METODI PER L'INFERENZA DELLA STORIA DELLE POPOLAZIONI UMANE

Dati genetici di partenza

1

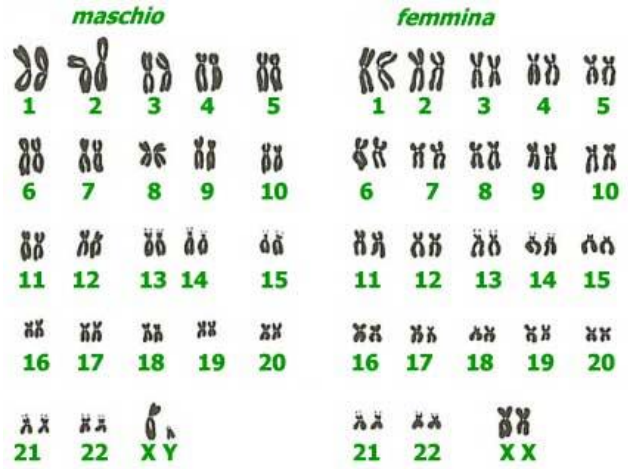
2

Sistemi aploidi

Sistemi diploidi

Sistemi aploidi e diploidi nell'uomo

L'essere umano è un organismo diploide:
22 coppie di autosomi + 1 coppia di cromosomi sessuali (XX nella femmina e XY nel maschio).

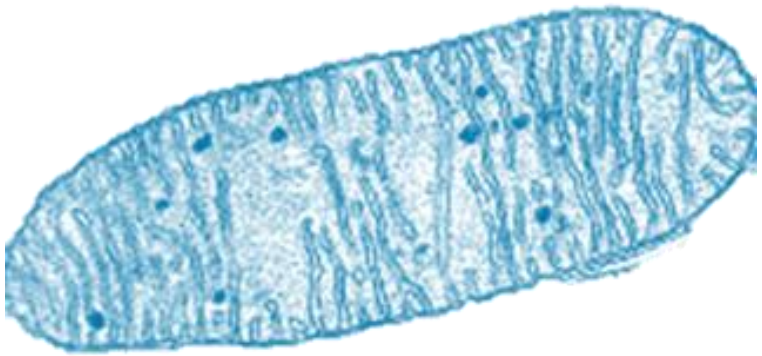


La maggior parte del genoma umano:
- È ereditato e trasmesso da entrambi i genitori;
- Va incontro a ricombinazione meiotica

Tuttavia, nell'uomo esistono due sistemi aploidi:

Il cromosoma Y

Il DNA mitocondriale (mtDNA)

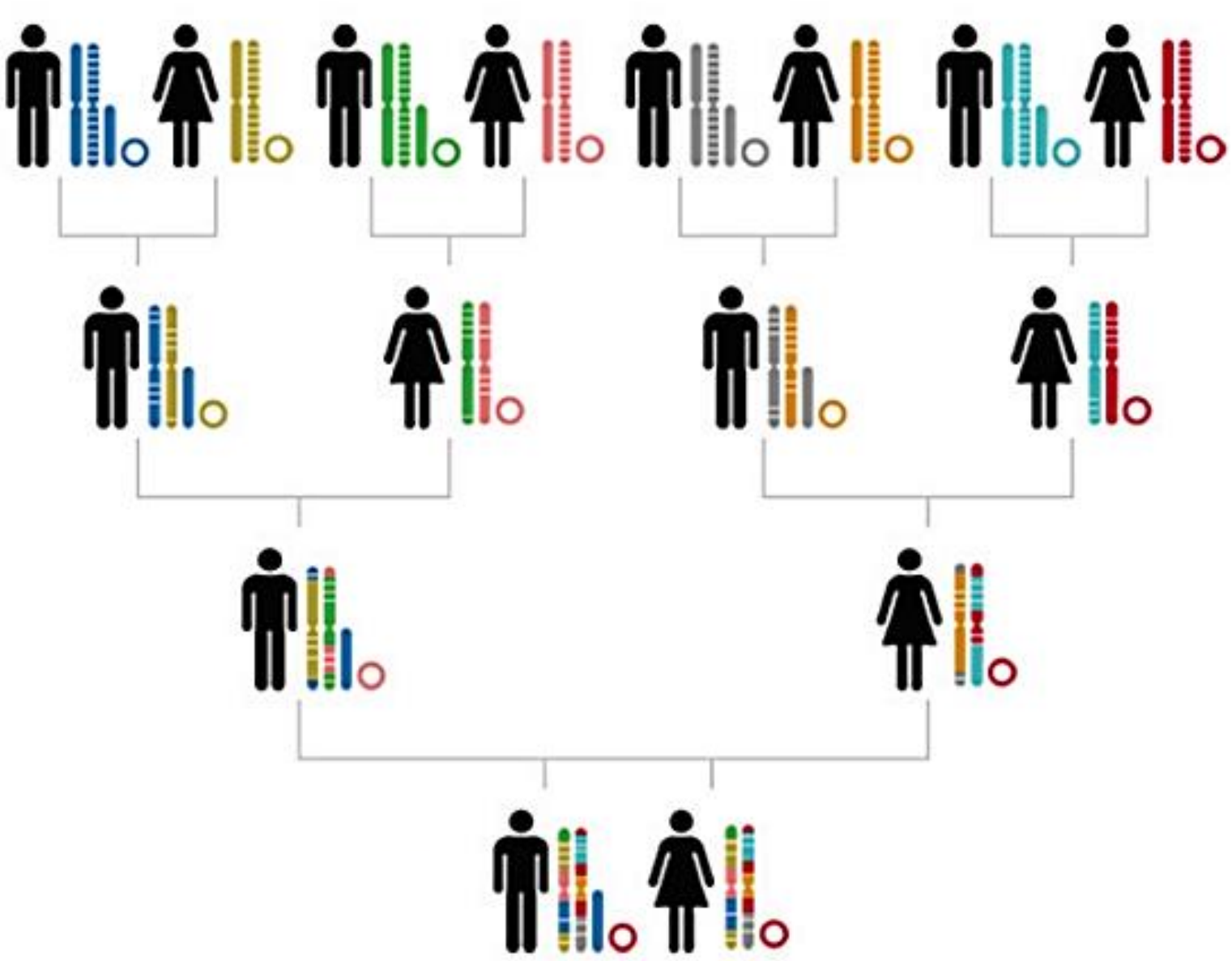


Caratteristiche generali dei sistemi aploidi umani

Caratteristiche	Cromosoma Y	mtDNA
Posizione	Nucleo	Mitocondrio
Lunghezza e forma	≈ 60 Mb, cromosoma lineare	≈ 16.5 kb, cromosoma circolare
Funzione	Determinazione del sesso: Geni essenziali per la fertilità maschile	Produzione di energia: Geni coinvolti nel pathway della fosforilazione ossidativa
Numero	1 per cellula (nei maschi; 0 nelle femmine)	Variabile: cellule che hanno un elevato bisogno di energia (per esempio, neuroni e muscoli) possono avere fino a migliaia di mitocondri, ognuno con 2-10 copie di mtDNA
Ricombinazione	NO	NO
Eredità	Patrilineare: trasmesso dal padre al figlio maschio	Matrilineare: trasmesso dalla madre a tutti i figli (maschi e femmine)
Principale fonte di variabilità	Accumulo sequenziale di SNP	Accumulo sequenziale di SNP
Tasso di mutazione per gli SNP	ordine di 10^{-9} per posizione nucleotidica per anno	ordine di 10^{-8} per posizione nucleotidica per anno

SNP: Single Nucleotide Polymorphism = Polimorfismo a singolo nucleotide

Variabilità e trasmissione di Y e mtDNA



Gli autosomi vanno incontro a crossing-over → mescolamento dei loci

Y e mtDNA vengono trasmessi inalterati da una generazione all'altra

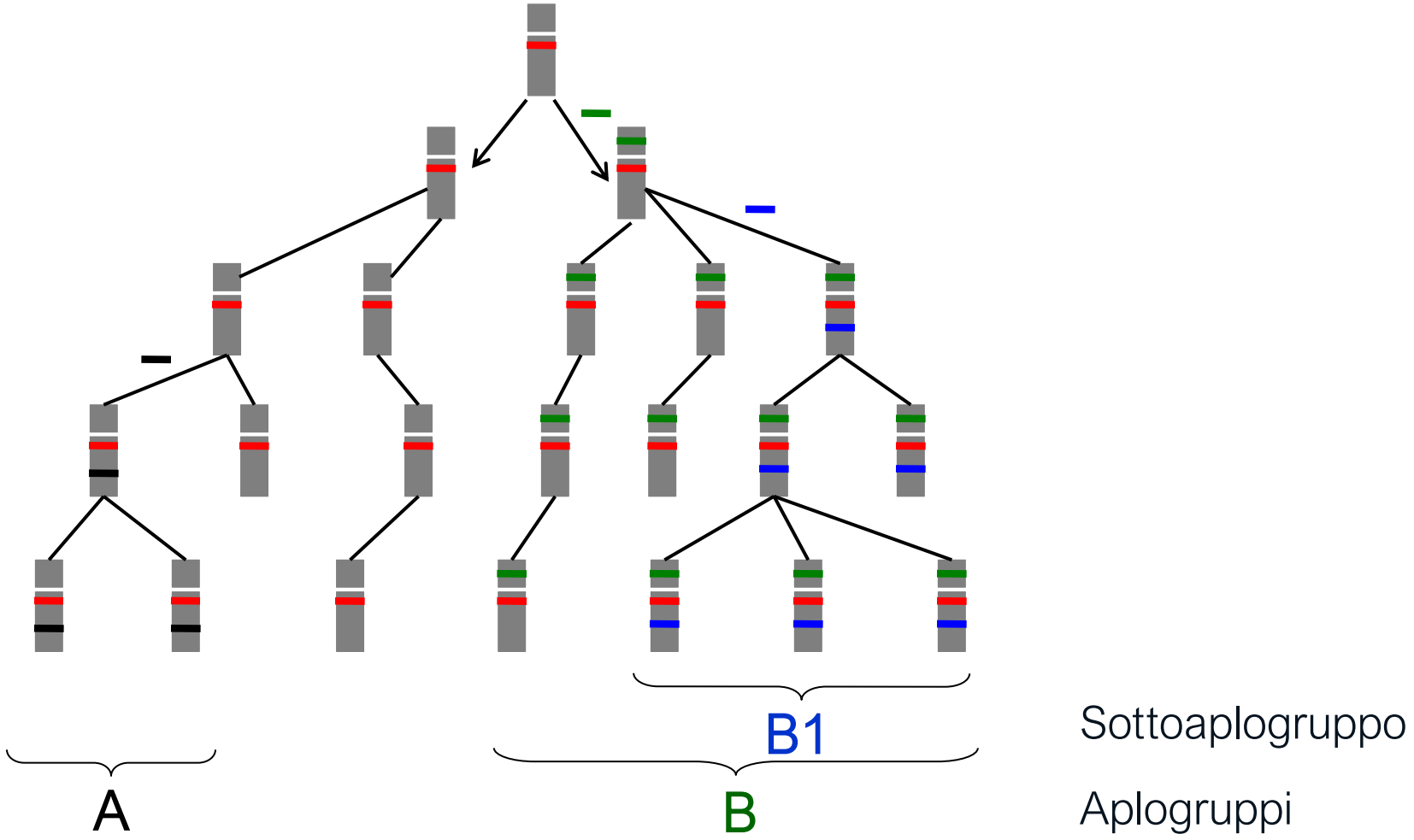
Inoltre, vengono trasmessi da uno solo dei due genitori:

- l'Y solo dal padre ai figli maschi;
- l'mtDNA solo dalla madre ai figli (maschi e femmine)

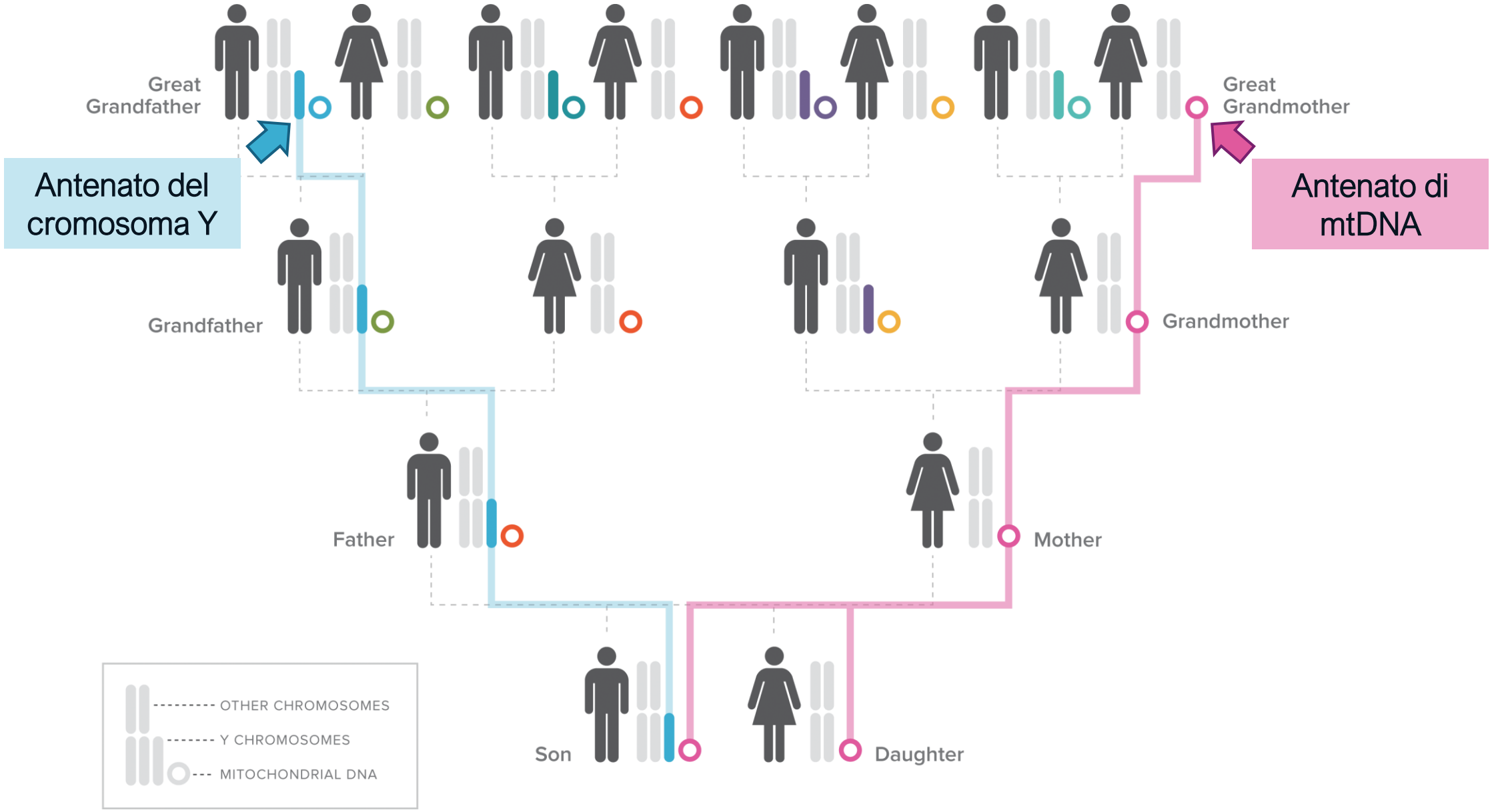
Nel corso delle generazioni, la principale fonte di variabilità di Y e mtDNA è **l'accumulo sequenziale di nuove mutazioni** (SNP), a causa dell'assenza di ricombinazione meiotica in questi sistemi

Il concetto di aplogruppo

Aplogruppo: gruppo di cromosomi definiti da un unico set di mutazioni acquisite da uno stesso antenato comune

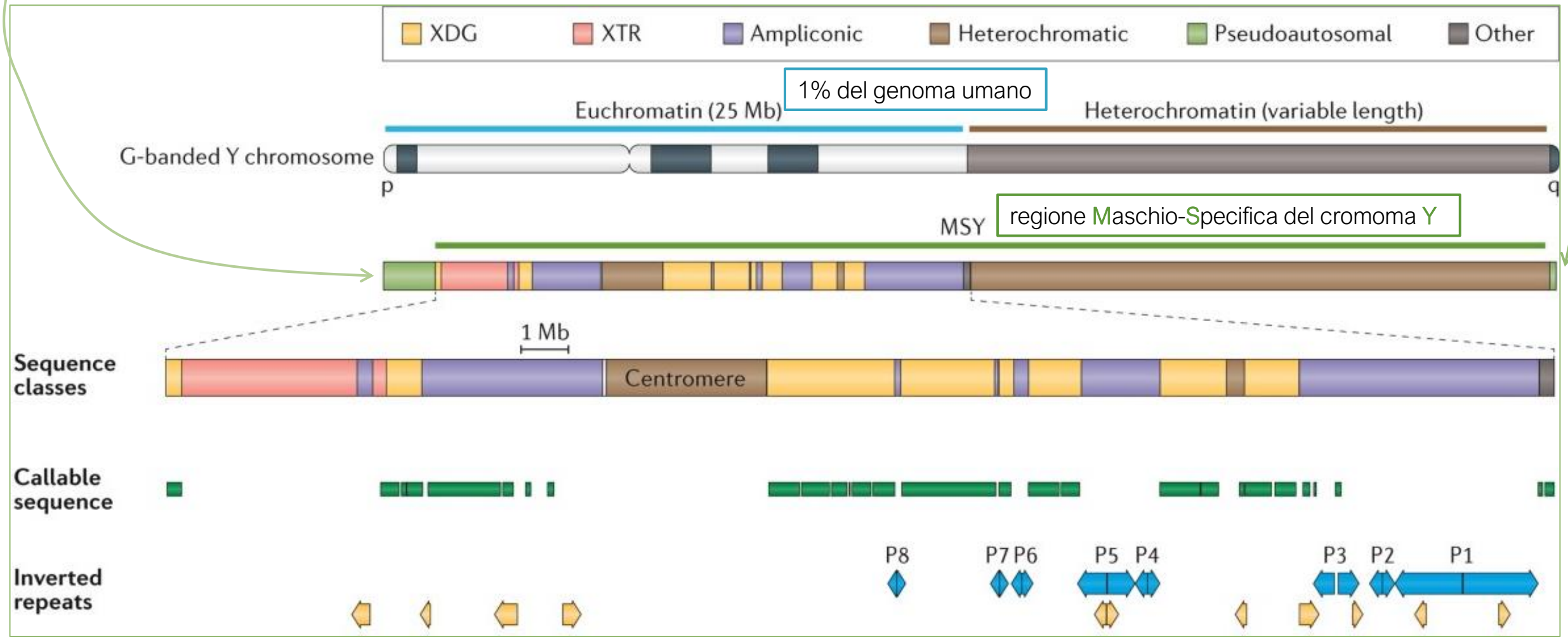


Le linee di ascendenza



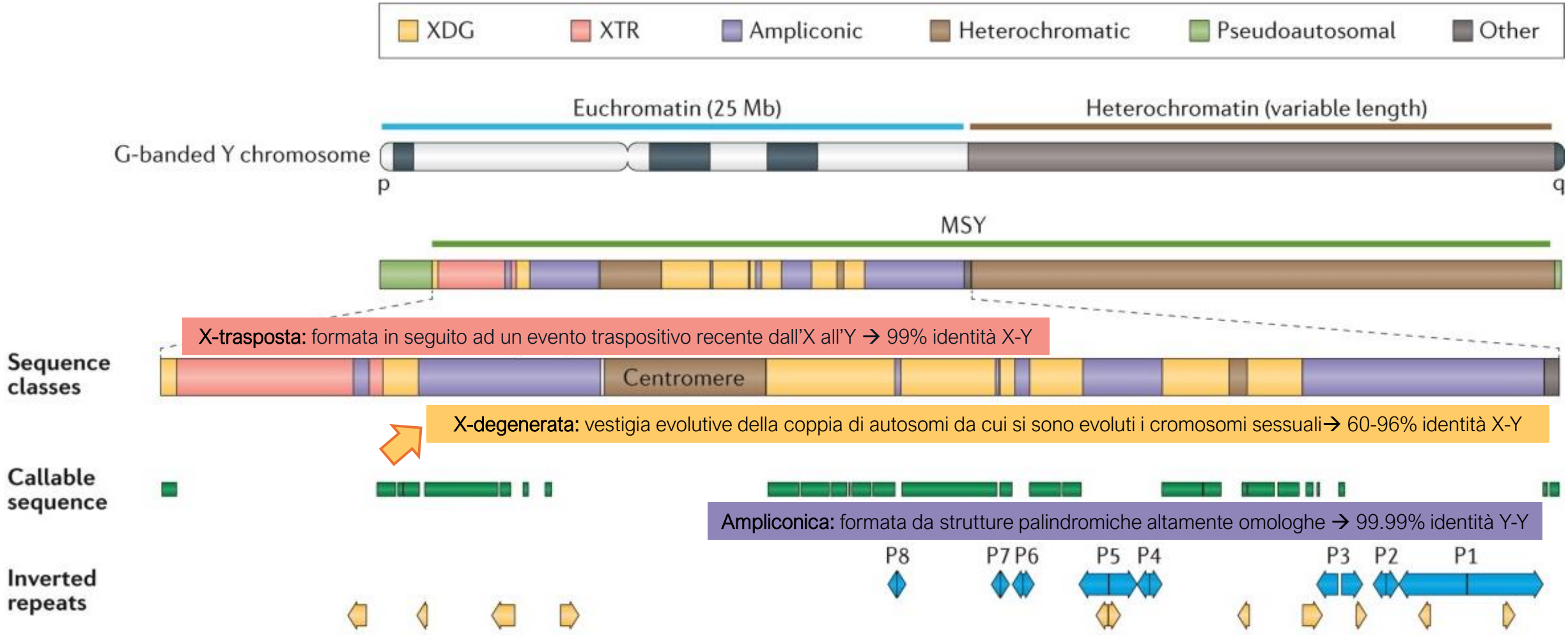
La struttura del cromosoma Y umano

PAR: Regioni Pseudo-Autosomiche → ricombinano con le regioni omologhe del cromosoma X durante la meiosi maschile: **trasmesse come le regioni autosomiche**
(il crossing over in PAR1 è obbligato per una corretta segregazione dei cromosomi sessuali)

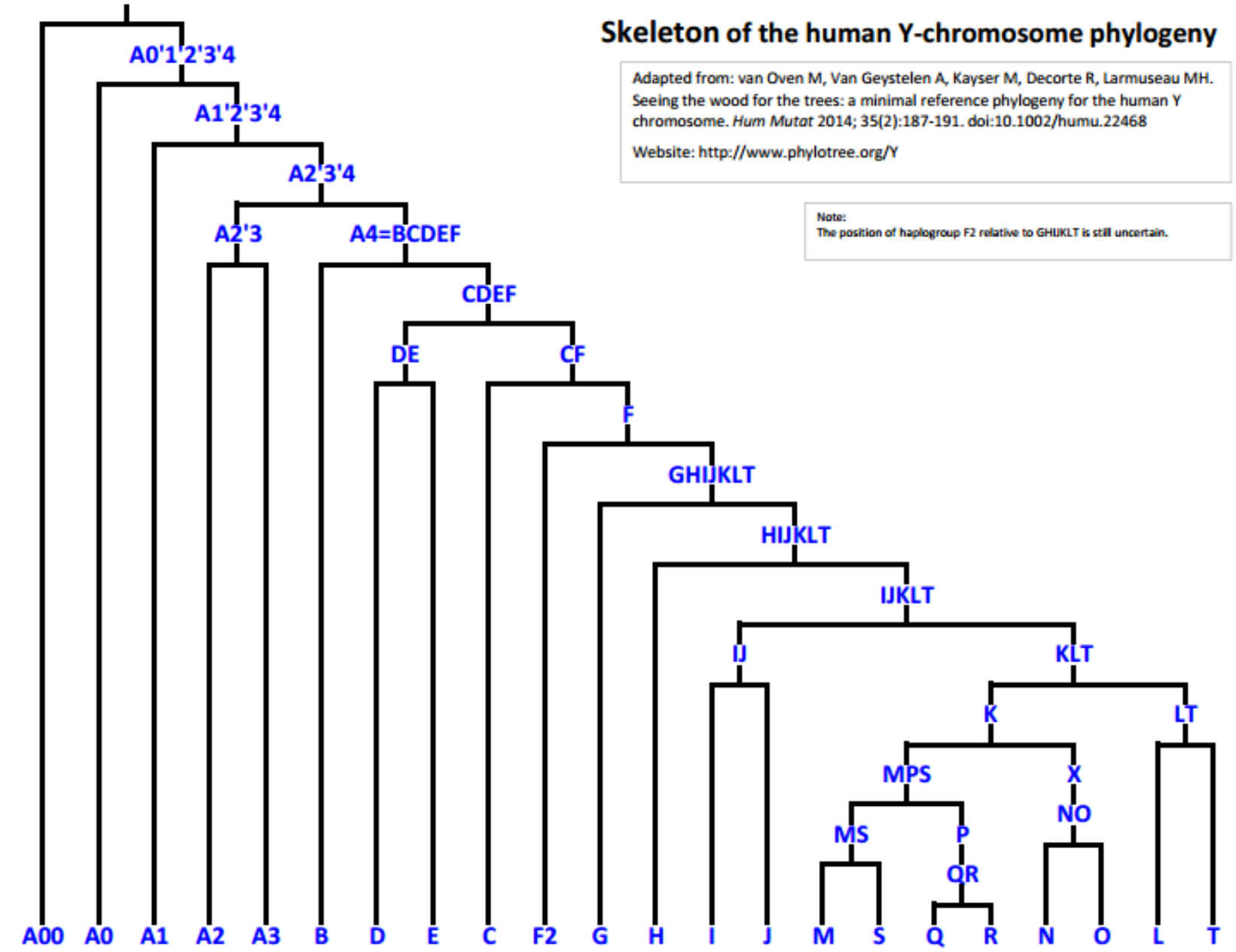


La porzione eucromatica del cromosoma Y

La porzione eucromatica del cromosoma Y è costituita da **tre classi di sequenze** → formate in momenti e con modalità diverse → diverso grado di identità di sequenza intra- o inter-cromosomico (con X).



Albero filogenetico del cromosoma Y



Nel corso del tempo, l'**identificazione di SNP del cromosoma Y umano** ha permesso di ricostruire le **relazioni filogenetiche** tra i vari cromosomi → **albero filogenetico** del cromosoma Y umano

Questo è solo lo "scheletro " dell'albero completo.

La **versione completa**:

1) Contiene **centinaia di linee** diverse;

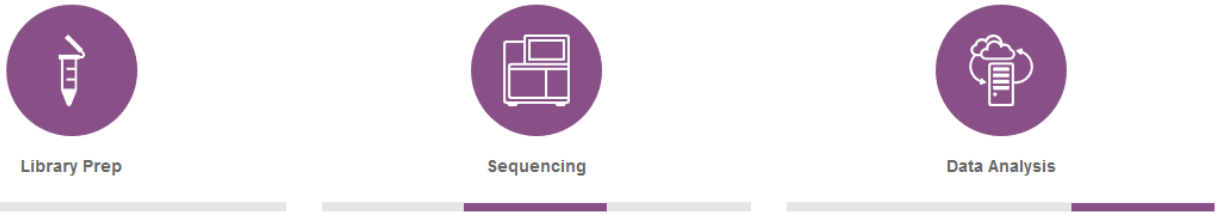
2) Le linee sono definite da **migliaia di SNP**;

3) È in costante aggiornamento grazie alla **grande quantità di dati provenienti dal sequenziamento di nuova generazione.**

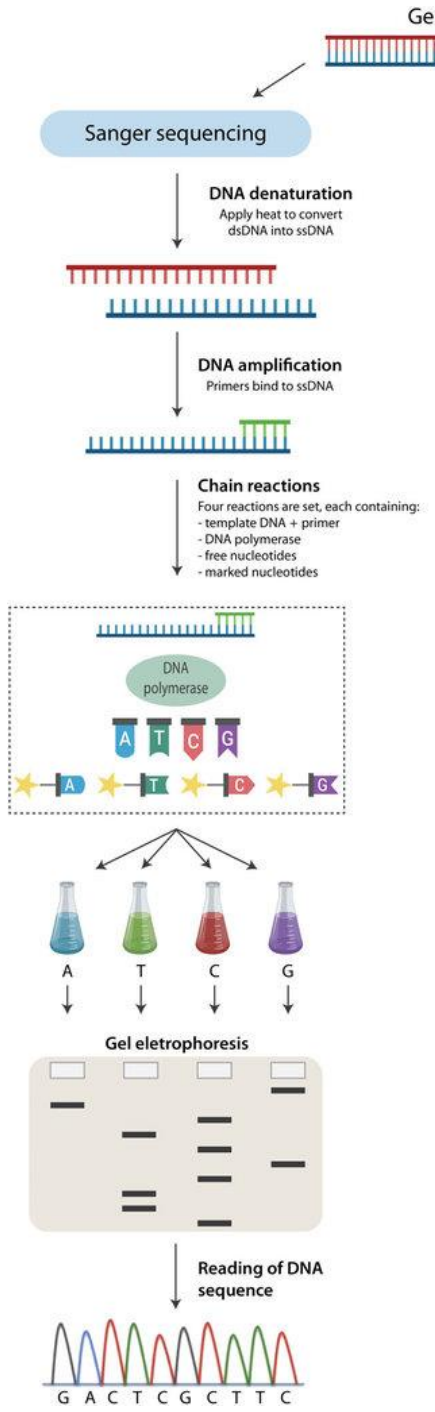
NGS: cenni generali

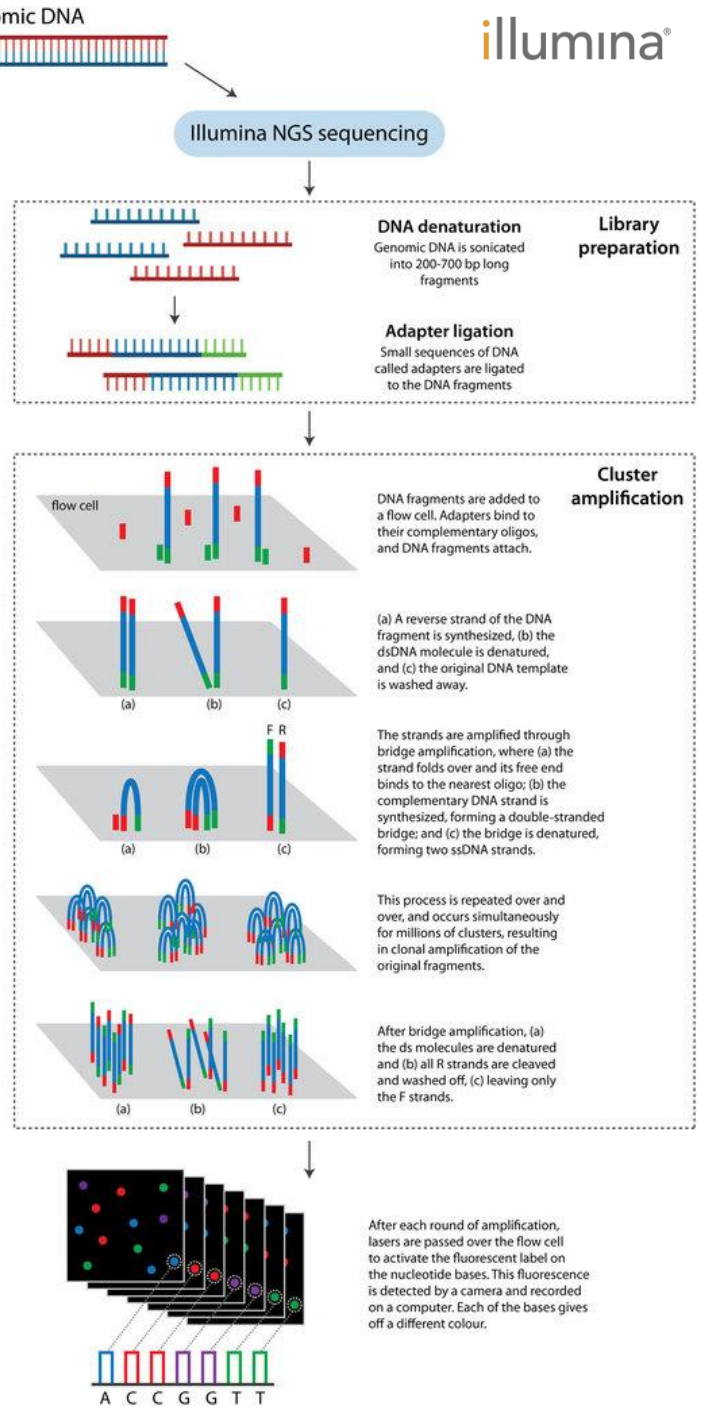
NGS = Next Generation Sequencing → non una tecnologia singola, ma un nome generico che include tecnologie di sequenziamento post-Sanger

Tecnologia NGS più usata → Illumina: miglior rapporto tra tempo di corsa, quantità di dati, qualità di dati e prezzo.



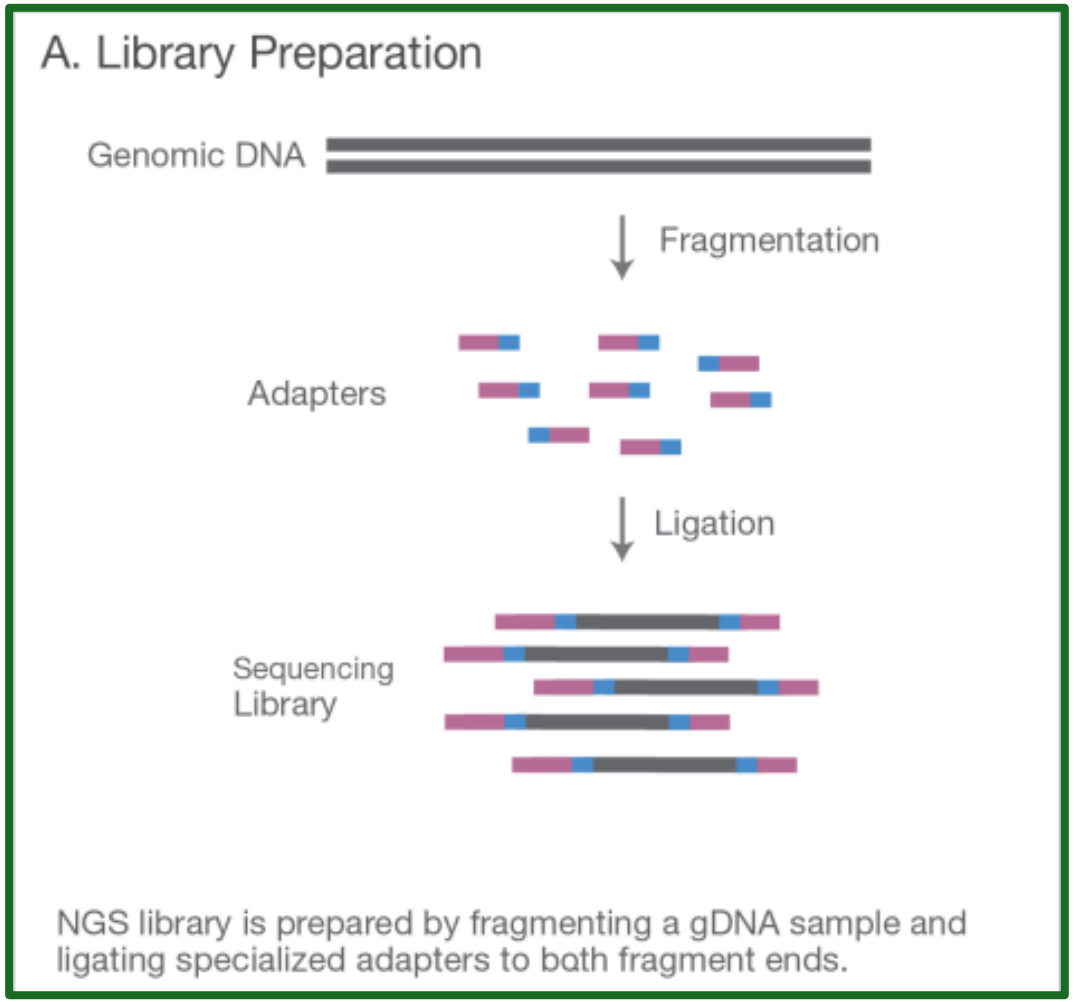
Platform	Platform cost	Running costs (per mb)	Analysis time: Wet laboratory	Analysis time: Run time	Analysis time: Dry laboratory	Data types	Read length	Output	Accuracy (consensus)	Key benefits and limitations
Sanger	$250 000	$500	8 h	20 min–3 h	5–10 min/per sample	ab1	<1000 bp	1 read per sequencing reaction	99.999%	Low throughput High accuracy
Illumina MiSeq	$125 000	$0.5	8–24 h	21–56 h	~15 min/1–50 samples	fastq	2 × 300 bp	15 Gb (~25 M reads)	99.99%	High-throughput Good accuracy Short reads



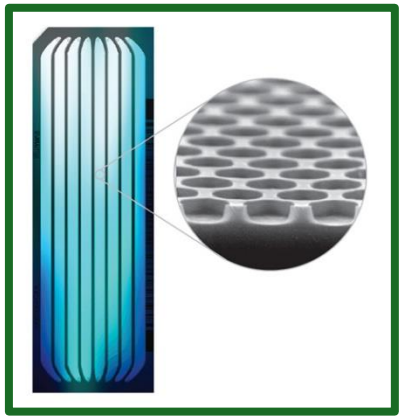


NGS: preparazione della library

Preparazione della library: frammentazione del DNA genomico in pezzetti di 100-200 bp → legame di corti oligonucletidi, detti **adapters**, alle due estremità di ciascun frammento



Caricamento della library in una flowcell: la flowcell è un supporto solido, simile ad un vetrino porta-oggetto, in cui sono scavati 8 canali, detti lane

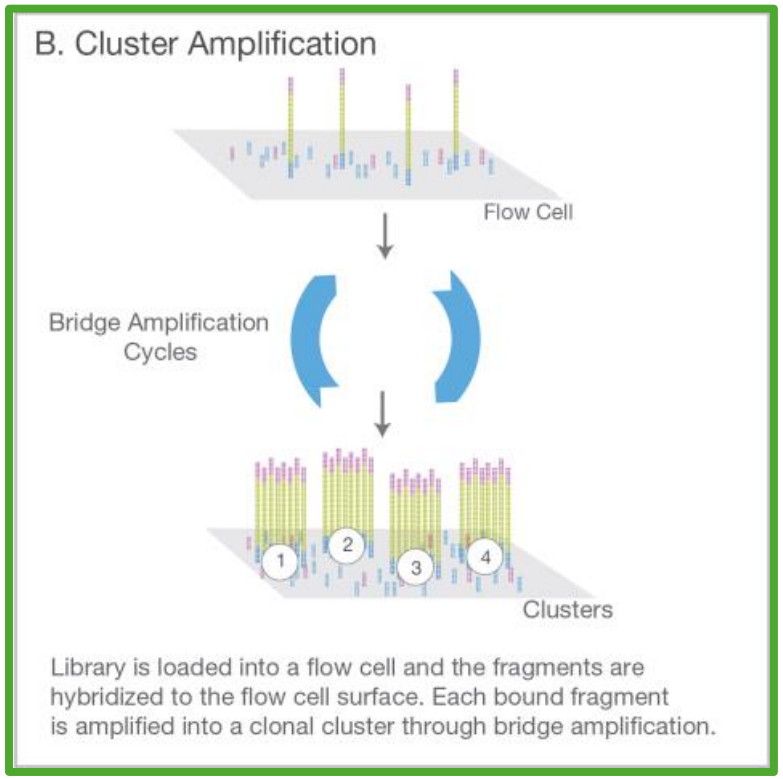


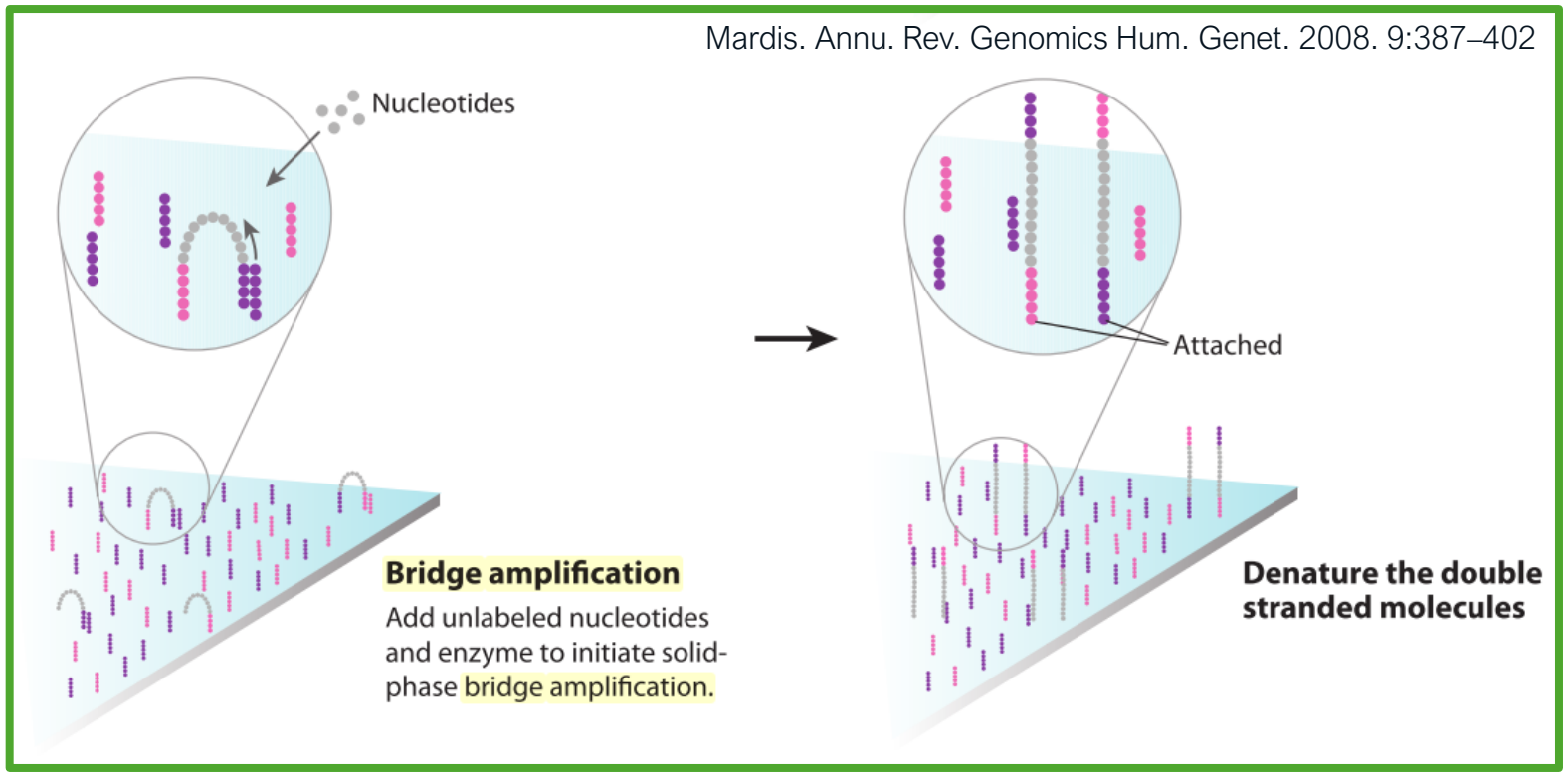
Ibridazione library + oligo: in ogni lane sono legati covalentemente degli oligonucleotidi complementari agli adapter

Video che spiega tutte le fasi del sequenziamento Illumina con animazioni (in inglese con possibilità di sottotitoli)
https://youtu.be/fCd6B5HRaZ8

NGS: generazione dei cluster

Generazione dei cluster: in ciascuna lane sono presenti oligonucleotidi complementari ad entrambi gli adapters (5' e 3') → formazione di strutture a ponte a singolo filamento → bridge amplification: usando l'adapter come primer ed il singolo filamento come stampo, viene sintetizzato il filamento complementare → formazione di strutture a ponte a doppio filamento → denaturazione dei due filamenti, ognuno fissato ad un adapter nella lane → cicli successivi di bridge amplification → formazione di cluster: migliaia di molecole clonate a partire dallo stesso frammento.

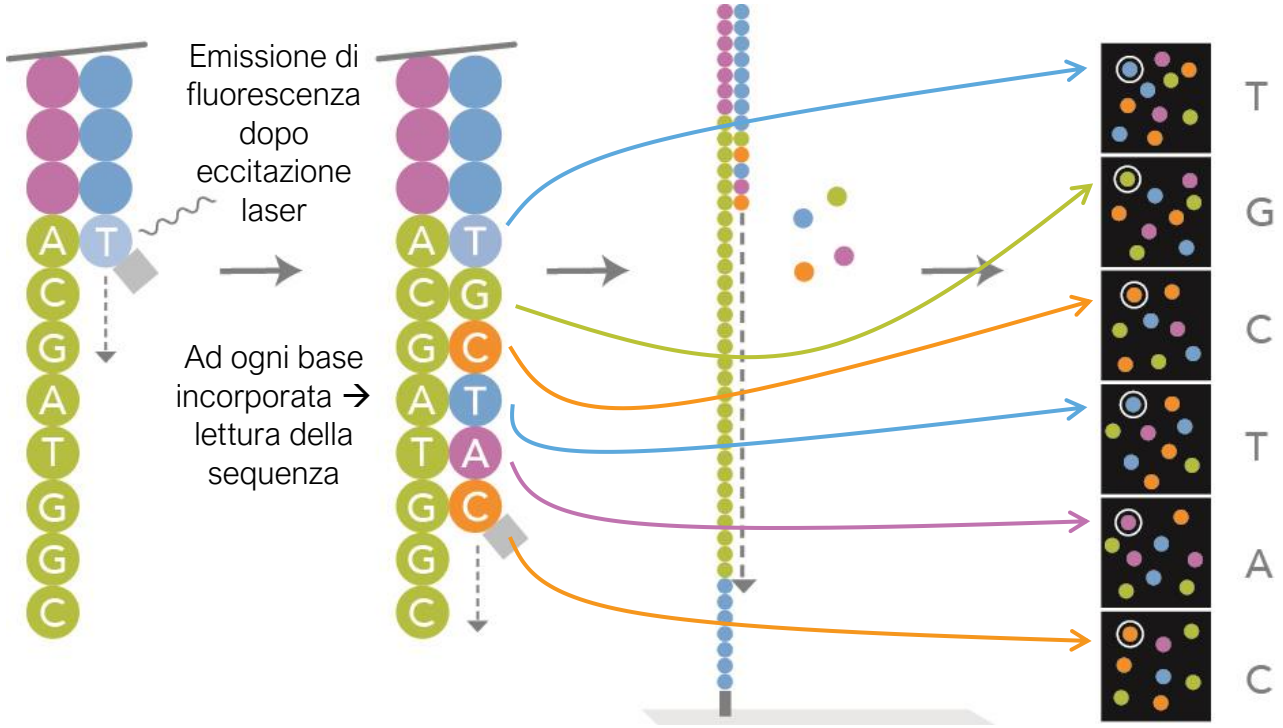


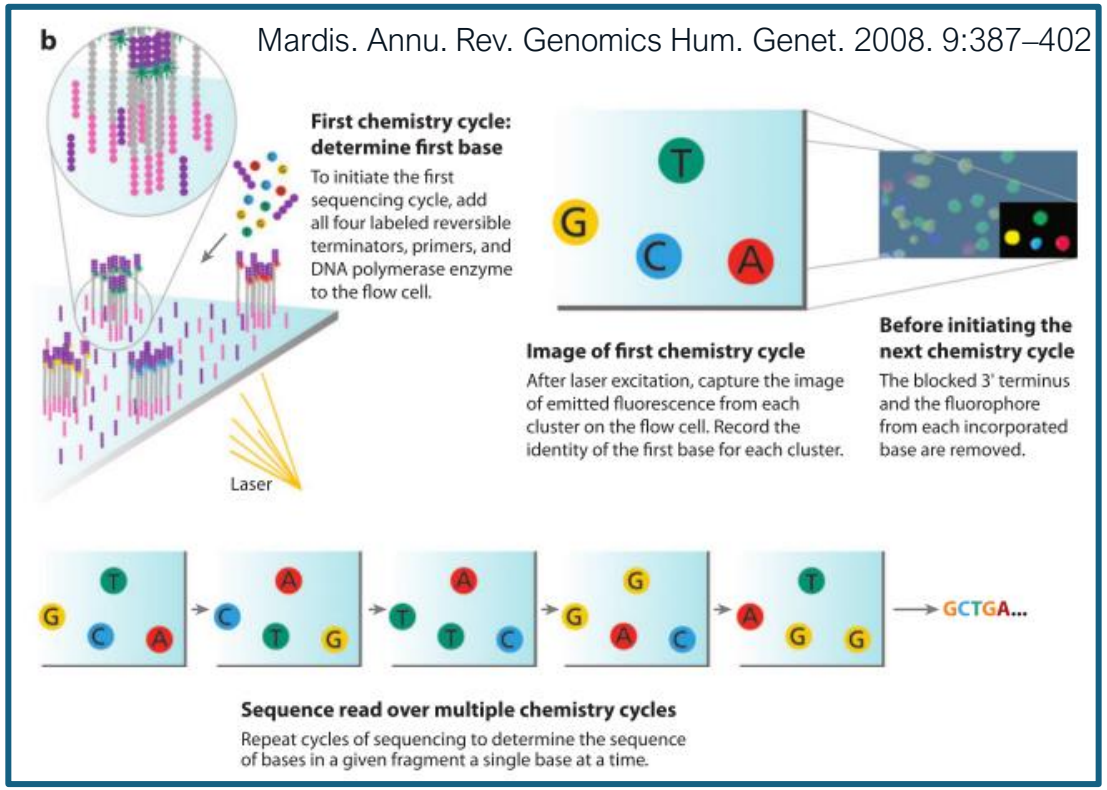


NGS: sequenziamento

Reagenti per il sequenziamento: inseriti nelle lane: primer complementare agli adapters, polimerasi, 4 nucleotidi → i nucleotidi hanno il 3'OH bloccato in modo reversibile e sono marcati da un fluoroforo

Reazione di sequenziamento: usando il primer come innesco, incorporato il primo nucleotide complementare alla prima base del frammento da sequenziare → stimolazione laser → rilascio di un segnale luminoso specifico da parte del fluoroforo → segnale catturato da una fotocamera → più cicli → ricostruzione della sequenza





NGS: allineamento

Allineamento: Allineamento delle corte (100-200 bp) letture ad una sequenza di riferimento

Muzzley et al. Curr Genet Med Rep 2015 3:158–165

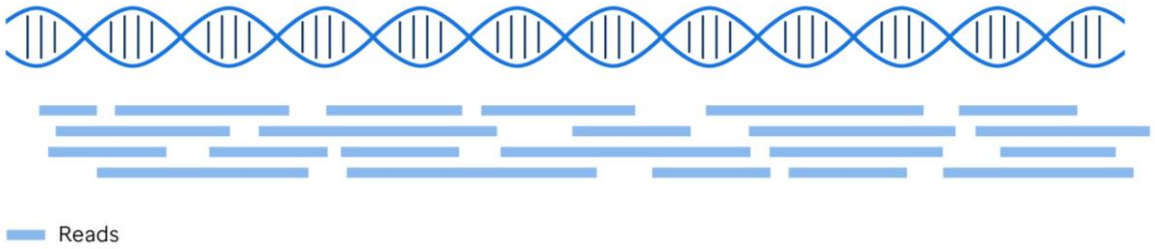
Depth: numero di volte in cui una base viene letta

SNP: identificati come differenza nucleotidica tra il campione e il riferimento

INDEL: inserzione/delezione di una singola base rispetto alla sequenza del riferimento

Possibili approcci di sequenziamento

Whole Genome Sequecing (WGS)

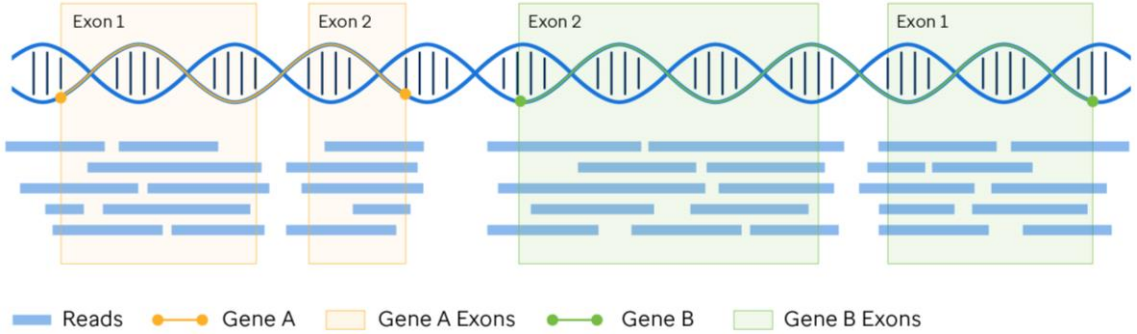


Sequenziamento dell'intero genoma

- depth: ~ 30 ×
- identificazione di nuove varianti

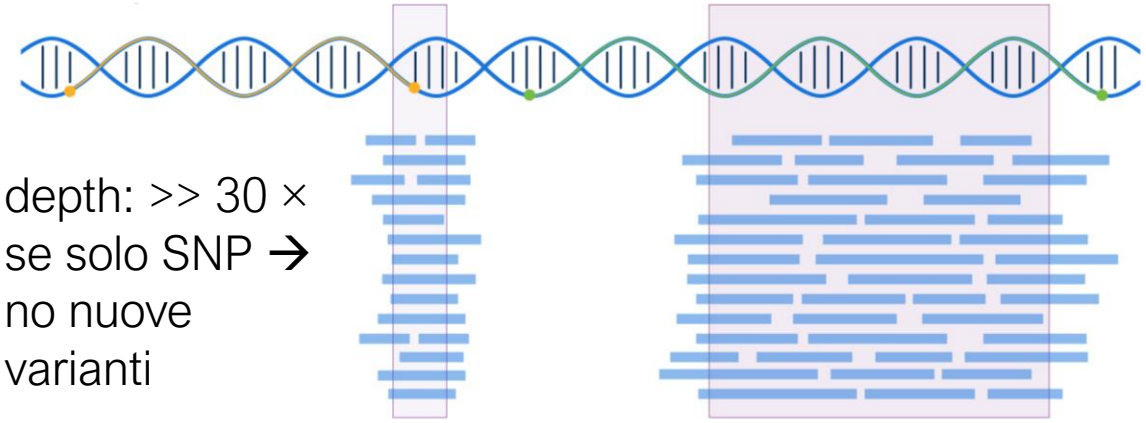
Sequenziamento di regioni selezionate

Whole Exome Sequencing (WES)



- depth: > 30 ×
- Identificazione di nuove varianti esoniche

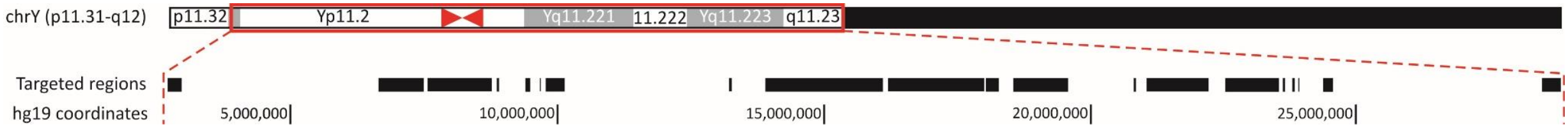
Target Sequencing (TS)



- depth: >> 30 ×
- se solo SNP → no nuove varianti

Custom target sequencing

Sequenziamento di regioni personalizzate: si fa ricorso a sonde disegnate appositamente per selezionare le regioni di interesse → es. **sequenziamento solo della regione X-degenerata del cromosoma Y**, evitando tutti gli elementi ripetuti



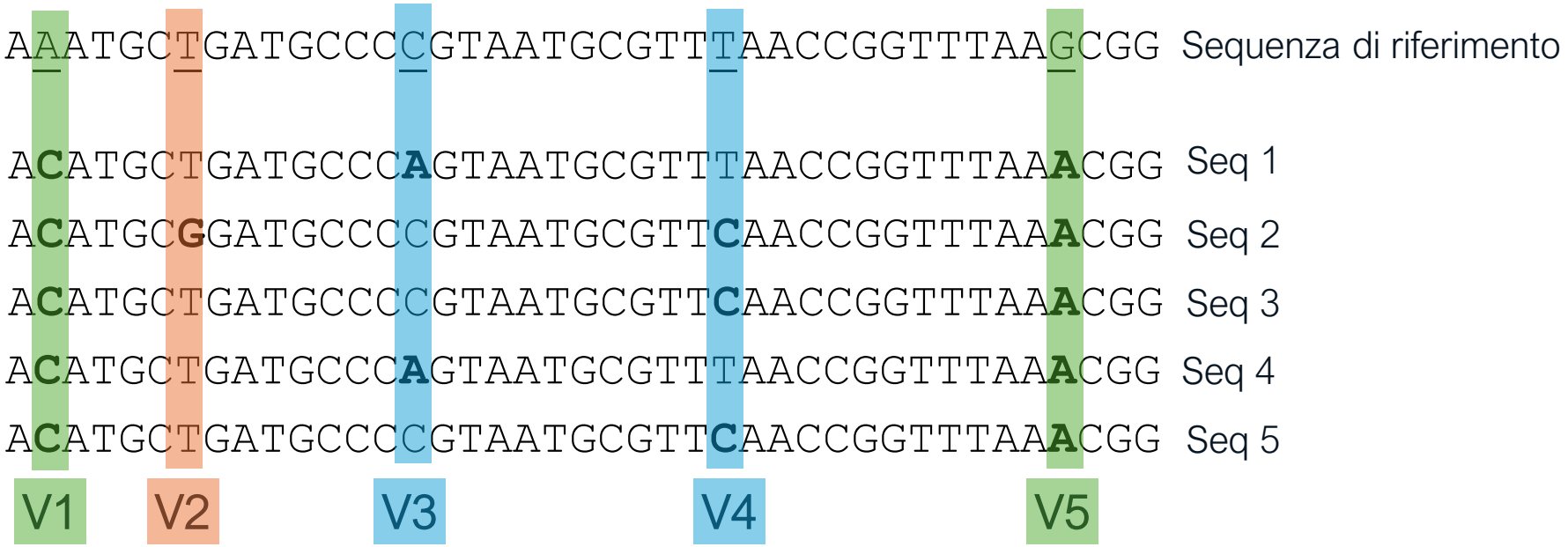
Perché si evitano gli elementi ripetuti?

1) Problemi di sequenziamento
2) Problemi di allineamento
3) Problemi di identificazione delle varianti

Perché solo regione X-degenerata?

Bassa identità di sequenza X-Y → relativamente facile identificare le varianti Y-specifiche

Costruzione di un albero filogenetico



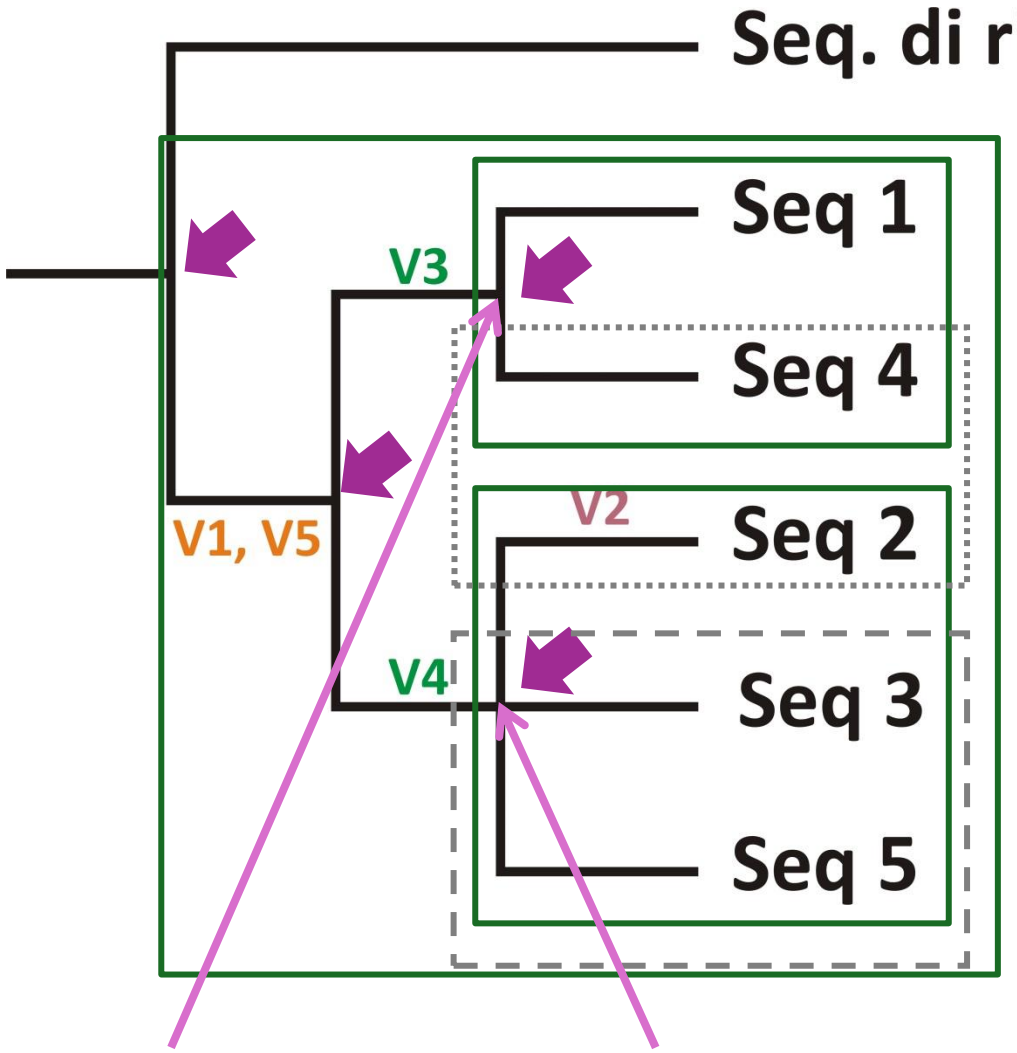
V1 e V5: condivise da tutte le sequenze analizzate, tutte diverse dalla sequenza di riferimento → V1 e V5 sono insorte nell'antenato comune dei 5 individui.

V3 e V4 sono condivise solo da un sottogruppo di sequenze: V3 è condivisa solo da Seq 1 e Seq 4, mentre V4 è condivisa solo Seq 2, Seq 3 e Seq 5 → sono insorte nell'antenato comune di alcune sequenze, ma non delle altre

V2 si trova solo in Seq 2: variante privata

L'approccio qui descritto segue il **principio di massima parsimonia.** Secondo questo principio, l'albero migliore è quello che richiede il **minor numero di eventi mutazionali per spiegare i dati osservati**.

Elementi di un albero



Rami

Nodi

Aplogruppo (o clade, taxon, etc): raggruppamento monofiletico di linee → le linee discendono da un antenato comune

Gruppo parafiletico: raggruppamento di linee discendenti da un antenato comune ma che ne esclude altre appartenenti allo stesso clade

Gruppo polifiletico: raggruppamento di linee appartenenti a diversi cladi

Biforcazione: da un nodo si dipartono due linee

Multiforcazione: da un nodo si dipartono tre o più linee

Datazione molecolare dei nodi dell'albero

La **variabilità genetica** lungo le linee dell'albero **si accumula in modo sequenziale e costante nel tempo** → la **variabilità genetica** può essere interpretata come una **misura del tempo** trascorso dalla separazione di due (o più) linee → **datazione molecolare dei nodi**

Orologio molecolare: l'evoluzione di sequenza di due linee filogenetiche avviene ad un tasso costante, così che è possibile correlare la variabilità genetica accumulata con il tempo in cui le due linee si sono separate dallo stesso nodo → stima del tempo dal più recente antenato comune (TMRCA: Time to the Most Recent Common Ancestor).

$$Misura\ di\ variabilità\ genetica = \mu \times t$$

Tasso di mutazione

tempo

Si usa tendenzialmente la variabilità genetica accumulata in due tipi di loci:

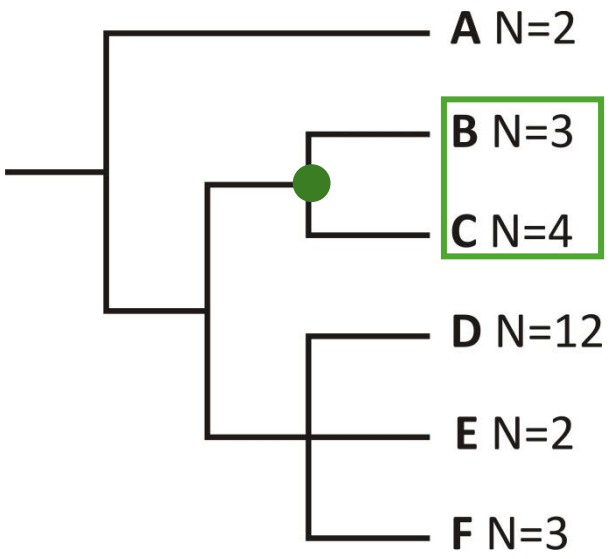
1) STR = Short Tandem Repeats;
2) SNP = Single Nucleotide Polymorphisms

Datazione usando gli STR: metodo ASD

Il metodo più usato per stimare la variabilità accumulata dagli STR è il metodo **ASD** = **A**verage **S**quared **D**istance

$$ASD = \frac{\sum_{i=1}^{N}(x_i - X)^2}{N}$$

x_i = numero di ripetizioni per un STR nell'i-esimo individuo
X = numero di ripetizioni più frequente (modale) per lo stesso STR
N = numero di individui appartenenti alle linee sottese al nodo da datare



Aplo e individui	STR1	STR2
B-Ind1	16	15
B-Ind2	16	20
B-Ind3	17	21
C-Ind1	16	20
C-Ind2	16	20
C-Ind3	12	20
C-Ind4	22	19

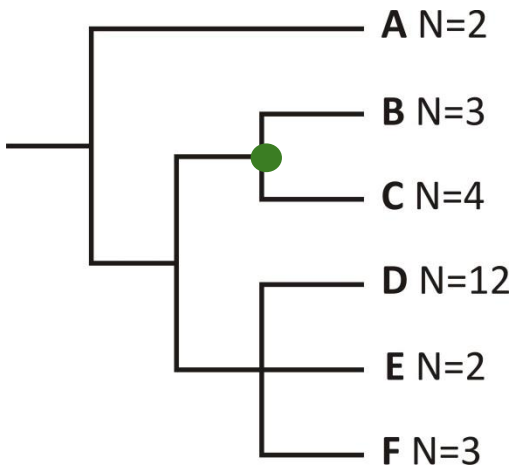
Ripetizioni modali:
$X_{STR1}=16$
$X_{STR2}=20$

N=7

$$ASD_{STR1} = \frac{(16-16)^2+(16-16)^2+(17-16)^2+(16-16)^2+(16-16)^2+(12-16)^2+(22-16)^2}{7} = 7{,}57$$

$$ASD_{STR2} = \frac{(15-20)^2+(20-20)^2+(21-20)^2+(20-20)^2+(20-20)^2+(20-20)^2+(19-20)^2}{7} = 3{,}86$$

Dall'ASD al tempo



$$\mathrm{ASD}_{\mathrm{STR1}} = 7{,}57$$

$$\mathrm{ASD}_{\mathrm{STR2}} = 3{,}86$$

$$\mathrm{ASD} = \frac{ASD_{STR1} + ASD_{STR2}}{2} = 5{,}72$$

$$\mathrm{ASD} = \mu \times t$$

Il tasso di mutazione dipende dai loci STR utilizzati.

In generale, il tasso di mutazione degli STR si calcola in due modi:

1) Sulla base di **dati raccolti da pedigree**: si misura la variabilità da una generazione alla successiva → **tasso genealogico**;
2) Sulla base di **dati popolazionistici** → **tasso evolutivo**

Tasso di mutazione genealogico ed evolutivo sono diversi:

1) Il **tasso genealogico è maggiore**: per deriva, si perde variabilità nel tempo → non considerata dal tasso evolutivo;
2) Il **tasso evolutivo è influenzato dalle dinamiche evolutive delle popolazioni**: per esempio, in una popolazione in forte espansione, la deriva è minore → si perde meno variabilità →il tasso evolutivo è maggiore, più simile al tasso genealogico

Problemi legati all'uso di STR

Questi loci mostrano un **tasso di mutazione elevato**:

1) Due individui con lo stesso numero di ripetizioni ad un locus STR non è detto che discendano da un antenato comune, ma possono averlo acquisito indipendentemente: gli alleli uguali in stato non è detto che siano uguali per discesa;
2) Gli eventi di **retromutazione** sono frequenti e ripristinano lo stato ancestrale → sottostima della variabilità accumulata;
3) Gli STR raggiungono un **plateau di variabilità**: il numero di ripetizioni non aumenta all'infinito ma raggiunge un equilibro tra mutazione (aggiunta di repeat) e retromutazione (diminuzione di repeat) → sottostima della variabilità accumulata.

Per questi motivi, **l'utilizzo degli SNP è preferibile**:

1) Tasso di mutazione (e retromutazione) molto più basso rispetto agli STR;
2) Tasso di mutazione meno soggetto alle dinamiche della popolazione

Differenze tra mtDNA e Y:

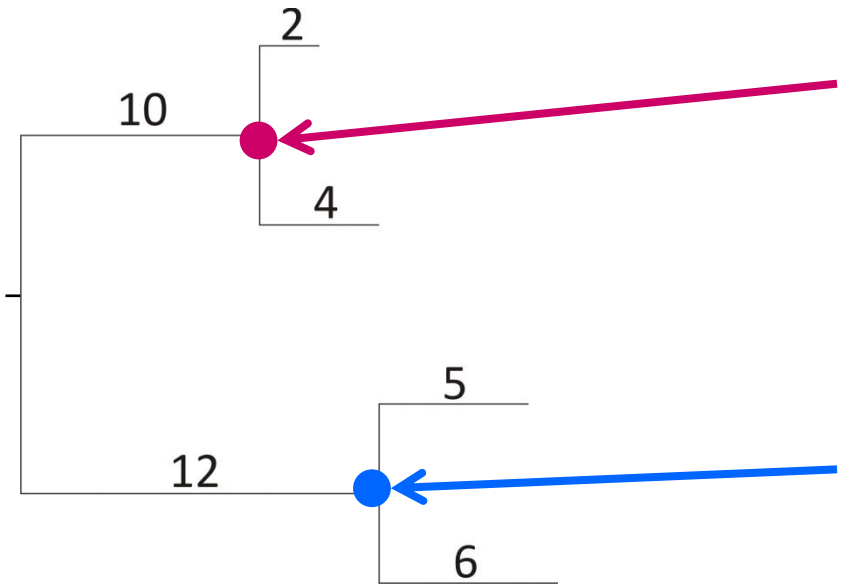
1) **mtDNA**: corto e tasso di mutazione maggiore → SNP facili da identificare: da sempre usati nelle datazioni;
2) **Y**: più lungo e tasso di mutazione minore → SNP difficili da indentificare, pochi: preferiti STR. Grazie all'avvento delle **nuove tecniche di sequenziamento (NGS: Next Generation Sequencing): molti SNP → utilizzati anche nell'Y**

Datazione utilizzando gli SNP: metodo Rho

Il metodo più usato per stimare la variabilità accumulata dagli SNP è il metodo **Rho (ρ)**

$$\text{Rho} = \frac{\sum_{i=1}^{N} x_i}{N}$$

x_i = numero di SNP lungo la i-esima linea sottesa al nodo
N = numero di linee sottese al nodo da datare



$$\rho_1 = \frac{2+4}{2} = 3 \qquad \rho_1 = \mu \times t_1$$

$\rho_1 < \rho_2 \rightarrow t_1 < t_2$

$$\rho_2 = \frac{5+6}{2} = 5{,}5 \qquad \rho_2 = \mu \times t_2$$

Quanto è μ?

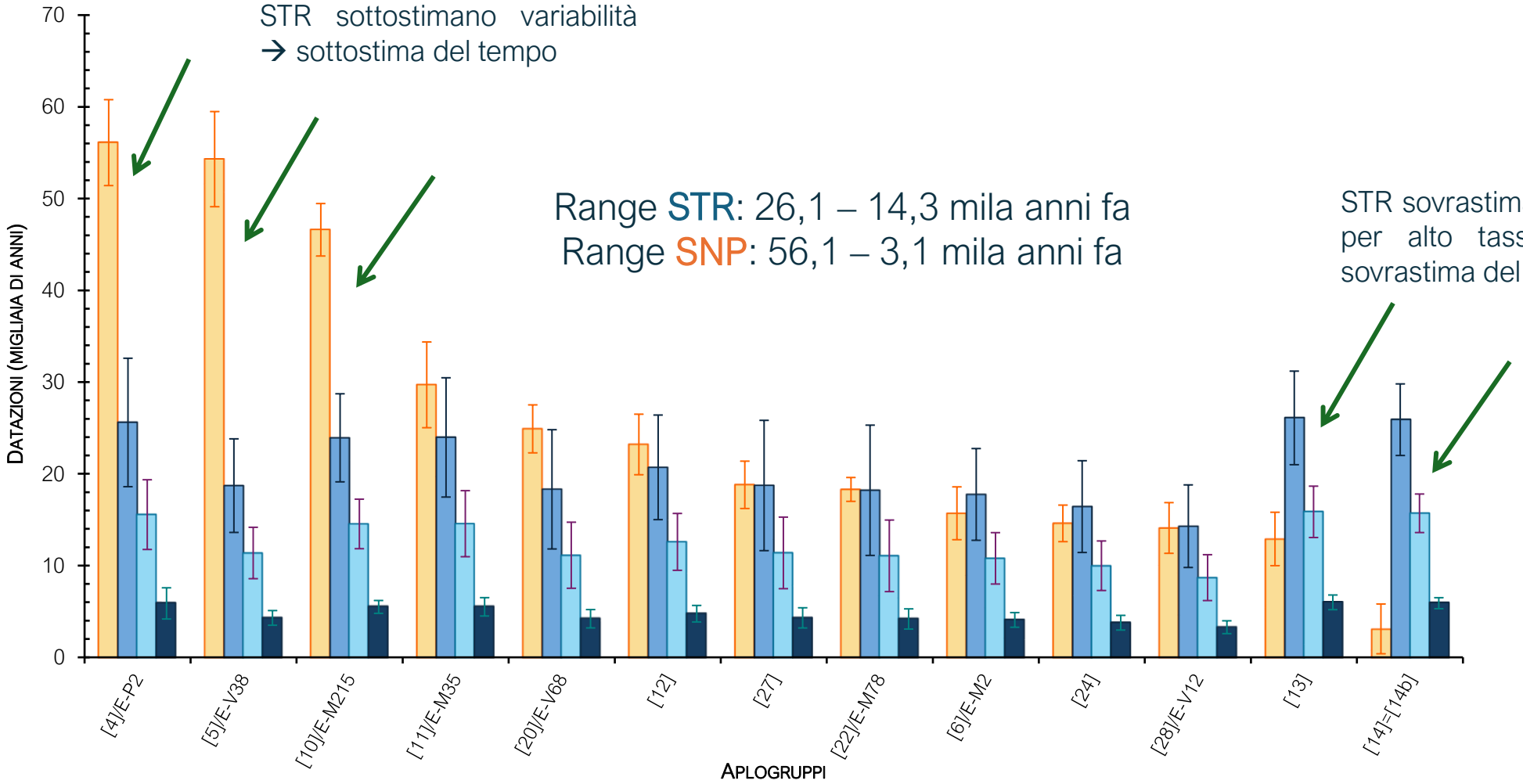
Sono stati adottati diversi metodi per il calcolo del tasso di mutazione del cromosoma Y → ognuno ha pro e contro.

Metodi di stima del tasso di mutazione

Metodo	Stima del tasso (per la regione X-degenerata dell'MSY)	Pro	Contro
Calcolato dai pedigree	0.9×10^{-9} mutazioni/posizione/anno	Calcolato direttamente	Non tiene conto della perdita di variabilità per effetto della selezione
Basato su dati antropologici e/o archeologici	0.75×10^{-9} mutazioni/posizione/anno	I resti archeologici sono datati con estrema precisione grazie a ^{14}C	La storia evolutiva di quella specifica linea rappresentata nei resti fossili potrebbe non essere rappresentativa della storia evolutiva del resto della filogenesi
Calcolato dal tasso di tutto il genoma	0.64×10^{-9} mutazioni/posizione/anno	Il tasso di tutto il genoma è basato su moltissime varianti ed è stato più volte confermato	Non è calcolato direttamente dall'Y

Confronto tra le datazioni STR vs SNP

- Datazioni in base agli **SNP**
- Datazioni in base agli **STR** con **tasso di mutazione evolutivo**
- Datazioni in base agli **STR** con **tasso di mutazione evolutivo per un modello di crescita esponenziale**
- Datazioni in base agli **STR** con **tasso di mutazione genealogico**



Datazioni in base agli SNP: metodo Rho
Datazioni in base agli STR: metodo ASD

Gli effetti della deriva genetica: cenni generali

Deriva genetica: fluttuazione nel tempo delle frequenze alleliche per solo effetto del caso → pesantemente influenzata dalla dimensione della popolazione.

$$H_t = H_0\left(1 - \frac{1}{2N}\right)^t$$

H_t = eterozigosità alla generazione t
H_0 = eterozigosità al tempo 0
N = dimensione della popolazione
t = numero di generazioni sotto deriva genetica

Quindi: **al diminuire di N, H_t diminuisce** → la deriva genetica porta ad una diminuzione dell'eterozigosità, fino al **fissaggio** di uno o l'altro allele.

In una **popolazione suddivisa** in sotto-popolazioni, di **piccole dimensioni** e sottoposte a **deriva genetica**, questa forza evolutiva agisce indipendentemente in ciascuna delle sotto-popolazioni → in alcune sottopopolazioni si fisserà un allele, in altre sottopopolazioni si fisserà l'altro allele → aumento della diversità tra sottopopolazioni = **aumento di F_{ST}**.

L'effetto finale della deriva genetica in una popolazione suddivisa è l'aumento della diversità tra le sottopopolazioni

Cromosoma Y e deriva genetica

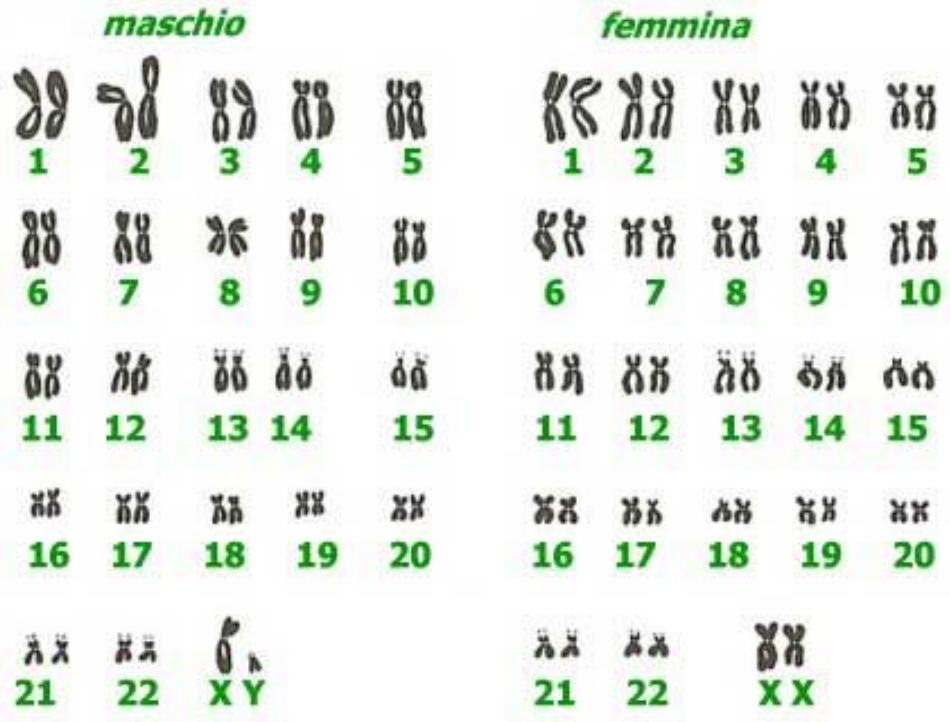
Dimensione efficace della popolazione: numero di individui che effettivamente constribuiscono al pool genico della generazione successiva → questo valore (più che la dimensione della popolazione in sè) ha effetti sulla deriva genetica

Considerando l'assetto cromosomico umano, in una **popolazione con N maschi e N femmine**:

- **4N assetti autosomici**: **2N** nei maschi **+ 2N** nelle femmine

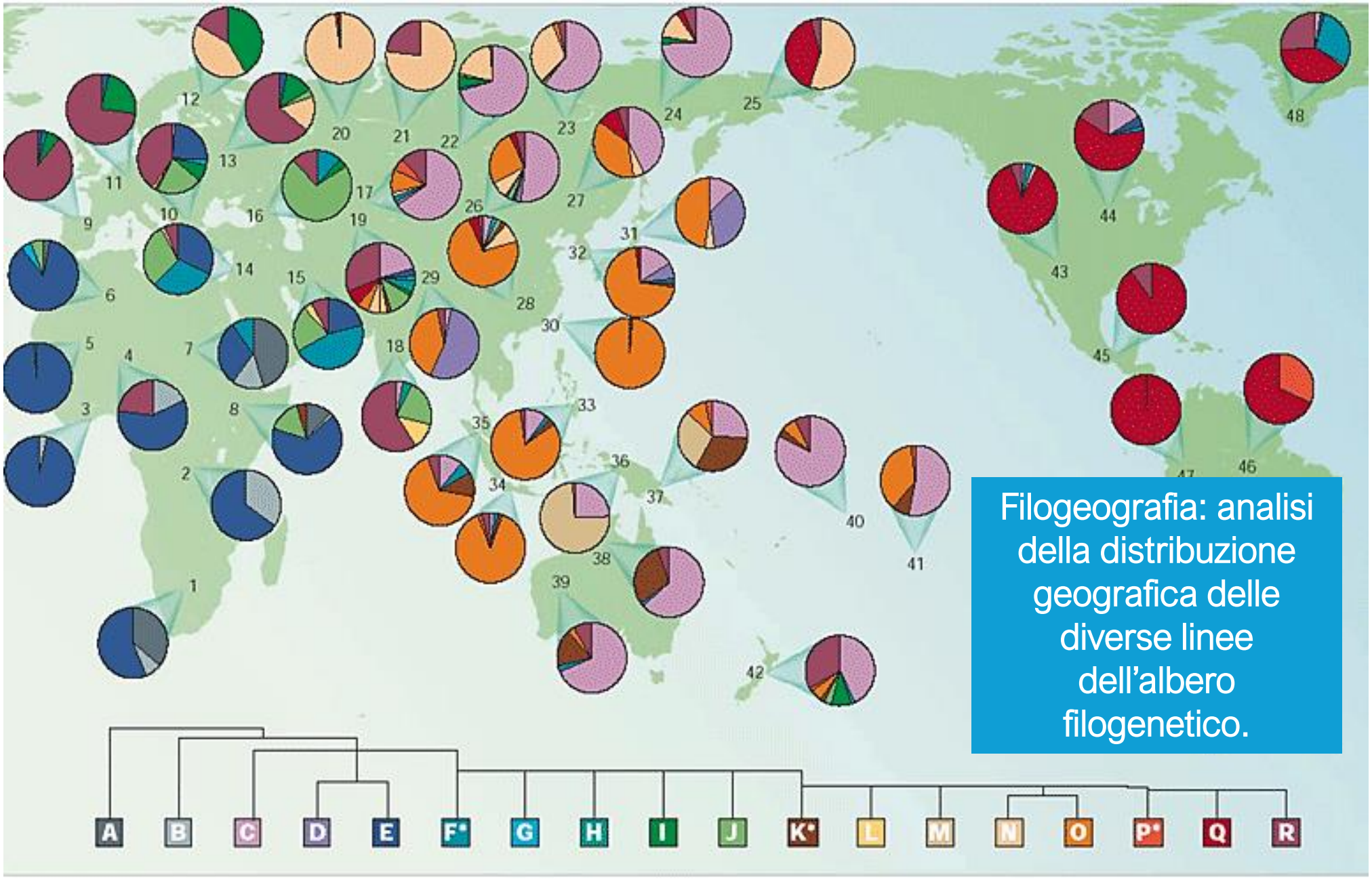
- **3N cromosomi X**: **N** nei maschi **+ 2N** nelle femmine

- **N cromosomi Y**: **N** nei maschi **+ 0** nelle femmine



Quindi la **dimensione efficace del cromosoma Y è ¼ rispetto a quella degli autosomi** → gli effetti della deriva genetica sono maggiori → la differenza interpopolazione è maggiore

La filogeografia

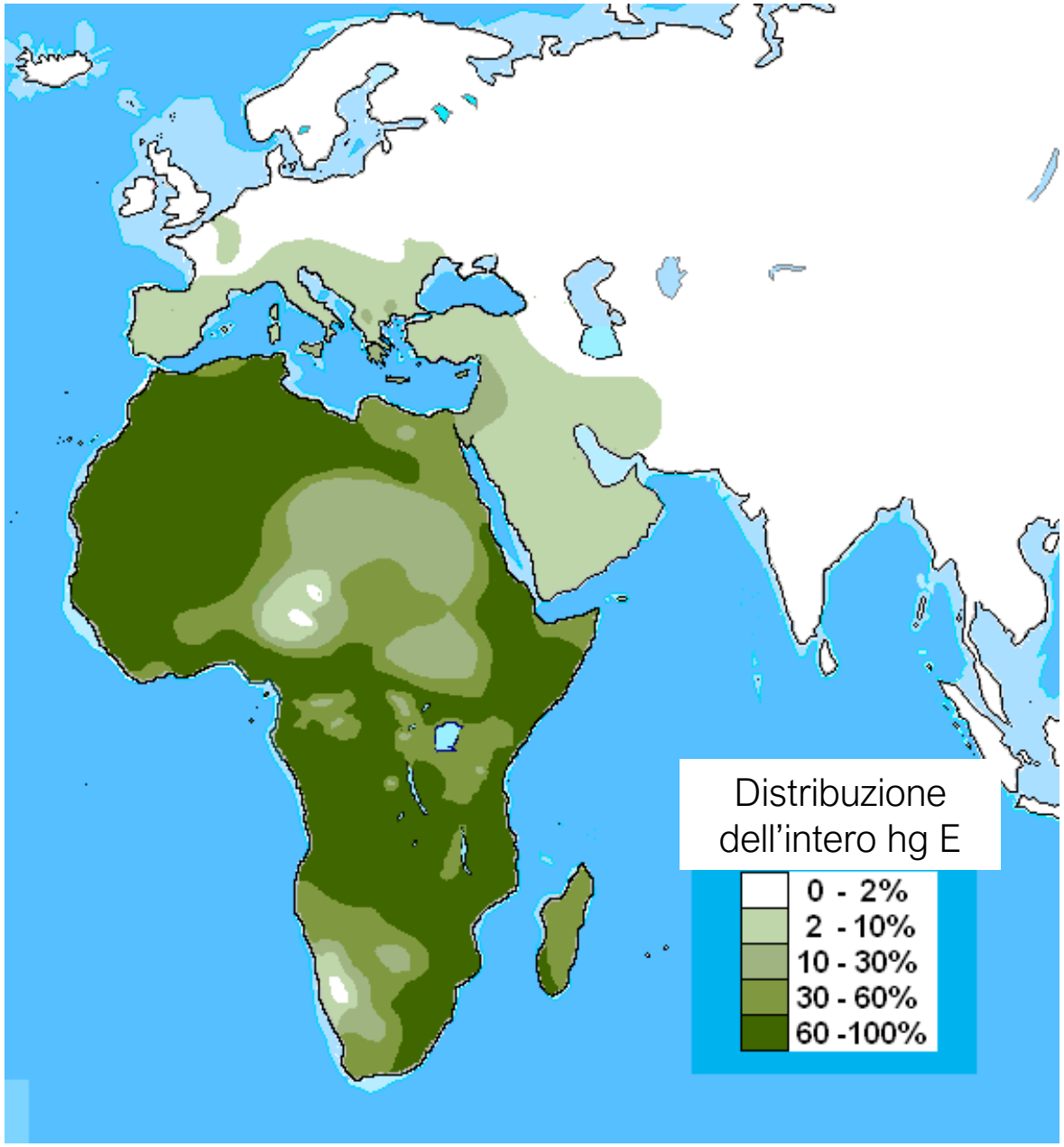


Filogeografia: analisi della distribuzione geografica delle diverse linee dell'albero filogenetico.

A causa dell'effetto intensificato della deriva genetica, **ciascuna popolazione** è caratterizzata dalla presenza di **aplogruppi specifici**, assenti (o presenti di rado) in altre popolazioni.

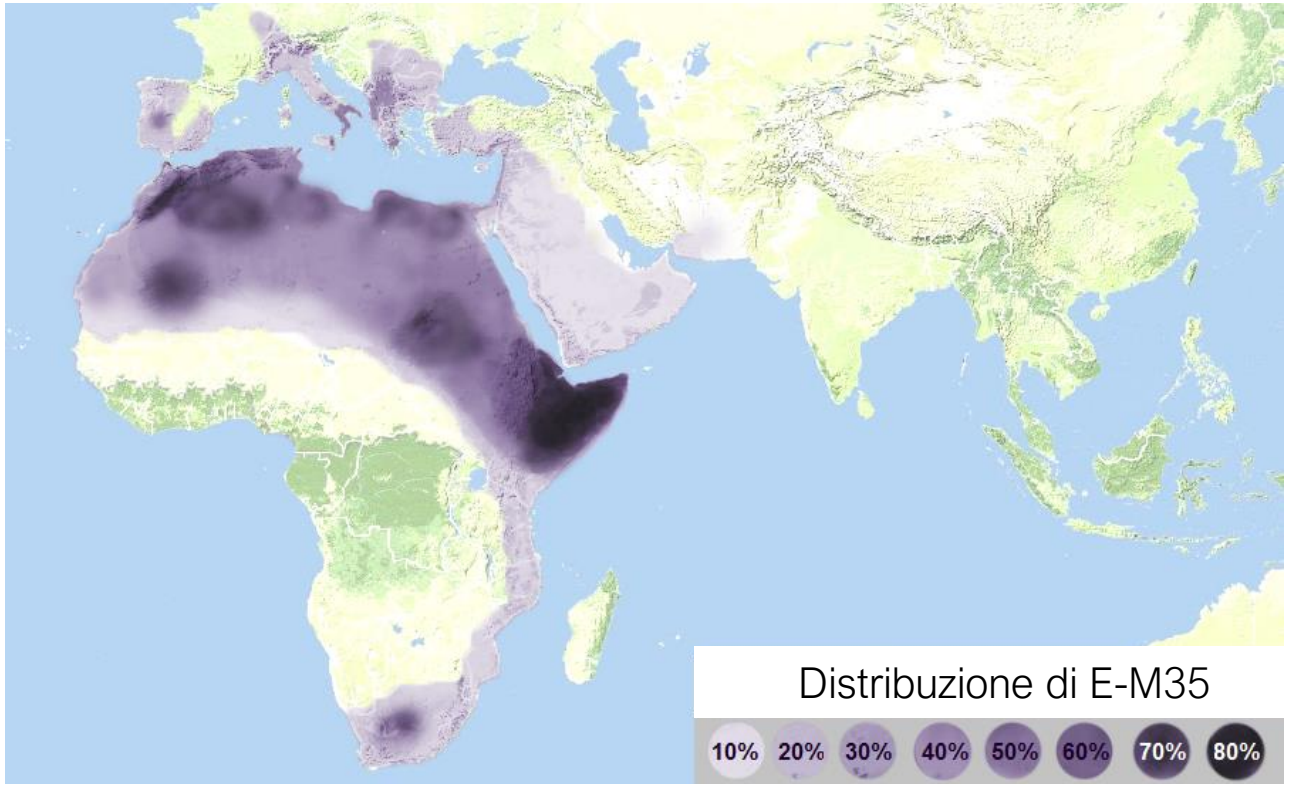
Di conseguenza, è possibile correlare le caratteristiche geografiche (distribuzione) e genetiche (aplogruppi) delle popolazioni.

La diffusione della pastorizia in Africa

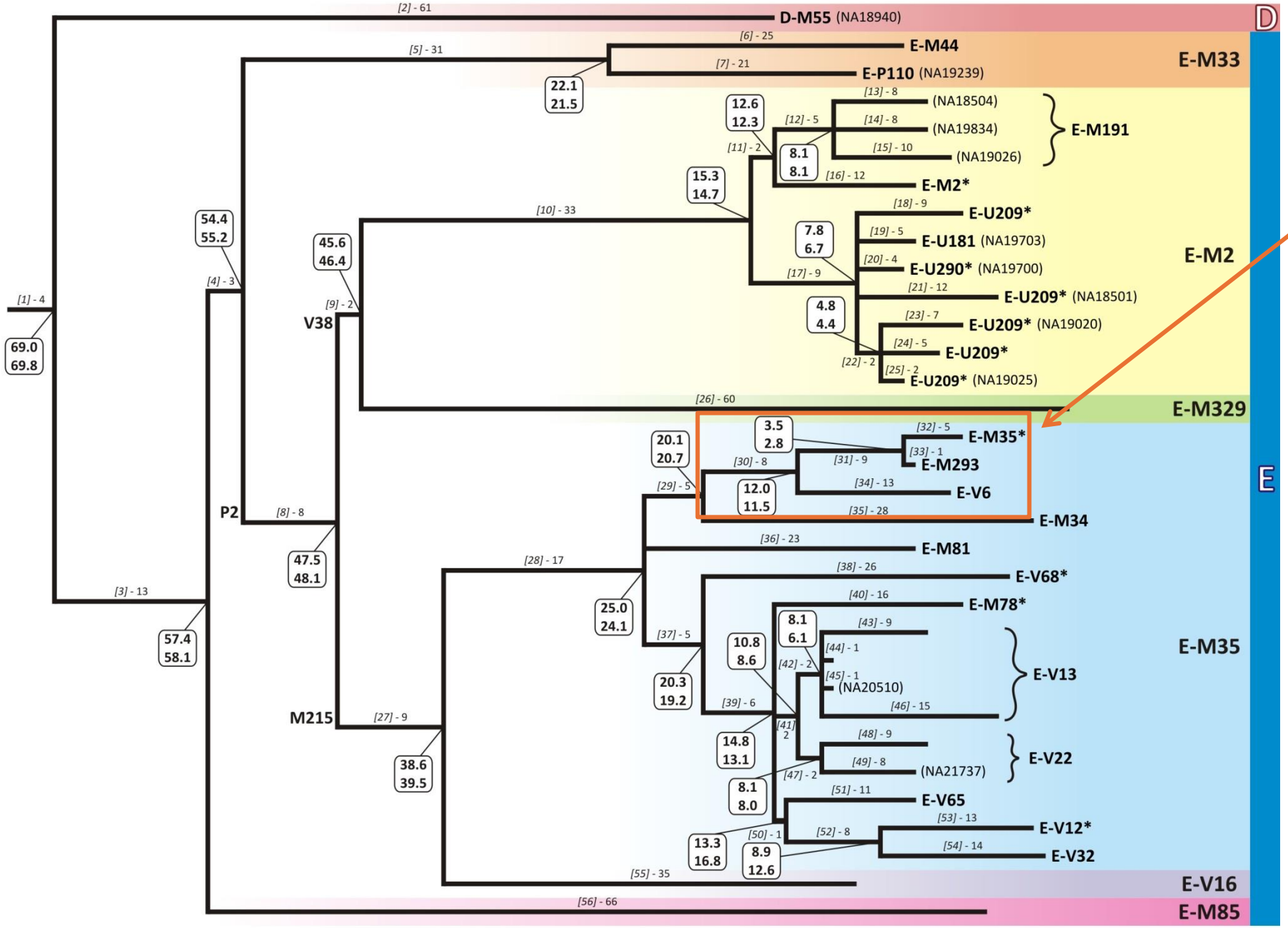


L'aplogruppo **E** è il **più frequente aplogruppo in Africa**

All'interno dell'E, la linea **E-M35** mostra un'**ampia distribuzione geografica** dentro e fuori il continente africano



Analisi di E-M35 tramite target sequencing



Linee frequenti in popolazioni pastorali dell'Africa orientale

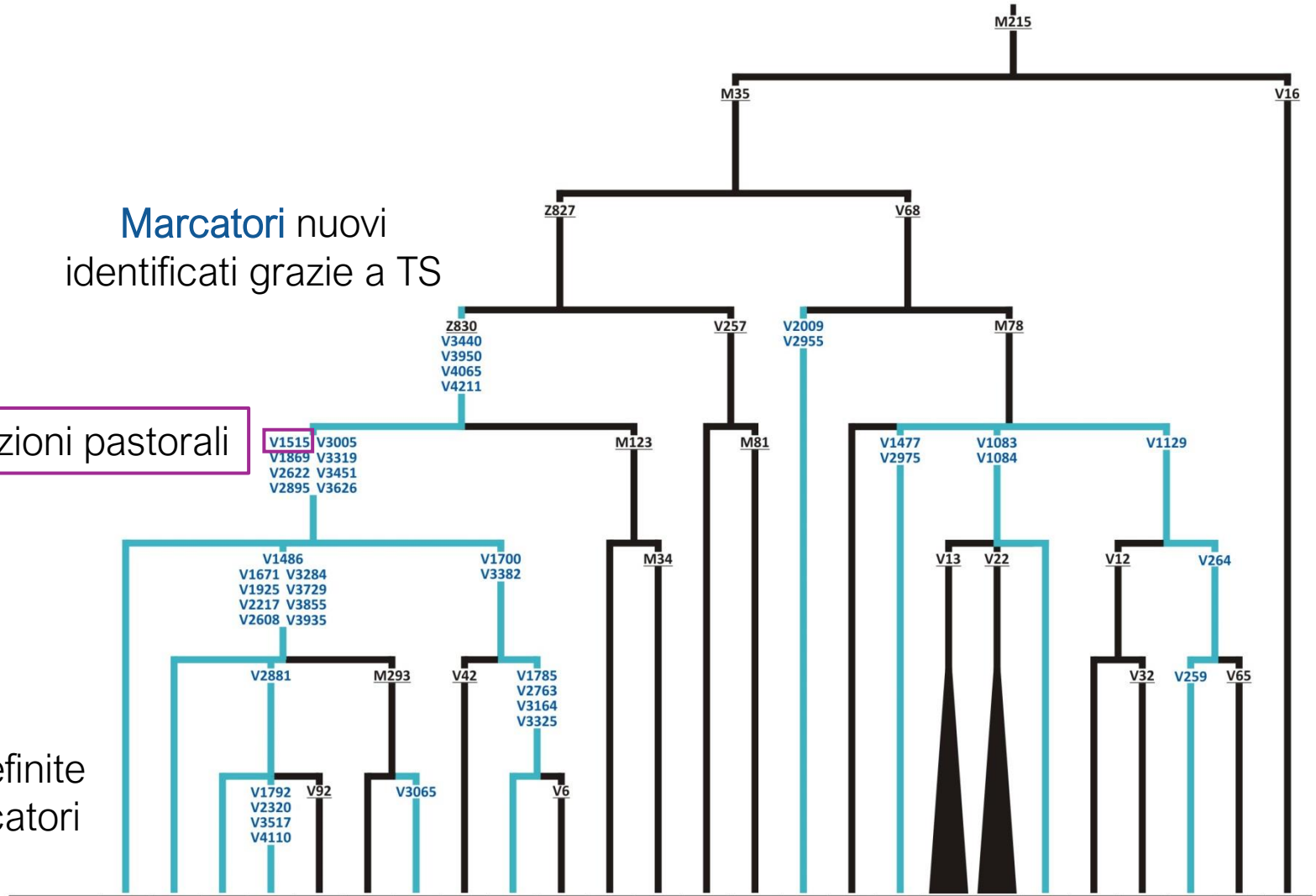
Datazioni coerenti con la storia nota dell'arrivo della pastorizia in Africa orientale

Caratterizzazione delle linee interne

Marcatori nuovi identificati grazie a TS

Linea in popolazioni pastorali

Linee nuove definite da questi marcatori

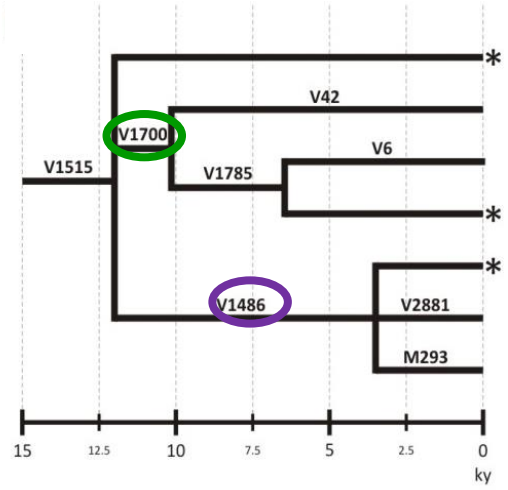


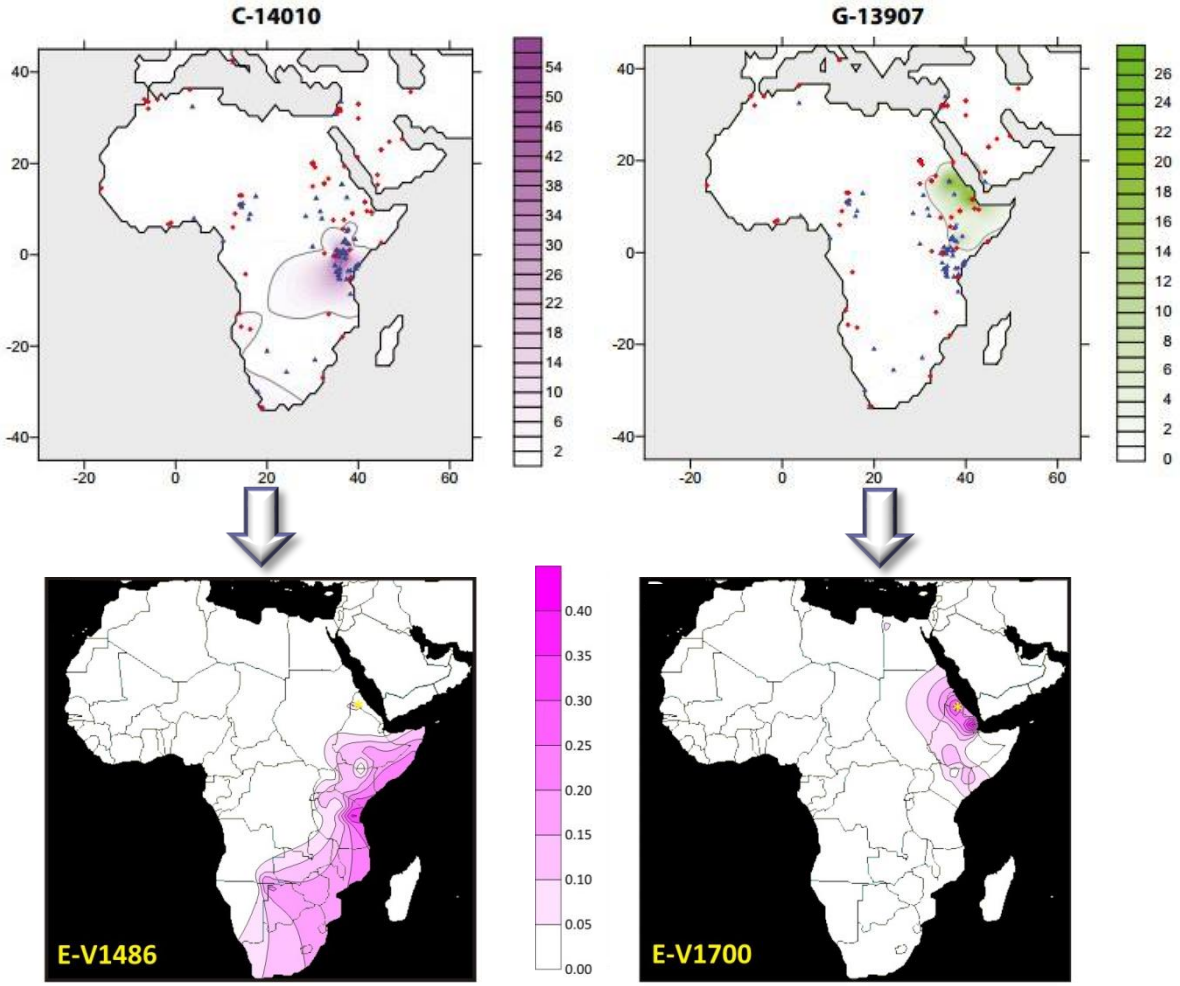
N	4	2	19	1	1	38	1	10	12	47	4	104	15	353	6	2	1	174	157	2	64	81	2	40	6
Former Hg by SNP	E-M35*	E-M35*	E-M35*	E-M35*	E-V92	E-M293	E-M293	E-V42	E-M35*	E-V6	E-M123*	E-M34	E-V257*	E-M81	E-V68*	E-M78*	E-M78*	E-V13	E-V22	E-M78*	E-V12*	E-V32	E-M78*	E-V65	E-V16
New Hg by SNP	E-V1515*	E-V1486*	E-V2881*	E-V1792	E-V92	E-M293*	E-V3065	E-V42	E-V1785*	E-V6	E-M123*	E-M34	E-V257*	E-M81	E-V2009	E-M78*	E-V1477	E-V13	E-V22	E-V1083*	E-V12*	E-V32	E-V259	E-V65	E-V16

Diffusione della pastorizia: E-V1515 e lattasi

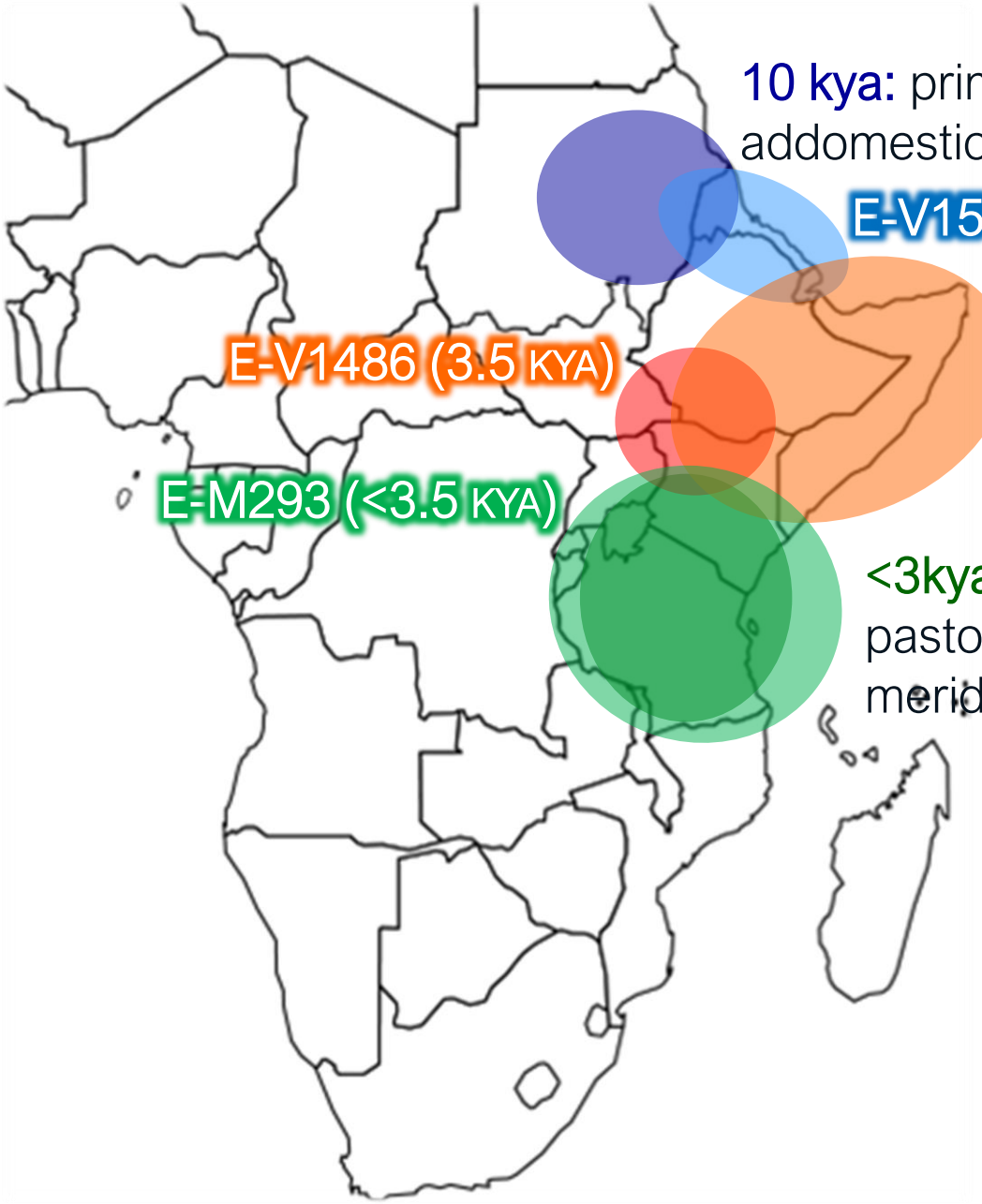
Identificati due SNP che causano la **persistenza della lattasi nell'adulto tipici dell'Africa orientale e presenti ad alte frequenze nelle popolazioni dedite alla pastorizia**

Corrispondenza tra le distribuzioni degli SNP della lattasi e le linee interne all'E-V1515





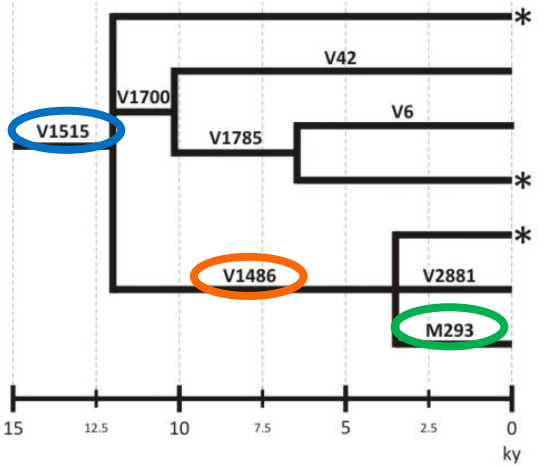
Diffusione della pastoriza: E-V1515 e archeologia



10 kya: prima addomesticazione di bestiame

4 kya: migrazione dei pastori verso l'altopiano etiope e il lago Turkana

<3kya: diffusione della pastorizia verso Kenya meridionale e Tanzania



Corrispondenza tra la rotta migratoria indicata dai nostri dati genetici e quella derivata dalle prove archeologiche

Analisi dei marcatori uniparentali: punti principali

1) I marcatori uniparentali permettono di studiare le linee di discendenza maschili (cromosoma Y) o femminili (mtDNA) → organizzate in **alberi filogenetici**

2) Le linee (aplogruppi) degli alberi filogenetici sono definite da marcatori stabili, spesso **SNP**

3) Le nuove tecnologie di sequenziamento (**NGS**) hanno permesso di identificare migliaia di SNP → definizione di centinaia di aplogruppi → alberi filogenetici ad altissima risoluzione

4) La variabilità genetica accumulata lungo le linee di discendenza può essere usata per **datare i nodi dell'albero** → marcatori usati: STR o SNP

5) Gli effetti della **deriva genetica** sono maggiori nell'Y rispetto agli autosomi → ciascun aplogruppo si trova in una o poche aree geografiche

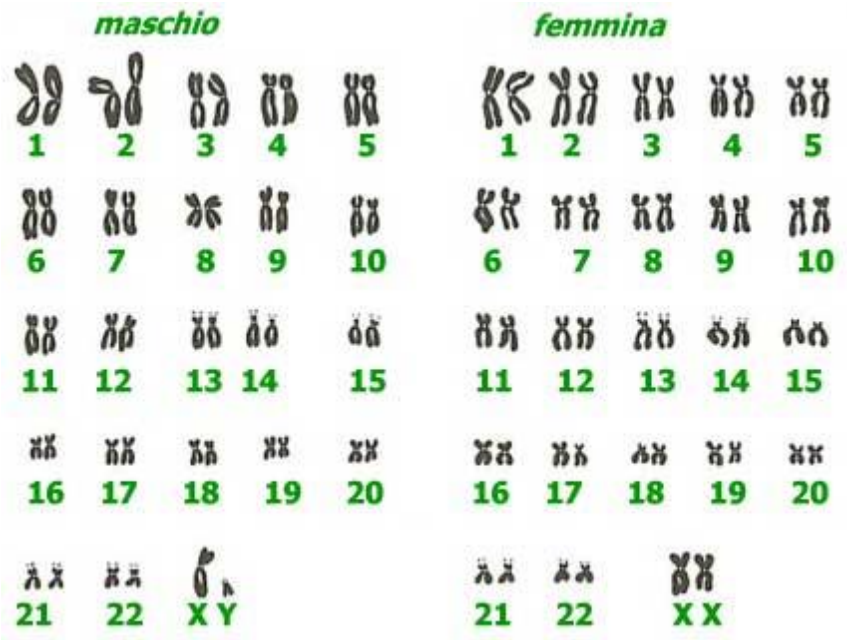
6) L'**approccio filogeografico** permette di correlare le caratteristiche genetiche e geografiche delle popolazioni umane

7) Datazioni dei nodi dell'albero + informazioni filogeografiche (+ informazioni paleoantropologiche, storiche, su altri marcatori genetici, etc.) → **storia delle popolazioni umane**

Caratteristiche generali degli autosomi umani

L'essere umano è un organismo diploide:
22 coppie di autosomi + 1 coppia di cromosomi sessuali (XX nella femmina e XY nel maschio).

Gli autosomi:
- sono ereditati e trasmessi da **entrambi i genitori**;
- vanno incontro a **ricombinazione meiotica**



Caso particolare: il cromosoma X è diploide nelle femmine ed aploide nei maschi

Progetto 1000 Genomi

ARTICLE OPEN
doi:10.1038/nature15393

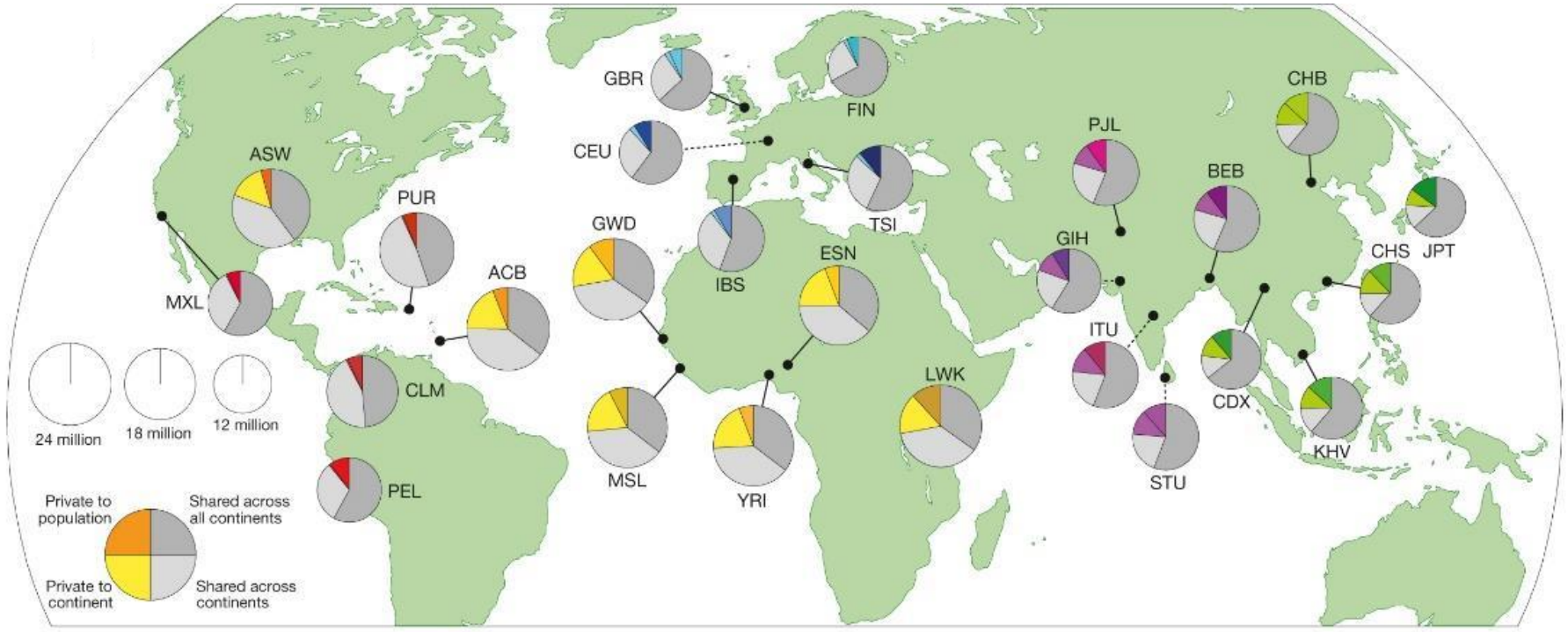
A global reference for human genetic variation

The 1000 Genomes Project Consortium*

Lo scopo del Progetto 1000 Genomi era l'identificazione della **maggior parte delle** varianti con una frequenza di almeno 1% nelle popolazioni studiate.

Il Progetto 1000 Genomi **ha sfruttato la tecnologia NGS**, che ha permesso una riduzione drastica dei costi di sequenziamento. I dati sono stati ottenuti con il metodo **Illumina**.

I dati finali (Phase 3) riguardano 2504 individui da 26 popolazioni mondiali→ database pubblico della **variabilità genetica umana a livello genomico**



Altri progetti sulla diversità genetica umana

ARTICLE

300 individui
142 popolazioni

Mallick et al 2016

doi:10.1038/nature18964

The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

RESEARCH ARTICLE SUMMARY

929 individui
54 popolazioni

HUMAN GENETICS

Insights into human genetic variation and population history from 929 diverse genomes

Anders Bergström*, Shane A. McCarthy†, Ruoyun Hui†, Mohamed A. Almarri†, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann‡, Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue§, Richard Durbin§, Chris Tyler-Smith§*

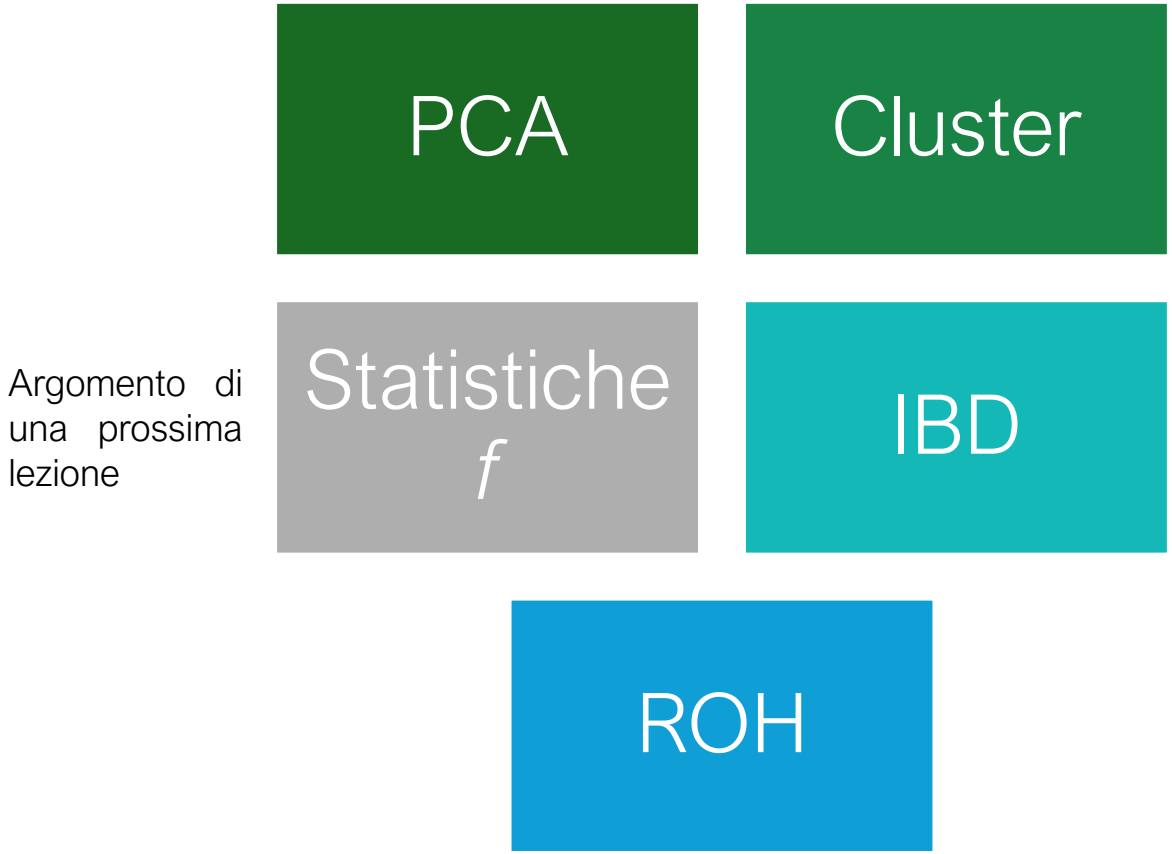
Conclusioni principali:

- Maggior diversità genetica in Africa
- Separazione più antica tra popolazioni in Africa

Panoramica dei metodi per le analisi genome-wide

La maggior parte delle analisi genomiche per ricostruire la storia delle popolazioni umane utilizza gli **SNP** come marcatori (sebbene esistano anche approcci basati su variant strutturali e CNV) identificati negli autosomi (marcatori uniparentali e cromosoma X vengono analizzati separatamente con approcci diversi).

Di solito, vengono **scartati gli SNP molto rari** nelle popolazioni in analisi (soglia 0,1-3%, a seconda degli studi).



Argomento di una prossima lezione

Relazioni tra individui e popolazioni: PCA

PCA: *Principal Component Analysis* (analisi delle component principali)

Analisi statistica di dati con molte variabili: trasformazione lineare che proietta le variabili originarie in un nuovo sistema cartesiano in cui la nuova variabile con la maggiore varianza viene proiettata sul primo asse, la nuova variabile seconda per dimensione della varianza sul secondo asse e così via → riduzione delle variabili di partenza ad un numero minore di componenti principali (PC: *principal component*).

Questo approccio si può applicare ai dati genetici, per i quali si hanno a disposizioni moltissime osservazioni

Osservazioni di partenza:

- SNP = osservazioni → metodo più comunemente usato, basato sulle frequenze alleliche
- Aplotipi, segmenti IBD, etc = osservazioni → metodi utilizzati per individuare relazioni e strutture particolari tra e intra gruppi

PCA: popolazioni mondiali

ARTICLE

Mallick et al 2016

doi:10.1038/nature18964

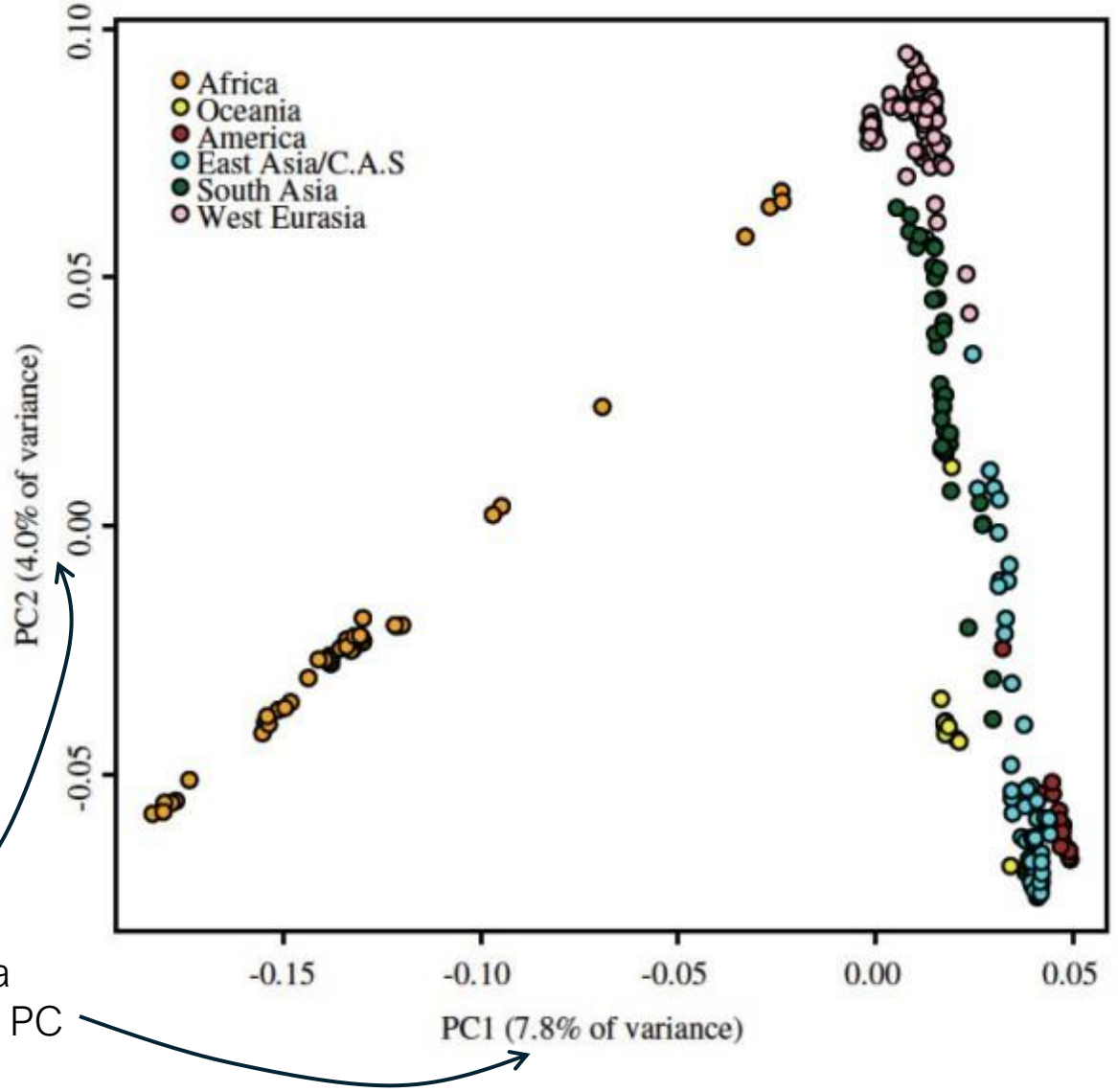
The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

Caratteristiche principali:

- Forma a V
- Popolazioni africane nettamente separate dalle altre

Interpretazioni:

- Ridotto mescolamento tra africani e non africani
- In Eurasia, diversi gruppi distinguibili sebbene parzialmente sovrapposti → mescolamento nel corso della storia umana
- Oceania: isolata
- Popolamento delle Americhe dall'Asia → native americani cadono con asiatici orientali



% di variabilità spiegata rispetto al totale di tutte le PC

PCA: popolazioni locali

Lorente-Galdos *et al. Genome Biology* (2019) 20:77
https://doi.org/10.1186/s13059-019-1684-5

Genome Biology

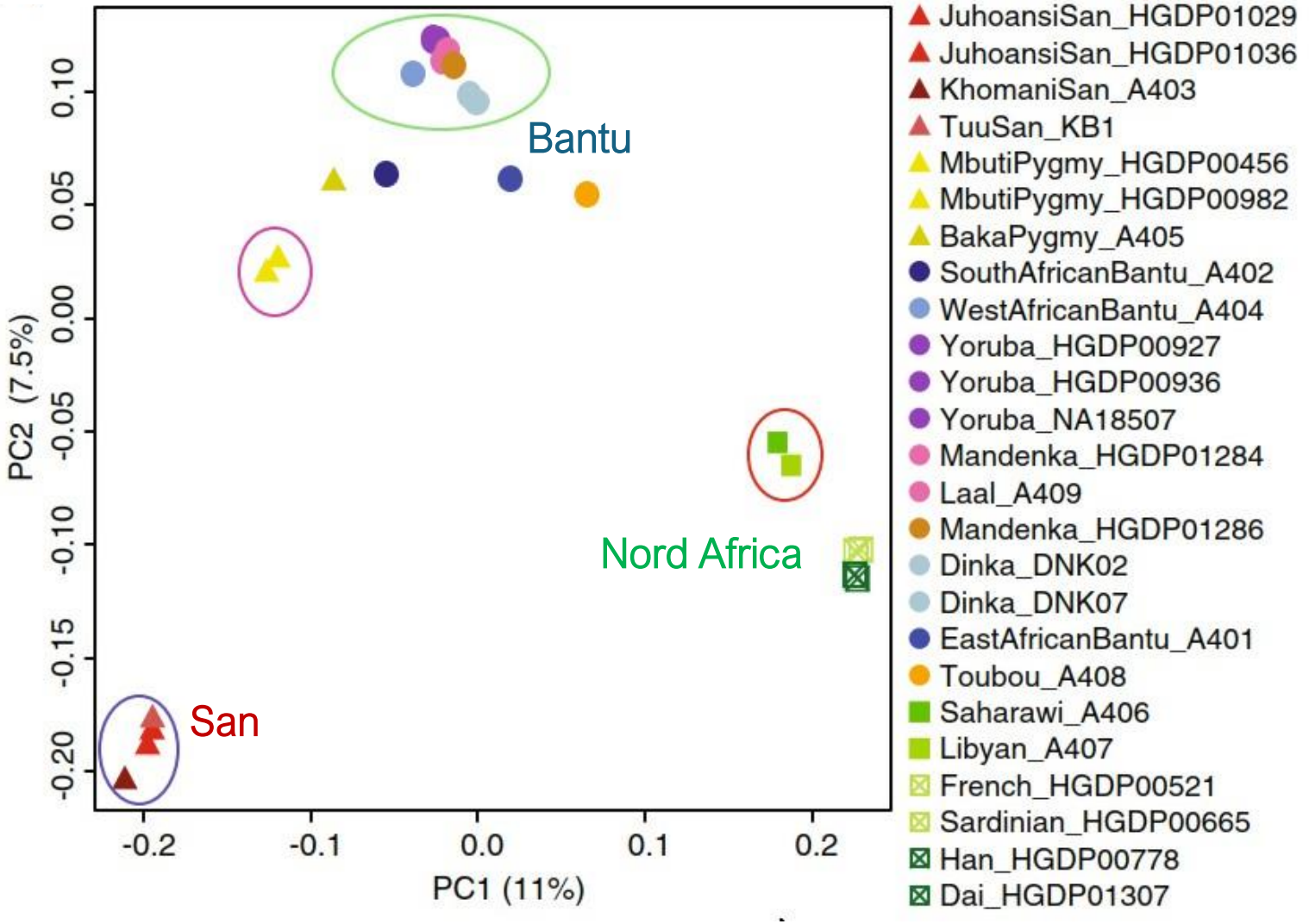
RESEARCH Open Access

Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations

Studio focalizzato sulla variabilità delle popolazioni africane:

- 21 individui
- 15 popolazioni africane

Cerchi: popolazioni utilizzate come rappresentati nell'analisi dei cluster



Caratteristiche principali:

- Netta separazione delle popolazioni San → cacciatori-raccoglitori, i primi che si separano dalle altre popolazioni africane
- Nordafricani spostati verso Eurasia → mescolamento nel corso della storia umana
- Bantu: posizione rispecchia la loro storia (migrazione dall'Africa centrale verso sud ed est da 1500 anni fa)

Definizione delle popolazioni: analisi dei cluster

Stima delle proporzioni di *ancestry* negli individui: partendo dai molti dati genotipici → raggruppamento in cluster ben separati, senza definizioni *a priori* delle popolazioni fonte → identificazioni delle K popolazioni (= cluster) presenti nel gruppo di individui di partenza

Ogni individuo viene modellato come un insieme di cluster ben distinti

Applicazioni:

- Assegnare ciascun individui ad una o più delle K popolazioni
- Identificare gli individui con *ancestry* mista (*admixed*)
- Identificare un'eventuale stratificazione della popolazione

Dati di partenza:

- SNP indipendenti
- Aplotipi: combinazioni di SNP associati lungo un segmento cromosomico

Basi dell'analisi dei cluster

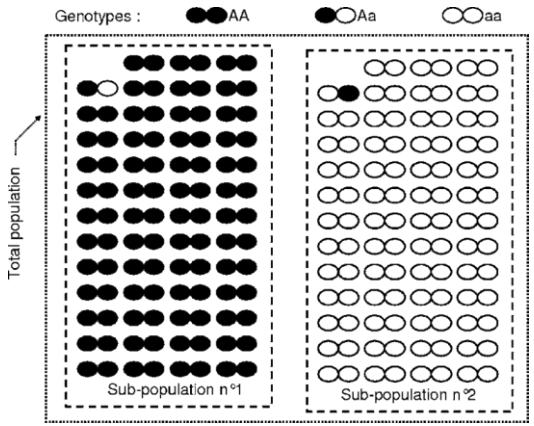
Popolazione suddivisa in sottopopolazioni

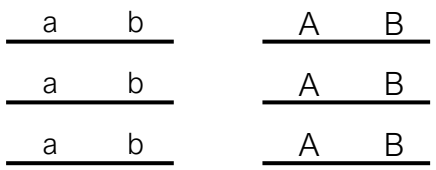
Diminuzione di eterozigoti nella popolazione (effetto Wahlund)

Hardy-Weinberg disequilibrium

Linkage disequilibrium

Associazione tra alleli in ciascuna sottopopolazione





Minimizzazione di entrambi → assegnazione ottimale di individui alle popolazioni

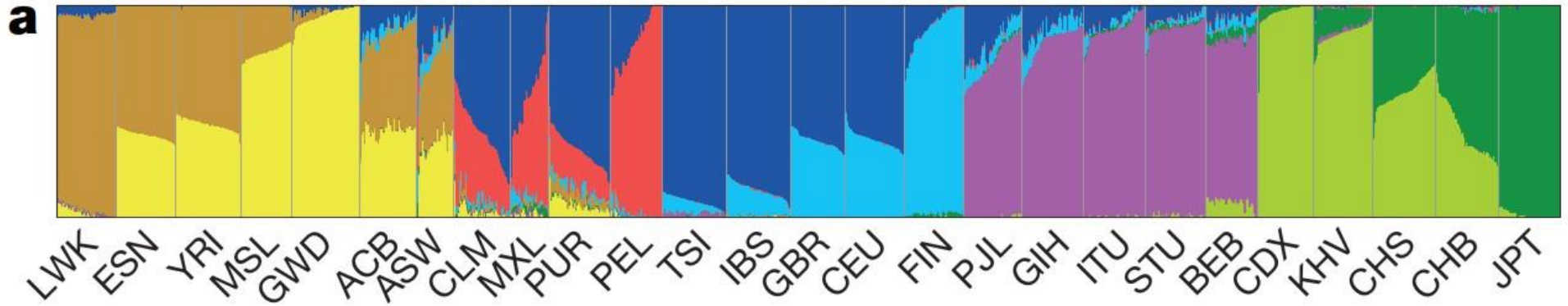
Analisi dei cluster: popolazioni globali

ARTICLE OPEN
doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

8 cluster (1 cluster = 1 colore)



Separazione dei 4 macro-gruppi continentali dei 1000 Genomi:

- Africa (giallo-marrone)
- Europa (blu-azzurro)
- Asia meridionale (viola)
- Asia orientale (Verde chiaro-scuro)

Individui americani admixed (colonizzazione europea e tratta trans-atlantica degli schiavi)

- Afro-americani: ACB e ASW, giallo-marrone + blu
- Meso- e sud-americani: blu + componente rossa assente altrove → componente nativo-americana

Analisi dei cluster: popolazioni locali

Lorente-Galdos *et al. Genome Biology* (2019) 20:77
https://doi.org/10.1186/s13059-019-1684-5

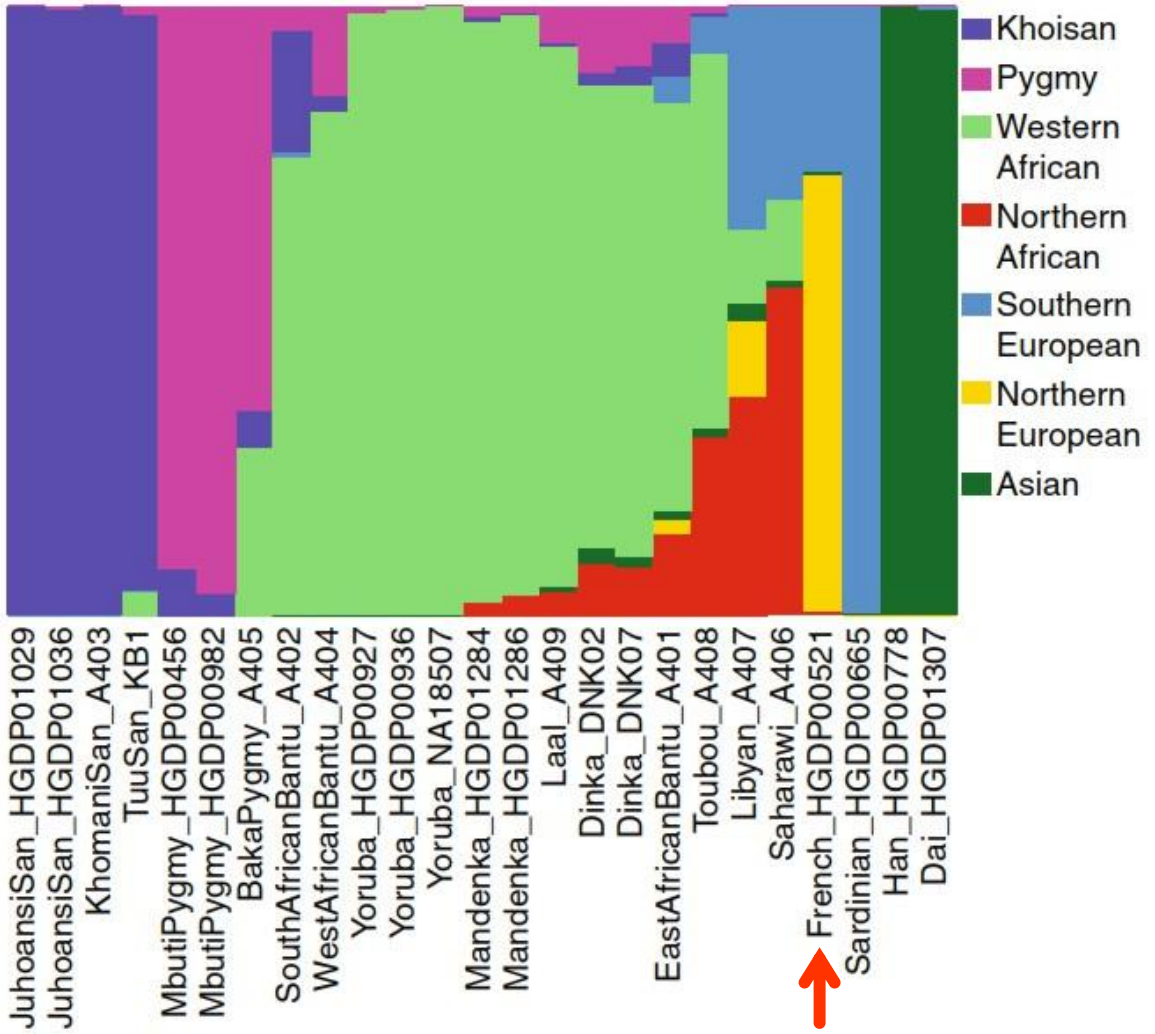
Genome Biology

RESEARCH Open Access

Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations

Differenziazione all'interno del continente africano più granulare rispetto alla precedente analisi su scala globale

7 cluster



Esempio: **interpretazione del pattern di *ancestry* del campione francese (freccia rossa)**

- **Componente nord-europea**
- **Componente sud-europea**
- **Componente asiatica** (minima)

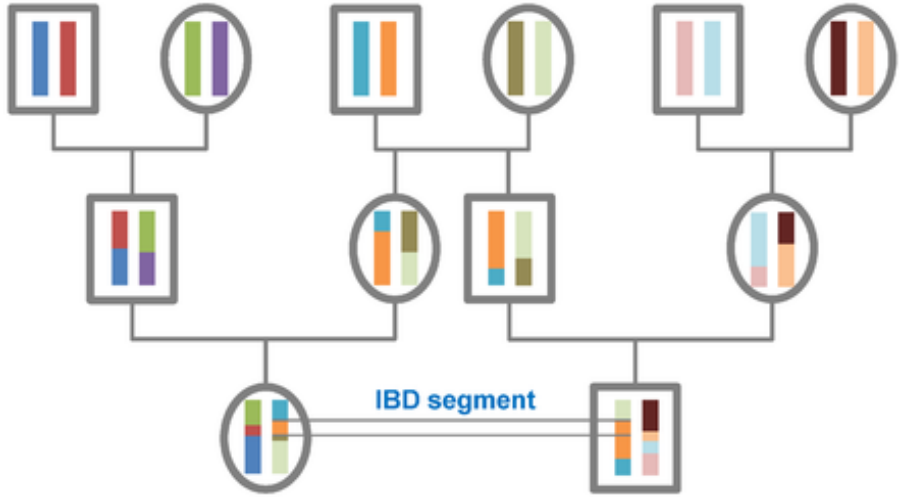
Campione nato dal mescolamento (*admixture*) di queste 3 componenti? NO

Analisi non va interpretata come prova dell'avvenuta *admixture* → evidenza di *ancestry* condivisa (per *admixture*, discendenza o altro)

Struttura della popolazione: IBD

IBD: Identità per discesa (*identity by descent*) → segmenti identici in due o più individui perchè ereditati da un antenato comune

Frammenti più corti → maggiore è il tempo trascorso (ricombinazione ha avuto più tempo per agire e frammentare gli IBD)



Applicazioni:

- Determinare le relazioni di parentela tra gli individui
- Determinare la dimensione efficace Ne di una popolazione (Ne piccola → più individui imparentati in media → IBD più lunghi)
- Informazioni su migrazione (IBD condivisi tra popolazioni diverse)
- Stima del tasso di mutazione (in base a differenze nucleotidiche negli IBD)
- Stima del tasso di ricombinazione (sulla base dei punti di rottura degli IBD)
- Individuare regioni sotto selezione (in media, IBD più condivisi)
- Analizzare la struttura della popolazione a livello più fine

IBD: struttura di una popolazione a livello locale

Studio su popolazione dei Paesi Bassi

PCA fatta su tutti i **segmenti IBD** (pannello a) e segmenti suddivisi per lunghezza (espressa in cM, pannello b)

nature COMMUNICATIONS

ARTICLE

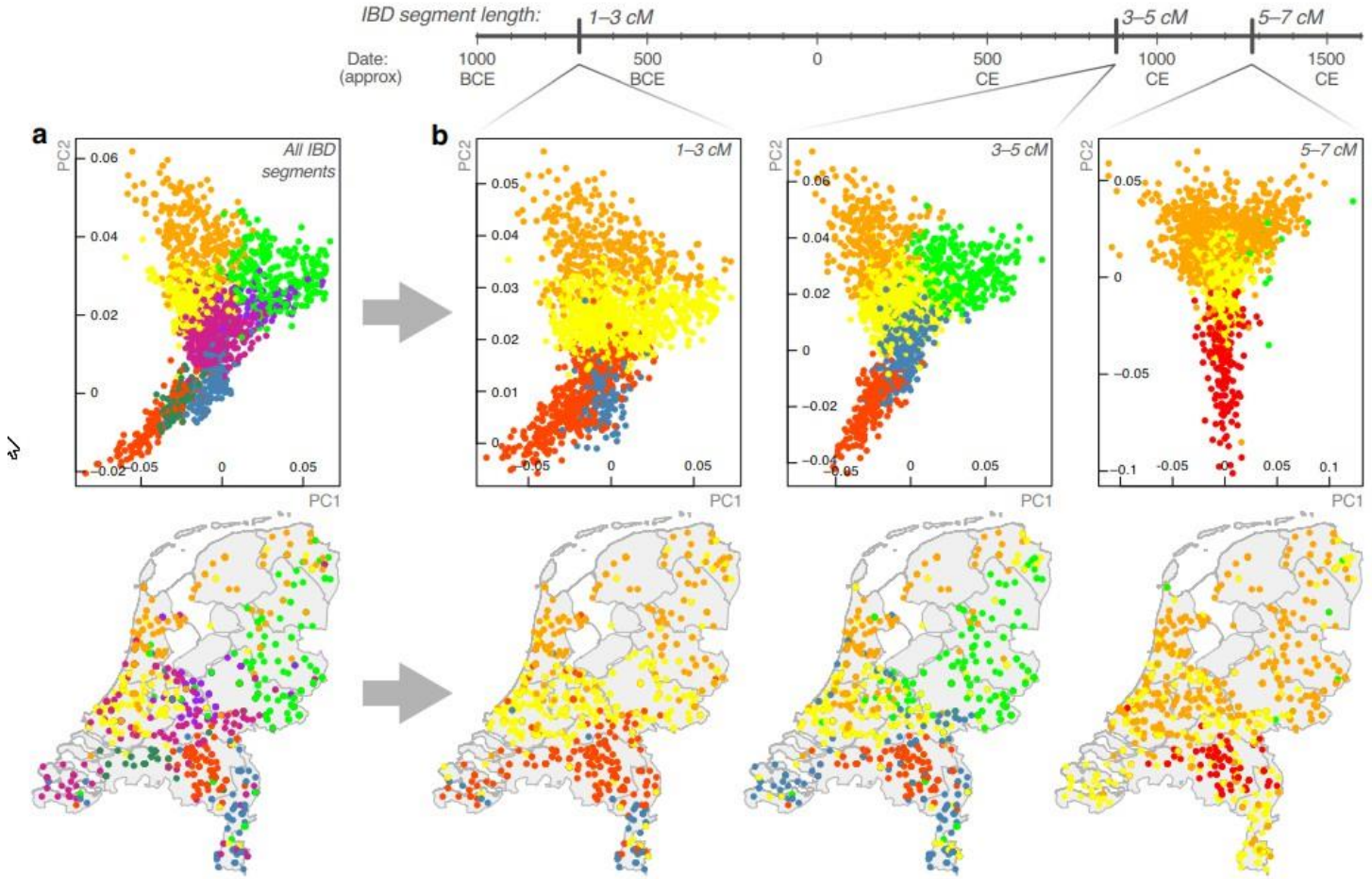
https://doi.org/10.1038/s41467-020-18418-4 OPEN

Dutch population structure across space, time and GWAS design

Ross P. Byrne[1&2], Wouter van Rheenen[2], Project MinE ALS GWAS Consortium*, Leonard H. van den Berg[2], Jan H. Veldink[2] & Russell L. McLaughlin[1&2]

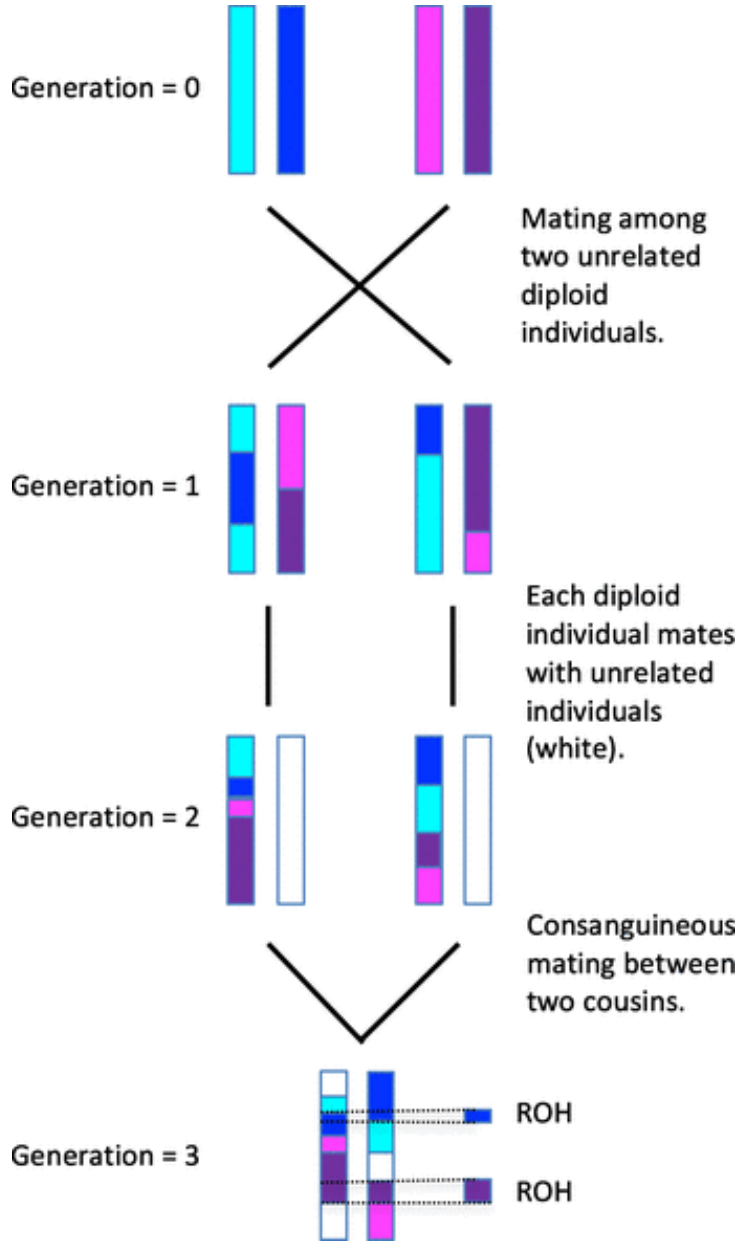
Struttura nord-sud evidente anche con i segmenti più corti e fino ai segmenti più lunghi: più antica (2700 anni fa circa) e mantenuta stabilmente nel tempo

Struttura est-ovest con i segmenti medi: più recente (1120 anni fa circa) e meno stabile nel tempo



Struttura della popolazione: ROH

ROH: *Runs of Homozygosity* → lunghi tratti cromosomici con genotipi omozigoti consecutive



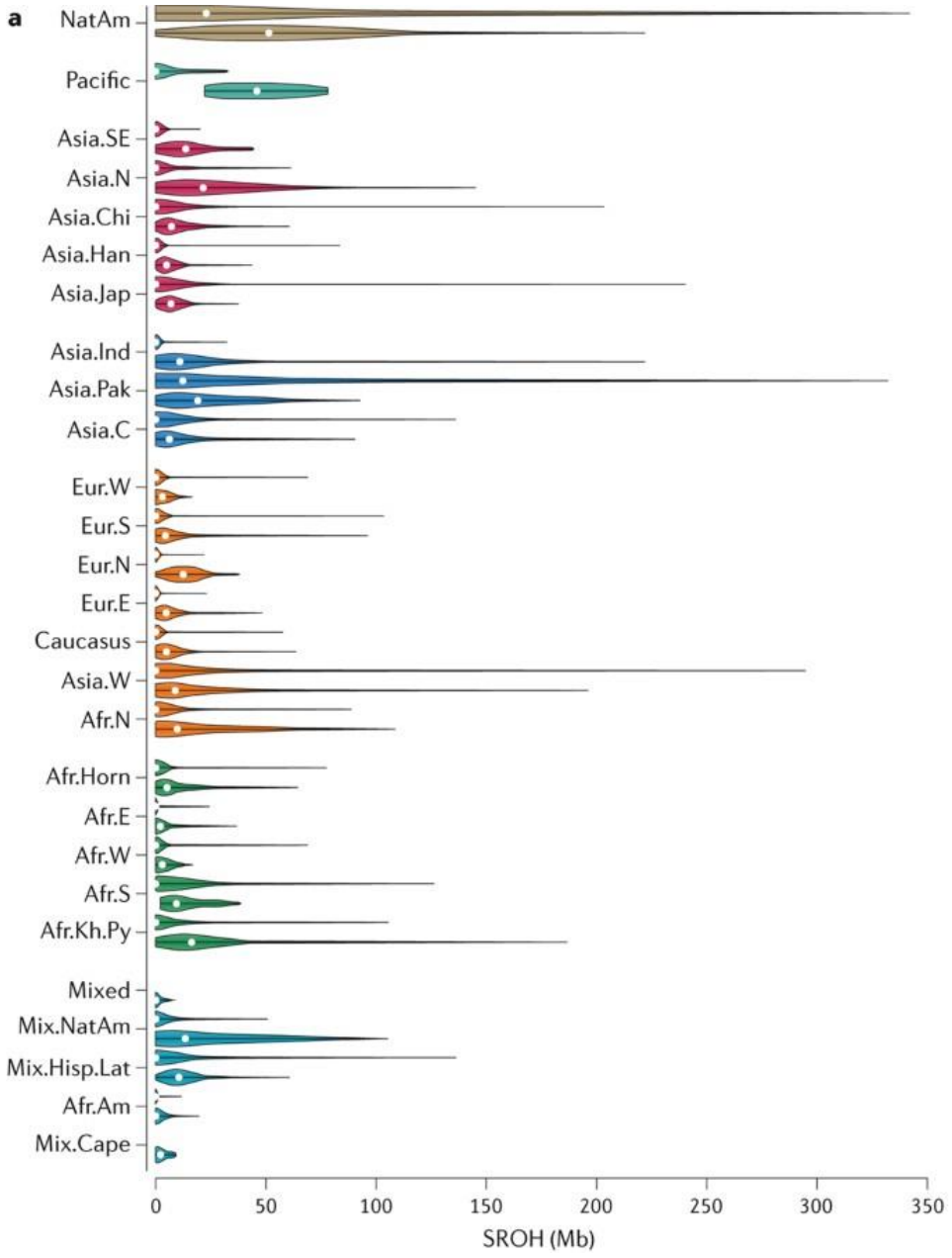
Caratteristiche:

- Ereditati da un antenato comune → segnale di ridotta diversità genetica nella popolazione: consaguineità, effetto del fondatore, collo di bottiglia, etc
- Tendenza ad accumulare mutazioni deleterie nei ROH

Differenza tra ROH e IBD:

- ROH: intra-individuo
- IBD: inter-individuo

ROH: distribuzione nelle popolazioni umane

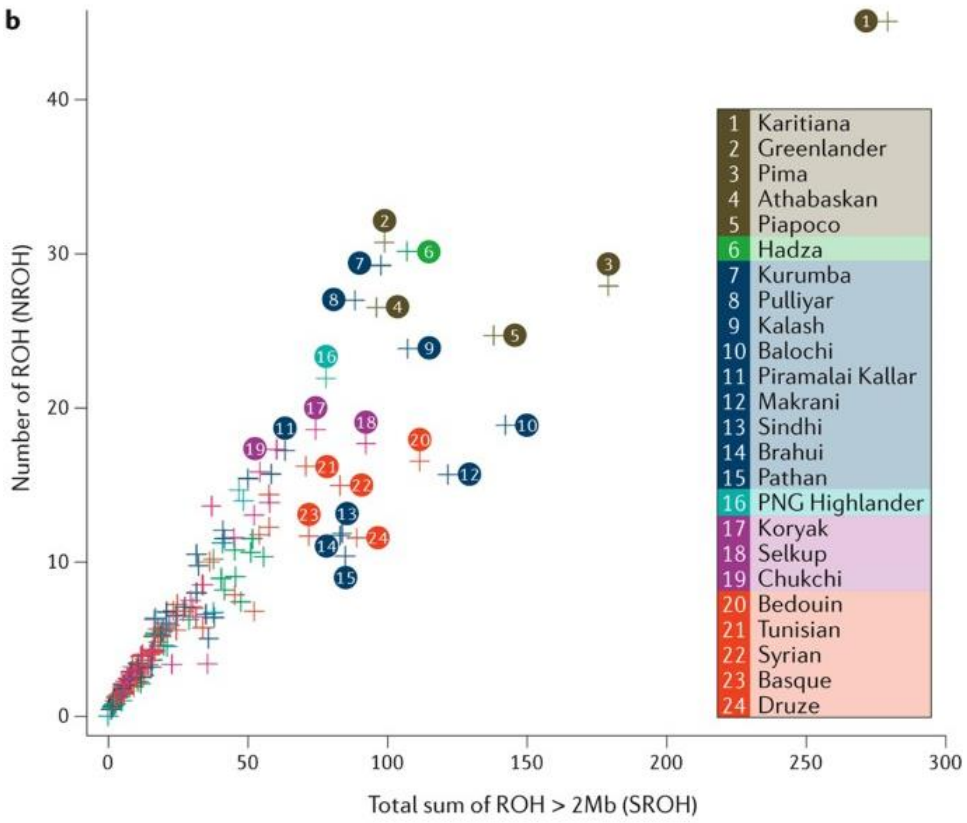


Per ogni popolazione nel pannello **a**, ROH lunghi (> 10 Mb) sopra e corti (2-5 Mb) sotto

nature reviews genetics

Runs of homozygosity: windows into population history and trait architecture

Francisco C. Ceballos[1,2], Peter K. Joshi[3], David W. Clark[3], Michèle Ramsay[1,4] and James F. Wilson[2,3]



NROH: numero totale dei ROH

SROH: somma della lunghezza dei ROH

Analisi degli autosomi: punti principali

1) Partendo da un elevato numero di SNP identificati grazie approcci *genome-wide* su tutti gli autosomi, si può condurre un'**analisi delle componenti principali: PCA** → relazione tra individui/popolazioni

2) Partendo da un elevato numero di SNP identificati grazie approcci *genome-wide* su tutti gli autosomi, si possono assegnare gli individui a cluster ben specifici e ricostruirne le componenti di *ancestry* → **analisi dei cluster**

3) Partendo da un elevato numero di SNP identificati grazie approcci *genome-wide* su tutti gli autosomi, le **statistiche *f*** possono dare informazioni su eventi di *admixture* tra le popolazioni analizzate

4) Partendo da un elevato numero di SNP identificati grazie approcci *genome-wide* su tutti gli autosomi, si possono identificare **segmenti IBD e segmenti RHO** → struttura ed eventi demografici della popolazione nel tempo