Ultima lezione (prima dei progetti)!

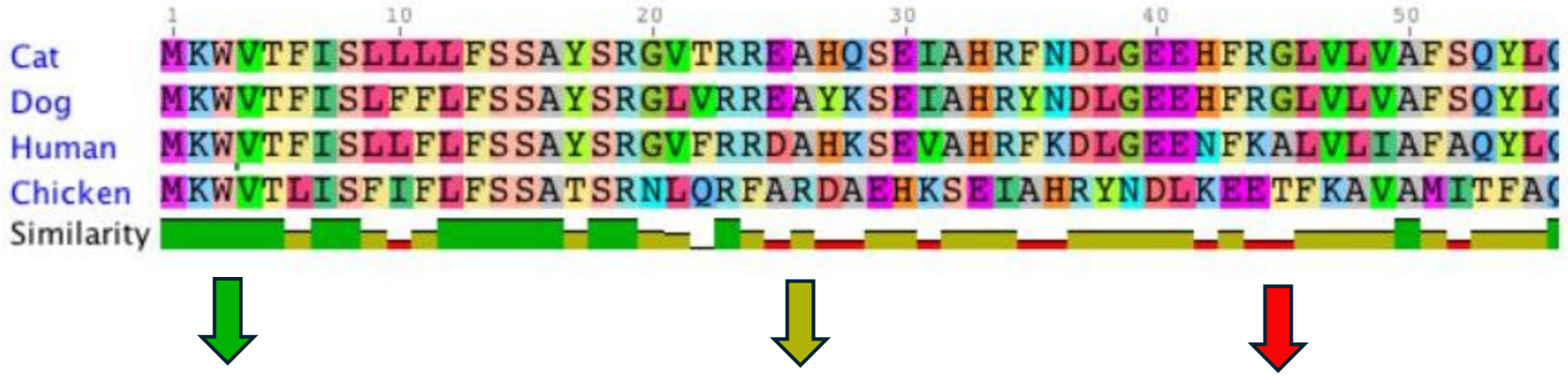
- *Phastcons* e predizioni di conservazione/patogenicità (l'avete chiesto voi)
- Algoritmi deterministici/probabilistici vs AI
- Refusi e ultimi concetti
- Progetti

Abbiamo un sacco di genomi! E ora che ci facciamo??

- Trovare regioni importanti nel genoma con nuove funzioni
- Trovare le mutazioni che causano malattie
- Trovare le mutazioni da selezionare o ingegnerizzare per creare batteri che degradano plastica, piante piu' resistenti ai patogeni o che producono di piu'
- Etc.

La conservazione genetica contiene informazione sulla funzionalità del genoma

Allineamento muliplo dei prima 55 amino acidi del gene dell'albumina



Molto conservate

Probabilmente funzionale e con mutazioni deleterie/patogeniche

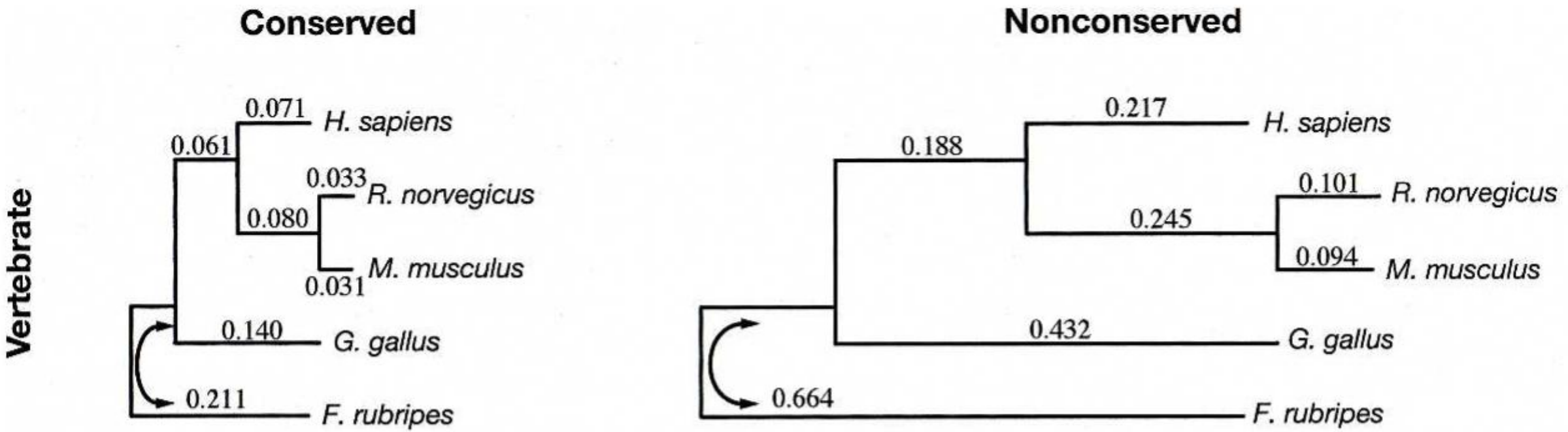
Non conservate

Probabilmente neutrali o con mutazioni che hanno poche o nessuna conseguenza

Moltissime differenze

Potenzialmente con evoluzione accelerata

Possiamo pensare alle regioni conservate come a degli “alberi filogenetici” piu’ corti



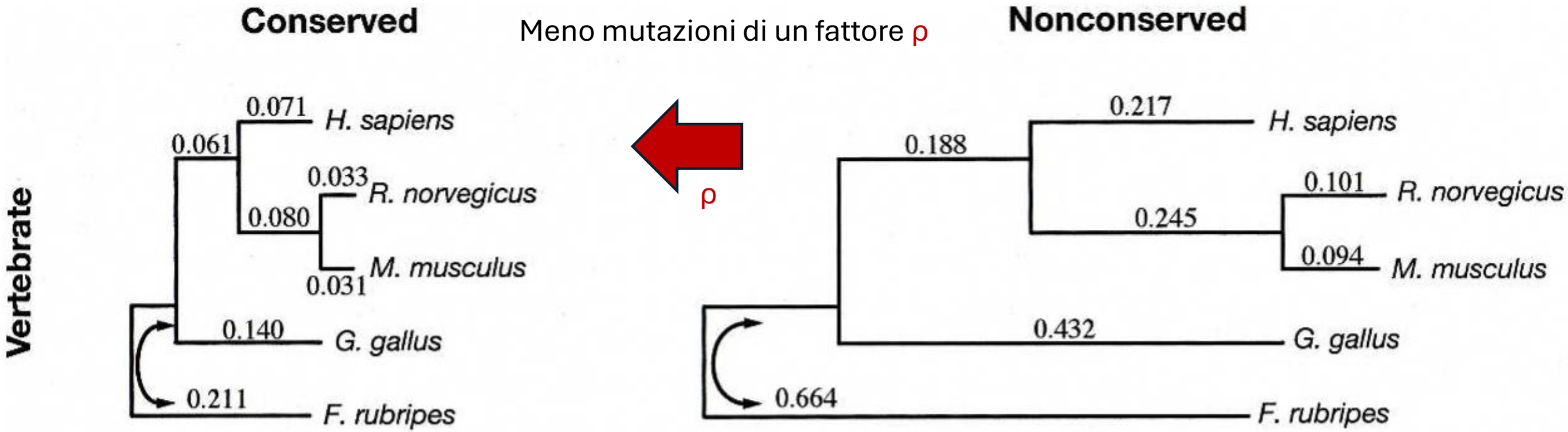
NB: Rami piu’ corti indicano semplicemente meno mutazioni

Possiamo pensare alle regioni conservate come a degli “alberi filogenetici” piu’ corti

ρ (rho): proporzione della lunghezza dei rami

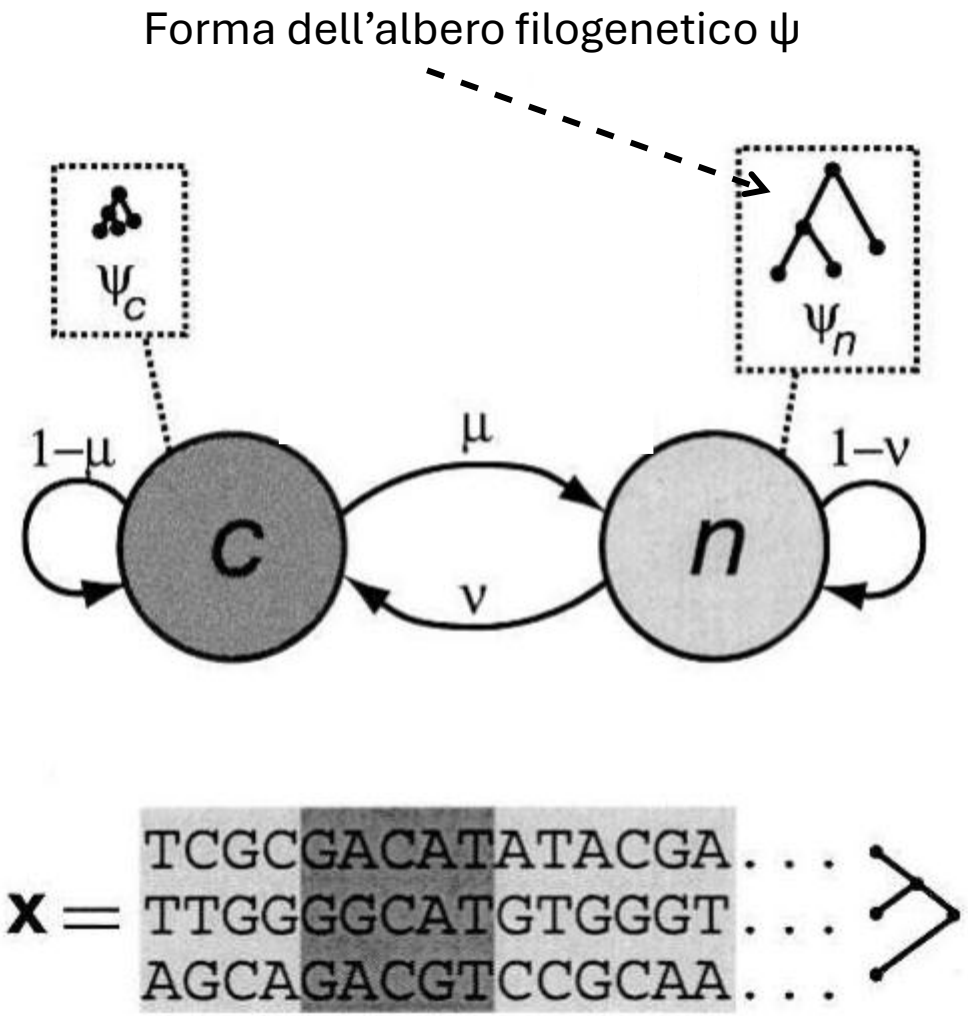
~

Meno mutazioni di un fattore ρ



NB: Rami piu’ corti indicano semplicemente meno mutazioni

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma



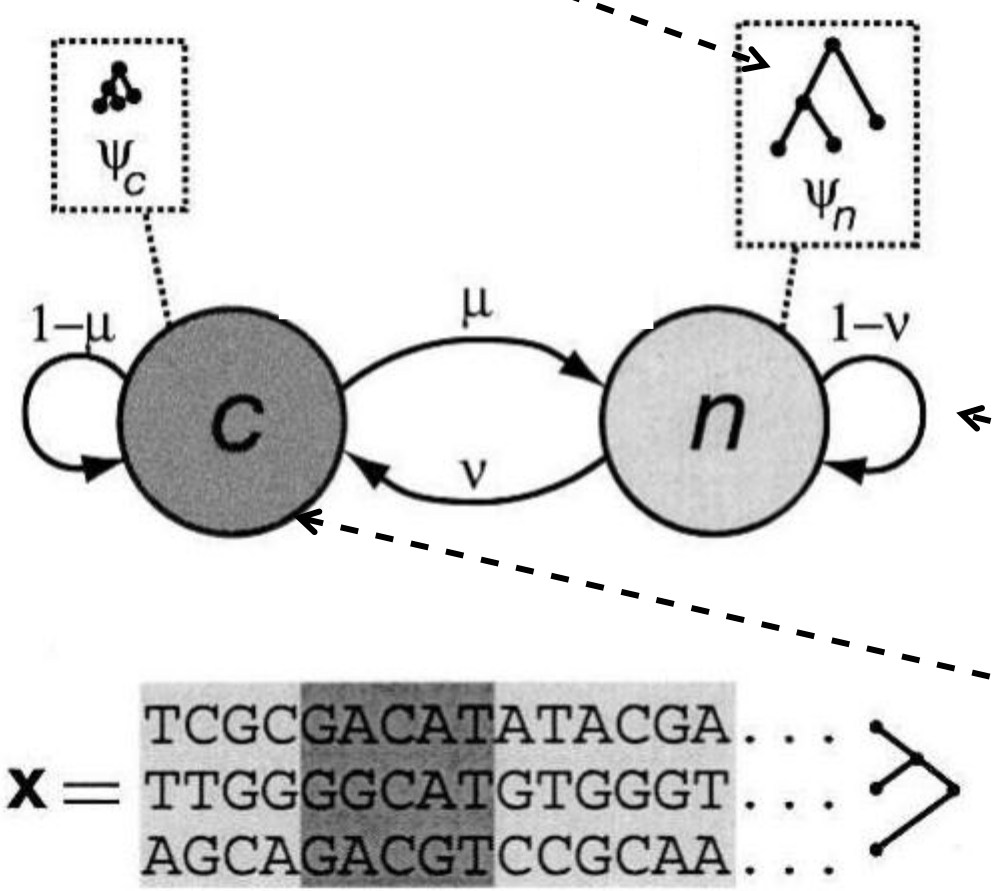
Due stati:

Conservato (c)
e
Non-Conservato (n)

Siepel et al., Genome Research, 2005

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma

Forma dell'albero filogenetico ψ



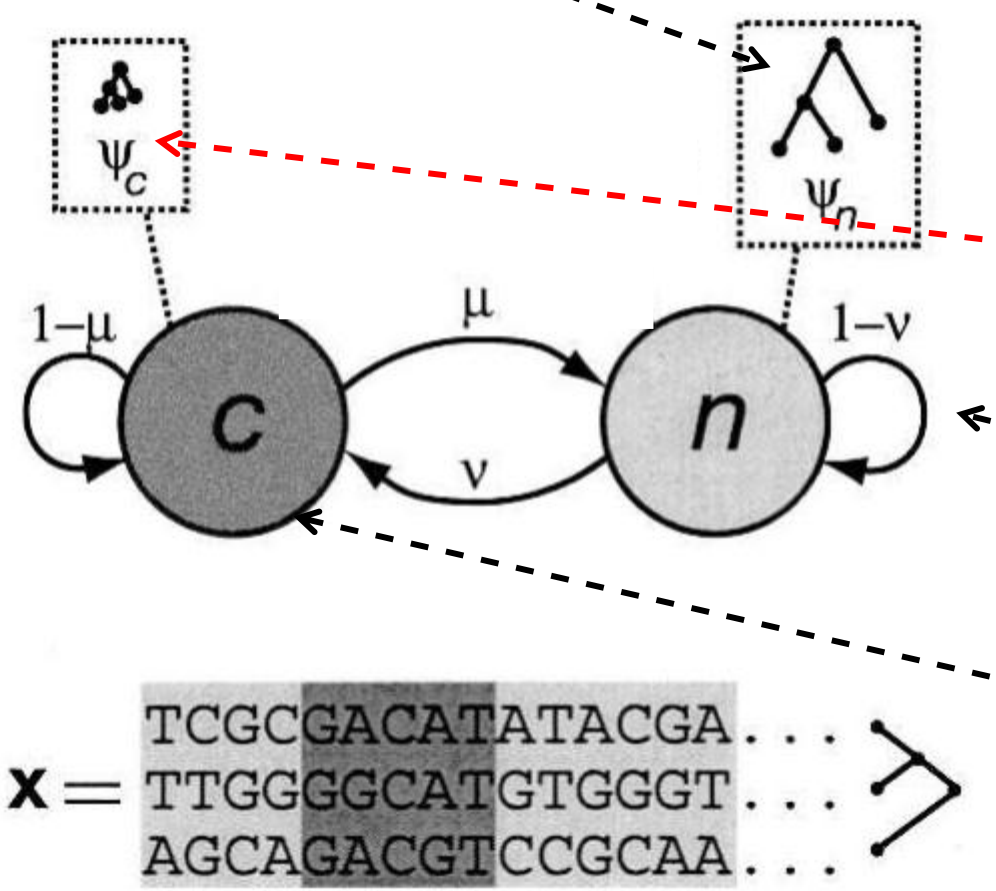
Parametri stimati

- v: probabilità di transizione dallo stato non-conservato a quello conservato
- μ: probabilità di transizione dallo stato conservato a quello non-conservato

Siepel et al., Genome Research, 2005

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma

Forma dell'albero filogenetico ψ



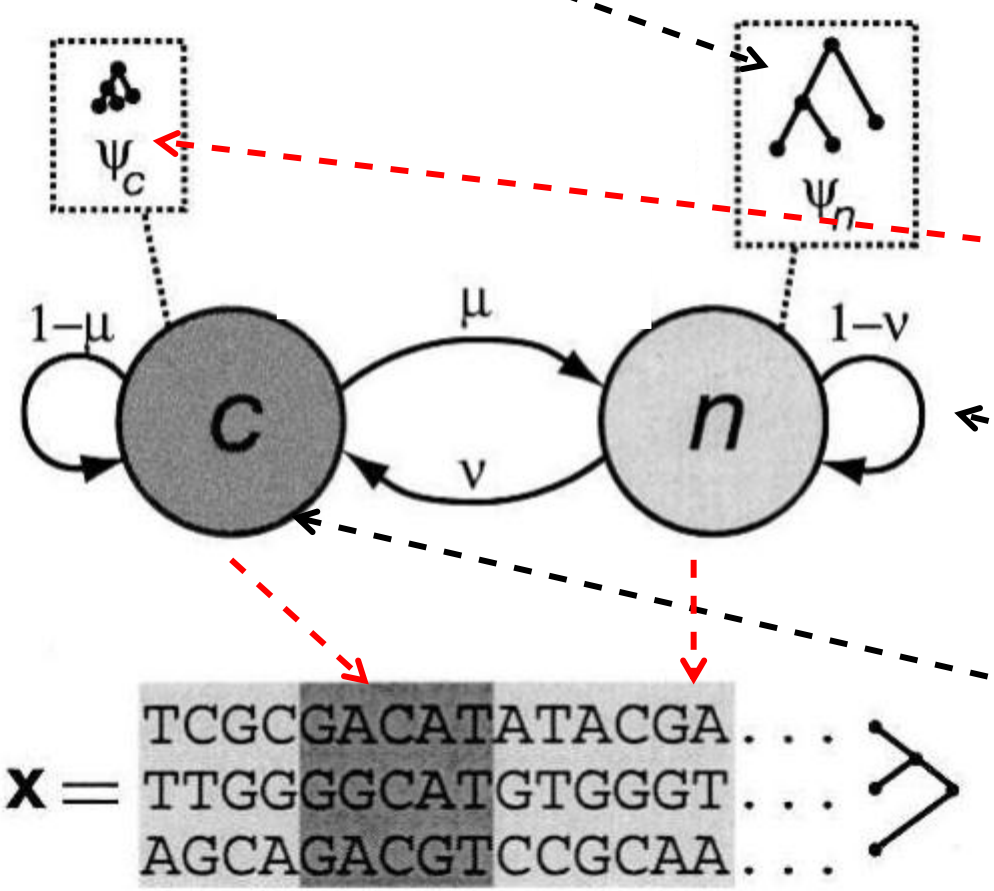
Parametri stimati

- ρ quante mutazioni in meno nelle regioni conservate
- v: probabilità di transizione dallo stato non-conservato a quello conservato
- μ: probabilità di transizione dallo stato conservato a quello non-conservato

Siepel et al., Genome Research, 2005

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma

Forma dell'albero filogenetico ψ



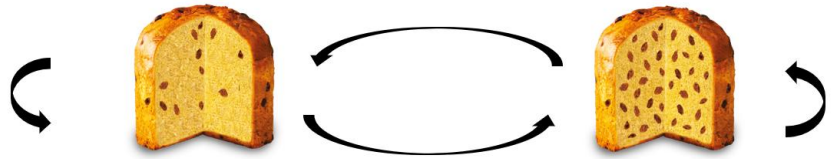
Parametri stimati

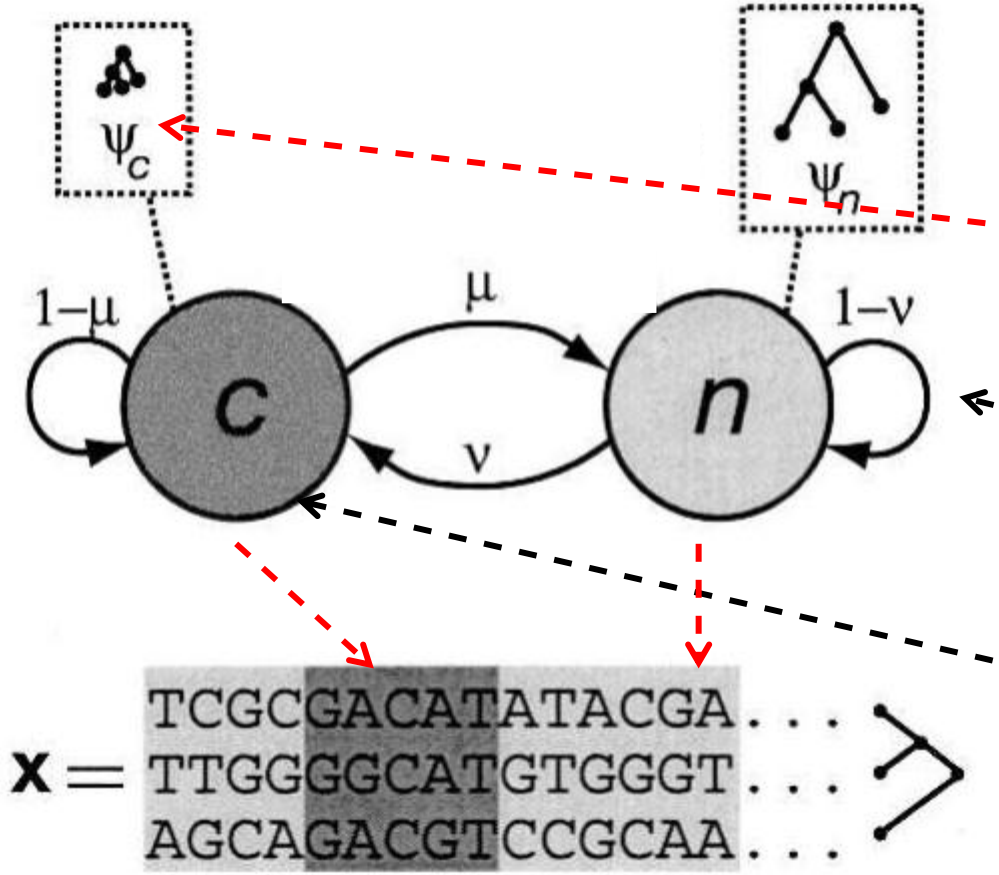
- ρ quante mutazioni in meno nelle regioni conservate
- v: probabilità di transizione dallo stato non-conservato a quello conservato
- μ: probabilità di transizione dallo stato conservato a quello non-conservato

Emissioni sono date da un **modello filogenetico**: Essenzialmente una combinazione (somma e prodotto) di varie **Poisson** per rappresentare il numero di mutazioni osservato ad ogni posizione della sequenza e lungo ogni ramo dell'albero filogenetico

Siepel et al., Genome Research, 2005

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma





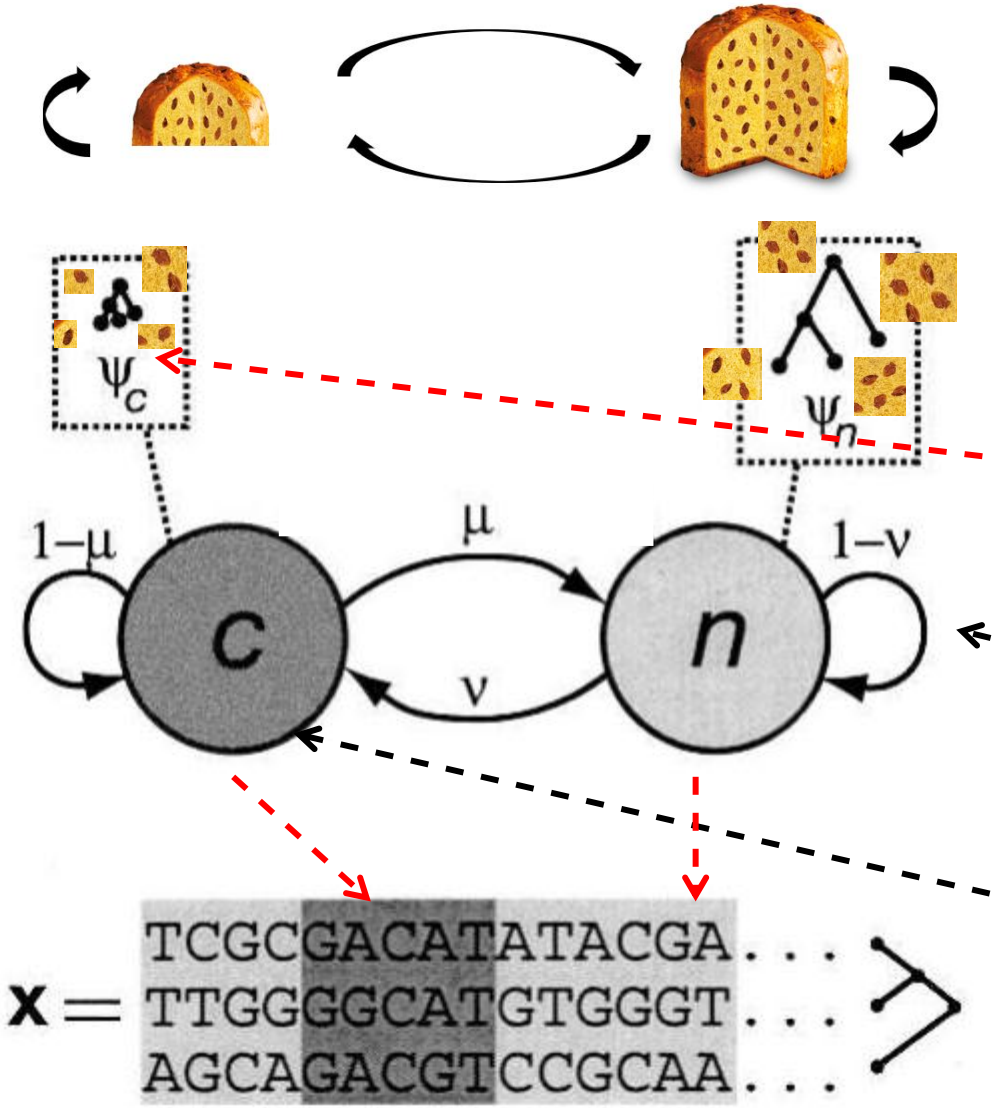
Parametri stimati

- ρ quante mutazioni in meno nelle regioni conservate
- v: probabilità di transizione dallo stato non-conservato a quello conservato
- μ: probabilità di transizione dallo stato conservato a quello non-conservato

Emissioni sono date da un **modello filogenetico**: Essenzialmente una combinazione (somma e prodotto) di varie **Poisson** per rappresentare il numero di mutazioni osservato ad ogni posizione della sequenza e lungo ogni ramo dell'albero filogenetico

Siepel et al., Genome Research, 2005

PhastCons: Un Hidden Markov Model per trovare regioni conservate nel genoma



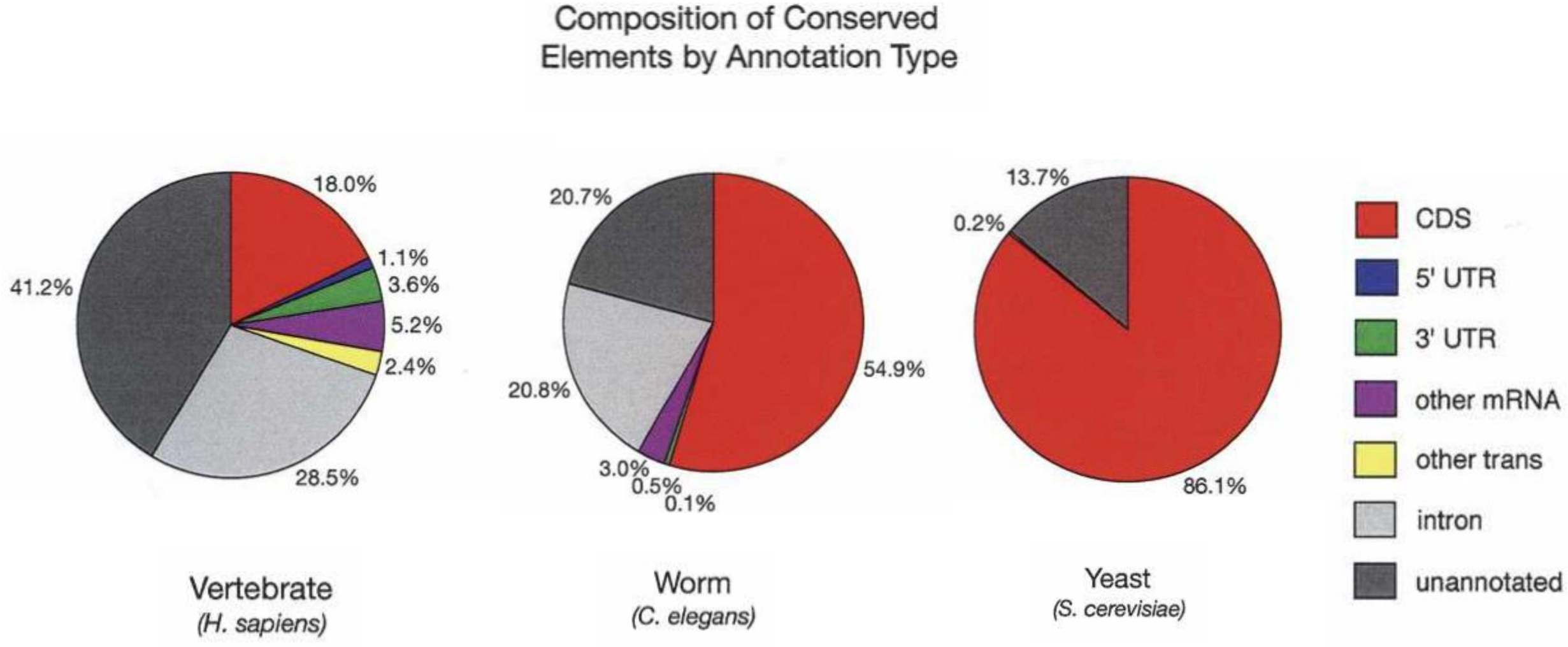
Parametri stimati

- ρ quante mutazioni in meno nelle regioni conservate
- v: probabilità di transizione dallo stato non-conservato a quello conservato
- μ: probabilità di transizione dallo stato conservato a quello non-conservato

Emissioni sono date da un **modello filogenetico**: Essenzialmente una combinazione (somma e prodotto) di varie **Poisson** per rappresentare il numero di mutazioni osservato ad ogni posizione della sequenza e lungo ogni ramo dell'albero filogenetico

Siepel et al., Genome Research, 2005

Le regioni conservate nel genoma umano non si trovano solo nelle regioni codificanti per proteine



Siepel et al., Genome Research, 2005

The perplexing figure behind a crucial virus database p. 332

Regulatory reforms to advance psychedelic therapies p. 347

A compact galaxy in the early Universe p. 416

Science

$15
28 APRIL 2023
SPECIAL ISSUE
science.org
AAAS

CRYPTOPROCTA FEROX

DAUBENTONIA MADAGASCARIENSIS

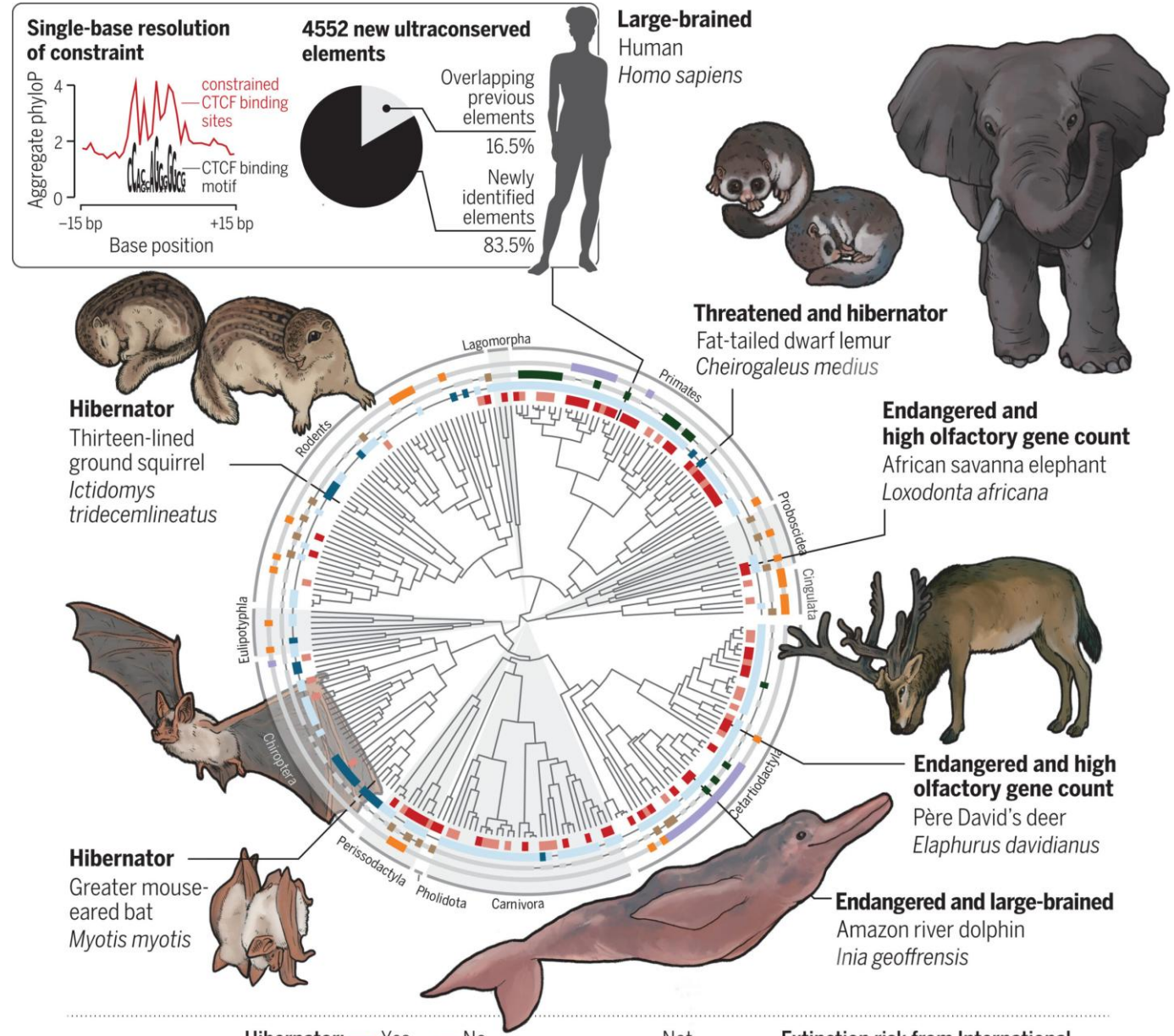
CHOLOEPUS HOFFMANNI

PHATAGINUS TRICUSPIS

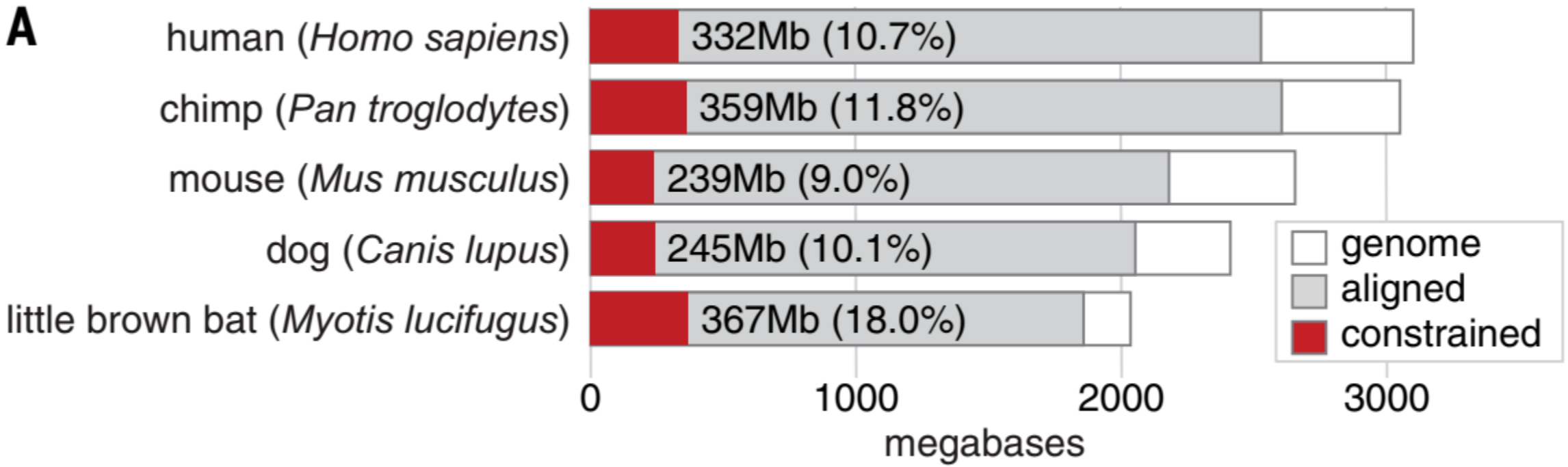
ECHINOPS TELFAIRI

ZOONOMIA

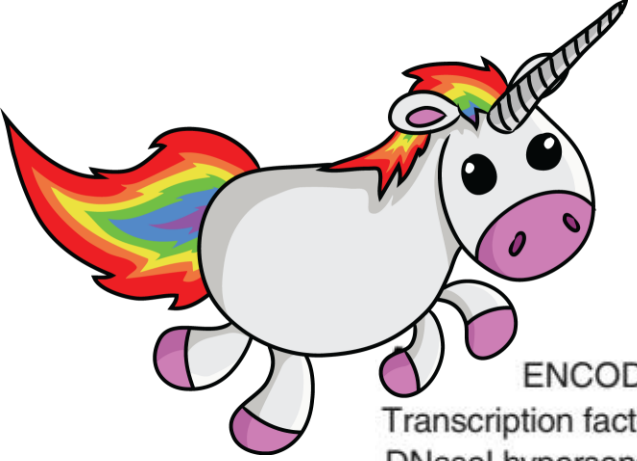
Diverse genomes reveal mammalian secrets p. 356



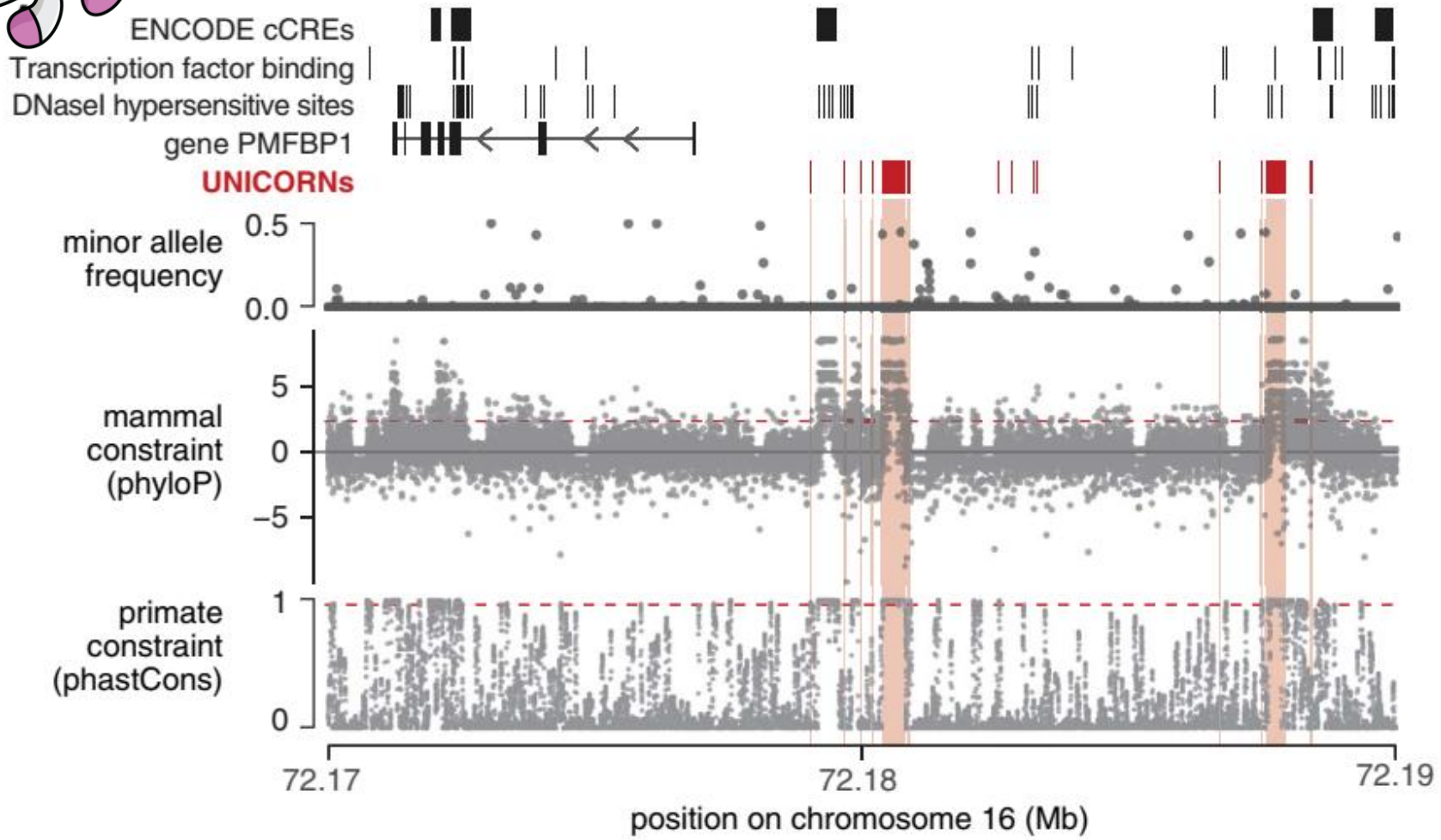
Modelli probabilistici con approcci simili (PhyloP) possono essere applicati a centinaia di genomi



Christmas et al.,Science, 2023

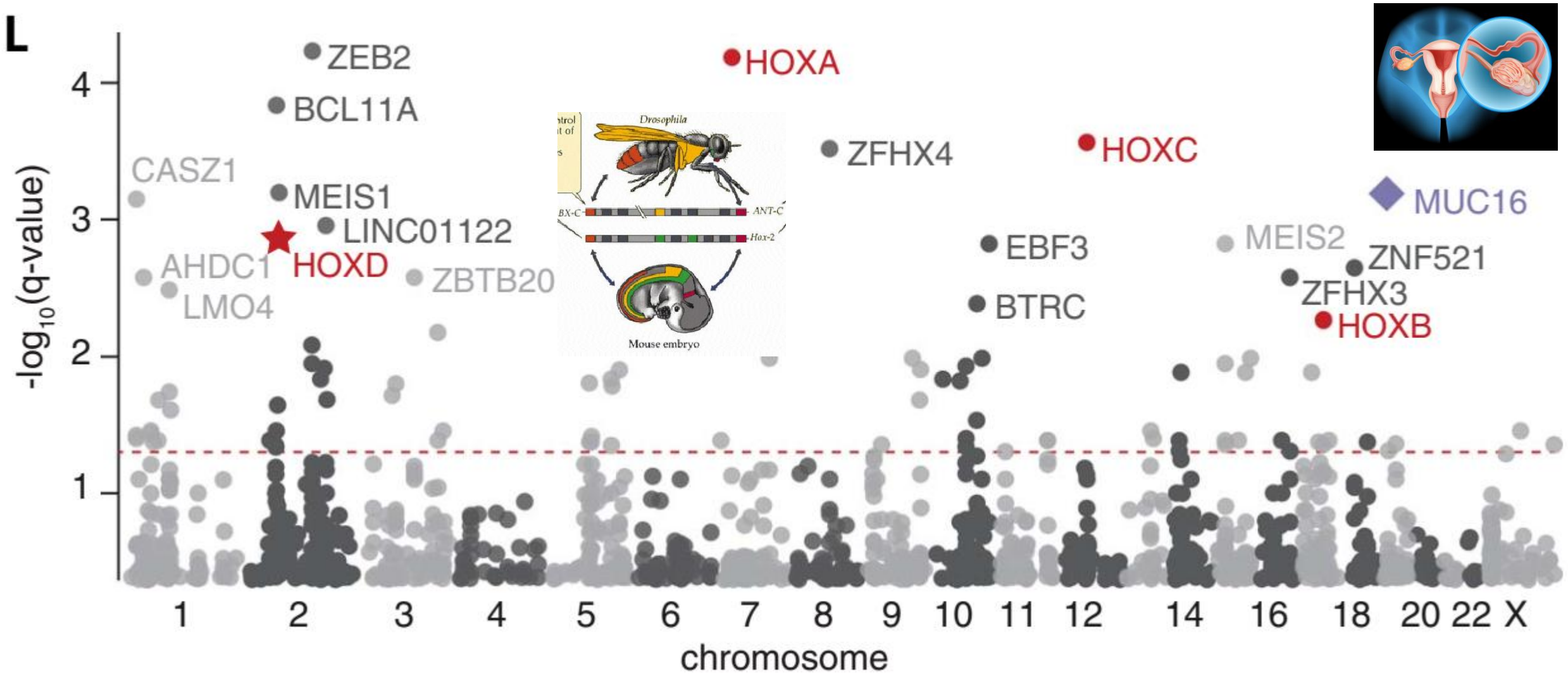


Questi modelli possono trovare UNICORNs (unannotated conserved regions)



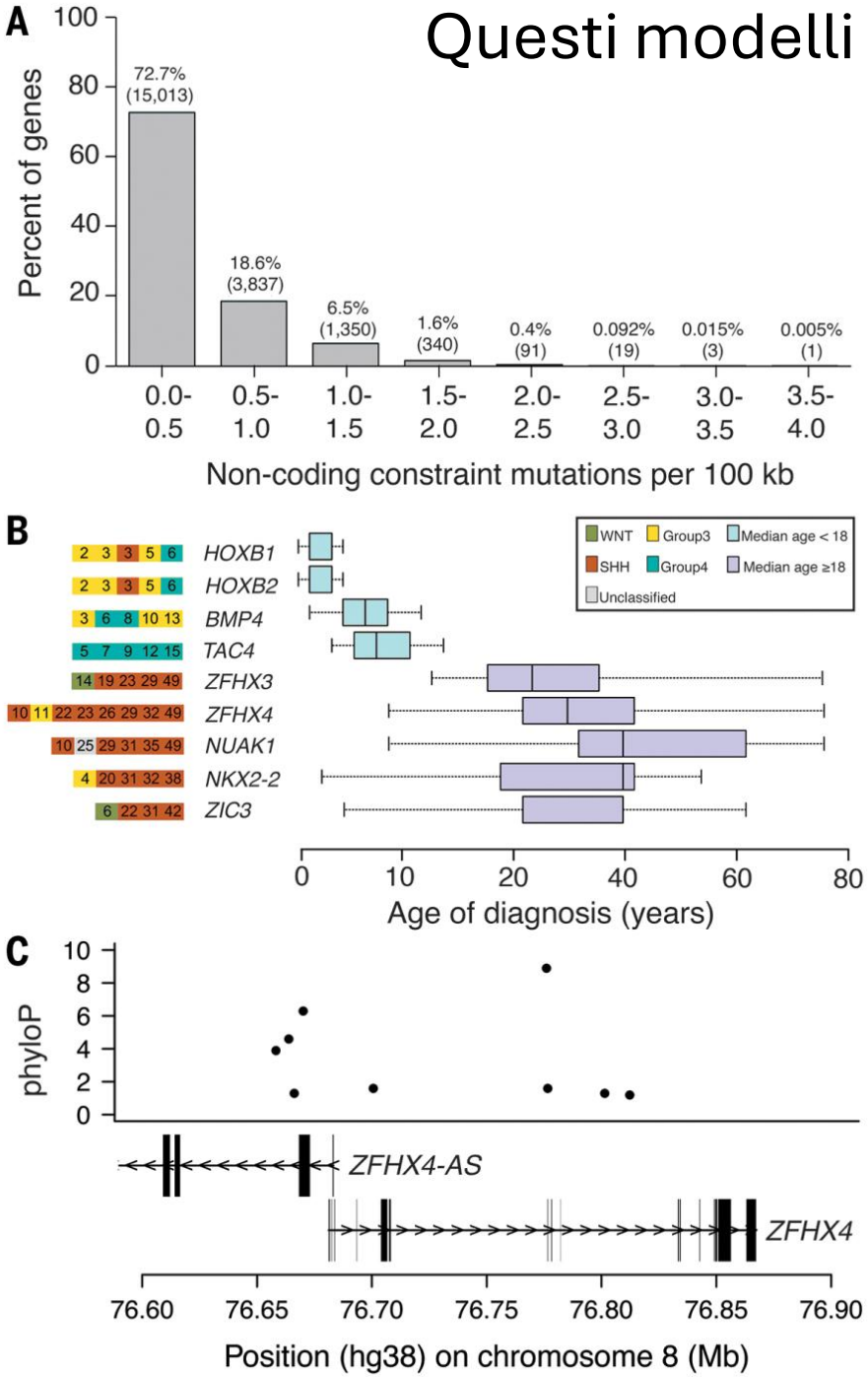
Christmas et al.,Science, 2023

Questi modelli possono trovare geni ultraconservati

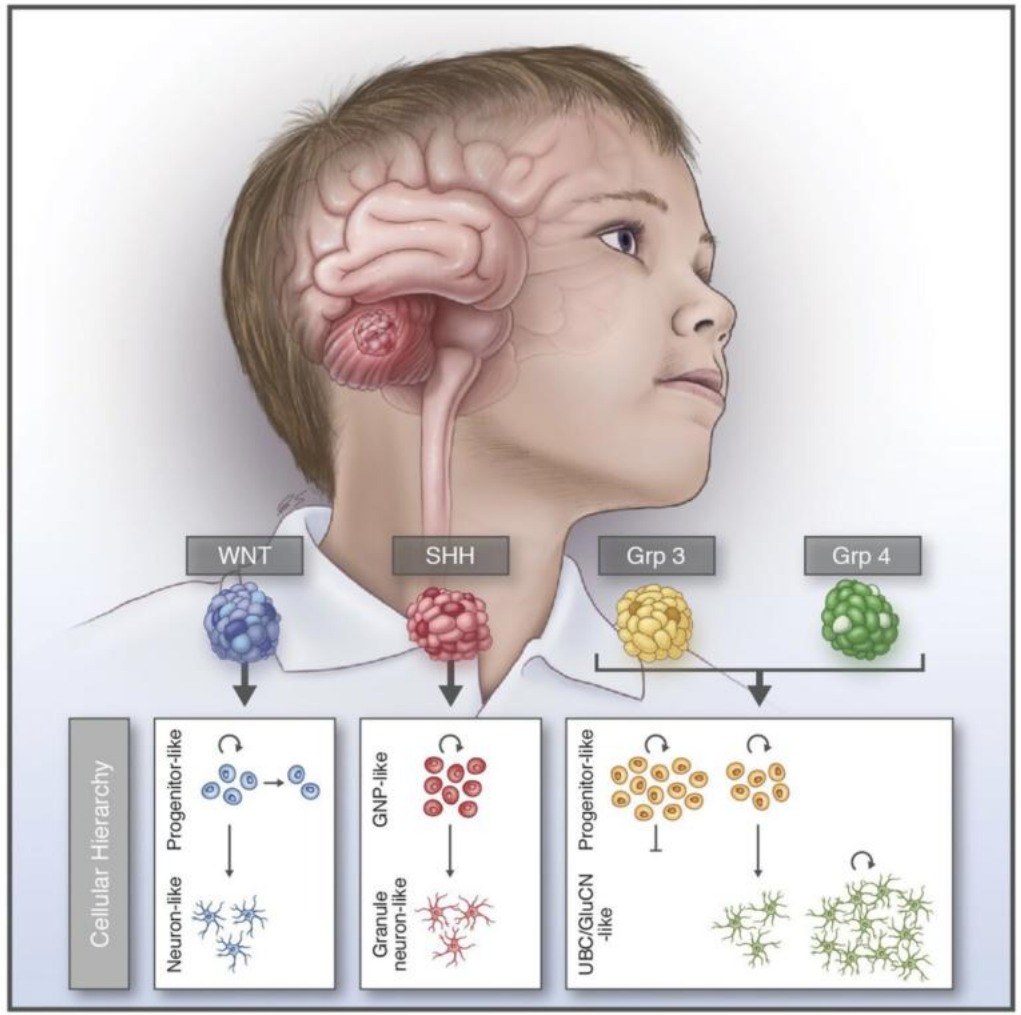


Christmas et al.,Science, 2023

Questi modelli possono trovare gene drivers di tumori



Medulloblastoma



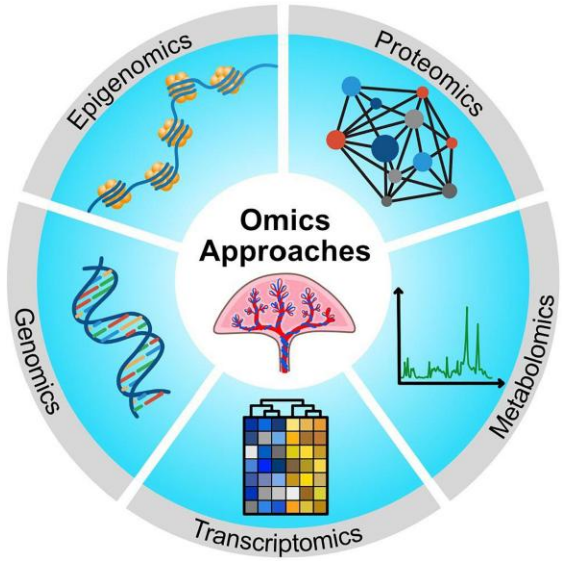
Sullivan et al.,Science, 2023

Machine learning nella predizione degli effetti delle mutazioni:

Sommario:

- Hidden Markov Models e modelli probabilistici
  - Phastcons
  - PhyloP
- Support Vector Machines
  - CADD (Kircher et al., Nat Gen, 2014)
- Feed Forward Neural Networks
  - DANN (Quang et al.,Bioinf.,2014)
- Language models
  - Gpn
- Prospettive future e riflessioni generali

Multi-Omics Integration



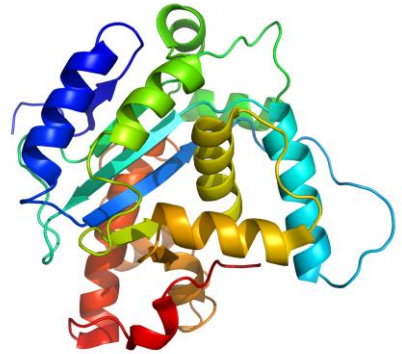
Genomics

- GC content
- Conservation scores
- …

Genomica

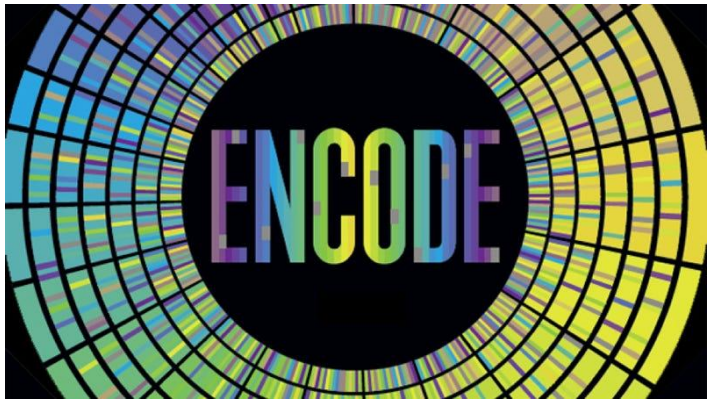
- GC content
- Conservation scores
- …

Strutture proteiche



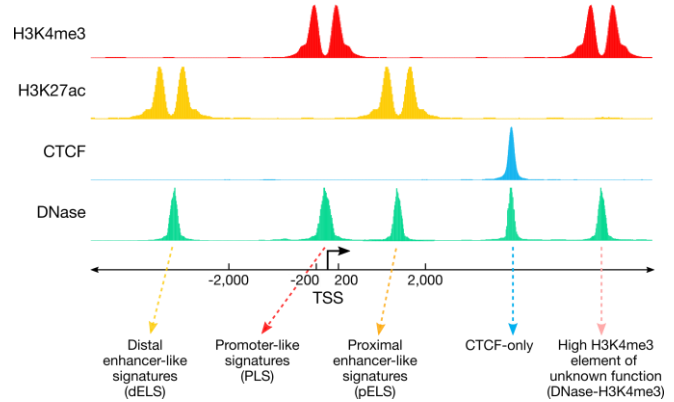
- SIFT scores
- Missense/synonymous
- …

Trascrittomica



- Gene expression in different tissues
- ATAC-seq
- …

Epigenomica

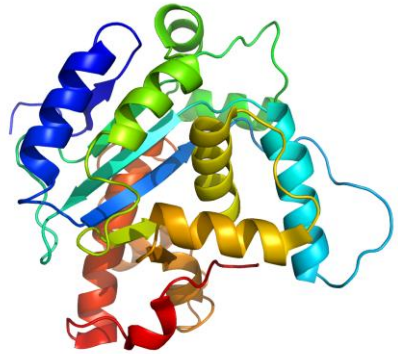


- Histone marks
- DNA methylation
- …

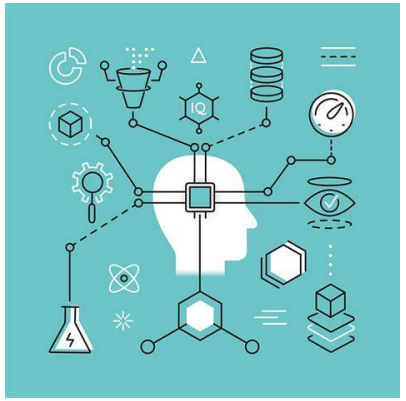
Genomica

- GC content
- Conservation scores
- …

Strutture proteiche

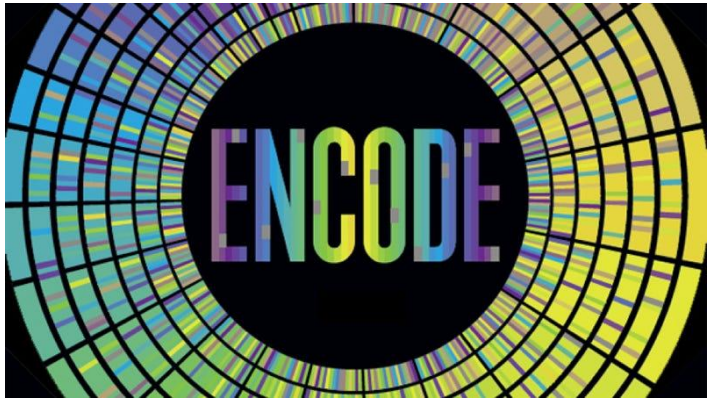


- SIFT scores
- Missense/synonymous
- …



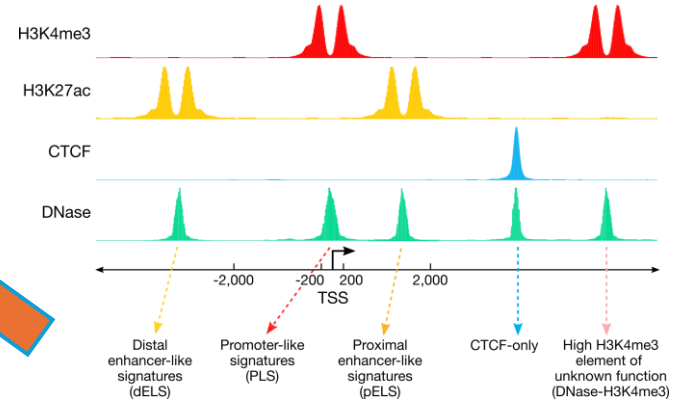
Machine learning model

Trascrittomica



- Gene expression in different tissues
- ATAC-seq
- …

Epigenomica

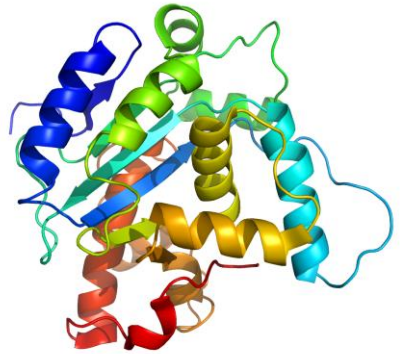


- Histone marks
- DNA methylation
- …

Genomica

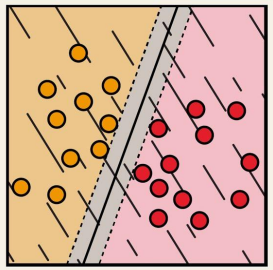
- GC content
- Conservation scores
- …

Strutture proteiche

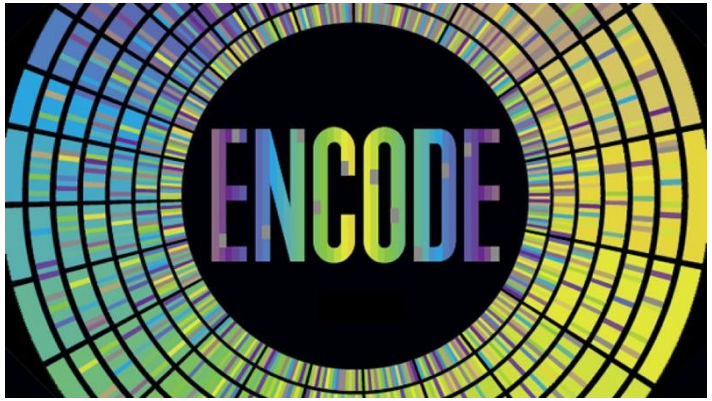


- SIFT scores
- Missense/synonymous
- …

Support Vector Machines

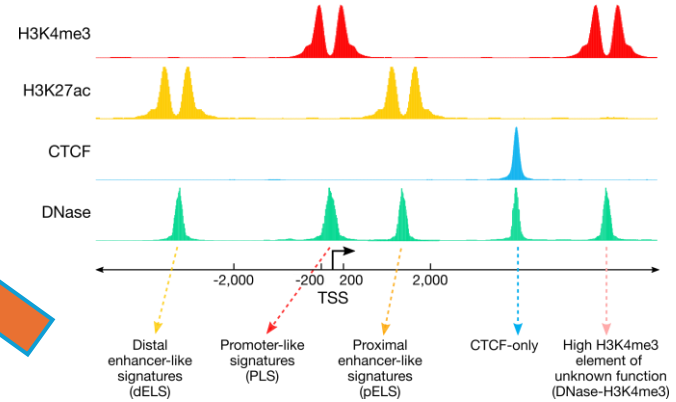


Trascrittomica



- Gene expression in different tissues
- ATAC-seq
- …

Epigenomica

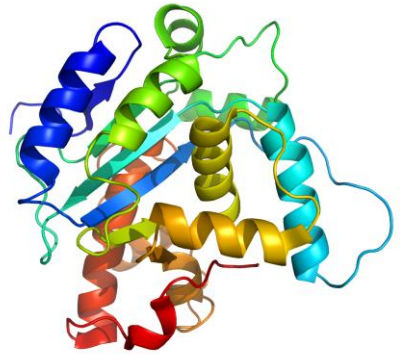


- Histone marks
- DNA methylation
- …

Genomica

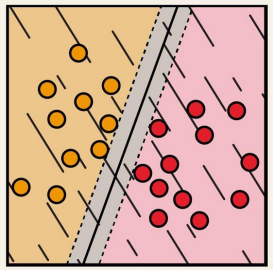
- GC content
- Conservation scores
- …

Strutture proteiche

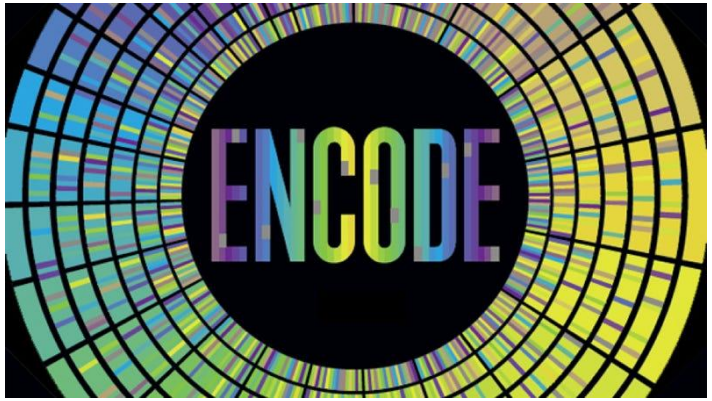


- SIFT scores
- Missense/synonymous
- …

Support Vector Machines

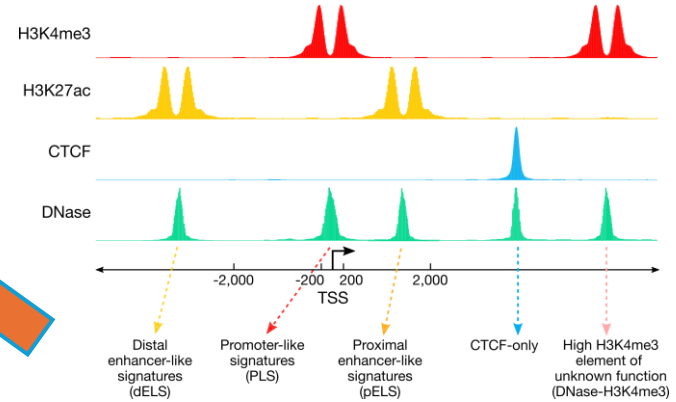


Trascrittomica



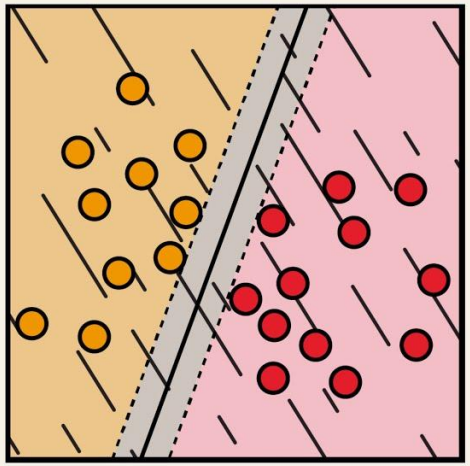
- Gene expression in different tissues
- ATAC-seq
- …

Epigenomica



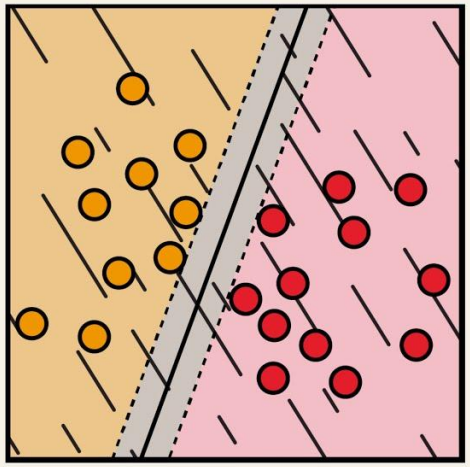
- Histone marks
- DNA methylation
- …

Support Vector Machines



- Le Support Vector Machines (SVMs) separano/classificano i dati combinando molti predittori (ad esempio genomica, transcriptomica)
- Simili concettualmente ad una regressione logistica multipla

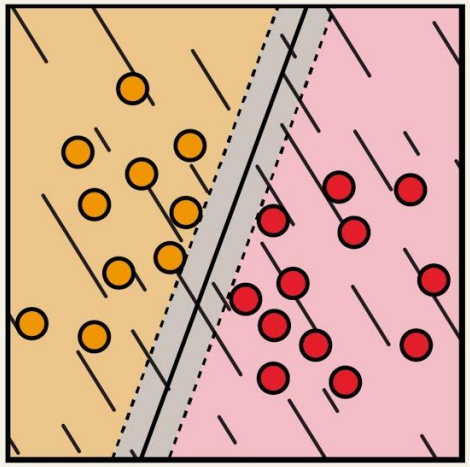
Support Vector Machines



- Le Support Vector Machines (SVMs) separano/classificano i dati combinando molti predittori (ad esempio genomica, transcriptomica)
- Sono una forma di supervised learning, quindi necessitano dei dati labelled (un set di dati “positivo” ed uno “negativo”, per imparare come distinguerli).

Proxy-neutrali/non patogeniche ? ? Proxy-deleterie/patogeniche

Support Vector Machines



- Le Support Vector Machines (SVMs) separano/classificano i dati combinando molti predittori (ad esempio genomica, transcriptomica)
- Sono una forma di supervised learning, quindi necessitano dei dati labelled (un set di dati “positivo” ed uno “negativo”, per imparare come distinguerli).

Proxy-neutrali/non patogeniche

diverse da

Proxy-deleterie/patogeniche

A → C

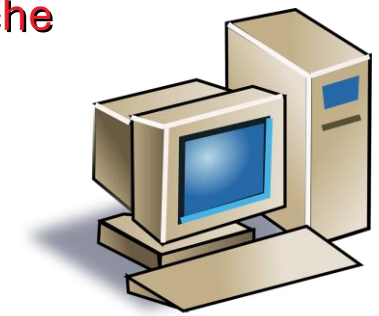
T

C

T → A

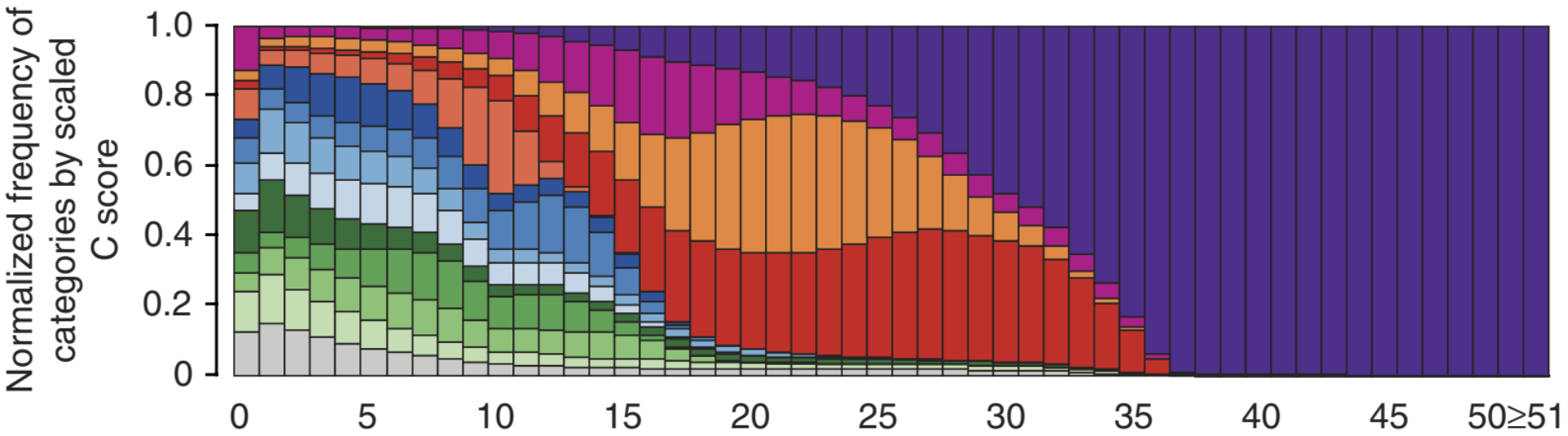
G

A



mutazioni simulate *de novo*

SVMs predicono effetti biologicamente plausibili per le mutazioni



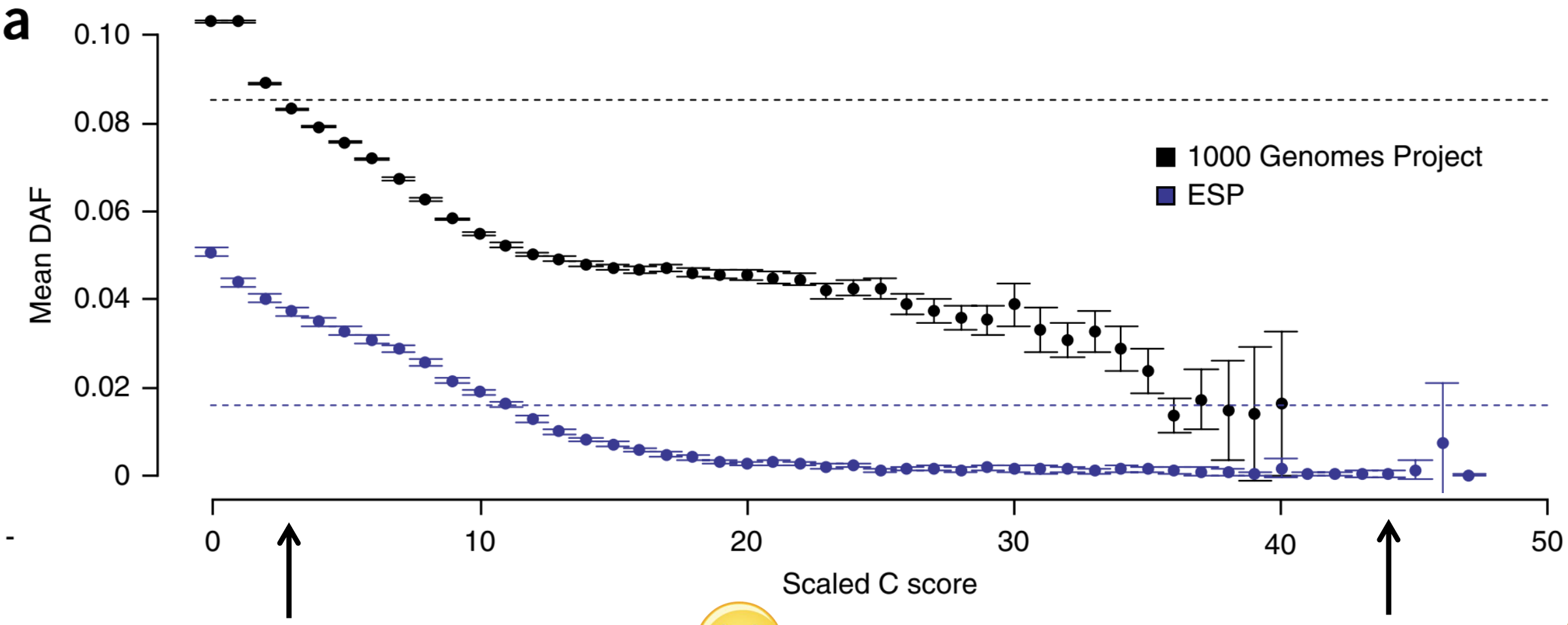
Neutrali

Molto deleterie

- Stop loss (11; 0–43)
- Stop gain (37; 0–99)
- Canonical splice (15; 0–37)
- Nonsynonymous (15; 0–39)
- Synonymous (7; 0–27)
- Noncoding (4; 0–35)
- Splice site (7; 0–35)
- Intronic (3; 0–39)
- Regulatory (5; 0–37)
- Downstream (3; 0–38)
- 3′ UTR (6; 0–32)
- 5′ UTR (5; 0–34)
- Upstream (3; 0–39)
- Intergenic (2; 0–39)

Kircher et al.,Nat Gen, 2014

La deleteriosità delle mutazioni predice la loro frequenza nelle popolazioni umane

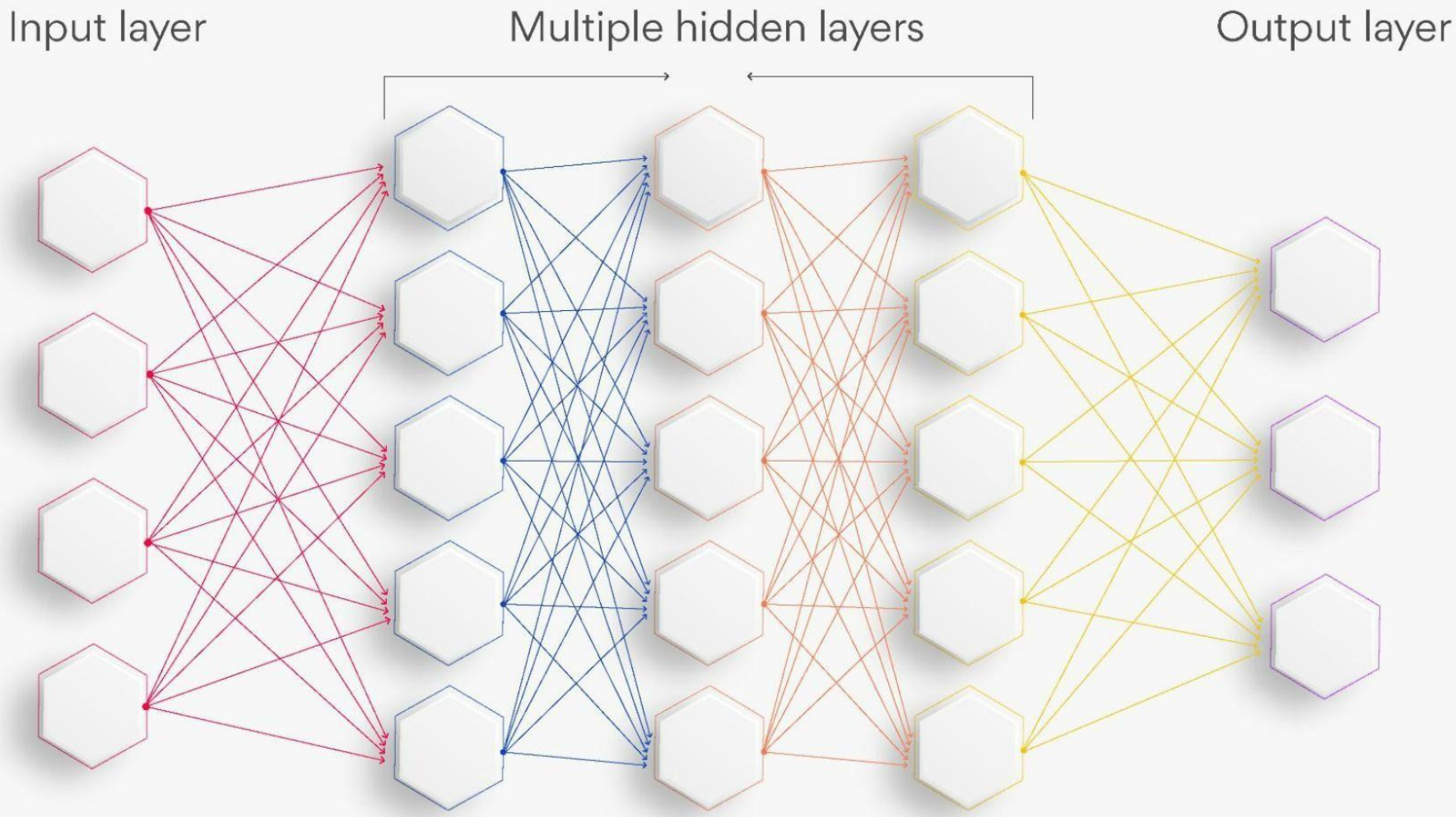


Varianti neutrali ad alta frequenza

Varianti deleterie rare

Kircher et al.,Nat Gen, 2014

Deep Neural Network

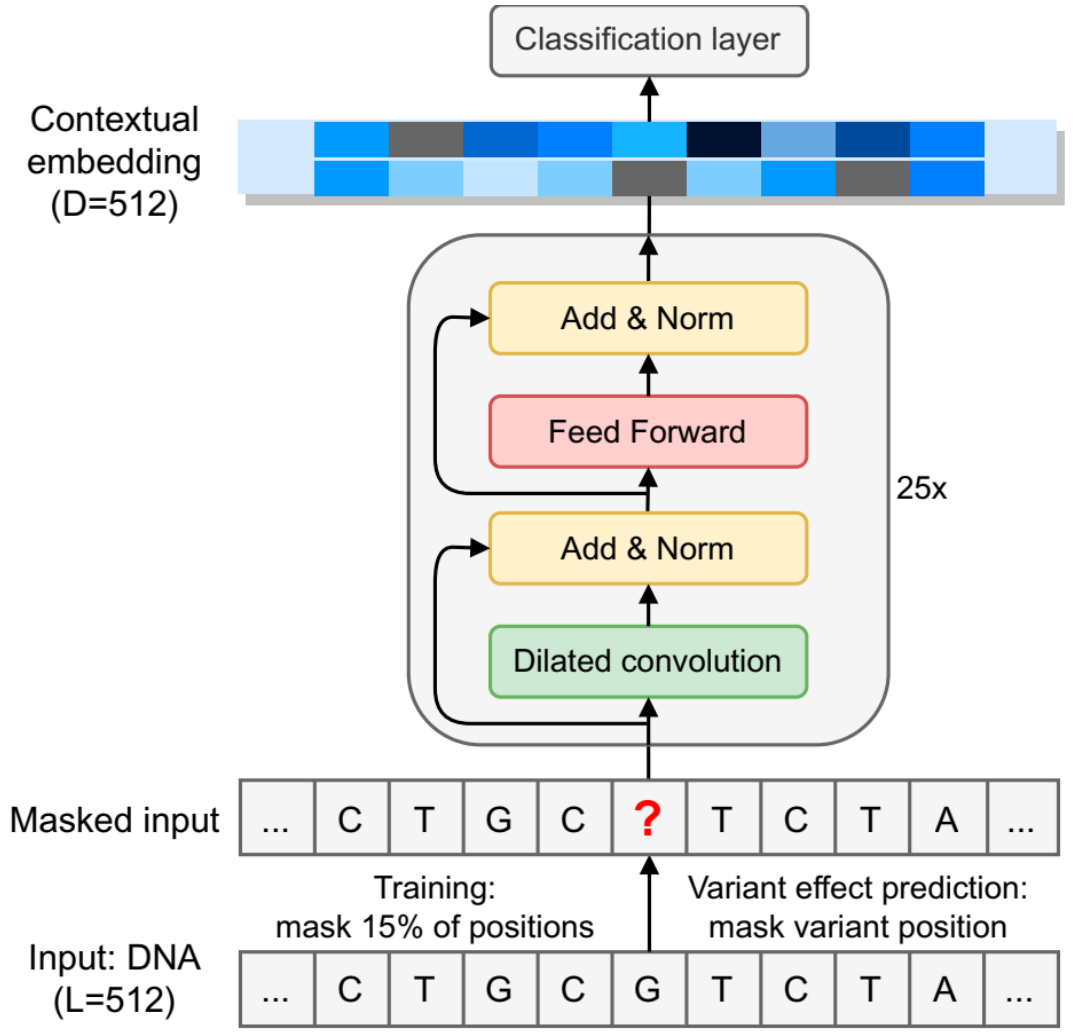


Turing

ChatGPT

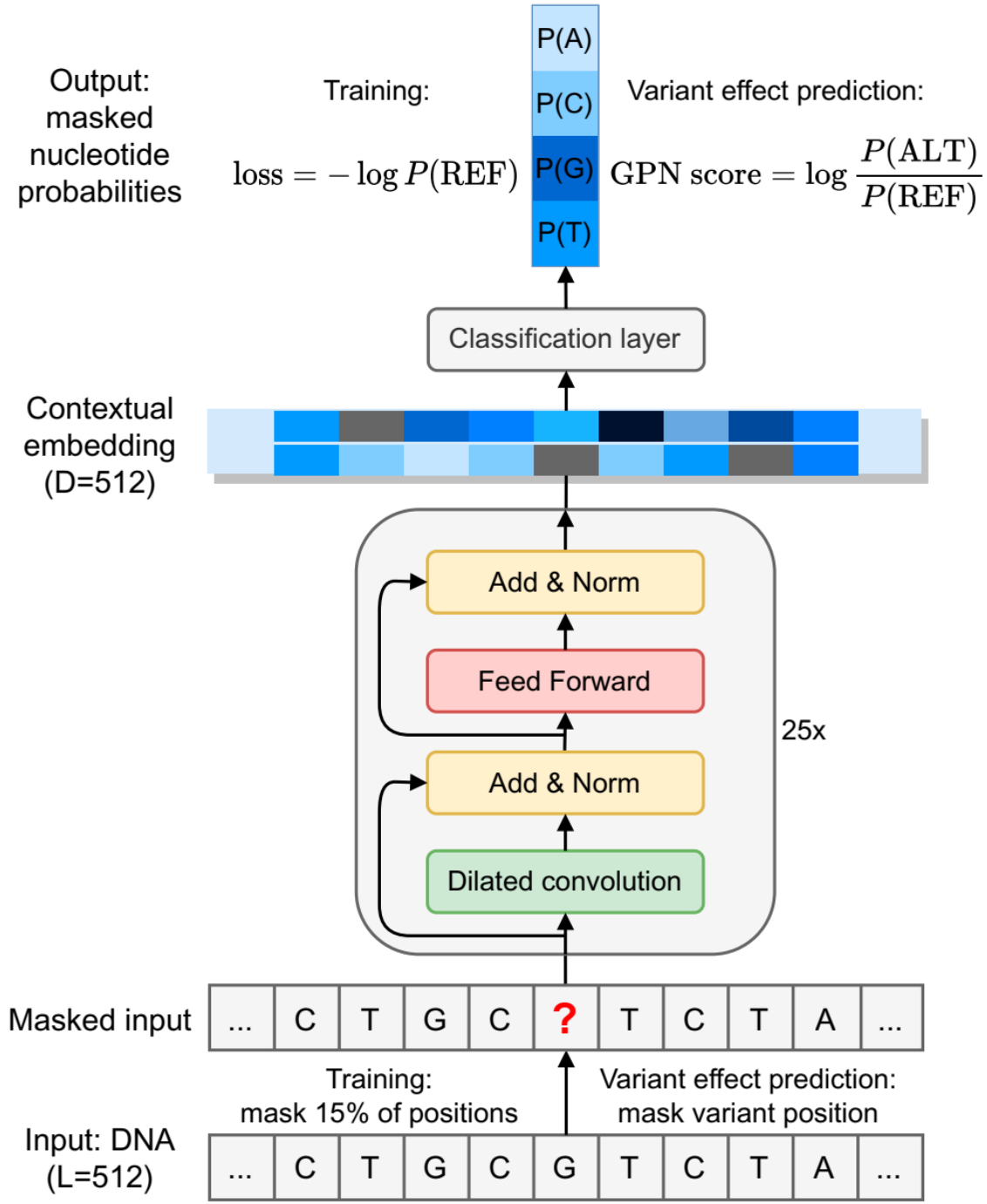
Large
Language
Models

Bard



Benegas et al.,PNAS, 2023

Arabidopsis thaliana



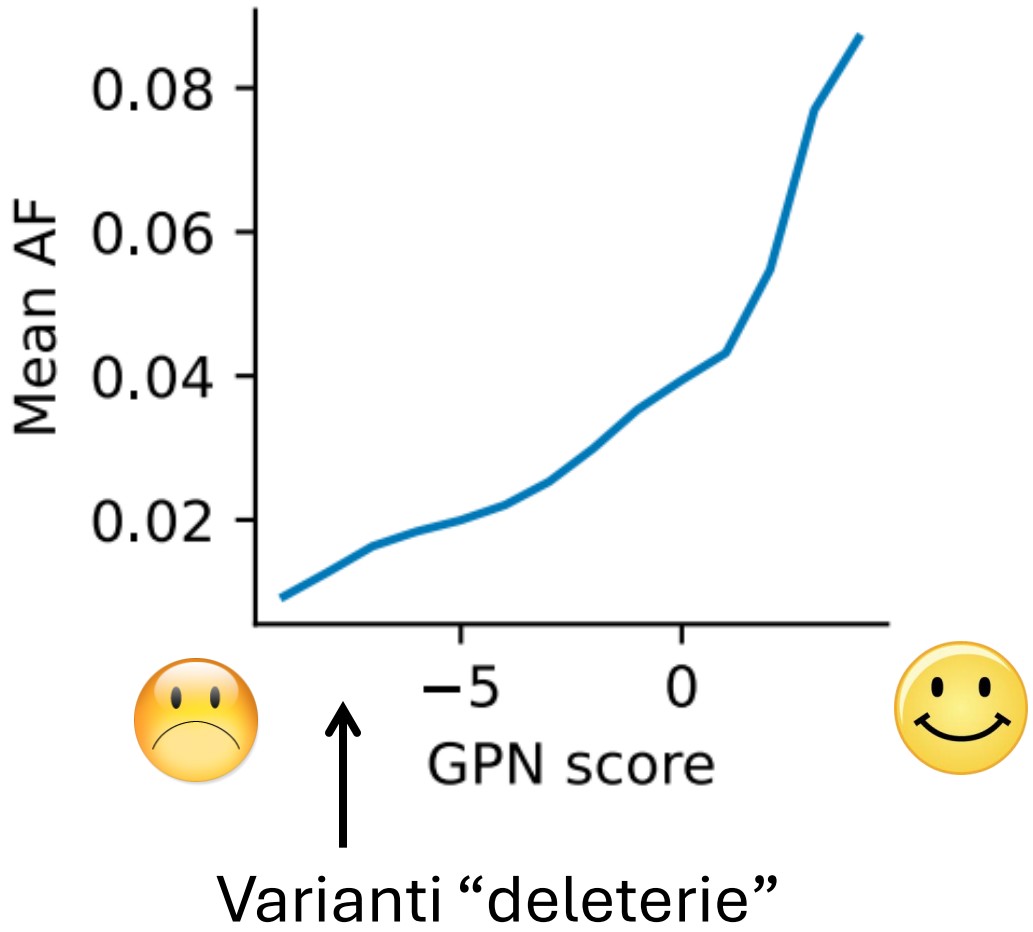
Benegas et al.,PNAS, 2023

Valore di 'Deleteriosità':
Le posizioni per le quali viene predetto con confidenza un nucleotide, ma ne viene osservato un altro (molto inatteso), sono predette come deleterie.

Piu' confidenza nella referenza REF, e piu' inatteso l'alternativo ALT, piu' è considerata deleteria la mutazione

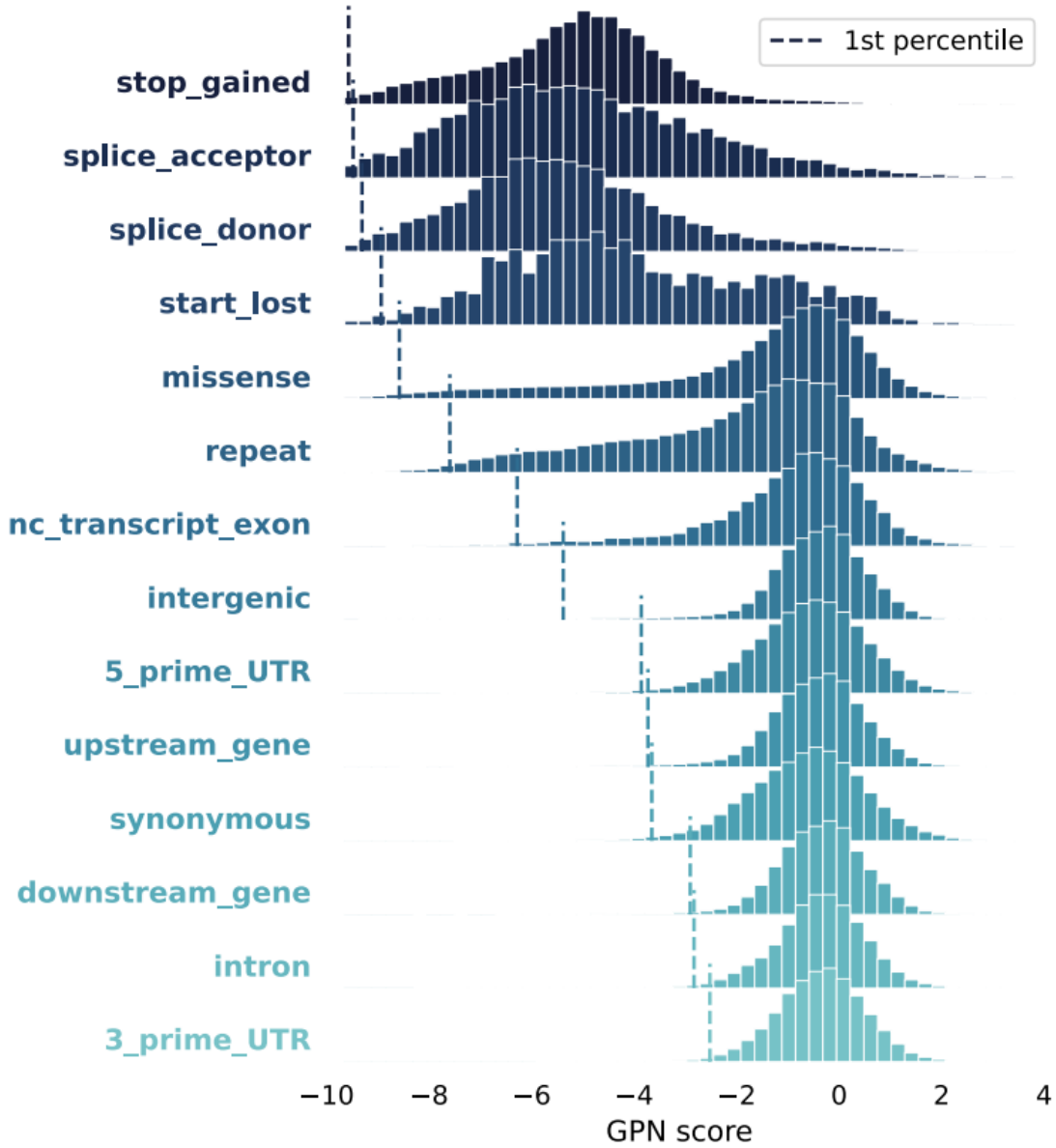
Arabidopsis thaliana

I DNA language models predicono la frequenza delle varianti



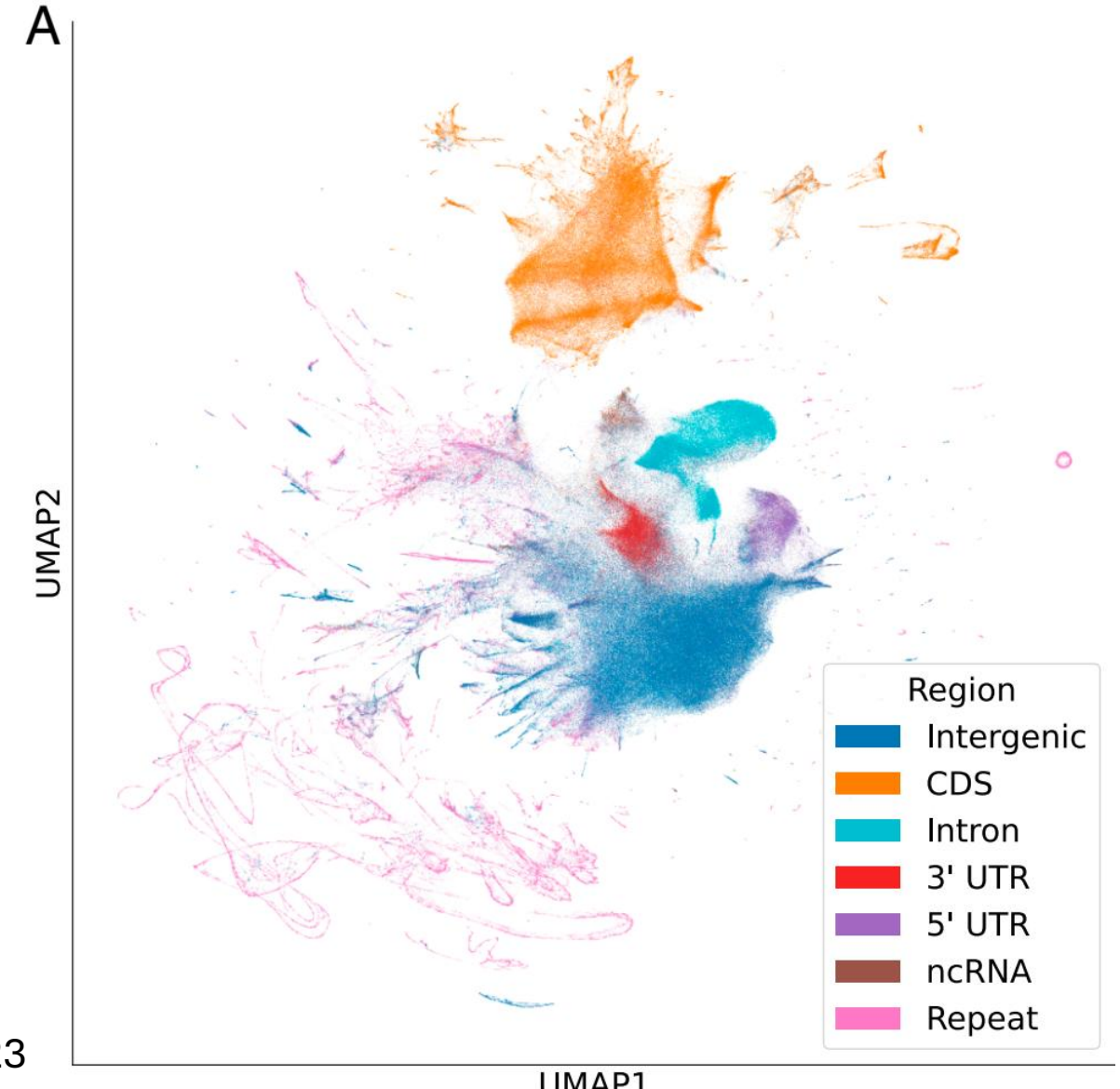
Benegas et al.,PNAS, 2023

I DNA language models predicono effetti biologicamente plausibili per le mutazioni



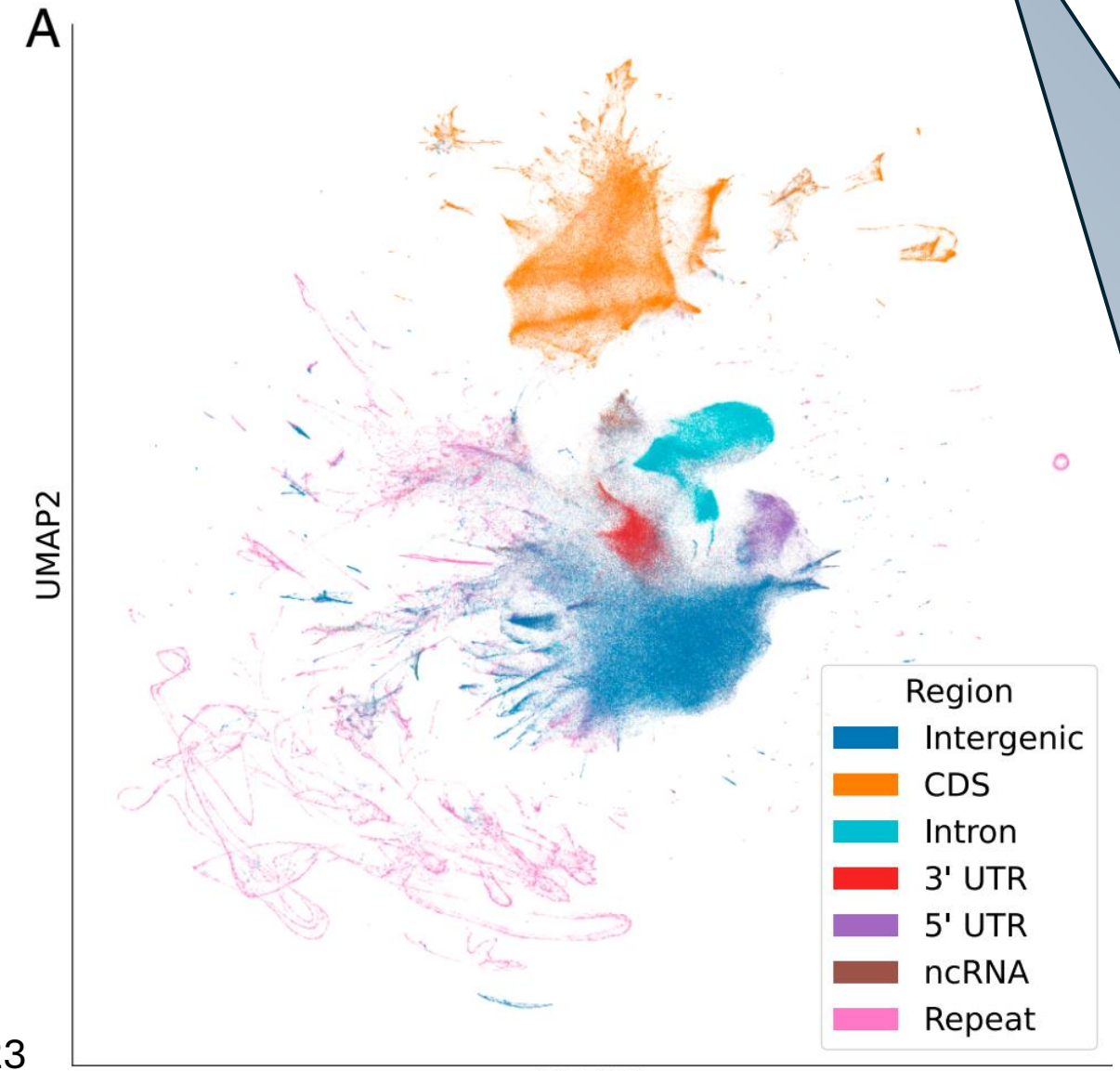
Benegas et al.,PNAS, 2023

I DNA language models possono classificare/“annotare” i genomi senza supervisione



Benegas et al.,PNAS, 2023

I DNA language models possono classificare/“annotare” i genomi senza supervisione



Benegas et al.,PNAS, 2023

Annotazione

L’annotazione genomica è il processo di identificazione degli elementi funzionali del genoma, come geni, regioni regolatorie e zone funzionali.

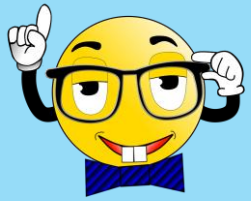
Le annotazioni per nuovi genomi vengono effettuate combinando molti tipi di dati (RNAseq, caratteristiche delle sequenze) e confrontando i genomi con altri genomi già annotati.

Pros and cons

Modelli probabilistici

Interpretabili e informativi

Teoria solida



Pochi dati richiesti

Rapidi

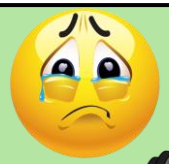
Poche risorse computazionali

Metodi ben caratterizzati

Meno flessibili

SVM/Deep learning

Difficili da interpretare



Poca teoria richiesta

(“facili” da implementare)

Data hungry

Non molto rapidi

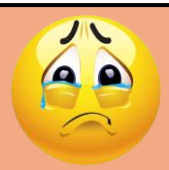
GPU è utile ma non necessaria

Bu librerie disponibili

Flessibili ma no informazioni meccanicistiche

Language models

Difficili da interpretare



oca teoria richiesta

ncora difficili da implementare)

Data hungry ma intelligenti

Lenti

Calcolatori necessari

Ancora nell’infanzia

Promettenti ma ancora una black box

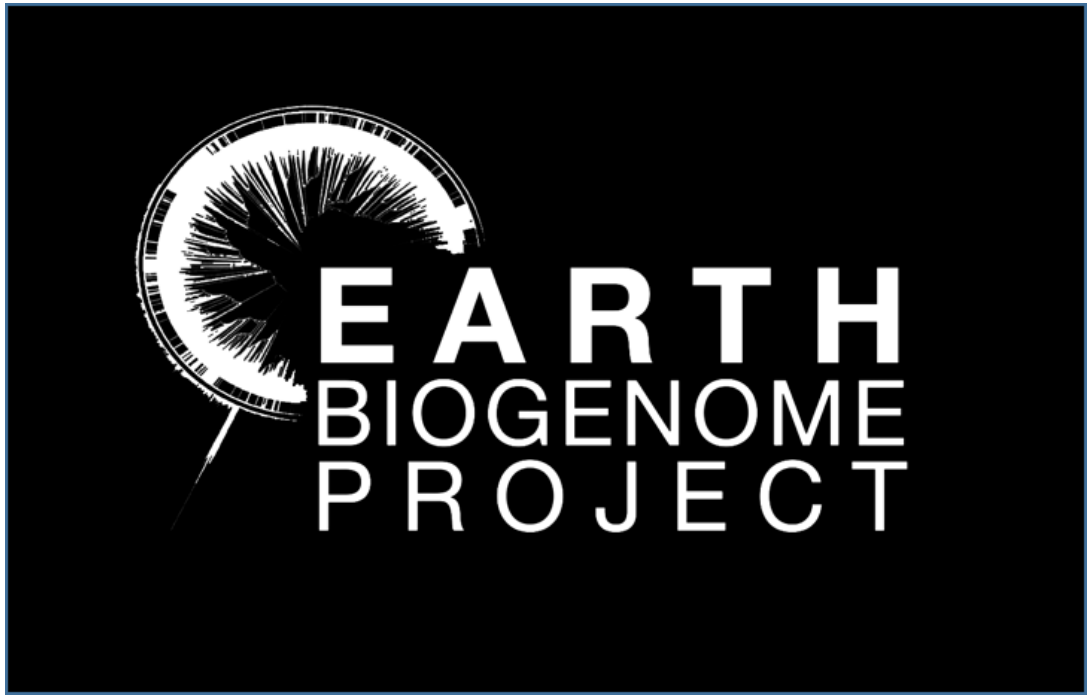
Prospettive future:

Ma non per tutto! Algoritmi deterministici e modelli probabilistici rimarranno sempre (probabilmente) piu’ interpretabili e piu’ efficienti per operazioni computazionalmente intense (mappaggio, assemblaggio, ..)

Anche gli approcci tradizionali hanno grande potere con “big data”

biobank uk
Improving the health of future generations

All of Us℠ | The Precision Medicine Initiative®
THE FUTURE OF HEALTH BEGINS WITH YOU



The perplexing figure behind a crucial virus database p. 332

Regulatory reforms to advance psychedelic therapies p. 347

A compact galaxy in the early Universe p. 416

Science

$15
28 APRIL 2023
SPECIAL ISSUE
science.org
AAAS

CRYPTOPROCTA FEROX

DAUBENTONIA MADAGASCARIENSIS

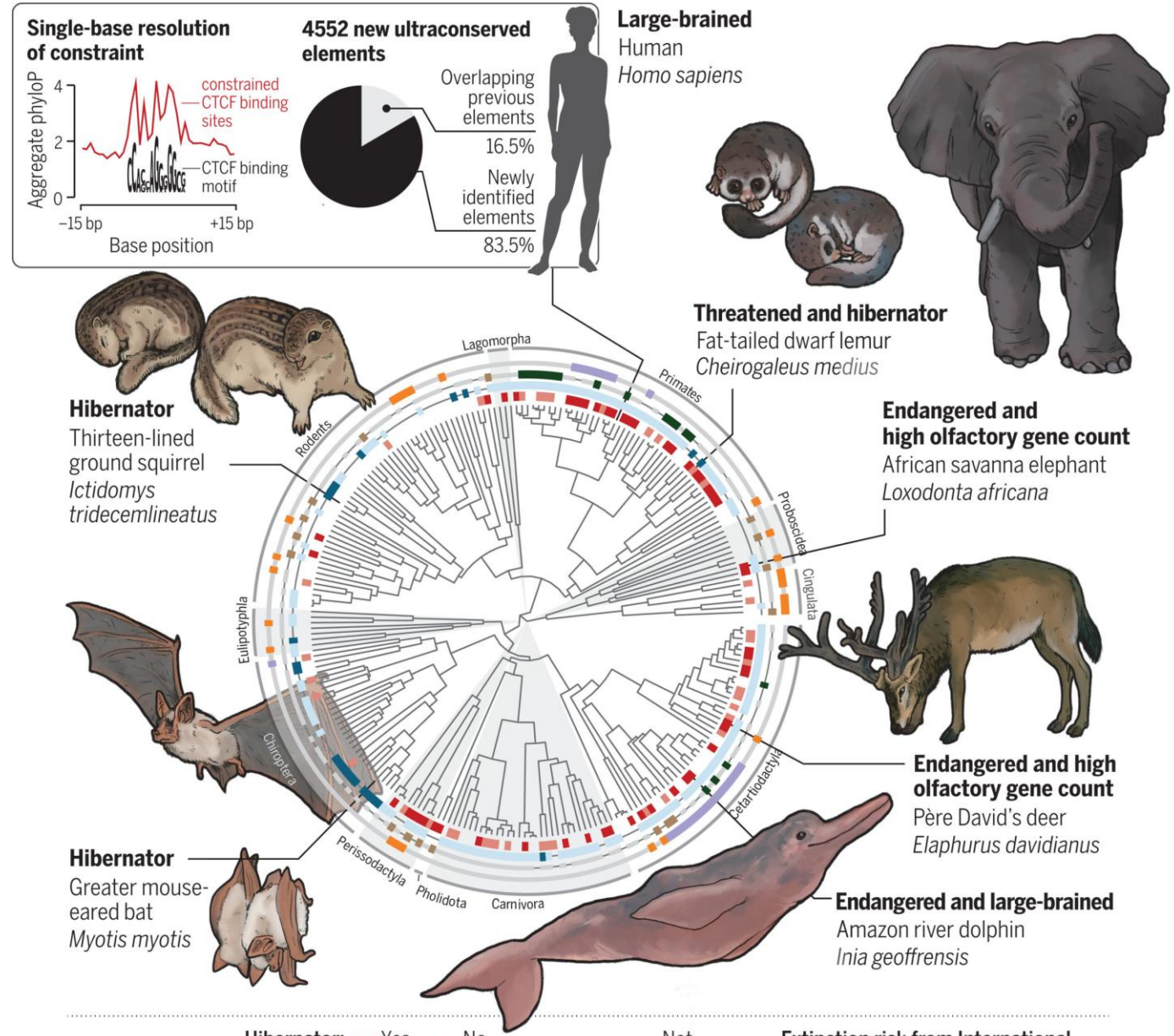
CHOLOEPUS HOFFMANNI

PHATAGINUS TRICUSPIS

ECHINOPS TELFAIRI

ZOONOMIA

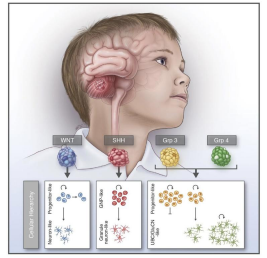
Diverse genomes reveal mammalian secrets p. 356



Conclusioni

Sia modelli probabilistici, machine learning tradizionale ed AI, in combinazione con molti dati (molti genomi e dati omici) possono/potranno:

- Migliorare la nostra conoscenza del genoma
- Informare studi clinici su tumori e varianti patogeniche
- Informare studi evolutive, biologia della conservazione e biodiversità



Domande?

Progetti

Argomenti per i progetti e conclusioni generali

Dal sequenziatore..

..al genoma..

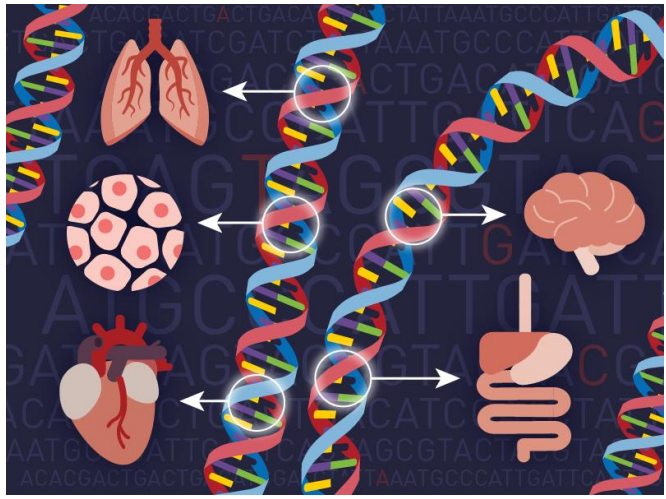
..alla funzione

sequenze

assemblaggio

mappatura
e
genotipizzazione

HMM



Modifiche epigenomiche
Ricerca di geni omologhi
Ricerca di promotori/Isole CpG
Stima di patogenicità
etc.

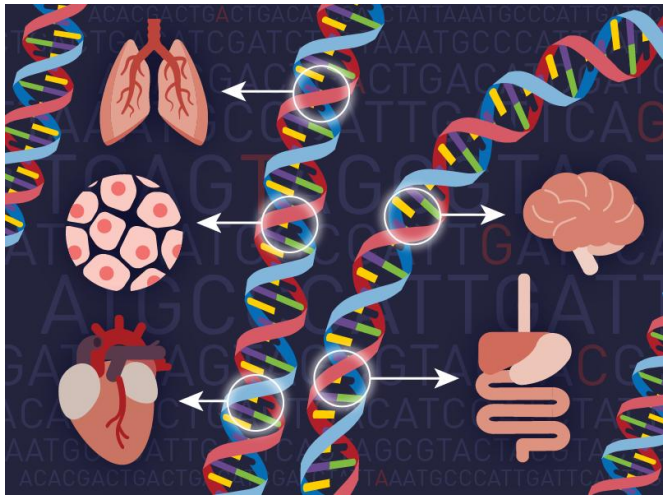
Argomenti per i progetti e conclusioni generali

Grafi di De Brujin

Grafi delle sovrapposizioni

assemblaggio

HMM



mappatura e genotipizzazione

minimizers vs kmers

Trasformata di Burrow-Wheeler

Competitive mapping

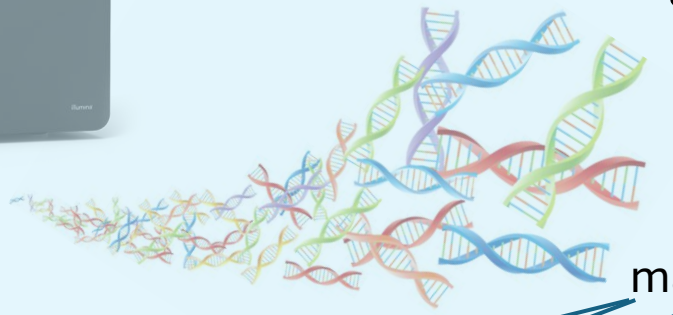
Algoritmi deterministici

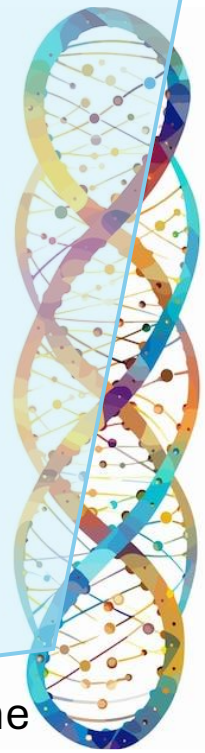
Teoria dei grafi

Grafi di De Brujin

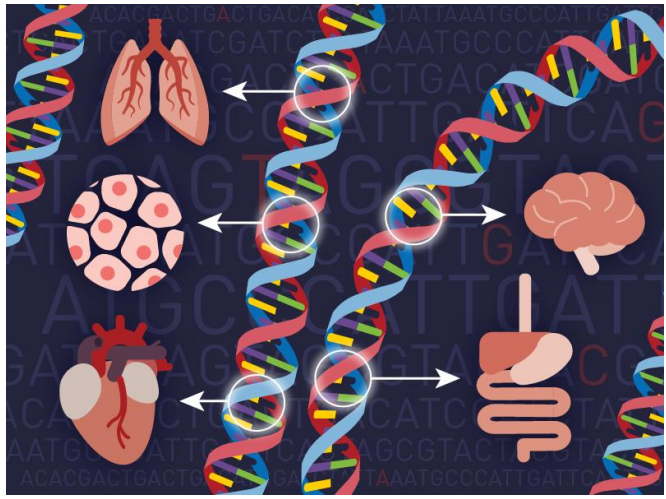
Grafi delle sovrapposizioni

assemblaggio





HMM



mappatura e genotipizzazione

minimizers vs kmers

Trasformata di Burrow-Wheeler

Competitive mapping

Teoria della compressione

Algoritmi deterministici

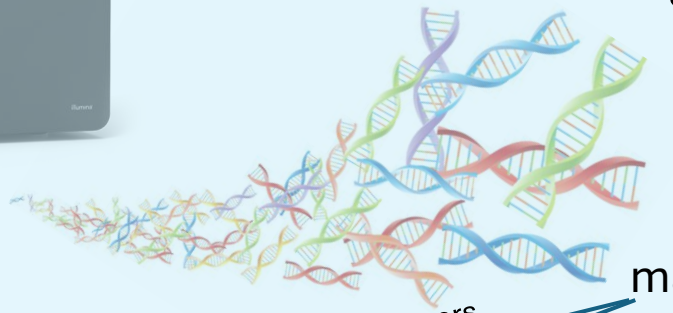
Modelli probabilistici

Teoria dei grafi

Grafi di De Brujin

Grafi delle sovrapposizioni

assemblaggio



mappatura

e

genotipizzazione

minimizers vs kmers

Trasformata di Burrow-Wheeler

Competitive mapping

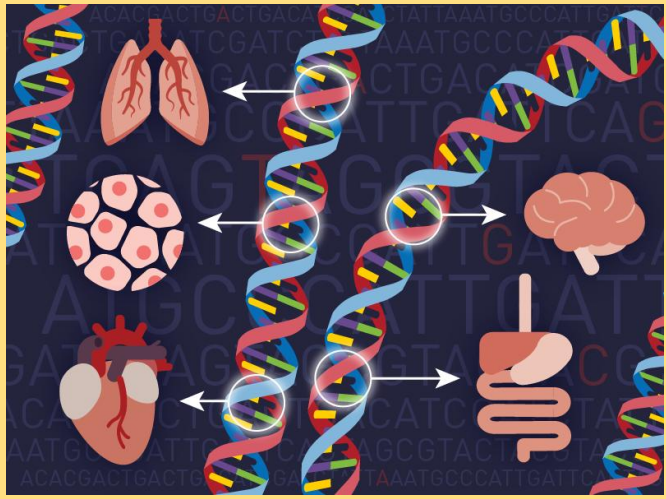
Teoria della compressione

Genotype likelihood

Baum-Welch

Algoritmo di Viterbi

HMM



Argomenti per i progetti e conclusioni generali

Dal sequenziatore..

..al genoma..

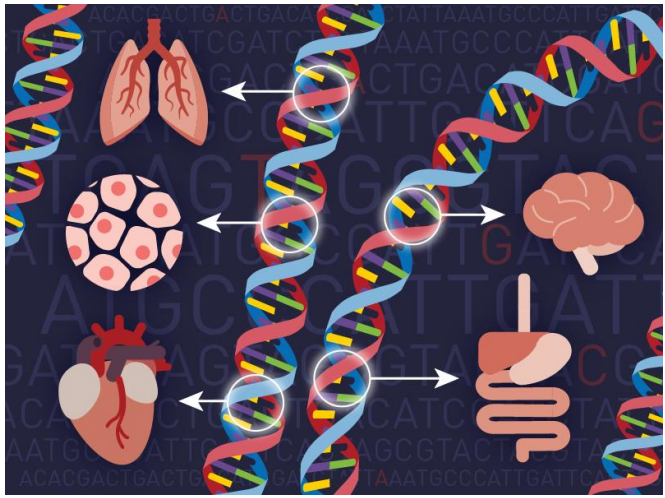
..alla funzione

sequenze

assemblaggio

mappatura
e
genotipizzazione

HMM



Formati di file bioinformatici

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

Argomenti per i progetti e conclusioni generali

Dal sequenziatore..

..al genoma..

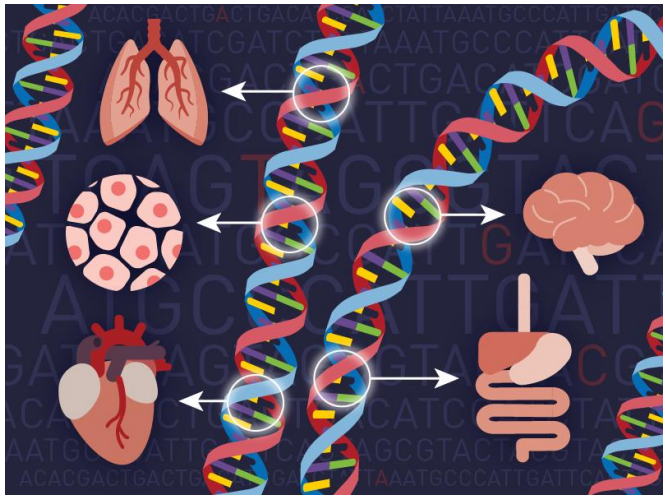
- Flye
- miniasm

- bowtie2
- bwa
- minimap2

- bcftools

- hmmratac
- CHROMHMM
- hmmer
- AuthentiCT
- HMMcopy
- Phastcons

..alla funzione

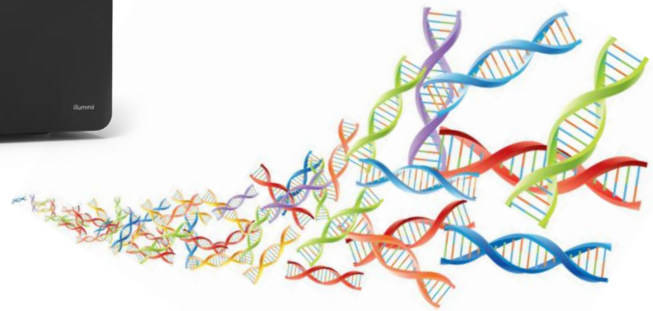


Formati di file bioinformatici

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

Argomenti per i progetti e conclusioni generali

Dal sequenziatore..



- cat, sed, awk

..al genoma..

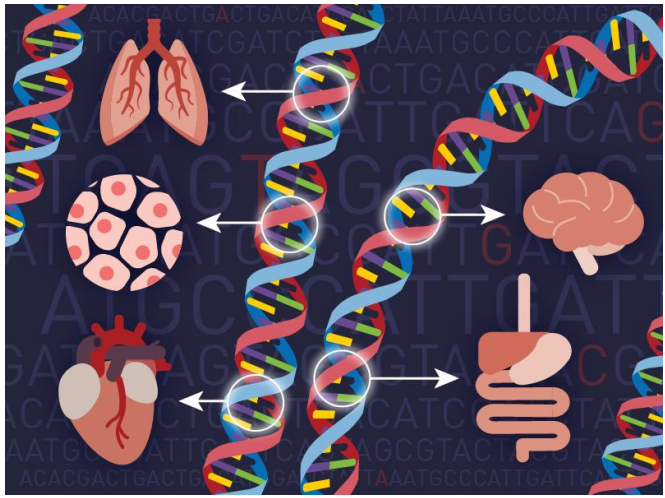
- Flye
- miniasm

- bowtie2
- bwa
- minimap2

- bcftools

..alla funzione

- hmmratac
- CHROMHMM
- hmmer
- AuthentiCT
- HMMcopy
- Phastcons



Formati di file bioinformatici

.fastq

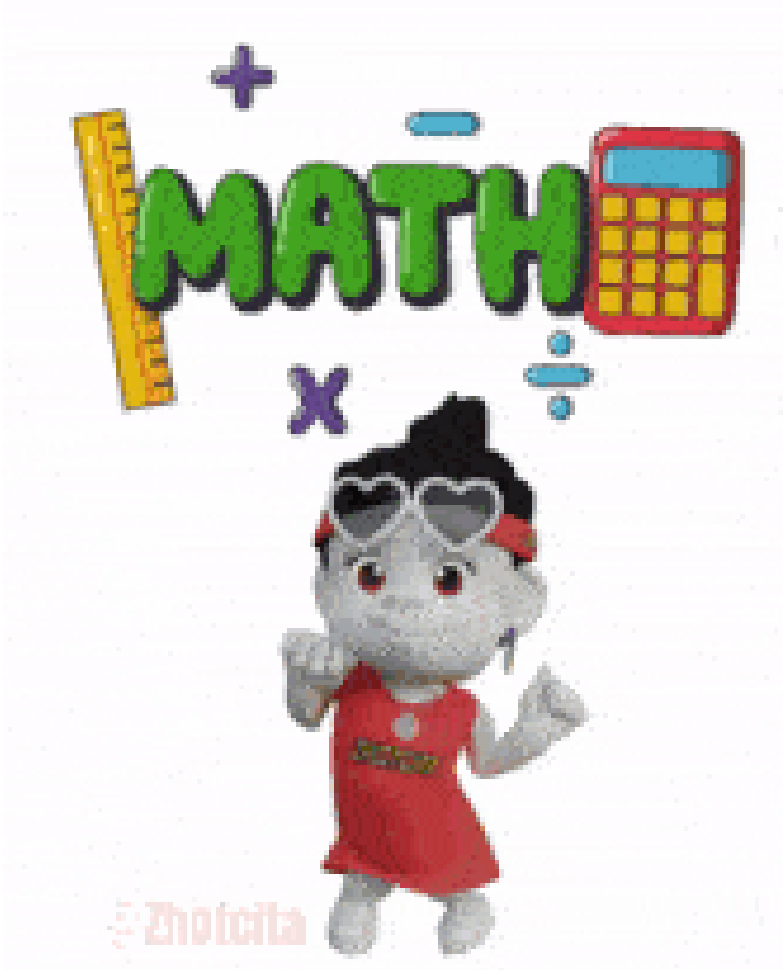
- samtools
- bcftools

.fasta, .bam, .vcf

- bedtools

.bed


Congrats!


MATH

Cosa ci possiamo fare con tutti questi strumenti?

ENCODE | Data | Encyclopedia | Materials & Methods | Help

Functional Genomics data
- Experiment search
- Experiment matrix
- ChIP-seq matrix
- Human and mouse body maps
- Functional genomics series
- Single-cell experiments

Functional Characterization data
- Experiment search
- Experiment matrix

Cloud Resources
- AWS Open Data

Collections
- RNA-protein interactions (ENCORE)
- Epigenomes from four individuals (ENTEx)
- Rush Alzheimer's disease study
- Stem cell differentiation
- Deeply profiled cell lines
- Human donor matrix
- Immune cells
- Human reference epigenomes
- Mouse reference epigenomes
- Mouse development matrix
- Protein knockdown (Degron)

Search by region

arch the ENCODE Portal

ENCODE SCREEN

Functional

characterization

Encyclopedia of elements

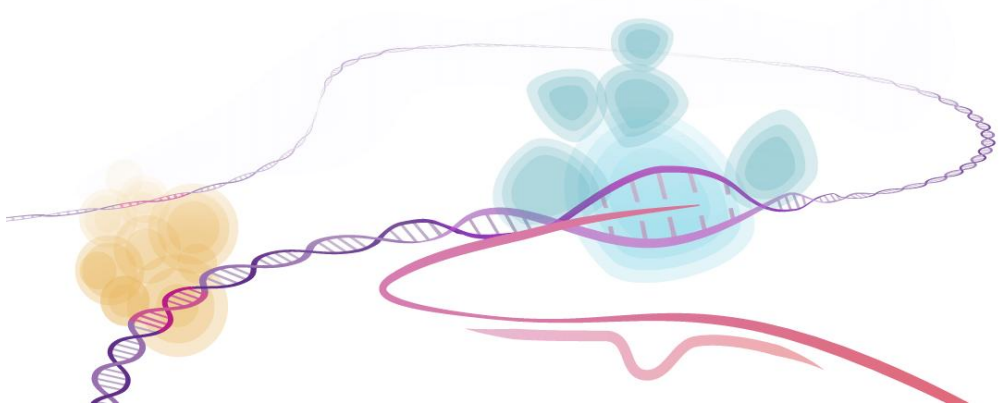
ENCODE GTEx

EN-TEx

Deeply profiled cell lines

Caratterizzate differenze nello stato epigenomico di tipi cellulari diversi

Cosa fareste?



Caratterizzate differenze nello stato epigenomico di tipi cellulari diversi

Andate su ENCODE e scegliete dei tipi cellulari di vostro interesse (ad esempio due o piu' tipi di linfociti; due diversi tipi di neurone), verificando come visto a lezione che abbiano dati comparabili e fastq/bam files disponibili

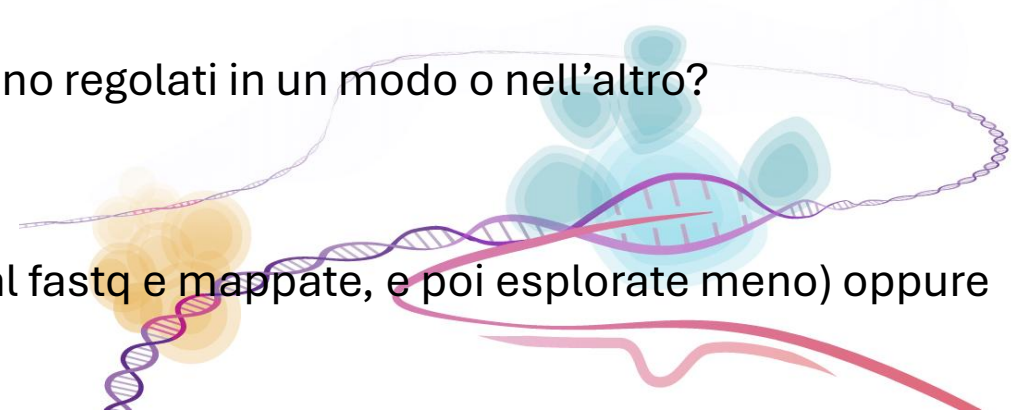
Caratterizzatene gli stati epigenetici con CHROMHMM e hmmratac

Usate CHROMHMM per segmentare il genoma in vari stati (provate vari numeri se avete molte tracce epigenetiche, ad esempio 4 stati per 6 tracce istoniche + ATACseq)

Esploratene:

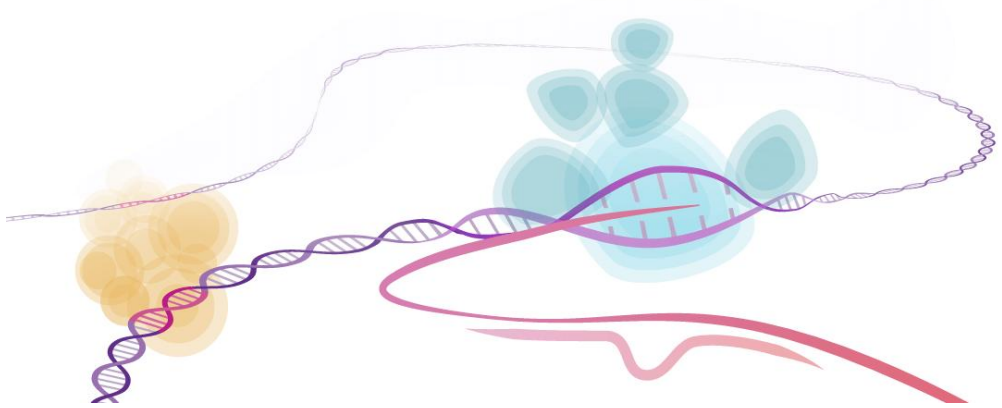
- il significato guardando alle probabilità di emissione
- le proporzioni del genoma occupate dai vari stati
- usando bedtools intersect, quali geni o regioni del genoma sono regolati in un modo o nell'altro?
- Che differenze ci sono tra i due tipi cellulari?

Potete scegliere se fare di piu' della parte preliminare (partite dal fastq e mappate, e poi esplorate meno) oppure partite direttamente dal .bam ed esplorate di piu'



Caratterizzate differenze nello stato epigenomico indotto da un tumore

Cosa fareste?



Caratterizzate differenze nello stato epigenomico indotto da un tumore

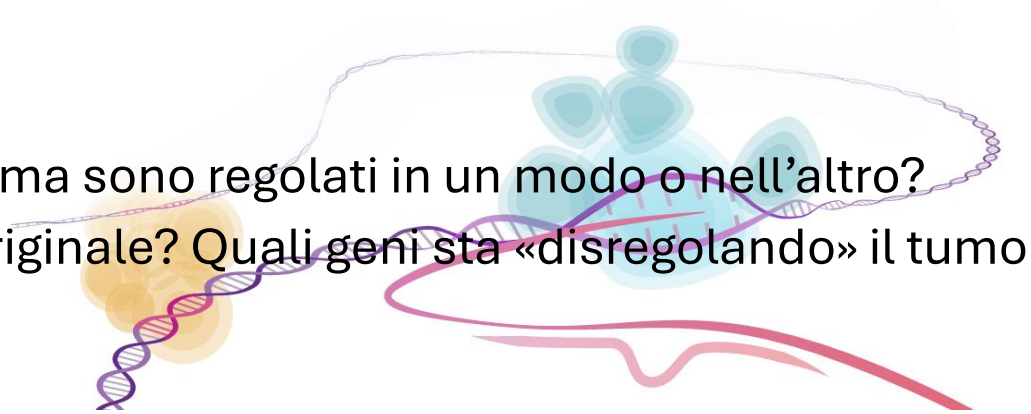
Andate su ENCODE e scegliete una linea cellulare o un tumore di vostro interesse ed il corrispondente tipo cellulare (oppure potete comparare due linee tumorali) verificando come visto a lezione che abbiano dati comparabili e fastq/bam files disponibili

Caratterizzatene gli stati epigenetici con CHROMHMM e hmmratac

Usate CHROMHMM per segmentare il genoma in vari stati (provate vari numeri se avete molte tracce epigenetiche, ad esempio 4 stati per 6 tracce istoniche + ATACseq)

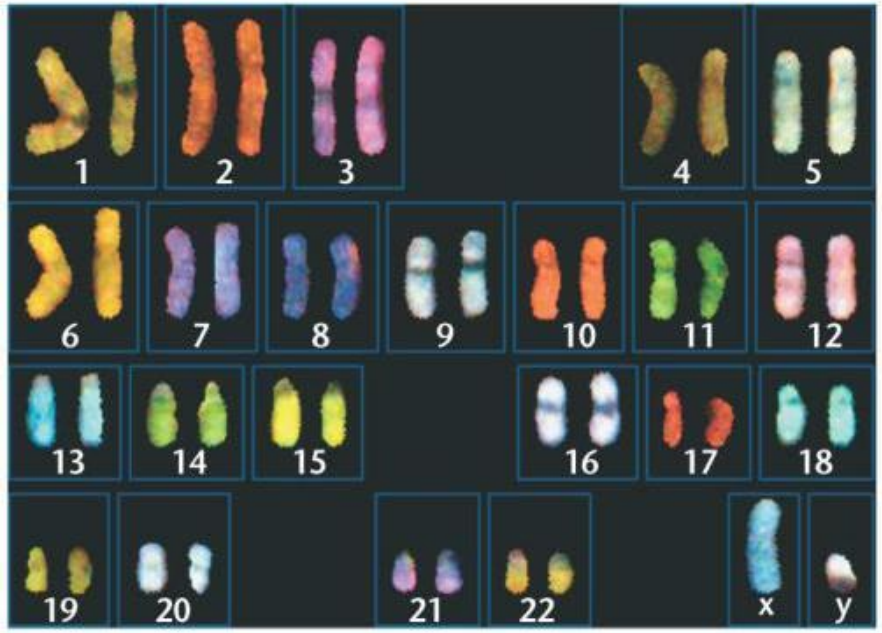
Esploratene:

- il significato guardando alle probabilità di emissione
- le proporzioni del genoma occupate dai vari stati
- usando bedtools intersect, quali geni o regioni del genoma sono regolati in un modo o nell'altro?
- Che differenze ci sono tra i il tumore e il tipo cellulare originale? Quali geni sta «disregolando» il tumore?

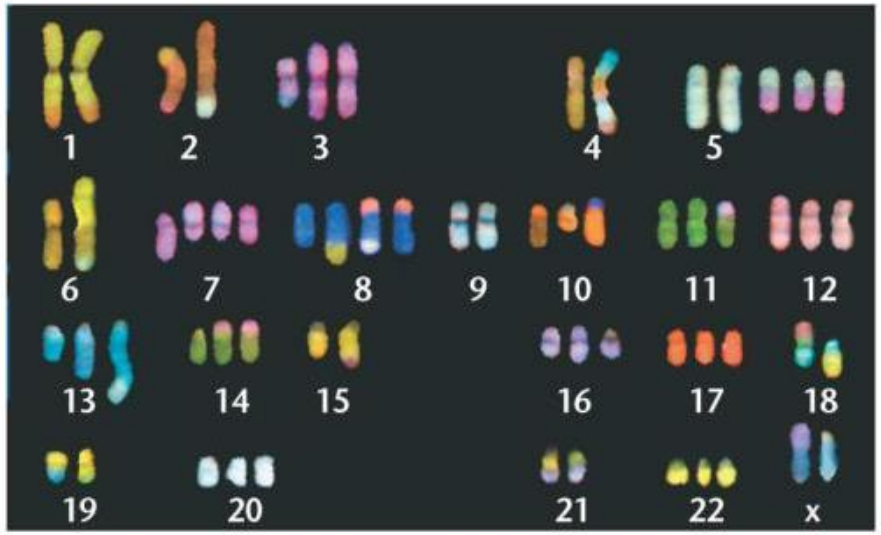


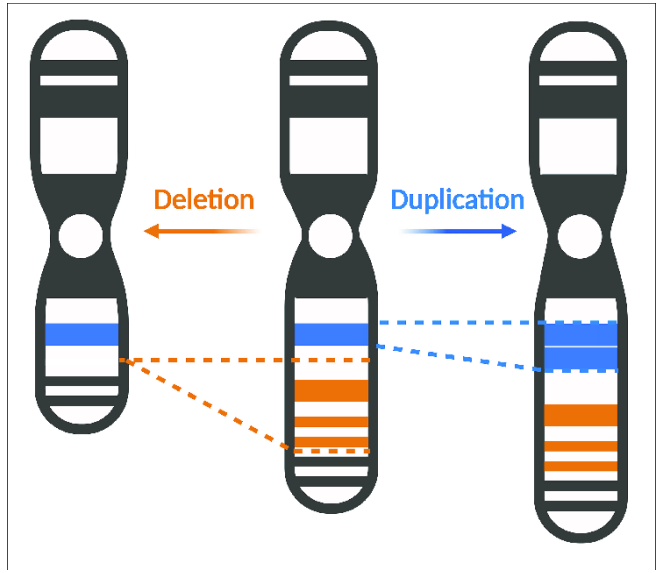
Esaminate campioni di tumore per trovare riarrangiamenti con HMMcopy/CNVpytor

(a) cariotipo normale



(b) cariotipo tumorale

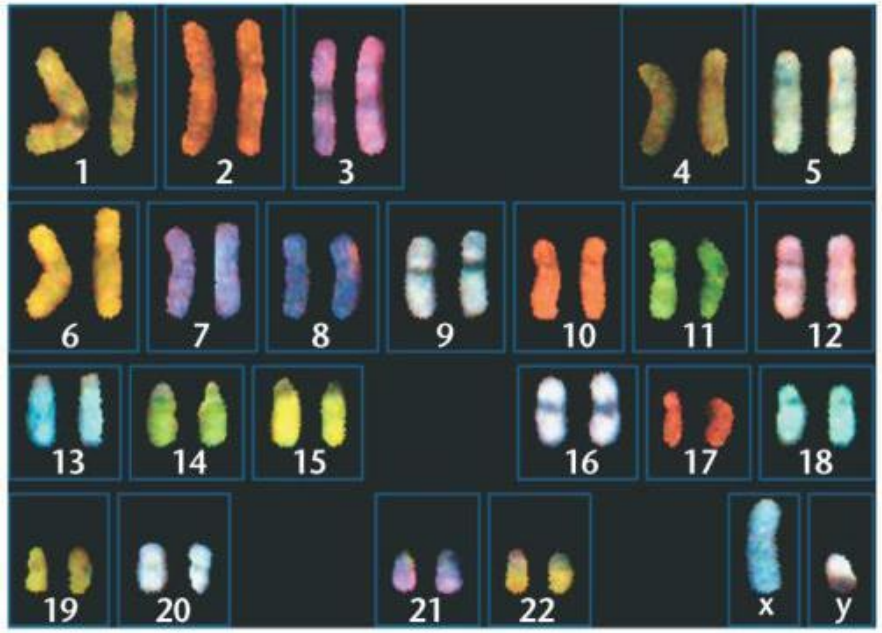




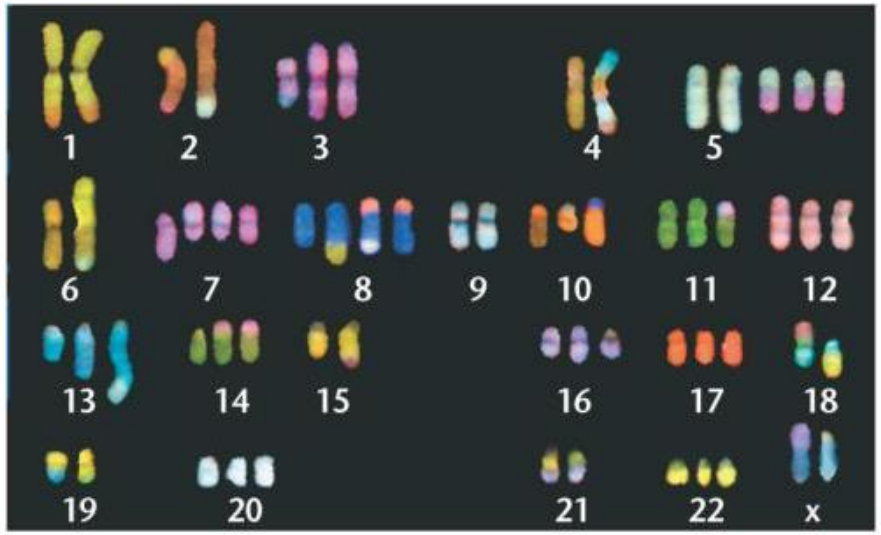
Esaminate campioni di tumore per trovare riarrangiamenti con HMMcopy

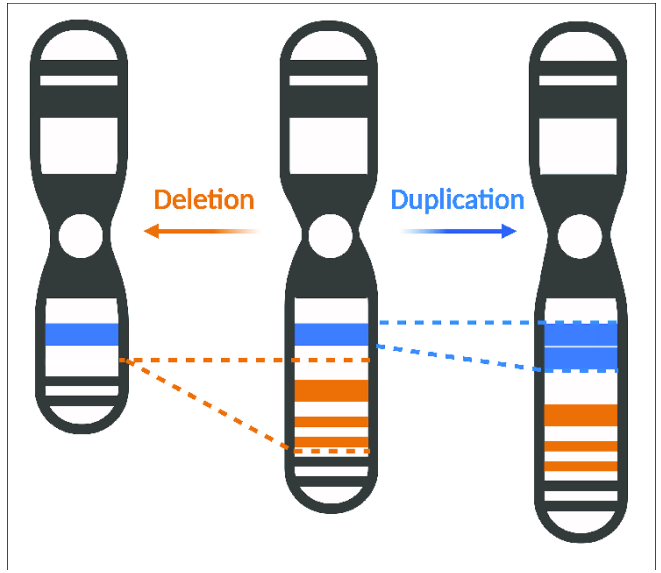
Che fareste?

(a) cariotipo normale



(b) cariotipo tumorale

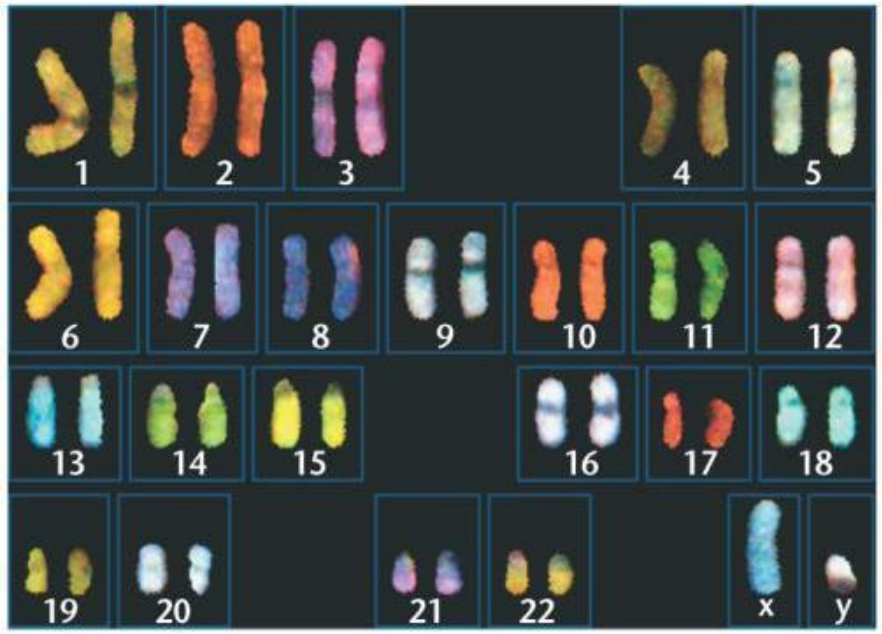




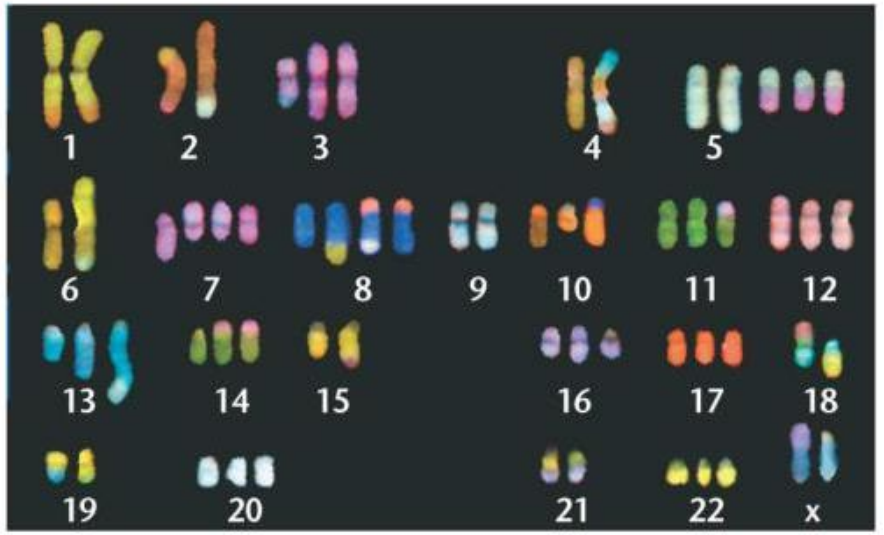
Esaminate campioni di tumore per trovare riarrangiamenti con HMMcopy

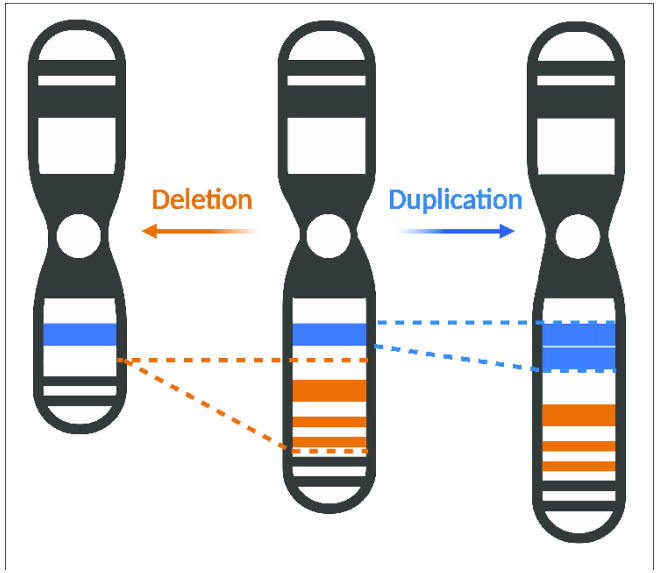
- Scaricate i vostri tumori di interesse: moltissimi dataset specializzati, ma potreste anche usare linee cellulari su ENCODE come **GM12878** (linfoblastoidi) o **K562** (Leucemia mieloide cronica, con note CNV).
- Mappate le reads e usate samtools per esplorare la variazione in coverage lungo il genoma
- Usate HMMcopy per individuare copy number variation (CNV)

(a) cariotipo normale



(b) cariotipo tumorale



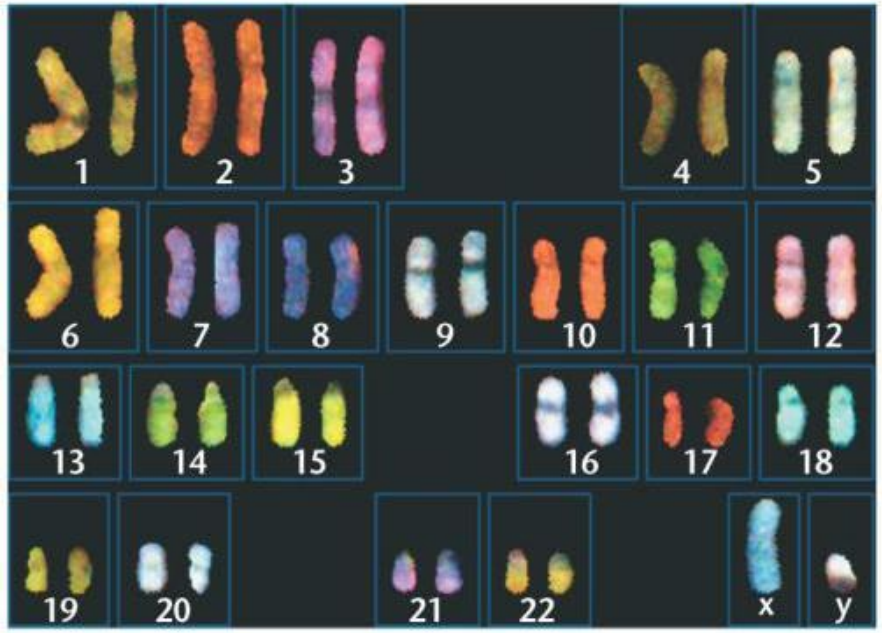


Esaminate campioni di tumore per trovare riarran… HMMcopy

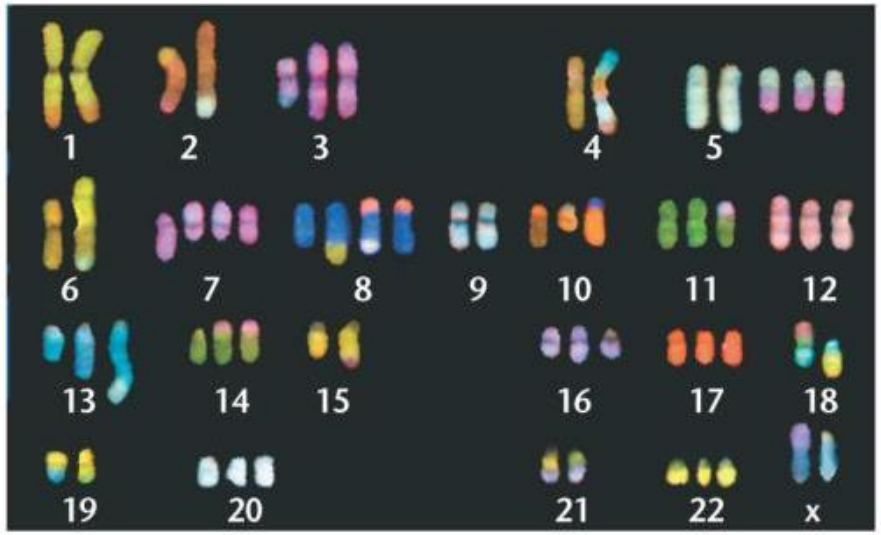
Agli interessati fornirò dati e scripts di aiuto

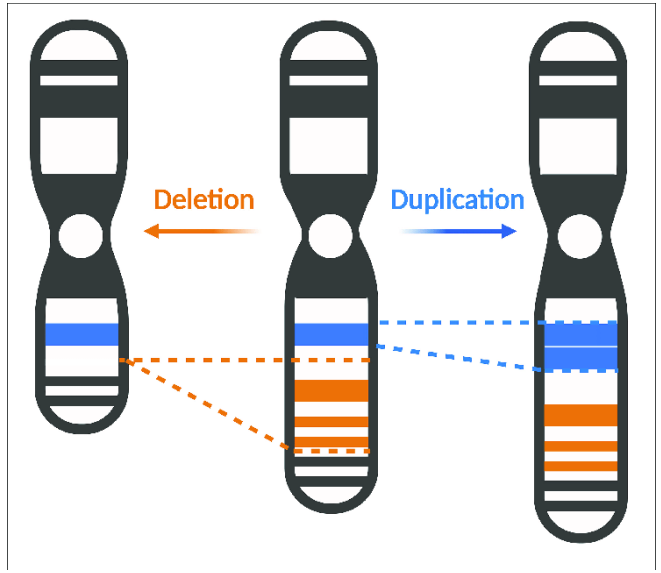
- Scaricate i vostri tumori di interesse: moltissimi dataset specializzati, ma potreste anche usare linee cellulari su ENCODE come **GM12878** (linfoblastoidi) o **K562** (Leucemia mieloide cronica, con note CNV).
- Mappate le reads e usate samtools per esplorare la variazione in coverage lungo il genoma
- Usate HMMcopy per individuare copy number variation (CNV)

(a) cariotipo normale



(b) cariotipo tumorale





Peste giustiniana

Abbiamo delle ossa potenzialmente attribuibili alla peste Giustiniana (541-542 d.C)

Che fareste?

Peste giustiniana

- Mappate le reads
- Esaminatene i pattern di C>T con AuthentiCT
- Competitive mapping per vedere se nei nostri campioni c'è *Yersinia pestis*
- Esaminatene il sesso del campione

Metagenomica di DNA antico dalle stalattiti del Salar Boliviano

DNA antico da stalattiti apparentemente di origine vegetale

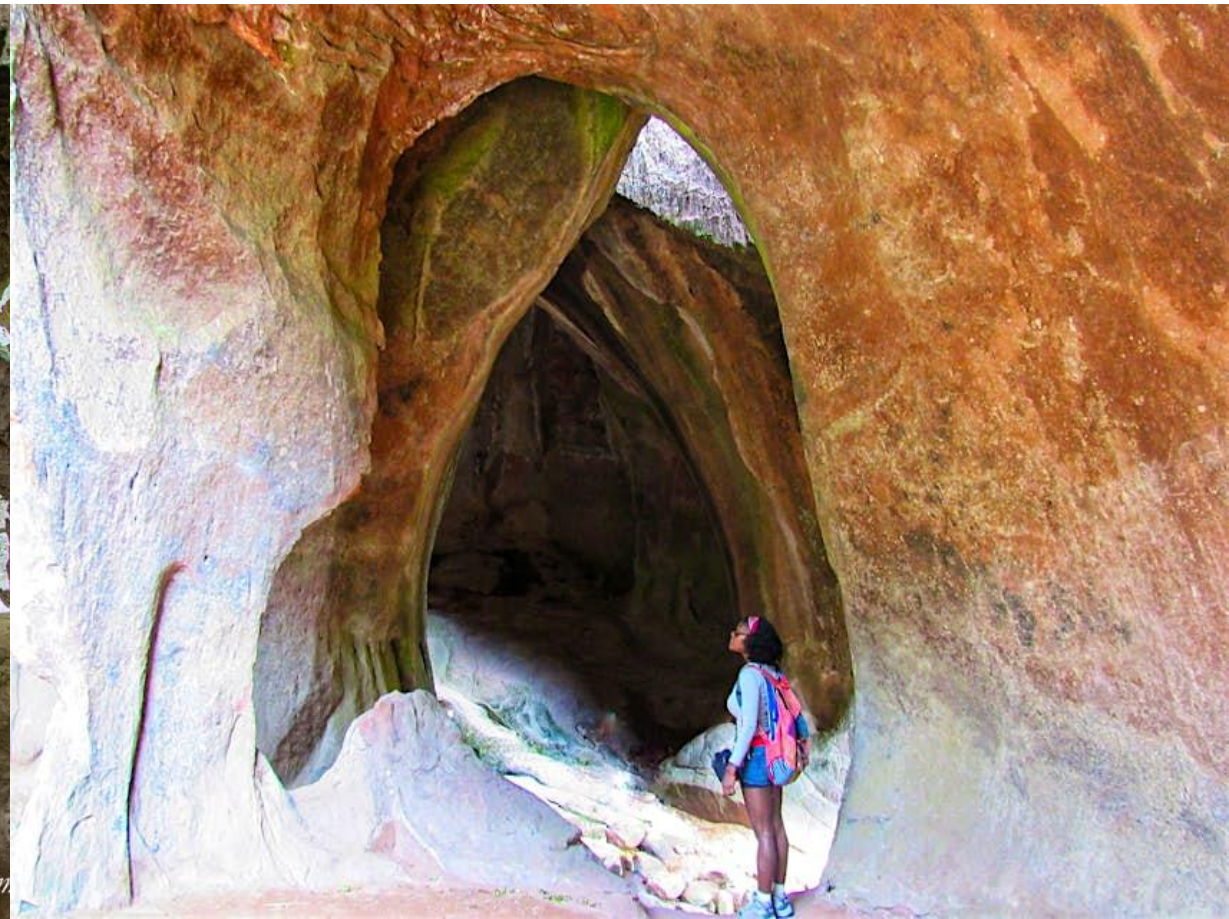
Metagenomica di DNA antico dalle stalattiti del Salar Boliviano

DNA antico da stalattiti apparentemente di origine vegetale

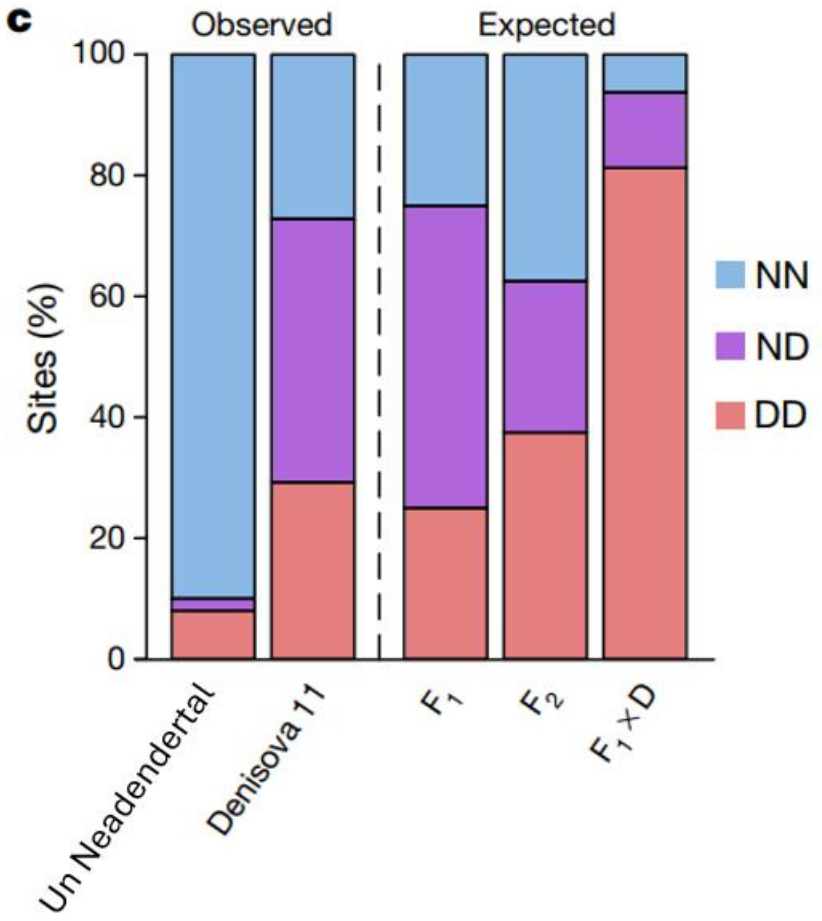
Metagenomica di DNA antico dalle stalattiti del Salar Boliviano

Progetto 1

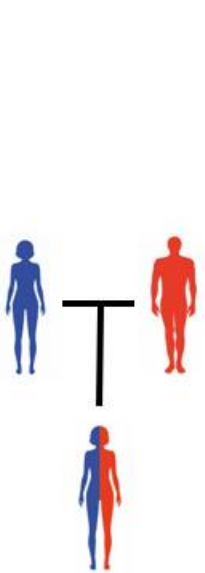
- Competitive mapping per trovare da quale organismo derivano
- Classificazione delle reads come endogene e contaminanti con AuthentiCT



Rianalizzate Denisova 11



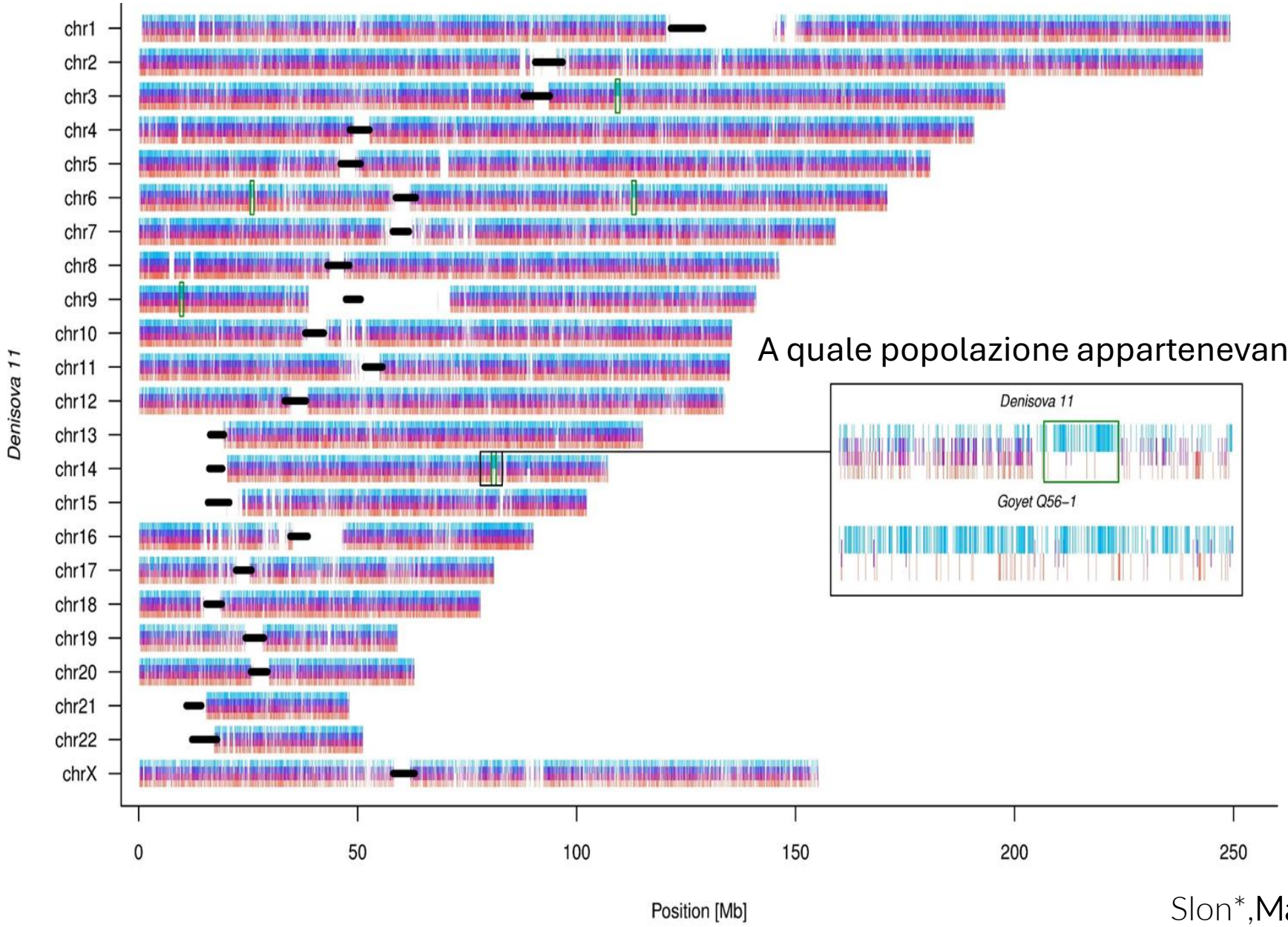
Campioniamo 2 reads



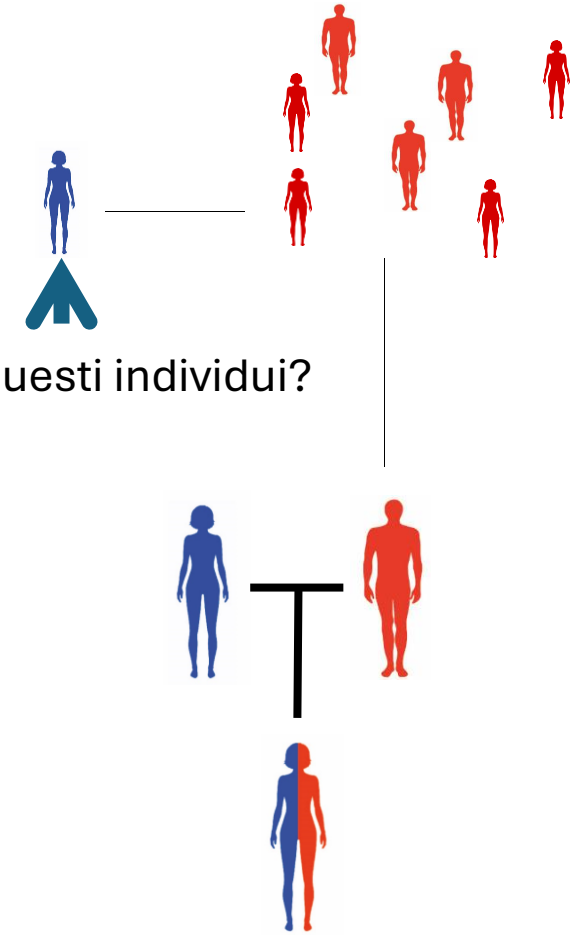
Rianalizzate Denisova 11

- Mappate le reads
- Esaminatene i pattern di C>T e la contaminazione con AuthentiCT
- Nel frattempo abbiamo generato altri genomi di Neandertal:

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

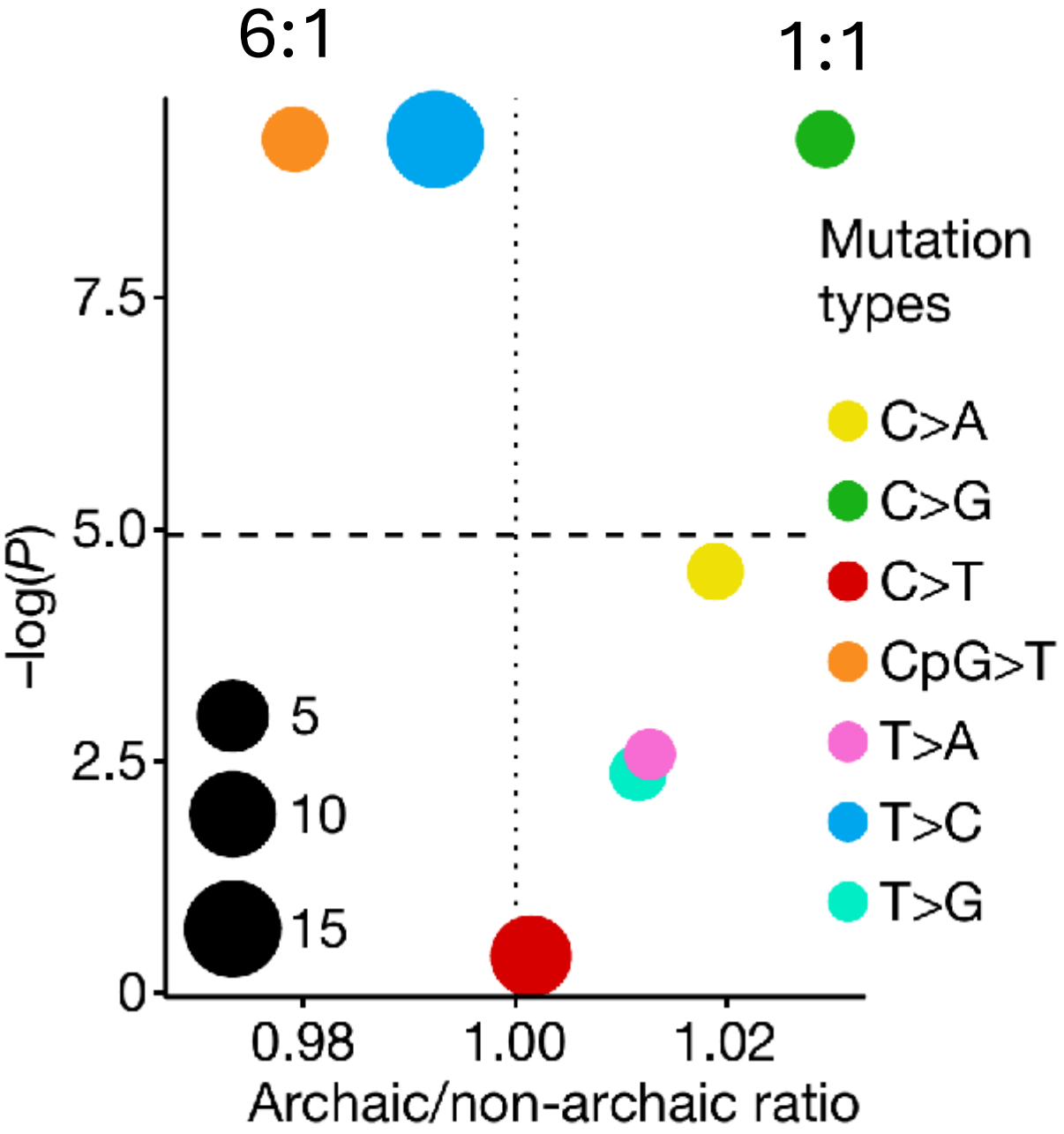


A quale popolazione appartenevano questi individui?



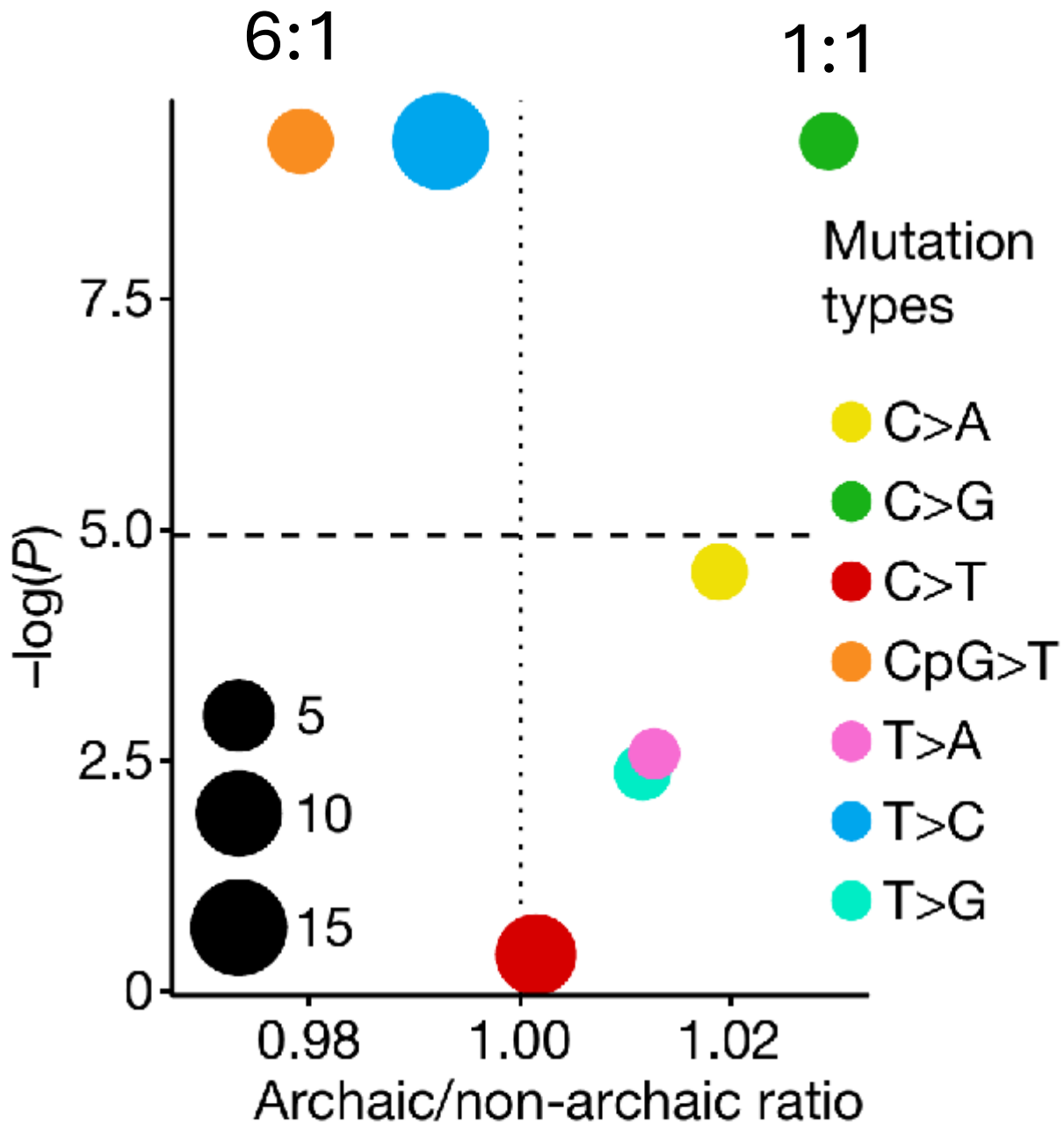
Slon*,**Mafessoni***,Vernot* et al,2018. *Nature*

Quanto erano «vecchie» le mamme Denisova?



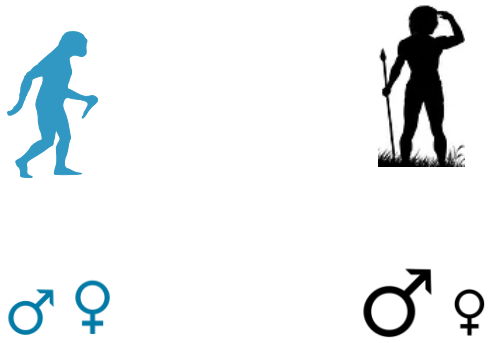
Skov et al., 2020, *Nature*

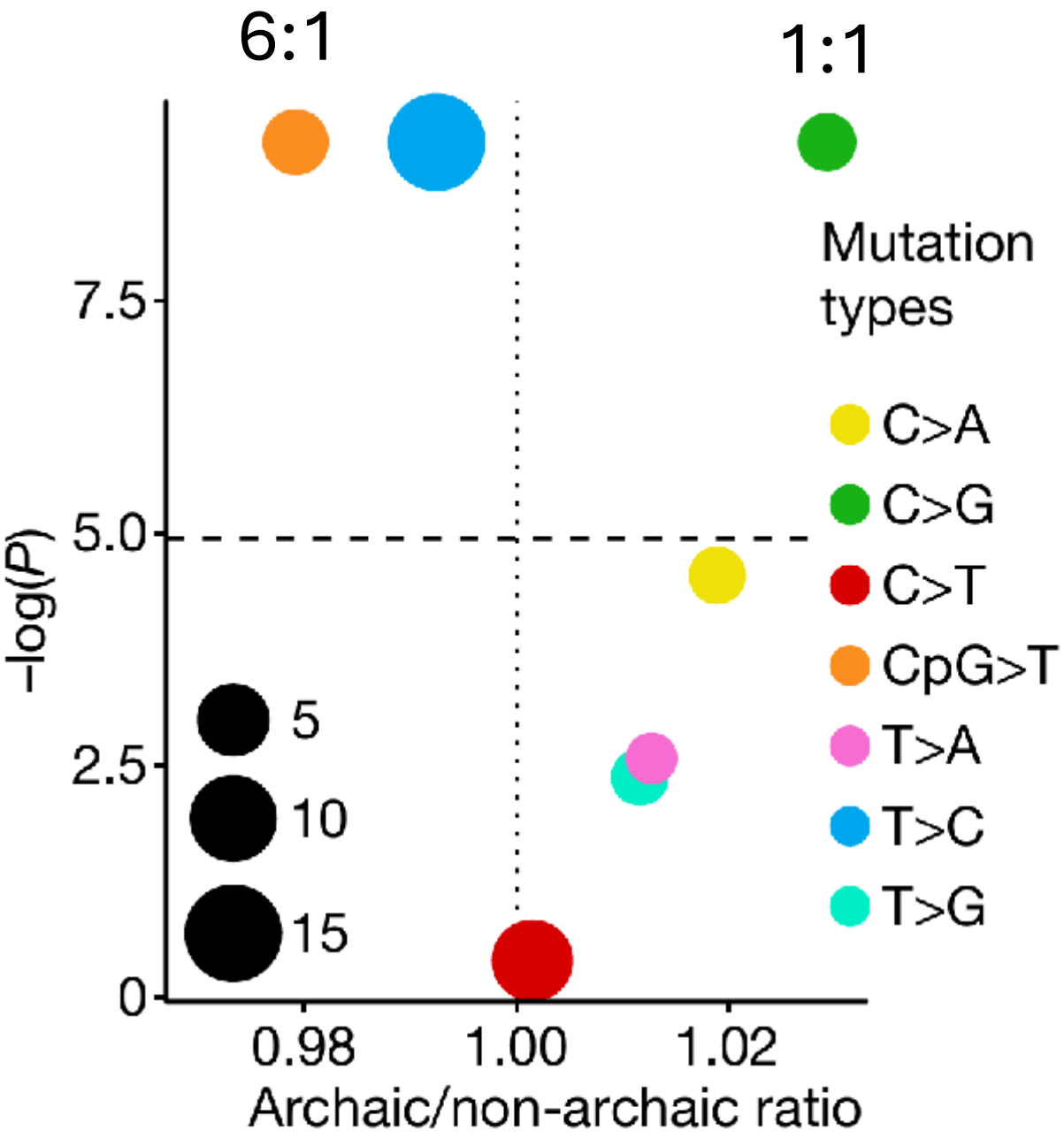
$$\frac{\text{paternal } ♂}{\text{maternal } ♀} \text{ mutation ratio}$$



Skov et al., 2020, *Nature*

$$\frac{\text{paternal } ♂}{\text{maternal } ♀} \text{ mutation ratio}$$

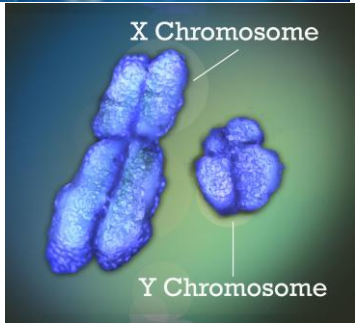




Skov et al., 2020, *Nature*

Mappate le reads fastq di un «Donatore anonimo» alla referenza *T2T* (e opzionalmente ad un altro genoma di referenza) e chiamatene i genotipi

Esploratene il cromosoma Y: quante varianti si trovano (e con quale qualità) nelle zone prima inaccessibili con la vecchia referenza del genoma umano?



Usate hmmer per classificare geni associati alla degradazione della plastica (in genomi fungini)

Item	Update
Microorganisms	875
Proteins	329
Last update	04/11/2025
Last addition	Protein 329

New Study on the Global Ocean Microbiome

PlasticDB was used to study the expression of putative plastic-degrading enzymes in the global ocean microbiome. The paper suggests that the expression of these enzymes do not correlate to plastic pollution! Check the paper at: doi.org/10.1186/s40793-024-00575-4.

Collaborate with PlasticDB

We invite researchers, developers, or enthusiasts to contribute to PlasticDB. We are planning on releasing new features and increasing our literature coverage. Please contact us at contact.plasticdb@gmail.com if you think you can help. Thank you! :)

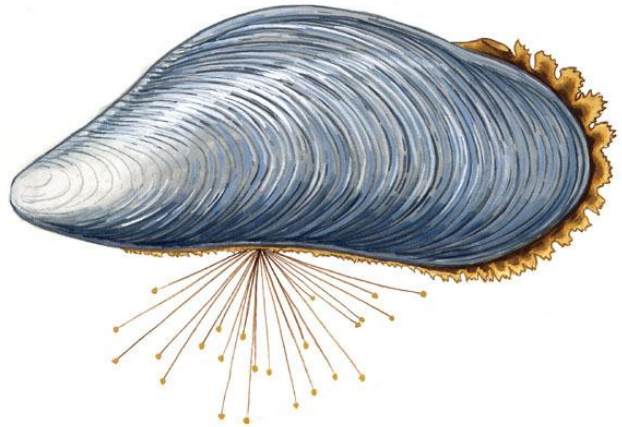
About PlasticDB

The database currently has 875 species of microorganisms that were reported in the scientific literature to have plastic-degrading capabilities and 329 proteins described to breakdown plastics.

The database is updated bimonthly to keep up with new publications and reports of plastic biodegradation. There are several options for using the data in the database:

- Get information about specific species (Microorganisms)
- Get information about the proteins that degrade each type of plastic (Proteins)
- Annotate a single protein and find out if it matches any of the proteins in the database (Annotate gene)
- Annotate a whole genome or metagenome (Annotate genome)
- Annotate and compare genomes or metagenomes (Compare genomes)

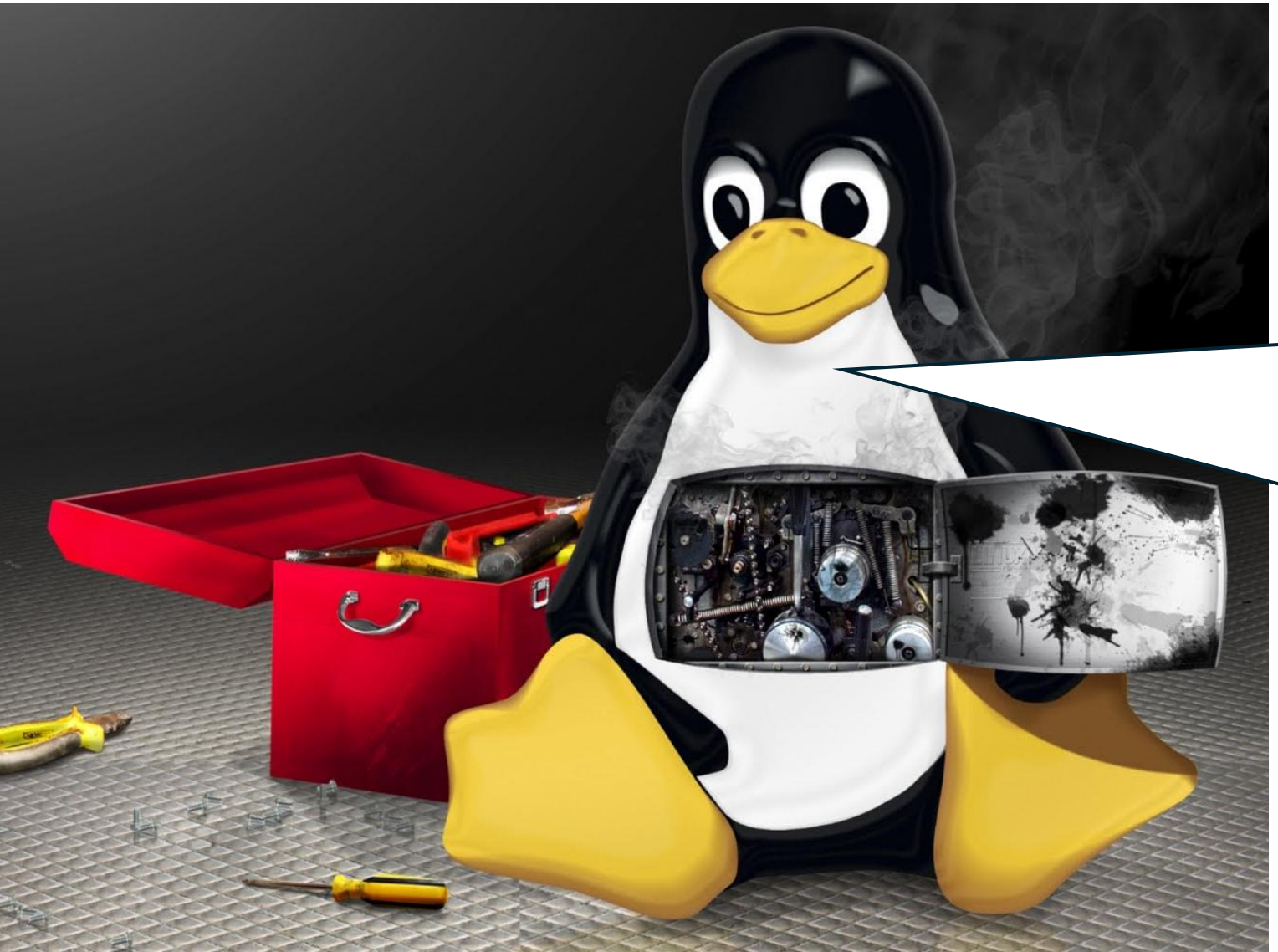
Con Dona



- Mappate ed analizzate le reads di larve di mitilo

- Dati di metabarcoding di cloroplasto per api, per vedere che tipo di piante utilizzando in ambienti ristorati vs non ristorati

Progetti di programmazione/algoritmi



Potete usare chatgpt ma commentate e descrivete bene i passaggi cosi che sia chiaro che abbiate capito ciò che fate

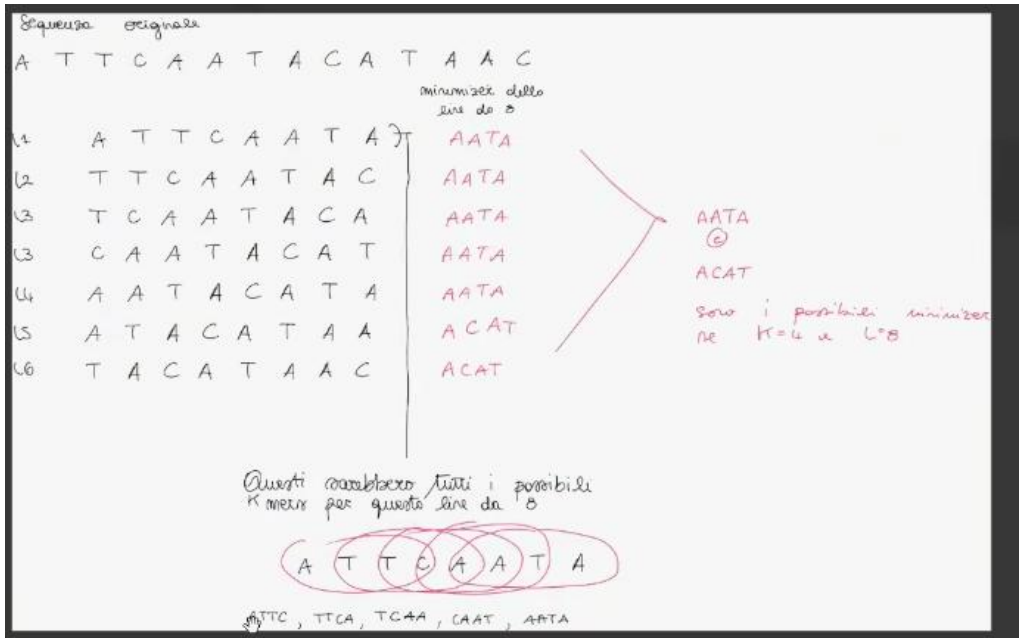
Progetti ambiziosi. Va bene anche che falliate!

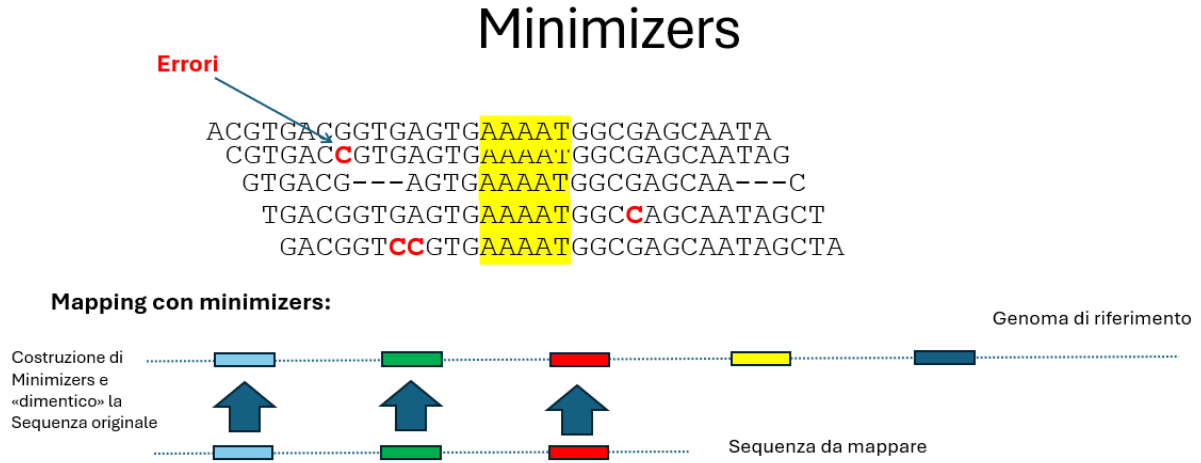
Scrivete uno script per comprimere e decomprimere un genoma con trasformata di Burrow-Wheeler

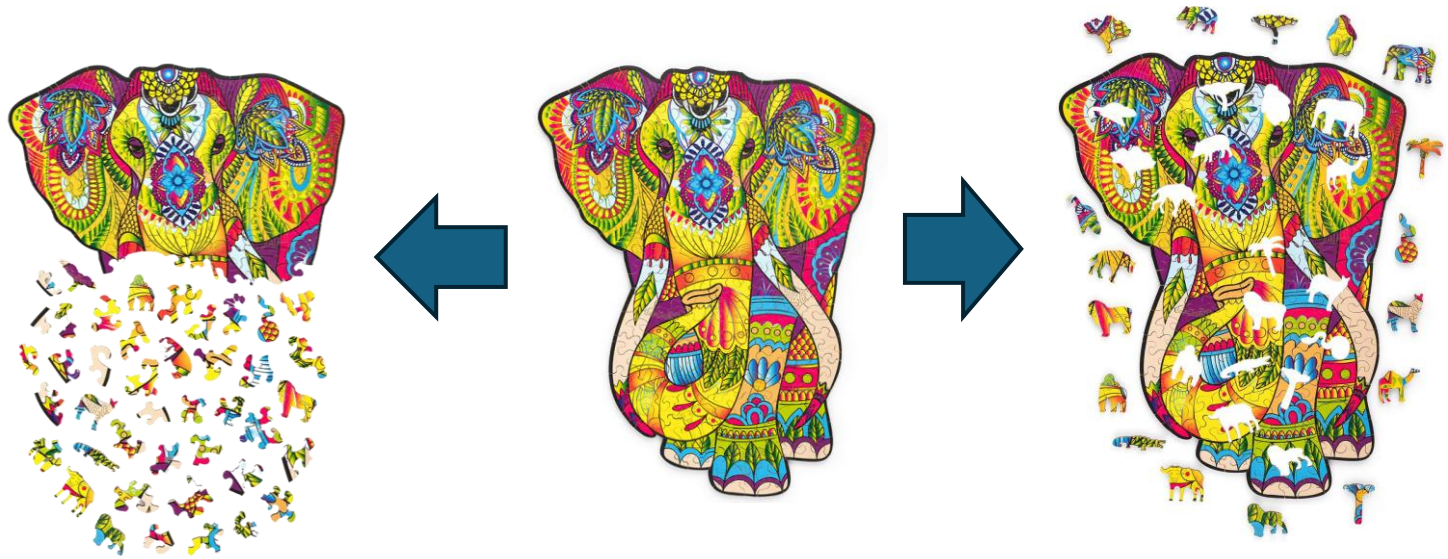
Potete usare chatgpt ma commentate e descrivete bene i passaggi cosi che sia chiaro che abbiate capito ciò che fate

$	p	a	n	a	m	a	b	a	n	a	n	a	s
a_1	b	a	n	a	n	a	s	$	p	a	n	a	m
a_2	m	a	b	a	n	a	n	a	s	$	p	a	n
a_3	n	a	m	a	b	a	n	a	n	a	s	$	p
a_4	n	a	n	a	s	$	p	a	n	a	m	a	b
a_5	n	a	s	$	p	a	n	a	m	a	b	a	n
a_6	s	$	p	a	n	a	m	a	b	a	n	a	n
b	a	n	a	n	a	s	$	p	a	n	a	m	a
m	a	b	a	n	a	n	a	s	$	p	a	n	a
n	a	m	a	b	a	n	a	n	a	s	$	p	a
n	a	n	a	s	$	p	a	n	a	m	a	b	a
n	a	s	$	p	a	n	a	m	a	b	a	n	a
p	a	n	a	m	a	b	a	n	a	n	a	s	$
s	$	p	a	n	a	m	a	b	a	n	a	n	a

Scrivete con awk un piccolo programma che trovi i minimizers e per due sequenze «mappi» usando solo i minimizers







A vostro piacimento!!

Se iniziate dai fastq files invece che dai .bam

Usate fastp o AdapterRemoval!!

- `fastp -i in.R1.fq.gz -o out.R1.fq.gz`
- `AdapterRemoval --file1 in.R1.fq --file2 in.R2.fq --output1 out.R1.fq --output2 out.R2.fq`

```
~$ ./create_report_sequences.sh report.txt
~$ cat report.txt
================
sequences1.fastq
----------------
number of reads
250
number of high quality (F) bases
22881
number of adaptors
9
================
sequences2.fastq
----------------
number of reads
250
number of high quality (F) bases
23064
number of adaptors
3
================
```

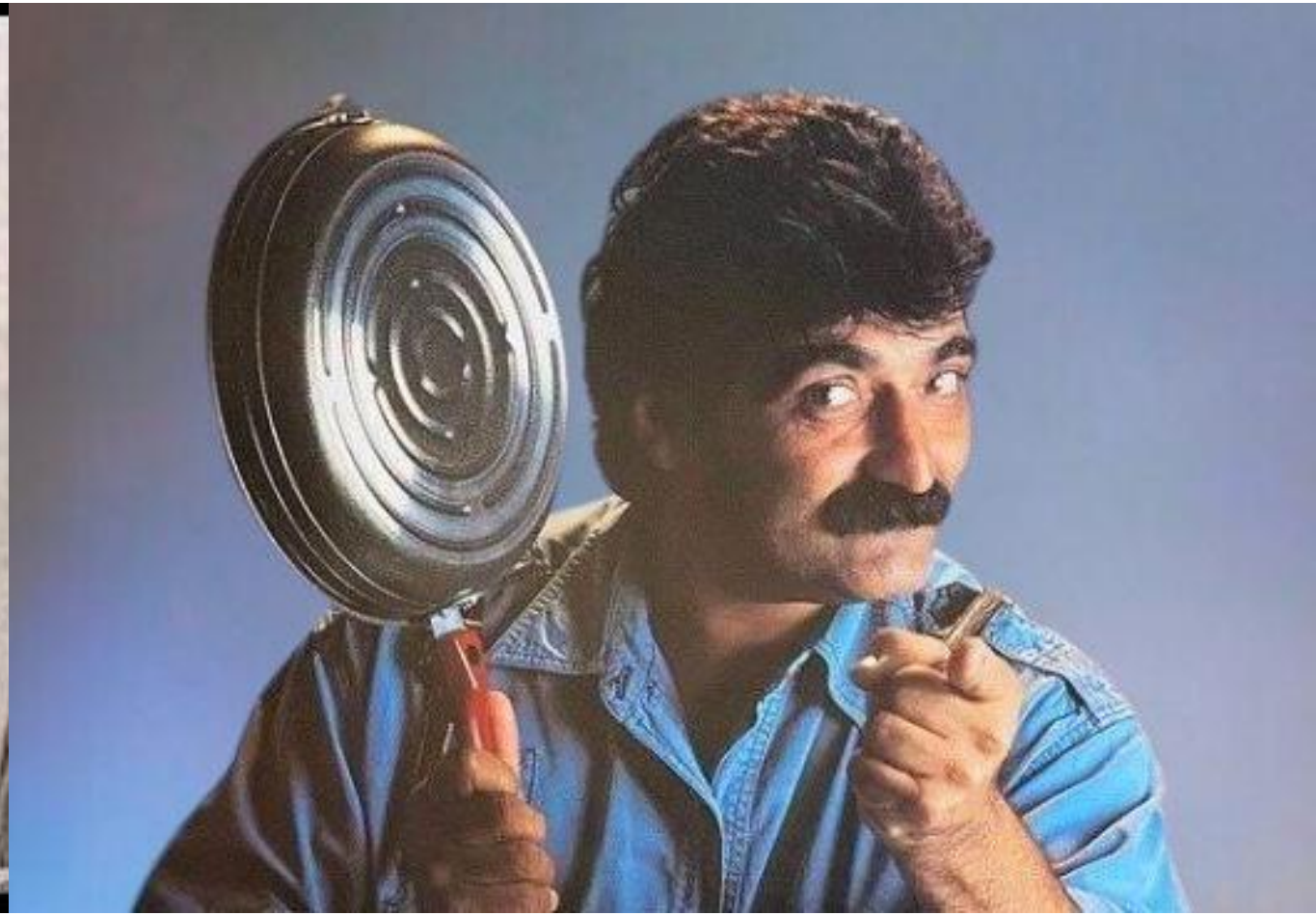
Istruzioni per il rapporto del progetto

- Sezioni:
  - **Abstract**: una piccola descrizione del progetto.
  - **Introduzione**: breve descrizione biologica del progetto e di ciò che avete fatto
  - **Metodi**: descrizione generale delle analisi. Non lesinatevi in dettagli. Scrivete perché avete fatto ciò che avete fatto, perché avete scelto certi programmi e non altri, etc.
  - **Risultati:** elencate i comandi che avete eserguito, i risultati ottenuti, etc.
  - **Discussione e Conclusioni:** descrivete cosa avete trovato, ma anche cosa avete notato e imparato metodologicamente.
  - **Appendice dei tentativi falliti:** non lesinatevi nel descrivere ciò che avete provato, cosa non ha funzionato e come lo avete corretto. Non necessariamente tutto (ad esempio «`cd .. ma ero nella cartella sbagliata`», ma mostratemi come avete ragionato. E' da qui che vedrò il vero lavoro!

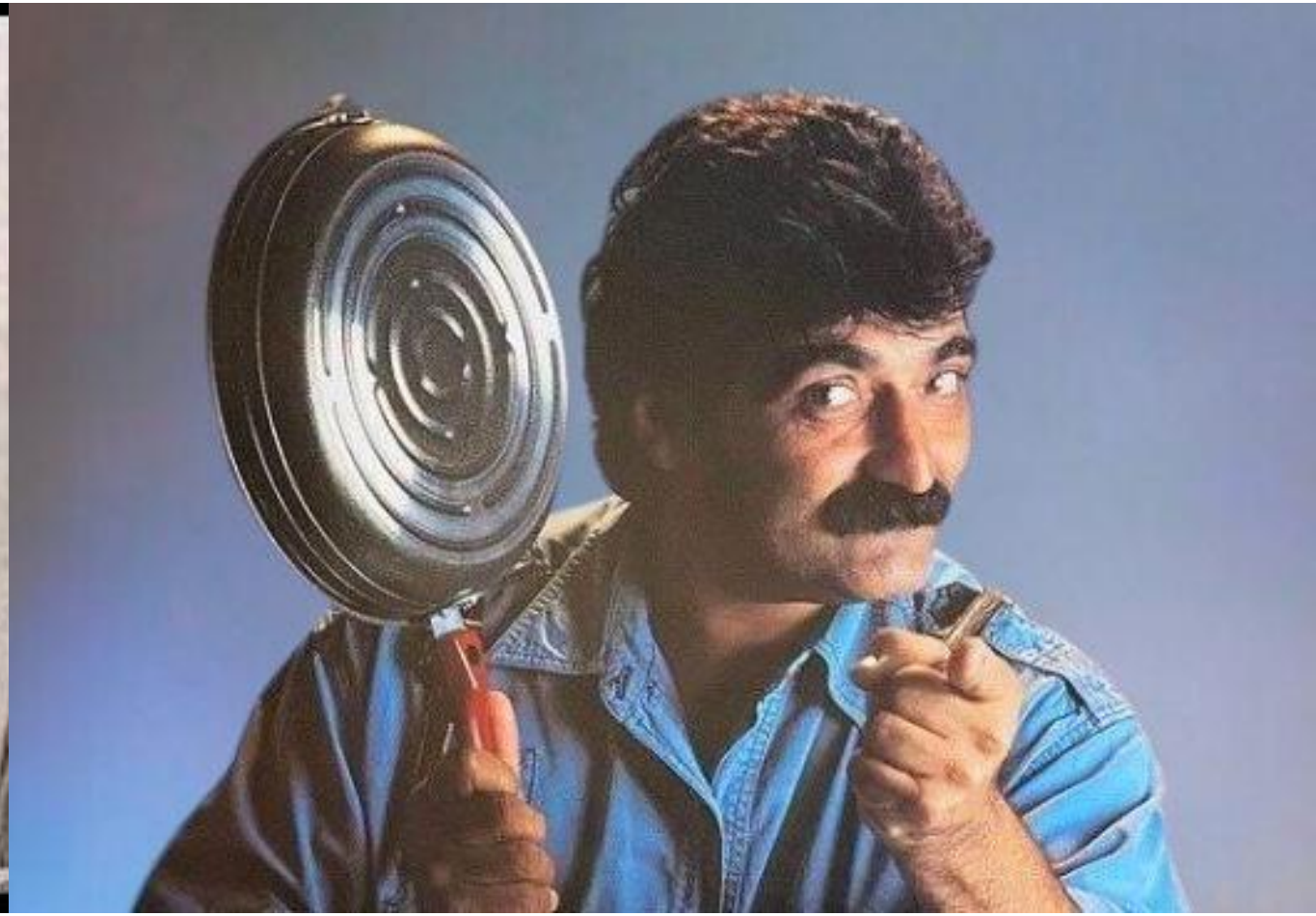
- Lingua: idealmente inglese, ma va benissimo anche in italiano.
- Sentite liberi di usare chatgpt ma con criterio – rischiate che vi faccia cose difficili per voi da spiegare e farvi fare piu' lavoro!! Il progetto è anche per imparare (e lo noto subito ciò che è fatto con chatgpt).
- Nessun limite (inferiore o superiore di pagine). Non lesinatevi sui tentativi falliti.
- Siamo qui per aiutarvi, usate l'opportunità per chiedere e imparare davvero!


Rai 1 HD
Siamo qui per aiutarvi, usate l’opportunità per chiedere e praticare assistiti!

E ora la pubblicità!



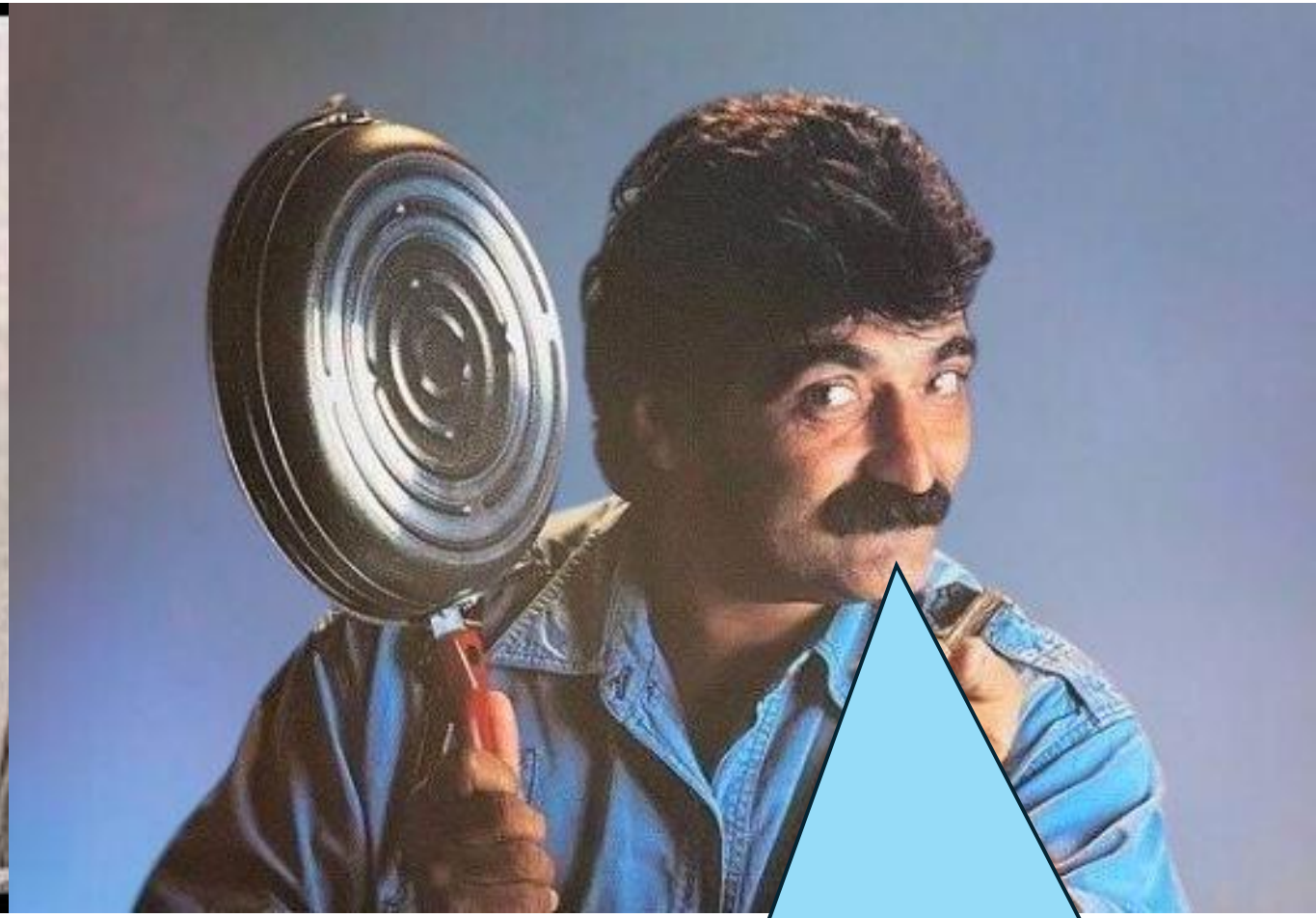
E ora la pubblicità!



Corso di **Biologia Evoluzionistica** con **pratica in R** nel secondo semestre per il terzo anno di STB:

- Alberi filogenetici
- Genetica di popolazione
- Come trovare regioni del genoma sotto selezione

E ora la pubblicità!



Siete benvenuti come **tesisti/e**!

Alcuni argomenti possibili:

- Biodiversità e Genomica Comparativa
- Riparazione del DNA (tumori, CRISPR)
- DNA antico ed evoluzione