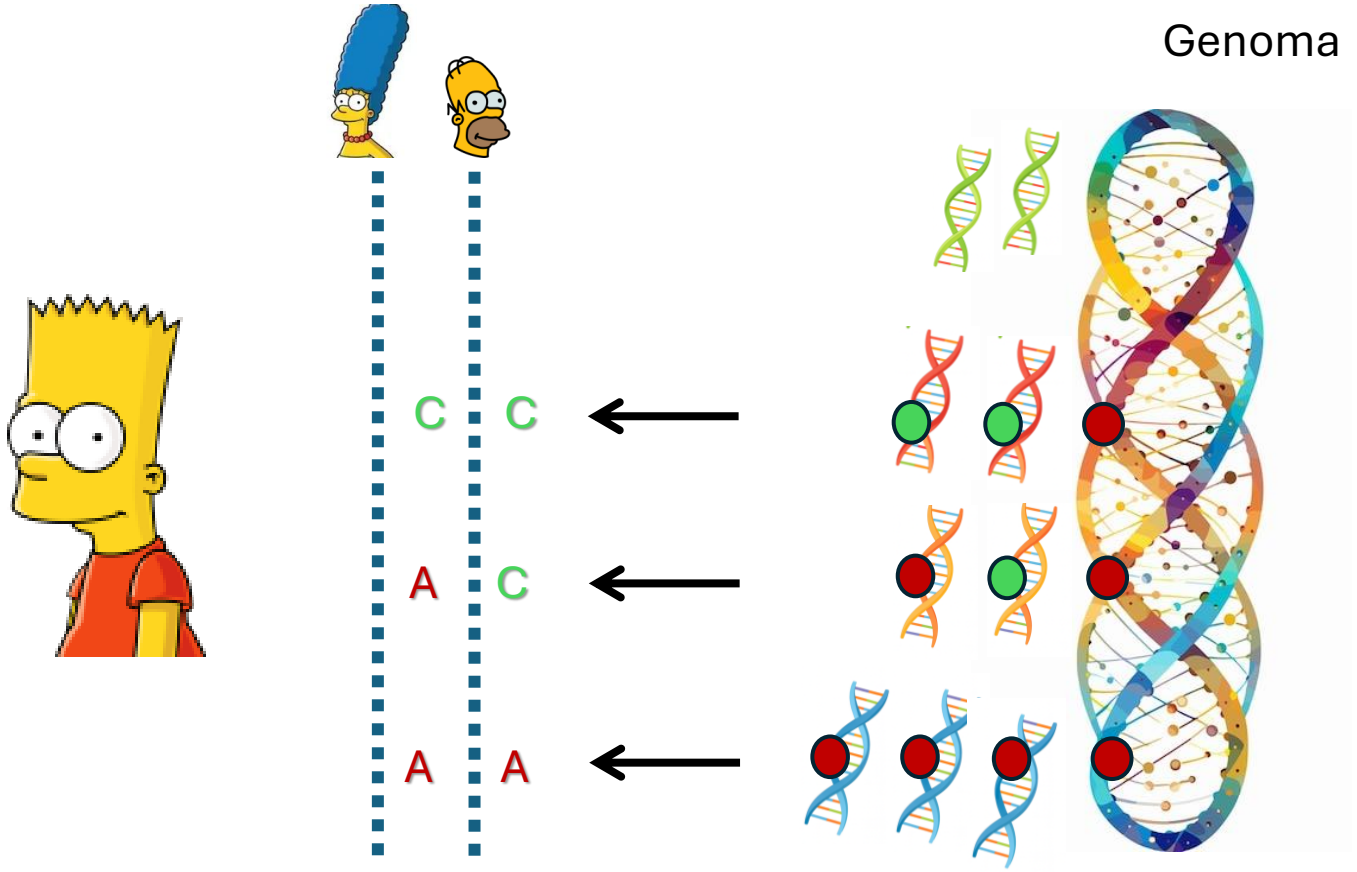
I genotipi (diploidi): un sito eterozigote



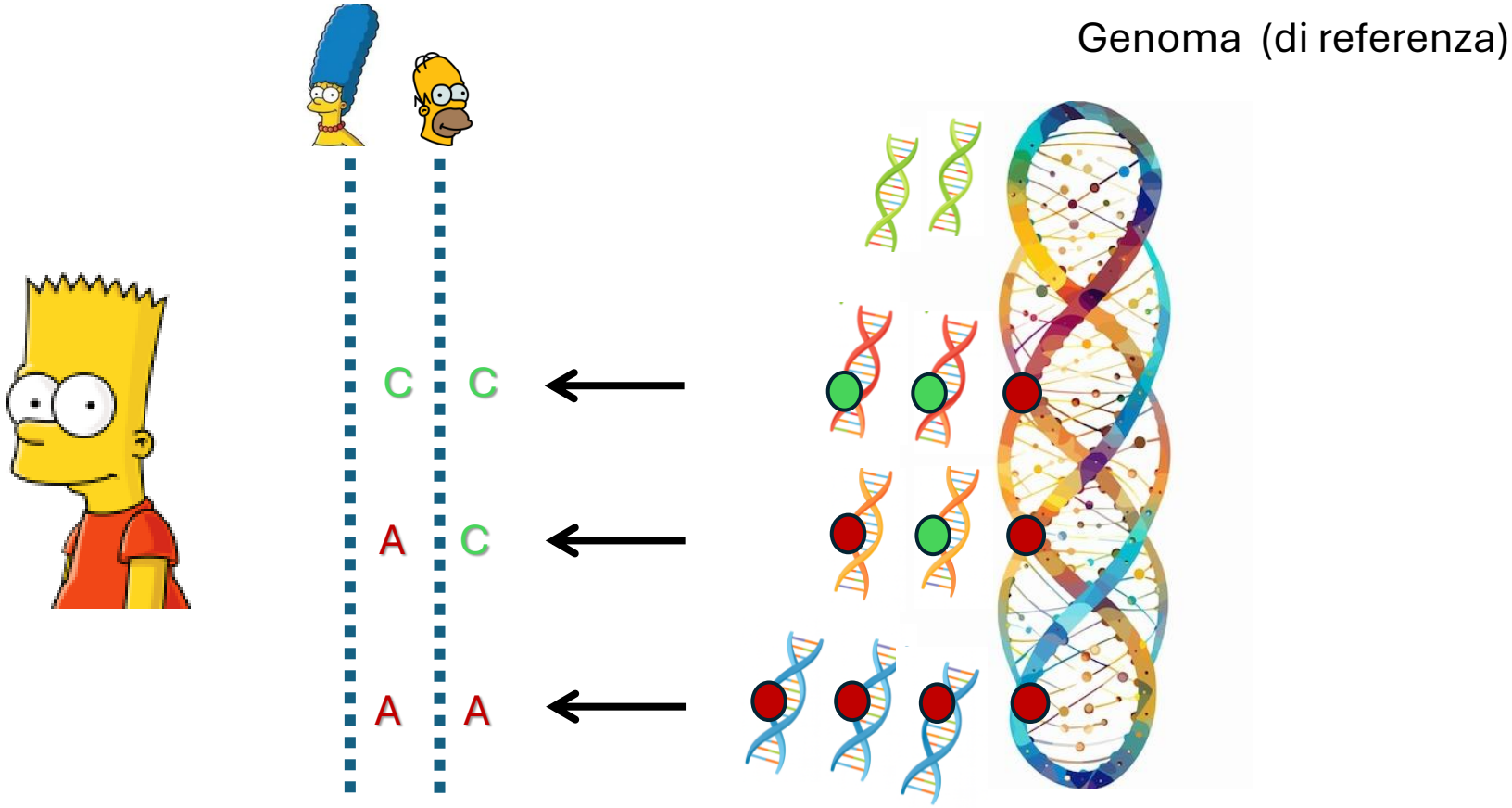
I genotipi (diploidi): un sito omozigote (diverso alla referenza)



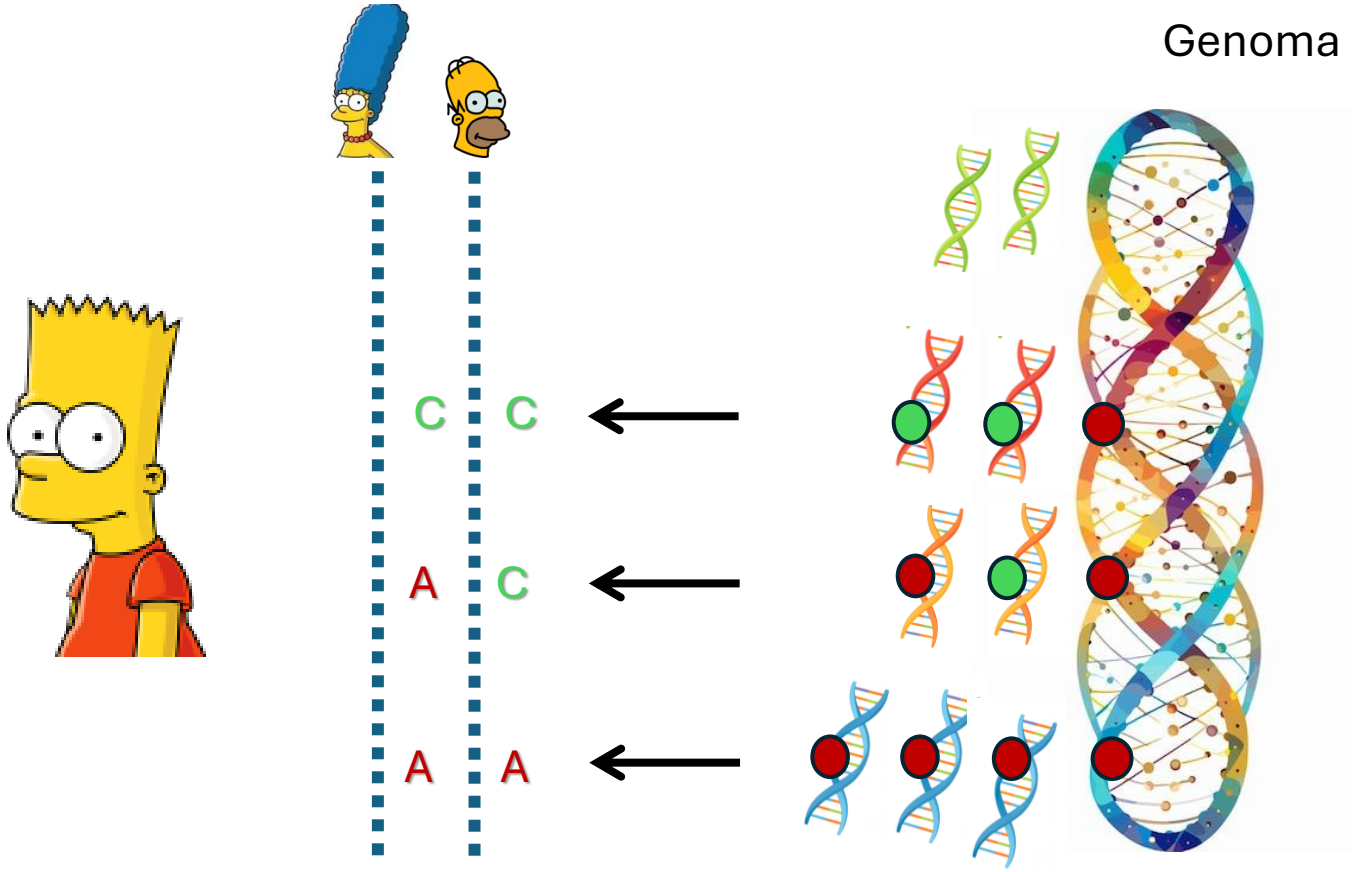
I genotipi (diploidi): un sito omozigote (diverso alla referenza)



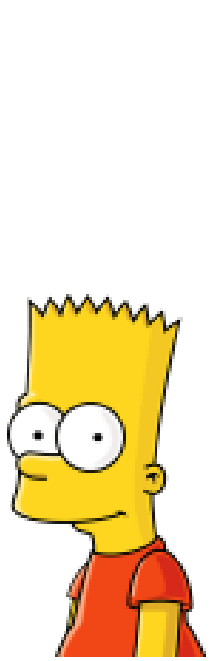
La genotipizzazione

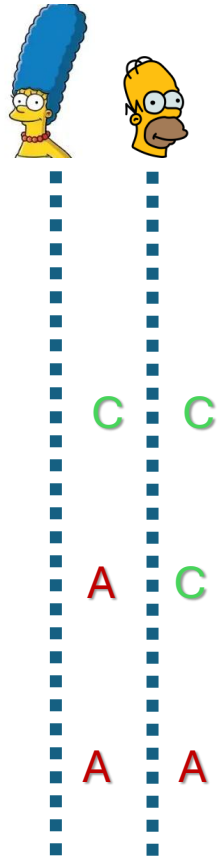


Quali problemi potrebbero esserci in questa ricostruzione?

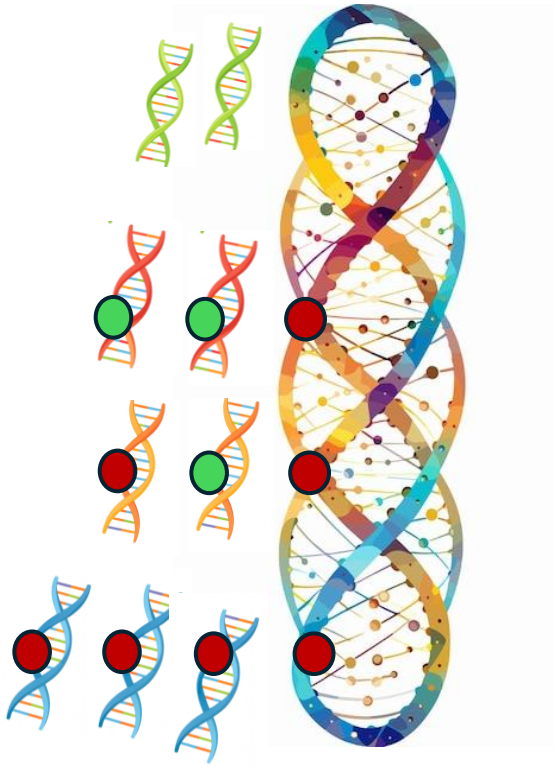


Quali problemi potrebbero esserci in questa ricostruzione?





Genoma (di referenza)

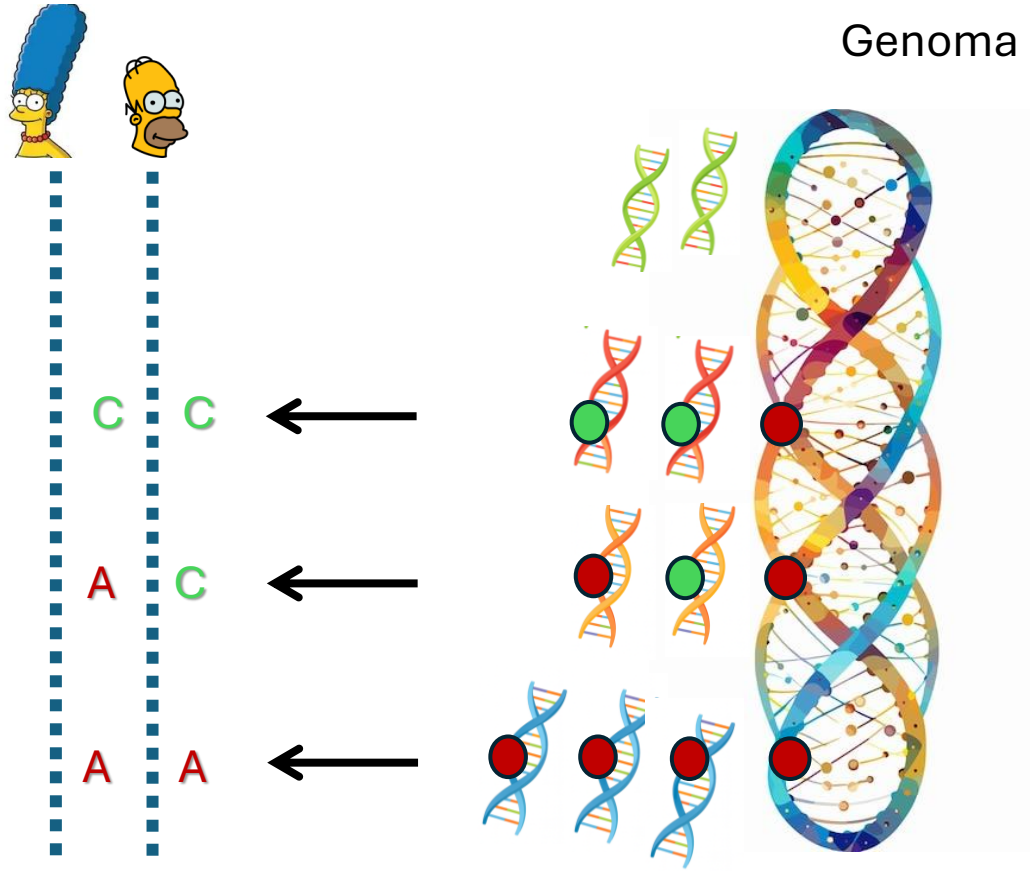


Non aver campionato/mappato
un allele verde diverso dalla referenza

Quali problemi potrebbero esserci in questa ricostruzione?

Uno dei due (soprattutto il verde)
potrebbe essere un errore

Non aver campionato/mappato
un allele verde diverso dalla referenza



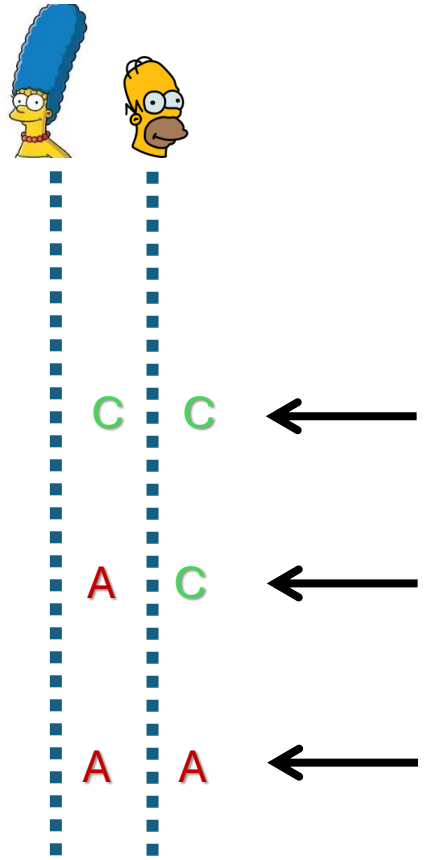
Genoma (di referenza)

Quali problemi potrebbero esserci in questa ricostruzione?

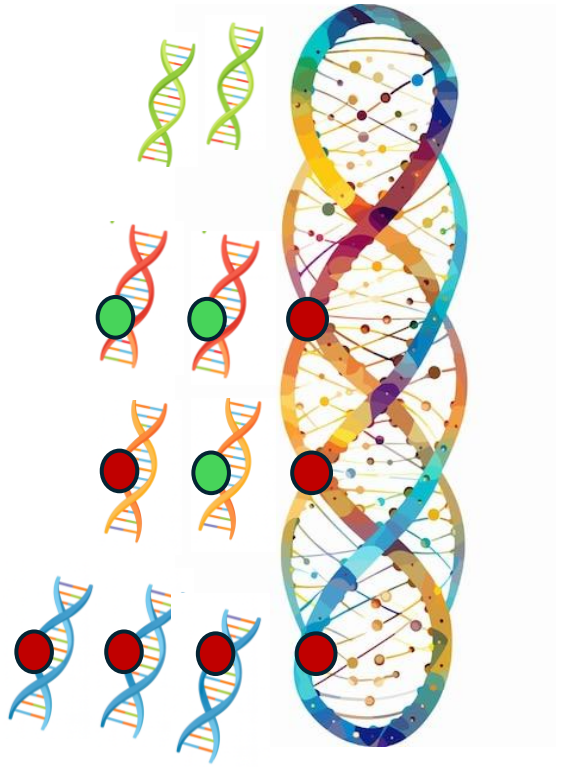
Non aver campionato/mappato
un allele verde diverso dalla referenza;
o un errore sistematico;
o una mappatura errata.

Uno dei due (soprattutto il verde)
potrebbe essere un errore

Non aver campionato/mappato
un allele verde diverso dalla referenza



Genoma (di referenza)

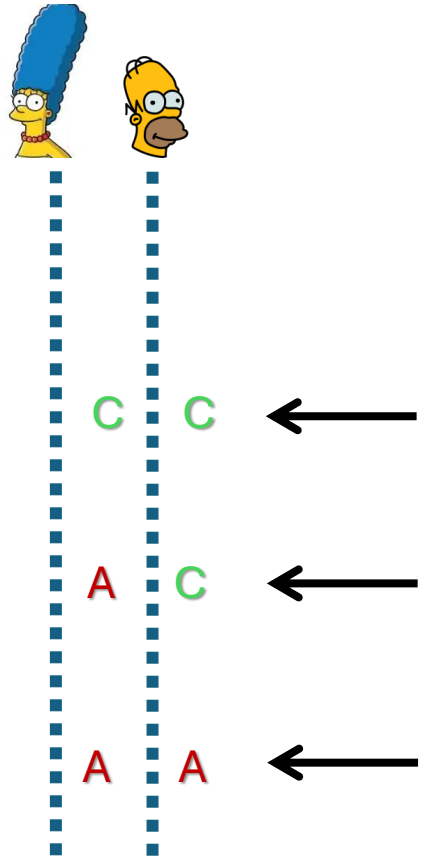


La genotipizzazione è il processo di ricostruzione del genotipo di un individuo a partire dalle varianti sequenziate

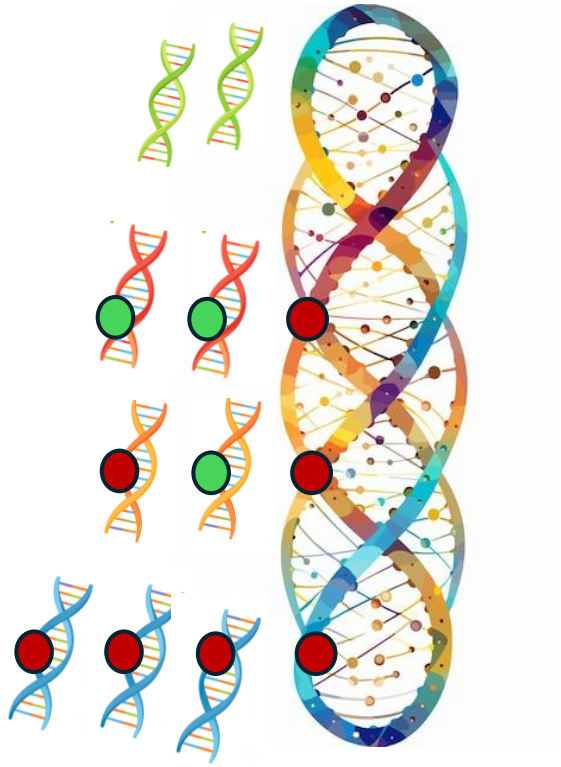
Non aver campionato/mappato
un allele verde diverso dalla referenza;
o un errore sistematico;
o una mappatura errata.

Uno dei due (soprattutto il verde)
potrebbe essere un errore

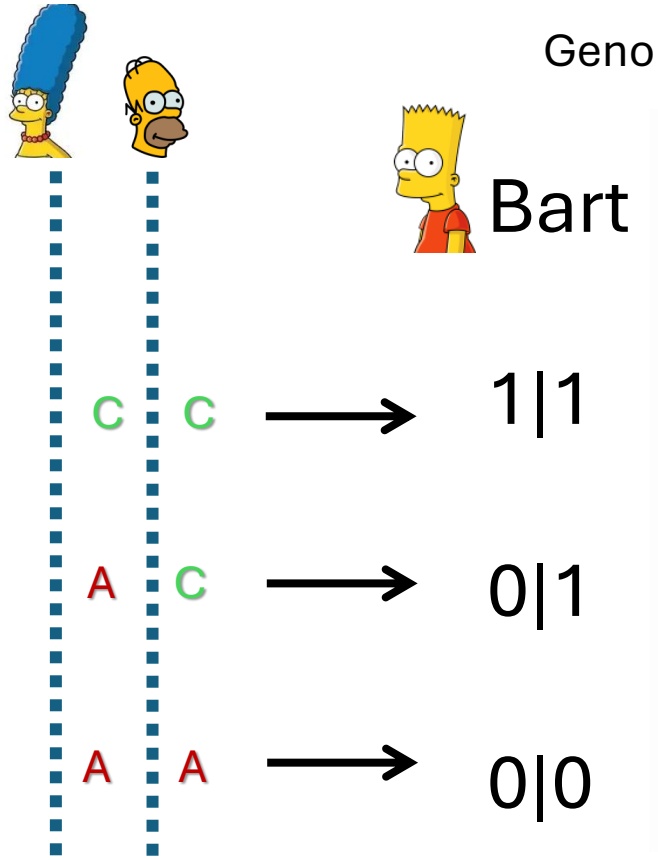
Non aver campionato/mappato
un allele verde diverso dalla referenza



Genoma (di referenza)



Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni



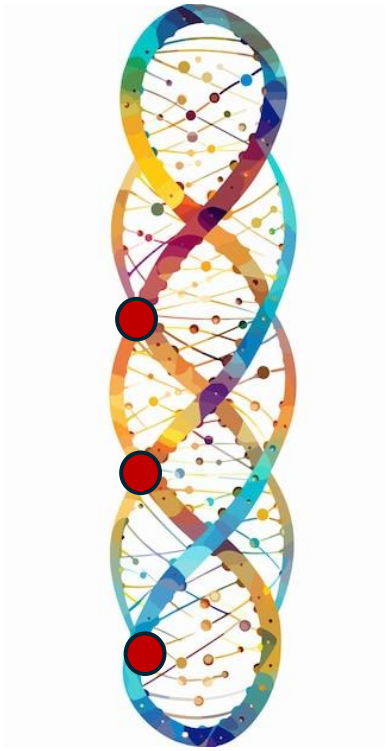
Bart

1|1

0|1

0|0

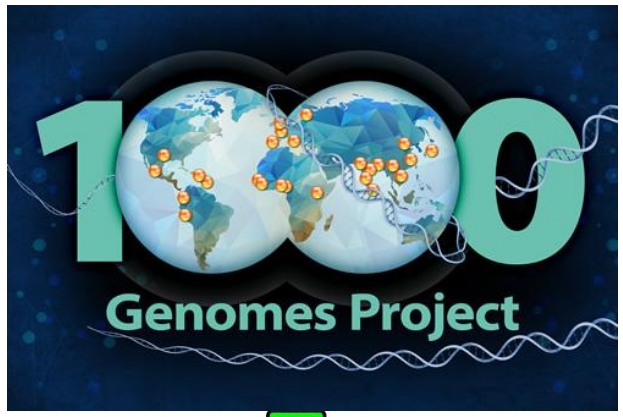
Genoma (di referenza)

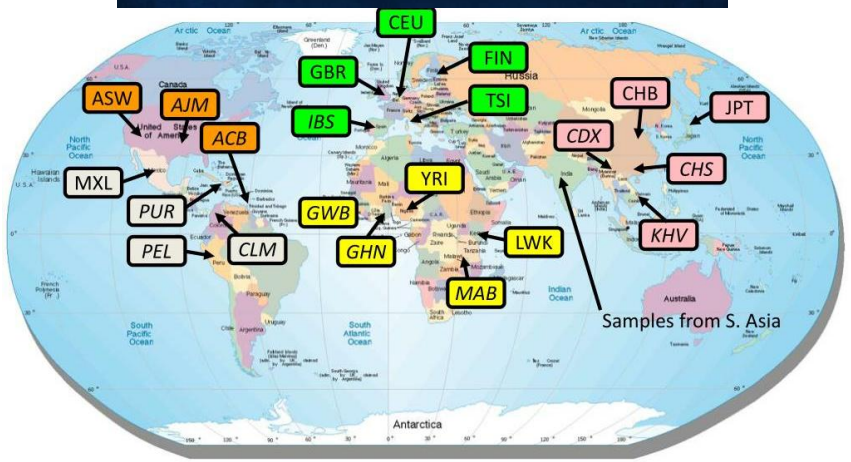


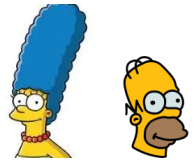
= allele presente nella referenza (**REF**) rappresentato con **0**

= allele non presente nella referenza o alternativo (**ALT**) rappresentato con **1** (o 2,3 etc. se piu' di un allele alternativo)

Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni

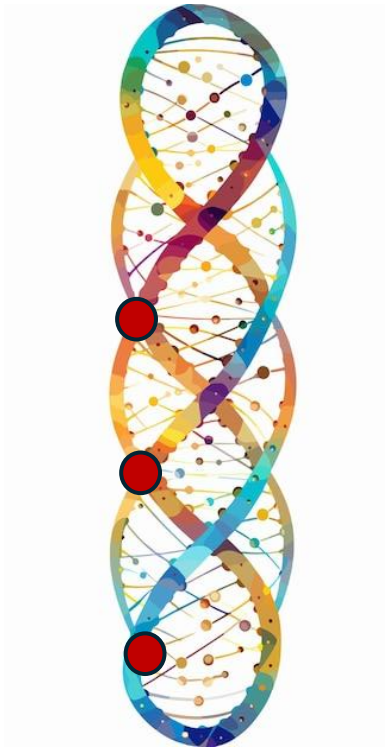






Bart

Genoma (di referenza)



C C → 1/1

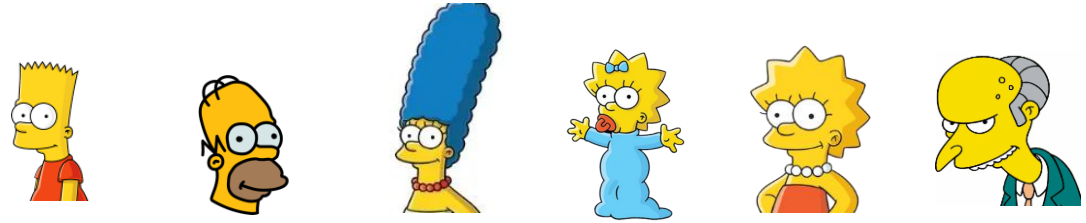
A C → 0/1

A A → 0/0

= allele presente nella referenza (**REF**) rappresentato con **0**

= allele non presente nella referenza o alternativo (**ALT**) rappresentato con **1** (o 2,3 etc. se piu' di un allele alternativo)

Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni



#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	...	0/1	1/1	0/0	1/0	0/0	0/0

Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni






#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	…	0/1	1/1	0/0	1/0	0/0	0/0
1	1422	A	C	50	PASS	…	0/1	0/1	0/0	0/0	0/0	0/0
1	2456	C	G	45	PASS	…	1/0	0/0	1/0	1/0	0/0	0/0
2	111	G	C,T	3	QD10	…	0/1	0/1	0/0	0/0	0/1	0/2

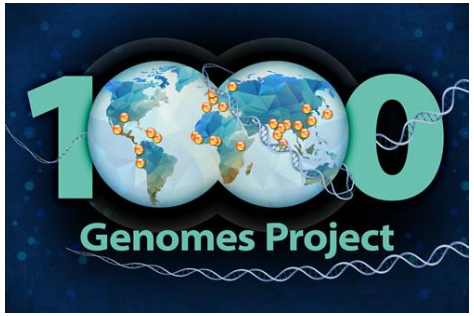
Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni

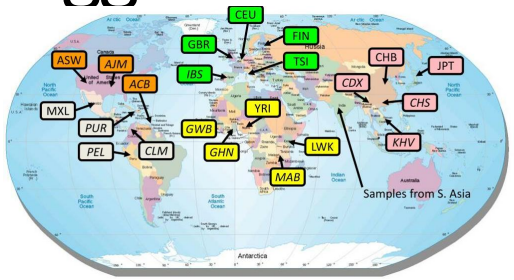





#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	…	0/1	1/1	0/0	1/0	0/0	0/0
1	1422	A	C	50	PASS	…	0/1	0/1	0/0	0/0	0/0	0/0
1	2456	C	G	45	PASS	…	1/0	0/0	1/0	1/0	0/0	0/0
2	111	G	C,T	3	QD10	…	0/1	0/1	0/0	0/0	0/1	0/2
2	921	T	TA	45	PASS	…	0/0	0/0	0/0	0/0	0/0	0/1
2	5002	A	.	51	PASS	…	0/0	0/0	0/1	0/1	0/1	0/0

Il formato VCF (Variant Call Format) è usato per rappresentare tutti i maggiori datasets di genomi di popolazioni





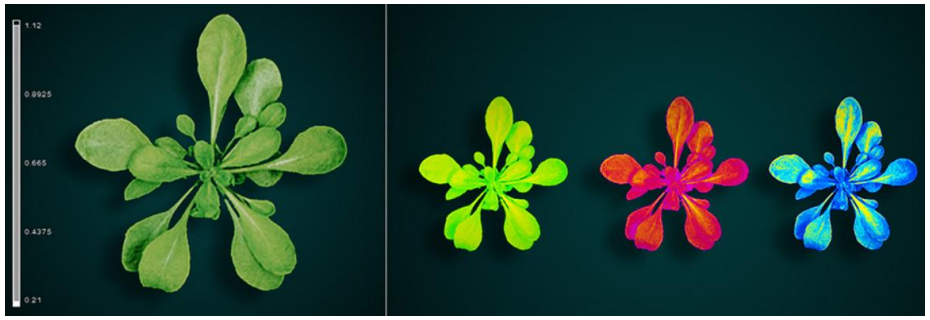
#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	...	0/1	1/1	0/0	1/0	0/0	0/0
1	1422	A	C	50	PASS	...	0/1	0/1	0/0	0/0	0/0	0/0
1	2456	C	G	45	PASS	...	1/0	0/0	1/0	1/0	0/0	0/0
2	111	G	C,T	3	QD10	...	0/1	0/1	0/0	0/0	0/1	0/2
2	921	T	TA	45	PASS	...	0/0	0/0	0/0	0/0	0/0	0/1
2	5002	A	.	51	PASS	...	0/0	0/0	0/1	0/1	0/1	0/0

Referenza						
…	…	…	…	…	…	…
C	C	C	C	C	C	C
C	C	C	C	C	C	C
A	A	A	A	A	A	A
T	**C**	T	**C**	T	T	T
G	G	G	G	G	G	G
C	C	C	C	C	C	C
A	A	A	A	A	A	A
C	C	C	C	C	C	C
G	G	G	G	G	**T**	G
A	A	A	A	A	A	A
T	T	T	T	T	T	T
A	A	A	A	A	A	A
C	C	C	C	C	C	C
T	T	T	T	T	T	T
A	**T**	A	**T**	A	A	A
T	T	T	T	T	T	T
T	T	T	T	T	T	T
T	T	T	T	T	T	T
C	C	C	C	C	C	C
C	C	C	C	C	C	C
…	…	…	…	…	…	…

Mappare rapidamente milioni di sequenze ad un genoma di referenza ha permesso la ricostruzione (economica) di genomi individuali e di ampi datasets di migliaia di genomi per i) studi clinici e individuare varianti patogeniche, ii) studi di popolazione ed evolutivi, iii) screening medico

E questi sono una minuscola parte!!!

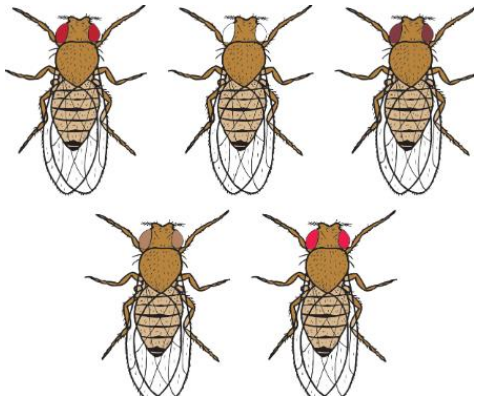
1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*



gnomAD

Genome Aggregation Database

Drosophila Genetic Reference Panel



biobank uk

Improving the health of future generations

500.000 individui con tanto di dati fenotipici e medici

Article

Sequence diversity lost in early pregnancy

deCODE genetics

https://doi.org/10.1038/s41586-025-09031-w
Received: 16 July 2024
Accepted: 16 April 2025
Published online: 21 May 2025

Gudny A. Arnadottir[1,19], Hakon Jonsson[1,19] ✉, Tanja Schlaikjær Hartwig[2,19], Jennifer R. Gruhn[3,19], Peter Loof Møller[4], Arnaldur Gylfason[1], David Westergaard[2], Andrew Chi-Ho Chan[3], Asmundur Oddsson[1], Lilja Stefansdottir[1], Louise le Roux[1], Valgerdur Steinthorsdottir[1], Kristjan H. Swerford Moore[1], Sigurgeir Olafsson[1], Pall I. Olason[1], Hannes P. Eggertsson[1], Gísli H. Halldórsson[1], G. Bragi Walters[1], Hreinn Stefansson[1], Sigurjon A. Gudjonsson[1], Gunnar Palsson[1], Brynjar O. Jensson[1], Run Fridriksdottir[1], Jesper Friis Petersen[3,5,6],

Scopo ultimo del corso: saper partire da sequenze di DNA fino ad assemblare interi genomi e caratterizzarli funzionalmente

Dal sequenziatore..

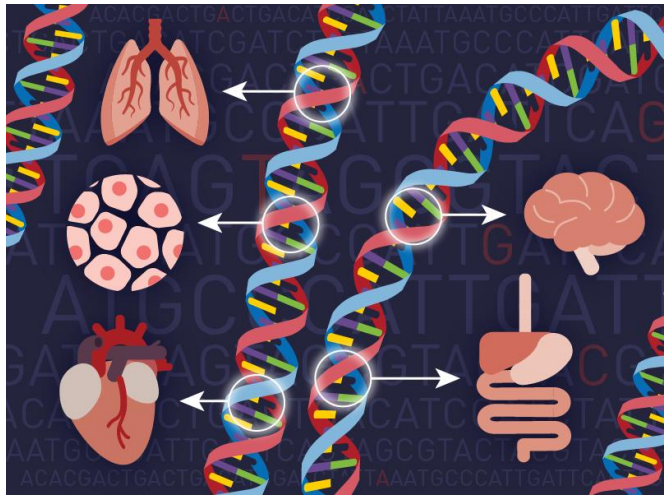
..al genoma..

..alla funzione

sequenze

assemblaggio

mappatura
e
genotipizzazione



Formati di file bioinformatici

.fastq **.fasta, .bam, .vcf** .bed, .gff

Errori di sequenziamento, mappatura e campionamento possono portare a errori di genotipizzazione

Genoma

di referenza

Non aver campionato/mappato
un allele verde diverso dalla referenza;
o un errore sistematico;
o una mappatura errata.

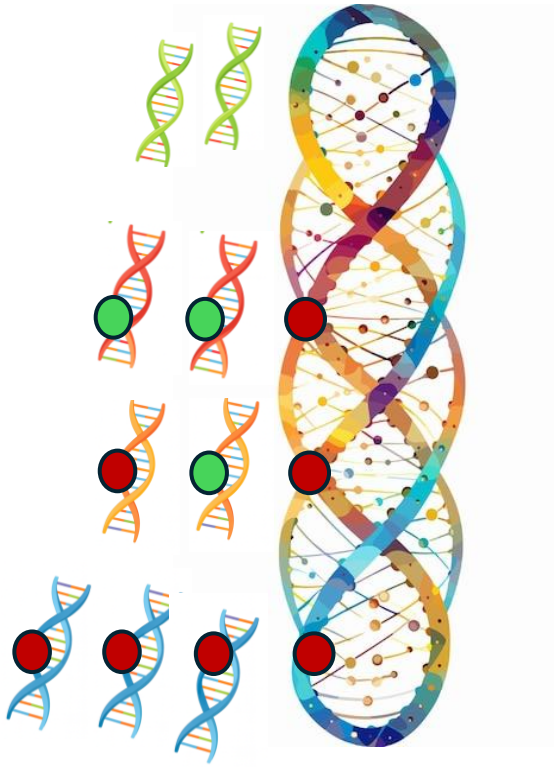
C C

Uno dei due (soprattutto il verde)
potrebbe essere un errore

A C

Non aver campionato/mappato
un allele verde diverso dalla referenza

A A



Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

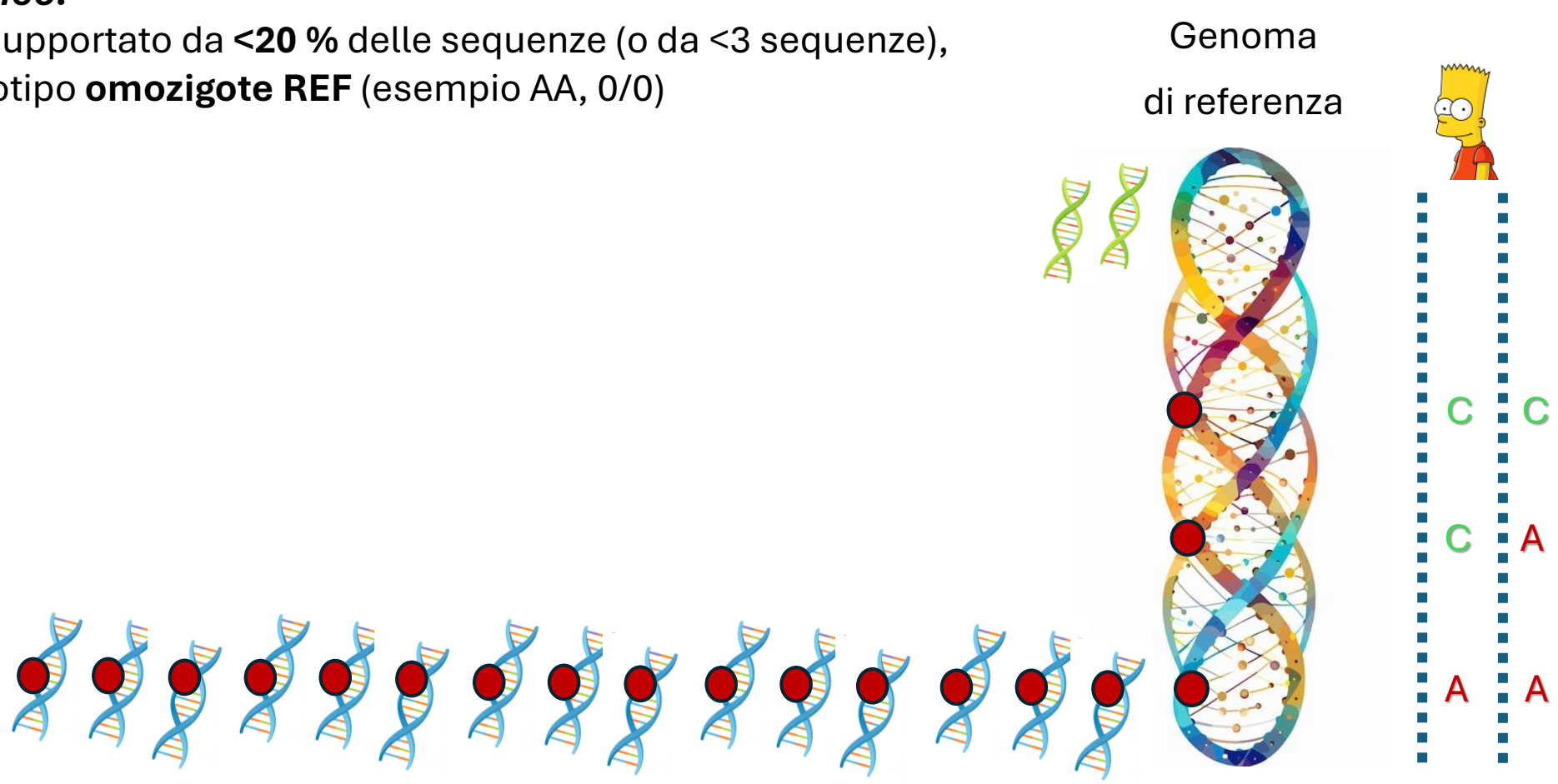
*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Un primo algoritmo euristico:

- Se l'allele alternativo è supportato da **<20 %** delle sequenze (o da <3 sequenze), allora considero il mio genotipo **omozigote REF** (esempio AA, 0/0)

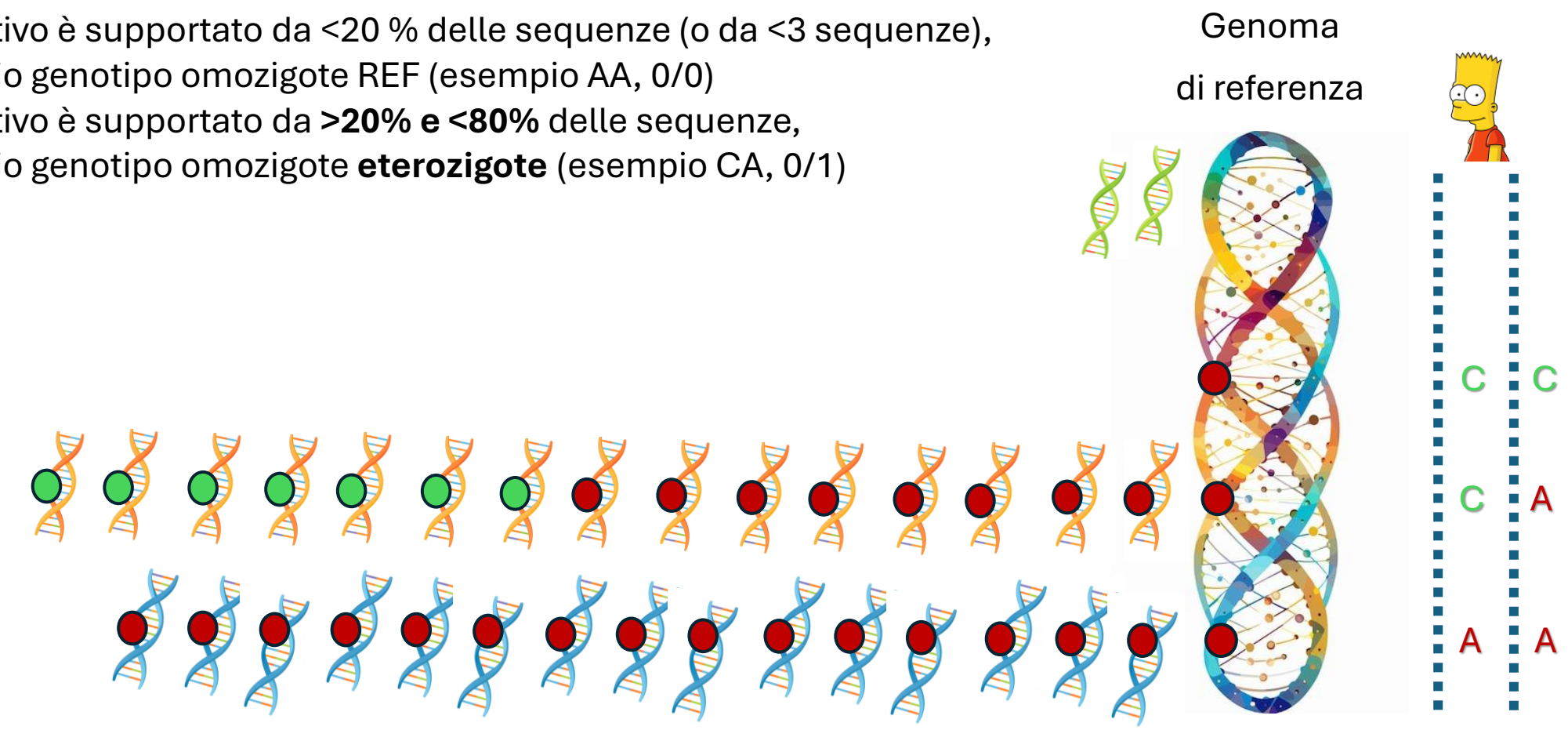


Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Un primo algoritmo euristico:

- Se l'allele alternativo è supportato da <20 % delle sequenze (o da <3 sequenze), allora considero il mio genotipo omozigote REF (esempio AA, 0/0)
- Se l'allele alternativo è supportato da **>20% e <80%** delle sequenze, allora considero il mio genotipo omozigote **eterozigote** (esempio CA, 0/1)

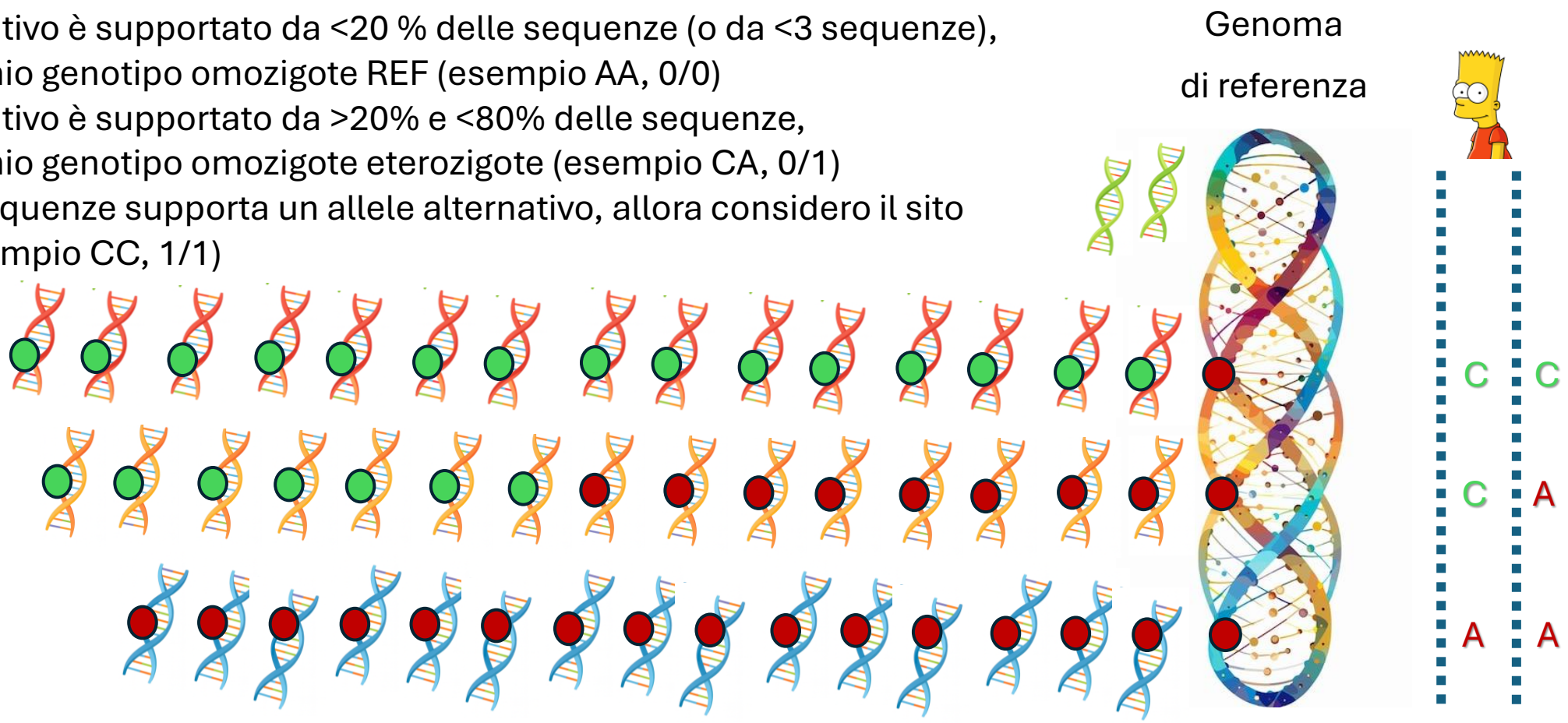


Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Un primo algoritmo euristico:

- Se l'allele alternativo è supportato da <20 % delle sequenze (o da <3 sequenze), allora considero il mio genotipo omozigote REF (esempio AA, 0/0)
- Se l'allele alternativo è supportato da >20% e <80% delle sequenze, allora considero il mio genotipo omozigote eterozigote (esempio CA, 0/1)
- Se **>80%** delle sequenze supporta un allele alternativo, allora considero il sito **omozigote ALT** (esempio CC, 1/1)

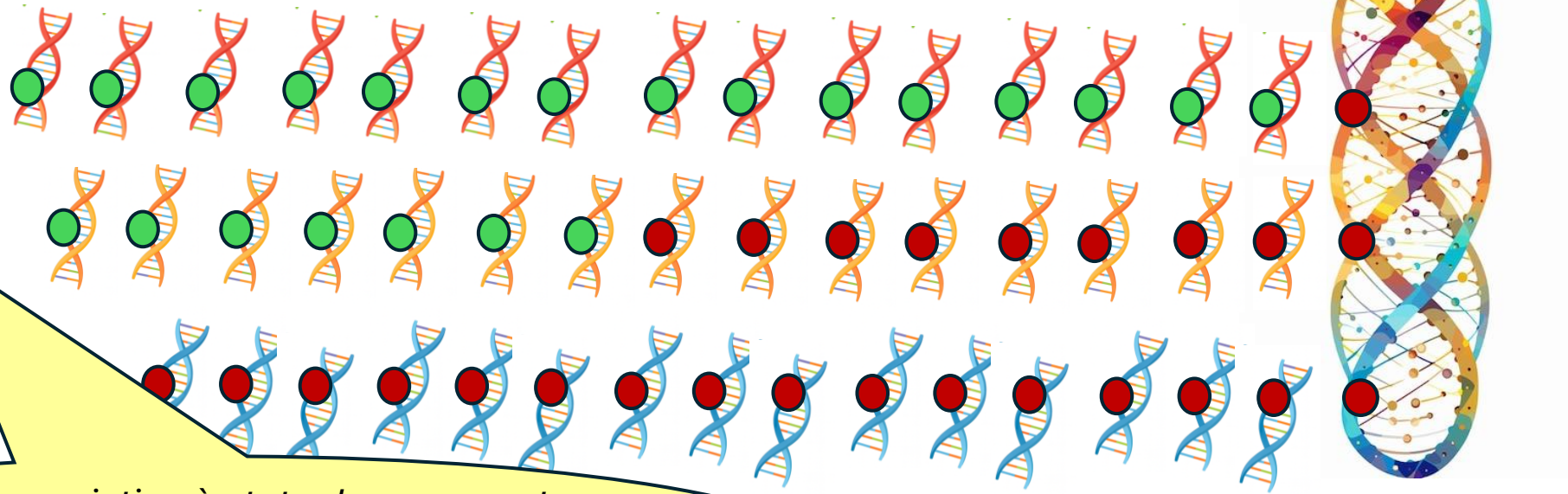


Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Un primo algoritmo euristico:

- Se l'allele alternativo è supportato da <20 % delle sequenze (o da <3 sequenze), allora considero il mio genotipo omozigote REF (esempio AA, 0/0)
- Se l'allele alternativo è supportato da >20% e <80% delle sequenze, allora considero il mio genotipo omozigote eterozigote (esempio CA, 0/1)
- Se **>80%** delle sequenze supporta un allele alternativo, allora considero il sito **omozigote ALT** (esempio CC, 1/1)



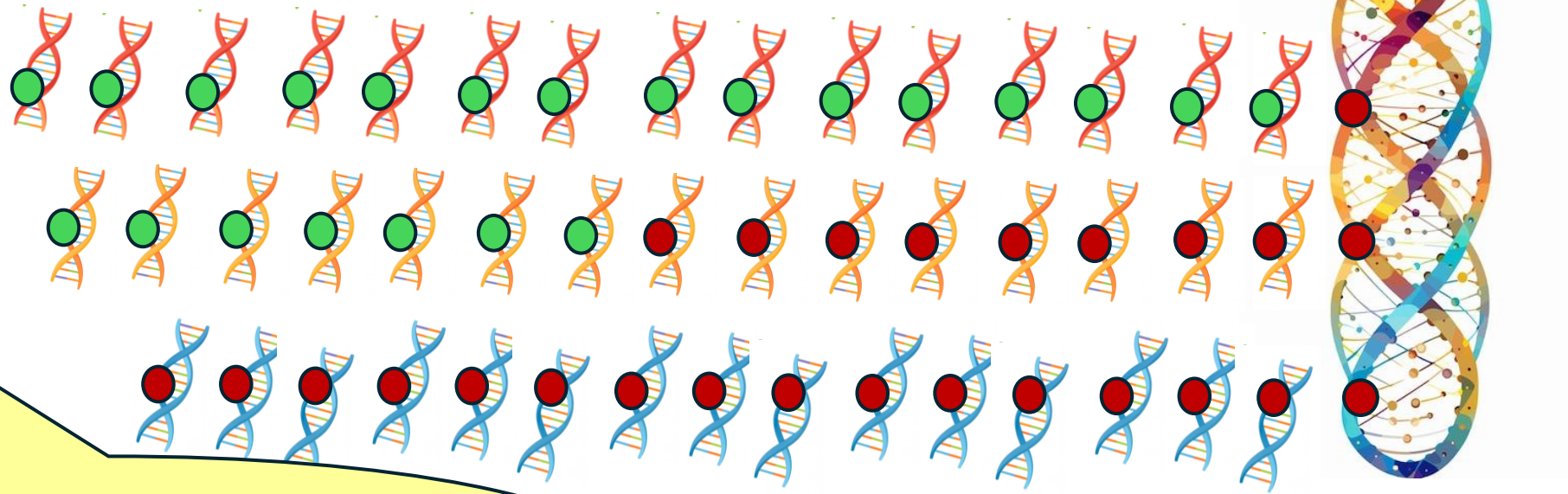
Questa euristica è stata davvero usata da alcuni software commerciali come CLC genomics. E funziona abbastanza bene quando coverage > 30x

Una prima soluzione: sequenziare tantissimo per avere alta coperatura*

*coperatura: cui spesso ci si riferisce anche come Coverage o Read Depth

Un primo algoritmo euristico:

- Se l'allele alternativo è supportato da <20 % delle sequenze (o da <3 sequenze), allora considero il mio genotipo omozigote REF (esempio AA, 0/0)
- Se l'allele alternativo è supportato da >20% e <80% delle sequenze, allora considero il mio genotipo omozigote eterozigote (esempio CA, 0/1)
- Se **>80%** delle sequenze supporta un allele alternativo, allora considero il sito **omozigote ALT** (esempio CC, 1/1)



Quest'approccio funziona piu' o meno bene con altissima coverage, ma ciuccia il calzino in condizioni realistiche

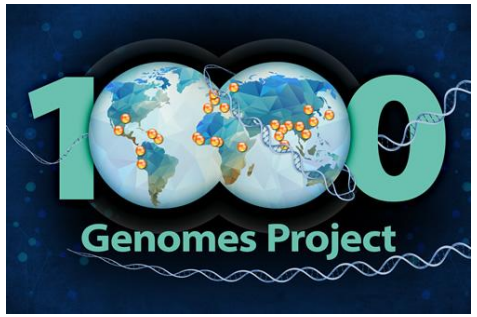
In moltissimi contesti/studi si ha coverage effettiva ben piu' bassa (spesso 2-10x)

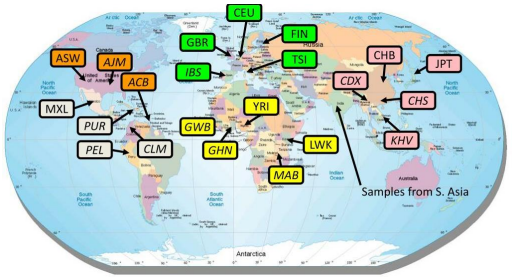
Perché?

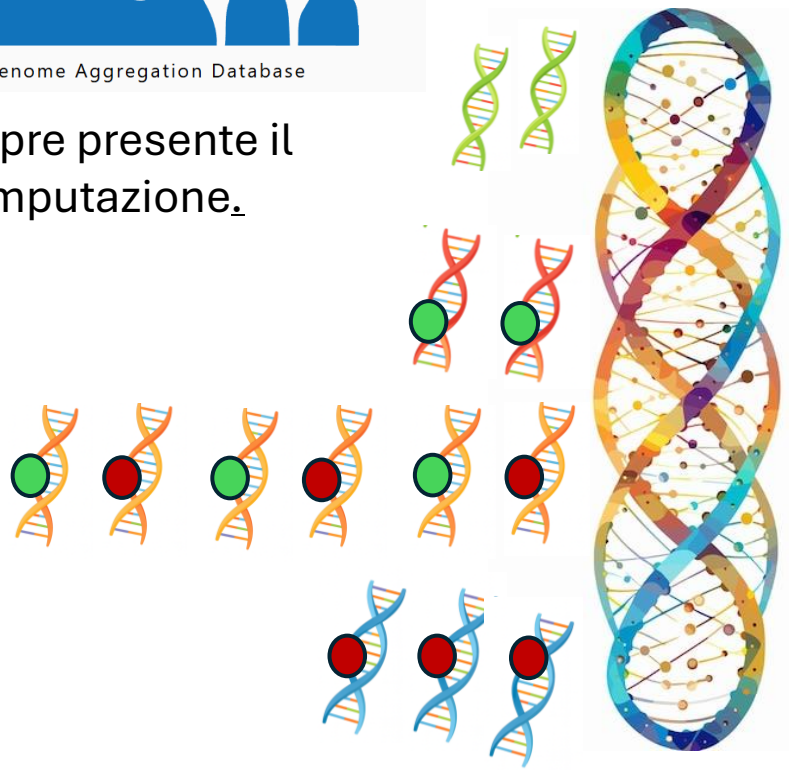
biobank uk
Improving the health of future generations

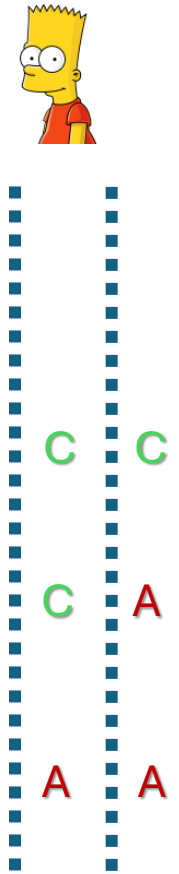
gnomAD
Genome Aggregation Database

- Anche se il costo del sequenziamento si abbatte, sempre presente il ***trade-off*** *piu' individui versus piu qualità.* Tutt'oggi forte imputazione.









In moltissimi contesti/studi si ha coverage effettiva ben piu' bassa (spesso 2-10x)

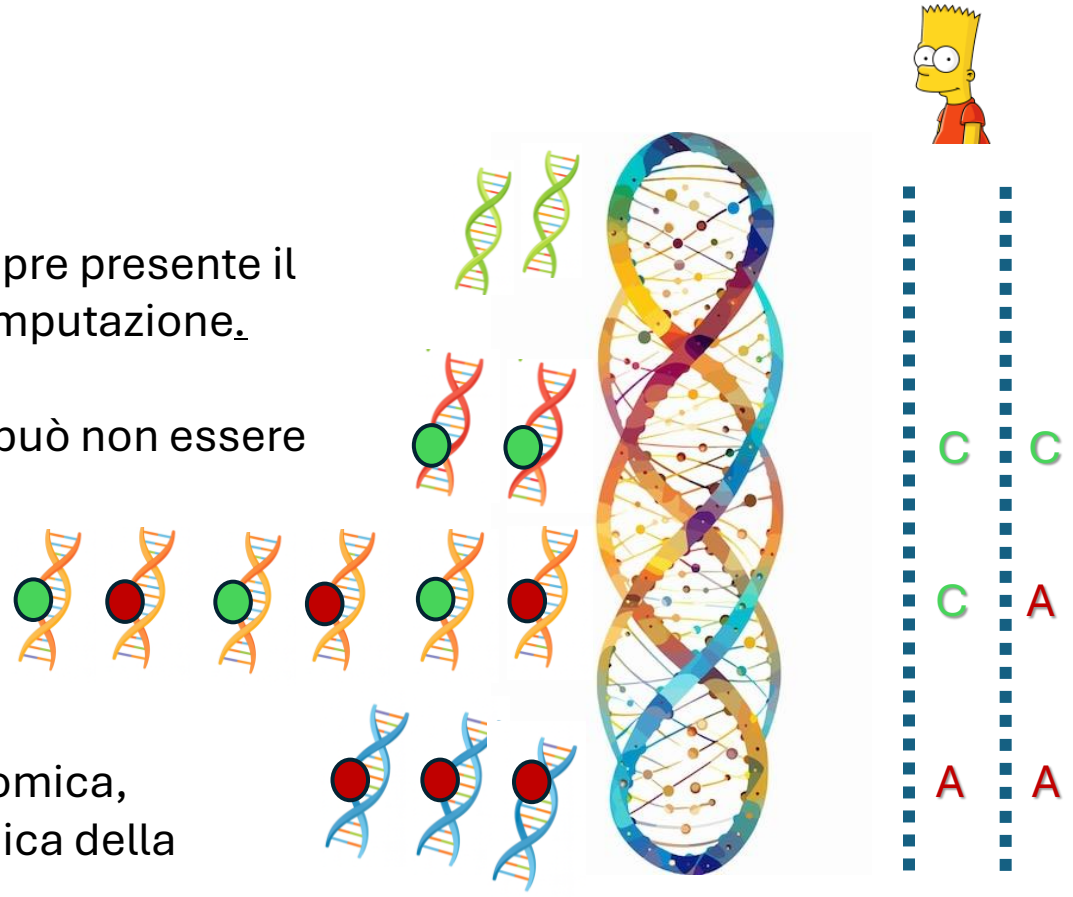
Perché?

- Anche se il costo del sequenziamento si abbatte, sempre presente il ***trade-off*** *piu' individui versus piu qualità.* Tutt'oggi forte imputazione.

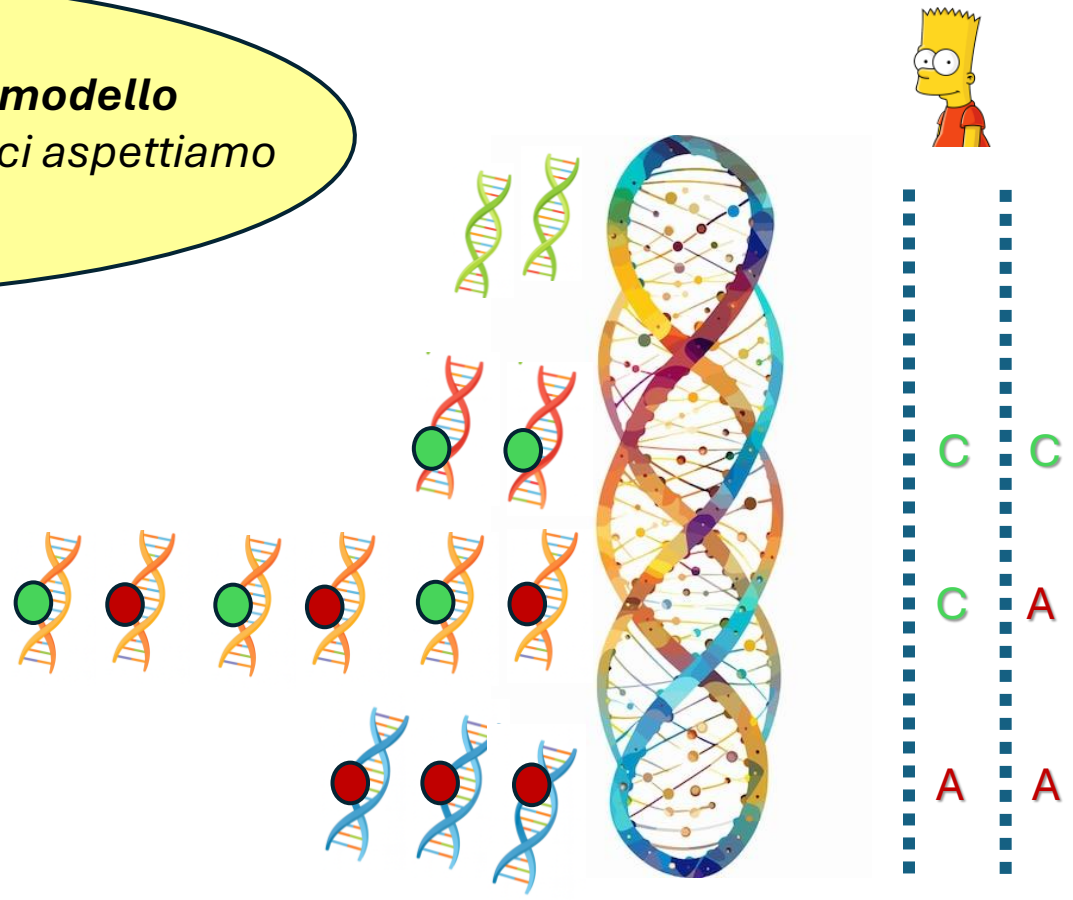
- Per zone complesse e varianti strutturali, anche 100x può non essere sufficiente per ricorrere ad euristiche

- Long-reads hanno generalmente coverage piu' bassa

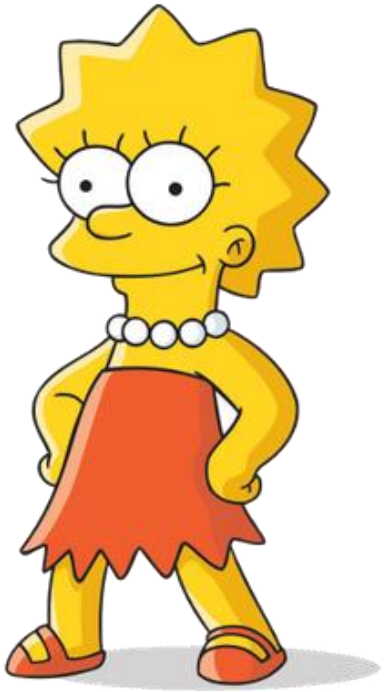
- Alcuni ambiti sono inerentemente esposti a limiti nel materiale (DNA antico, single cell sequencing, metagenomica, mosaicismo) o nelle risorse per sequenziamento (genomica della conservazione, trio sequencing per malattie rare)



In moltissimi contesti/studi si ha coverage effettiva ben piu' bassa (spesso 2-10x)

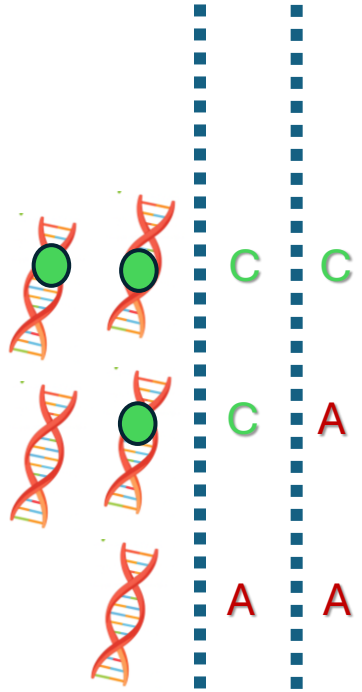


Un modello probabilistico di genotipi molto semplicistico



Probabilità (Pr) di osservare una read 

- da un genotipo C C = 1
- da un genotipo C A = 1/2
- da un genotipo A A = 0

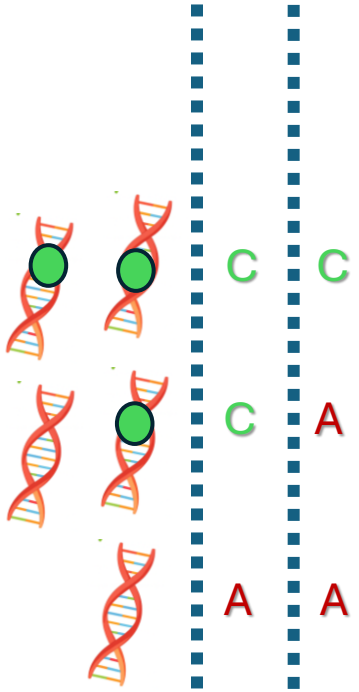


Un po’ di terminologia: la probabilità condizionata

Queste sono le **probabilità condizionate** $Pr(d_i|g_i)$, dove d sono le sequenze osservate e g i genotipi sul sito i.
Esempio: Pr(C|CA)=1/2

Probabilità (Pr) di osservare una read 

- da un genotipo C C = 1
- da un genotipo C A = 1/2
- da un genotipo A A = 0



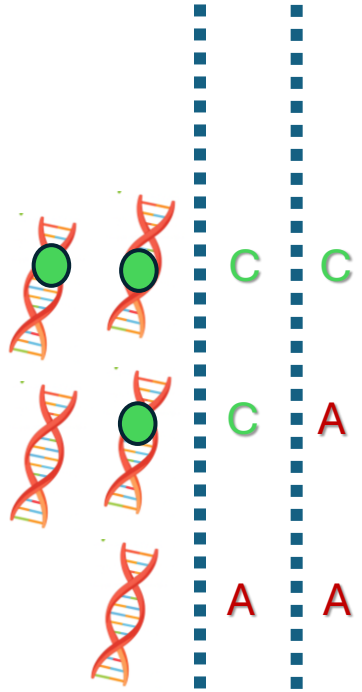
Un po' di terminologia: la probabilità condizionata

La probabilità di un evento ***y*** (esempio: che nevichi) condizionata ad ***x*** (esempio: 0°C) non è altro la probabilità di Y= ***y*** dato che X=***x*** sia accaduto. Ad esempio la probabilità totale che nevichi può essere bassa, ma se condizionata ad ***x*** può essere ben piu' alta.

Queste sono le **probabilità condizionate** $Pr(d_i|g_i)$, dove *d* sono le sequenze osservate e *g* i genotipi sul sito *i*.
Esempio: Pr(C|CA)=1/2

Probabilità (Pr) di osservare una read 

- da un genotipo C C = 1
- da un genotipo C A = 1/2
- da un genotipo A A = 0



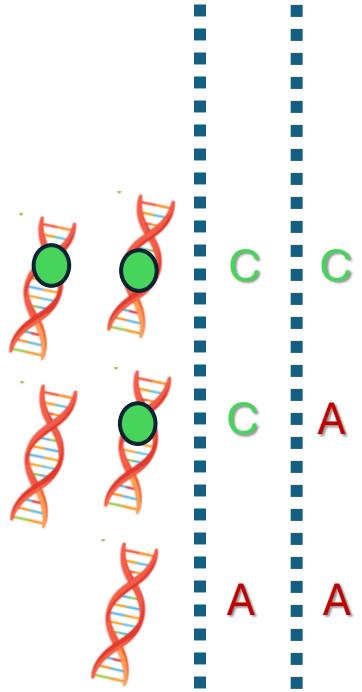
Un po' di terminologia: la probabilità condizionata

La probabilità di un evento ***y*** (esempio: che nevichi) condizionata ad ***x*** (esempio: 0°C) non è altro la probabilità di Y= ***y*** dato che X=***x*** sia accaduto. Ad esempio la probabilità totale che nevichi può essere bassa, ma se condizionata ad ***x*** può essere ben piu' alta.
Definiamo la probabilità «generale» che nevichi come la **probabilità totale**.

Queste sono le **probabilità condizionate** $Pr(d_i|g_i)$, dove *d* sono le sequenze osservate e *g* i genotipi sul sito *i*.
Esempio: Pr(C|CA)=1/2

Probabilità (Pr) di osservare una read

- da un genotipo C C = 1
- da un genotipo C A = 1/2
- da un genotipo A A = 0



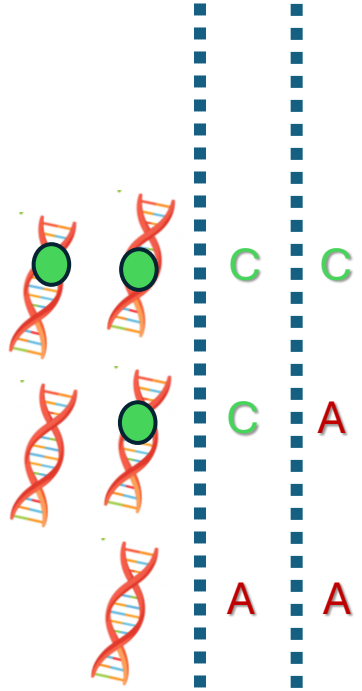
Un po' di terminologia: la probabilità condizionata

Notate la differenza tra probabilità condizionata $Pr(d_i|g_i)$ e **probabilità incondizionata** $Pr(d_i)$, che può essere ottenuta tramite la **legge della proabilità totale**: $Pr(d_i)=\sum_i Pr(di|gi)$
Esempio: $Pr(C)=Pr(C|CA)Pr(CA)+Pr(C|AA)Pr(AA)+Pr(C|CC)Pr(CC)+...$

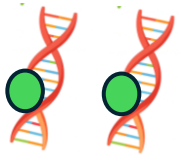
Queste sono le **probabilità condizionate** $Pr(d_i|g_i)$, dove *d* sono le sequenze osservate e *g* i genotipi sul sito *i*.
Esempio: Pr(C|CA)=1/2

Probabilità (Pr) di osservare una read 

- da un genotipo C C = 1
- da un genotipo C A = 1/2
- da un genotipo A A = 0



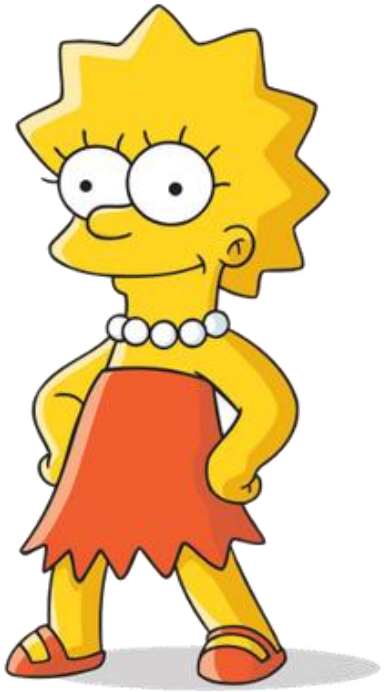
Un modello probabilistico di genotipi molto semplicistico

Probabilità (Pr) di osservare due reads 

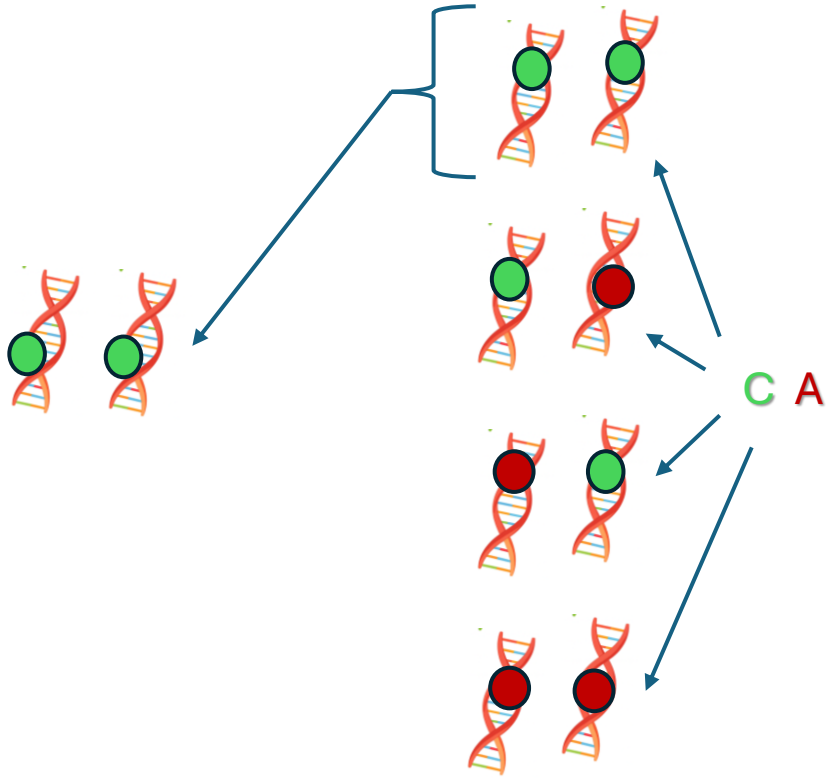
- da un genotipo C C = 1

- da un genotipo A A = 0

Un modello probabilistico di genotipi molto semplicistico

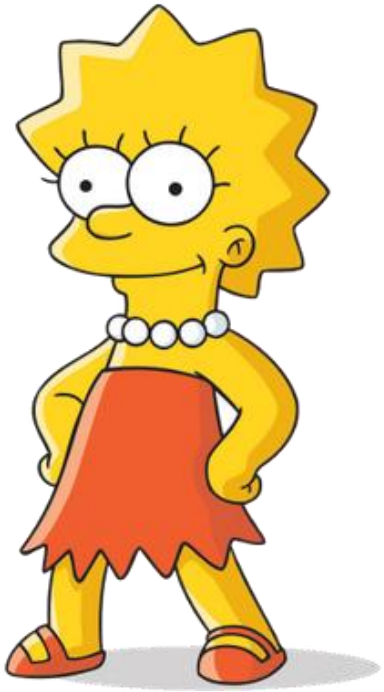


Probabilità (Pr) di osservare due reads

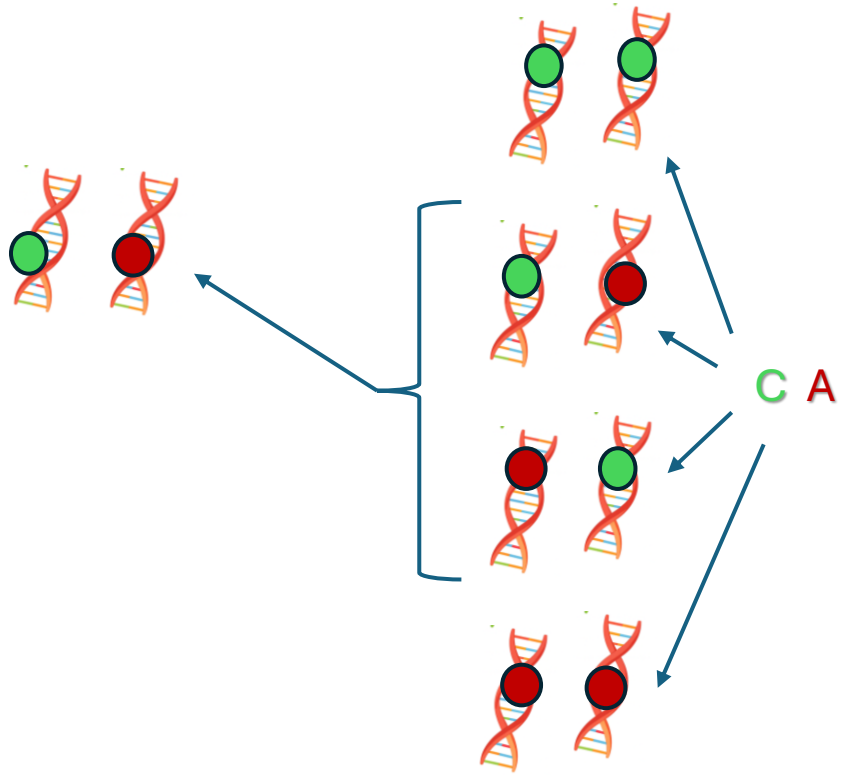


- da un genotipo C C = 1
- da un genotipo C A = ½ * ½=1/4
- da un genotipo A A = 0

Un modello probabilistico di genotipi molto semplicistico

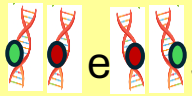


Probabilità (Pr) di osservare due reads (ignorandone l'ordine)



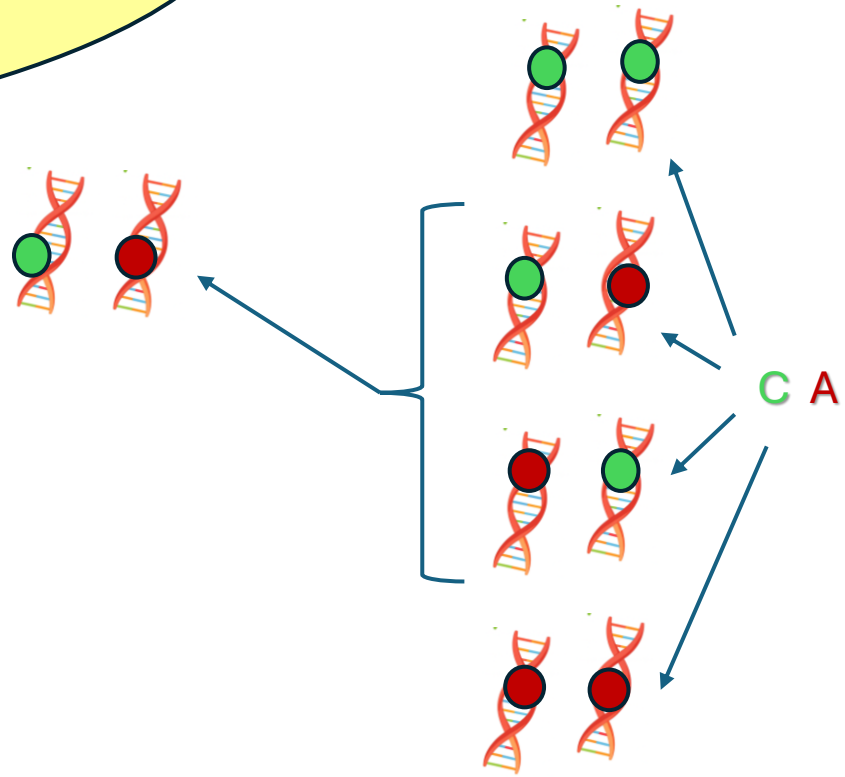
- da un genotipo C A = ½ * ½*2=1/2

Un modello probabilistico di genotipi molto semplicistico

Notate che ottenere due reads  e sono due (micro)eventi distinti, ognuno con la stessa probabilità ½*½=1/4.

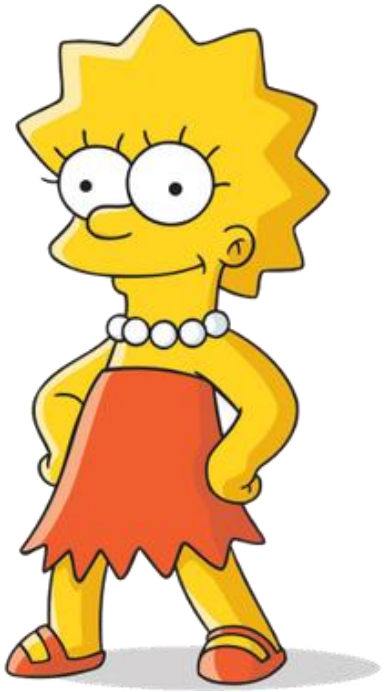
Tuttavia le nostre sequenze sono state sequenziate insieme da un pool di DNA molto grande, pertanto «biologicamente» i due micro(eventi) sono indistinguibili e rappresentano lo stesso evento: due reads di cui una è una **C** e l'altra una **A**.

Probabilità (Pr) di osservare due reads **(ignorandone l'ordine)**

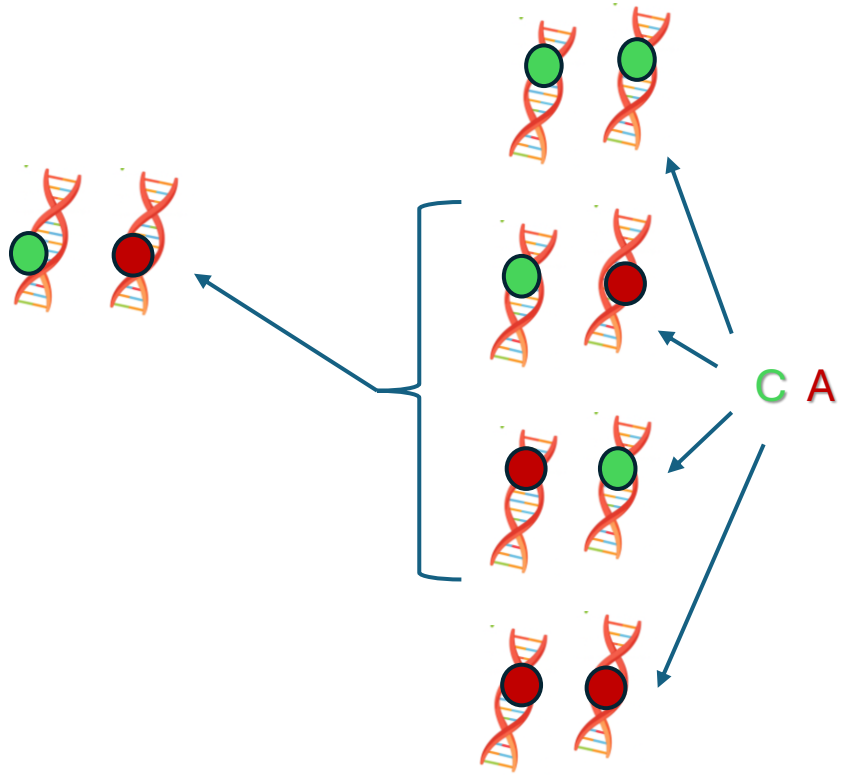


- da un genotipo C A = ½ * ½*2=1/2

Un modello probabilistico di genotipi molto semplicistico

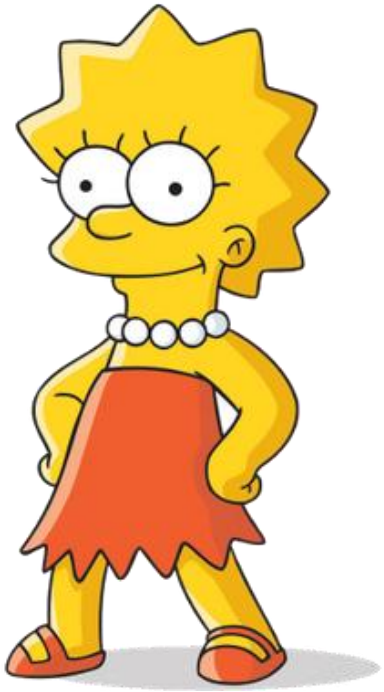


Probabilità (Pr) di osservare due reads (ignorandone l'ordine)



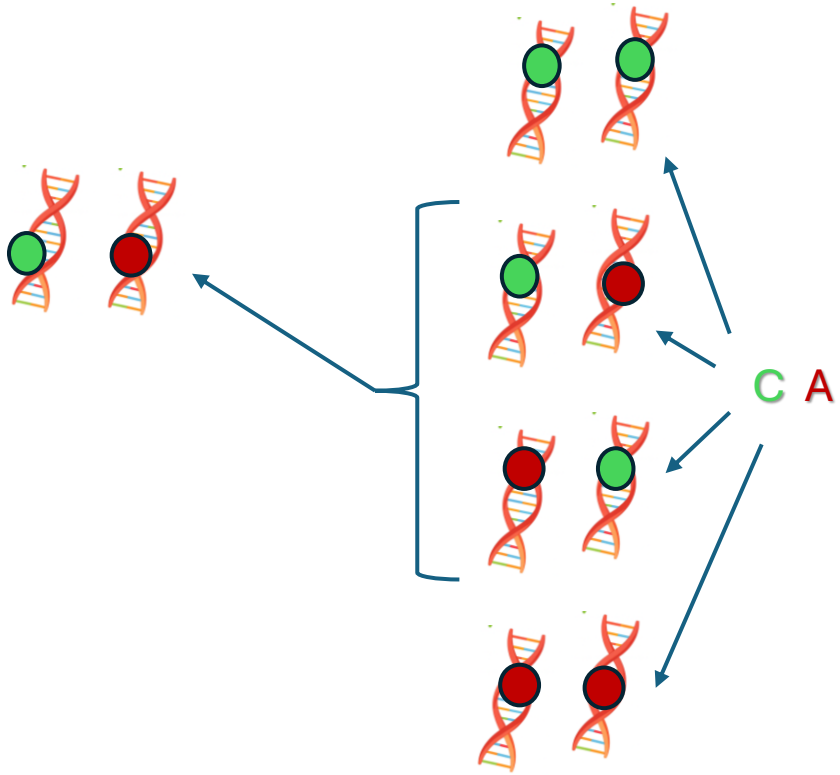
- da un genotipo C A = ½ * ½*2=1/2

Un modello probabilistico di genotipi molto semplicistico



Probabilità (Pr) di osservare due reads (ignorandone l'ordine)

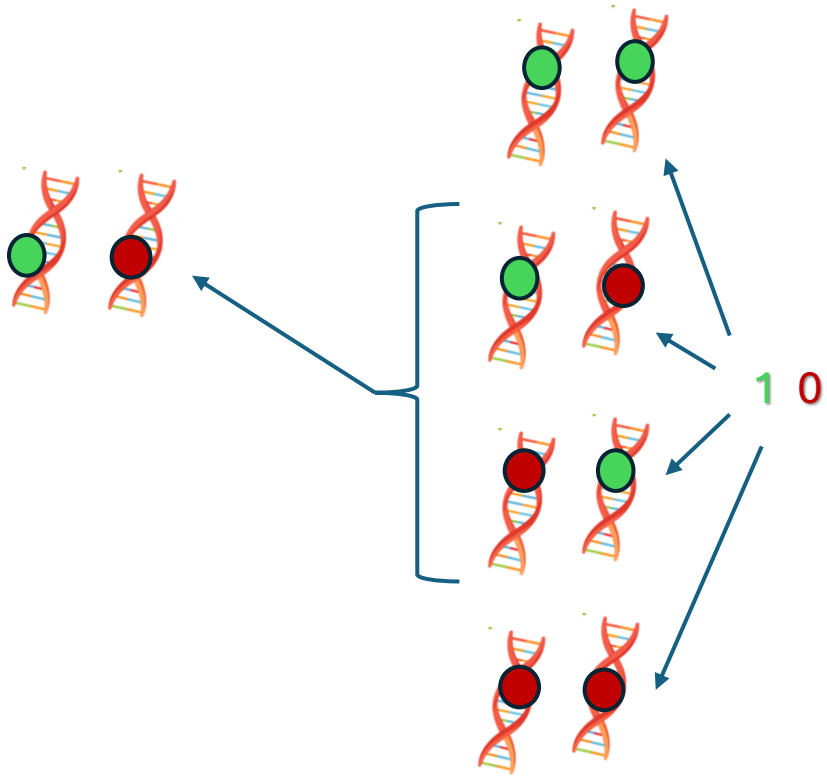
- da un genotipo C C = 0
- da un genotipo C A = ½ * ½*2=1/2
- da un genotipo A A = 0



Generalizziamo usando la notazione VCF di 1 per alternativo e 0 per referenza

Probabilità (Pr) di osservare due reads (ignorandone l'ordine)

- da un genotipo 1 1 = 0
- da un genotipo 1 0 = ½ * ½*2
- da un genotipo 0 0 = 0



La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

La distribuzione binomiale

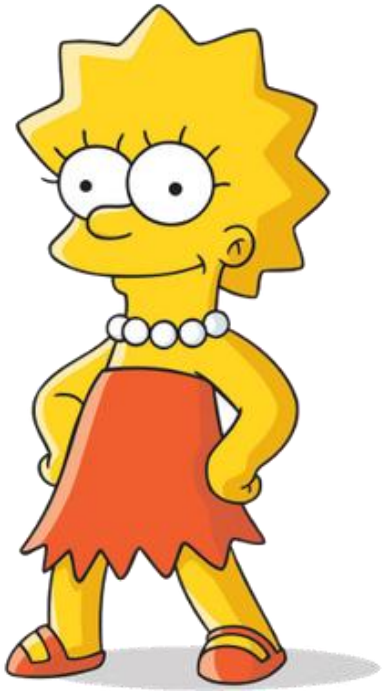
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

Fattoriali
n!=n*(n-1)*(n-2)....*1
Esempio:
3!=3*2=6



$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

dove $\binom{n}{k}=n!/(k!(n-k)!)$ è il coefficiente binomiale e descrive il numero di possibili combinazioni di alleli ALT e REF

La distribuzione binomiale

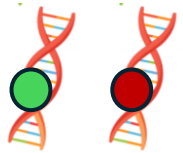
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

Fattoriali
n!=n*(n-1)*(n-2)....*1
Esempi:
3!=3*2*1=6
5!=5*4*3*2*1=120

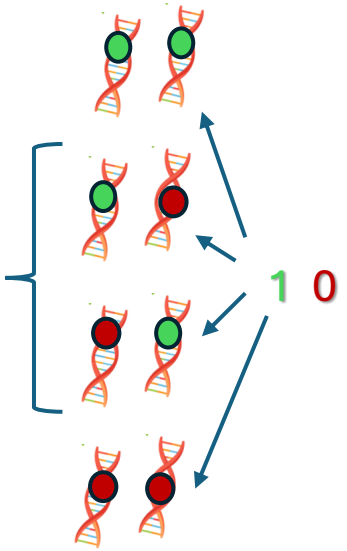
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

dove $\binom{n}{k}=n!/(k!(n-k)!)$ è il coefficiente binomiale e descrive il numero di possibili combinazioni di alleli ALT e REF

Esempio: k=1 e n=2



2 combinazioni visto che $\binom{2}{1}$ = 2!/(1!1!)=2



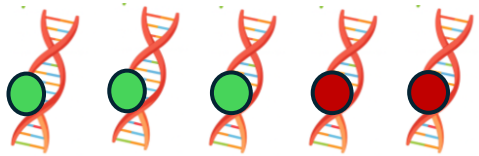
La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

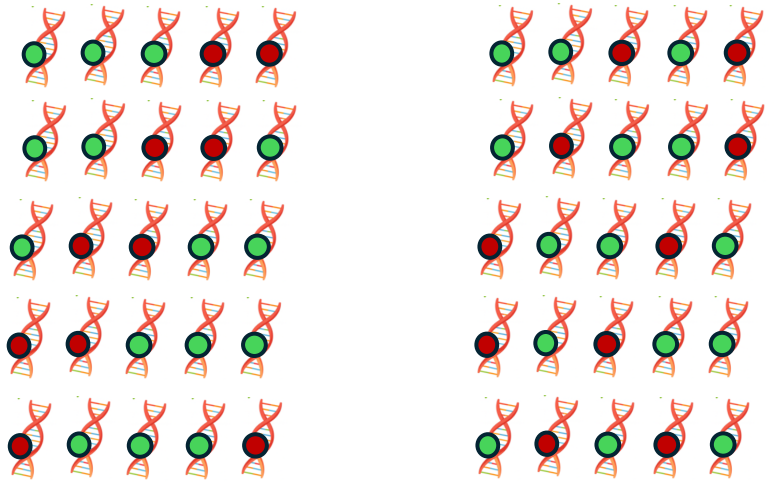
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti



dove $\binom{n}{k}=n!/(k!(n-k)!)$ è il coefficiente binomiale e descrive il numero di possibili combinazioni di alleli ALT e REF

10 combinazioni visto che $\binom{5}{3} = 5!/(3!*2!)=5*4/2!=10$

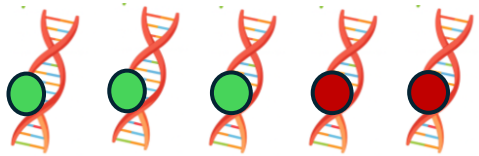
La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

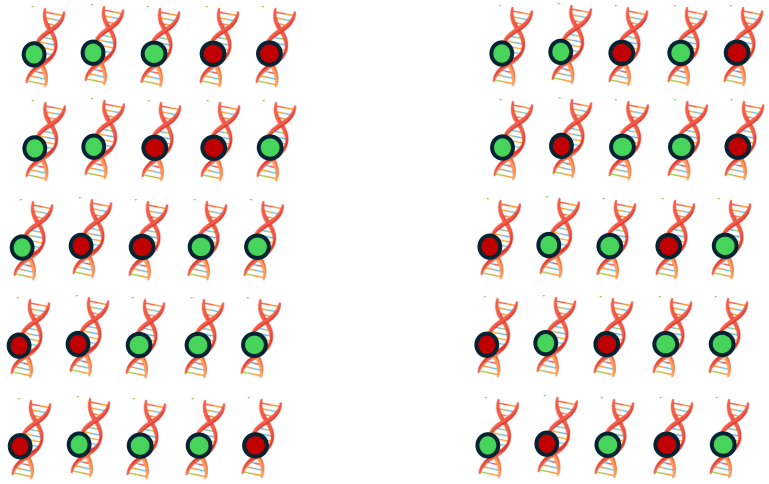
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



La distribuzione binomiale

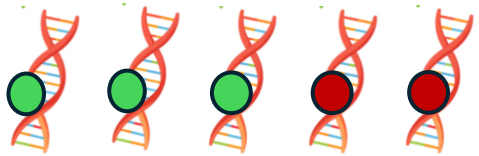
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

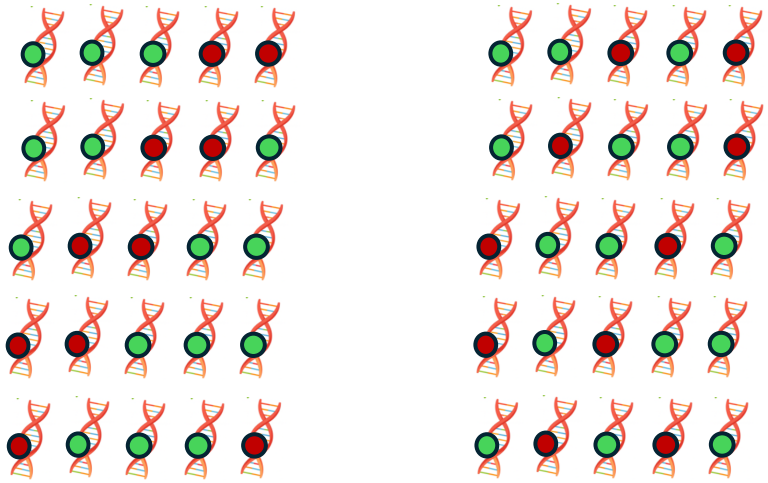
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



→ p*p*(1-p)*p*(1-p)

La distribuzione binomiale

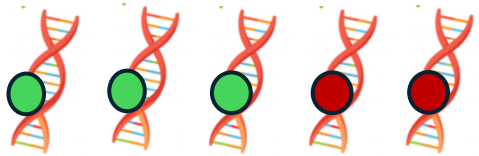
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

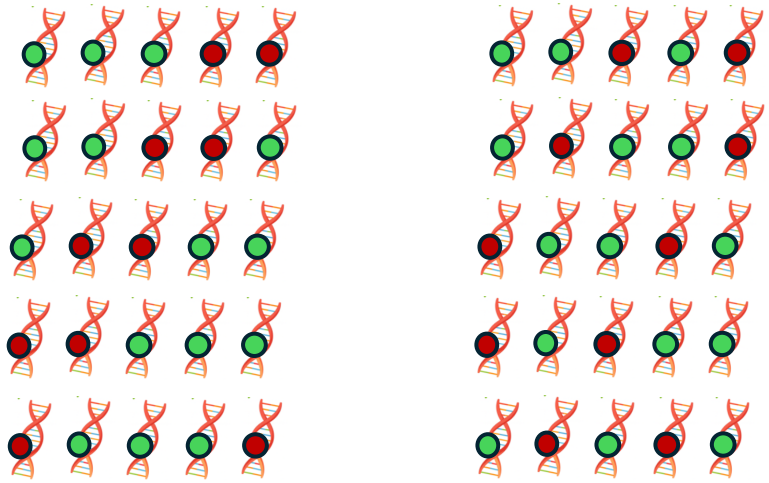
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



→ $p*p*(1-p)*p*(1-p) = p^3(1-p)^5$

La distribuzione binomiale

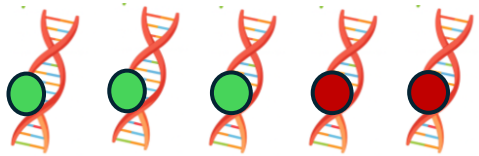
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

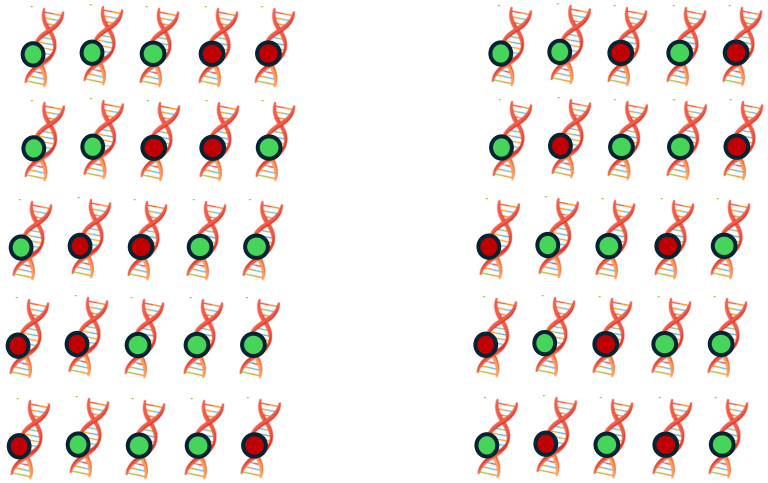
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



→ $p*p*(1-p)*p*(1-p) = p^3* (1-p)^5$

→ $p*(1-p)*p*p*(1-p)$

La distribuzione binomiale

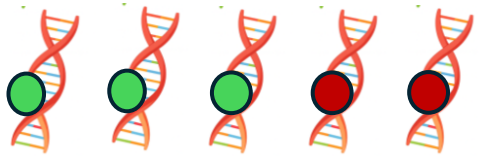
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

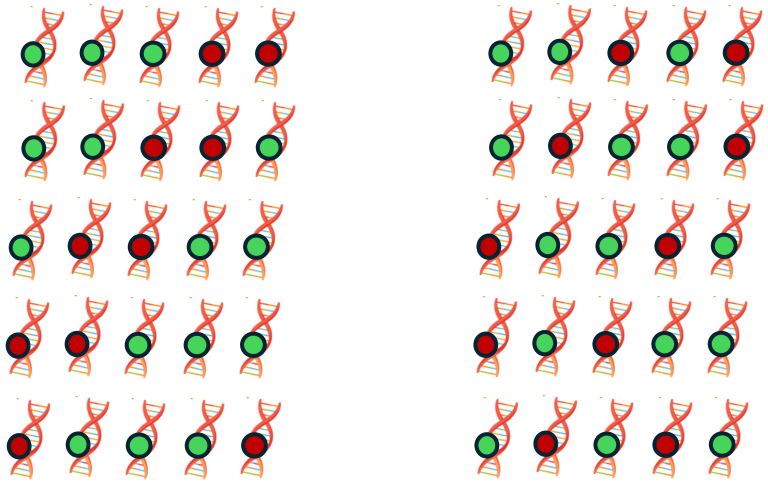
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



→ $p*p*(1-p)*p*(1-p) = p^3* (1-p)^5$

→ $p*(1-p)*p*p*(1-p) = p^3* (1-p)^5$

La distribuzione binomiale

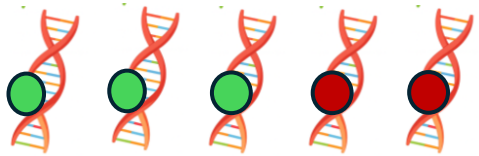
La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

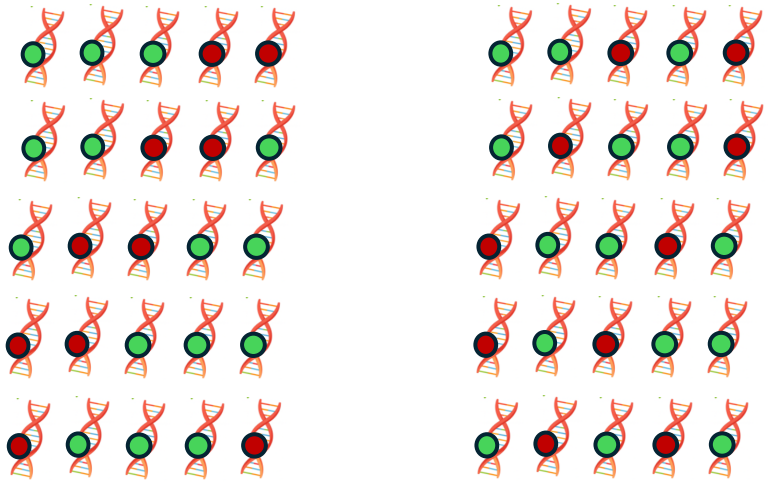
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: k=3 e n=5

reads osservate



Possibili combinazioni equivalenti ed equiprobabili



→ $p*p*(1-p)*p*(1-p) = p^3* (1-p)^5$

→ $p*(1-p)*p*p*(1-p) = p^3* (1-p)^5$

→ $(1-p)*p*p*(1-p)*p = p^3* (1-p)^5$

→ $(1-p)*p*(1-p)*p*p = p^3* (1-p)^5$

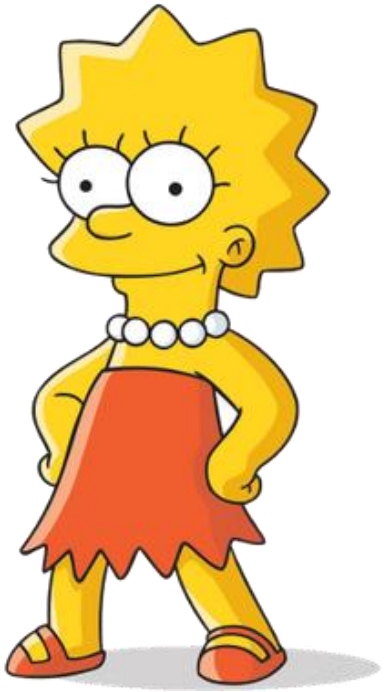
→ $p*(1-p)*p*(1-p)*p = p^3* (1-p)^5$

La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

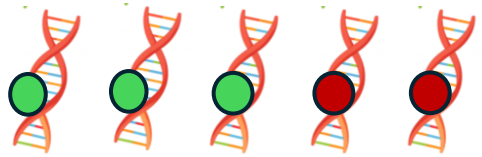
p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$



Esempio: k=3 e n=5:

reads osservate



Se il genotipo è **0/0** (quindi p=0):
p*p*p*(1-p)*(1-p)=0*0*0*1*1=0

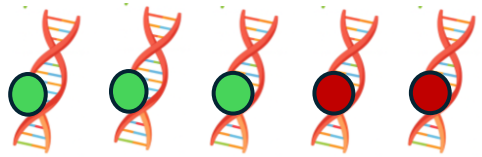
La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

reads osservate



Esempio: k=3 e n=5:

Se il genotipo è **0/0** (quindi p=0):
p*p*p*(1-p)*(1-p)=0*0*0*1*1=0

Se il genotipo è **1/1** (quindi p=0):
p*p*p*(1-p)*(1-p)=1*1*1*0*0=0

Se il genotipo è **0/1** e p=1/2:
p*p*p*(1-p)*(1-p)=1/2^5=1/32

La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

p è la probabilità di osservare alleli alternativi (o 1) a partire dal nostro genotipo

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

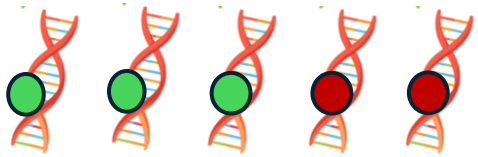
Esempio: k=3 e n=5:

Se il genotipo è **0/0**:
p*p*p*(1-p)*(1-p)=0*0*0*1*1=0

Se il genotipo è **1/1**:
p*p*p*(1-p)*(1-p)=1*1*1*0*0=0

Se il genotipo è **0/1** e p=1/2:
p*p*p*(1-p)*(1-p)=1/2^5=1/32

reads osservate



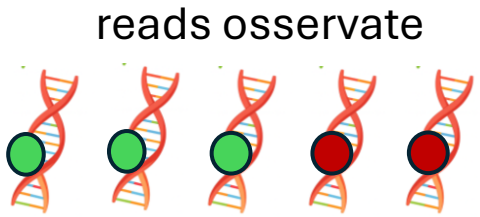
Possibile in presenza di *reference bias*: quando l'allele alternativo mappa meno efficientemente a causa delle differenze con la referenza

Se il genotipo è **0/1** e p=0.48:
0.48^3*0.52^2≃1/33.5

La distribuzione binomiale

La distribuzione binomiale descrive la probabilità di osservare numero di successi *k* su un totale *n* di eventi

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$



Esempio: k=3 e n=5:

Se il genotipo è **0/1** e p=1/2:

$$Pr(3)=\binom{5}{3} p^3 (1-p)^{5-3} = 0.3125$$

10 combinazioni

1/32

La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

p per i vari genotipi

Pr(1| 0/0)=ε

Pr(1| 0/1)=1/2

Pr(1| 1/1)=1-ε

La probabilità di un genotipo in presenza di errori

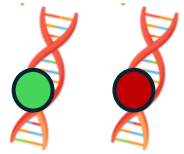
Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

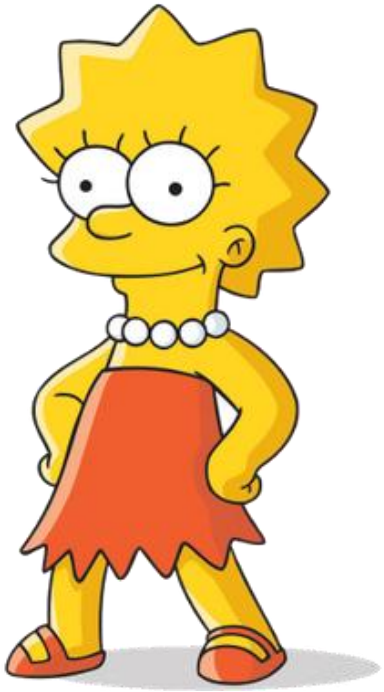
$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

p per i vari genotipi

Pr(1| 0/0)=ε
Pr(1| 0/1)=1/2
Pr(1| 1/1)=1-ε

reads osservate





Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

$$Pr(k)=\binom{n}{k} p^k (1-p)^{n-k}$$

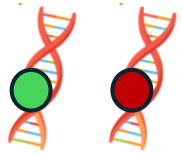
p per i vari genotipi

Pr(1| 0/0)=ε

Pr(1| 0/1)=1/2

Pr(1| 1/1)=1-ε

reads osservate



Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

Pr(k=1,n=2|0/1)=0.5

Pr(k=1,n=2|0/0)=0.0198

Pr(k=1,n=2|1/1)=0.0198

La probabilità di un genotipo in presenza di errori

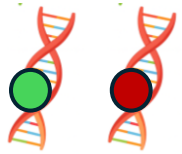
Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

Cosa ci può dire questo sul nostro rischio di commettere errori nel chiamare i genotipi?

Notate che abbiamo delle **probabilità condizionate** ai vari genotipi. Per trovare la probabilità totale dobbiamo quindi fare sapere quanti genotipi 0/0, 0/1 e 1/1 abbiamo. Assumiamo che il 99% dei siti siano 0/0 e l'1% sia 0/1.

Esempio: k=1, n=2 e ε=0.01:

reads osservate



Se il genotipo è:

$Pr(k=1,n=2|0/1)=0.5$

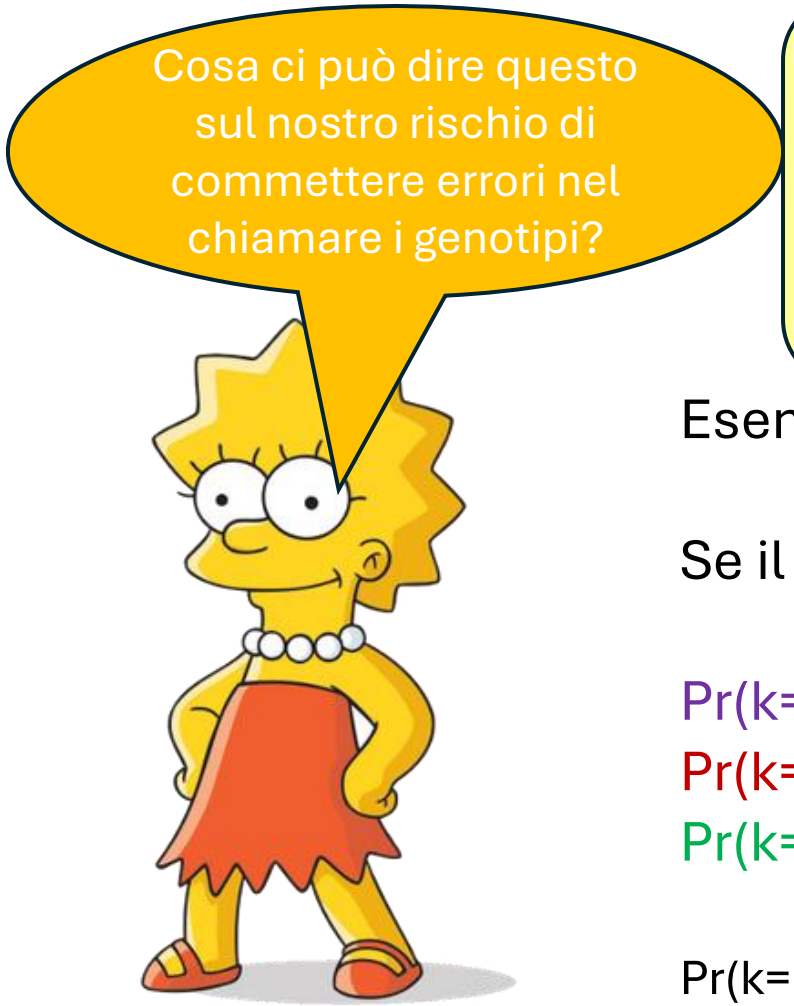
$Pr(k=1,n=2|0/0)=0.0198$

$Pr(k=1,n=2|1/1)=0.0198$

$Pr(k=1,n=2)=Pr(k=1,n=2|0/1)Pr(0/1)+Pr(k=1,n=2|0/0)Pr(0/0)+Pr(k=1,n=2|1/1)Pr(1/1)=$

La probabilità di un genotipo in presenza di errori

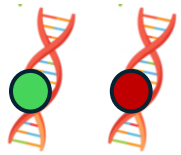
Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro



Notate che abbiamo delle **probabilità condizionate** ai vari genotipi. Per trovare la probabilità totale dobbiamo quindi fare sapere quanti genotipi 0/0, 0/1 e 1/1 abbiamo. Assumiamo che il 99% dei siti siano 0/0 e l'1% sia 0/1.

Esempio: k=1, n=2 e ε=0.01:

reads osservate



Se il genotipo è:

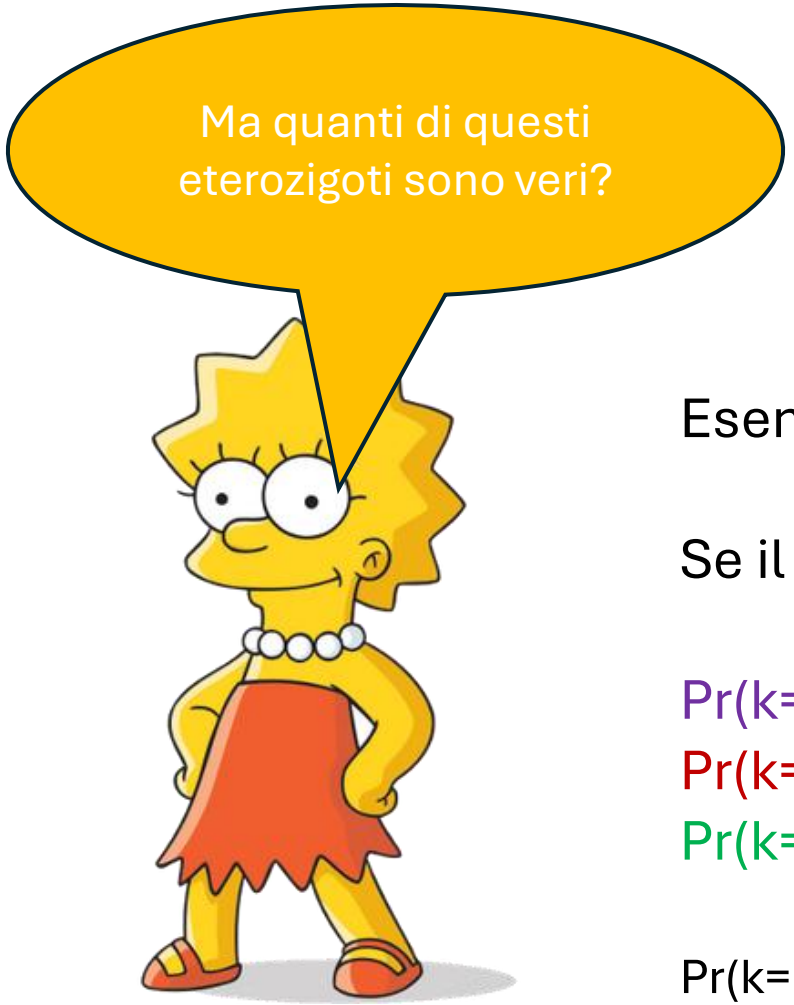
Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

	0.5	0.01	0.0198	0.99	0.0198	0
Pr(k=1,n=2)=	Pr(k=1,n=2|0/1)	Pr(0/1)+	Pr(k=1,n=2|0/0)	Pr(0/0)+	Pr(k=1,n=2|1/1)	Pr(1/1)=

=0.024602 Probabilità totale di vedere un eterozigote

La probabilità di un genotipo in presenza di errori

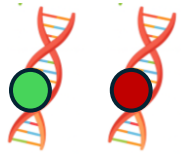
Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro



Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

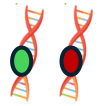
reads osservate



Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

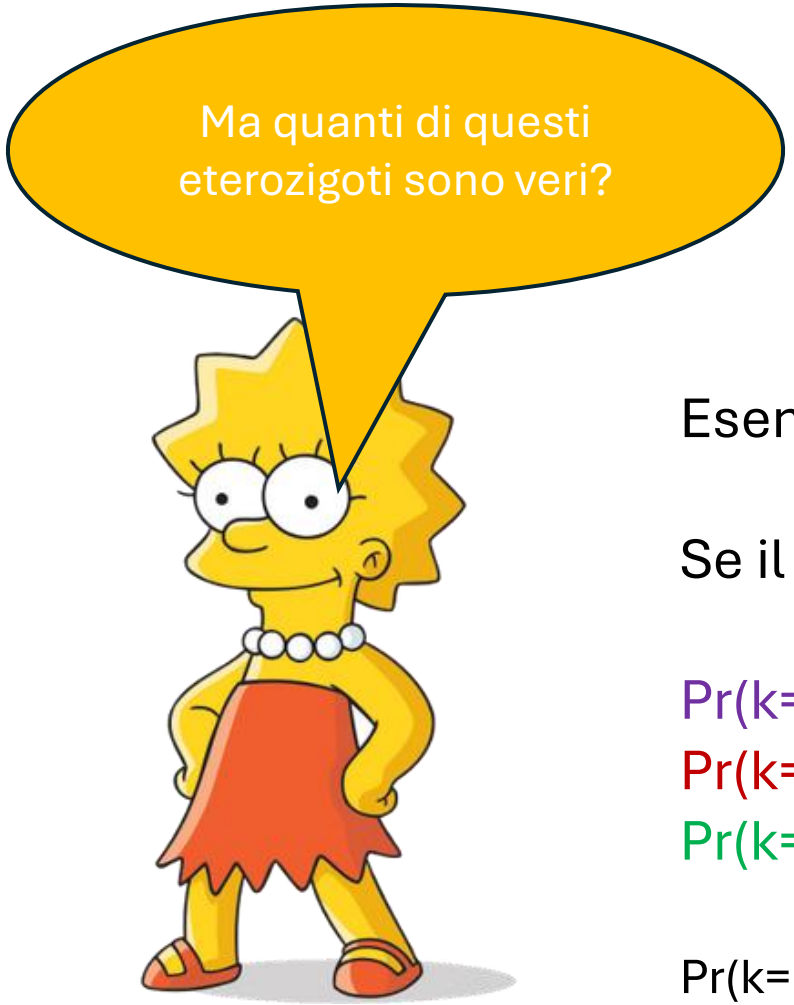
	0.5	0.01	0.0198	0.99	0.0198	0
Pr(k=1,n=2)=	Pr(k=1,n=2|0/1)	Pr(0/1)+	Pr(k=1,n=2|0/0)	Pr(0/0)+	Pr(k=1,n=2|1/1)	Pr(1/1)=

=0.024602 Probabilità totale di vedere un eterozigote



La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro



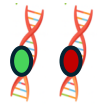
Esempio: k=1, n=2 e ε=0.01:

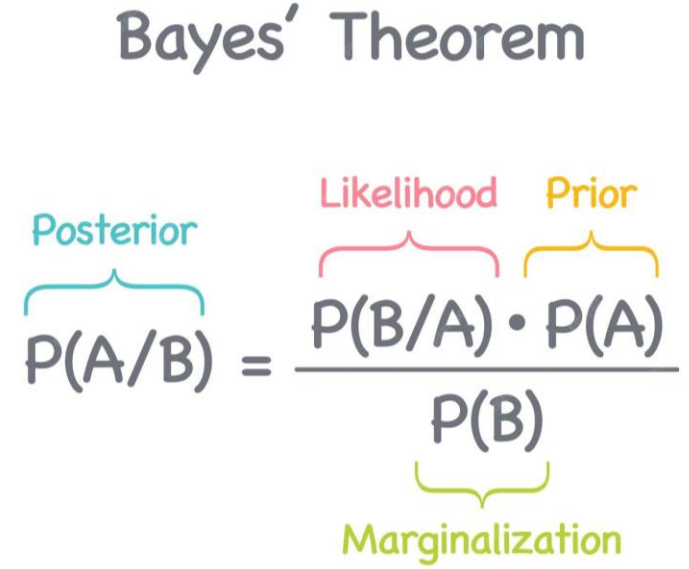
Se il genotipo è:

Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

$$\Pr(k=1,n=2) = \underset{0.5}{\Pr(k=1,n=2|0/1)}\underset{0.01}{\Pr(0/1)} + \underset{0.0198}{\Pr(k=1,n=2|0/0)}\underset{0.99}{\Pr(0/0)} + \underset{0.0198}{\Pr(k=1,n=2|1/1)}\underset{0}{\Pr(1/1)} =$$

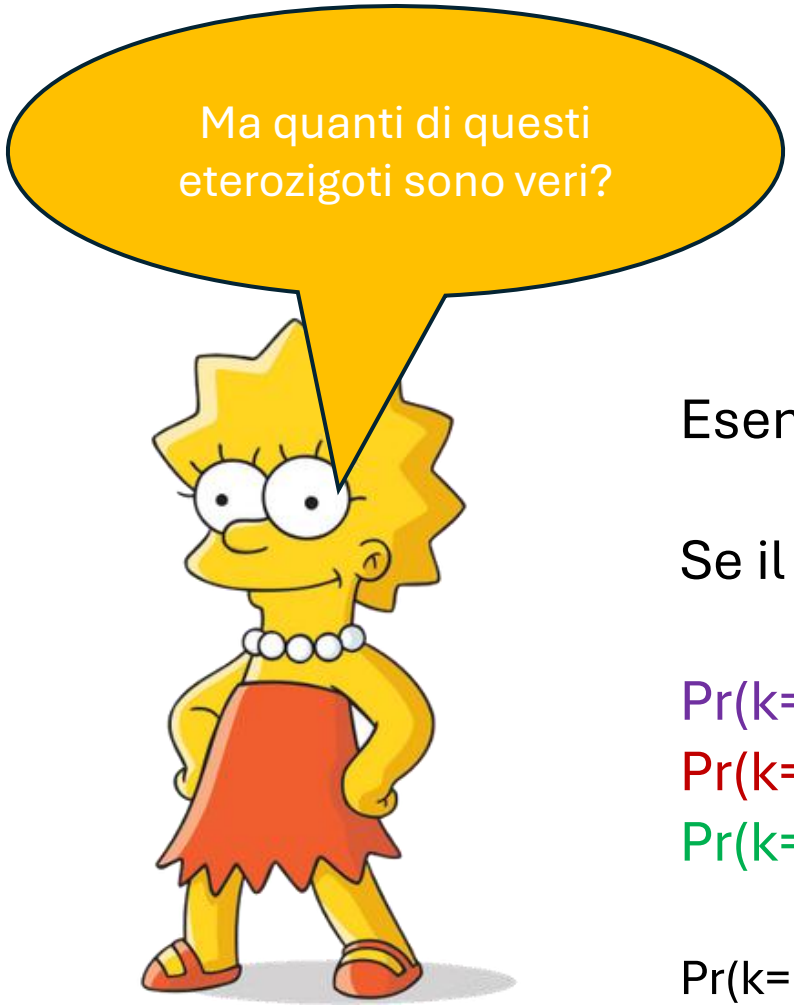
=0.024602 Probabilità totale di vedere un eterozigote





La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro



Probabilità di avere un vero eterozigote dato che osservo =

0/1 |

Likelihood di osservare 1 read dato 0/1 * Probabilità di avere un sito che sia vero eterozigote 0/1

Probabilità di osservare

Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

Pr(k=1,n=2)=Pr(k=1,n=2|0/1)Pr(0/1)+Pr(k=1,n=2|0/0)Pr(0/0)+Pr(k=1,n=2|1/1)Pr(1/1)=

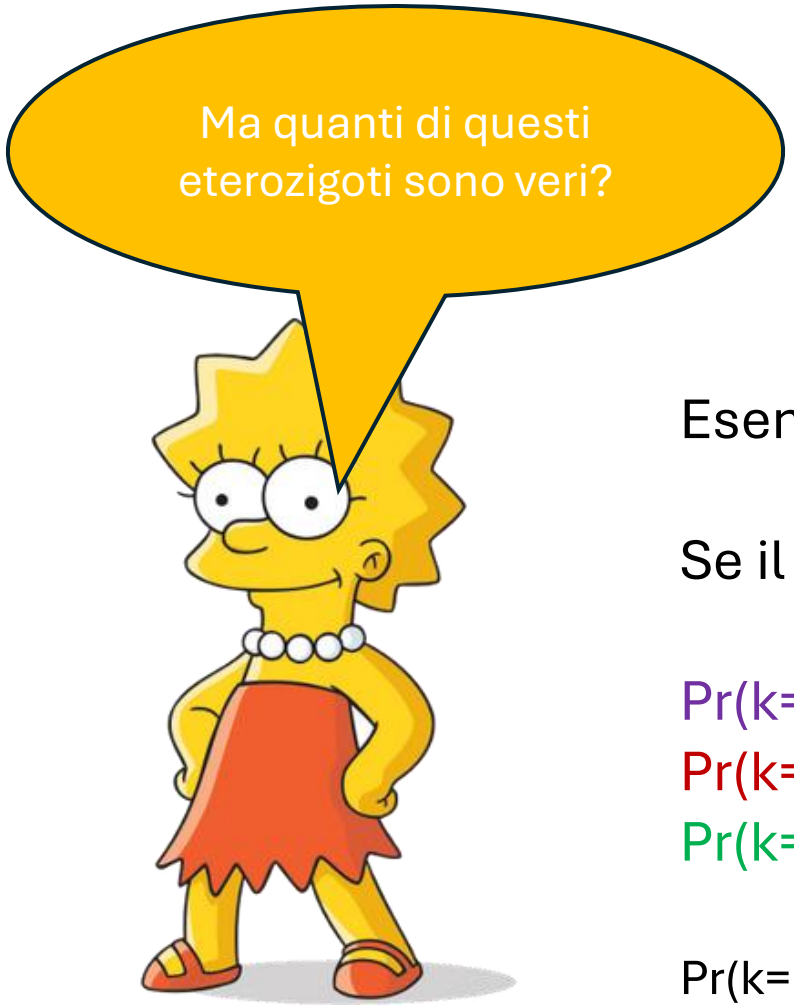
0.5 0.01 0.0198 0.99 0.0198 0

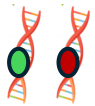
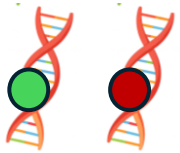
=0.024602

Probabilità totale di vedere un eterozigote

La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro



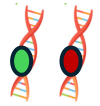
Probabilità di avere un vero eterozigote dato che osservo 0/1 |  = $\frac{0.5*0.01}{\text{Probabilità di osservare}}$ 

Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

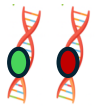
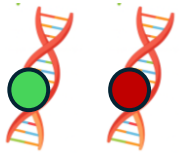
Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

$$\Pr(k=1,n=2)=\underset{0.5}{\Pr(k=1,n=2|0/1)}\underset{0.01}{\Pr(0/1)}+\underset{0.0198}{\Pr(k=1,n=2|0/0)}\underset{0.99}{\Pr(0/0)}+\underset{0.0198}{\Pr(k=1,n=2|1/1)}\underset{0}{\Pr(1/1)}=$$

=0.024602 Probabilità totale di vedere un eterozigote 

La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

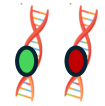
Probabilità di avere un vero eterozigote dato che osservo 0/1 |  = $\frac{0.5*0.01}{\text{Probabilità di osservare}}$ 

Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

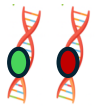
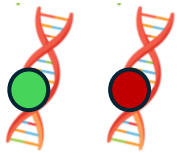
Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

$$\Pr(k=1,n=2)=\underset{0.5}{\Pr(k=1,n=2|0/1)}\underset{0.01}{\Pr(0/1)}+\underset{0.0198}{\Pr(k=1,n=2|0/0)}\underset{0.99}{\Pr(0/0)}+\underset{0.0198}{\Pr(k=1,n=2|1/1)}\underset{0}{\Pr(1/1)}=$$

=0.024602 Probabilità totale di vedere un eterozigote 

La probabilità di un genotipo in presenza di errori

Aggiungiamo una probabilità di errore ε che un allele possa essere letto come un altro

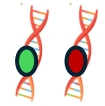
Probabilità di avere un vero eterozigote dato che osservo 0/1 |  = $\frac{0.5*0.01}{\text{Probabilità di osservare}}$ 

Esempio: k=1, n=2 e ε=0.01:

Se il genotipo è:

Pr(k=1,n=2|0/1)=0.5
Pr(k=1,n=2|0/0)=0.0198
Pr(k=1,n=2|1/1)=0.0198

$$\Pr(k=1,n=2)=\underset{0.5}{\Pr(k=1,n=2|0/1)}\underset{0.01}{\Pr(0/1)}+\underset{0.0198}{\Pr(k=1,n=2|0/0)}\underset{0.99}{\Pr(0/0)}+\underset{0.0198}{\Pr(k=1,n=2|1/1)}\underset{0}{\Pr(1/1)}=$$

=0.024602 Probabilità totale di vedere un eterozigote 

Un modello piu' complesso di errore: errori specifici per read

- Abbiamo visto che in un fastq e nei .bam files c'è anche informazione riguardo alla probabilità che ogni specifica base sia errata.

- Questa informazione può essere utilizzata per «considerare» meno le basi piu' incerte

```
~$ head sequences1.fastq
@A00619:220:HYK7GDRX3:1:2101:13440:1000 1:N:0:ACTATAGA+TGCATACC
NCCCGTGAGATGATGCGCGTCATCCTCAACAAGGCCAAATCTGACCCCAAGCGACTCGTTCTCGCTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00619:220:HYK7GDRX3:1:2101:18629:1000 1:N:0:ACTATAGA+TGCATACC
NGCTGTGGACTCCACCCGAGTTCCGGATCGAAACTGGTGAACTTATCCAGCAGCTCTCGGTCGATCCCCGGGAACTCCTCCAACACCACTGTTGGAAGACT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00619:220:HYK7GDRX3:1:2101:20745:1000 1:N:0:ACTATAGA+TGCATACC
NGGATATCAATCAACTCCACGCAACTCCAACTTCCCCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTATAGAATCTCGGTTTGCGTCTTATG
```

Le sequenze di DNA: il formato fastq

```
~$ head sequences1.fastq
@A00619:220:HYK7GDRX3:1:2101:13440:1000 1:N:0:ACTATAGA+TGCATACC
NCCCGTGAGATGATGCGCGTCATCCTCAACAAGGCCAAATCTGACCCCAAGCGACTCGTTCTCGCTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00619:220:HYK7GDRX3:1:2101:18629:1000 1:N:0:ACTATAGA+TGCATACC
NGCTGTGGACTCCACCCGAGTTCCGGATCGAAACTGGTGAACTTATCCAGCAGCTCTCGGTCGATCCCCGGGAACTCCTCCAACACCACTGTTGGAAGACT
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00619:220:HYK7GDRX3:1:2101:20745:1000 1:N:0:ACTATAGA+TGCATACC
NGGATATCAATCAACTCCACGCAACTCCAACTTCCCCTCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACACTATAGAATCTCGGTTTGCGTCTTATG
```

- Riga1: @ riga con I nomi delle sequenze ed informazioni sul sequenziamento
- Riga2: Sequenza ACGT..
- Riga3: +, un semplice separatore
- Riga4: Qualità delle sequenze

Riga 4

NCCCGTGAGATGATGCGCGTCATCCTCAACAAGGCCAAATCTGACCCCAAGCGACTCGTTCTCGCTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

+

#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Qualità codificata come ASCII **Phred+33** → character → Q score → probabilità di erore.

N all'inizio → base **not confidently called** (incerta, "può essere qualsiasi cosa").

Esempi:

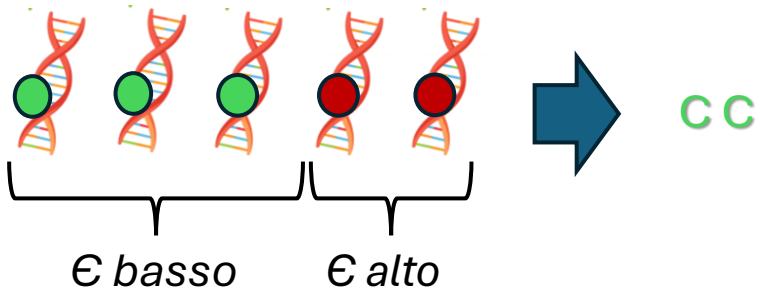
→ ASCII 35 → Q=2 → ~63% accuratezza (bassa, in pratica inutile).

F → ASCII 70 → Q=37 → error probability ≈ 0.0002 (~1 errore in 5000 basi, molto alta).

: → ASCII 58 → Q=25 → errore ≈ 0.003 (~1 in 316 basi).

Un modello piu' complesso di errore: errori specifici per read

```
NCCCGTGAGATGATGCGCGTCATCCTCAACAAGGCCAAATCTGACCCCAAGCGACTCGTTCTCGCTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```



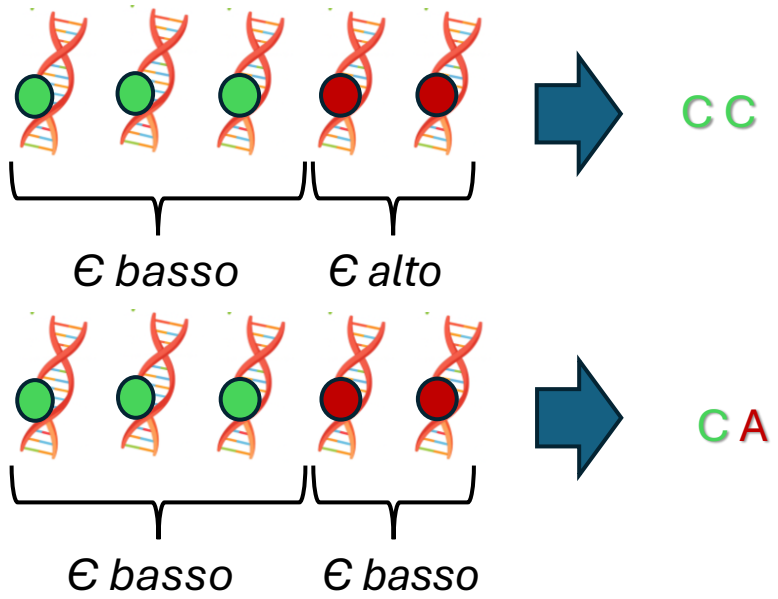
Per ogni read *j* ho una qualita' ϵ_j

(la probabilita' di errore per ogni base sul FASTQ)

che puo' essere utile per inferire quali reads dobbiamo considerare

Un modello piu' complesso di errore: errori specifici per read

```
NCCCGTGAGATGATGCGCGTCATCCTCAACAAGGCCAAATCTGACCCCAAGCGACTCGTTCTCGCTGAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```



Per ogni read *j* ho una qualita' ϵ_j

(la probabilita' di errore per ogni base sul FASTQ)

che puo' essere utile per inferire quali reads dobbiamo considerare

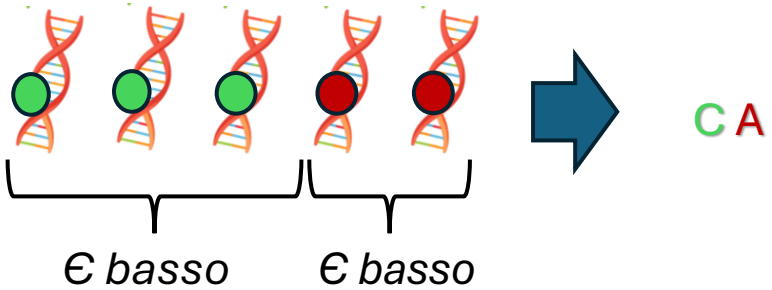
Un modello piu' complesso di errore: errori specifici per read

ALT con errore REF senza errore ALT senza errore REF con errore

$$\mathcal{L}(g)=\frac{1}{m^k}\prod_{j=1}^{l}\left[(m-g)\epsilon_j+g(1-\epsilon_j)\right]\prod_{j=l+1}^{k}\left[(m-g)(1-\epsilon_j)+g\epsilon_j\right]$$

Reads che supportano REF Reads che supportano ALT

Niente paura, non dovete saperla! E' sulle slides solo per referenza se siete curiosi come si calcola!



Per ogni read *j* ho una qualita' ϵ_j

(la probabilita' di errore per ogni base sul FASTQ)

che puo' essere utile per inferire quali reads dobbiamo considerare

Veramente, fate un respiro e andate oltre se volete!

ALT con errore REF senza errore ALT senza errore REF con errore

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

Reads che supportano REF — Reads che supportano ALT

- m è la ploidia, 2 per diploidi.
- g è il genotipo, o numero di reference alleles 2 per 0/0, 1 per 0/1 e 0 per 1/1.
- ϵ_j è l'errore per una specifica posizione/read *j*.
- *l* reads REF su un totale *k* di reads

Veramente, fate un respiro e andate oltre se volete!

$$\mathcal{L}(g) = \prod_{j=1}^{l} \left[(m-g)\epsilon_j \right]$$

ALT con errore

- $(2\text{-}0)\epsilon_j = 2\,\epsilon_j$ — 1/1 con errori e che risultano in REF
- $(2\text{-}1)\epsilon_j = \epsilon_j$ — 0/1 con errori e che risultano in REF
- $(2\text{-}2)\epsilon_j = 0$ — 0/0 con errori e che risultano in REF

Reads che supportano REF

Probabilità rappresentata

- m è la ploidia, 2 per diploidi.
- g è il genotipo, o numero di reference alleles 2 per 0/0, 1 per 0/1 e 0 per 1/1.
- ϵ_j è l'errore per una specifica posizione/read *j*.
- *l* reads REF su un totale *k* di reads

Veramente, fate un respiro e andate oltre se volete!

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j \right]$$

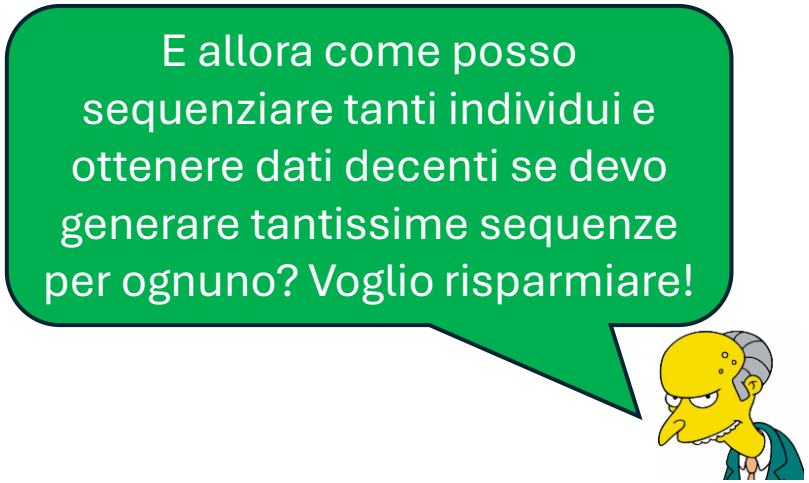
ALT con errore

Reads che supportano REF

		Probabilità rappresentata
$(2\text{-}0)\epsilon_j = 2\,\epsilon_j$	ϵ_j	1/1 con errori e che risultano in REF
$(2\text{-}1)\epsilon_j = \epsilon_j$	$\epsilon_j/2$	0/1 con errori e che risultano in REF
$(2\text{-}2)\epsilon_j = 0$	0	0/0 con errori e che risultano in REF

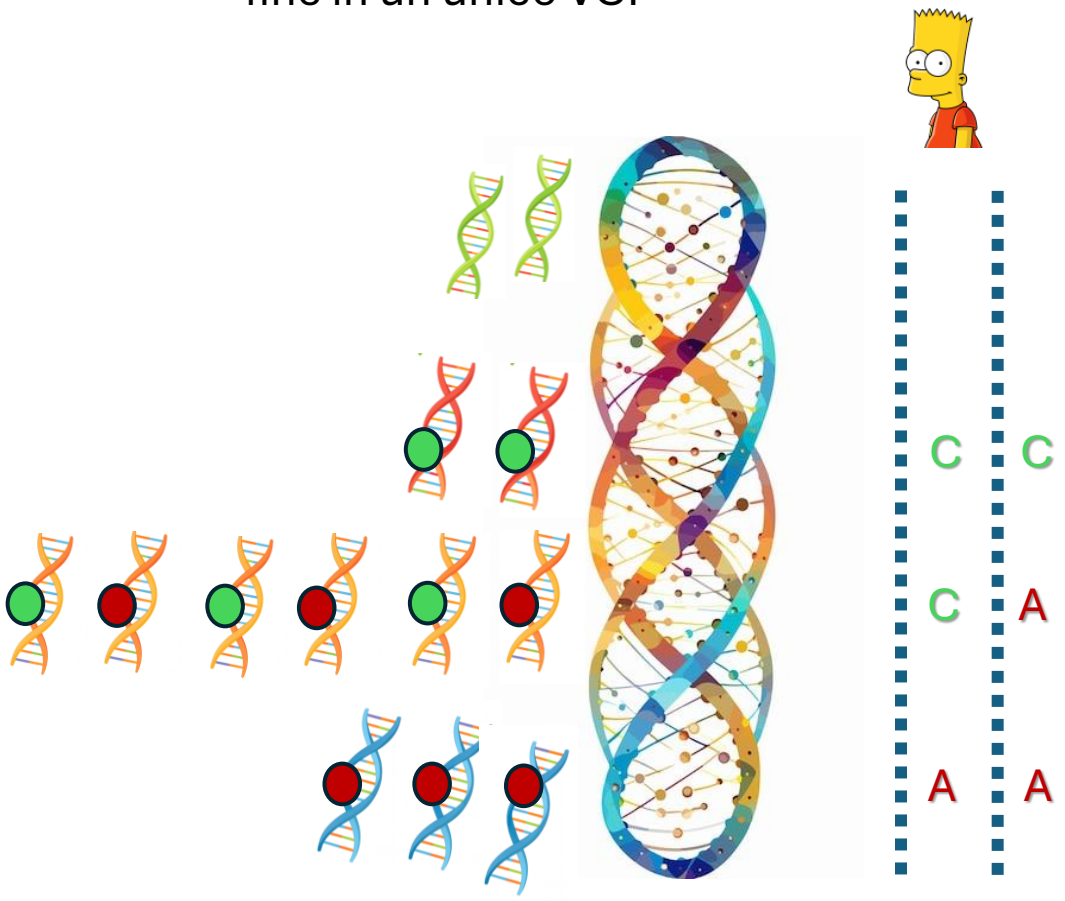
- m è la ploidia, 2 per diploidi.
- g è il genotipo, o numero di reference alleles 2 per 0/0, 1 per 0/1 e 0 per 1/1.
- ϵ_j è l'errore per una specifica posizione/read *j*.
- *l* reads REF su un totale *k* di reads

La probabilità di un genotipo in presenza di errori



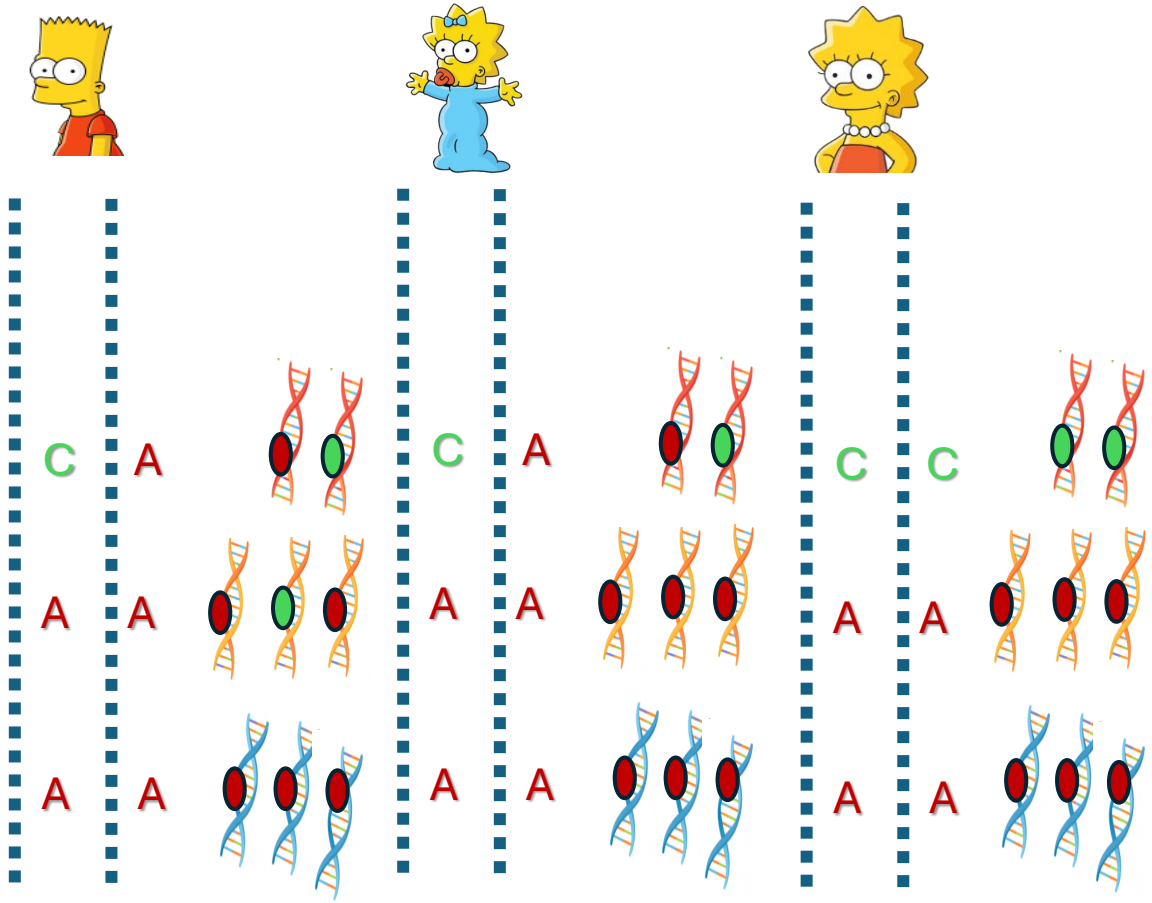
Genotipizzazione a singolo campione

- Tutti i campioni vengono genotipizzati uno per uno
- I genotipi vengono uniti alla fine in un unico VCF



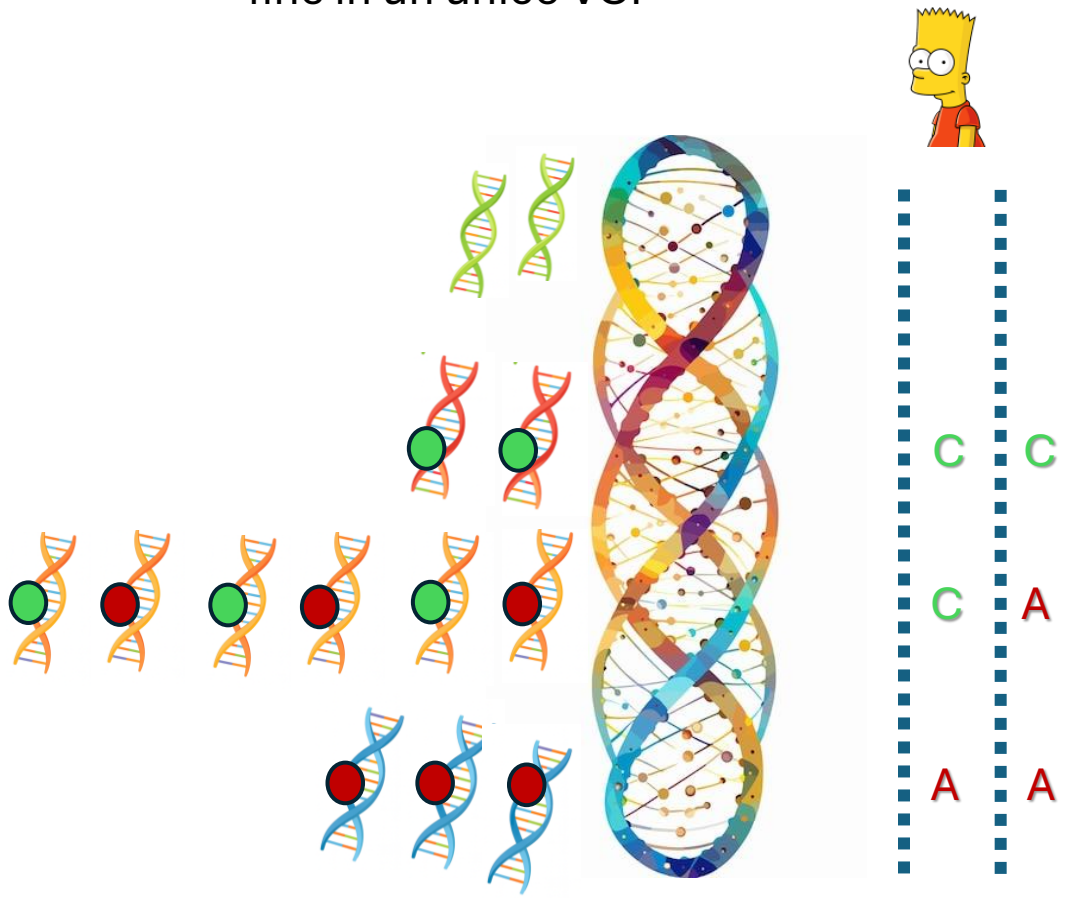
Genotipizzazione multi-campione

- L’informazione sui siti variabili e quelli dove è piu’ improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



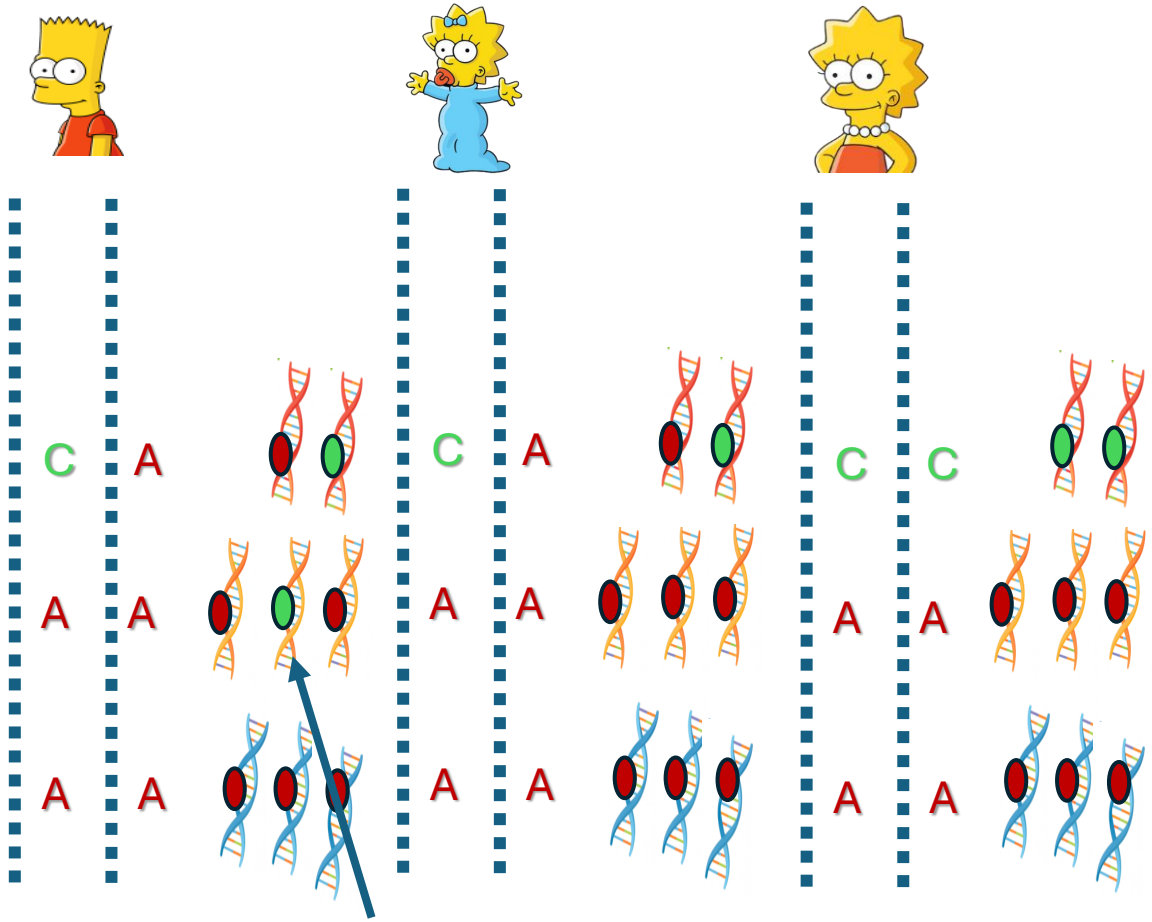
Genotipizzazione a singolo campione

- Tutti i campioni vengono genotipizzati uno per uno
- I genotipi vengono uniti alla fine in un unico VCF



Genotipizzazione multi-campione

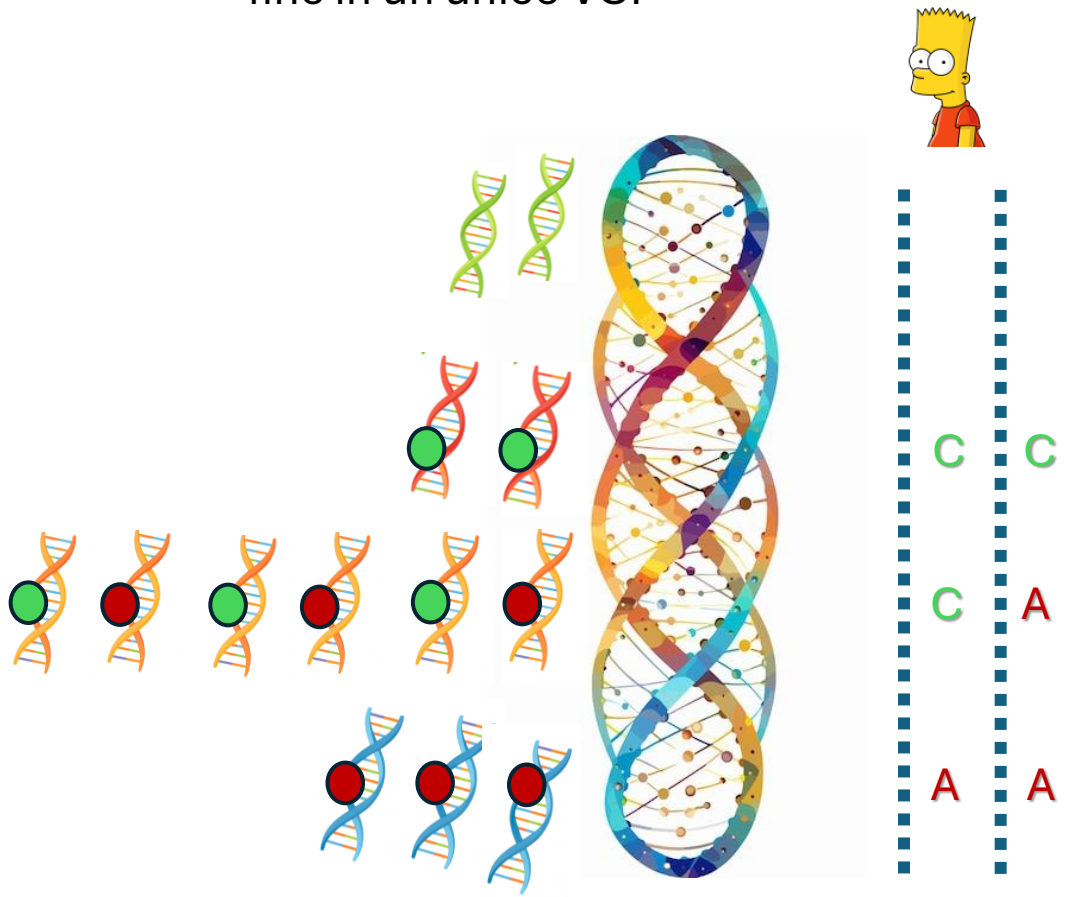
- L’informazione sui siti variabili e quelli dove è piu’ improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

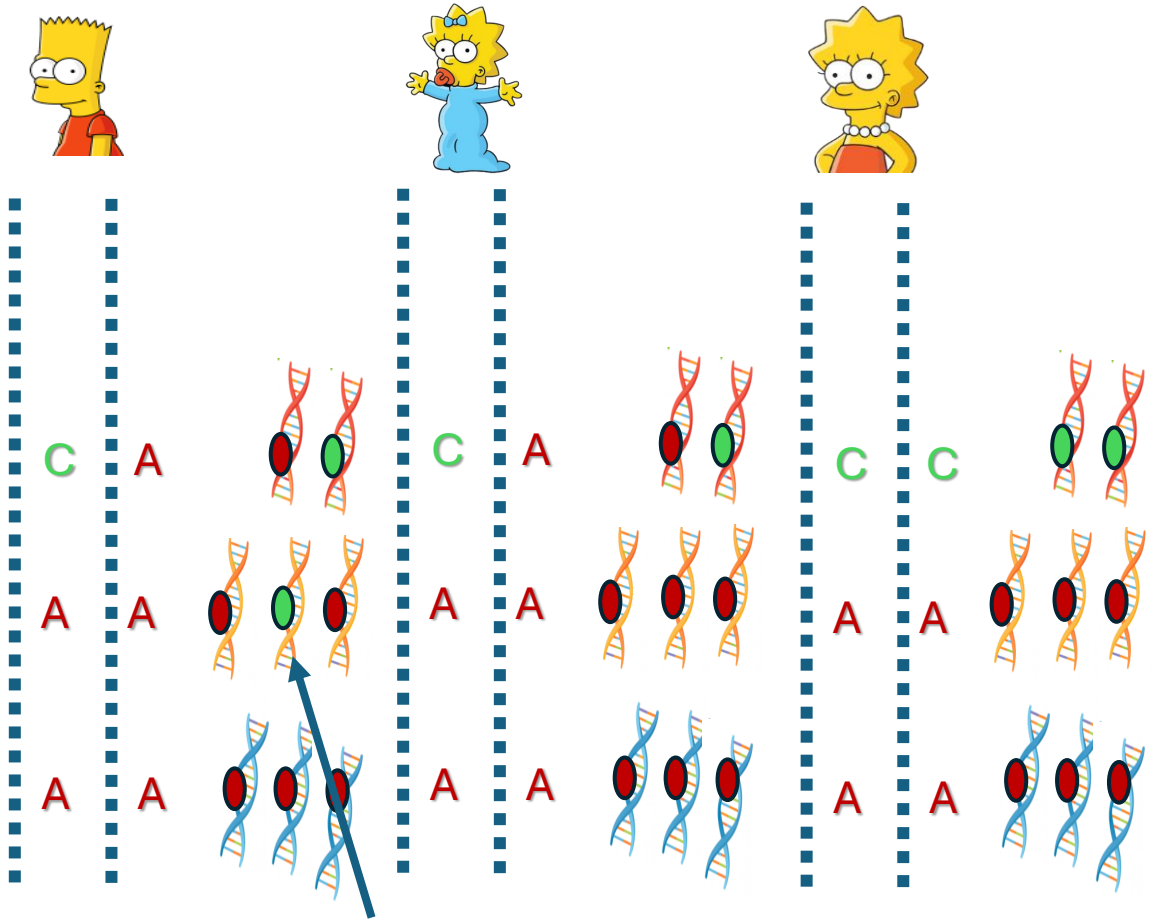
Genotipizzazione a singolo campione

- Tutti i campioni vengono genotipizzati uno per uno
- I genotipi vengono uniti alla fine in un unico VCF



Genotipizzazione multi-campione

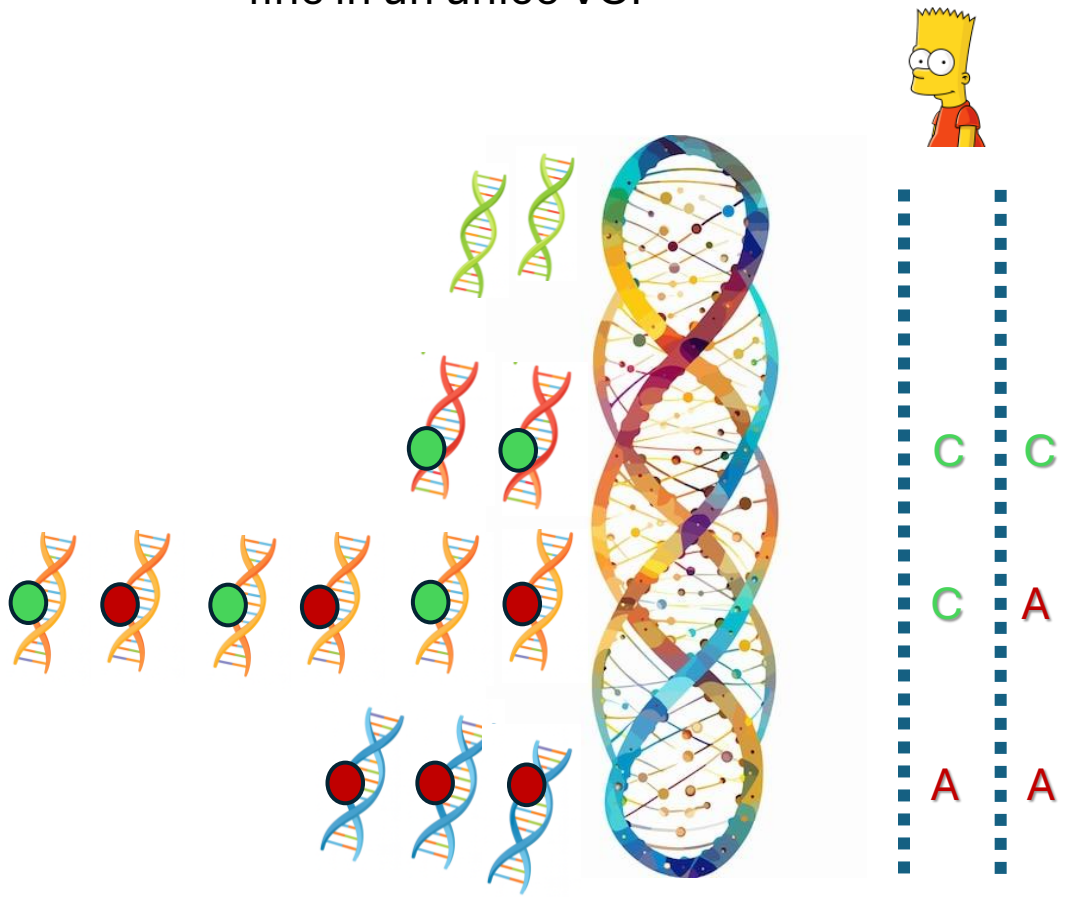
- L’informazione sui siti variabili e quelli dove è piu’ improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

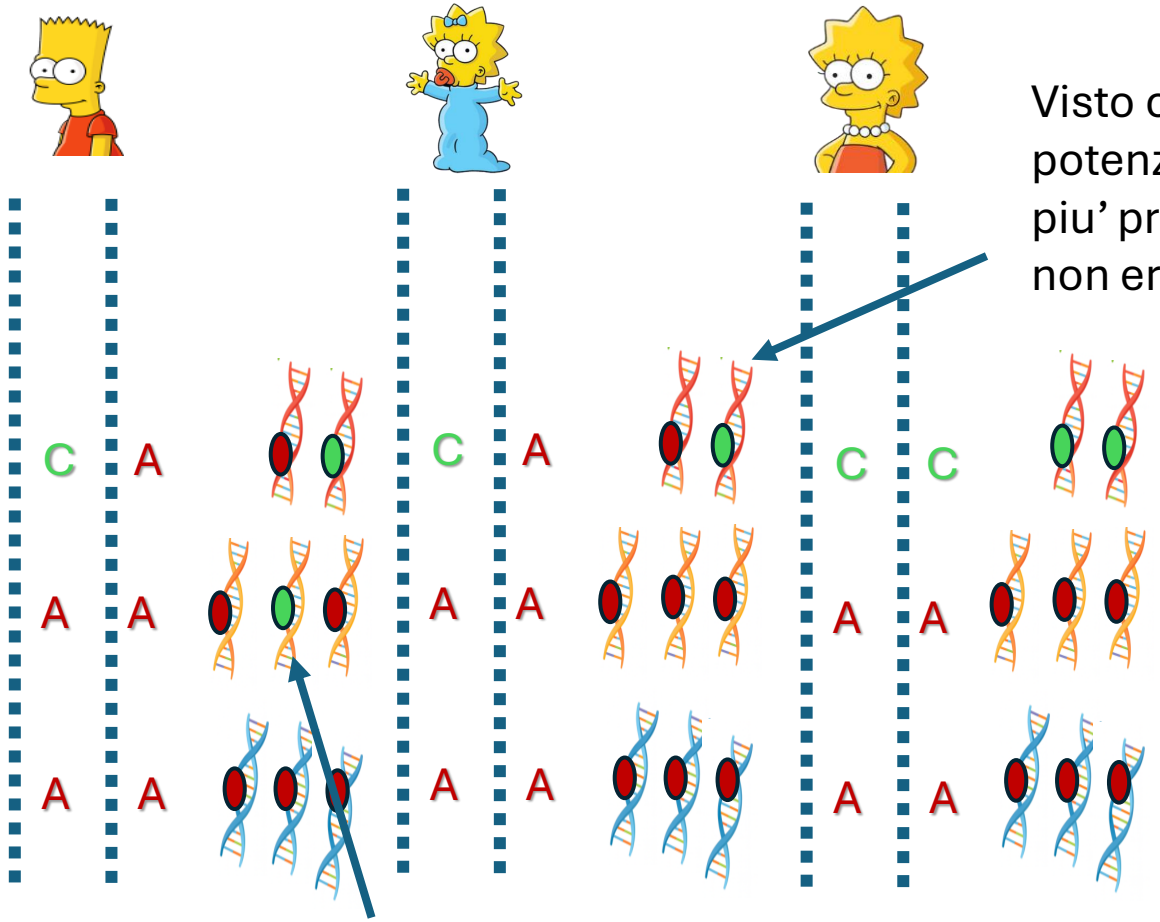
Genotipizzazione a singolo campione

- Tutti i campioni vengono genotipizzati uno per uno
- I genotipi vengono uniti alla fine in un unico VCF



Genotipizzazione multi-campione

- L’informazione sui siti variabili e quelli dove è piu’ improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri

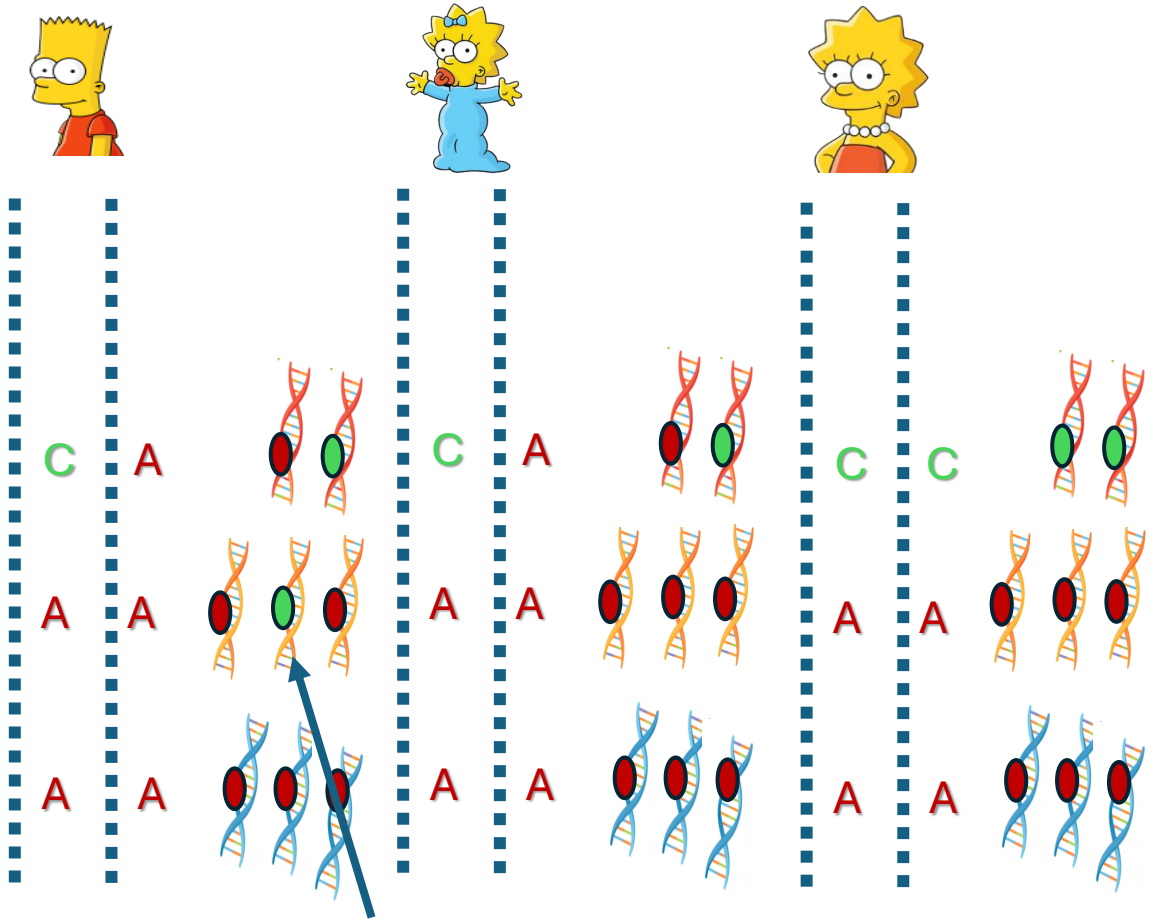


Visto che si osservano vari potenziali eterozigoti è piu’ probabile che siano reali e non errori

Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

Genotipizzazione multi-campione

- L'informazione sui siti variabili e quelli dove è piu' improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

Calcola i genotipi individuali

Usa i genotipi individuali per ottenere una stima ψ delle frequenze di C e A (REF e ALT).

Usa l'equilibrio di Hardy-Weinberg per stimare la proporzione attesa dei genotipi data ψ

Hardy-Weinberg equilibrium

If there are only 2 alleles for a trait in a population, then:

$$p^2 + 2pq + q^2 = 1$$

p^2: frequency of homozygous dominant genotype

$2pq$: frequency of heterozygous genotype

q^2: frequency of homozygous recessive genotype

Purple is dominant to Pink

Hardy Weinberg Theorem

ALLELES

Probability of A = p p + q = 1

Probability of a = q

GENOTYPES

AA: p x p = p^2

Aa: p x q + q x p = 2pq

aa: q x q = q^2

$p^2 + 2pq + q^2 = 1$

Eggs

Sperm

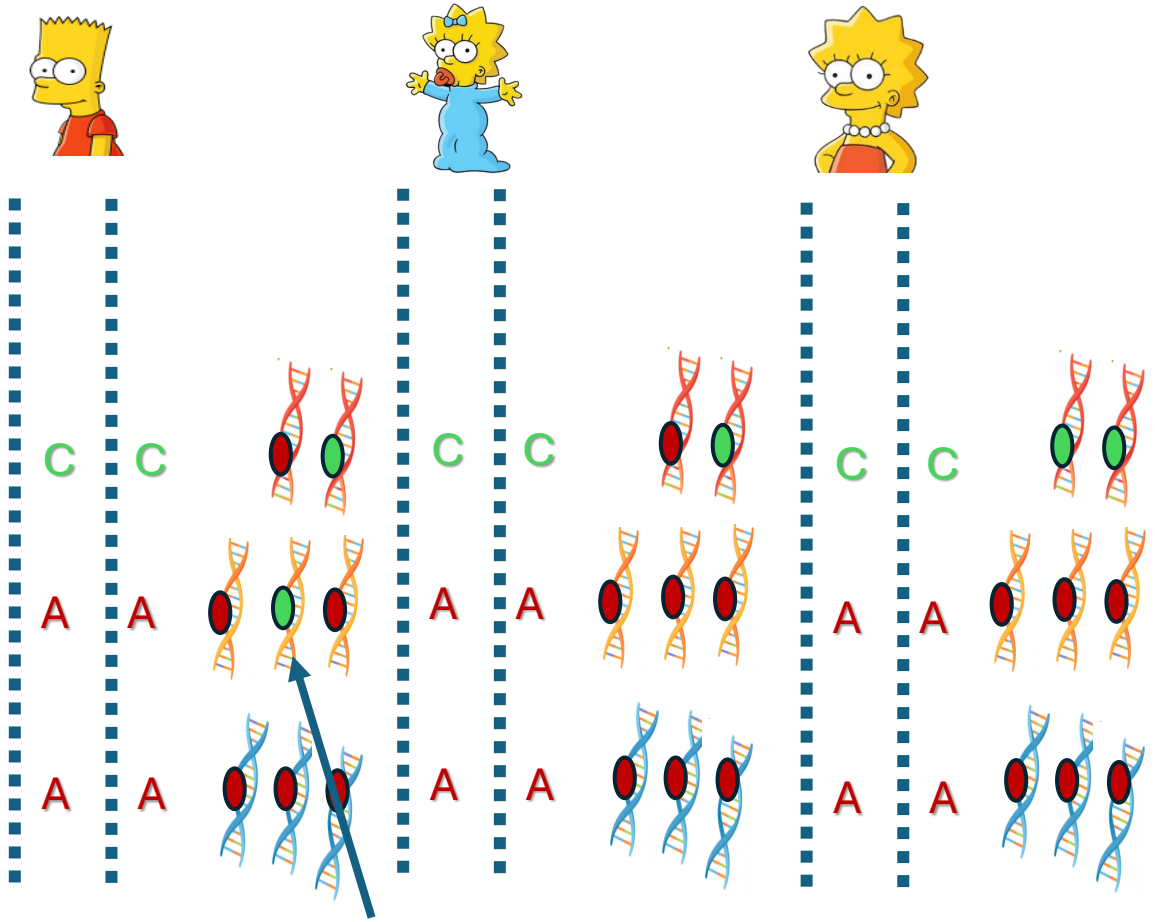
	A	a
A	AA	Aa
a	aA	aa

Zygotes

Figure 6-4 Evolutionary Analysis, 4/e

Genotipizzazione multi-campione

- L'informazione sui siti variabili e quelli dove è piu' improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

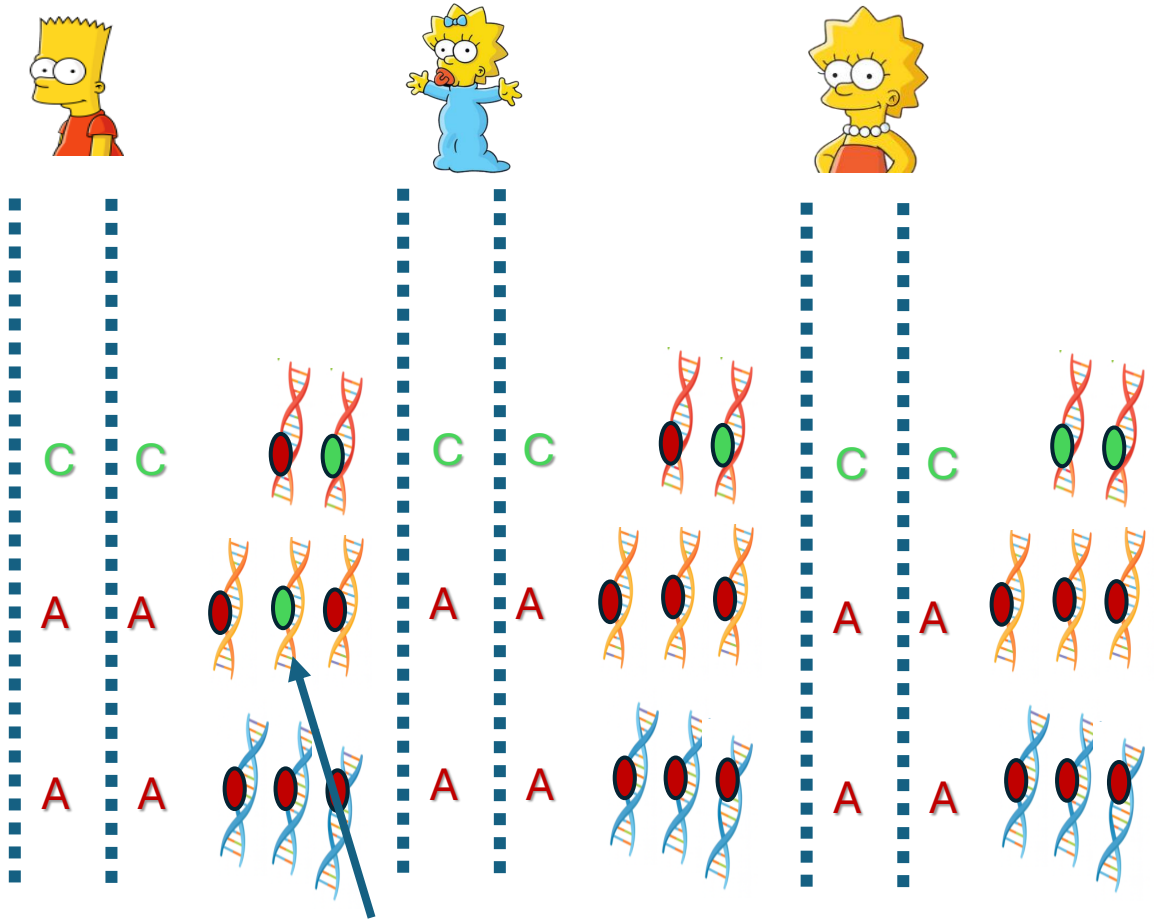
Calcola i genotipi individuali

Usa i genotipi individuali per ottenere una stima ψ delle frequenze di C e A (REF e ALT).

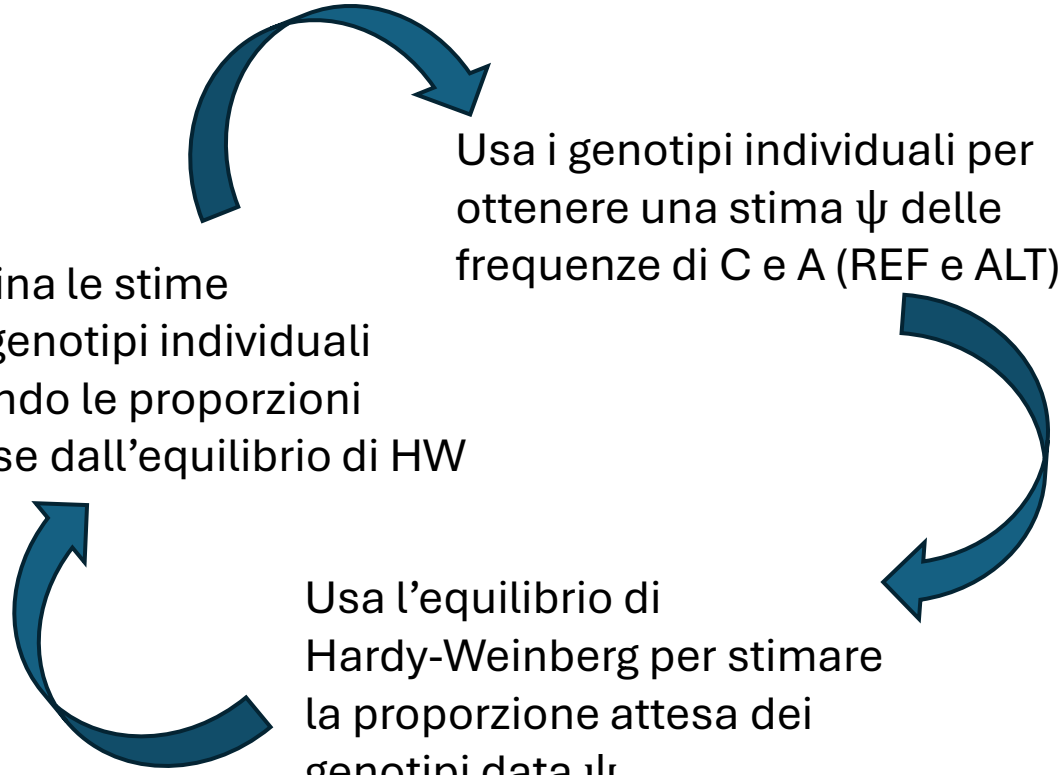
Usa l'equilibrio di Hardy-Weinberg per stimare la proporzione attesa dei genotipi data ψ

Genotipizzazione multi-campione

- L'informazione sui siti variabili e quelli dove è piu' improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri

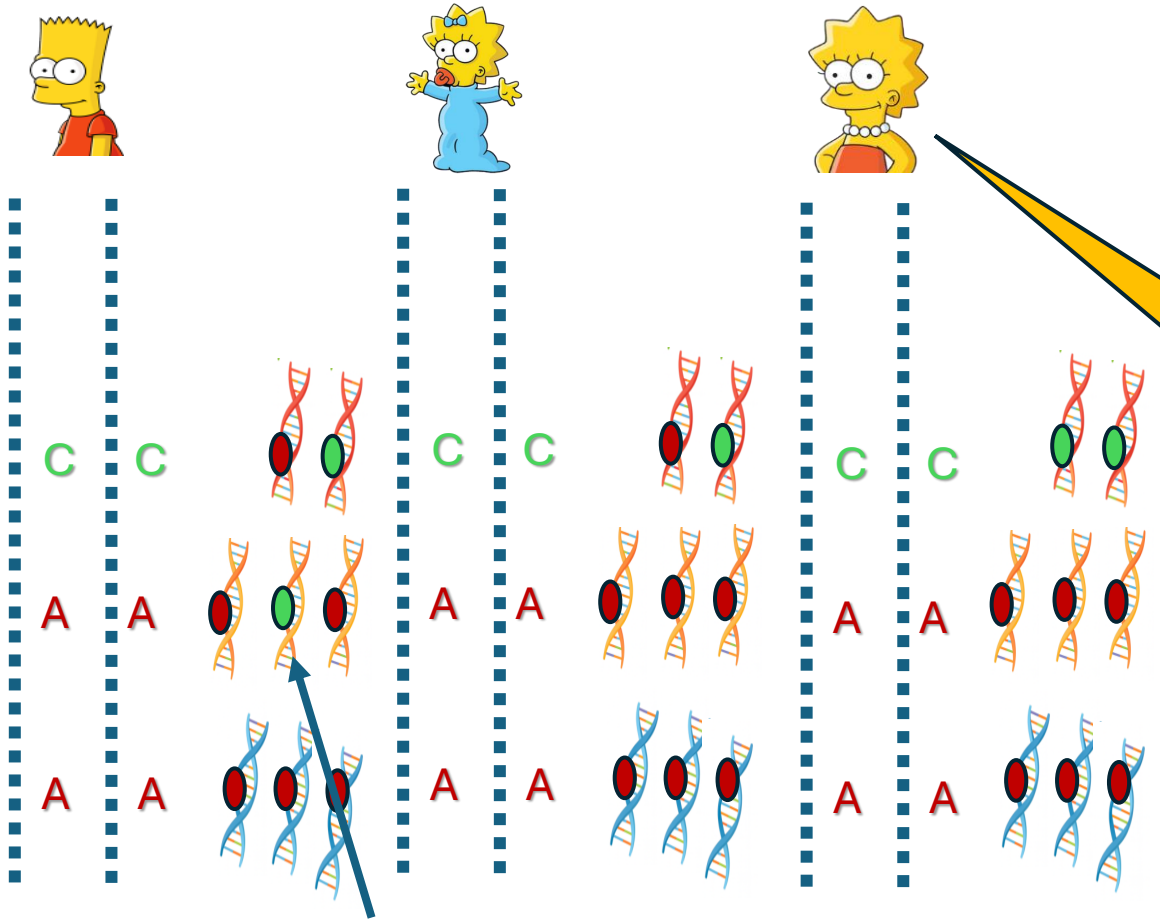


Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore



Genotipizzazione multi-campione

- L'informazione sui siti variabili e quelli dove è piu' improbabile che ci siano reali varianti viene combinata
- Non sono solo VCF combinati, ma i genotipi di un individuo sono informati dagli altri



Visto che nessun altro individuo mostra un allele alternativo, probabilmente si tratta di un errore

Usa i genotipi individuali per ottenere una stima ψ delle frequenze di C e A (REF e ALT).

Usa l'equilibrio di Hardy-Weinberg per stimare la proporzione attesa dei genotipi data ψ

Raffina le stime dei genotipi individuali Usando le proporzioni attese dall'equilibrio di HW

Nelle prossime lezioni vedremo come implementare questi algoritmi iterativi (Expectation-Maximization e Hidden Markov Models)

Genotipizzazione multi-campione

Prima avevamo calcolato la likelihood dei genotipi individuali

$$\mathcal{L}(g) = \frac{1}{m^k} \prod_{j=1}^{l} \left[(m-g)\epsilon_j + g(1-\epsilon_j) \right] \prod_{j=l+1}^{k} \left[(m-g)(1-\epsilon_j) + g\epsilon_j \right]$$

Niente paura, anche questa è solo per referenza se siete curiosi come si calcola!

Genotipizzazione multi-campione

Prima avevamo calcolato la likelihood dei genotipi individuali

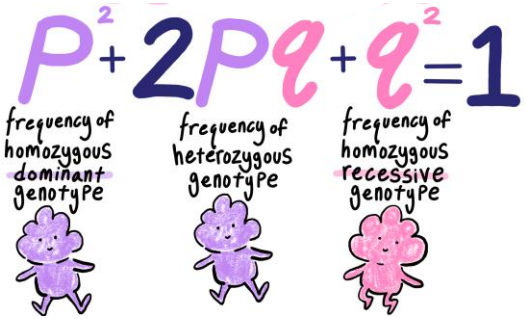
$$\mathcal{L}(g)=\frac{1}{m^k}\prod_{j=1}^{l}\left[(m-g)\epsilon_j+g(1-\epsilon_j)\right]\prod_{j=l+1}^{k}\left[(m-g)(1-\epsilon_j)+g\epsilon_j\right]$$

Ora non facciamo che usare questa formula condizionando alla distribuzione dei genotipi data dalla distribuzione (binomiale) prevista dall'equilibrio di Hardy-Weinberg

$$\mathcal{L}(\psi)=\sum_{g_1}\cdots\sum_{g_n}\prod_{i}\Pr\{d_i,g_i|\psi\}=\prod_{i=1}^{n}\sum_{g=0}^{m_i}\mathcal{L}_i(g)f(g;m_i,\psi)$$

dove ψ sono le frequenze alleliche

$$f(g;m,\psi)=\binom{m}{g}\psi^g(1-\psi)^{m-g}$$



Imputazione

Non solo l'informazione di altri individui può essere utilizzata per migliorare l'accuratezza dei genotipi, ma anche quella di siti vicini.

Varianti vicine tra loro lungo il genoma sono spesso correlate a causa della ricombinazione

Questa informazione può essere sfruttata con l'utilizzo di Hidden Markov Models (che vedremo nelle lezioni successive)






#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1
1	1422	A	C	50	PASS	…	./.	0/1	0/0	0/0	0/1	1/1
1	2456	C	C	45	PASS	…	./.	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1

Imputazione

Sembra essere presente un aplotipo (una serie di alleli correlati) di alleli Alternativi in questa regione del genoma. Tutti gli individui con l'allele alternativo in posizione 122 e 3563 sembrano avere allele alternativo anche in posizione 1422 e 2456.






#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1
1	1422	A	C	50	PASS	…	./.	0/1	0/0	0/0	0/1	1/1
1	2456	C	C	45	PASS	…	./.	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1

Imputazione

Sembra essere presente un aplotipo (una serie di alleli correlati) di alleli Alternativi in questa regione del genoma. Tutti gli individui con l'allele alternativo in posizione 122 e 3563 sembrano avere allele alternativo anche in posizione 1422 e 2456. Probabilmente è cosi anche per Bart!!

Genotipi imputati






#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1
1	1422	A	C	50	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1
1	2456	C	C	45	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	…	0/1	0/1	0/0	0/0	0/1	1/1

Genotipizzazione multi-sample e imputazione (qui con Beagle) migliorano l'accuratezza dei genotipi chiamati

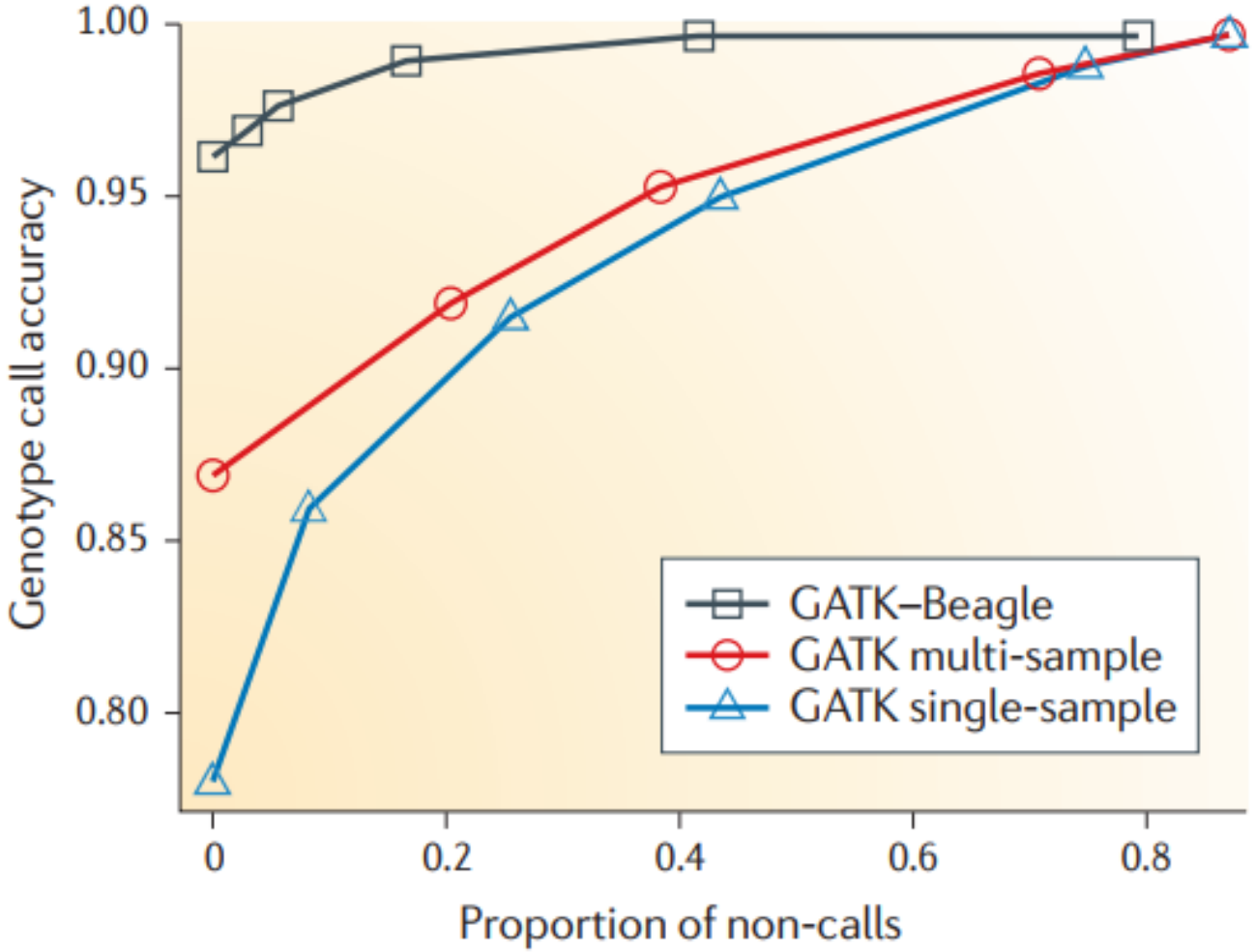


Figure 2 | **A comparison of three genotype callers.** A subset of the data (chromosome 20, bases 20,000,000–25,000,000) for the 62 CEU individuals in both the HapMap Public Release no. 28 and the 1000 Genomes Pilot Project

Nielsen et al.,2010. *Nat Rev Gen*

google/**deepvariant**

DeepVariant is an analysis pipeline that uses a deep neural network to call genetic variants from next-generation DNA sequencing data.

32 Contributors

4 Used by

4k Stars

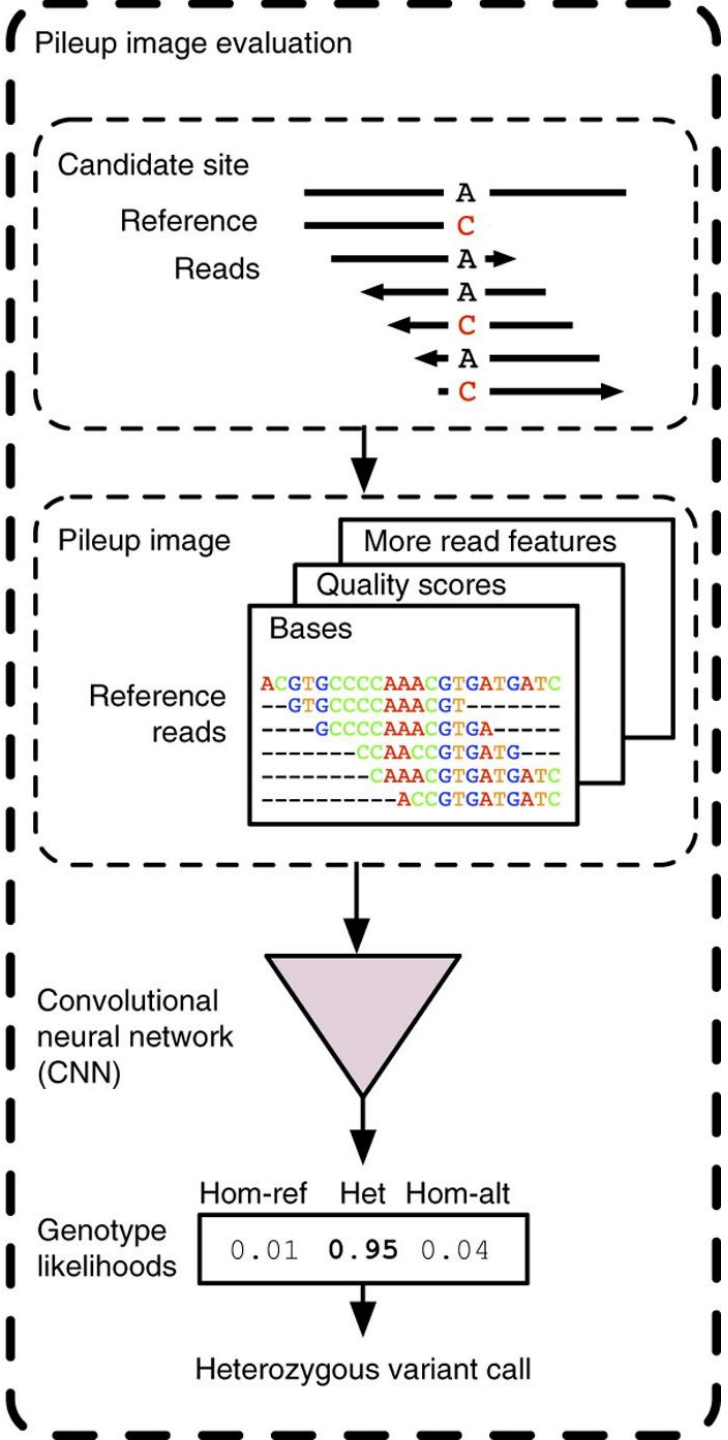
763 Forks

nature biotechnology

A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

- Usa reti neurali (CNN) che interpretano come immagini i dati delle reads e delle loro qualità
- La rete neurale non usa un modello probabilistico ma il suo output sono delle «genotype likelihood»!
- Queste genotype likelihood vengono poi convertite in genotipi da metodi probabilistici



nature biotechnology

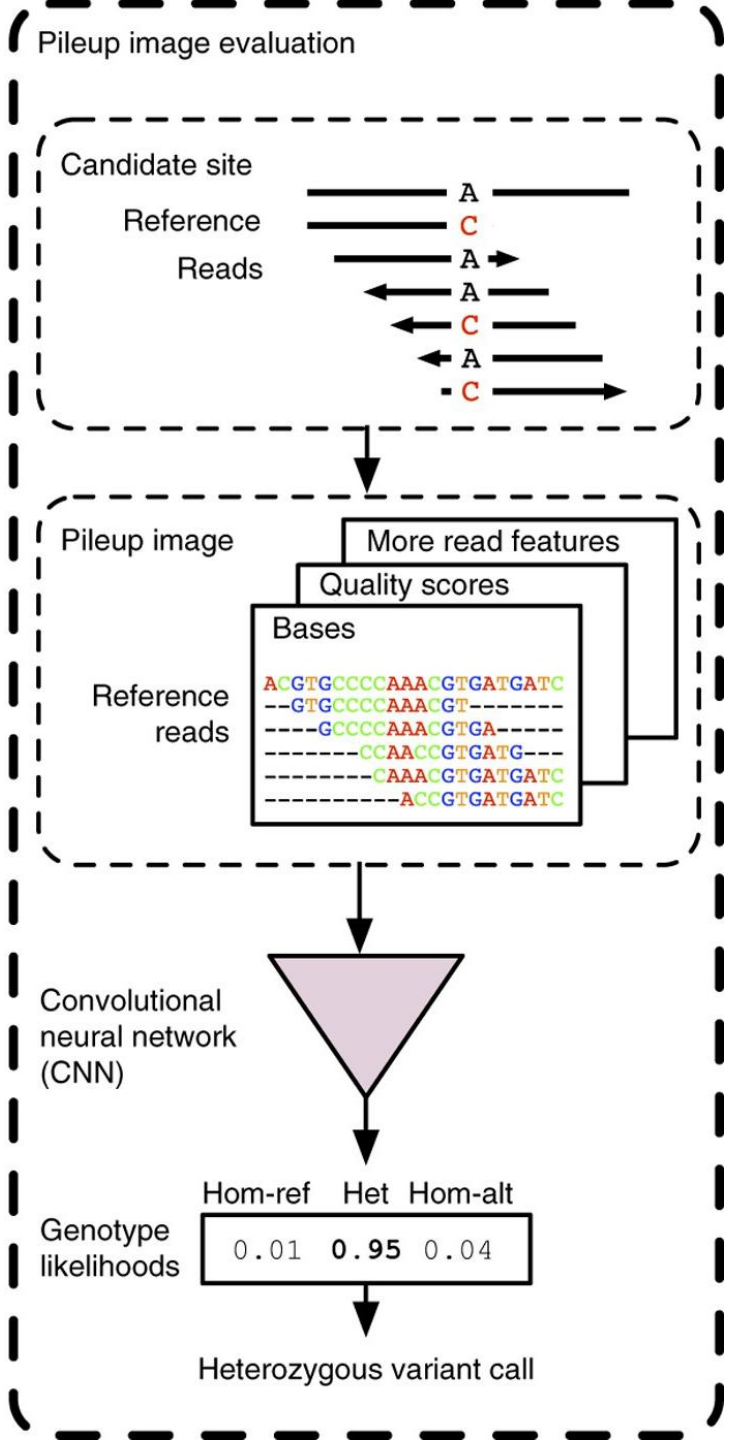
A universal SNP and small-indel variant caller using deep neural networks

Ryan Poplin[1,2], Pi-Chuan Chang[2], David Alexander[2], Scott Schwartz[2], Thomas Colthurst[2], Alexander Ku[2], Dan Newburger[1], Jojo Dijamco[1], Nam Nguyen[1], Pegah T Afshar[1], Sam S Gross[1], Lizzie Dorfman[1,2], Cory Y McLean[1,2] & Mark A DePristo[1,2]

Table 1 Evaluation of several bioinformatics methods on the high-coverage, whole-genome sample NA24385

Method	Type	F1	Recall	Precision	TP	FN	FP	FP.gt	FP.al	Version
DeepVariant (live GitHub)	Indel	0.99507	0.99347	0.99666	357,641	2350	1,198	217	840	Latest GitHub v0.4.1-b4e8d37d
GATK (raw)	Indel	0.99366	0.99219	0.99512	357,181	2810	1,752	377	995	3.8-0-ge9d806836
Strelka	Indel	0.99227	0.98829	0.99628	355,777	4214	1,329	221	855	2.8.4-3-gbe58942
DeepVariant (pFDA)	Indel	0.99112	0.98776	0.99450	355,586	4405	1,968	846	1,027	pFDA submission May 2016
GATK (VQSR)	Indel	0.99010	0.98454	0.99573	354,425	5566	1,522	343	909	3.8-0-ge9d806836
GATK (flt)	Indel	0.98229	0.96881	0.99615	348,764	11227	1,349	370	916	3.8-0-ge9d806836
FreeBayes	Indel	0.94091	0.91917	0.96372	330,891	29,100	12,569	9,149	3,347	v1.1.0-54-g49413aa
16GT	Indel	0.92732	0.91102	0.94422	327,960	32,031	19,364	10,700	7,745	v1.0-34e8f934
SAMtools	Indel	0.87951	0.83369	0.93066	300,120	59,871	22,682	2,302	20,282	1.6
DeepVariant (live GitHub)	SNP	0.99982	0.99975	0.99989	3,054,552	754	350	157	38	Latest GitHub v0.4.1-b4e8d37d
DeepVariant (pFDA)	SNP	0.99958	0.99944	0.99973	3,053,579	1,727	837	409	78	pFDA submission May 2016
Strelka	SNP	0.99935	0.99893	0.99976	3,052,050	3,256	732	87	136	2.8.4-3-gbe58942
GATK (raw)	SNP	0.99914	0.99973	0.99854	3,054,494	812	4,469	176	257	3.8-0-ge9d806836
16GT	SNP	0.99583	0.99850	0.99318	3,050,725	4,581	20,947	3,476	3,899	v1.0-34e8f934
GATK (VQSR)	SNP	0.99436	0.98940	0.99937	3,022,917	32,389	1,920	80	170	3.8-0-ge9d806836
FreeBayes	SNP	0.99124	0.98342	0.99919	3,004,641	50,665	2,434	351	1,232	v1.1.0-54-g49413aa
SAMtools	SNP	0.99021	0.98114	0.99945	2,997,677	57,629	1,651	1,040	200	1.6
GATK (flt)	SNP	0.98958	0.97953	0.99983	2,992,764	62,542	509	168	26	3.8-0-ge9d806836

The dataset used in this evaluation is the same as in the precisionFDA Truth Challenge (pFDA). Several methods are compared, including the DeepVariant callset as submitted to the contest and the most recent DeepVariant version from GitHub. Each method was run according to the individual authors' best-practice recommendations and represents a good-faith effort to achieve best results. Comparisons to the Genome in a Bottle truth set for this sample were performed using the hap.py software, available on GitHub at http://github.com/Illumina/hap.py, using the same version of the GIAB truth set (v3.2.2) used by pFDA. The overall accuracy (F1, sort order within each variant type), recall, precision, and numbers of true positives (TP), false negatives (FN) and false positives (FP) are shown over the whole genome. False positives are further divided by those caused by genotype mismatches (FP.gt) and those cause by allele mismatches (FP.al). Finally, the version of the software used for each method is provided. We present three GATK callsets: GATK (raw), the unfiltered calls emitted by the HaplotypeCaller; GATK (VQSR), the callset filtered with variant quality score recalibration (VQSR); and GATK (flt), the raw GATK callset filtered with run-flt in CHM-eval. See **Supplementary Note 7** for more details.



AI is becoming more and more widely used for genotyping. Do you think old probabilistic methods are totally irrelevant now? Does it make sense to still teach them to introduce probabilistic modeling? or should one just teach AI/deep learning?

- *"Excellent and fundamental question. This gets to the heart of pedagogy in the age of AI. My perspective, with appropriate uncertainty: the old probabilistic methods are more relevant than ever, and teaching them is crucial." DeepSeek, an AI*

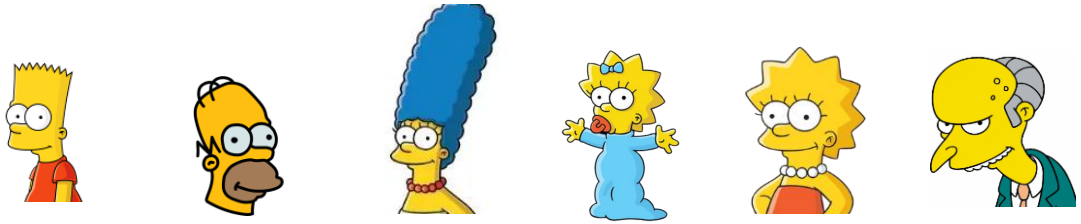
- **Why Probabilistic Methods Remain Essential:**
- Probabilistic models give explicit confidence scores (P(genotype|data))
- DL models are often "black boxes" - hard to debug when they fail
- Probabilistic models work well with low-coverage data or rare variants
- Most state-of-the-art tools now combine both (e.g., DeepVariant uses DL but outputs calibrated probabilities)
- **The Risk of Skipping Foundations:**
- Not understand what the models are actually approximating
- Struggle with model failures or edge cases

AI is becoming more and more widely used for genotyping. Do you think old probabilistic methods are totally irrelevant now? Does it make sense to still teach them to introduce probabilistic modeling? or should one just teach AI/deep learning?

- *"Excellent and fundamental question. This gets to the heart of pedagogy in the age of AI. My perspective, with appropriate uncertainty: the old probabilistic methods are more relevant than ever, and teaching them is crucial." DeepSeek, an AI*

- **Why Probabilistic Methods Remain Essential:**
- Probabilistic models give explicit confidence scores (P(genotype|data))
- DL models are often "black boxes" - hard to debug when they fail
- Probabilistic models work well with low-coverage data or rare variants
- Most state-of-the-art tools now combine both (e.g., DeepVariant uses DL but outputs calibrated probabilities)
- **The Risk of Skipping Foundations:**
- Not understand what the models are actually approximating
- Struggle with model failures or edge cases

Conclusioni



- Modelli probabilistici ci permettono di stimare l'incertezza nelle nostre inferenze e stimare parametri sconosciuti massimizzando la likelihood
- Algoritmi di genotipizzazione possono tenere conto degli errori di sequenziamento e permettere di utilizzare una coverage relativamente bassa per sequenziare migliaia di individui o correggere (parzialmente) i nostri errori quando i dati sono limitati
- Genotipizzazione multi-campione e imputazione aumentano l'accuratezza della genotipizzazione
- Il formato VCF è il formato standard dei dati di genomica di popolazione