

VCF ed esercizi di genotipizzazione

Fabrizio Mafessoni, Genomica Computazionale, Corso di Laurea Specialistica in Genomica Funzionale, Università di Trieste



Esercizio 1: compariamo genotipizzazione a singolo campione e multi-campione

1) Prima di tutto generiamo le chiamate per tutti gli individui

```
for i in NA12878 HG00096 HG00110;do
echo $i
MYBAM=${i}.chrom20.ILLUMINA.bwa.*.low coverage.*.bam
bcftools mpileup -Ou -f chr20.fa \
  -a FORMAT/DP,FORMAT/AD,FORMAT/ADF,FORMAT/ADR \
  ${MYBAM} | bcftools call -m -v -Oz -o ${i}.single.recall.vcf.gz
bcftools index ${i}.single.recall.vcf.gz
done
```

Ora compariamo tutti questi vcf individualmente alle chiamate dei genotipi fatti su tutti e tre gli individui simultaneamente. Prima di tutto notiamo che per tutti, sono molte piu' le varianti chiamate per i genotipi chiamati simultaneamente.

Tuttavia questo potrebbe essere dovuto sia al fatto che nel file trio ci sono piu' individui, sia ad un maggiore potere nel chiamare le varianti simultaneamente, quindi filtriamo le varianti per tenere solo quelle effettivamente supportate dal nostro individuo *i*

```
$for i in NA12878 HG00096 HG00110;do
echo ${i}
VAR_SINGLE=$( cat isec_out/${i}/0000.vcf | grep -v '#' | wc -l)
VAR_TRIO=$( cat isec_out/${i}/0001.vcf | grep -v '#' | wc -l)
VAR_TRIO_FILTERED=$( bcftools view -s ${i} isec_out/${i}/0001.vcf | awk '{if (substr($10,1,3)!="0/0" && substr($10,1,3)!="./."){print}}' | wc -l
)
echo "Varianti presenti solo in ${i} ${VAR_SINGLE}"
echo "Varianti presenti solo nel file trio ${VAR_TRIO}"
echo "Varianti presenti in $i e chiamate nel file trio ${VAR_TRIO_FILTERED}"
done
NA12878
Varianti presenti solo in NA12878 2224
Varianti presenti solo nel file trio 60698
Varianti presenti in NA12878 e chiamate nel file trio 12752
HG00096
Varianti presenti solo in HG00096 2980
Varianti presenti solo nel file trio 59122
Varianti presenti in HG00096 e chiamate nel file trio 11641
HG00110
Varianti presenti solo in HG00110 1711
Varianti presenti solo nel file trio 53054
Varianti presenti in HG00110 e chiamate nel file trio 9110
```

Conclusioni esercizio 1:

- Chiamare multi-campione da piu' potere alle nostre chiamate

```
$for i in NA12878 HG00096 HG00110;do
echo ${i}
VAR_SINGLE=$( cat isec_out/${i}/0000.vcf | grep -v '#' | wc -l)
VAR_TRIO=$( cat isec_out/${i}/0001.vcf | grep -v '#' | wc -l)
VAR_TRIO_FILTERED=$( bcftools view -s ${i} isec_out/${i}/0001.vcf | awk '{if (substr($10,1,3)!="0/0" && substr($10,1,3)!="./."){print}}' | wc -l
)
echo "Varianti presenti solo in ${i} ${VAR_SINGLE}"
echo "Varianti presenti solo nel file trio ${VAR_TRIO}"
echo "Varianti presenti in ${i} e chiamate nel file trio ${VAR_TRIO_FILTERED}"
done
NA12878
Varianti presenti solo in NA12878 2224
Varianti presenti solo nel file trio 60698
Varianti presenti in NA12878 e chiamate nel file trio 12752
HG00096
Varianti presenti solo in HG00096 2980
Varianti presenti solo nel file trio 59122
Varianti presenti in HG00096 e chiamate nel file trio 11641
HG00110
Varianti presenti solo in HG00110 1711
Varianti presenti solo nel file trio 53054
Varianti presenti in HG00110 e chiamate nel file trio 9110
```

Esercizio 2: esploriamo la differenza tra samtools mpileup e bcftools mpileup

Ora esploriamo piu' in dettaglio come vengono ottenute queste genotype likelihood. Per questo, esaminiamo i file samtools mpileup e compariamoli con i bcftools mpileup. Esaminiamo le prime 2 posizioni

```
$samtools mpileup NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam | head -2
[mpileup] 1 samples in 1 input files
20      59993      N      4      ^>G^>G^>G^>G      KLLK
20      59994      N      4      TTTT      H?FA
$
$bcftools mpileup -f chr20.fa NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam | grep -v '#' | head -2
[mpileup] 1 samples in 1 input files
[mpileup] maximum number of reads per input file set to -d 250
20      59993      .      N      G,<*>      0      .      DP=4;I16=0,0,4,0,0,0,170,7226,0,0,116,3364,0,0,96,2316;QS=0,1,0;VDB=0.00701511;SGB=-0.556411;MQ0F=0  P
L      88,12,0,88,12,88
20      59994      .      N      T,<*>      0      .      DP=4;I16=0,0,4,0,0,0,138,4814,0,0,116,3364,0,0,97,2359;QS=0,1,0;VDB=0.00835905;SGB=-0.556411;MQ0F=0  P
L      88,12,0,88,12,88
```

Per la prima posizione (chr20:59993) notate nei campi PL:

88,12,0,88,12,88

i quali si riferiscono alle PL di genotipi 0/0,0/1,1/1,0/2,1/2,2/2 con allele referenza N e alternativo G (notate che la N risulta referenza perché non avevamo definito una referenza di input nel comando mpileup, quindi G/G corrisponde ad 1/1)

Esercizio 2: esploriamo la differenza tra samtools mpileup e bcftools mpileup

Ora esploriamo piu' in dettaglio come vengono ottenute queste genotype likelihood. Per questo, esaminiamo i file samtools mpileup e compariamoli con i bcftools mpileup. Esaminiamo le prime 2 posizioni

```
$samtools mpileup NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam | head -2
[mpileup] 1 samples in 1 input files
20      59993      N      4      ^>G^>G^>G^>G      KLLK
20      59994      N      4      TTTT      H?FA
$
$bcftools mpileup -f chr20.fa NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam | grep -v '#' | head -2
[mpileup] 1 samples in 1 input files
[mpileup] maximum number of reads per input file set to -d 250
20      59993      .      N      G,<*>      0      .      DP=4;I16=0,0,4,0,0,0,170,7226,0,0,116,3364,0,0,96,2316;QS=0,1,0;VDB=0.00701511;SGB=-0.556411;MQ0F=0 P
L      88,12,0,88,12,88
20      59994      .      N      T,<*>      0      .      DP=4;I16=0,0,4,0,0,0,138,4814,0,0,116,3364,0,0,97,2359;QS=0,1,0;VDB=0.00835905;SGB=-0.556411;MQ0F=0 P
L      88,12,0,88,12,88
```

Per la prima posizione (chr20:59993) notate nei campi PL:

88,12,0,88,12,88

i quali si riferiscono alle PL di genotipi 0/0,0/1,1/1,0/2,1/2,2/2 con allele referenza N e alternativo G (notate che la N risulta referenza perché non avevamo definito una referenza di input nel comando mpileup, quindi G/G corrisponde ad 1/1)



BORING

Guardiamo la differenza su un sito eterozigote



INTERESTING

Guardiamo la differenza su un sito eterozigote

bcftools mpileup

```
$cat isec_out/NA12878/0003.vcf | grep '0/1' | head -1
```

```
20      61795      .      G      T      265.217      .  
=0;AC=3;AN=6;DP4=9,10,4,6;MQ=58 GT:PL:DP:ADF:ADR:AD
```

```
DP=30;VDB=0.522432;SGB=-1.94008;RPBZ=-1.10228;MQBZ=0.725476;MQSBZ=-0.901388;BQBZ=-0.484284;SCBZ=0;MQ0F  
0/1:135,0,161:11:5,2:1,3:6,5 1/1:171,15,0:5:0,2:0,3:0,5 0/0:0,39,255:13:4,0:9,0:13,0
```



INTERESTING

Guardiamo la differenza su un sito eterozigote

bcftools mpileup

```
$cat isec_out/NA12878/0003.vcf | grep '0/1' | head -1
20      61795      .      G      T      265.217      .      DP=30;VDB=0.522432;SGB=-1.94008;RPBZ=-1.10228;MQBZ=0.725476;MQSBZ=-0.901388;BQBZ=-0.484284;SCBZ=0;MQ0F
=0;AC=3;AN=6;DP4=9,10,4,6;MQ=58 GT:PL:DP:ADF:ADR:AD      0/1:135,0,161:11:5,2:1,3:6,5      1/1:171,15,0:5:0,2:0,3:0,5      0/0:0,39,255:13:4,0:9,0:13,0
```

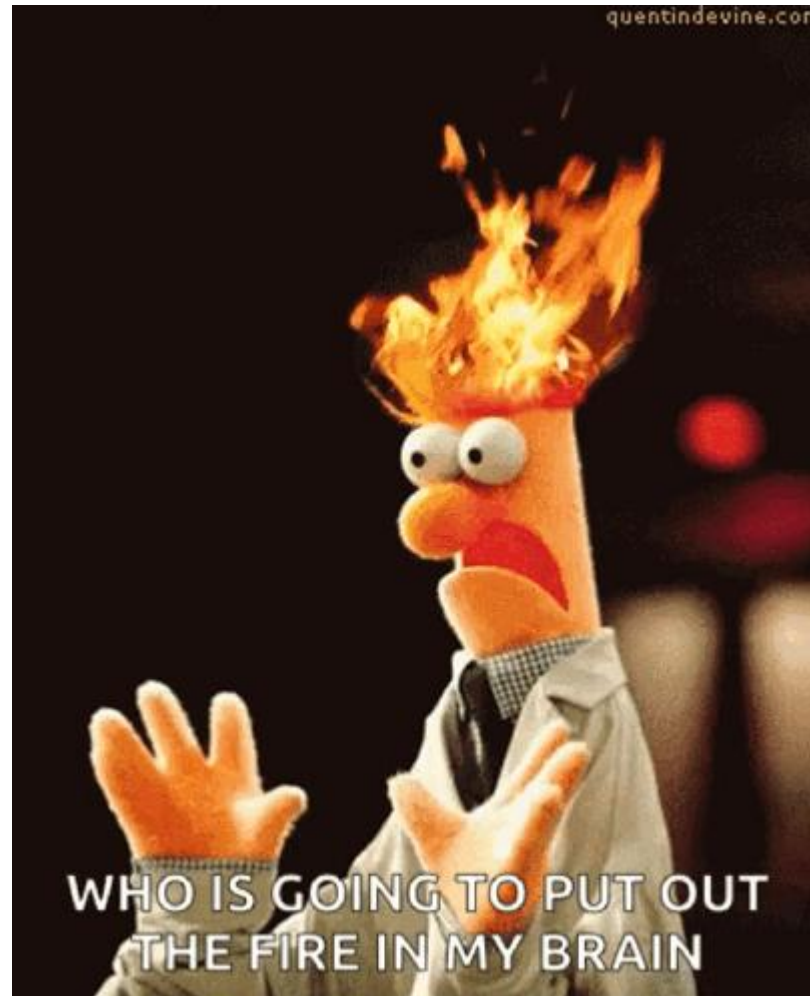
samtools mpileup

```
$for i in NA12878 HG00096 HG00110;do
echo ${i}
samtools mpileup ${i}.chrom20.ILLUMINA.bwa.*.low_coverage.*.bam | awk -v SNP=61795 '{if ($2==SNP){print;exit}}' | tr 'acgt' 'ACGT'
done
NA12878
[mpileup] 1 samples in 1 input files
20      61795      N      11      GTGGGGTTTGT      KJLJKKI?IDF
HG00096
[mpileup] 1 samples in 1 input files
20      61795      N      5      TTTT      GHP?/
HG00110
[mpileup] 1 samples in 1 input files
20      61795      N      13      GGGGGGGGGGGG      HFK8KHDIGJIGE
```



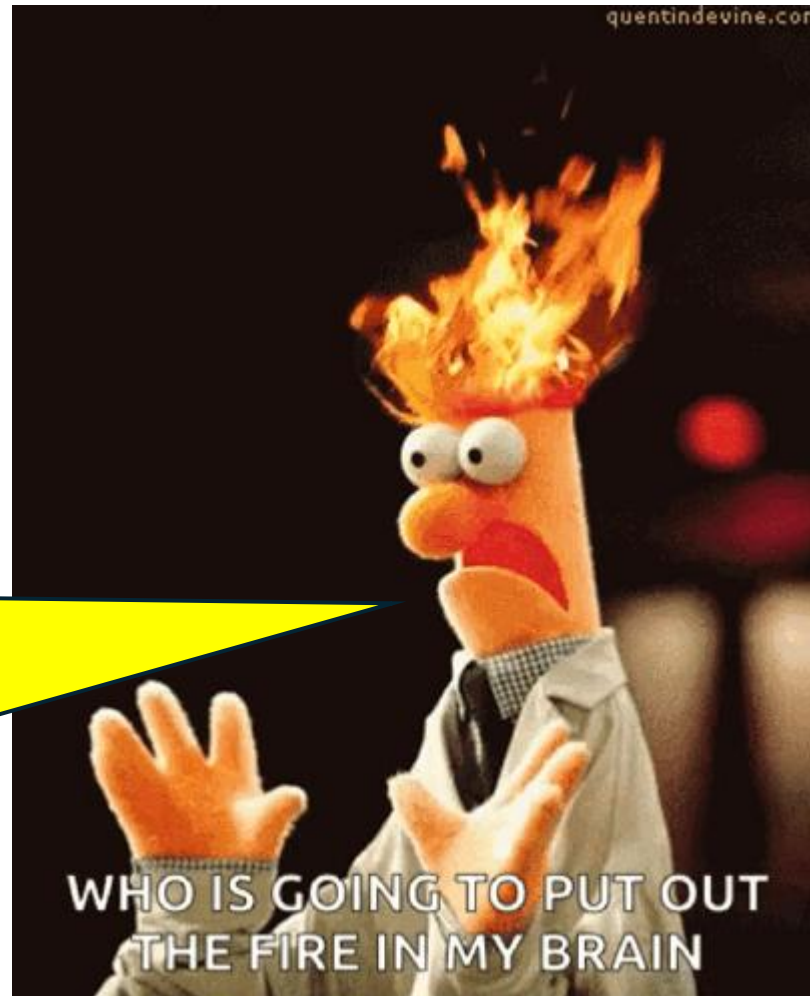
INTERESTING

Costruiamo un programma per genotipizzazione/SNP calling!!!



Costruiamo un programma per genotipizzazione/SNP calling!!!

**Suggerimenti?
Cosa usiamo?**



La distribuzione binomiale

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

Fattoriali

$n! = n * (n-1) * (n-2) * \dots * 1$

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 probabilità di un genotipo in presenza di errori

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

p per i vari genotipi

$$\Pr(1 \mid 0/0) = \varepsilon$$

$$\Pr(1 \mid 0/1) = 1/2$$

$$\Pr(1 \mid 1/1) = 1 - \varepsilon$$

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



La probabilità di un genotipo in presenza di errori

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

p per i vari genotipi

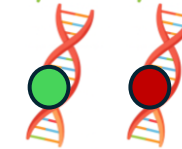
$$\Pr(1 \mid 0/0) = \varepsilon$$

$$\Pr(1 \mid 0/1) = 1/2$$

$$\Pr(1 \mid 1/1) = 1 - \varepsilon$$

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

reads osservate



Esempio: $k=1$, $n=2$ e $\varepsilon=0.01$:

Se il genotipo è:

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

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

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



Costruiamo un programma per genotipizzazione/SNP calling!!!

Eliminiamo anche caratteri speciali come quelli di inizio reads o indel con `sed (sed -E 's/\^./ /g; s/\$/ /g; s/[+-][0-9]+[ACGTNacgtn]+//g' )`.

Costruiamo un programma per genotipizzazione/SNP calling!!!

Eliminiamo anche caratteri speciali come quelli di inizio reads o indel con `sed` (`sed -E 's/\^./ /g; s/\$/ /g; s/[+-][0-9]+[ACGTNacgtn]+//g'`).

Creiamo poi un piccolo script `awk` chiamato `count_ref_alt.awk` (in appendice) che semplicemente conti gli alleli referenza e alternativi, di modo che dia un output che dia cromosoma, posizione, allele REF, conto REF, allele ALT, conto ALT, conto totale (o Read Depth), e poi un PASS quando il sito è biallelico (per semplicità evitiamo di calcolare le genotype likelihood per siti triallelici):

```
samtools mpileup
NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam -f
chr20.fa | sed -E 's/\^./ /g; s/\$/ /g; s/[+-][0-
9]+[ACGTNacgtn]+//g' | awk -f count_ref_alt.awk | head
#20 59993 N 0 G 1 1 PASS
#20 59994 N 0 T 4 4 PASS
```

A questo punto aggiungiamo uno piccolo script awk `simple_genotyper.awk` (nell'appendice), che calcoli i genotype likelihood assumendo un semplice modello binomiale con errore costante dell'1%. Il cuore del genotyper consiste delle seguenti righe:

$$p_{RR} = e$$

$$p_{RA} = 0.5$$

$$p_{AA} = 1 - e$$

$$\ln L_{RR} = k * \log(p_{RR}) + (n-k) * \log(1-p_{RR})$$

$$\ln L_{RA} = k * \log(p_{RA}) + (n-k) * \log(1-p_{RA})$$

$$\ln L_{AA} = k * \log(p_{AA}) + (n-k) * \log(1-p_{AA})$$

Le prime 3 definiscono il nostro modello di errore: quella che definivamo come p nelle slides, e che non è altro che la probabilità di campionare un allele alternativo (0.5 per un allele eterozigote, ed invece “1-e” ed “e” per gli omozigoti, con “e” il tasso di errore).

Le seconde tre definiscono la distribuzione binomiale. Notate che non consideriamo qui il coefficiente binomiale (visto che per ogni sito nel genoma cambia, ma è uguale per tutti i genotipi e quindi si può semplificare) e che calcoliamo la $\log(\text{likelihood})$

```
samtools mpileup
NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam -f
chr20.fa | sed -E 's/\^./ /g; s/\$/ /g; s/[+-][0-
9]+[ACGTNacgtn]+//g' | awk -f count_ref_alt.awk | awk -f
simple_genotyper.awk | grep 61795
```

```
#20          61795      G          T          6          2          8          0/1
G/T          GL=16.18,-0.00,96.00
```

```
samtools mpileup
NA12878.chrom20.ILLUMINA.bwa.CEU.low_coverage.20121211.bam -f
chr20.fa | sed -E 's/\^./ /g; s/\$/ /g; s/[+-][0-
9]+[ACGTNacgtn]+//g' | awk -f count_ref_alt.awk | awk -v
e=0.001 -f simple_genotyper.awk | grep 61795
```

#20	61795	G	T	6	2	8	0/1
G/T	GL=35.94,-0.00,155.93						

- 3) Ora compariamo i nostri genotipi con quelli presenti nel dataset dei 1000 genomi.
Estraiamo solo i nostri tre individui e solo le varianti per cui almeno uno di questi sia eterozigote o alternativo.

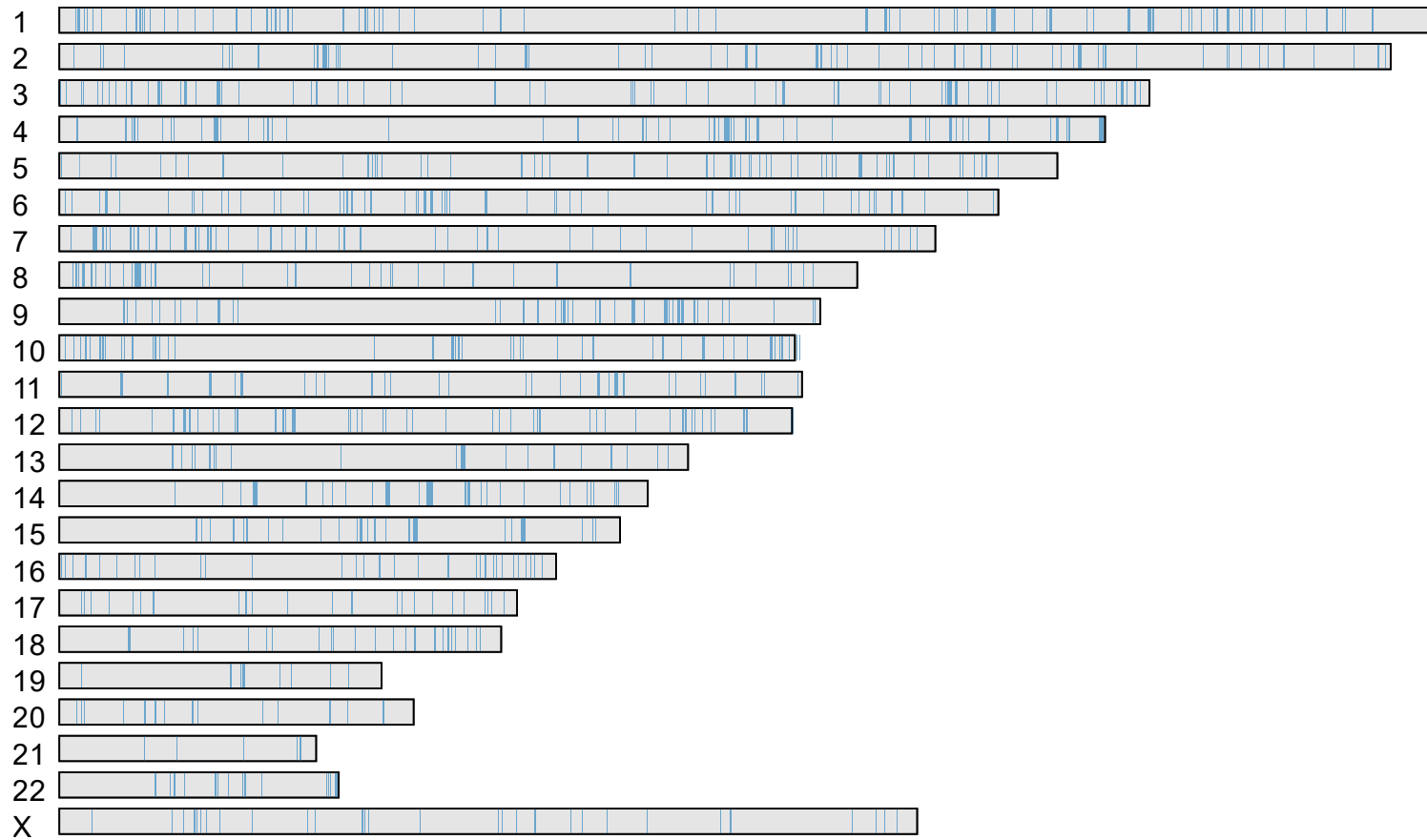
```
bcftools view -s NA12878,HG00096,HG00110  
ALL.chr20.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz | bcftools view -i 'COUNT(GT="het" ||  
GT="alt")>0' | bgzip -f > trio1000g.chr20.vcf.gz  
  
tabix trio1000g.chr20.vcf.gz
```

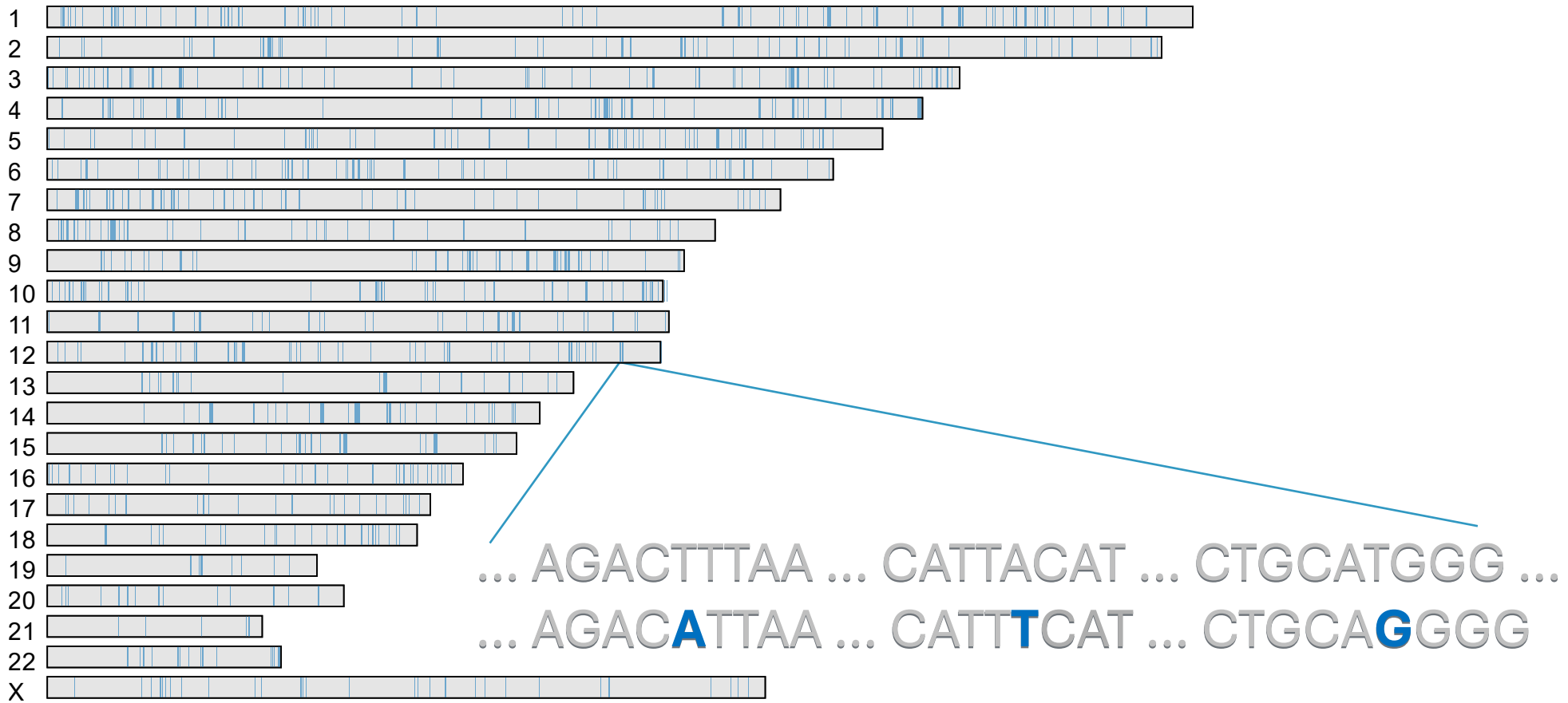
Ora compariamo nuovamente con `bcftools isec`

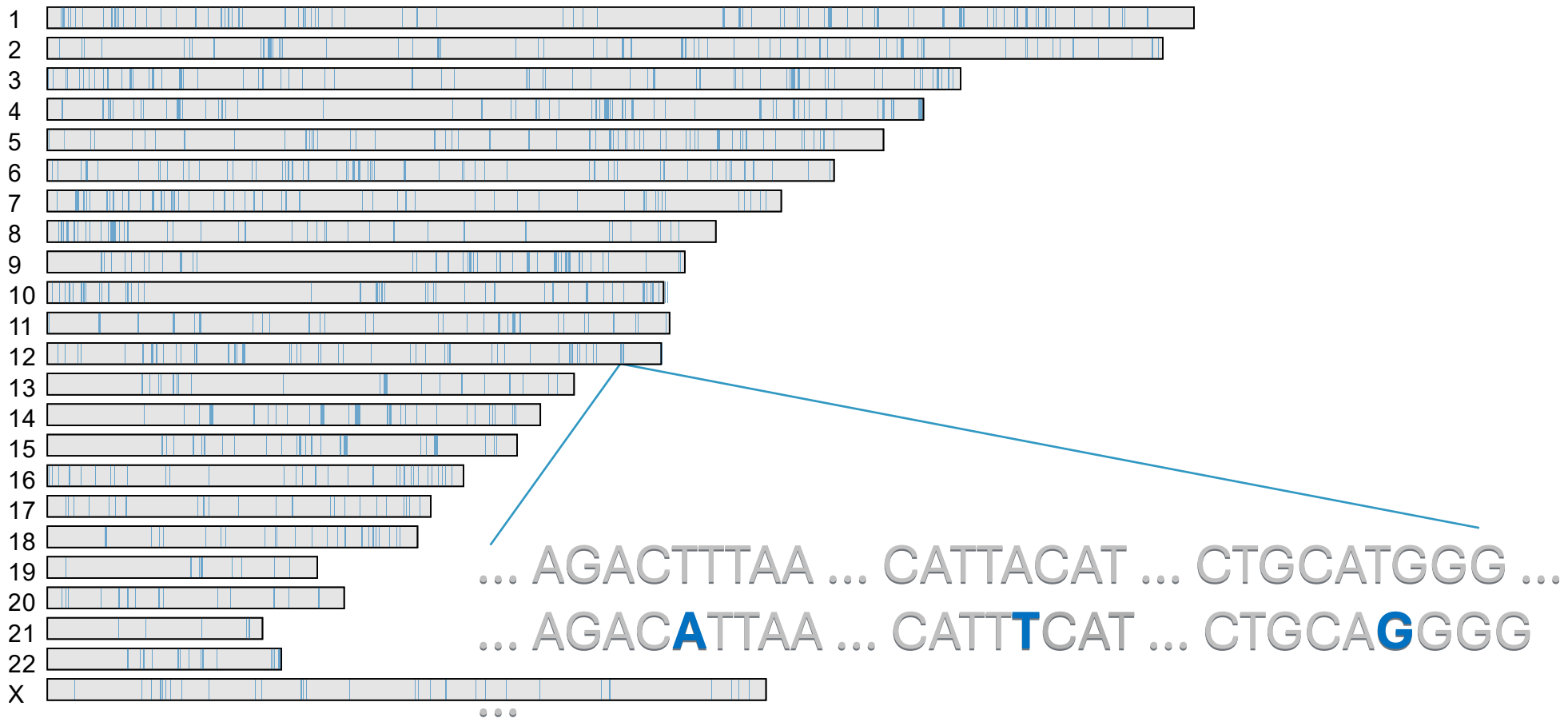
```
mkdir -p isec_out/1000g  
  
bcftools isec -p isec_out/1000g trio1000g.chr20.vcf.gz  
trio.recall.vcf.gz
```

Notiamo che ora la maggior parte delle posizioni sono condivise tra il nostro trio e quelle del trio chiamate anche nel 1000 genomes dataset, mentre meno varianti sono uniche. Sono leggermente piu' quelle chiamate nel 1000 genomes. Ora, come effettuato prima, potrebbe essere interessante vedere come appaiono quelle esclusive dei due datasets.

```
wc -l isec_out/1000g/000*.vcf  
  
#    30332 isec_out/1000g/0000.vcf  
#    27672 isec_out/1000g/0001.vcf  
#   101940 isec_out/1000g/0002.vcf  
#   101801 isec_out/1000g/0003.vcf
```







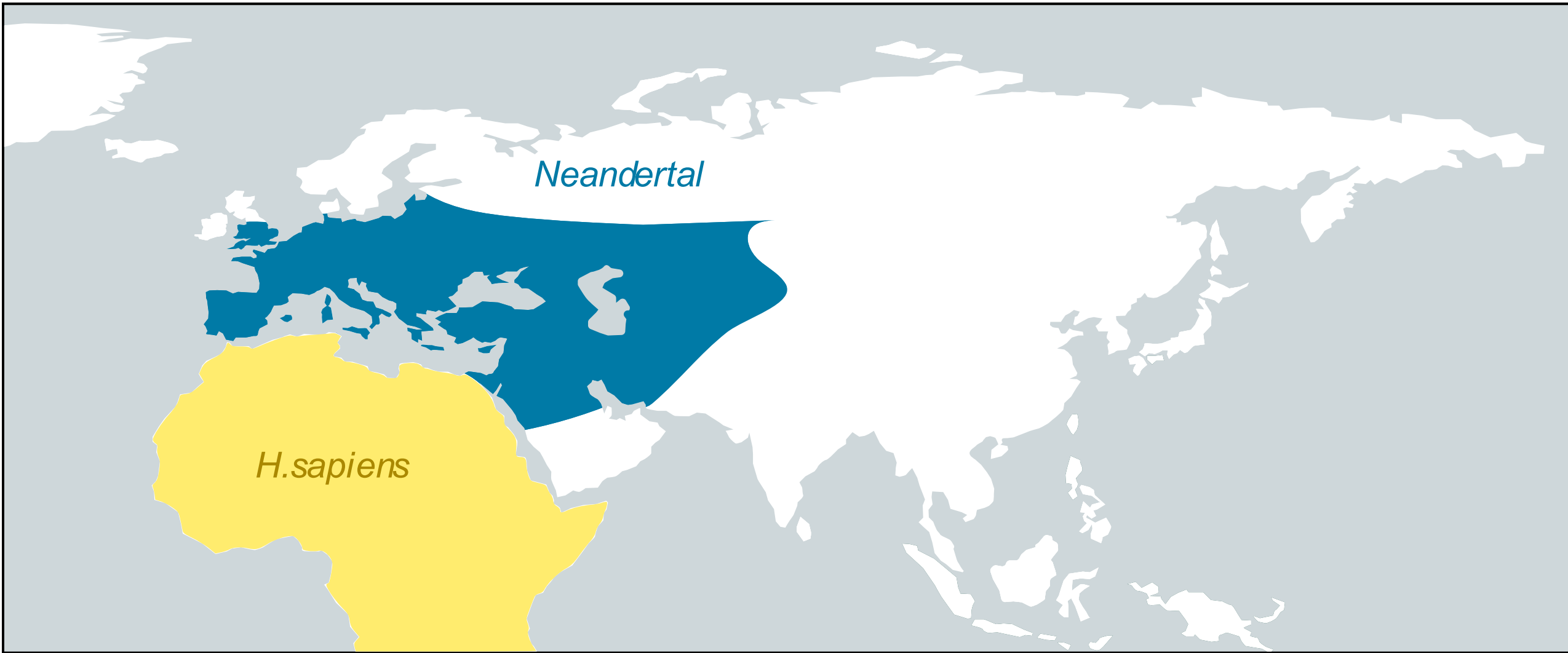
... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

...

... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG



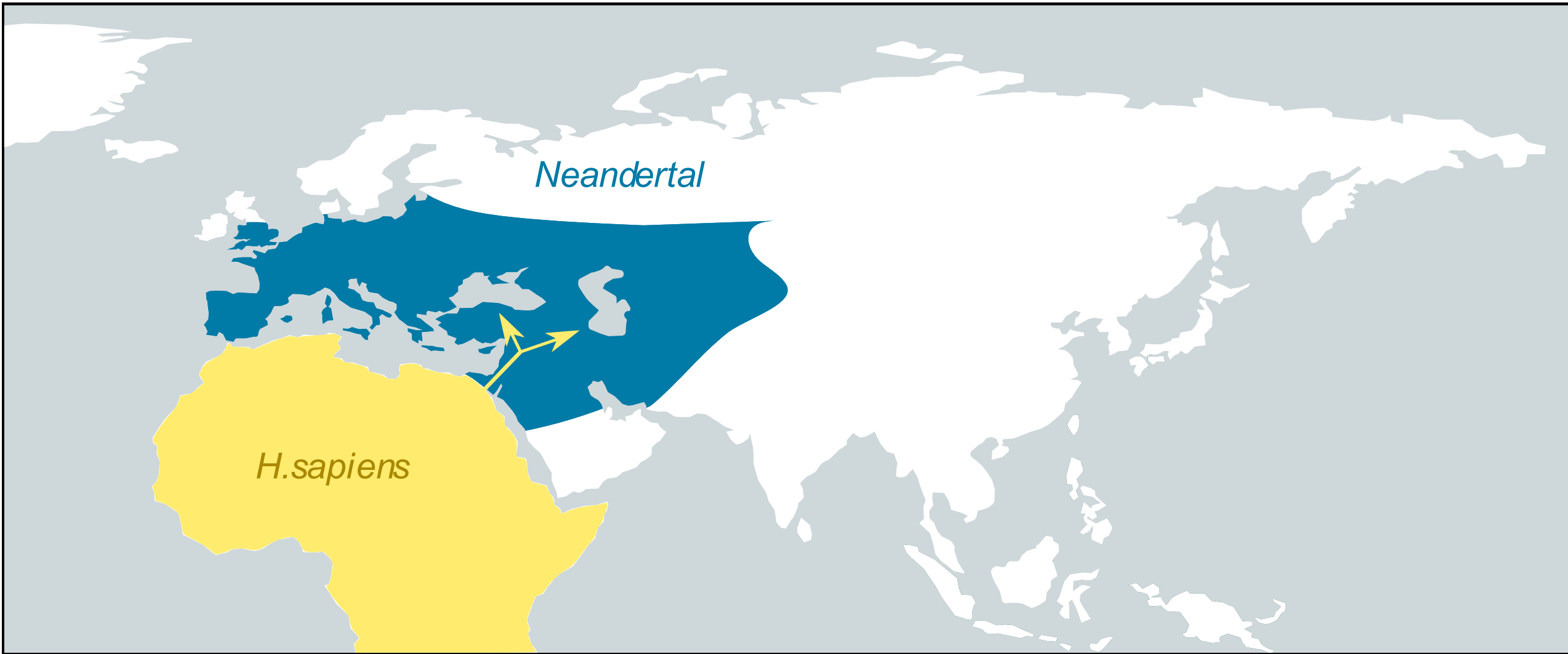
150.000 anni fa



Adattato da Mafessoni 2018, *Nat.Ecol.Evol.*

Green et al., 2010, *Science*

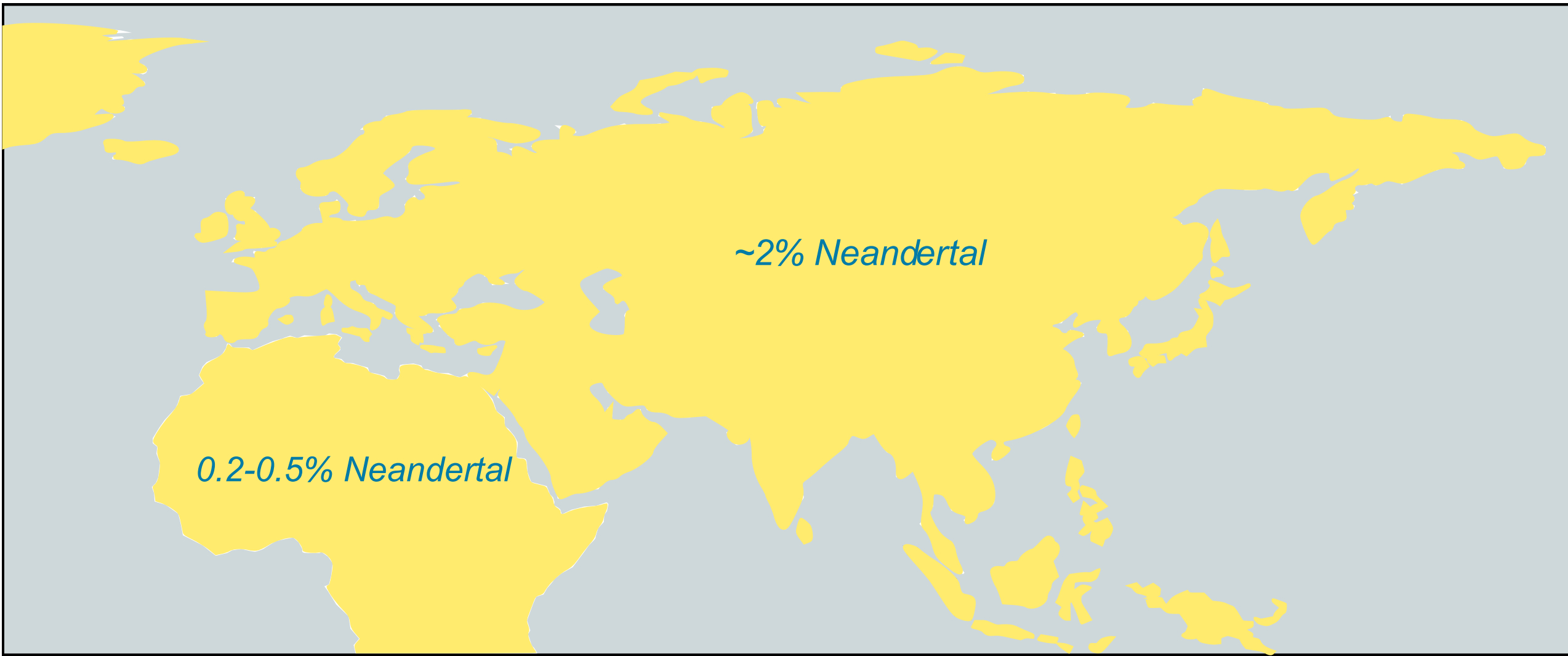
45-100.000 anni fa



Adattato da Mafessoni 2018, *Nat.Ecol.Evol.*

Green et al., 2010, *Science*

oggi



Adattato da Mafessoni 2018, *Nat.Ecol.Evol.*

Chen et al., 2020, *Cell*
Schaefer et al., 2021, *Sci.Ad.*

Che frazione del genoma di Neandertal potremmo ricostruire a partire dai genomi di moltissimi (diciamo migliaia, o tutta Trieste) di individui?

- ☐ 0%
- ☐ ~2%
- ☐ ~10%
- ☐ ~50%
- ☐ >90%

Che frazione del genoma di Neandertal potremmo ricostruire a partire dai genomi di moltissimi (diciamo migliaia, o tutta Trieste) di individui?

- ☐ 0%
- ☐ ~2%
- ☐ ~10%
- ☒ ~50%
- ☐ >90%

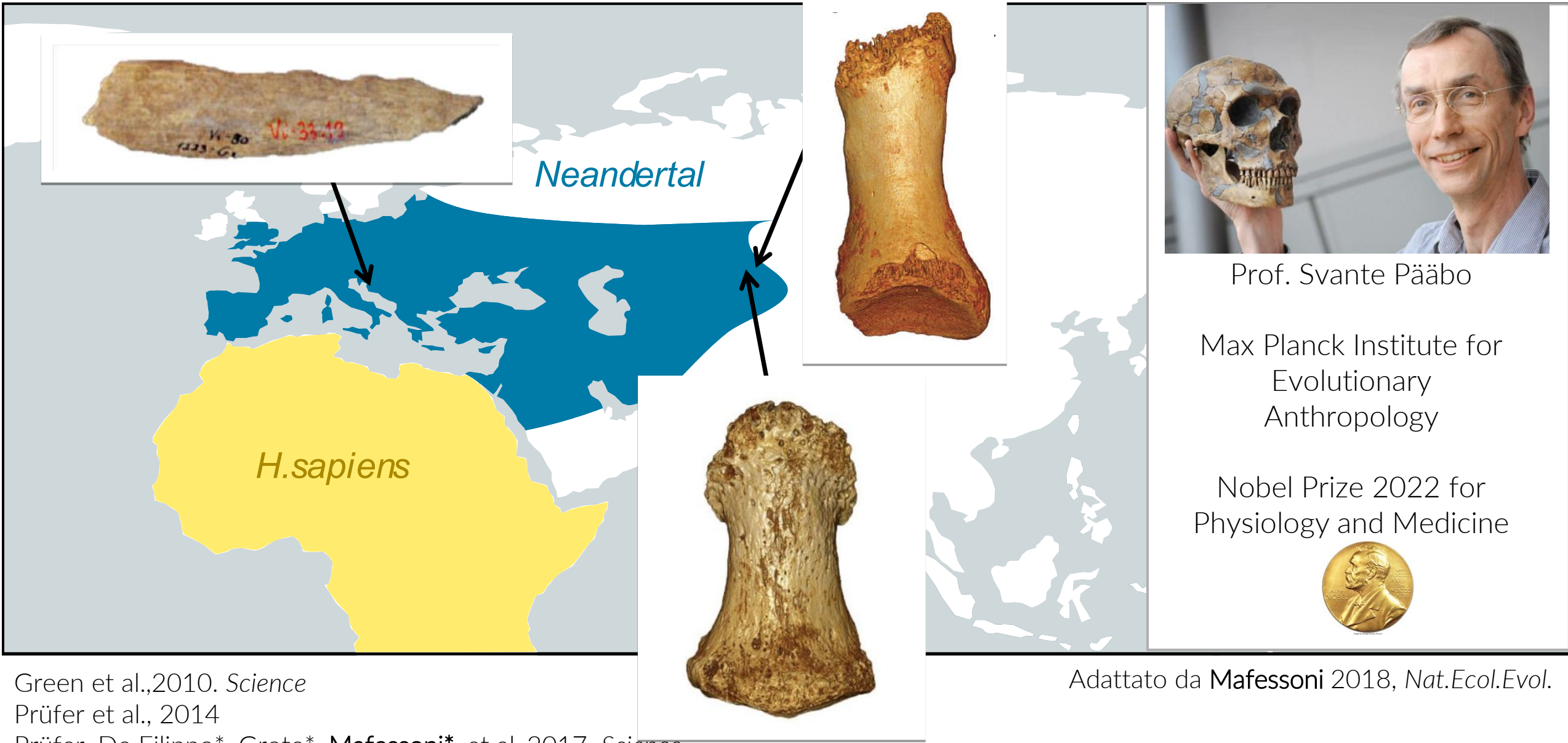
Aplotipi Neanderthal in 27,566 genomi di Islandesi



About 50% of the Neanderthal genome can be reconstructed using 27,566 genomes from present-day Islandic citizens

Adattato da from Skov, Call-Macia, Sveinbjörnsson , **Mafessoni** et al., 2020, *Nature*

Tre genomi di Neanderthal ad alta qualità (coverage > 25x)



Prof. Svante Pääbo

Max Planck Institute for
Evolutionary
Anthropology

Nobel Prize 2022 for
Physiology and Medicine



Adattato da **Mafessoni** 2018, *Nat.Ecol.Evol.*

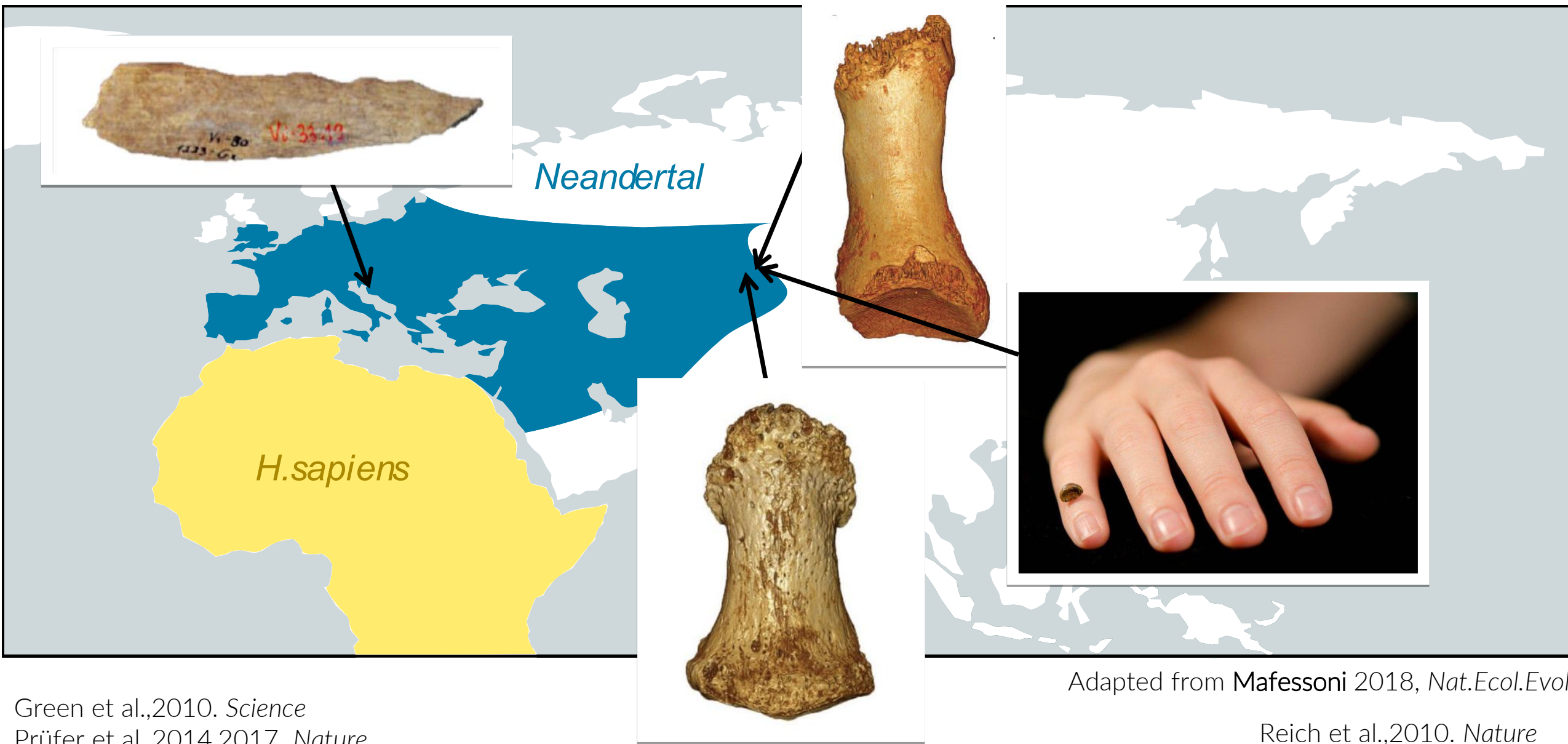
Green et al., 2010. *Science*

Prüfer et al., 2014

Prüfer, De Filippo*, Grote*, **Mafessoni***, et al., 2017. *Science*

Mafessoni et al., 2020. *PNAS*

Un genoma appartenente ad un ominide sconosciuto: i Denisova

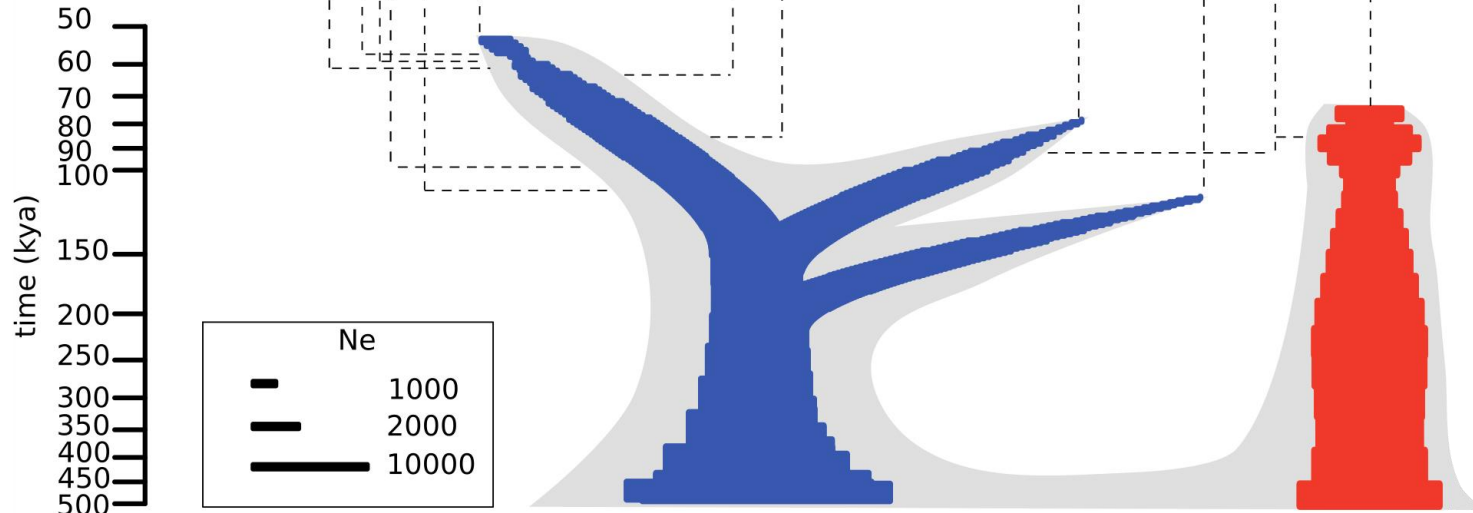
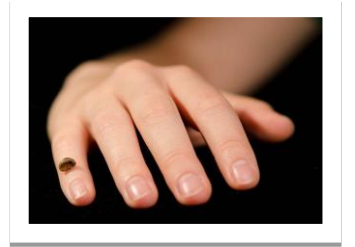


Adapted from Mafessoni 2018, *Nat.Ecol.Evol.*

Green et al.,2010. *Science*
Prüfer et al.,2014,2017. *Nature*
Mafessoni et al.,2020. *PNAS*

Reich et al.,2010. *Nature*
Meyer et al.,2012, *Science*

I **Denisova** furono una popolazione separata dai **Neanderthal** per centinaia di migliaia di anni







Meyer et al.,2011, *Science*
Zhang et al., 2020 , *Science*
Browning et al., 2018, *Cell*

L'antenato delle popolazioni umani del Sahul e dell'Asia "scambiò geni" con i Denisova



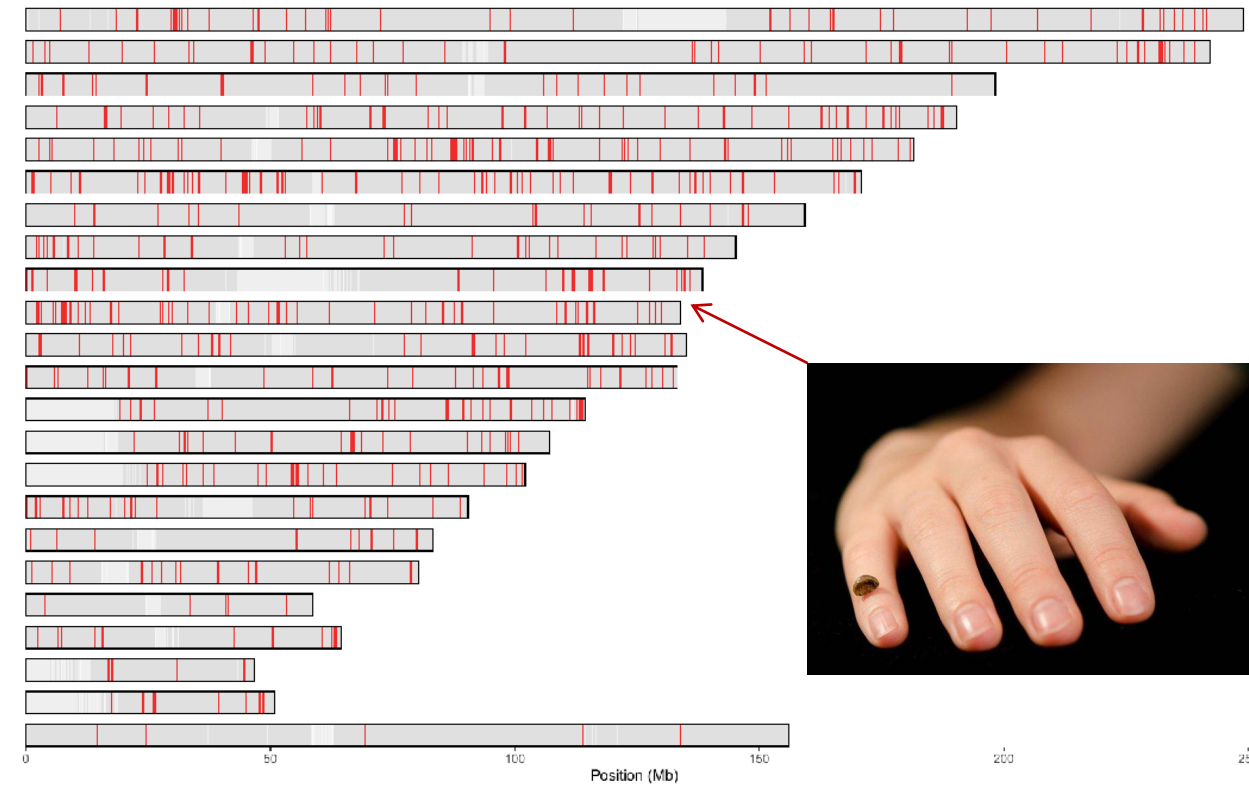
Browning et al., 2018, *Cell*
Texeira et al., 2019, *PNAS*
Zhang et al., 2020, *Science*

Adapted from Mafessoni 2018, *Nat.Ecol.Evol.*

Frammenti di DNA dal genoma di Denisova si trovano anche negli Europei



www.guidetoiceland.imgix.net



Adapted from Skov, Call-Macia, Sveinbjörnsson , **Mafessoni** et al., 2020,
Metamorphosis

Il nostro passato è stato costellato da molteplici eventi di flusso genico (aka making babies)



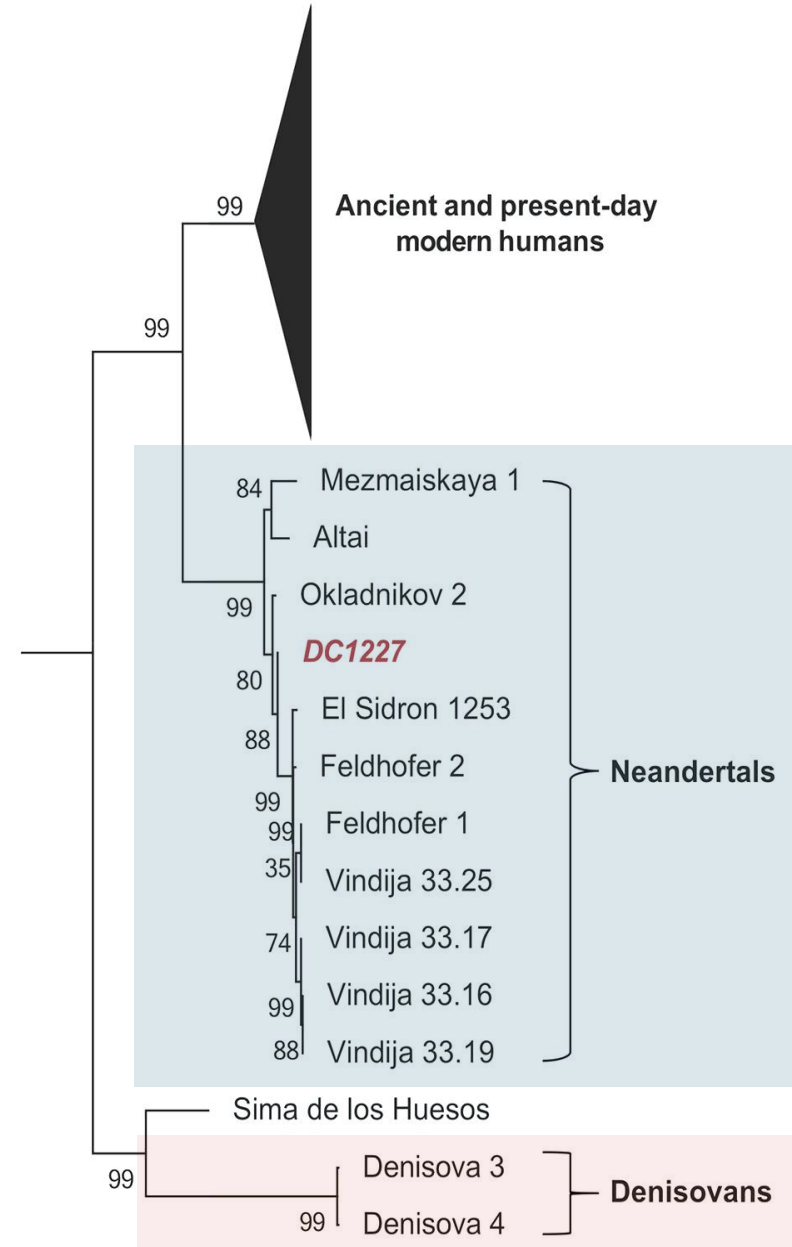
Browning et al., 2018, *Cell*
Texeira et al., 2019, *PNAS*
Zhang et al., 2020, *Science*

Adapted from Mafessoni 2018, *Nat.Ecol.Evol.*

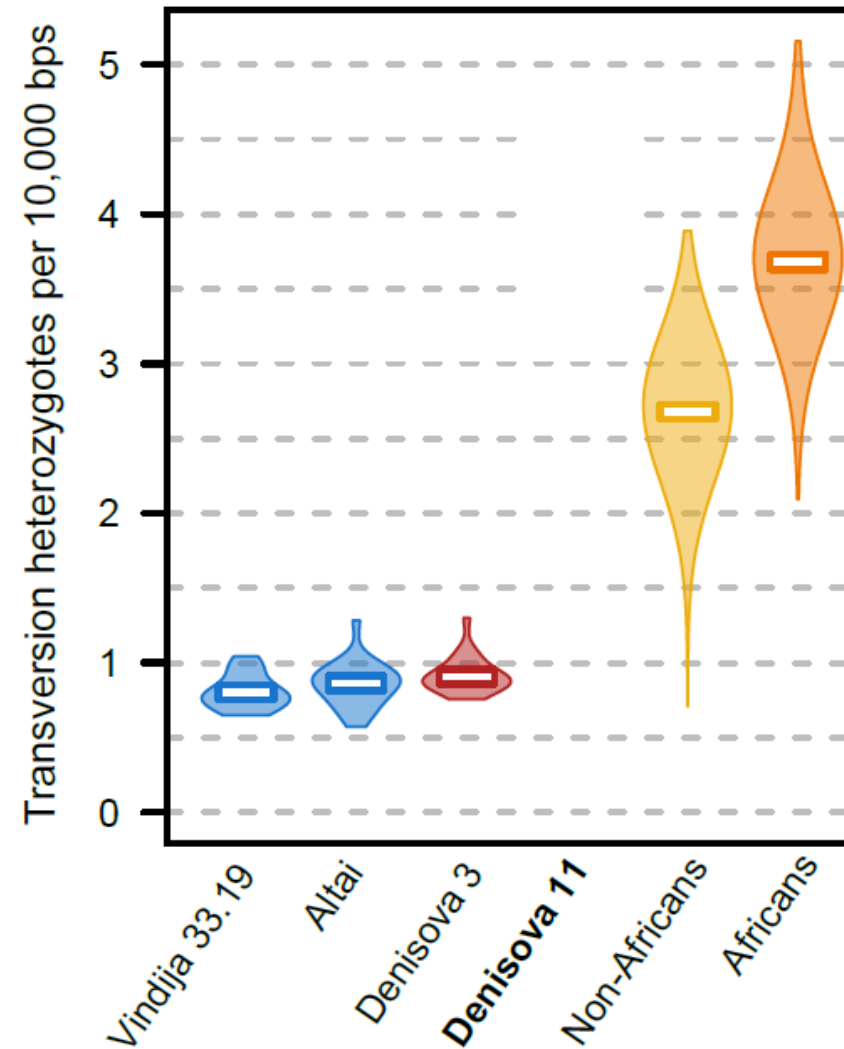
Il DNA mitocondriale di *Denisova 11* indica che ebbe una mamma Neanderthal



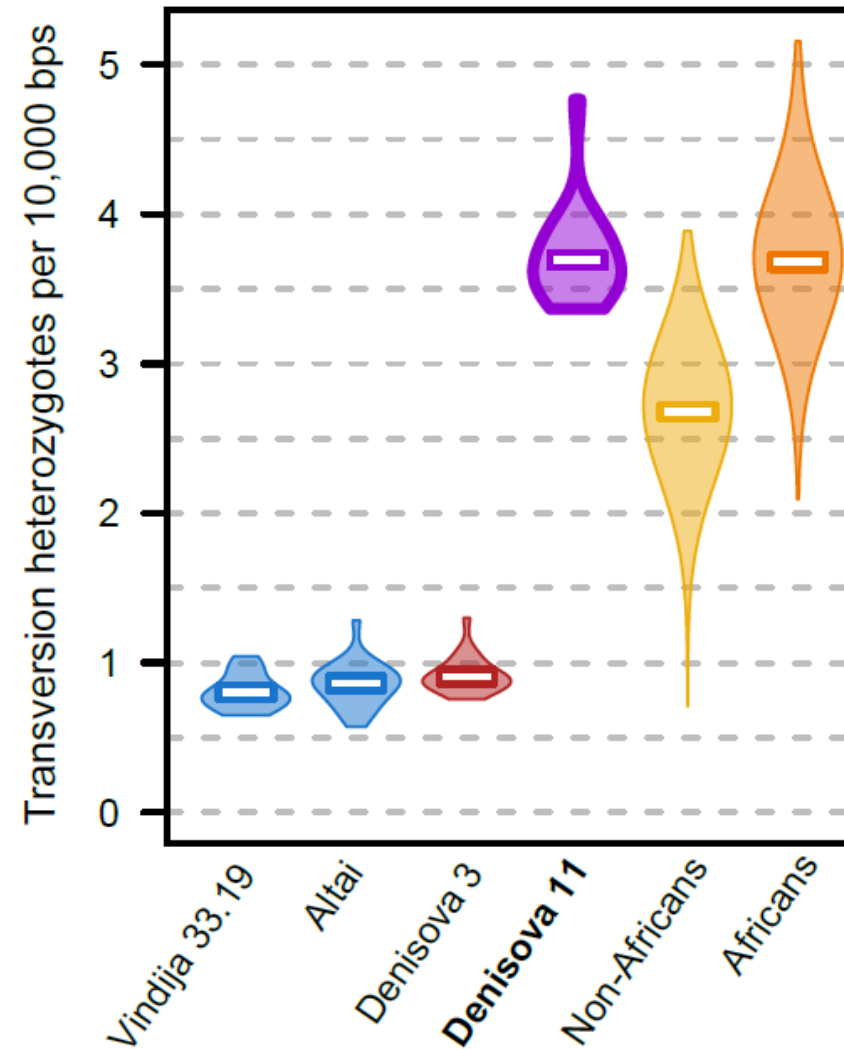
Denisova 11
(Sample identified as “human”
via Zooms)



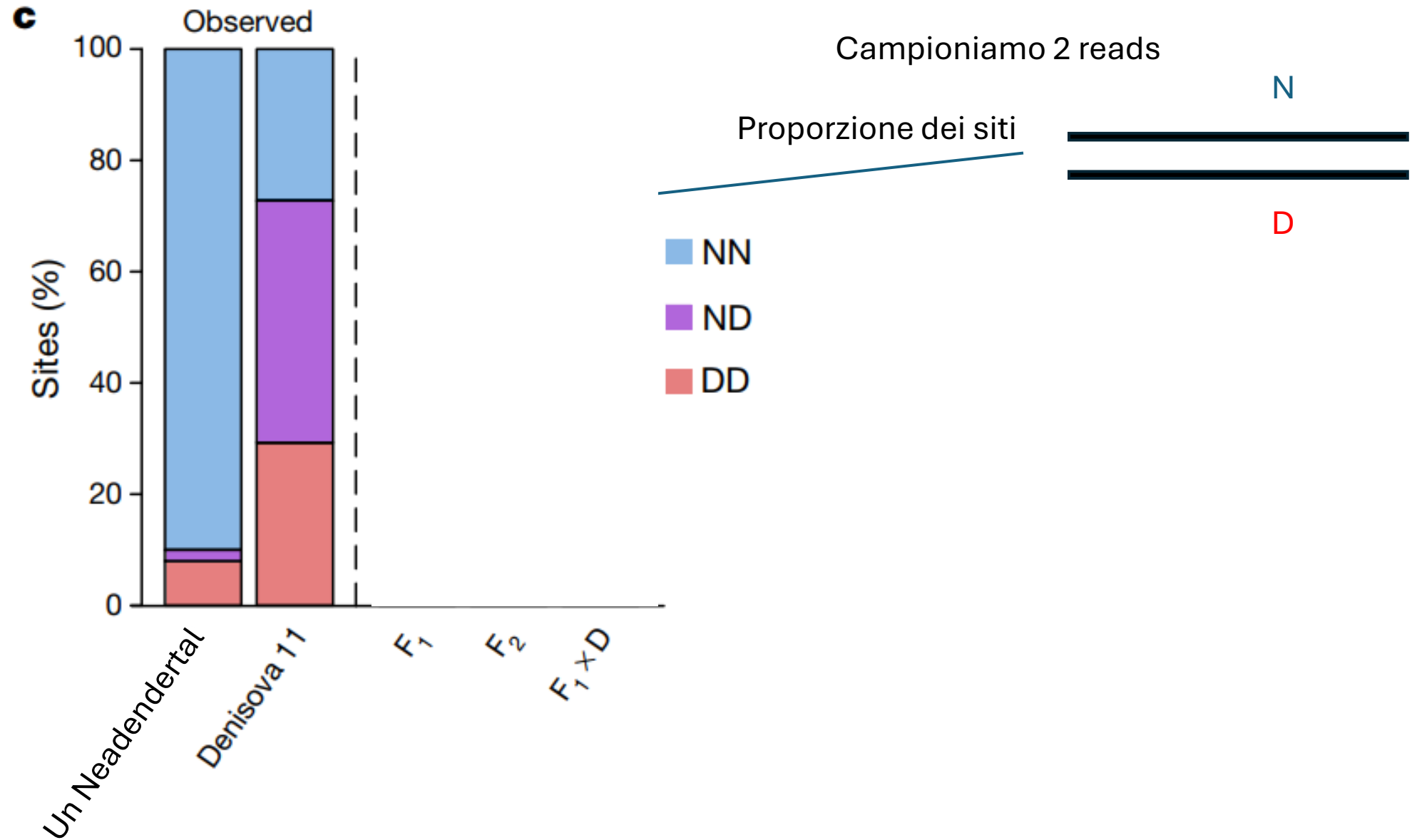
Il genoma nucleare (coverage 2.7x) di *Denisova 11* indica moltissimi siti eterozigoti



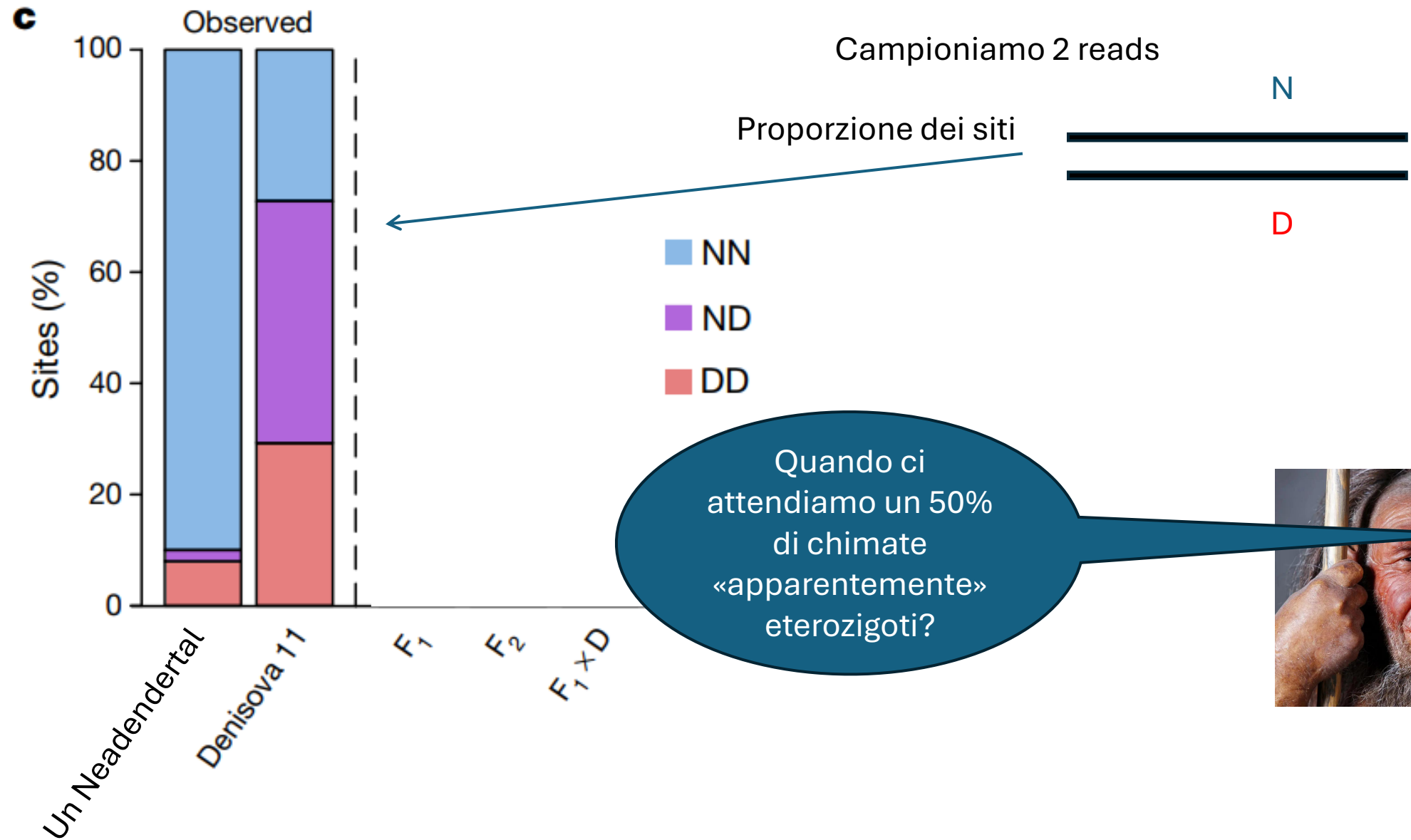
Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



La nostra distribuzione binomiale!

Fattoriali

$$n! = n * (n-1) * (n-2) * \dots * 1$$

Esempi:

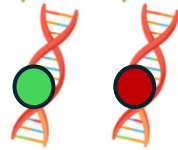
$$3! = 3 * 2 * 1 = 6$$

$$5! = 5 * 4 * 3 * 2 * 1 = 120$$

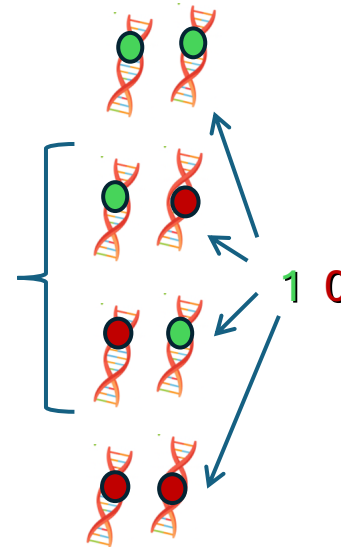


$$\Pr(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Esempio: $k=1$ e $n=2$



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



dove $\binom{n}{k} = n! / (k!(n-k)!)$

è il coefficiente
binomiale e descrive il
numero di possibili
combinazioni di alleli
ALT e REF

Un modello probabilistico di genotipi molto semplicistico

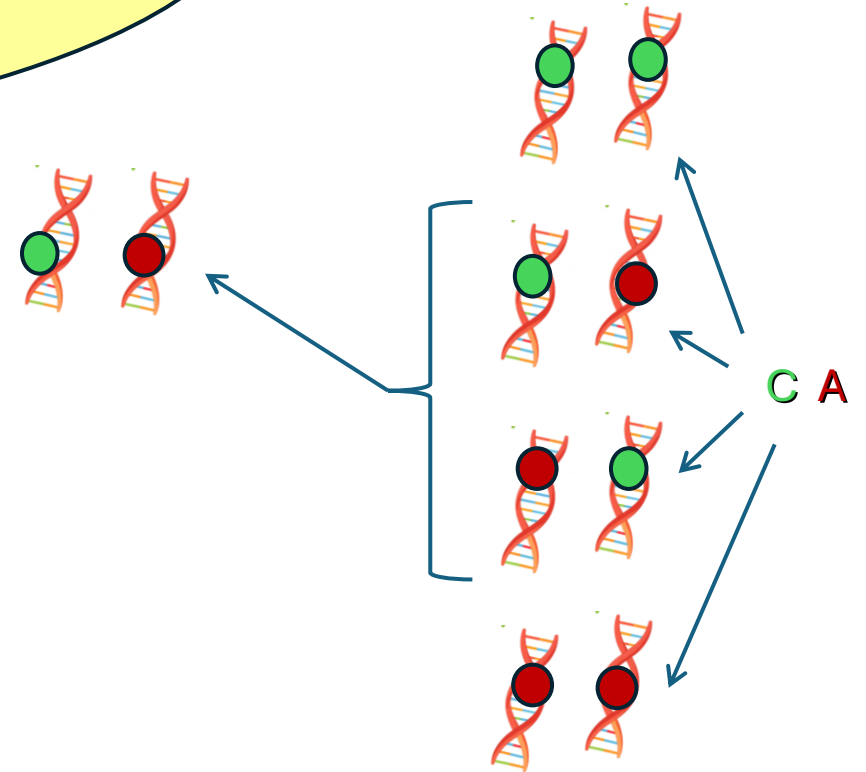
Notate che ottenere due reads  e  sono due (micro)eventi distinti, ognuno con la stessa probabilità $\frac{1}{2} * \frac{1}{2} = 1/4$.

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**.

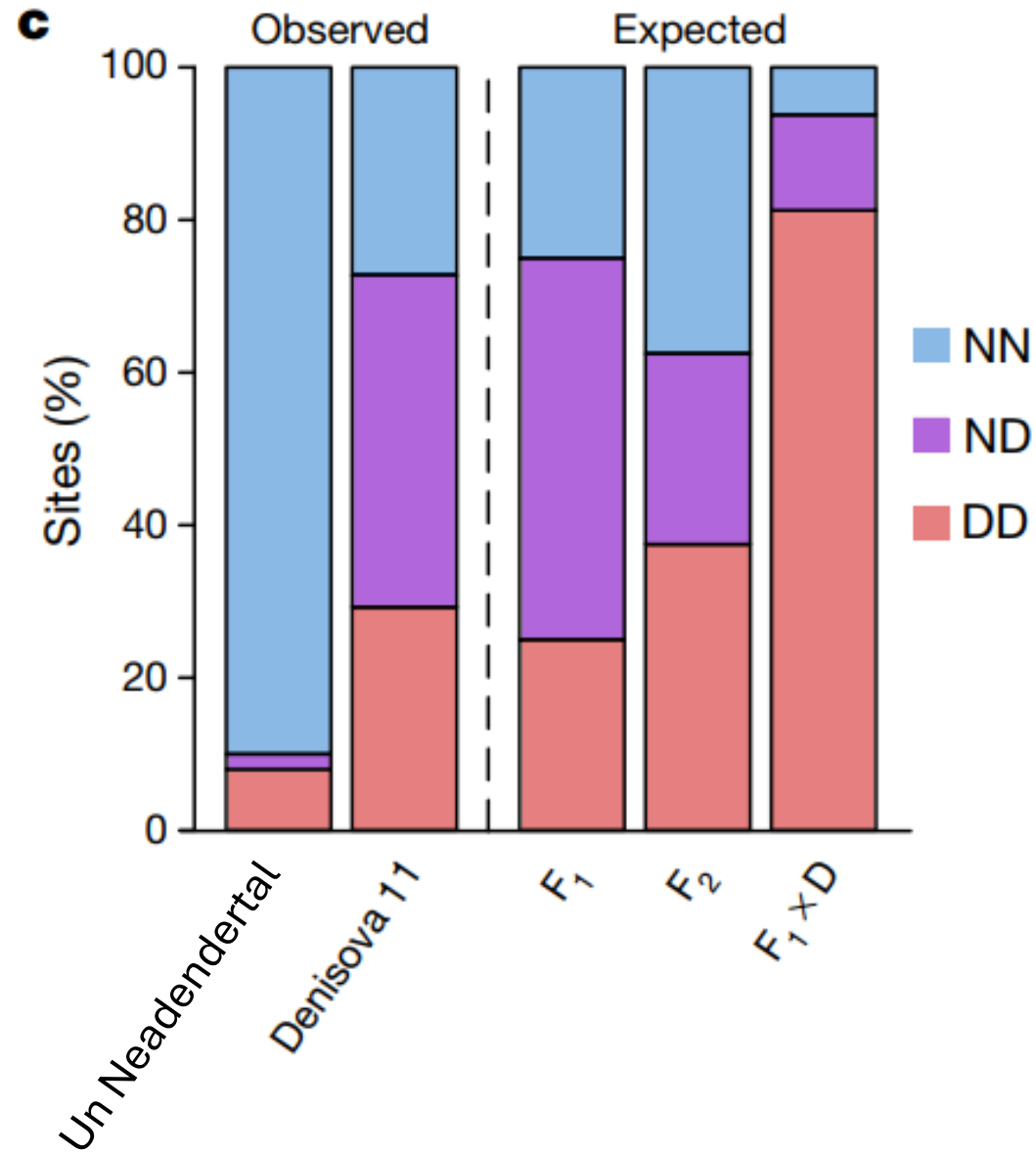


Probabilità (Pr) di osservare due reads   (ignorandone l'ordine)

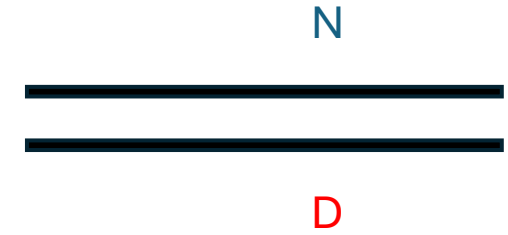
- da un genotipo **C A** = $\frac{1}{2} * \frac{1}{2} * 2 = 1/2$



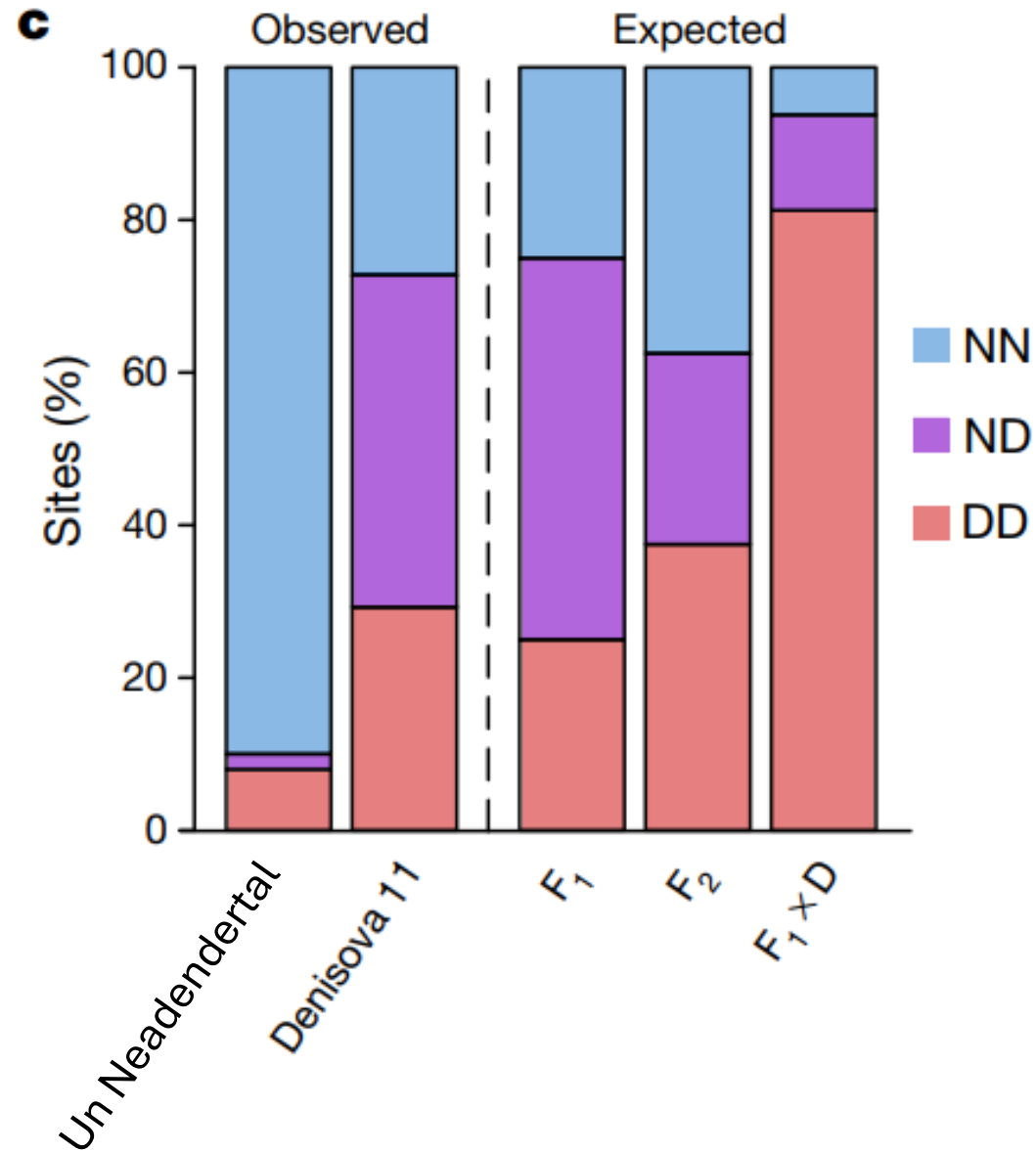
Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



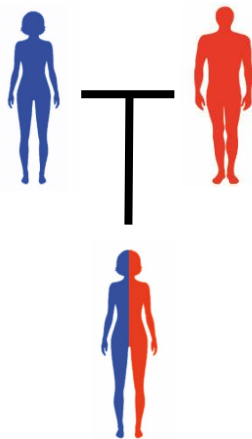
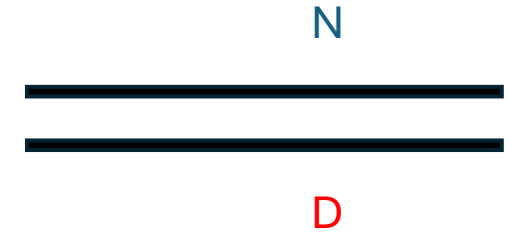
Campioniamo 2 reads



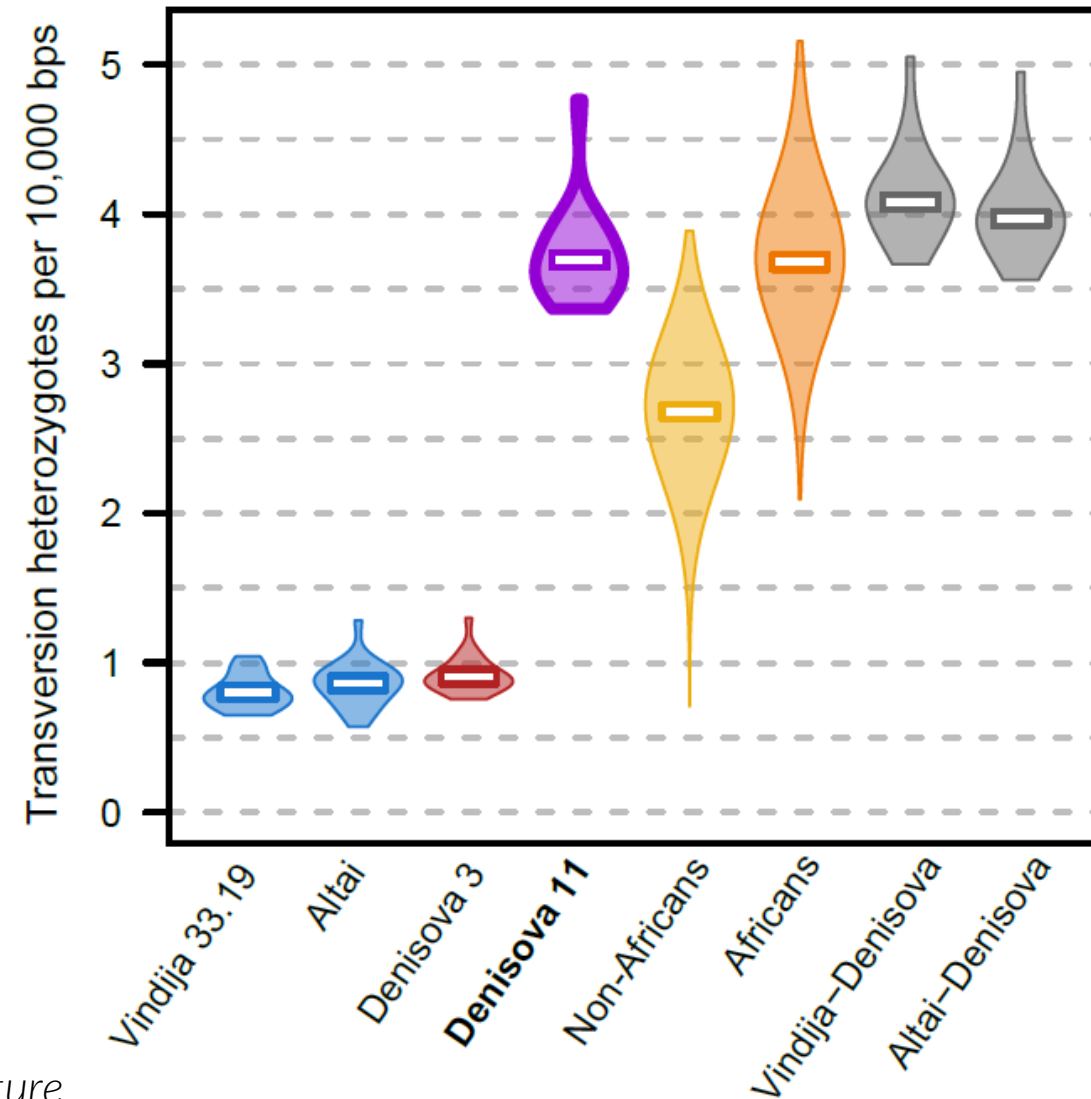
Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



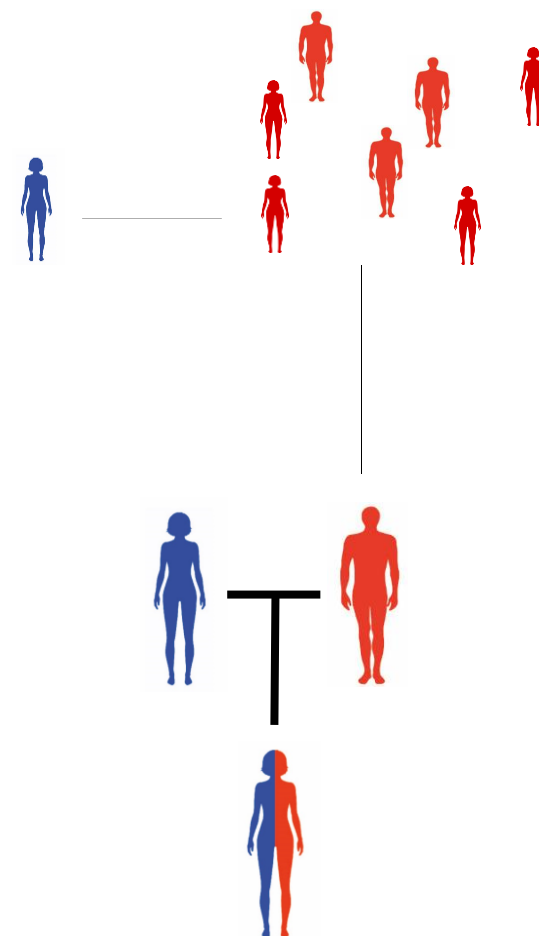
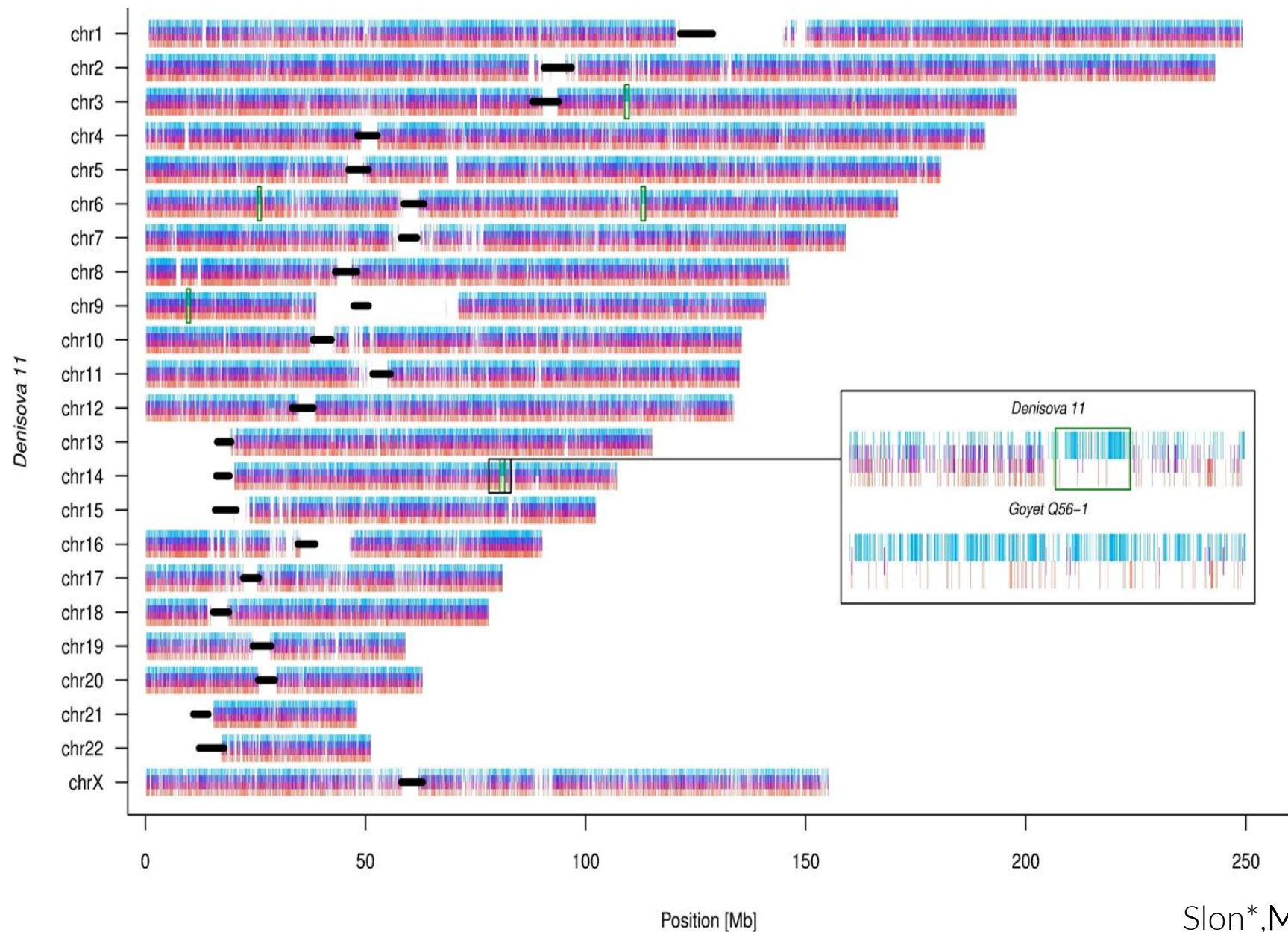
Campioniamo 2 reads



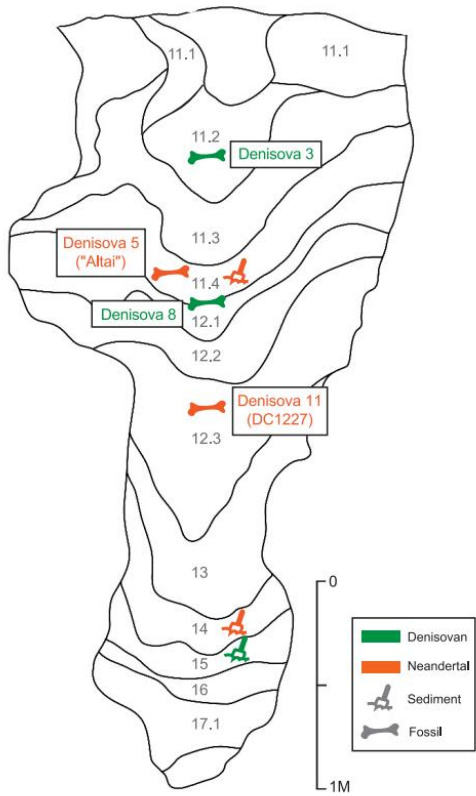
Il genoma nucleare di *Denisova 11* indica moltissimi siti eterozigoti



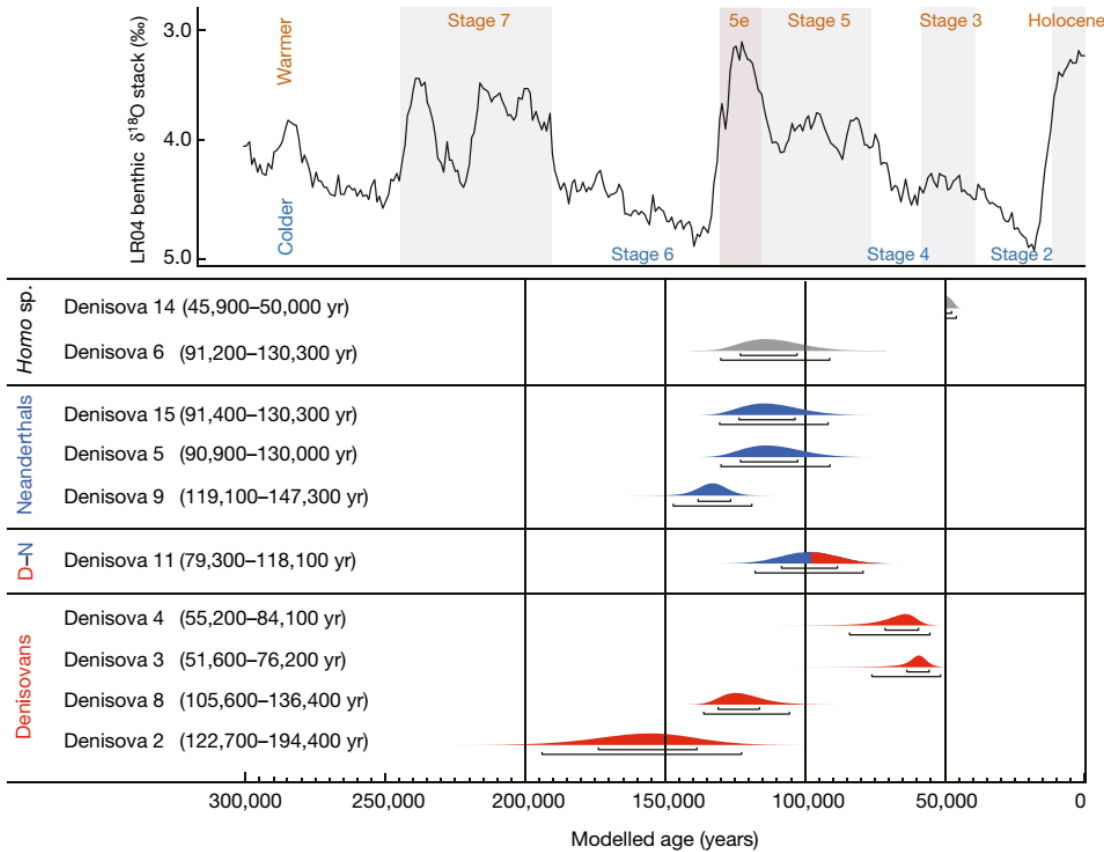
Denisova 11 was the daughter of a Neanderthal mother and a Denisovan father



Various evidence shows that Neanderthal and Denisova repeatedly succeeded each other in Denisova cave

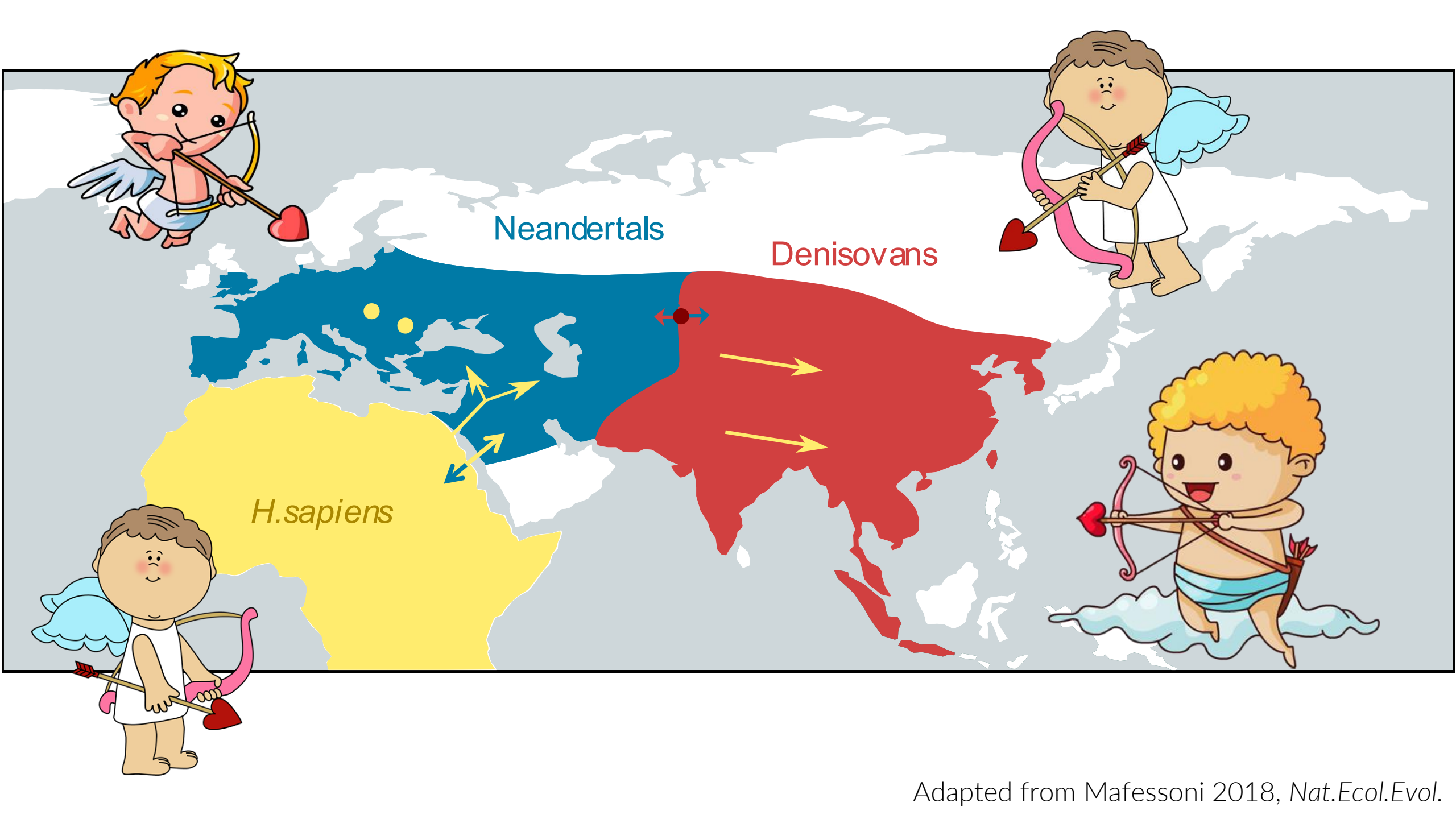


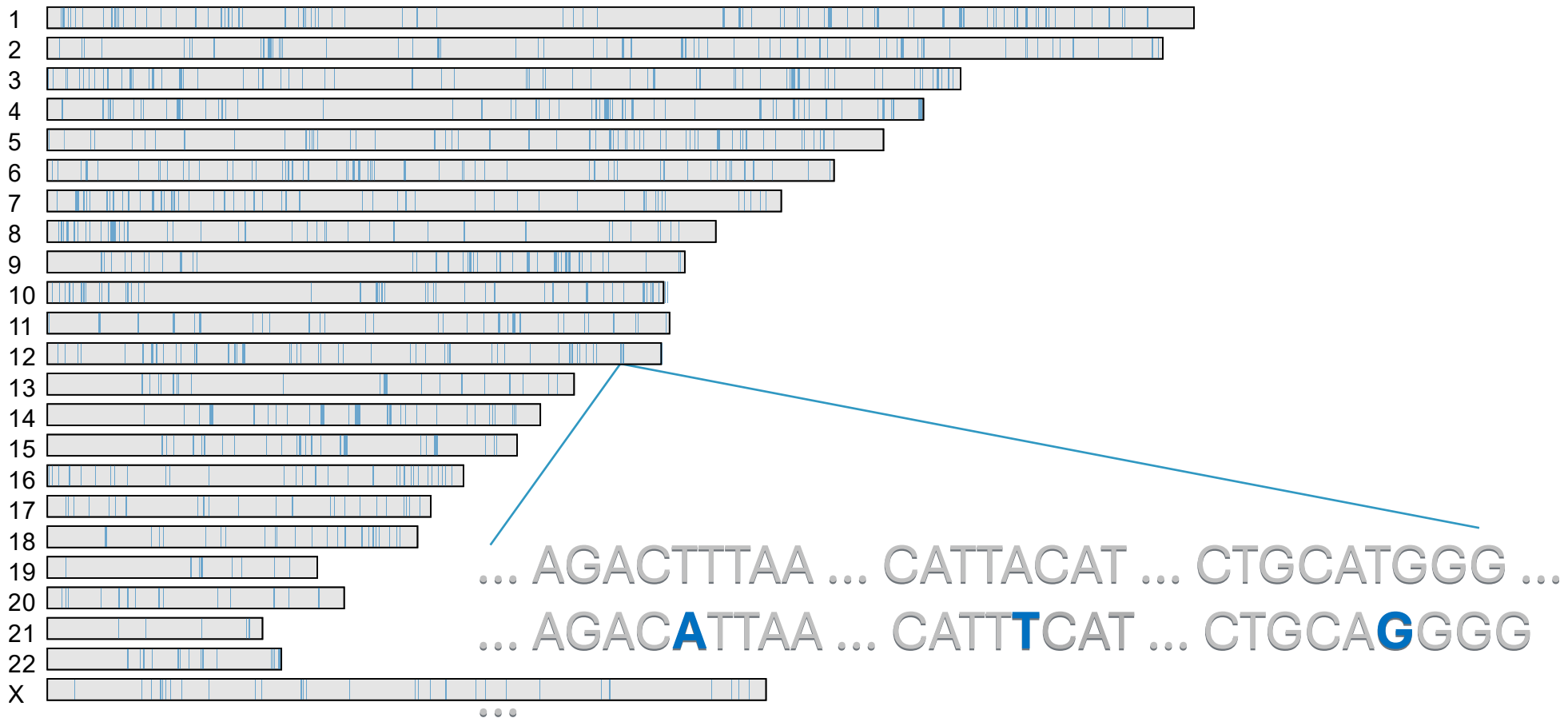
DNA from soil



Bayesian model with chronoarcheological and genetic evidence

Slon, Hopfe, Weiss, **Mafessoni** et al., 2017, *Science*
Vernot, Zavala, Gomez, Jacobs, Slon, **Mafessoni**, et al., 2021, *Science*
Douka et al., 2019, *Nature*





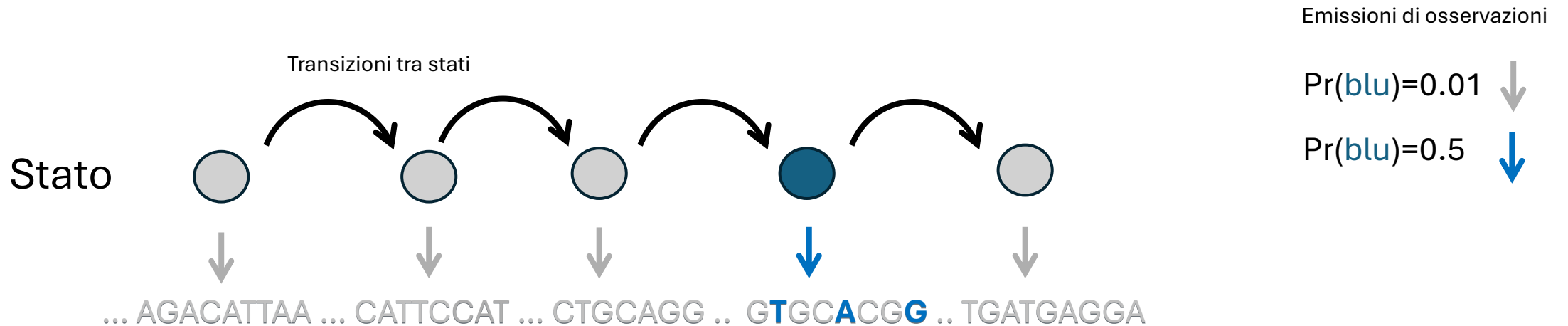
... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

...

... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

Hidden-Markov Models

Classe di algoritmi probabilistici utilizzatissimi in genomica in quanto rappresentano la naturale struttura del genoma in una serie di elementi che si susseguono linearmente



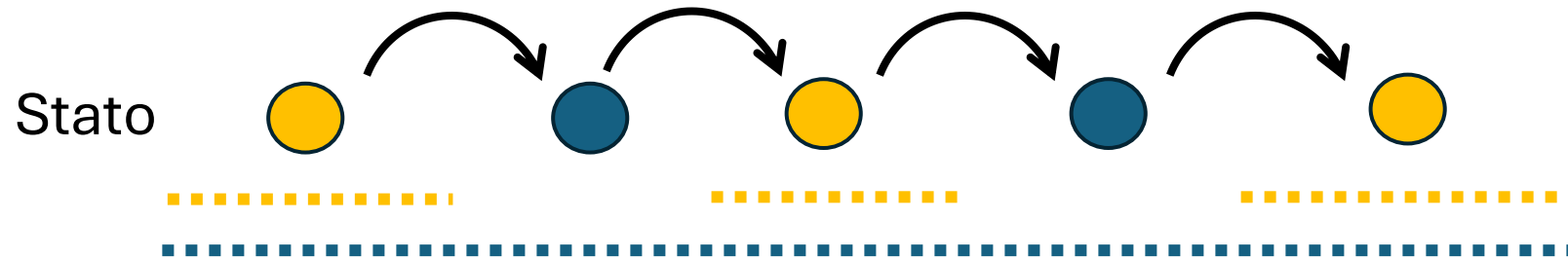
Esempio:

Regione non-metilata ●

Regione metilata ○

Hidden-Markov Models

Classe di algoritmi **probabilistici** utilizzatissimi in genomica in quanto rappresentano la naturale struttura del genoma in una serie di elementi che si susseguono linearmente



Esempio:

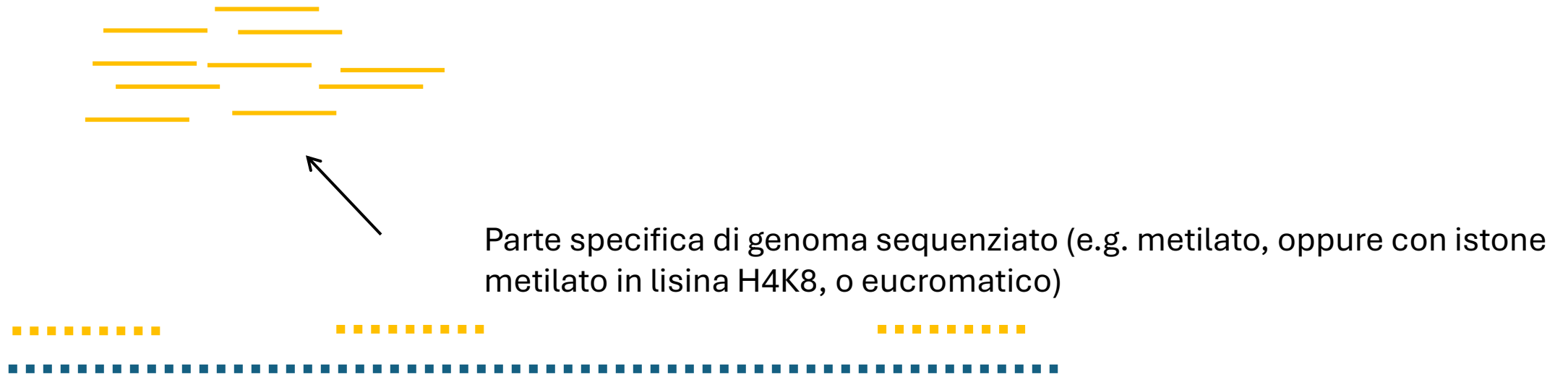
Regione non-metilata



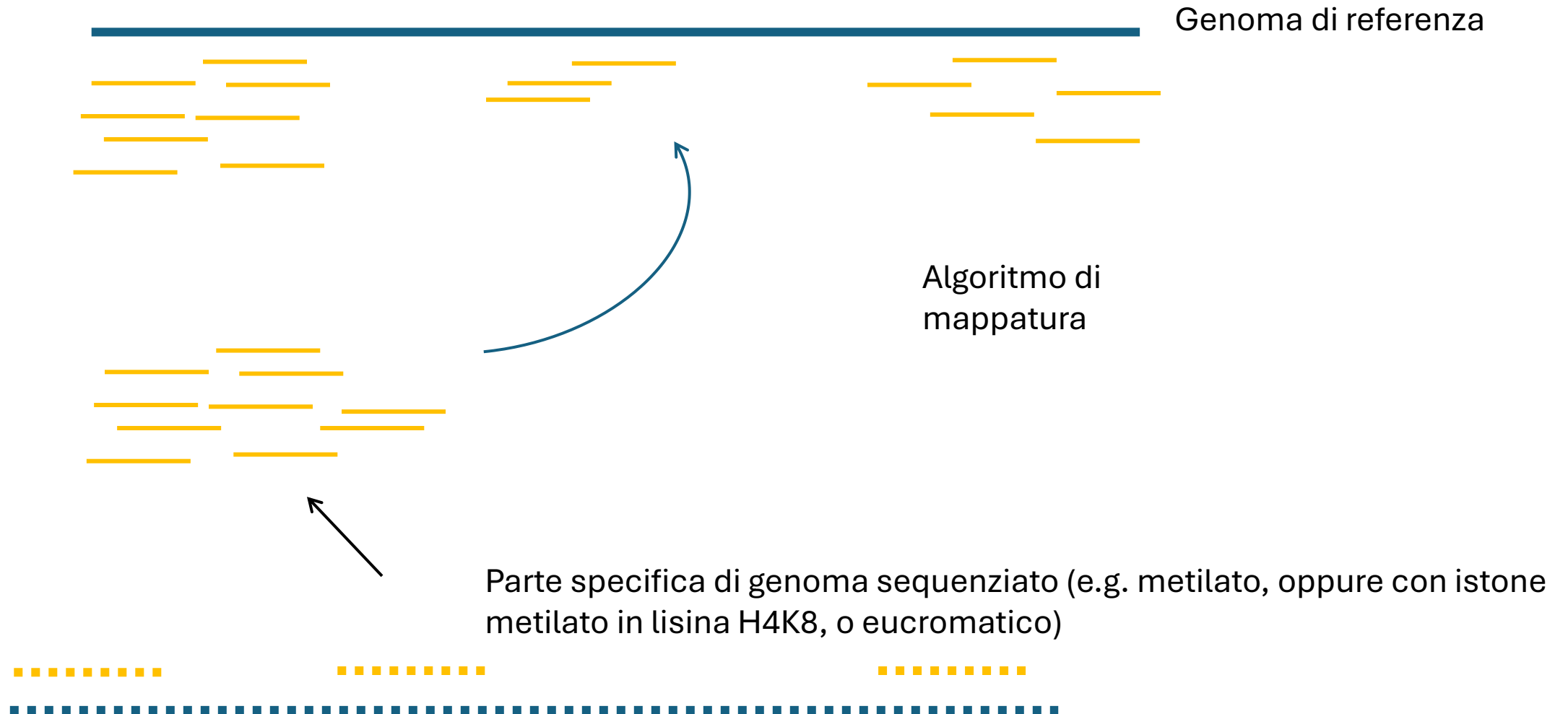
Regione metilata



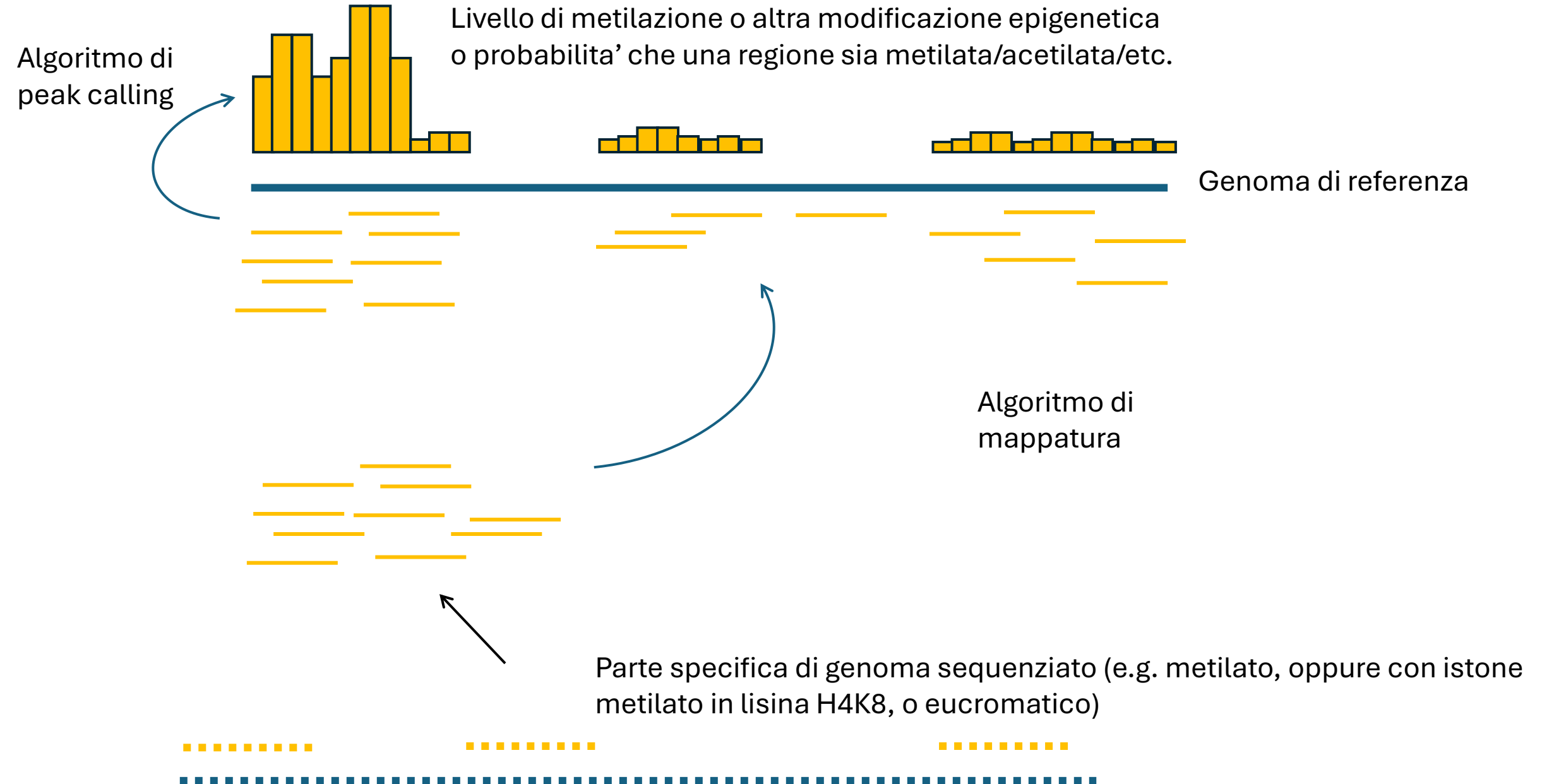
Peak calling per epigenomica



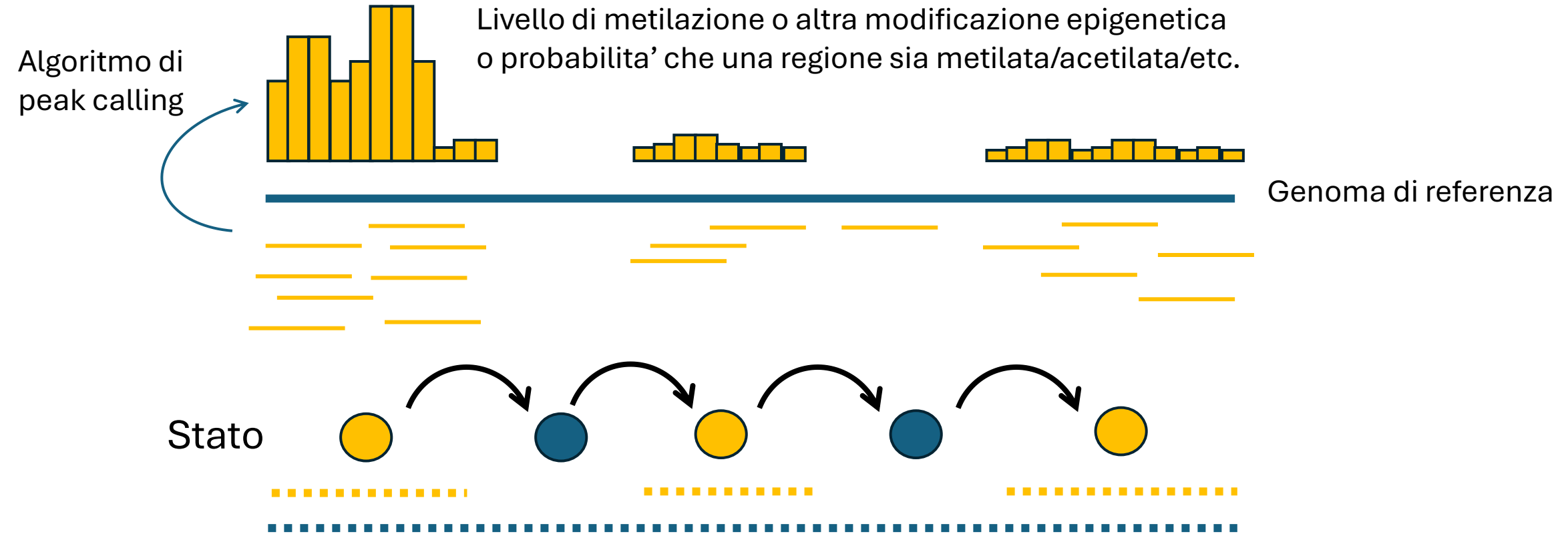
Peak calling per epigenomica



Peak calling per epigenomica



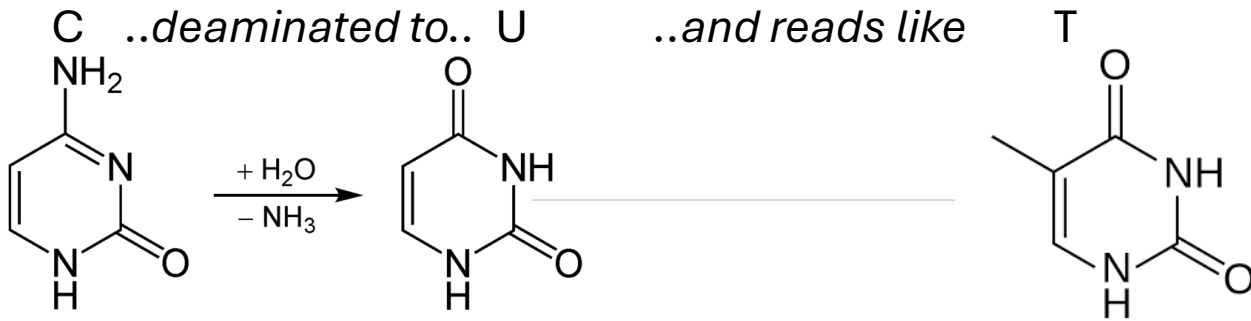
Peak calling per epigenomica



Esempio:

- Regione non-metilata 
- Regione metilata 

Deamination can be exploited in ancient epigenomics to infer regions differentially methylated between species



MODERN HUMAN

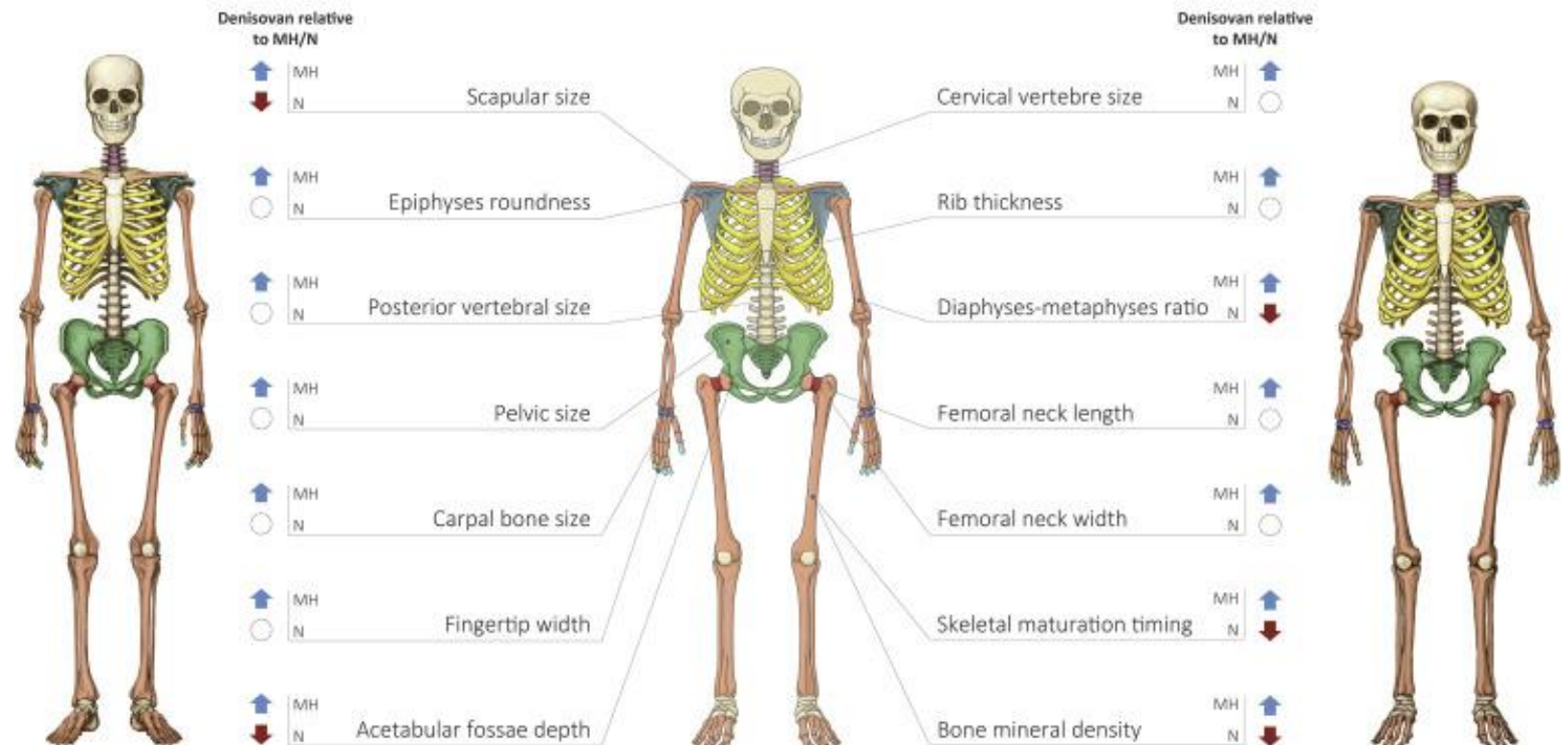
DENISOVAN

NEANDERTHAL

Faster for methylated C.



We can identify methylated regions in ancient genomes!!



Profile HMM e detezione di motivi proteici e geni con omologie remote

Pfam

