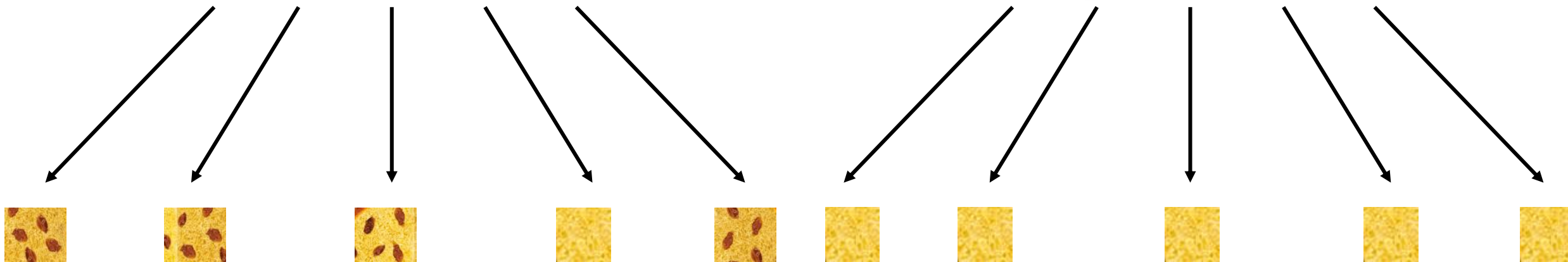
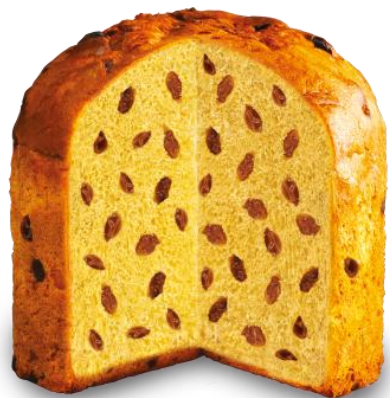


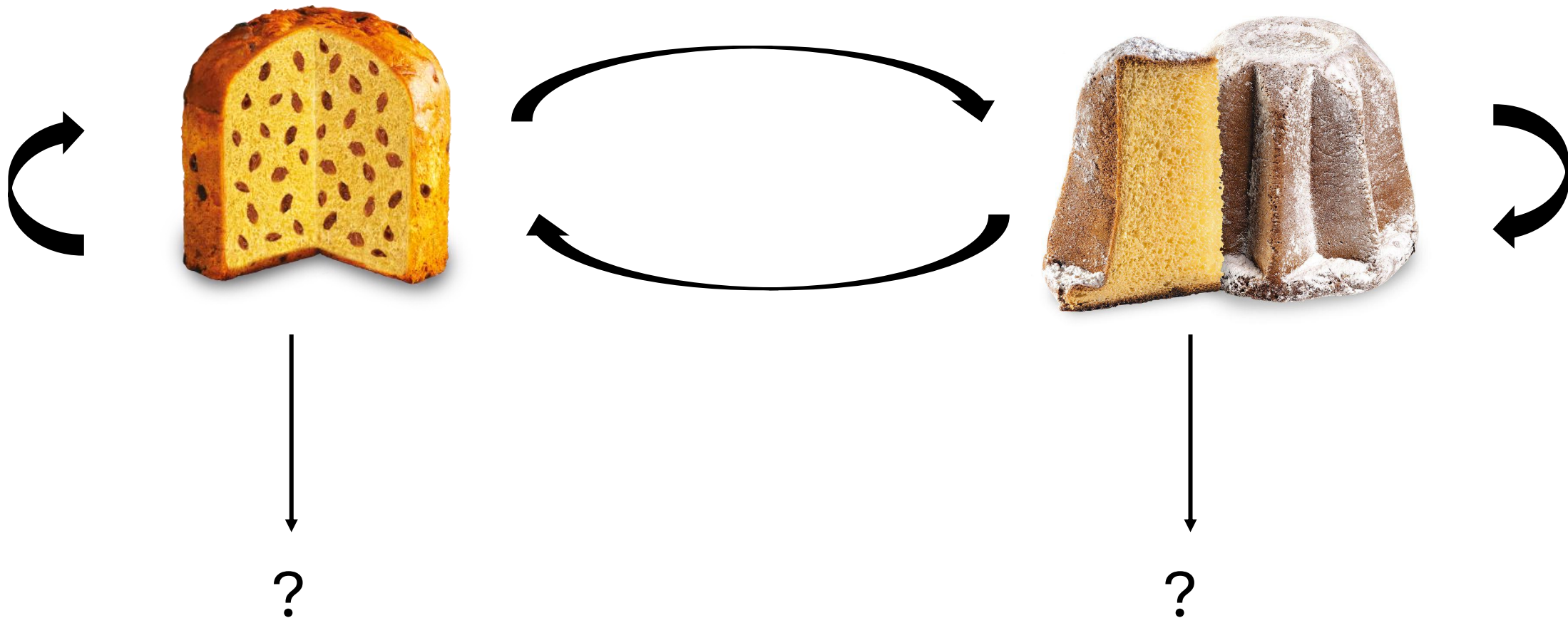
APPLICAZIONI DEGLI HIDDEN MARKOV MODELS IN GENOMICA

FABRIZIO MAFESSONI, GENOMICA COMPUTAZIONALE 2025/2026
CORSO DI LAUREA SPECIALISTICA IN GENOMICA FUNZIONALE, UNIVERSITÀ DI TRIESTE



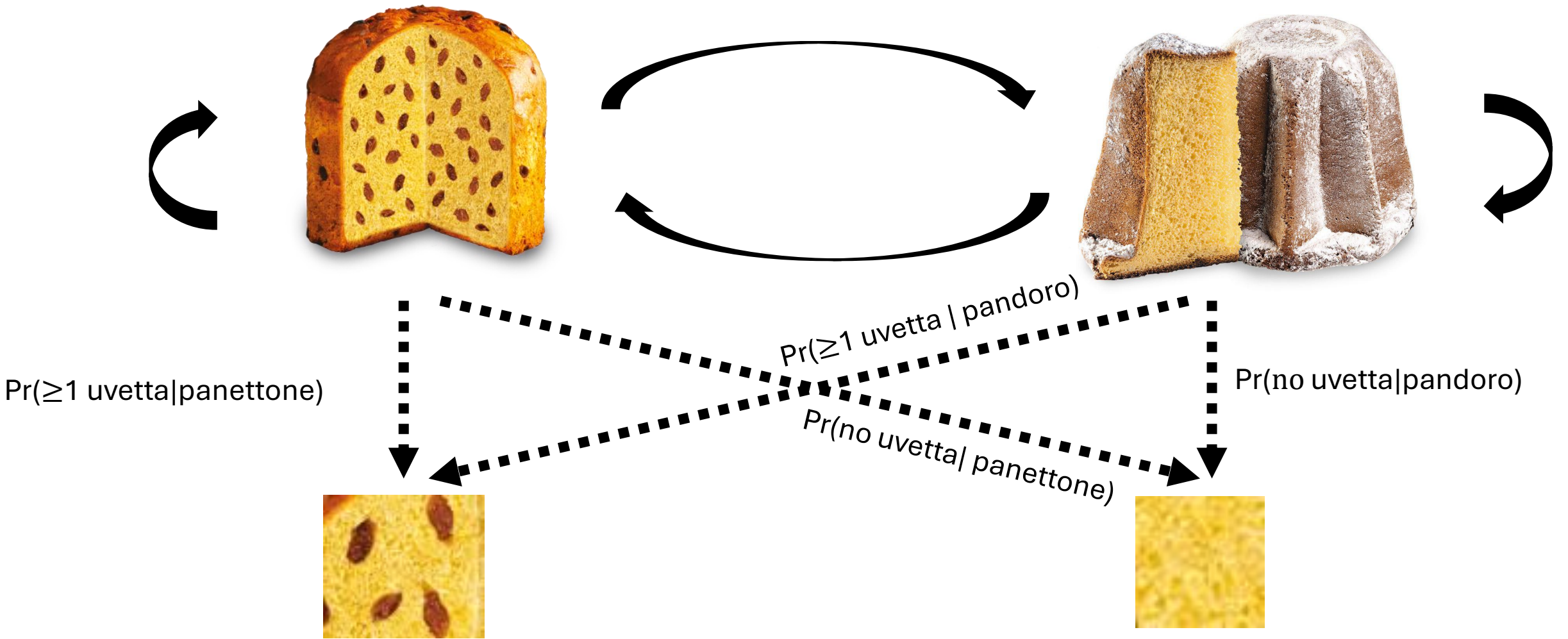


Assumiamo che tra di voi ci sia uno studente americano (che quindi non sa distinguere l'impasto), e folgorato dalla bontà dei dolci della tradizione natalizia italiana si ingozzi mangiando un sacco di fette (quindi ogni tanto dobbiamo aprire un panettone o un pandoro nuovo), e voglia poi capire con esattezza da quale dolce veniva ogni fetta. Come formalizzereste il problema?



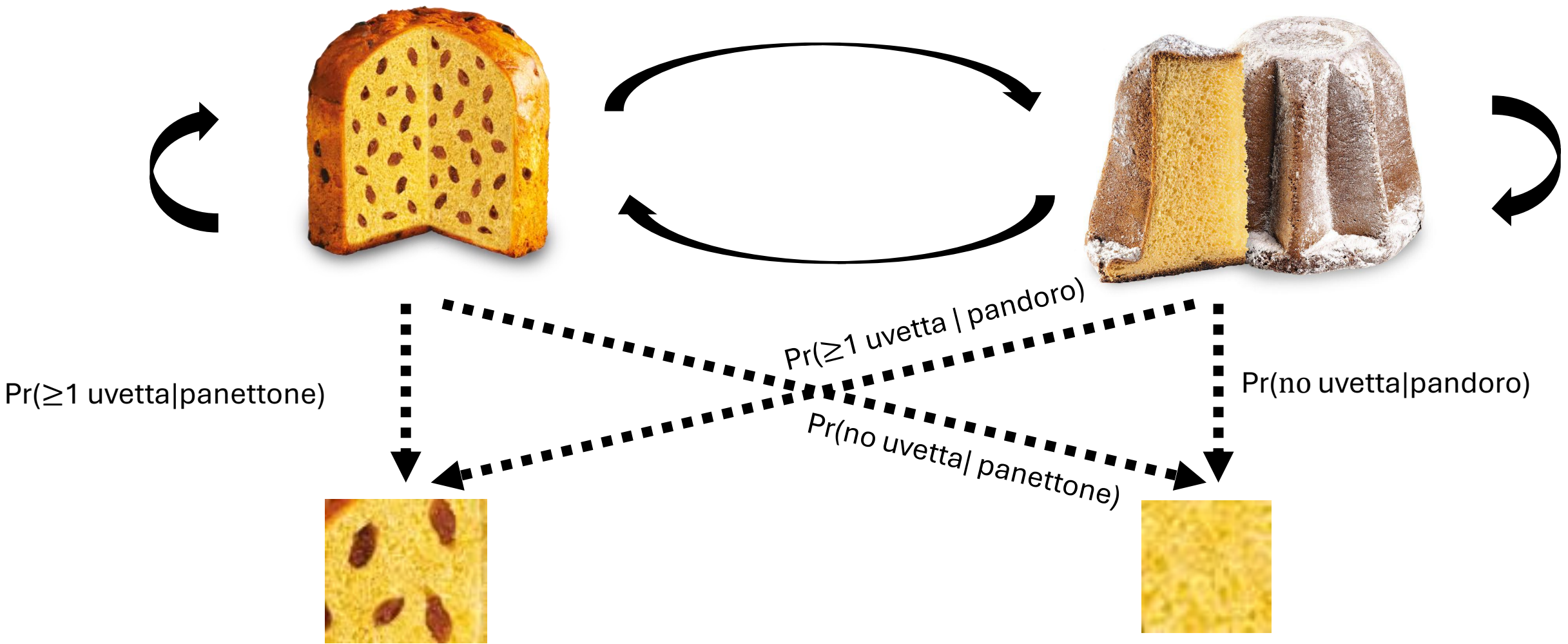
Assumiamo che tra di voi ci sia uno studente americano (che quindi non sa distinguere l'impasto), e folgorato dalla bontà dei dolci della tradizione natalizia italiana si ingozzi mangiando un sacco di fette (quindi ogni tanto dobbiamo aprire un panettone o un pandoro nuovo), e voglia poi capire con esattezza da quale dolce veniva ogni fetta. Come formalizzereste il problema?

Che tipo di **emissioni/probabilità di emissione**?

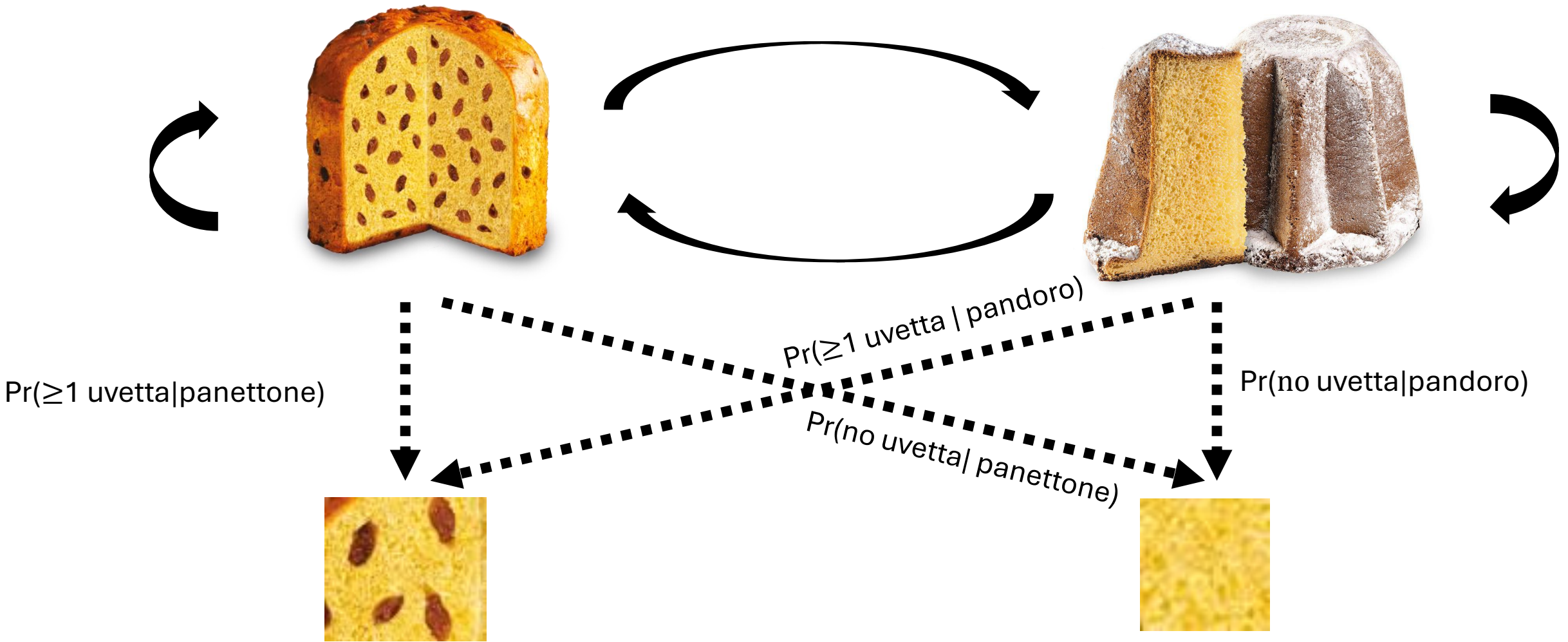


Assumiamo che tra di voi ci sia uno studente americano (che quindi non sa distinguere l'impasto), e folgorato dalla bontà dei dolci della tradizione natalizia italiana si ingozzi mangiando un sacco di fette (quindi ogni tanto dobbiamo aprire un panettone o un pandoro nuovo), e voglia poi capire con esattezza da quale dolce veniva ogni fetta. Come formalizzereste il problema?

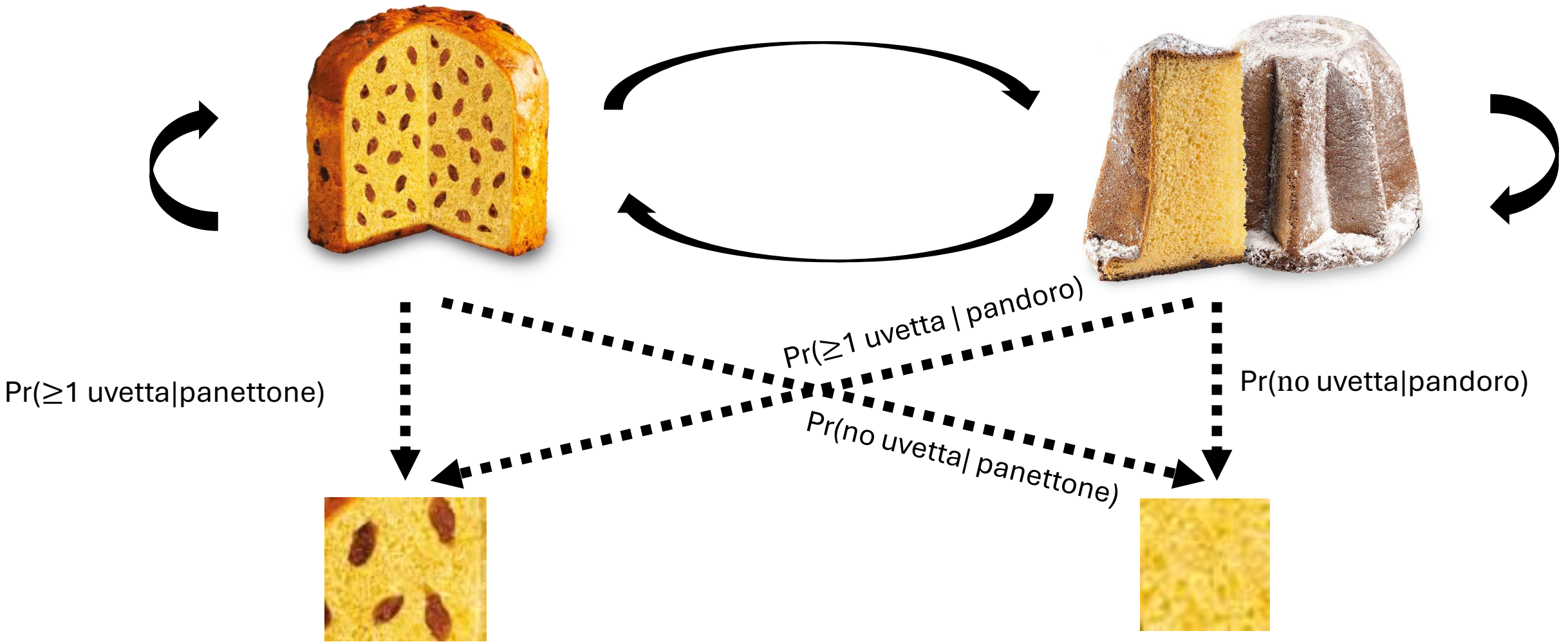
Che tipo di **emissioni/probabilità di emissione**? Emissioni binarie.



Quali problemi/applicazioni abbiamo visto con emissioni binarie?



Quali problemi/applicazioni abbiamo visto con emissioni binarie?
E a quali problemi potreste pensare in questi termini?



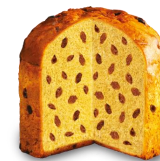
Quali problemi/applicazioni abbiamo visto con emissioni binarie?
E a quali problemi potreste pensare in questi termini?



Isole CpG

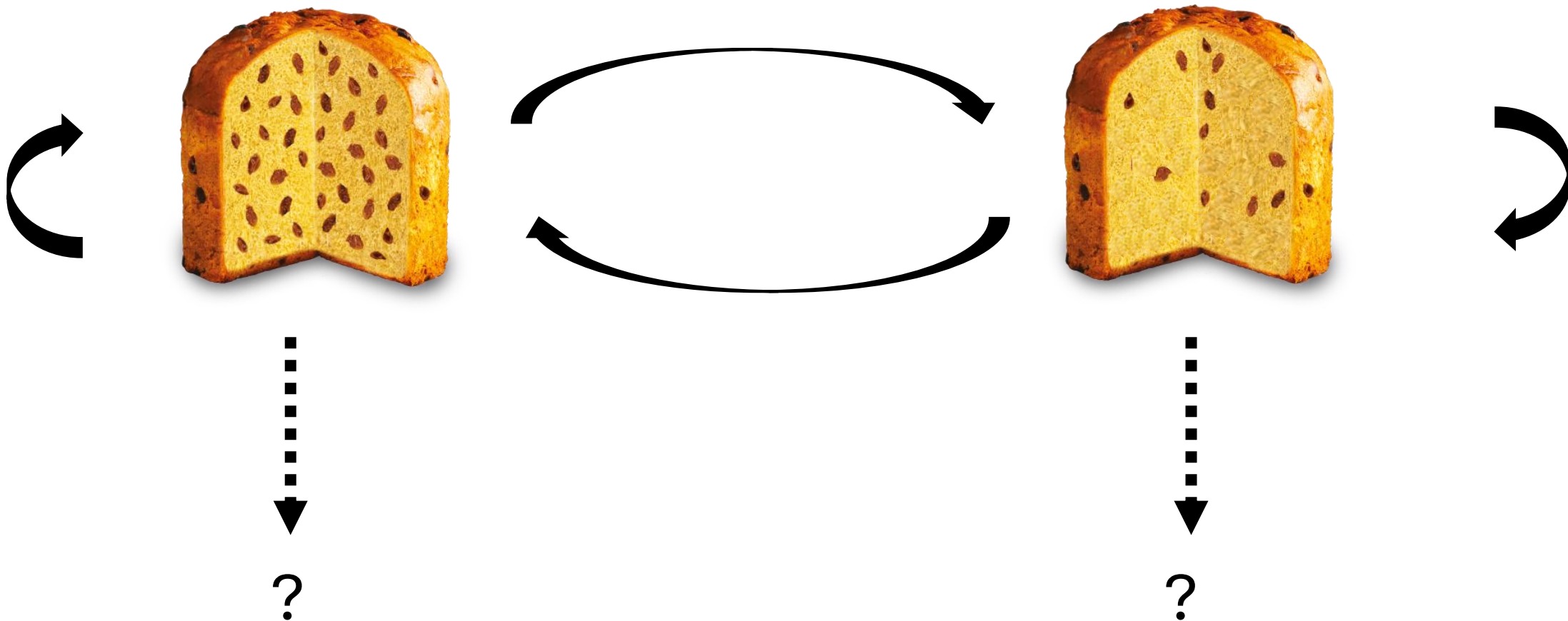


CHROMHMM



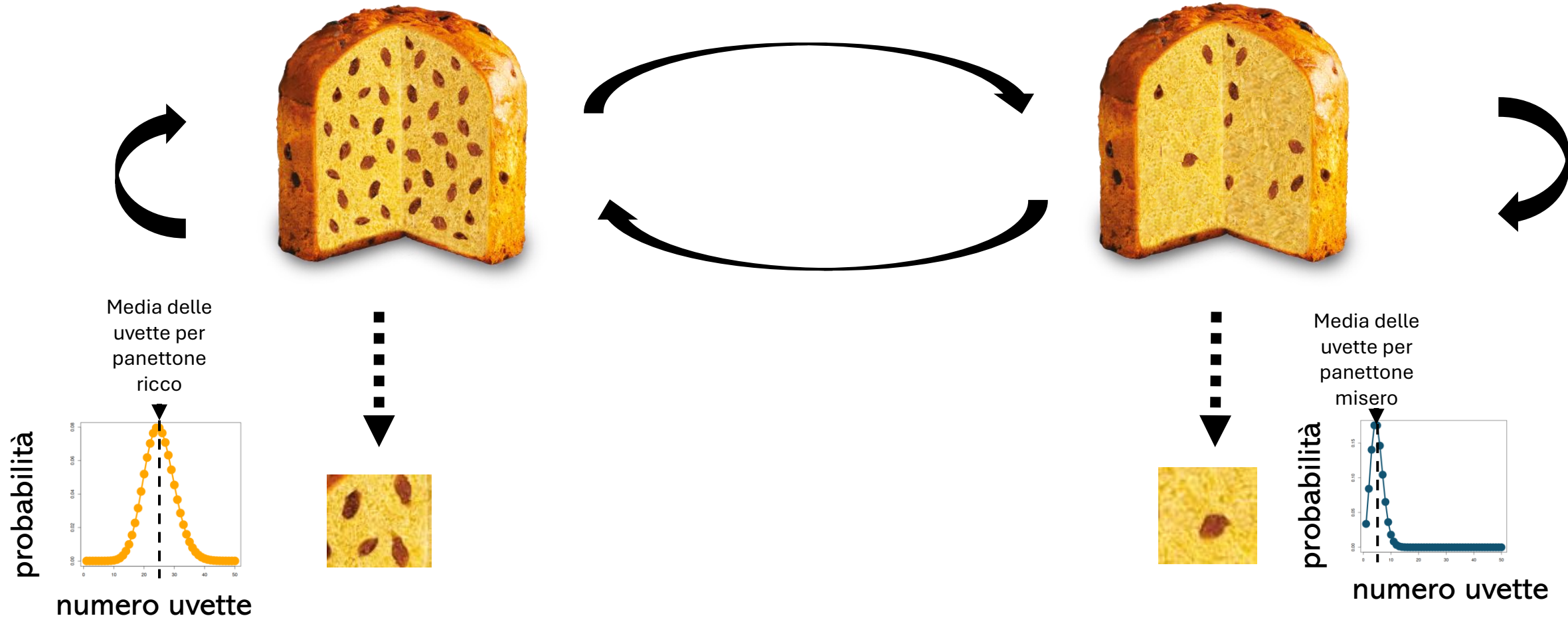
Se fossimo Leo la formica



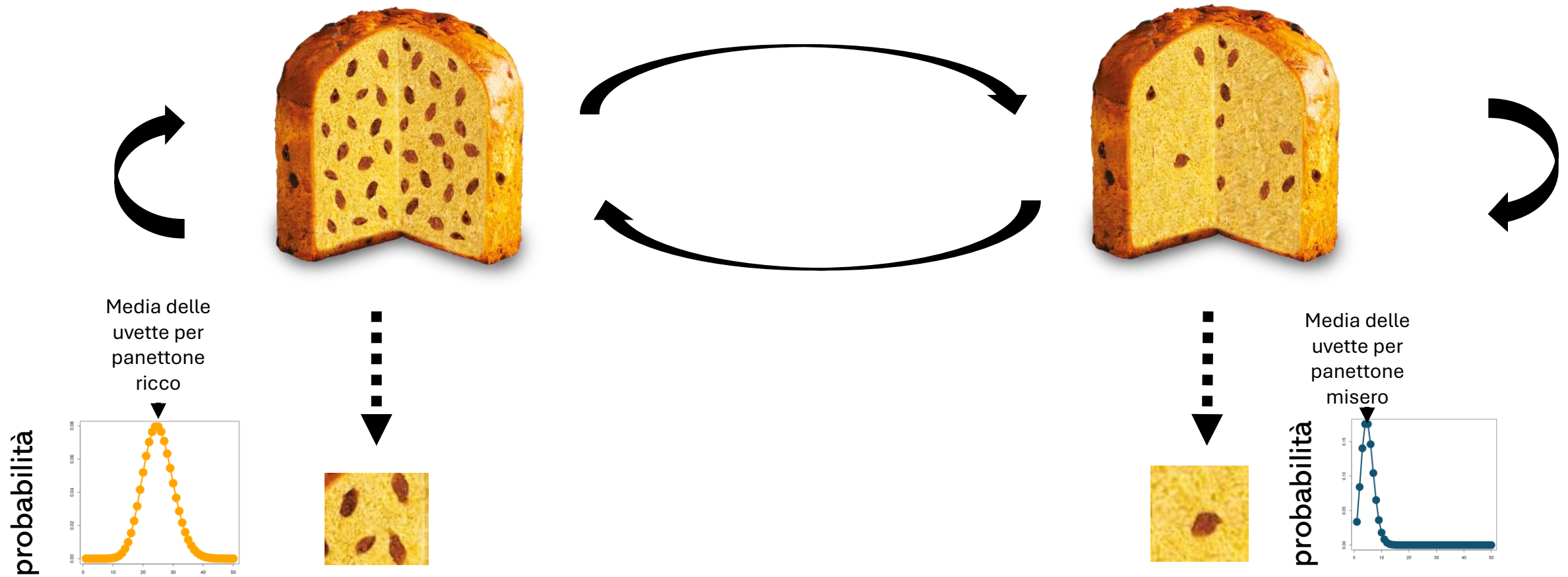


E se invece volessimo distinguere tra due tipi diversi di panettone, uno di buona marca e ricco d'uvetta, e l'altro a marca risparmio?

Che tipo di emissioni/**probabilità di emissione** useremmo?



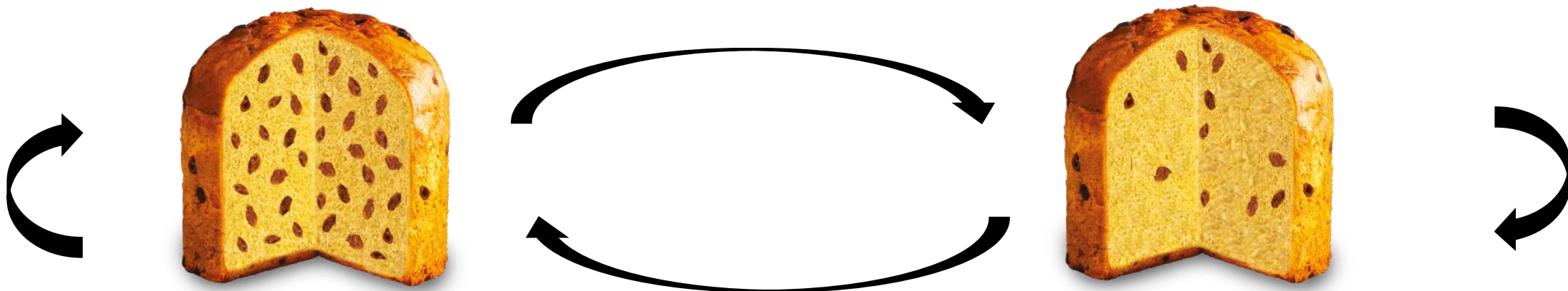
Vogliamo usare il **numero delle uvette**.



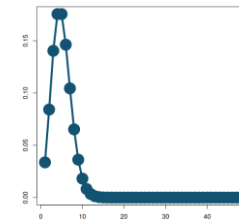
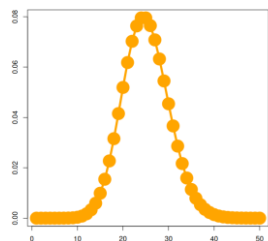
$$\Pr(\text{numero di uvette} \mid \text{tipo di panettone}) \approx \text{Poisson}(\text{densità media delle uvette})$$

Vogliamo usare il **numero delle uvette**. Perché usare una **Poisson**?

- La Poisson è semplicemente la distribuzione di probabilità di campionare un certo numero di successi (uvette) dato un rate o una densità costanti.
- Ha una formula bruttina ($e^{-\lambda} \lambda^k / k!$) su wikipedia ma non ci interessa!
- E' simile ad una Gaussiana (forma a campana) ma per eventi discreti.



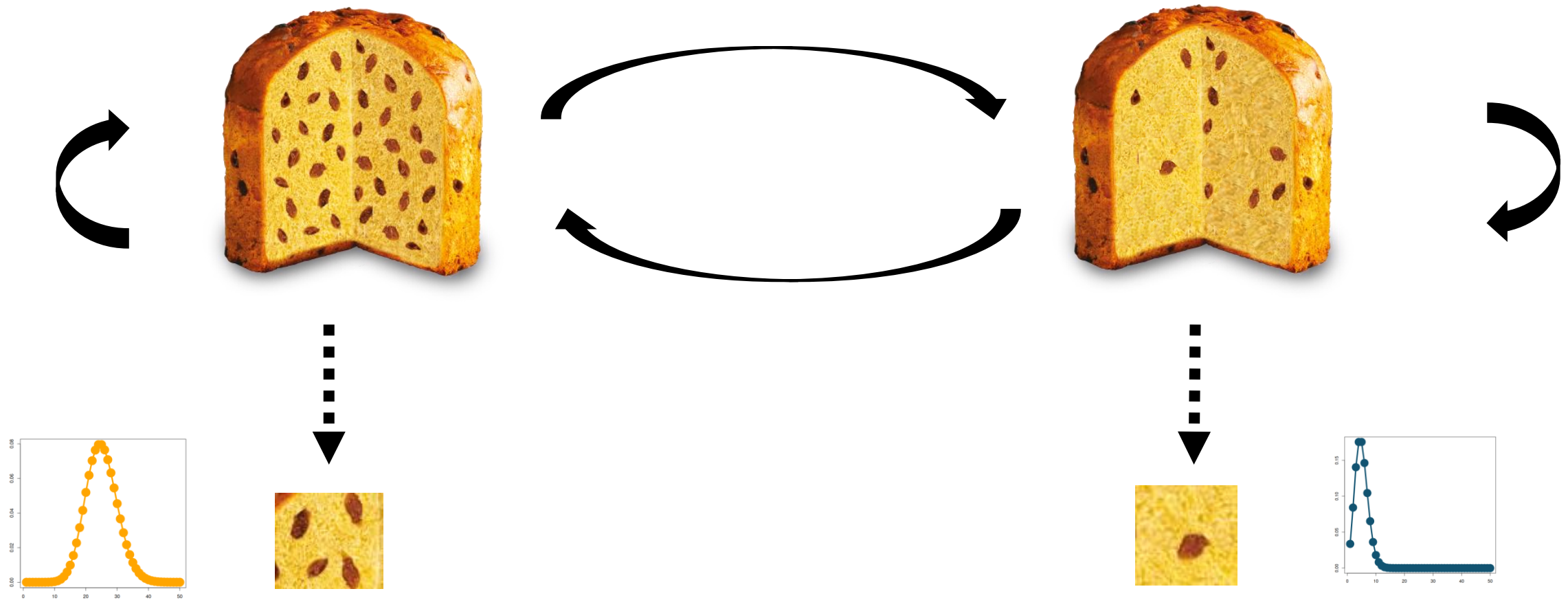
*piu' alta la media, piu' alta la varianza!



$$\text{Pr}(\text{numero di uvette} \mid \text{tipo di panettone}) \approx \text{Poisson}(\text{densità media delle uvette})$$

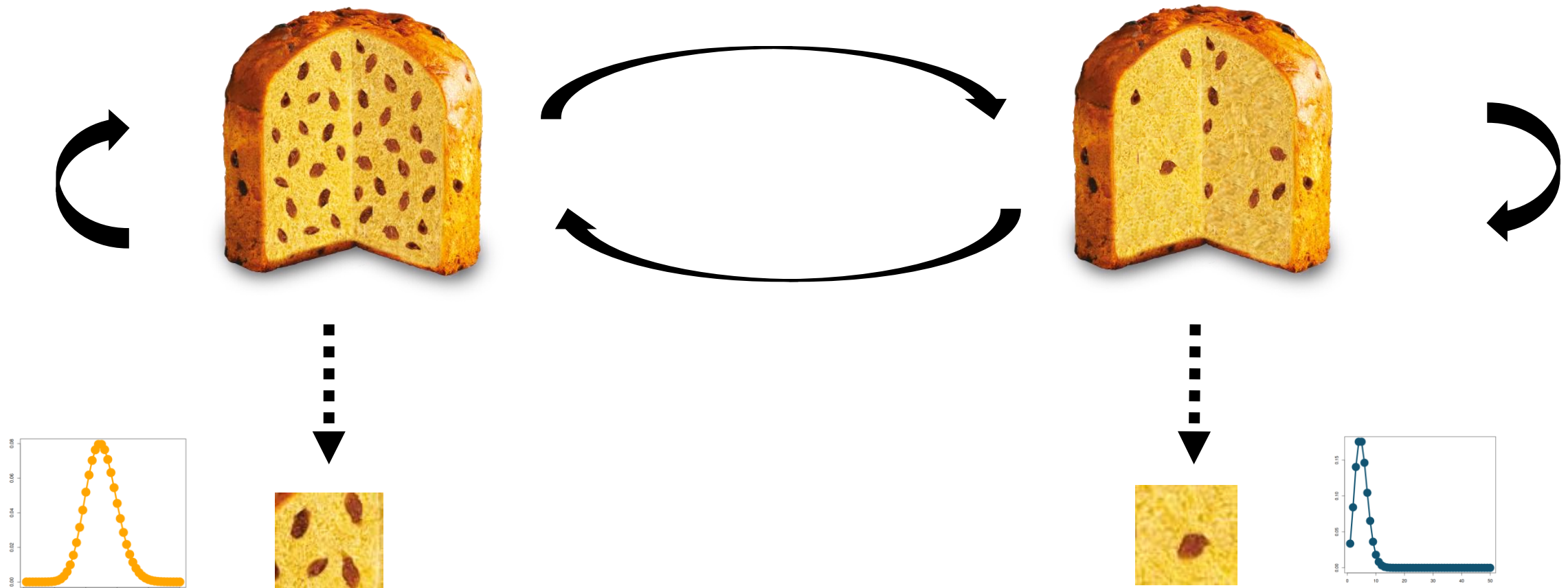
Vogliamo usare il **numero delle uvette**. Perché usare una **Poisson**?

- La Poisson è semplicemente la distribuzione di probabilità di campionare un certo numero di successi (uvette) dato un rate o una densità costanti.
- Ha una formula bruttina ($e^{-\lambda} \lambda^k / k!$) su wikipedia ma non ci interessa!
- E' simile ad una Gaussiana (forma a campana) ma per eventi discreti.
- E' persino piu' semplice di una Gaussiana perché la Gaussiana ha due parametri (media e varianza), la Poisson solo uno (media/rate e varianza sono uguali!), e grazie a questo rappresenta il fenomeno della varianza intorno alla media*



$\text{Pr}(\text{numero di uvette} \mid \text{tipo di panettone}) \simeq \text{Poisson}(\text{densità media delle uvette})$

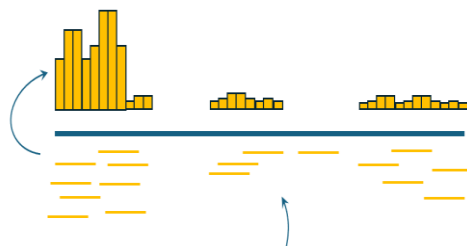
Che tipo di applicazioni ed HMM immaginate con questa struttura?



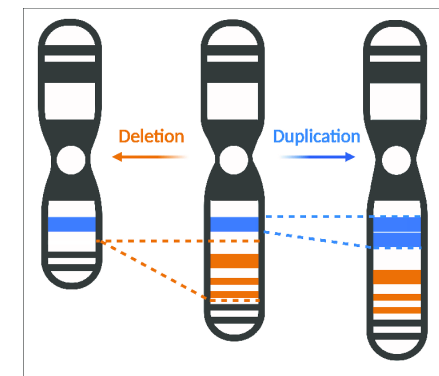
$\Pr(\text{numero di uvette} \mid \text{tipo di panettone}) \approx \text{Poisson}(\text{densità media delle uvette})$

Che tipo di applicazioni ed HMM immaginate con questa struttura?

Numero delle reads in **peak calling**
(chip-seq, ATACseq,
Methylated DNA Immunoprecipitation
sequencing, etc.)

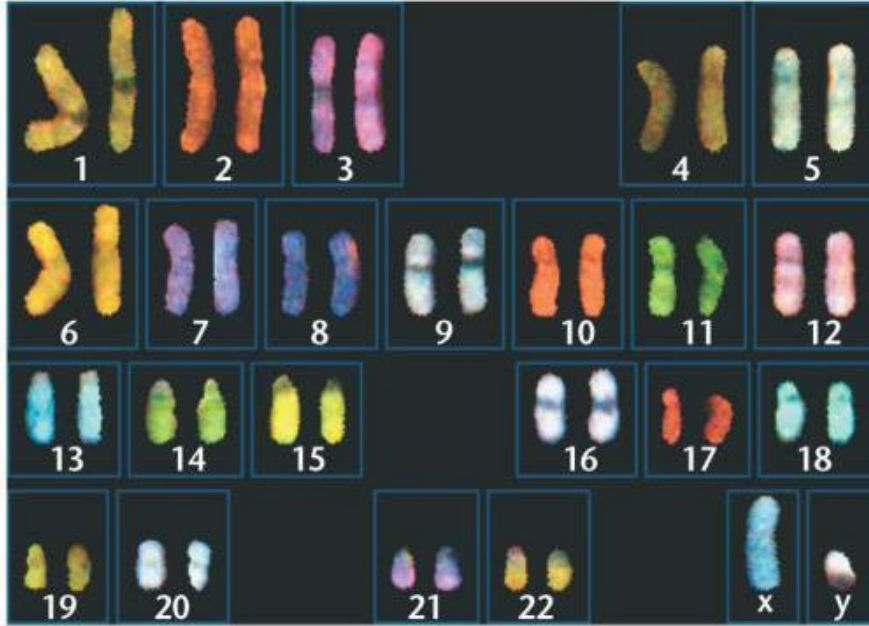


Numero delle reads per
copy number variation o
CNV (HMMcopy per
delezioni e duplicazioni)

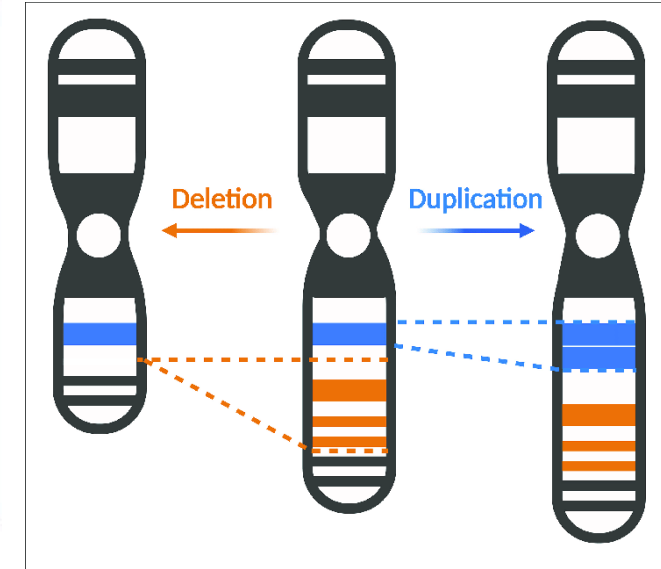
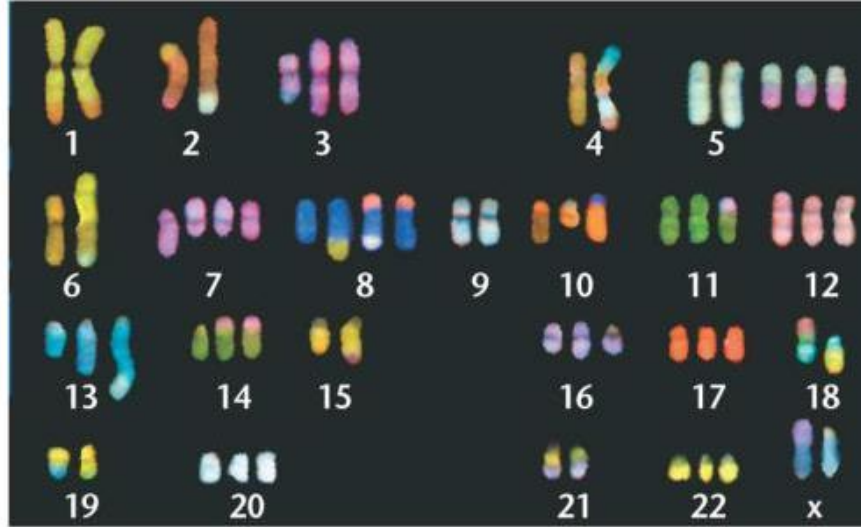


Copy number variation (CNV) e tumori

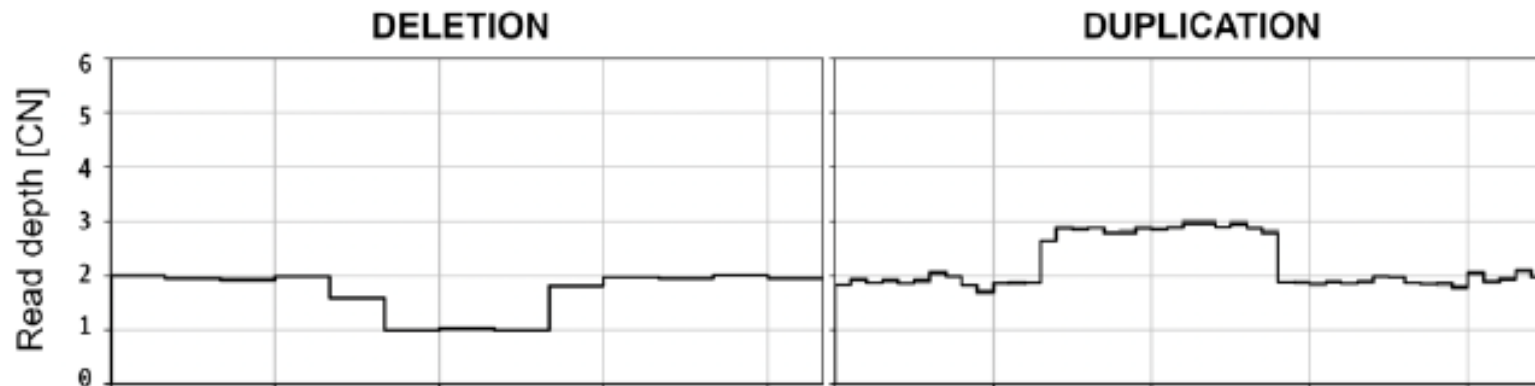
(a) cariotipo normale

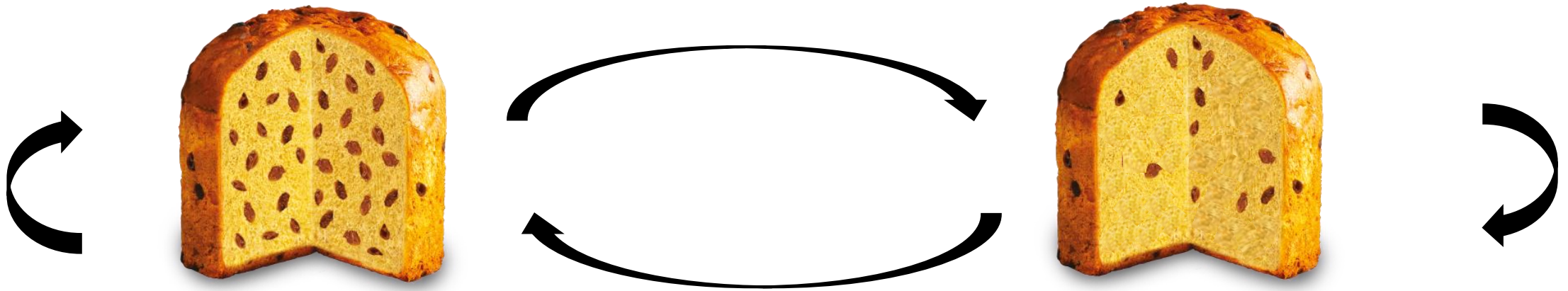


(b) cariotipo tumorale

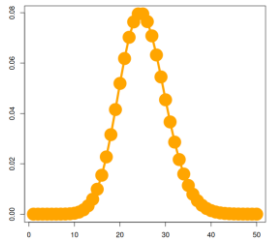


Copyright © 2006 Pearson Prentice Hall, Inc.

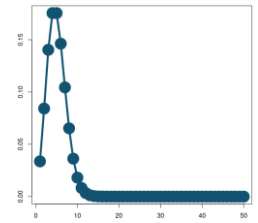




*piu' alta la media, piu' alta la varianza!



$$\text{Pr}(\text{numero di uvette} \mid \text{tipo di panettone}) \simeq \text{Poisson}(\text{densità media delle uvette})$$



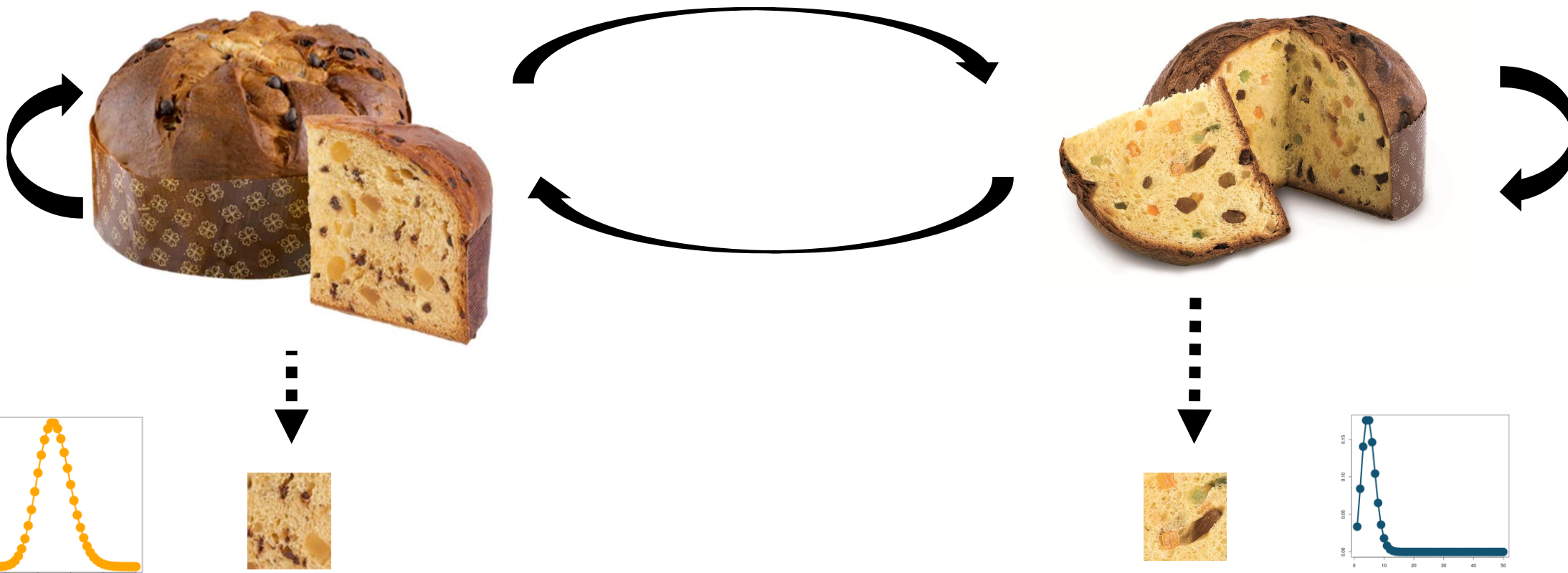
NON IMPORTANTE: Esistono varianti della Poisson che potrete incontrare in futuro con varie proprietà statistiche (ad esempio la **negative binomial** per dare permettere piu' varianza) ma vedetele solo come Poisson un pochino piu' sofisticate, nulla piu'!



- E' persino piu' semplice di una Gaussiana perché la Gaussiana ha due parametri (media e varianza), la Poisson solo uno (media/rate e varianza sono uguali!), e grazie a questo rappresenta il fenomeno della varianza intorno alla media*

Ma nei panettoni ci sono anche i canditi!!





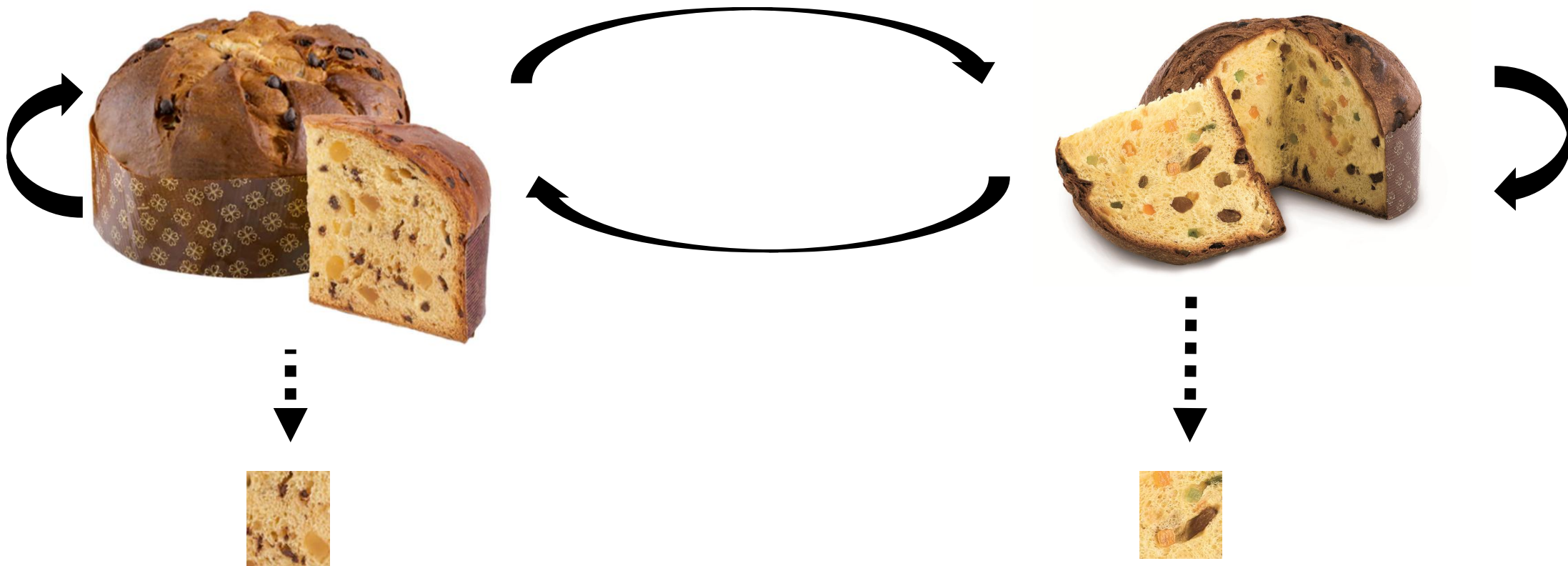
*piu' alta la media, piu' alta la varianza!

$$\text{Pr}(\text{numero di uvette} \mid \text{tipo di panettone}) \approx \text{Poisson}(\text{densità media delle uvette})$$

NON IMPORTANTE: Esistono varianti della Poisson che potrete incontrare in futuro con varie proprietà statistiche (ad esempio la **negative binomial** per dare permettere piu' varianza) ma vedetele solo come Poisson un pochino piu' sofisticate, nulla piu'!

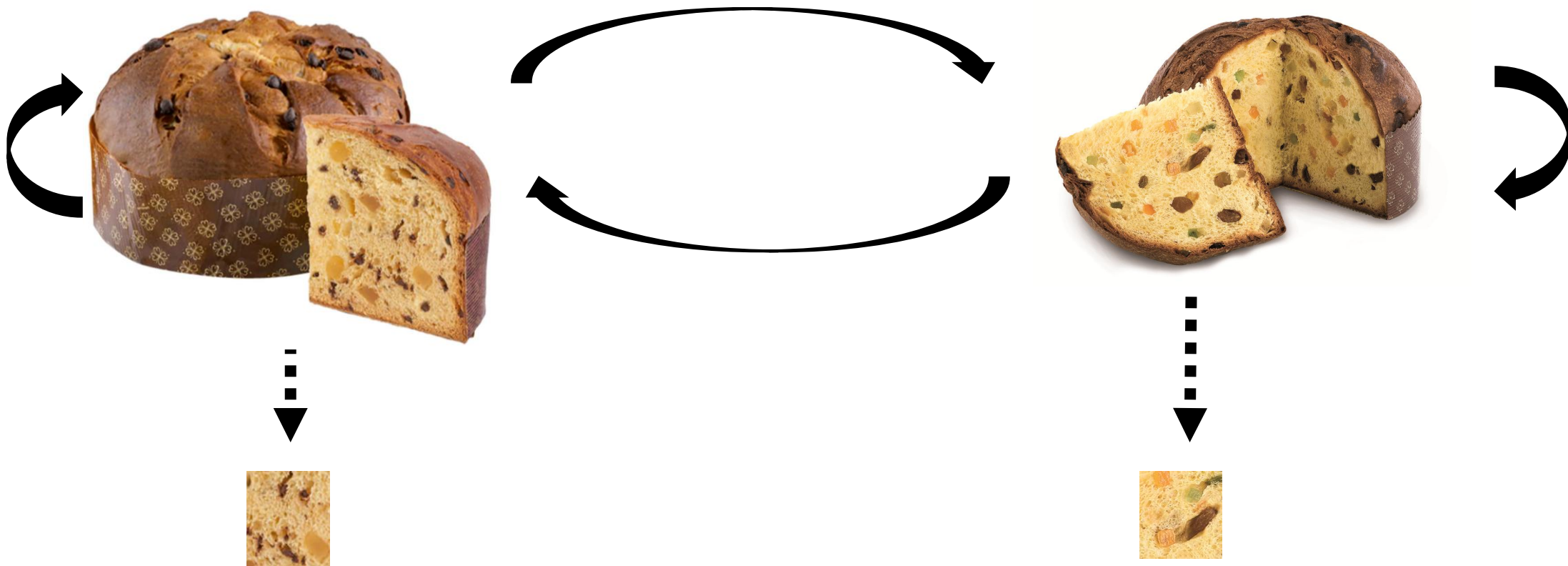


- E' persino piu' semplice di una Gaussiana perché la Gaussiana ha due parametri (media e varianza), la Poisson solo uno (media/rate e varianza sono uguali!), e grazie a questo rappresenta il fenomeno della varianza intorno alla media*



Diverse proporzioni di uvetta verso canditi in due marche diverse di panettoni.

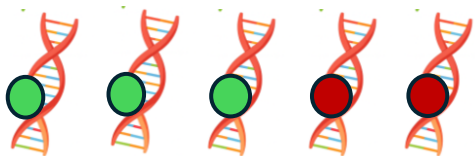
Come li distinguereste ora?



Diverse proporzioni di uvetta verso canditi in due marche diverse di panettoni.

Come li distinguereste ora?

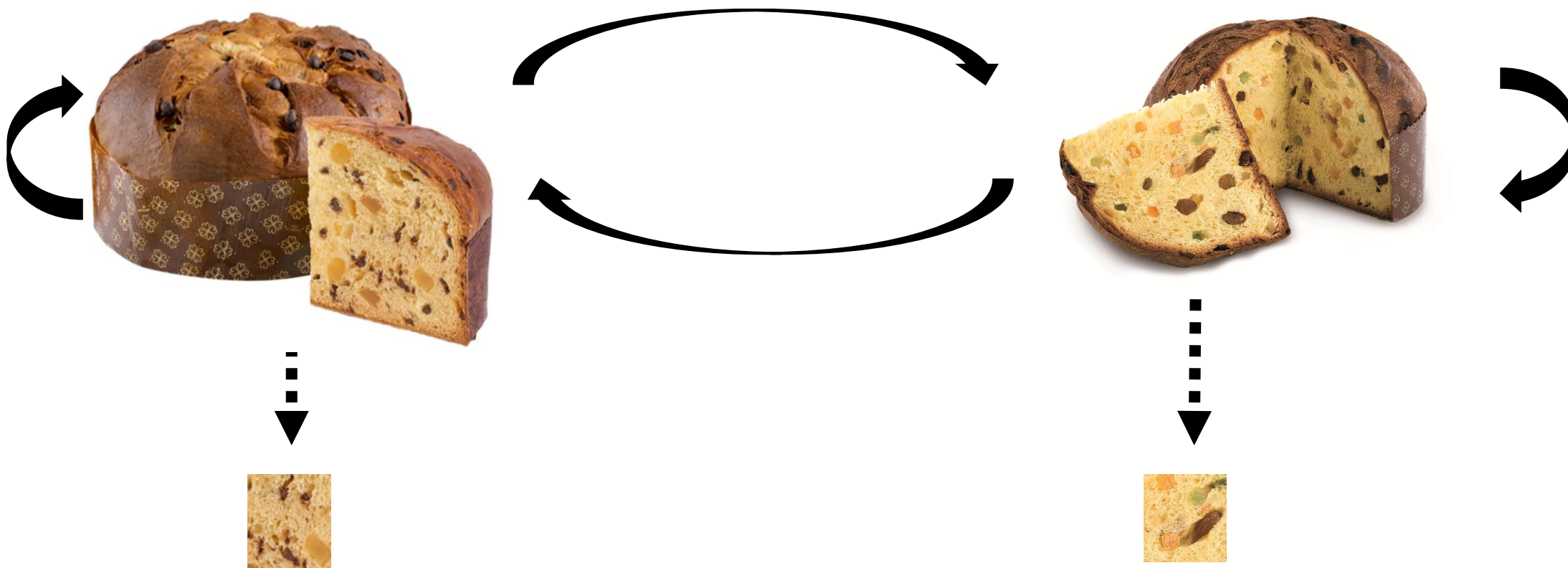
Una binomiale!!!!



uvetta



canditi



Diverse proporzioni di uvetta verso canditi in due marche diverse di panettoni.

Metilazione

Bisulfite sequencing:

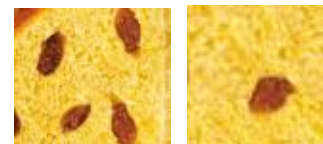
Differenti proporzioni di C e T in posizioni dove sul genoma di referenza ci attenderemo una C



Methylated DNA

Immunoprecipitation sequencing:

Piu' simile ad una chip-seq. Abbiamo picchi di molte reads vs poche reads



Metilazione del DNA e bisulfite sequencing

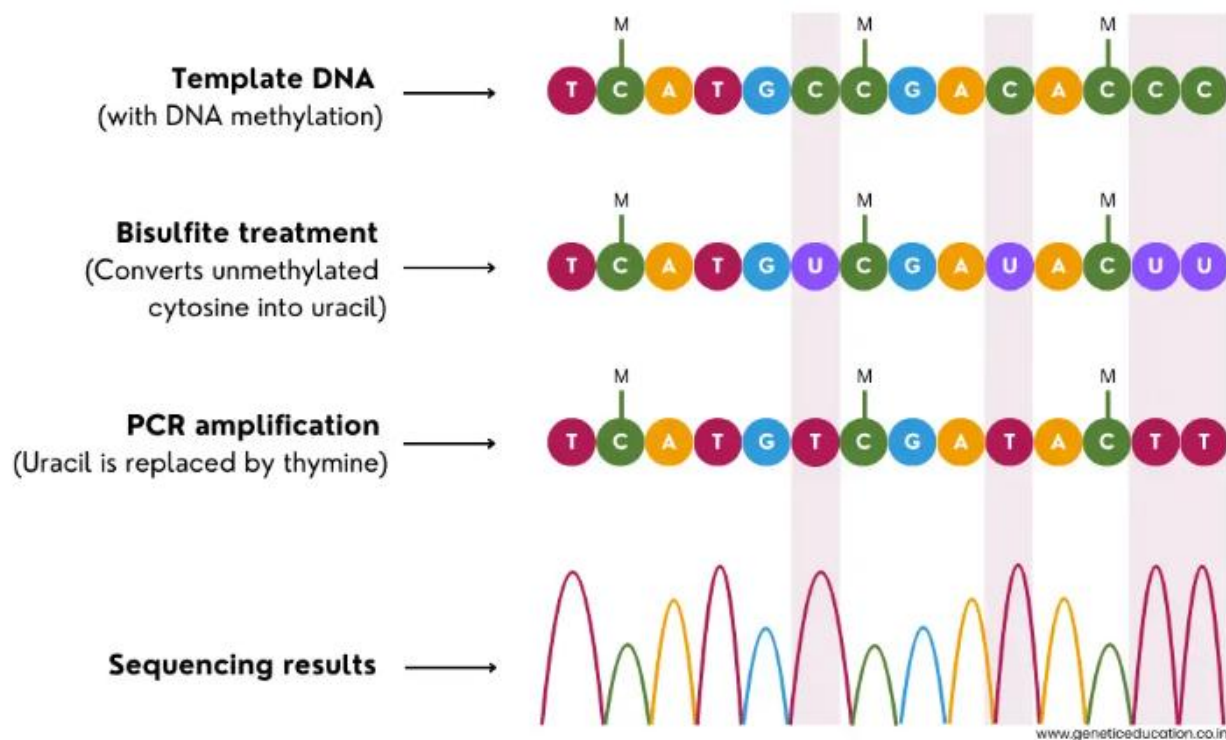
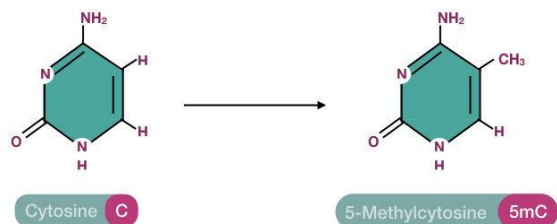
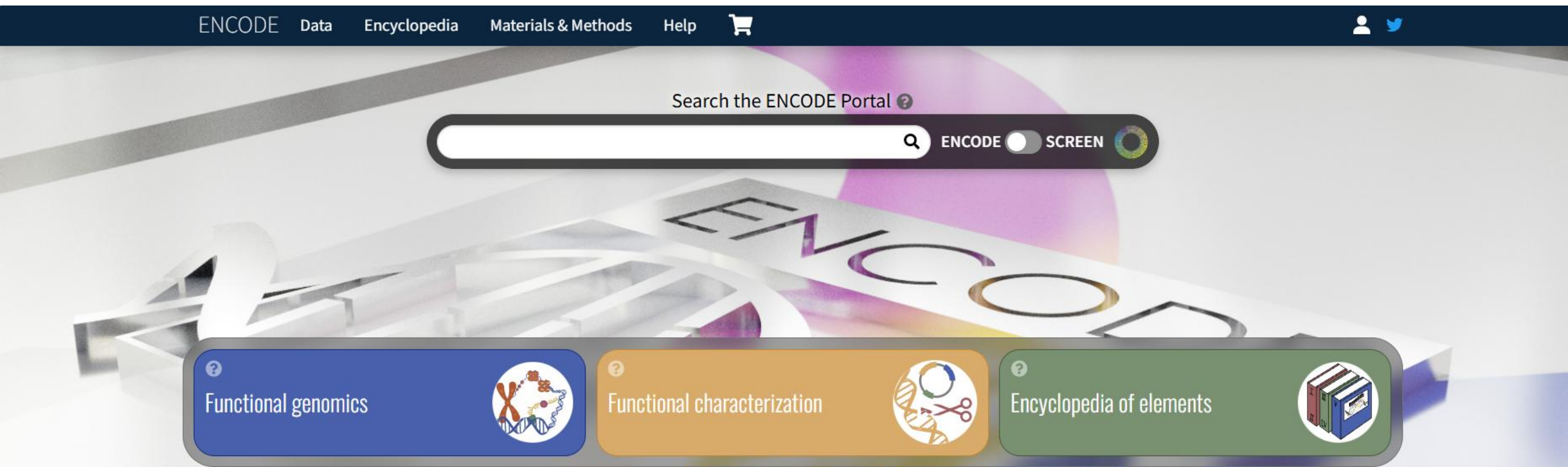


Illustration of the complete bisulfite sequencing process.

In regioni metilate avremo una proporzione piu alta di sostituzioni C>T nelle reads mappate sulla referenza

Utilizzo del portale **ENCODE** per ottenere dati epigenomici e non solo

- The Encyclopedia of DNA Elements (ENCODE) contiene dati per oltre 10.000 esperimenti che investigano l'espressione genica e l'epigenoma umano e di altre specie modello
- Collegamenti diretti al The Cancer Genome Atlas (TCGA) e altri
- Centinaia di tipi cellulari e tessuti (ad esempio, anche solo per il topo, piu' di 100)





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

Search the ENCODE Portal ?



ENCODE



SCREEN



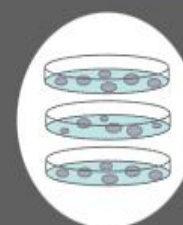
Functiona



Characterization



Encyclopedia of elements

EN-TE_x

Deeply profiled cell lines

Biosample

Organism

Homo sapiens14750

Mus musculus2675

Drosophila melanogaster1058

Caenorhabditis elegans612

Drosophila pseudoobscura4

Biosample classification

Biosample

Organ

Cell

Systems

Sex

Life stage

Disease

Biosample treatment

Cellular component

Assay

Assay type

Assay title

Search

TF ChIP-seq4690

Histone ChIP-seq3694

DNase-seq3425

total RNA-seq980

Mint-ChIP-seq895

polyA plus RNA-seq784

ATAC-seq464

microRNA-seq412

snRNA-seq395

snATAC-seq364

Number of displayed results:

25

50

100

200

Add all items to cart

DNase-seq in activated naive CD4-positive, alpha-beta T cell

Experiment

Homo sapiens activated naive CD4-positive, alpha-beta T cell male adult (43 years) treated with 10 ng/mL Interleukin-2 for 5 days, anti-CD3 and anti-CD28 coated beads for 7 days

ENCSR856UND

released

Lab: John Stamatoyannopoulos, UW

Project: ENCODE

Treatments: 10 ng/mL Interleukin-2 (UniProtKB:P60568) for 5 days, anti-CD3 and anti-CD28 coated beads for 7 days

polyA plus RNA-seq in GM20787

Experiment

Homo sapiens GM20787

ENCSR150XQV

Analysis

Genome assembly

Available file types

fastq4158

bam4140

bed narrowPeak4133

bigWig4133

bigBed narrowPeak4126

Read length (nt)

Mapped read length (nt)

Potete scegliere organismo, tipo di esperimento e tipo cellulare, oltre che formato di file

Experiment search

Clear all selections ✕

Assay

Assay type

Assay title

Search

Histone ChIP-seq18

ATAC-seq10

WGBS10

in situ Hi-C9

total RNA-seq9

Control ChIP-seq6

ATAC-seq ✕

Histone ChIP-seq ✕

All ✕

Target category

Target of assay

Search

H3K27ac6

H3K4me16

H3K4me36

Showing 25 of 28 results

Report

Experiment matrix

Download

Visualize

{;}

Number of displayed results:

25

50

100

200

Add all items to cart

Histone ChIP-seq in motor neuron

Homo sapiens from a donor with amyotrophic lateral sclerosis; motor neuron

Target: H3K4me1

Lab: Michael Snyder, Stanford

Project: ENCODE

Experiment

ENCSR698HPK

released

33

Histone ChIP-seq in motor neuron

Homo sapiens from a donor with amyotrophic lateral sclerosis; motor neuron

Target: H3K27ac

Lab: Michael Snyder, Stanford

Project: ENCODE

Experiment

ENCSR425FUS

released

13

Histone ChIP-seq in motor neuron

Homo sapiens from a donor with amyotrophic lateral sclerosis; motor neuron

Target: H3K4me1

Lab: Michael Snyder, Stanford

Project: ENCODE

Experiment

ENCSR358AOC

released

4

ENCFF732BWD  	bam	unfiltered alignments	1	100	GRCh38	2021-08-18	6.56 GB	 released
ENCFF722KEW  	bam	unfiltered alignments	2	100	GRCh38	2021-08-18	4.92 GB	 released
ENCFF869FUW  	bam	alignments	2	100	GRCh38	2021-08-18	3.22 GB	 released

ENCFF742DEP

Open link in new tab

Open link in new window

Open link in InPrivate window

Open link in split screen window

Save link as

Copy link

Visual Search

Alt+Shift+S

Toucan - Language Learning

More tools

Inspect

GRCh38

2021-08-18

3.69 MB

released

GRCh38

2021-08-18

3.54 MB

released

GRCh38

2021-08-18

1.31 GB

released

GRCh38

2021-08-18

1.76 MB

released

100

GRCh38

2021-08-18

6.56 GB

released

100

GRCh38

2021-08-18

4.92 GB

released

100

GRCh38

2021-08-18

3.22 GB

released

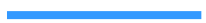
Caratteristiche del DNA antico



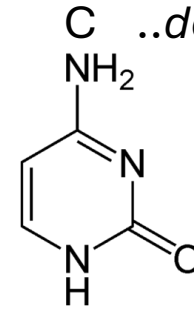
DNA endogeno (Neandertal)



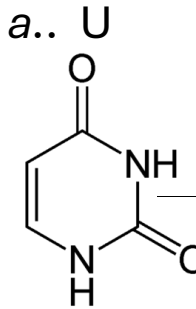
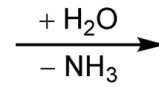
DNA contaminante (*H.sapiens*)



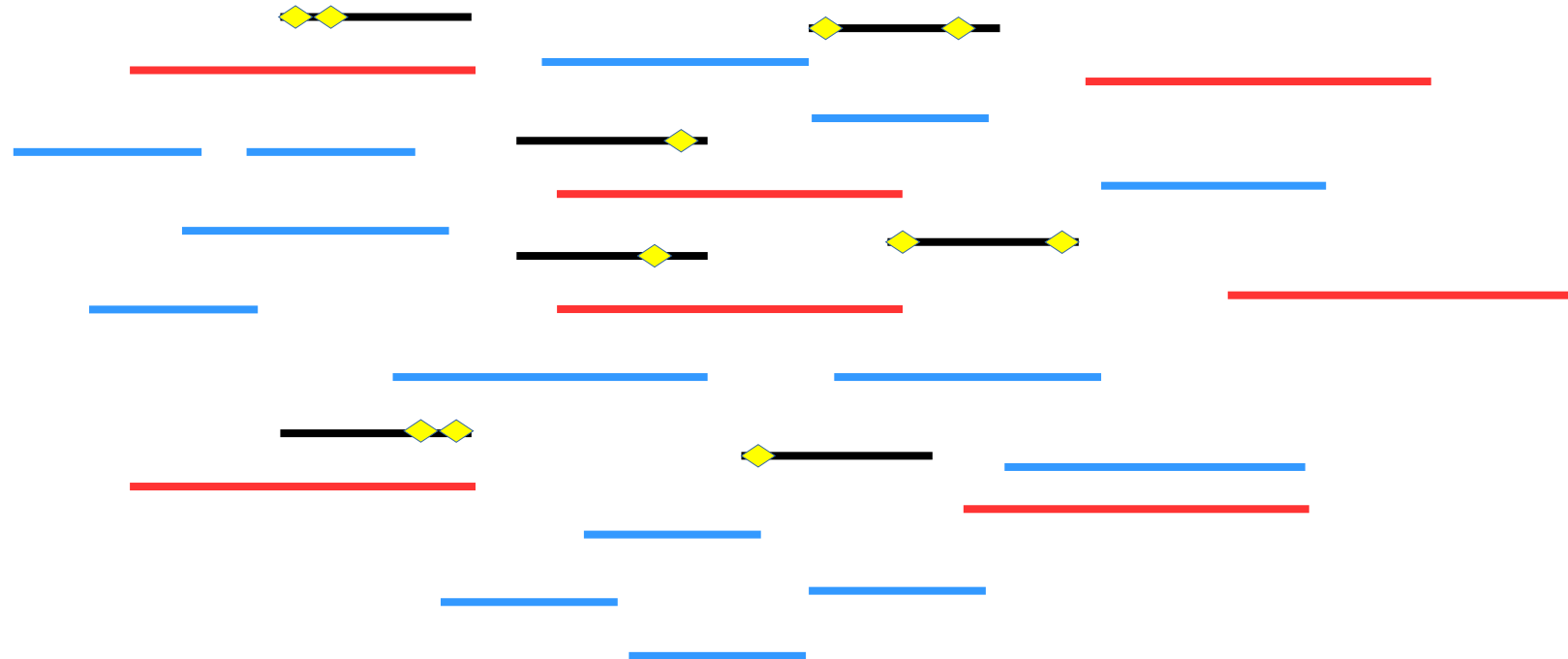
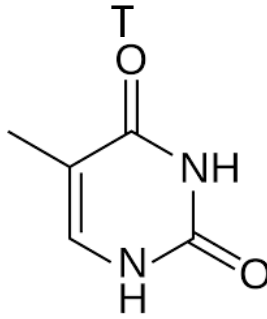
DNA batterico (ed altro)



..deaminato a.. U



..e letto come..



Referenza



...

C

C

A

T

C

C

A

C

G

A

T

A

C

T

A

T

T

T

C

C

...

...

T

T

A

T

T

C

A

C

G

A

T

G

C

T

A

T

T

T

C

C

...

...

C

C

A

T

C

C

A

G

G

A

T

A

C

T

A

T

T

T

C

C

...

5' (inizio sequenza)

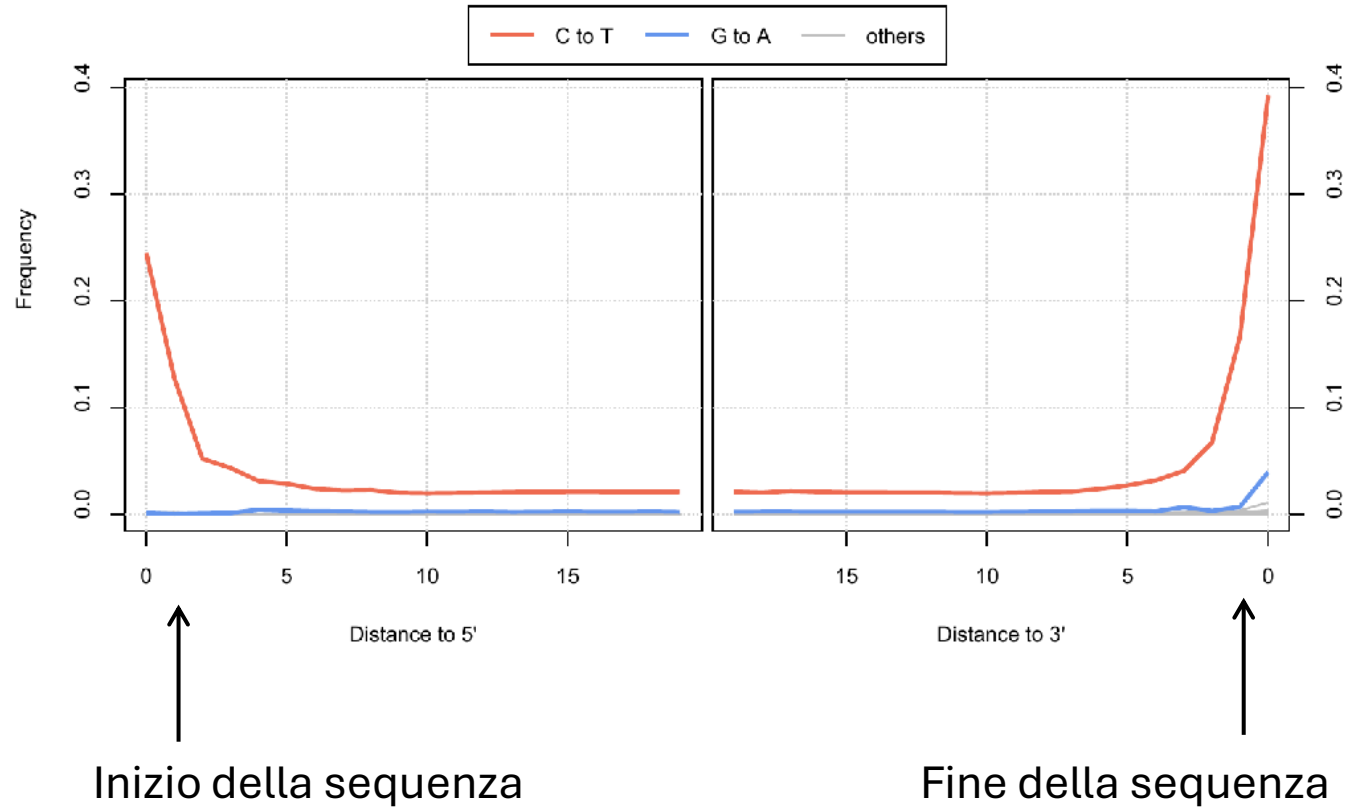
Legenda:

Deaminazioni C->T in DNA antico

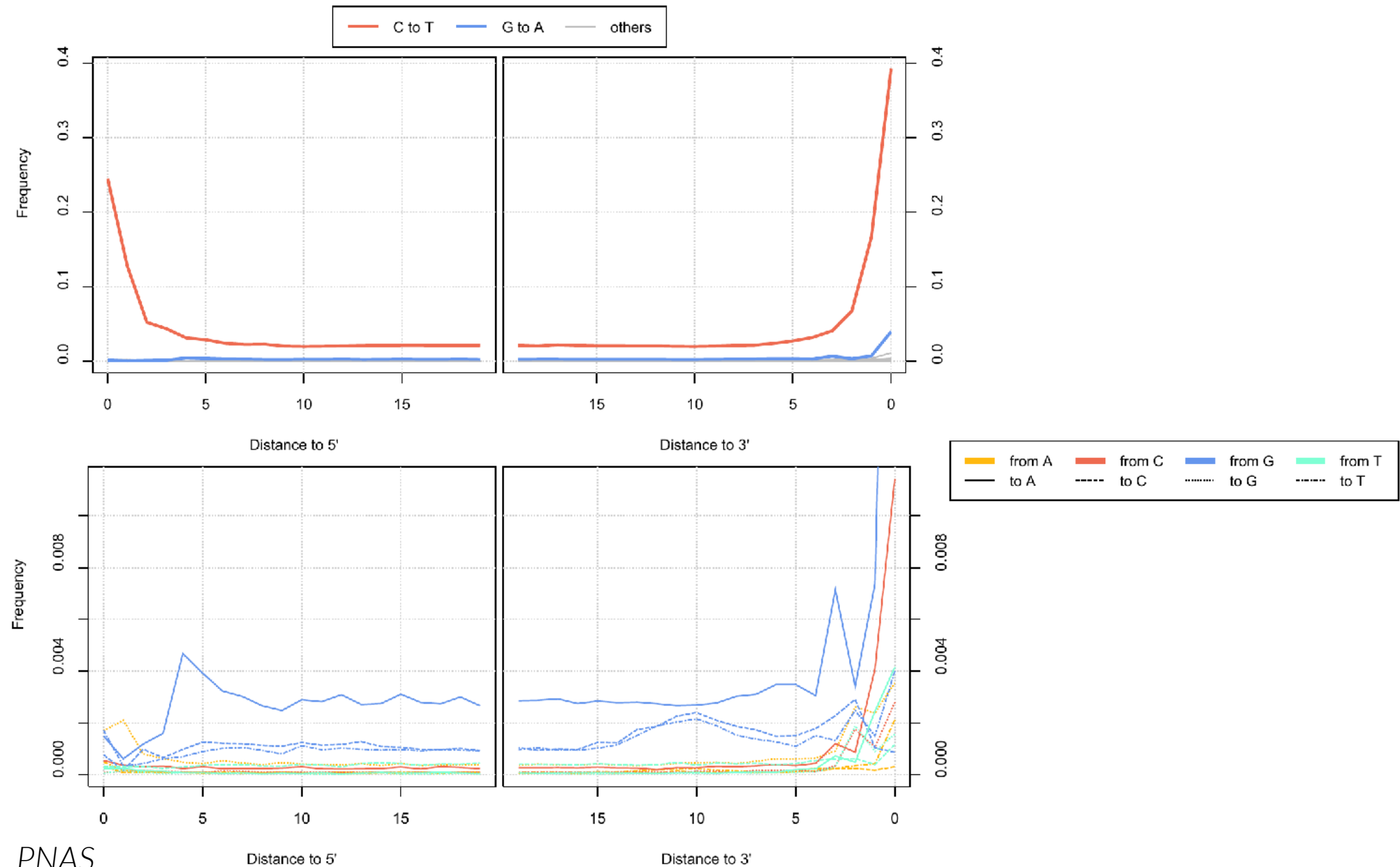
Altre sostituzioni

Deaminazione del DNA antico

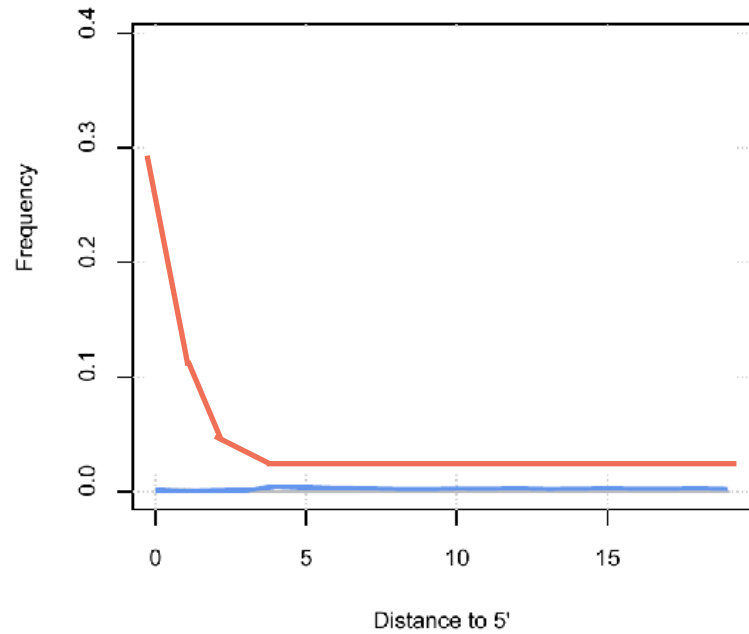
Proporzione di posizioni per le quali la referenza ha una **C** e il campione una **T**



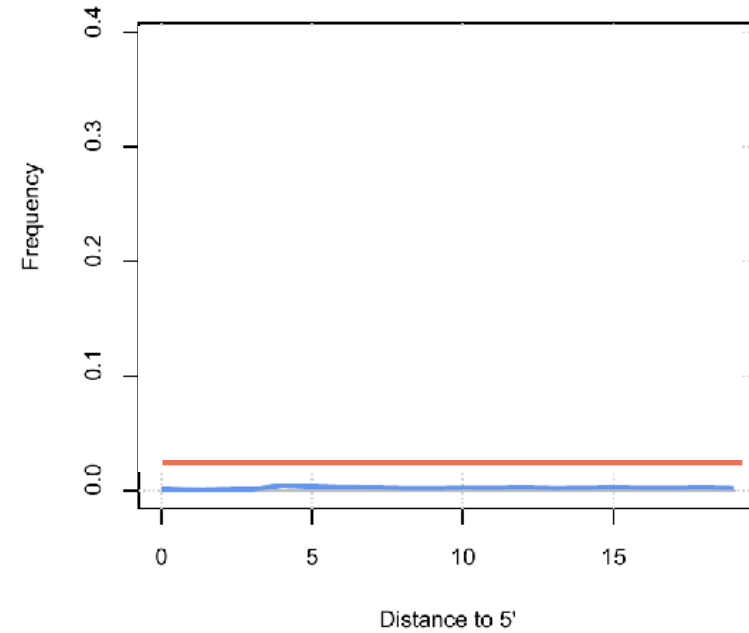
Deaminazione del DNA antico



La deaminazione permette di riconoscere il DNA antico



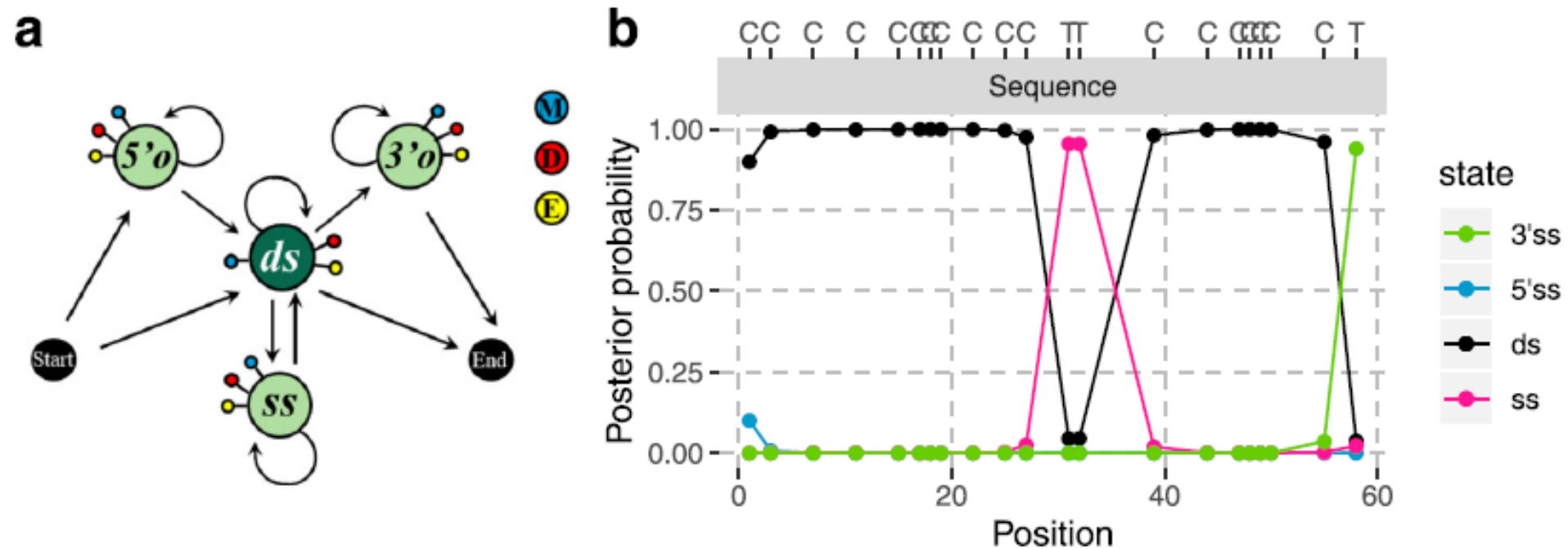
DNA antico
endogeno



DNA moderno
contaminante



Modello di AuthentiCT (Peyregne et al., 2020, *Genome Biology*)



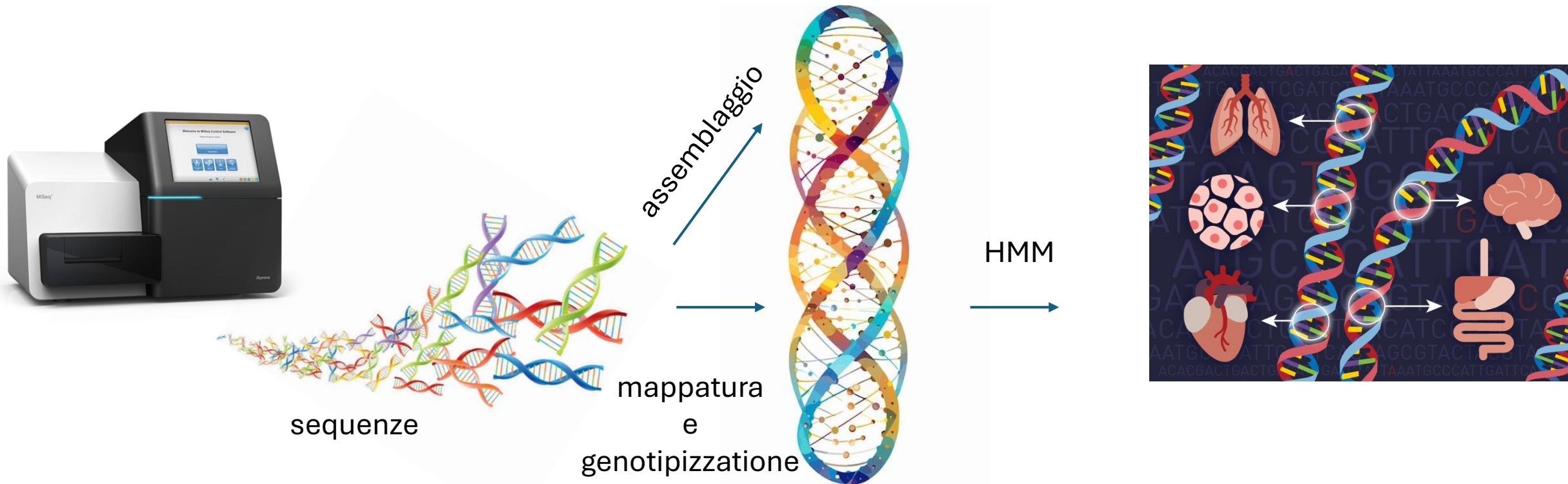
HMM a 4 stati, dove il DNA a singolo filamento (3'-terminale, 5-terminale o interno) viene definito come 3 stati (con maggiore probabilità di osservare una C>T) in aggiunta al DNA a doppio filamento.

Complimenti, potete costruire genomi e genotipi di piu' individui partire a partire da sequenze di DNA!

Dal sequenziatore..

..al genoma..

..alla funzione



Formati di file bioinformatici

.fastq

.fasta, .bam, .vcf

.bed

Il formato .bed

Cromosoma	inizio	fine	Campi addizionali
chr21	0	8219800	E1
chr21	8219800	8220200	E2
chr21	8220200	8234200	E1
chr21	8234200	8234800	E2
chr21	8234800	8420000	E1
chr21	8420000	8420600	E2
chr21	8420600	14026800	E1
chr21	14026800	14027800	E2
chr21	14027800	14080400	E1
chr21	14080400	14081400	E2
chr21	14081400	14216000	E1
chr21	14216000	14216400	E2
chr21	14216400	14381800	E1
chr21	14381800	14385000	E2
chr21	14385000	14451800	E1
chr21	14451800	14455400	E2
chr21	14455400	14477200	E1
chr21	14477200	14477800	E2

Il formato .bed

Cromosoma	inizio	fine	Campi addizionali
chr21	0	8219800	E1
chr21	8219800	8220200	E2
chr21	8220200	8234200	E1
chr21	8234200	8234800	E2
chr21	8234800	8420000	E1
chr21	8420000	8420600	E2
chr21	8420600	14026800	E1
chr21	14026800	14027800	E2
chr21	14027800	14080400	E1
chr21	14080400	14081400	E2
chr21	14081400	14216000	E1
chr21	14216000	14216400	E2
chr21	14216400	14381800	E1
chr21	14381800	14385000	E2
chr21	14385000	14451800	E1
chr21	14451800	14455400	E2
chr21	14455400	14477200	E1
chr21	14477200	14477800	E2

La posizione di inizio è definita 0-based, cioè come POSIZIONE ATTUALE -1. Ad esempio, un singolo SNP in posizioni 10 del cromosoma 6 è definito come:

chr6 9 10

In generale, una posizione 1-based (come quella della fine) è definita come

POSIZIONE_0_BASED= POSIZIONE_1_BASED-1

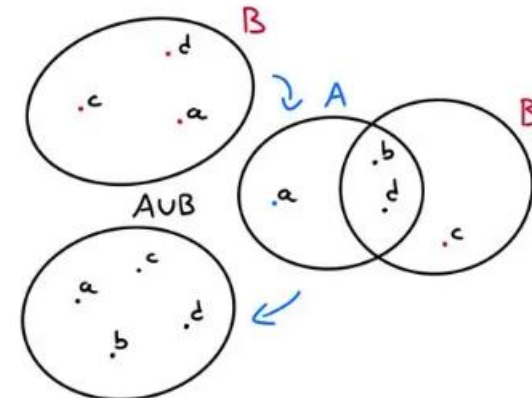
Il formato .bed

Cromosoma	inizio	fine	Campi addizionali
chr21	0	8219800	E1
chr21	8219800	8220200	E2
chr21	8220200	8234200	E1
chr21	8234200	8234800	E2
chr21	8234800	8420000	E1
chr21	8420000	8420600	E2
chr21	8420600	14026800	E1
chr21	14026800	14027800	E2
chr21	14027800	14080400	E1
chr21	14080400	14081400	E2
chr21	14081400	14216000	E1
chr21	14216000	14216400	E2
chr21	14216400	14381800	E1
chr21	14381800	14385000	E2
chr21	14385000	14451800	E1
chr21	14451800	14455400	E2
chr21	14455400	14477200	E1
chr21	14477200	14477800	E2

Come sono utili?

Possono definire degli intervalli

Permettono agilmente di fare operazioni di intersezione, unione e sottrazione tra varianti





bedtools - the swiss army knife for genome arithmetic

- Permette di intersecare, sottrarre e unire .bed.
- Utilizzi tipici:

```
bedtools intersect -a file1.bed -b file2.bed
```

e

```
bedtools subtract -a file1.bed -b file2.bed
```

che è uguale a

```
bedtools intersect -a file1.bed -b file2.bed -v
```

