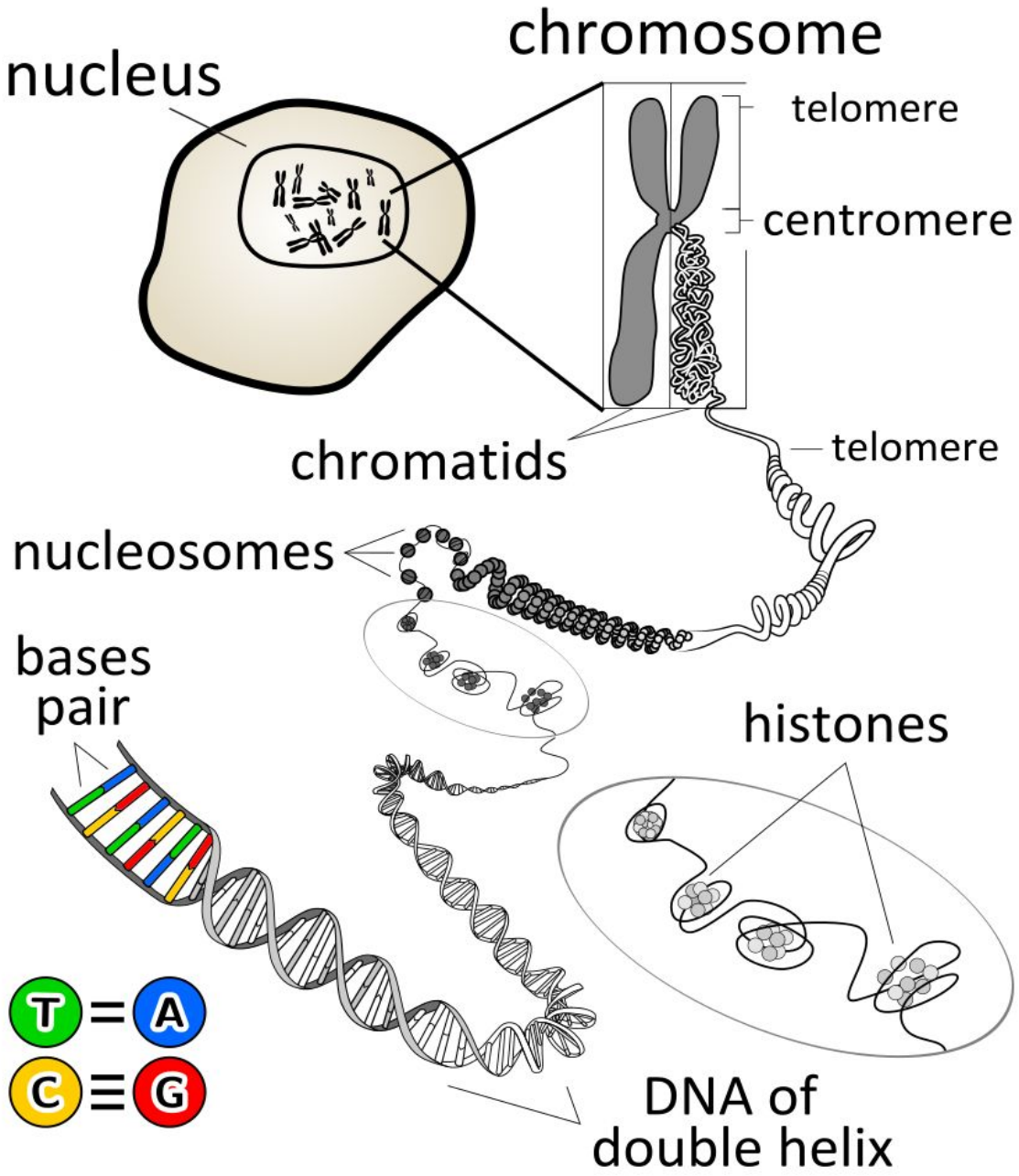

DTU

DTU Health Technology
Bioinformatics

Alignment post-processing and variant calling

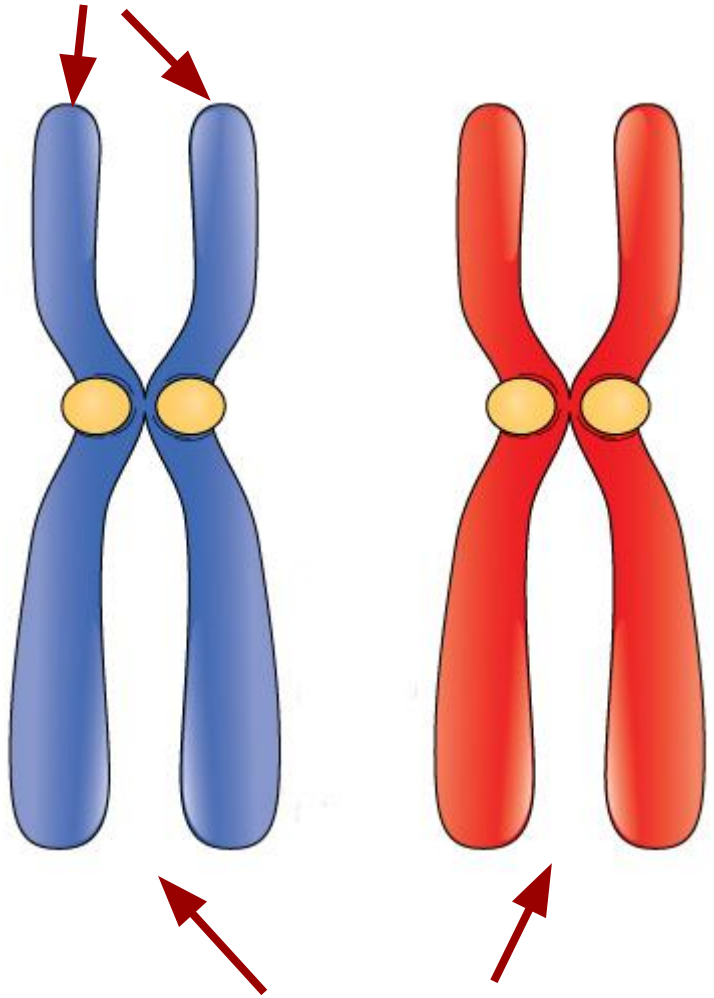
Gabriel Renaud
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gabriel.reno@gmail.com

A brief reminder



A brief reminder

sister chromatids



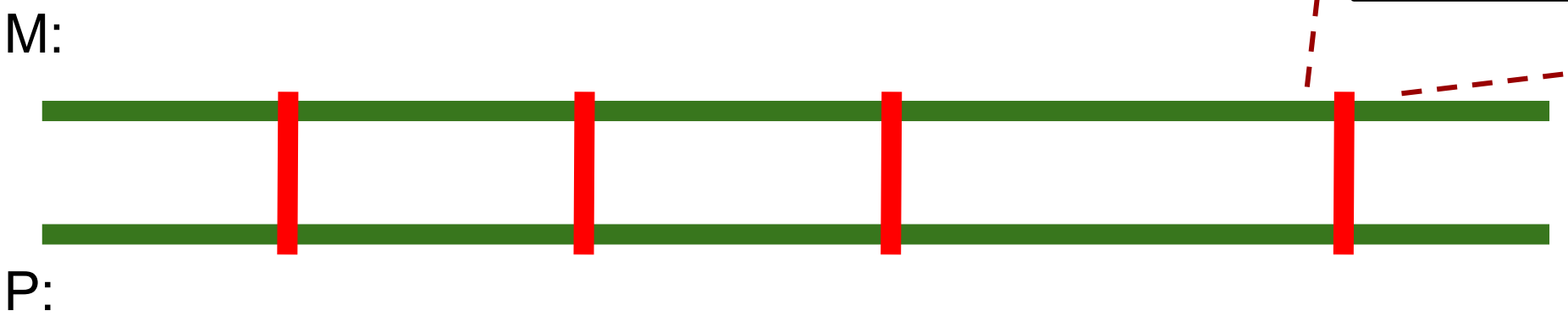
Paternal

Maternal

homologous chromosomes

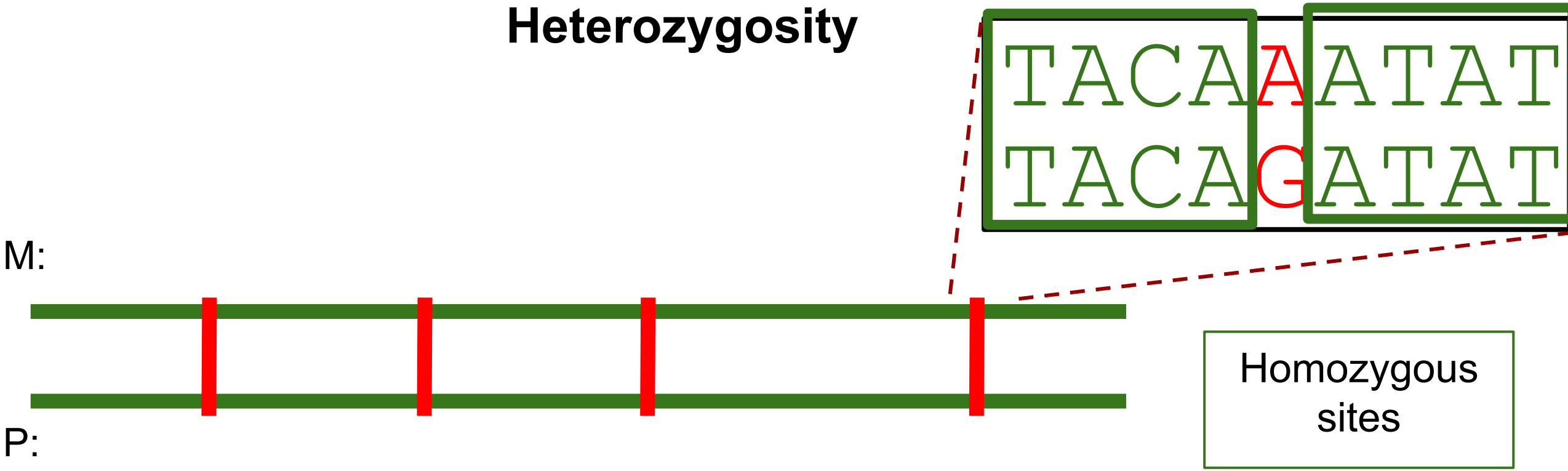
Heterozygosity

TACAAATAT
TACAGATAT

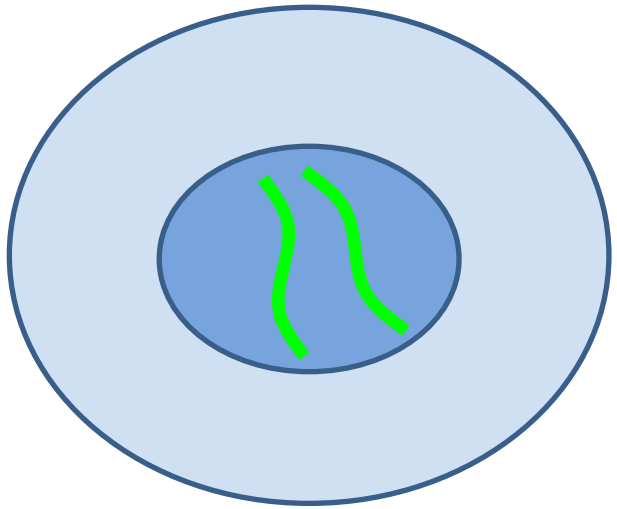


Heterozygous sites

Heterozygosity



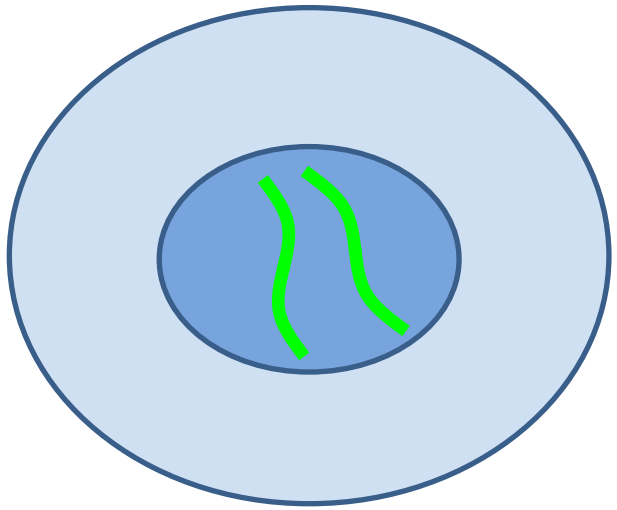
ind#A



M: **TACAAATAT**

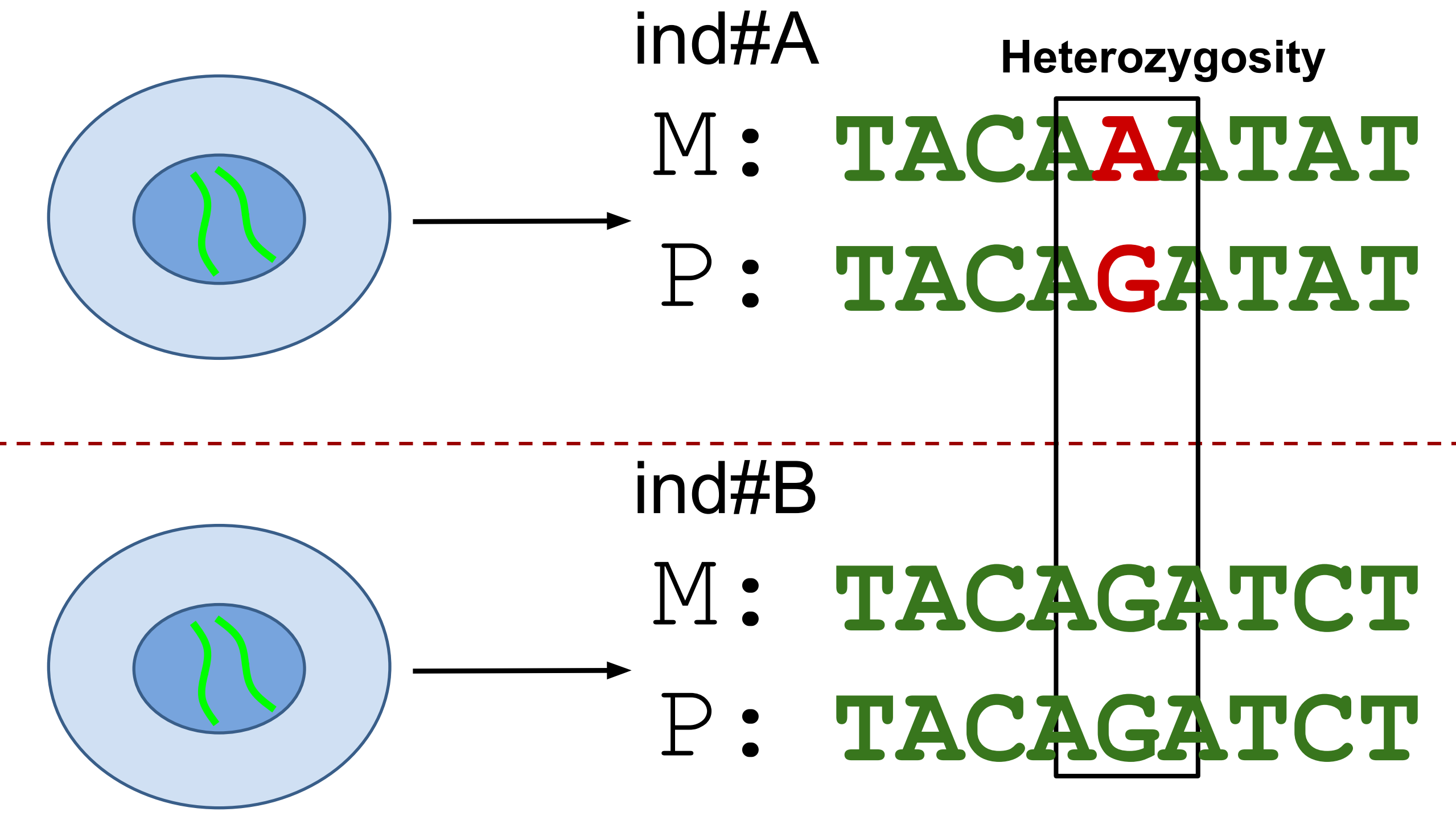
P: **TACAGATAT**

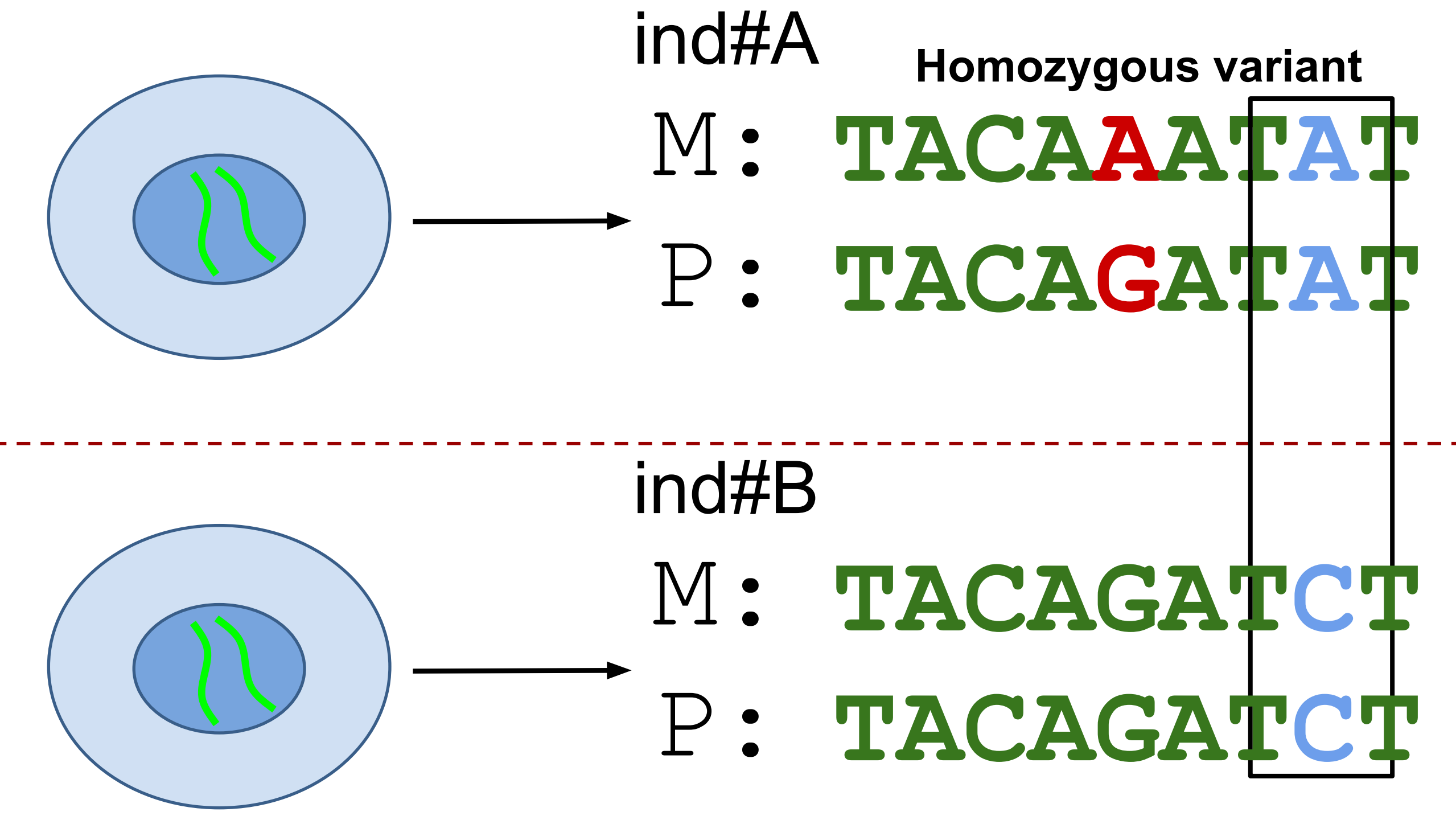
ind#B

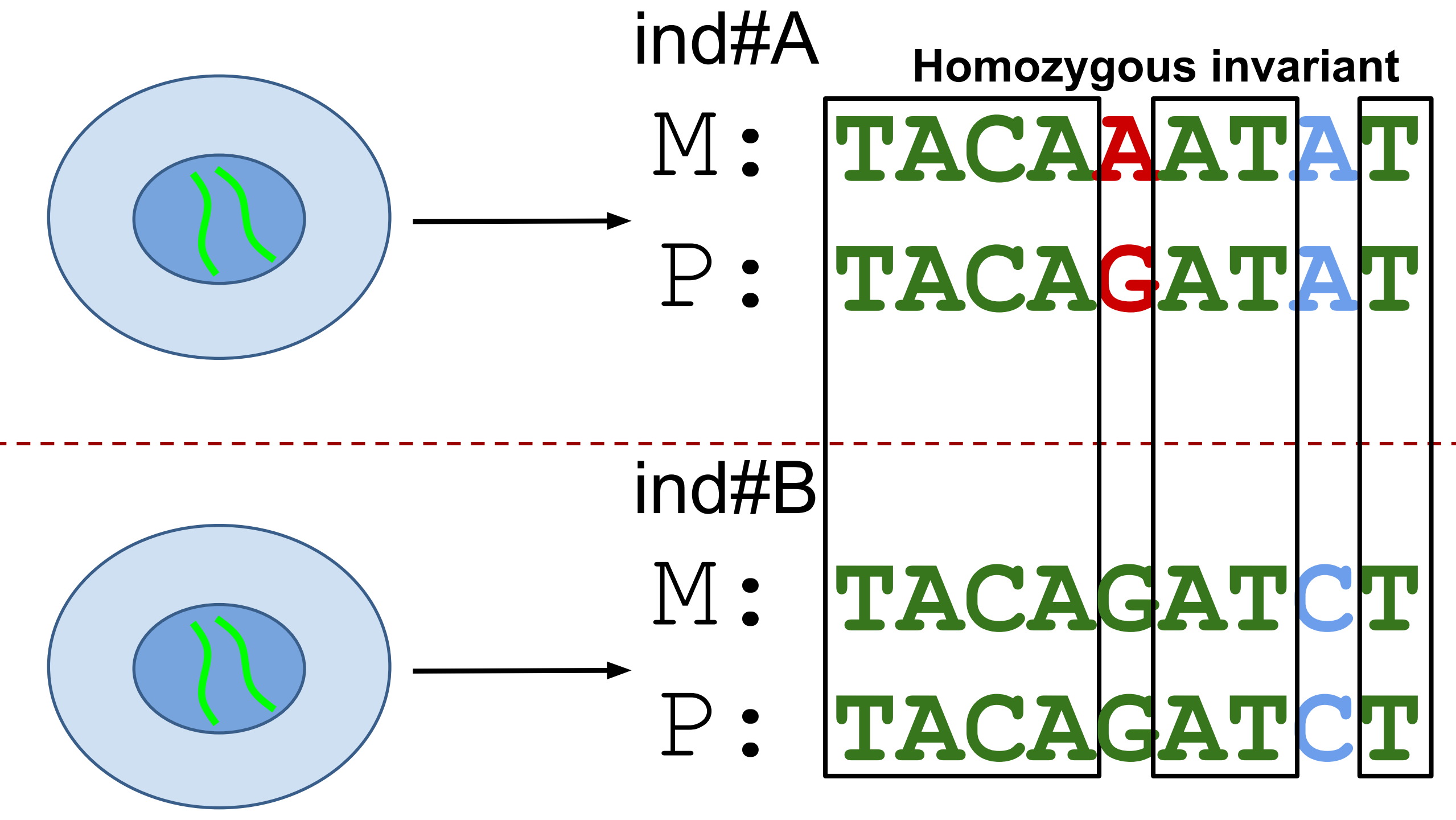


M: **TACAGATCT**

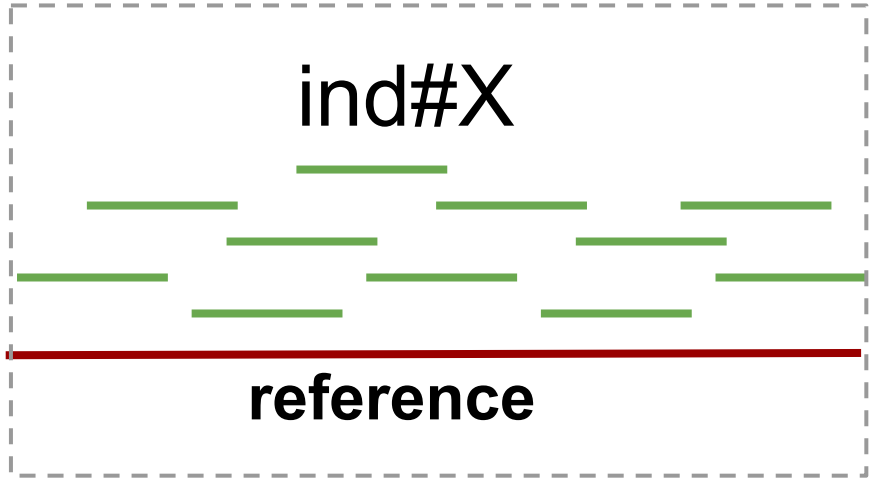
P: **TACAGATCT**


ind#A
Heterozygosity
M: TACAAATAT
P: TACAGATAT
ind#B
M: TACAGATCT
P: TACAGATCT


ind#A
Homozygous variant
M: TACAAATAT
P: TACAGATAT
ind#B
M: TACAGATCT
P: TACAGATCT


ind#A
Homozygous invariant
M:
P:
TACAAATAT
TACAGATAT
ind#B
M:
P:
TACAGATCT
TACAGATCT

Genotyping

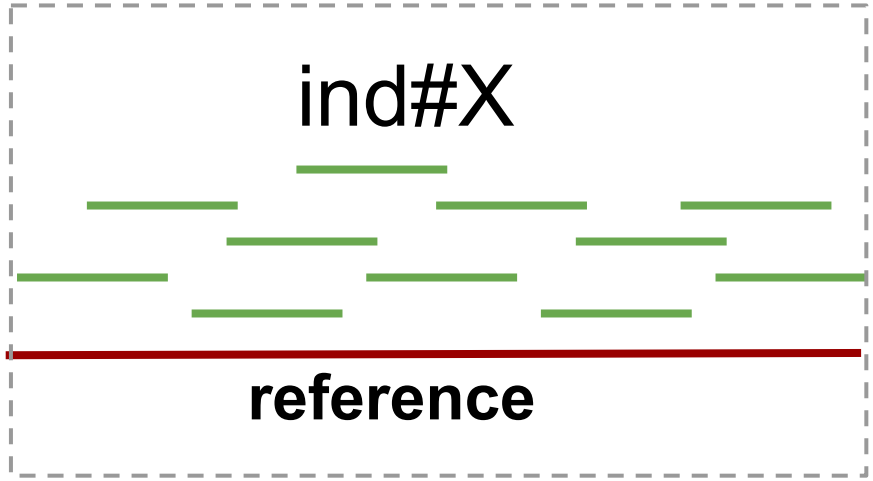


TACAAATAT
TACAGATAT

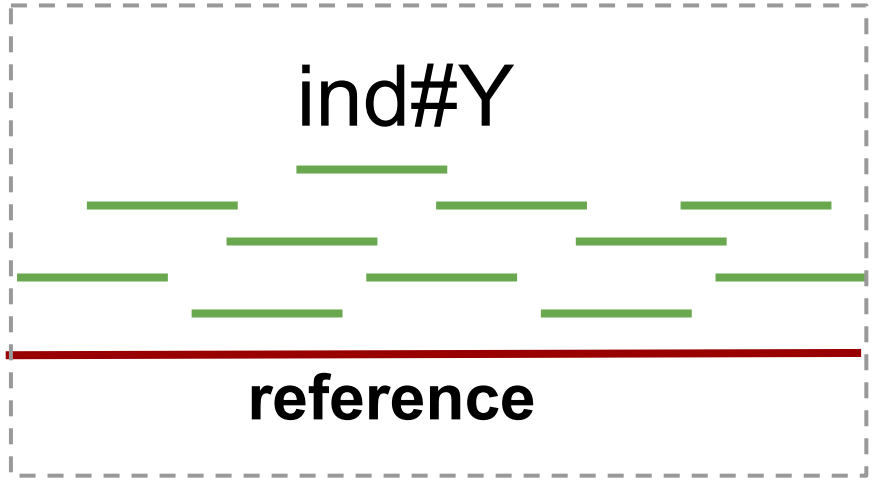
Which of the 10 possible genotypes is the most likely?

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Joint Genotyping



TACAAATAT
TACAGATAT

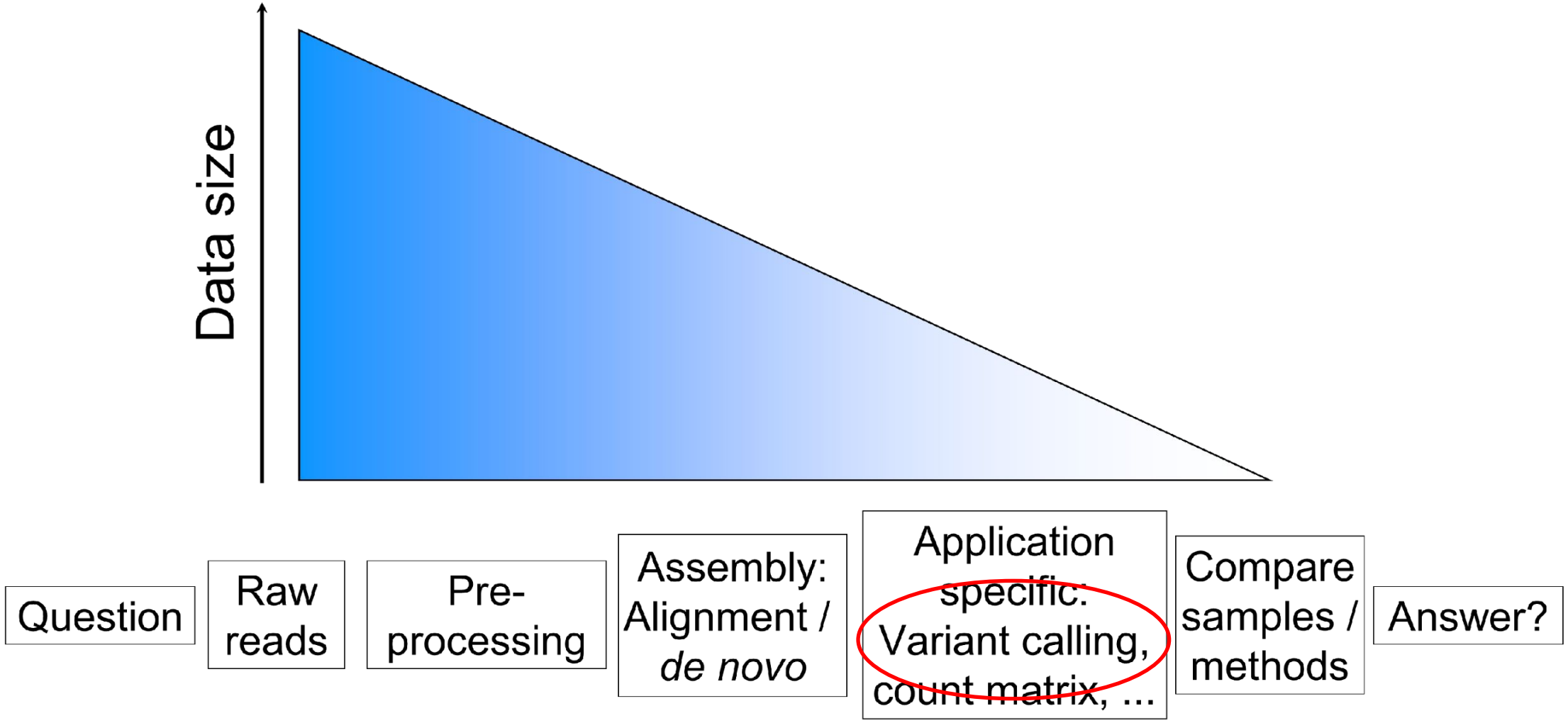


TACAGATCT
TACAGATCT

Menu

- Introduction
- From aligned reads to genomic variation
- Alignment post-processing
- Variant effect

Generalized NGS analysis

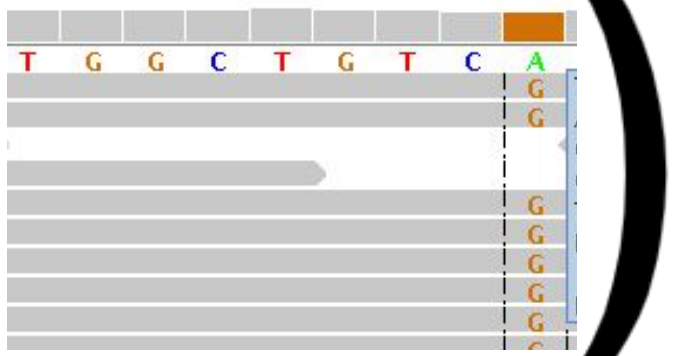


What is genotyping?

Genotyping is determining which genotype maximizes:

What is genotyping?

Genotyping is determining which genotype maximizes:

$$P(G \mid$$  $$)$$

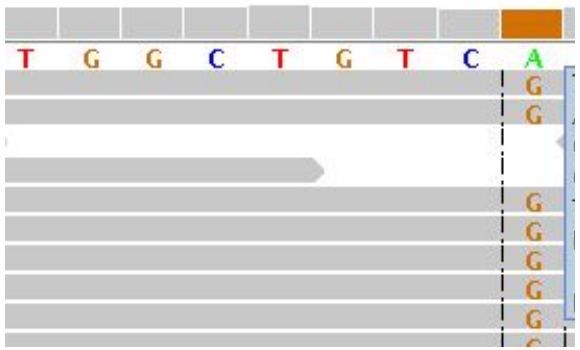
genotype

data

What is genotyping?

Genotyping is determining which genotype maximizes:

$$P(\texttt{TA} \mid \text{data})$$



genotype

data

What is genotyping?

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

What is genotyping?

prior: what is the probability of the genotype to begin with?

likelihood: What is the probability of seeing the data given the genotype?

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

What is genotyping?

prior: what is the probability of the genotype to begin with?

likelihood: What is the probability of seeing the data given the genotype?

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

evidence: What is the probability of generating that data to begin with?

$$P(D) = \sum_{G \in \mathbb{G}} P(G)P(D|G)$$

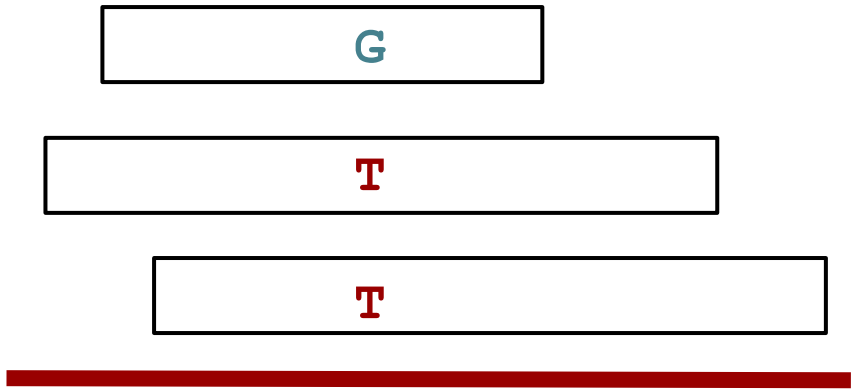
The likelihood

$$P(D|G) = \prod_{b \in READS} P(b|G)$$

i.e. each reads is an independent observation

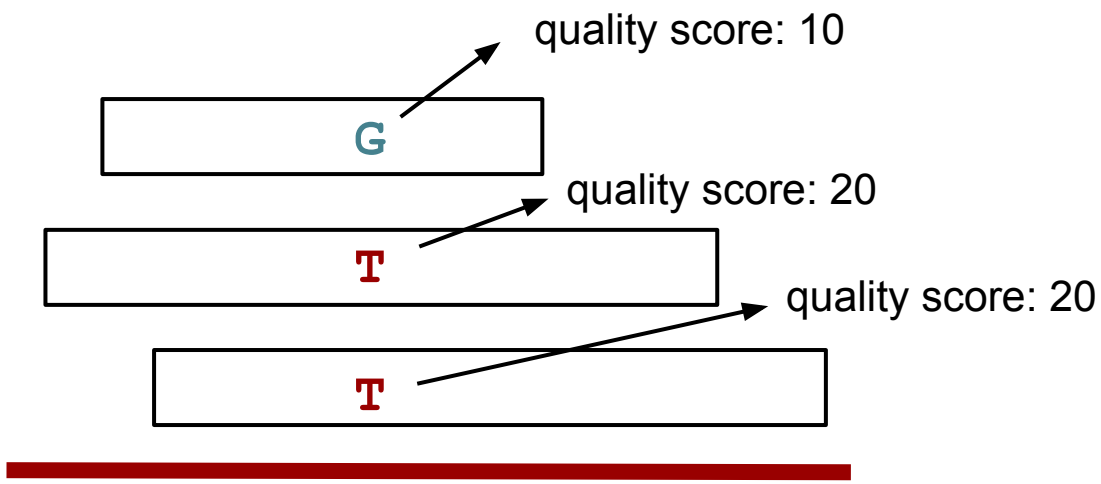
The likelihood $P(D|G)$

Toy example:



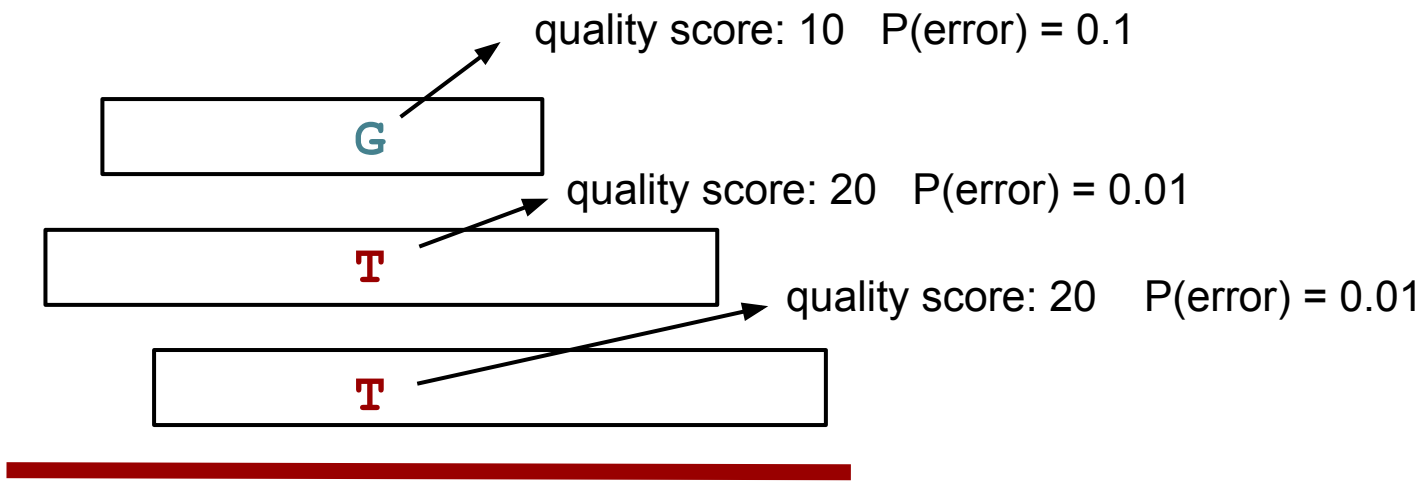
The likelihood $P(D|G)$

Toy example:



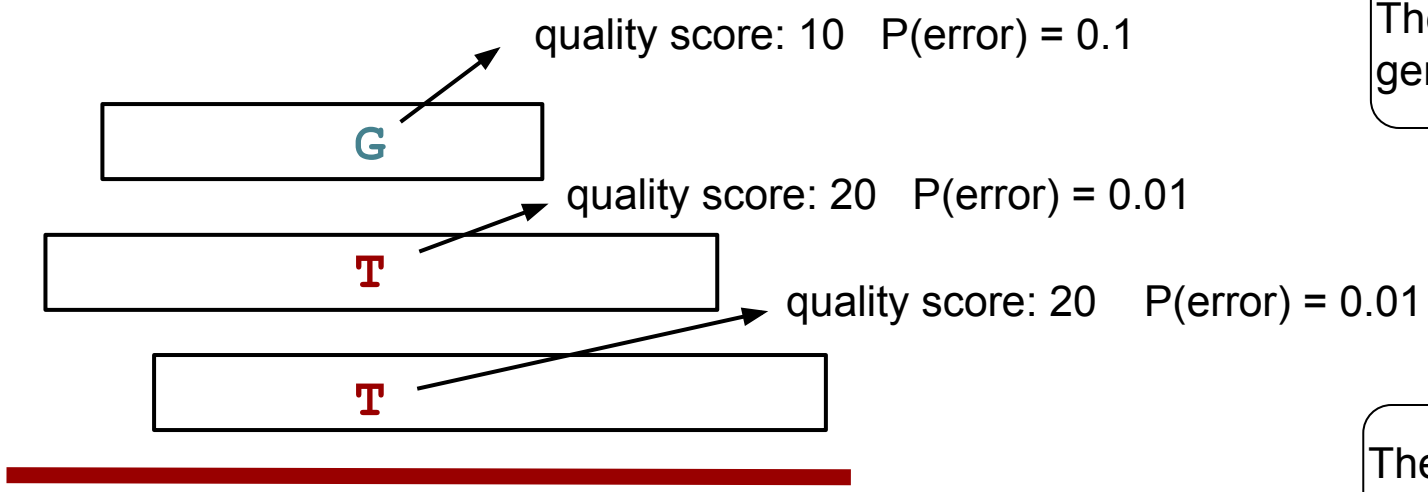
The likelihood $P(D|G)$

Toy example:



The likelihood $P(D|G)$

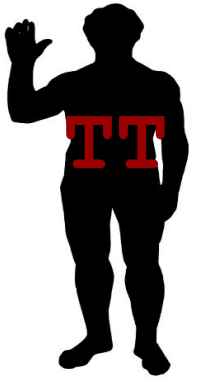
Toy example:



The 2 Ts are sequencing errors! The genotype is GG

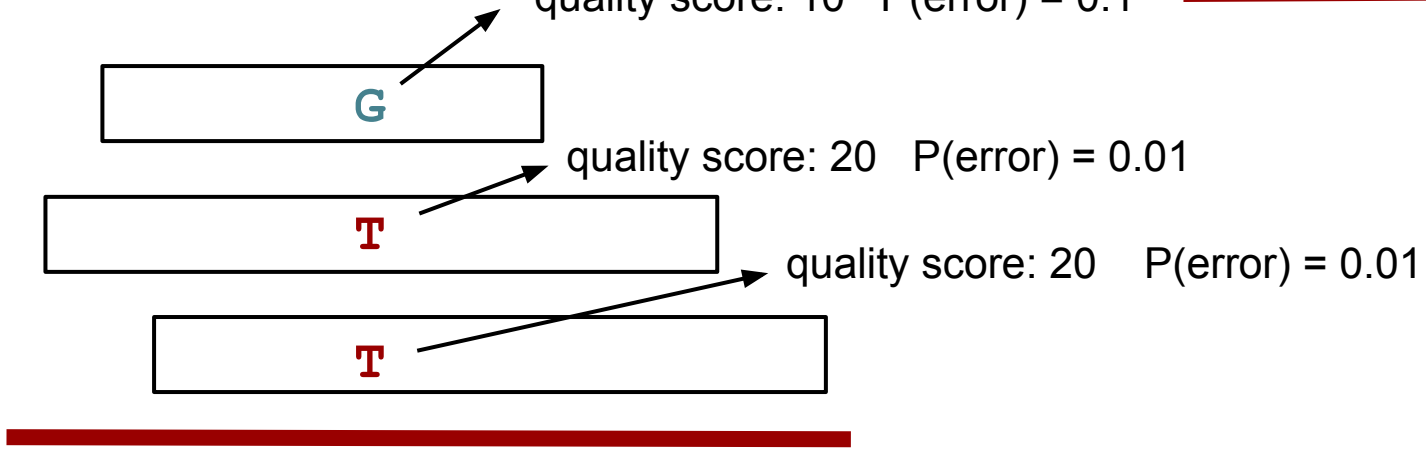
They are all correct and the genotype is GT

The G is a sequencing error! TT is the genotype



The likelihood $P(D|G)$

Toy example:



Error model

What I think is the base

probability of the data given the base

$$P(G|A) = 0.1\frac{1}{3}$$

$$P(G|C) = 0.1\frac{1}{3}$$

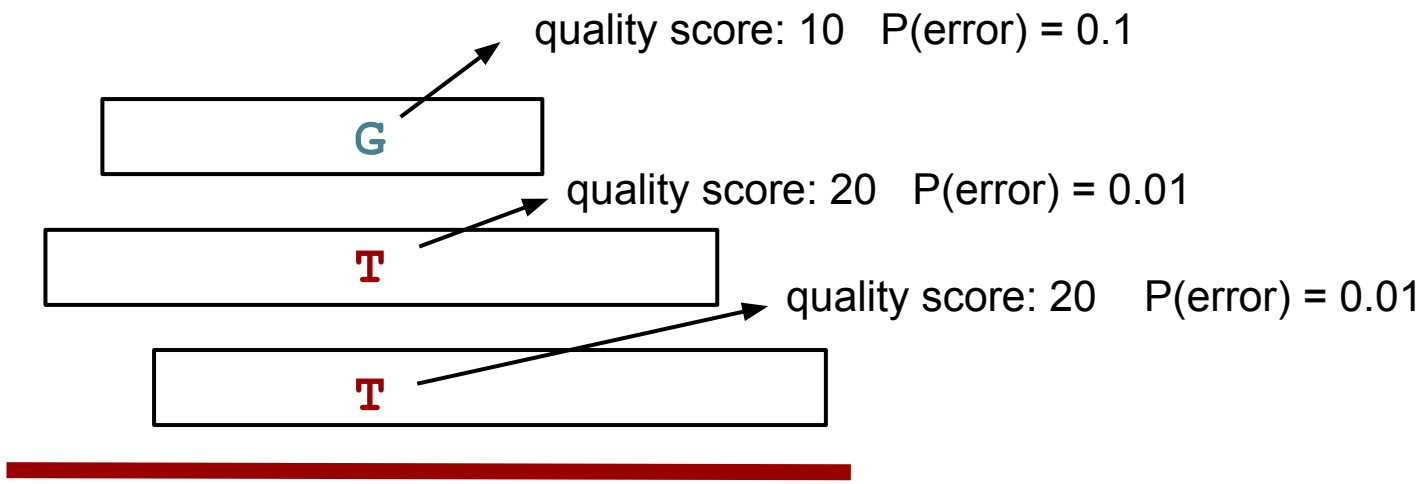
$$P(G|G) = 0.9$$

$$P(G|T) = 0.1\frac{1}{3}$$

The likelihood $P(D|G)$

Let's evaluate 3 possible genotypes:

Toy example:



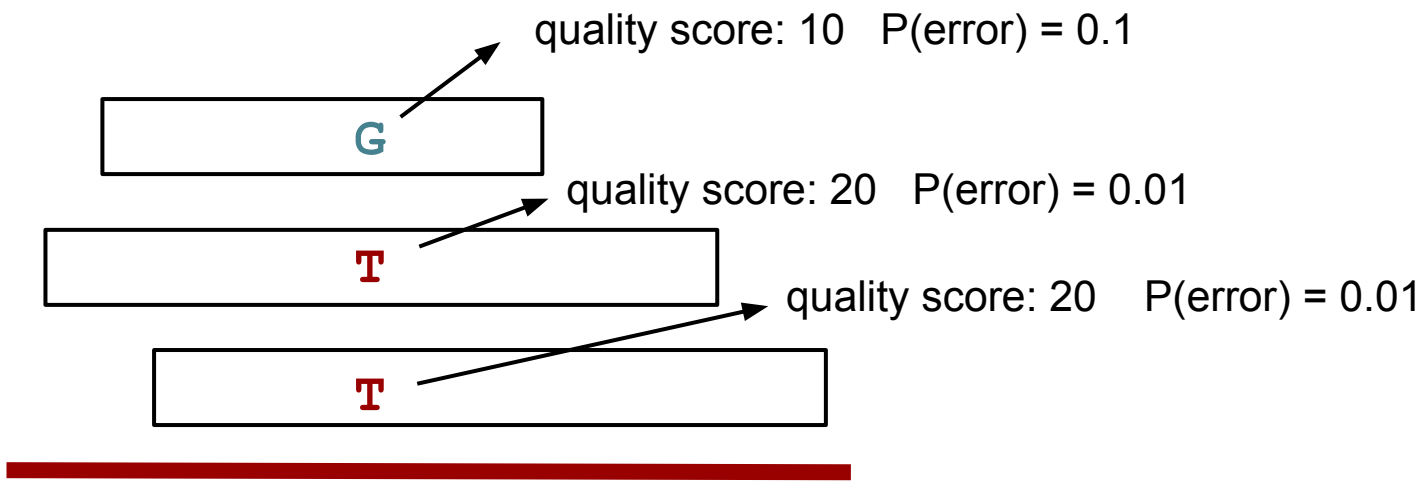
GG

GT

TT

$P(D|GG)$

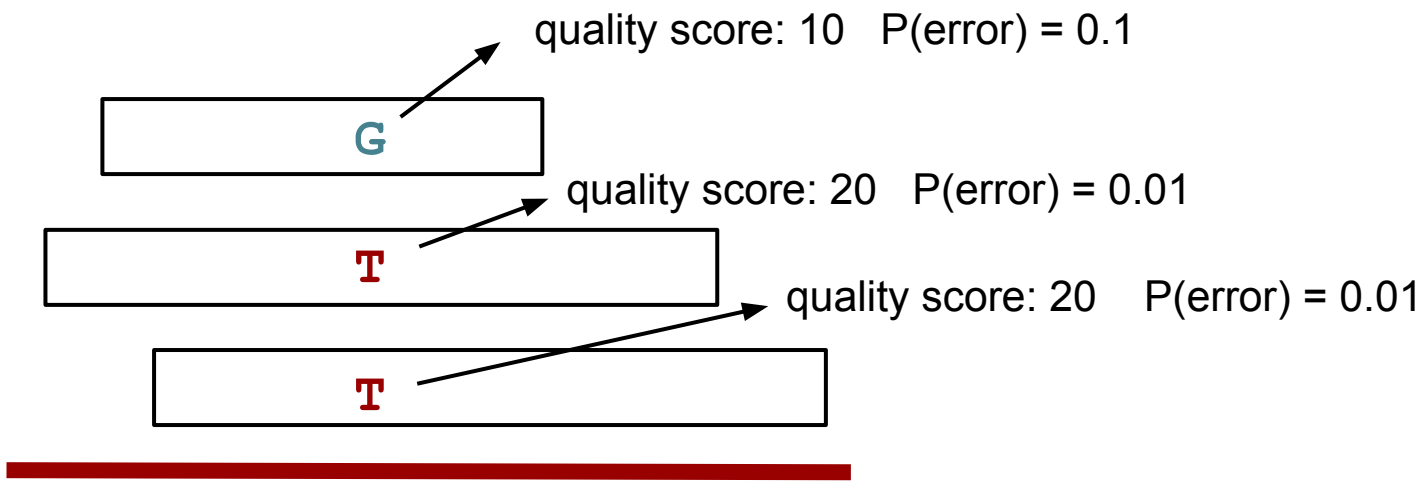
The likelihood $P(D|G)$



The likelihood $P(D|G)$

$$P(D|\mathtt{GG})$$

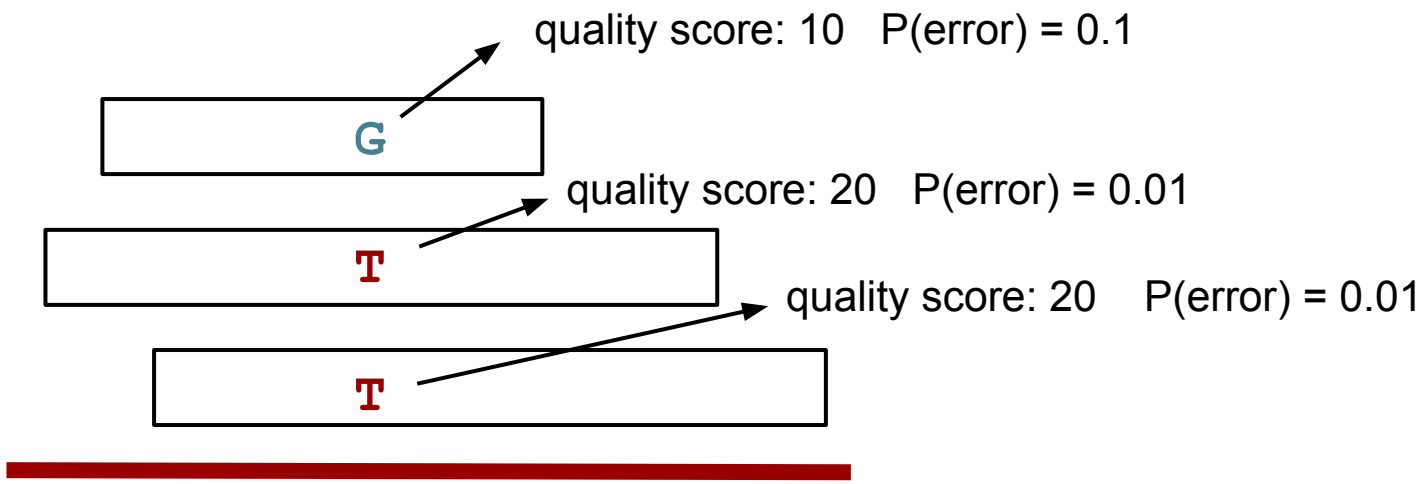
$\frac{1}{2}$ G $\frac{1}{2}$ G



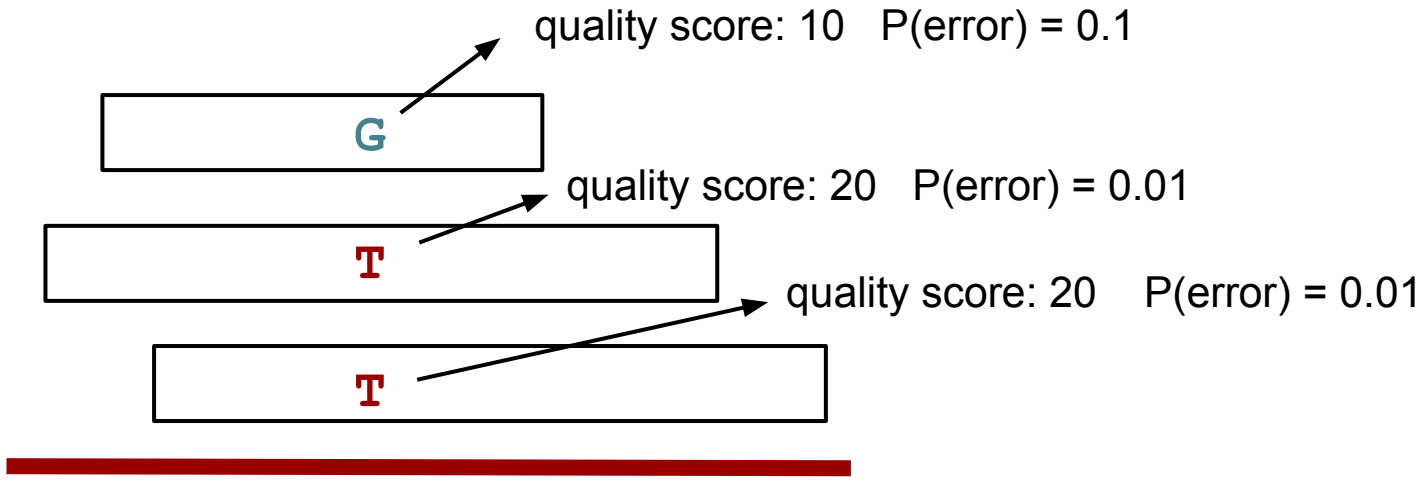
The likelihood $P(D|G)$

$$P(D|\mathtt{GG})$$

$\frac{1}{2}$ G ✓ 0.9 $\frac{1}{2}$ G ✓ 0.9



The likelihood $P(D|G)$



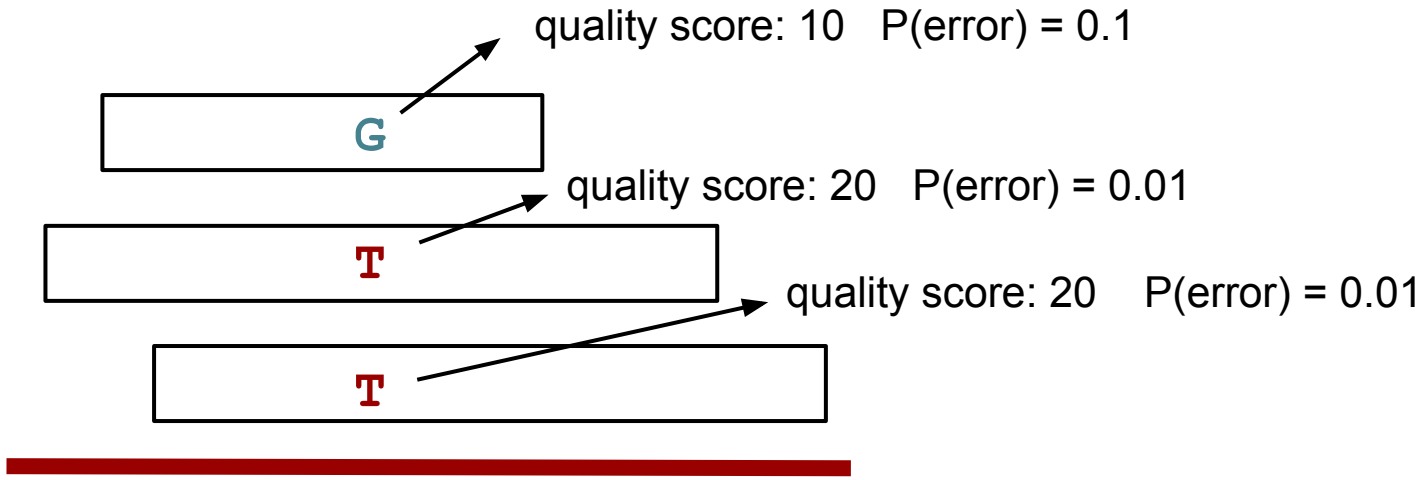
$$P(D|\mathtt{GG})$$

$\frac{1}{2}$ G $\frac{1}{2}$ G

✓ 0.9 ✓ 0.9

✗ $\frac{0.01}{3}$ ✗ $\frac{0.01}{3}$

The likelihood $P(D|G)$



$$P(D|\mathrm{GG})$$

$\frac{1}{2}$ G $\frac{1}{2}$ G

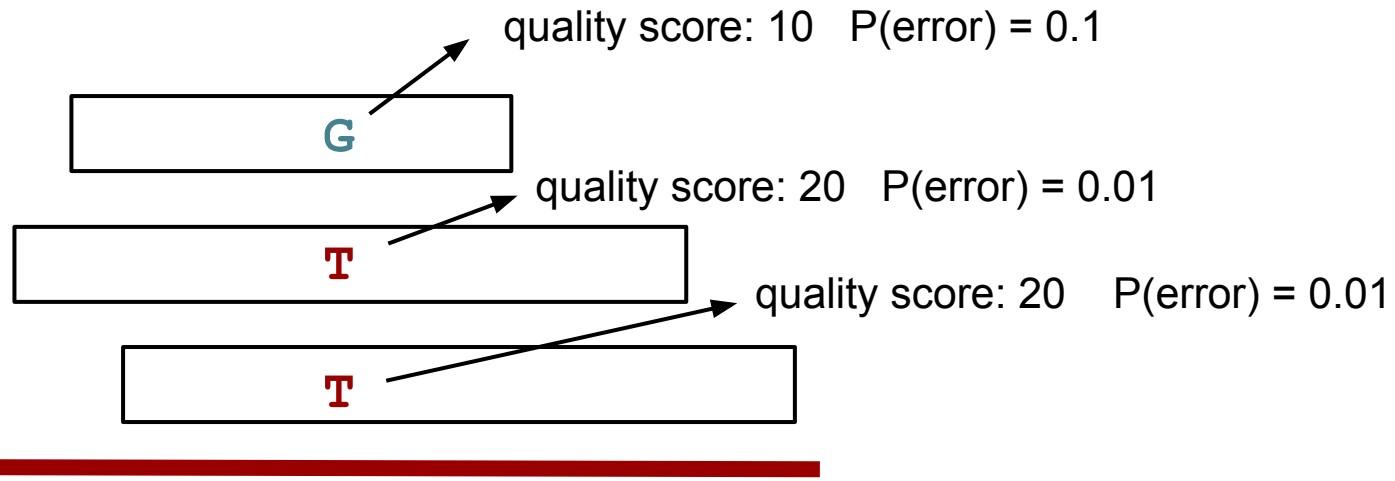
✓ 0.9 ✓ 0.9

✗ $\frac{0.01}{3}$ ✗ $\frac{0.01}{3}$

✗ $\frac{0.01}{3}$ ✗ $\frac{0.01}{3}$

The likelihood $P(D|G)$

$$P(D|\mathtt{GG})$$

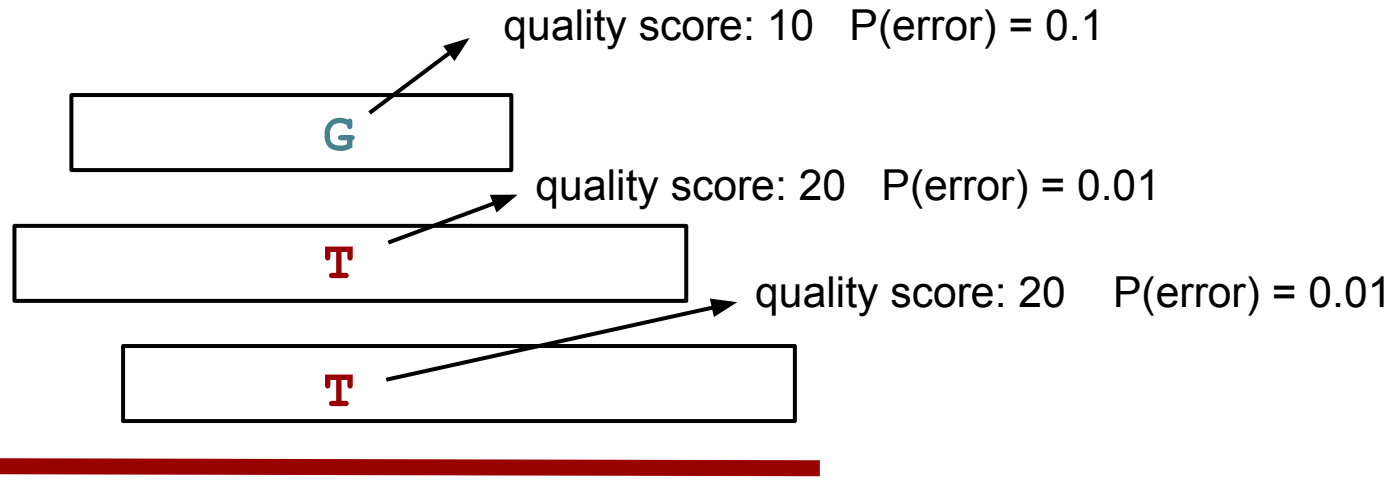


$\frac{1}{2}$ G		$\frac{1}{2}$ G	
✓	0.9	✓	0.9
✗	$\frac{0.01}{3}$	✗	$\frac{0.01}{3}$
✗	$\frac{0.01}{3}$	✗	$\frac{0.01}{3}$

$$(\tfrac{1}{2}0.9 + \tfrac{1}{2}0.9)(\tfrac{1}{2}\tfrac{0.01}{3} + \tfrac{1}{2}\tfrac{0.01}{3})(\tfrac{1}{2}\tfrac{0.01}{3} + \tfrac{1}{2}\tfrac{0.01}{3}) = 0.00001$$

The likelihood $P(D|G)$

$$P(D|\mathtt{GT})$$



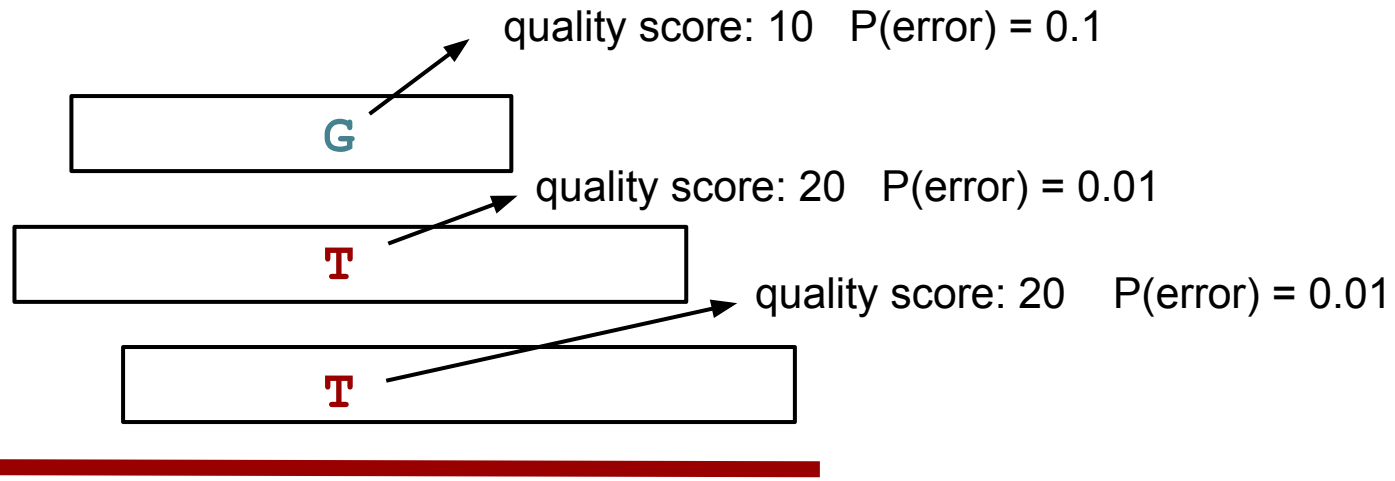
$\frac{1}{2}$ G $\frac{1}{2}$ T

✓	0.9	✗	$\frac{0.1}{3}$
✗	$\frac{0.01}{3}$	✓	0.99
✗	$\frac{0.01}{3}$	✓	0.99

$$(\frac{1}{2}0.9 + \frac{1}{2}\frac{0.1}{3})(\frac{1}{2}\frac{0.01}{3} + \frac{1}{2}0.99)(\frac{1}{2}\frac{0.01}{3} + \frac{1}{2}0.99) = 0.1151163$$

The likelihood $P(D|G)$

$$P(D|\texttt{TT})$$



$$\frac{1}{2}\,\texttt{T} \qquad \frac{1}{2}\,\texttt{T}$$

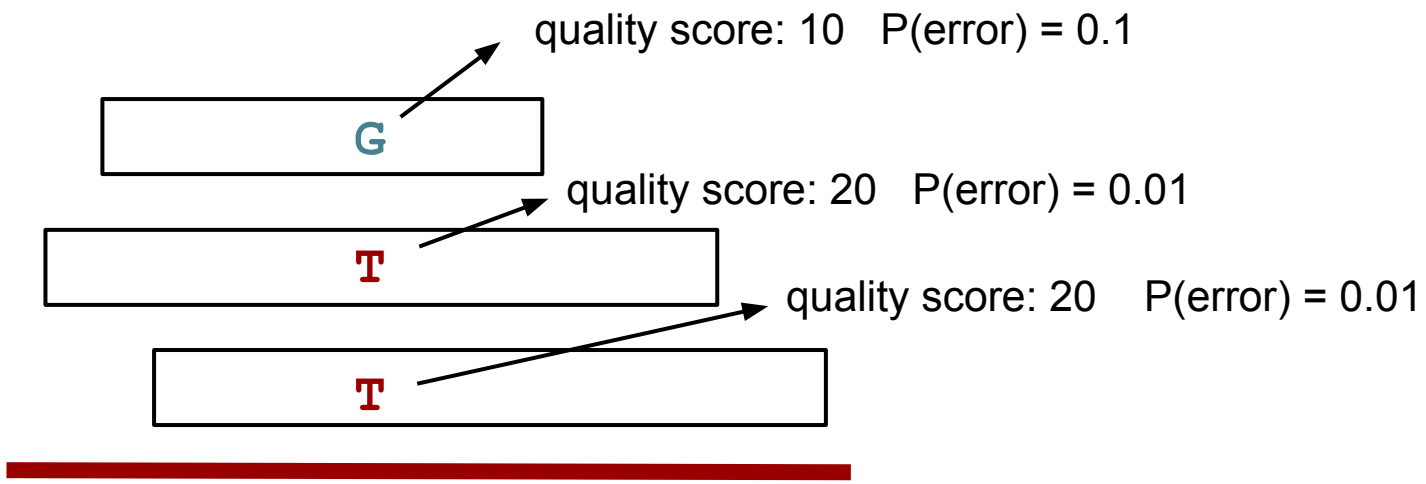
$$\times \frac{0.1}{3} \qquad \times \frac{0.1}{3}$$

$$\checkmark\ 0.99 \qquad \checkmark\ 0.99$$

$$\checkmark\ 0.99 \qquad \checkmark\ 0.99$$

$$(\tfrac{1}{2}\tfrac{0.1}{3} + \tfrac{1}{2}\tfrac{0.1}{3})(\tfrac{1}{2}0.99 + \tfrac{1}{2}0.99)(\tfrac{1}{2}0.99 + \tfrac{1}{2}0.99) = 0.03267$$

The likelihood $P(D|G)$



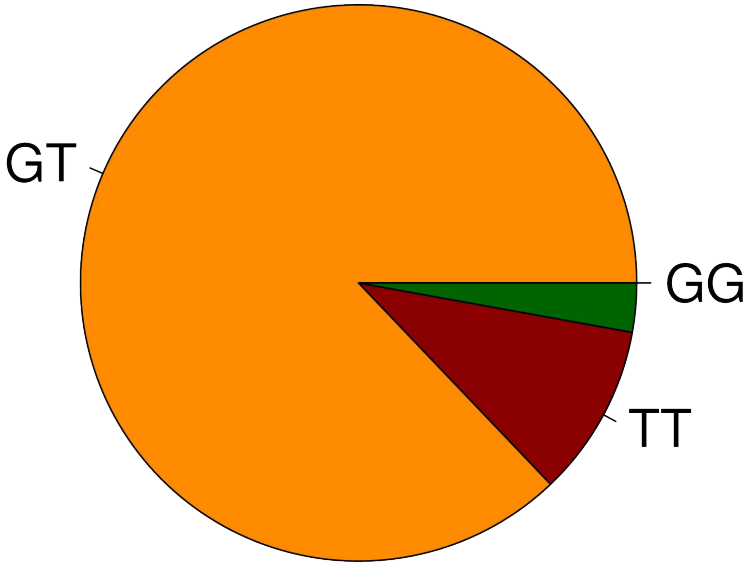
$$P(D|\mathtt{GG}) = 0.00001$$

$$P(D|\mathtt{GT}) = 0.11511$$

$$P(D|\mathtt{TT}) = 0.0327$$

The likelihood $P(D|G)$

A likelihood in itself is not meaningful, you need to compare it to other models



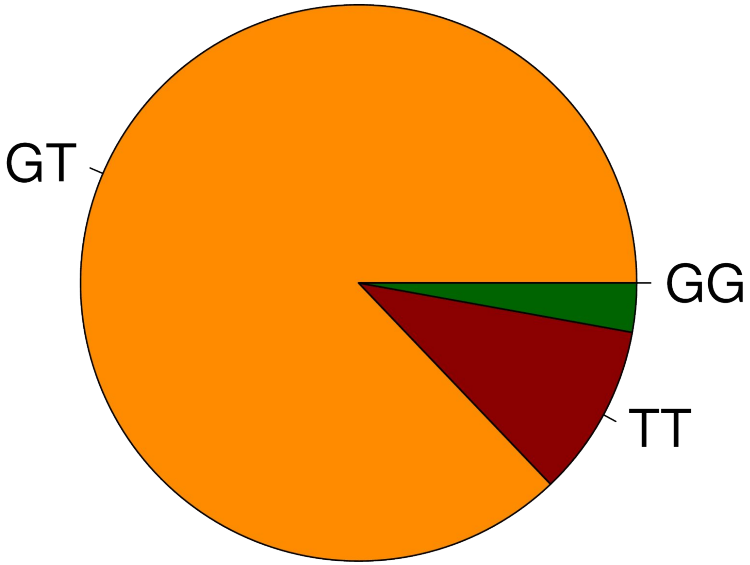
$$P(D|\mathtt{GG}) = 0.00001$$

$$P(D|\mathtt{GT}) = 0.11511$$

$$P(D|\mathtt{TT}) = 0.0327$$

$$P(D) = P(\mathtt{GG})P(D|\mathtt{GG}) + P(\mathtt{GT})P(D|\mathtt{GT}) + P(\mathtt{TT})P(D|\mathtt{TT})$$

The likelihood $P(D|G)$



$$P(D|\mathtt{GG}) = 0.00001$$

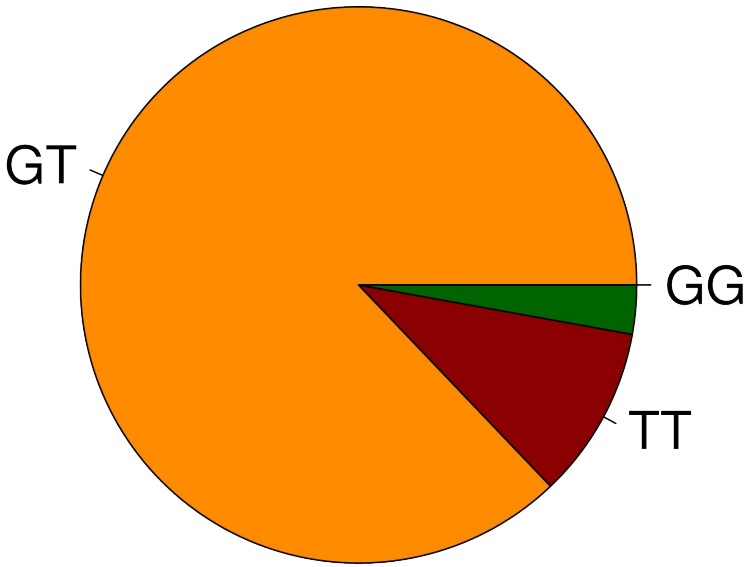
$$P(D|\mathtt{GT}) = 0.11511$$

$$P(D|\mathtt{TT}) = 0.0327$$

We will neglect the genotype prior this time

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}$$

The likelihood $P(D|G)$

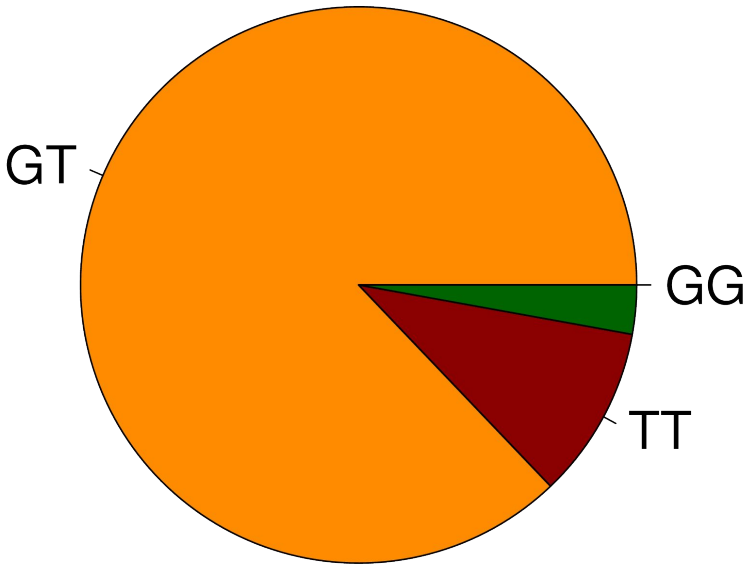


$$P(\mathtt{GG}|D) = \mathtt{6.7e\text{-}05}$$

$$P(\mathtt{GT}|D) = \mathtt{0.77888}$$

$$P(\mathtt{TT}|D) = \mathtt{0.22104}$$

The likelihood $P(D|G)$



$$P(\mathtt{GG}|D) = \mathtt{6.7e\text{-}05}$$

$$P(\mathtt{GT}|D) = \mathtt{0.77888}$$

$$P(\mathtt{TT}|D) = \mathtt{0.22104}$$

Important point: More coverage $\longrightarrow$ More multiplications $\longrightarrow$ The relative difference between models become larger

The likelihood $P(D|G)$

	PHRED	PHRED-scaled
$P(GG|D)$ =	6.7e-05	40.60
$P(GT|D)$ =	0.77888	0.00
$P(TT|D)$ =	0.22104	5.47

Details I did not cover

- Error model
  - Most genotypers do not simply use raw quality scores

Most common genotypers

- GATK
- SAMtools/BCFtools
- Graphtyper
- FreeBayes

Deep Learning and genotyping?

Published: 24 September 2018

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

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

Nature Biotechnology **36**, 983–987 (2018) | Cite this article

26k Accesses | **196** Citations | **319** Altmetric | Metrics

Abstract

Despite rapid advances in sequencing technologies, accurately calling genetic variants present in an individual genome from billions of short, errorful sequence reads remains challenging. Here we show that a deep convolutional neural network can call genetic variation in aligned next-generation sequencing read data by learning statistical

Accurate, scalable cohort variant calls using DeepVariant and GLnexus

Taedong Yun, Helen Li, Pi-Chuan Chang, Michael F Lin, Andrew Carroll, Cory Y McLean

Bioinformatics, Volume 36, Issue 24, 15 December 2020, Pages 5582–5589,
https://doi.org/10.1093/bioinformatics/btaa1081
Published: 05 January 2021 **Article history**

Abstract

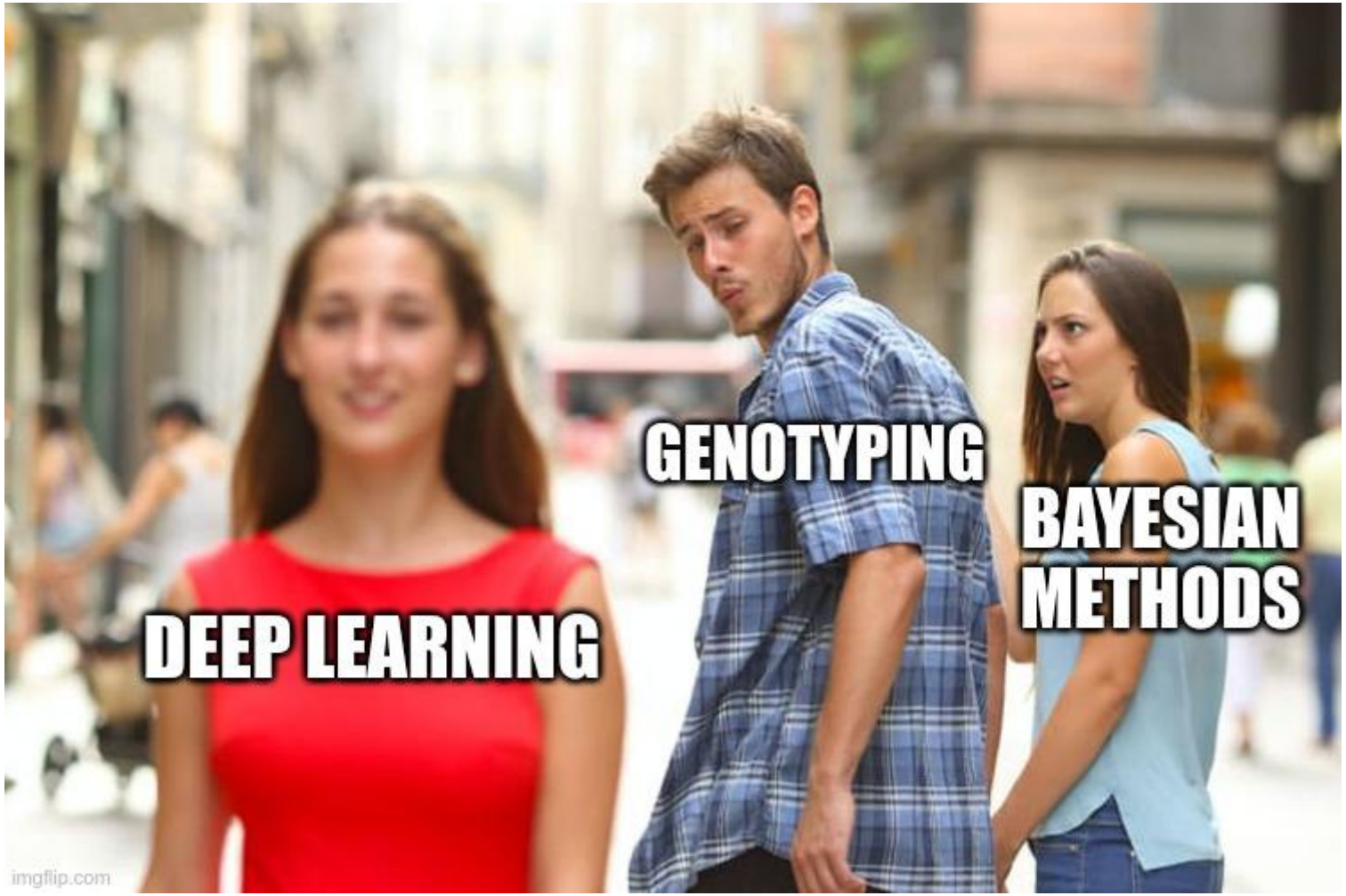
Motivation

Population-scale sequenced cohorts are foundational resources for genetic analyses, but processing raw reads into analysis-ready cohort-level variants remains challenging.

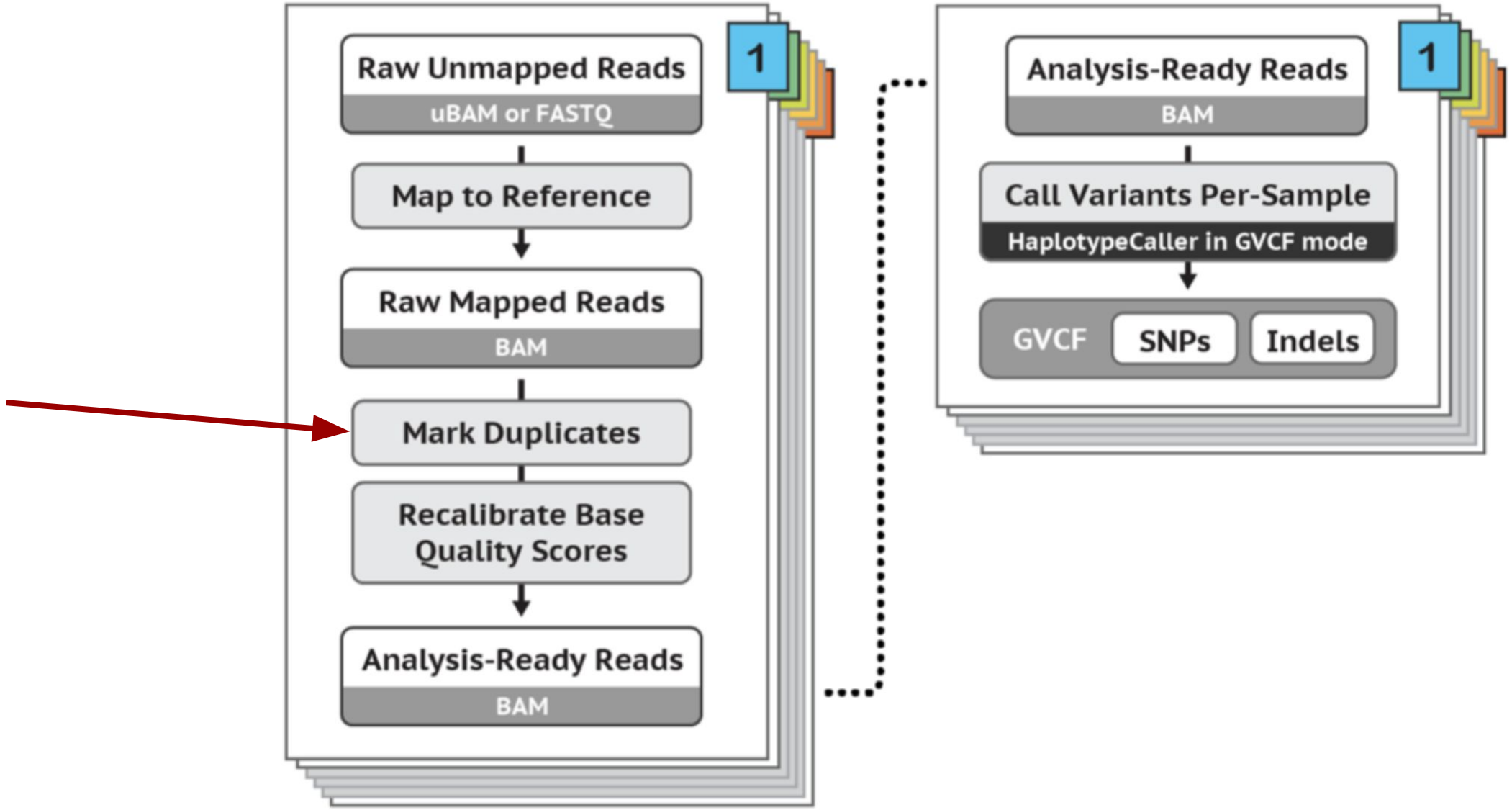
Results

We introduce an open-source cohort-calling method that uses the highly accurate caller DeepVariant and scalable merging tool GLnexus. Using callset

Deep Learning and genotyping?

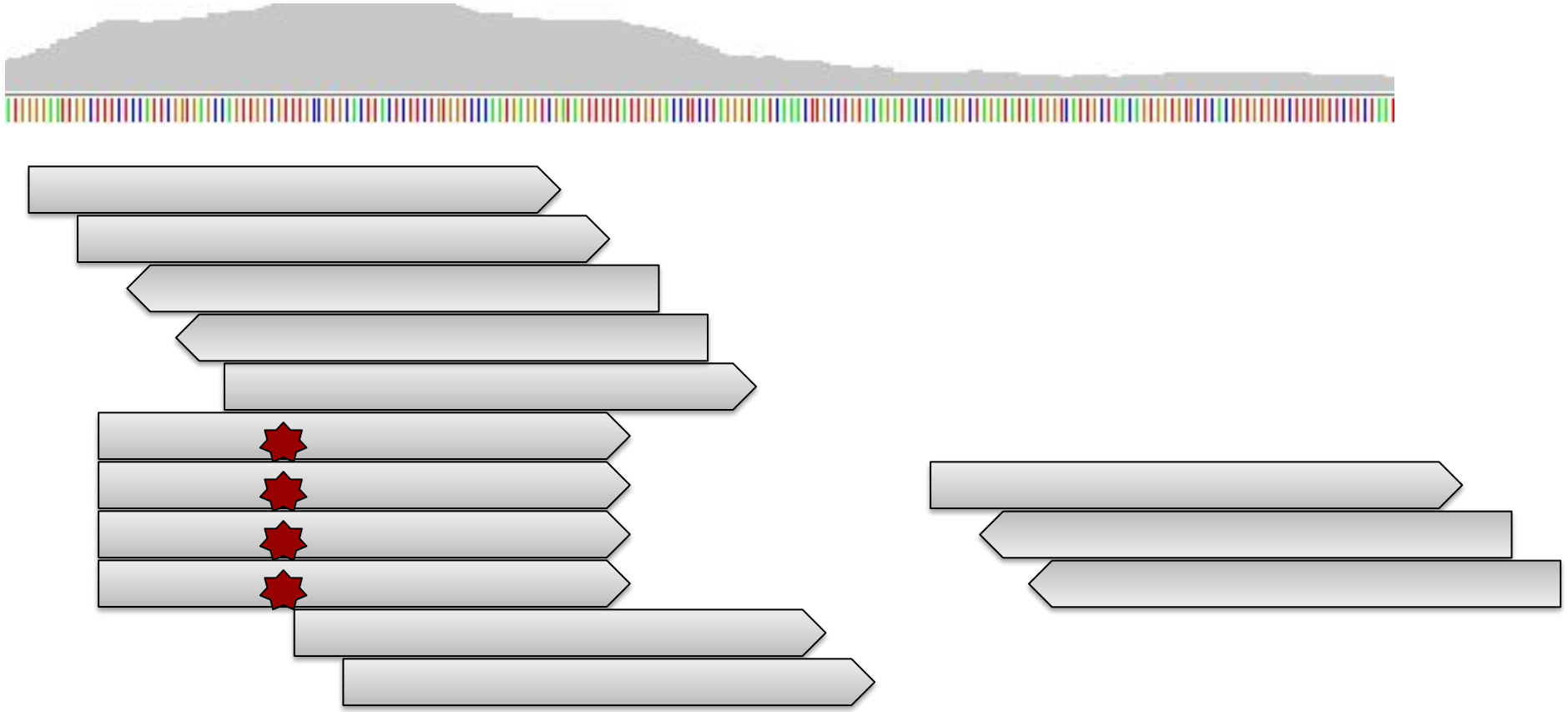


GATK's recommended workflow



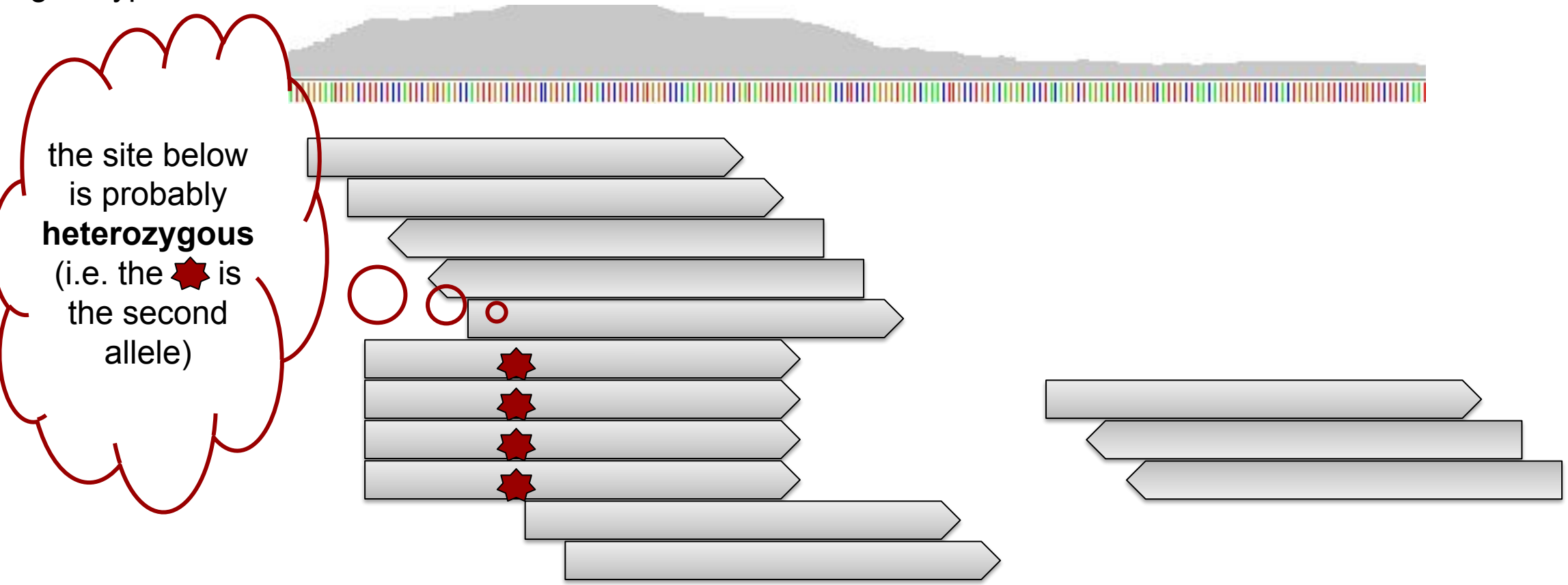
https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

The PCR amplification step included in the majority of NGS library construction techniques can introduce duplicates in the data.



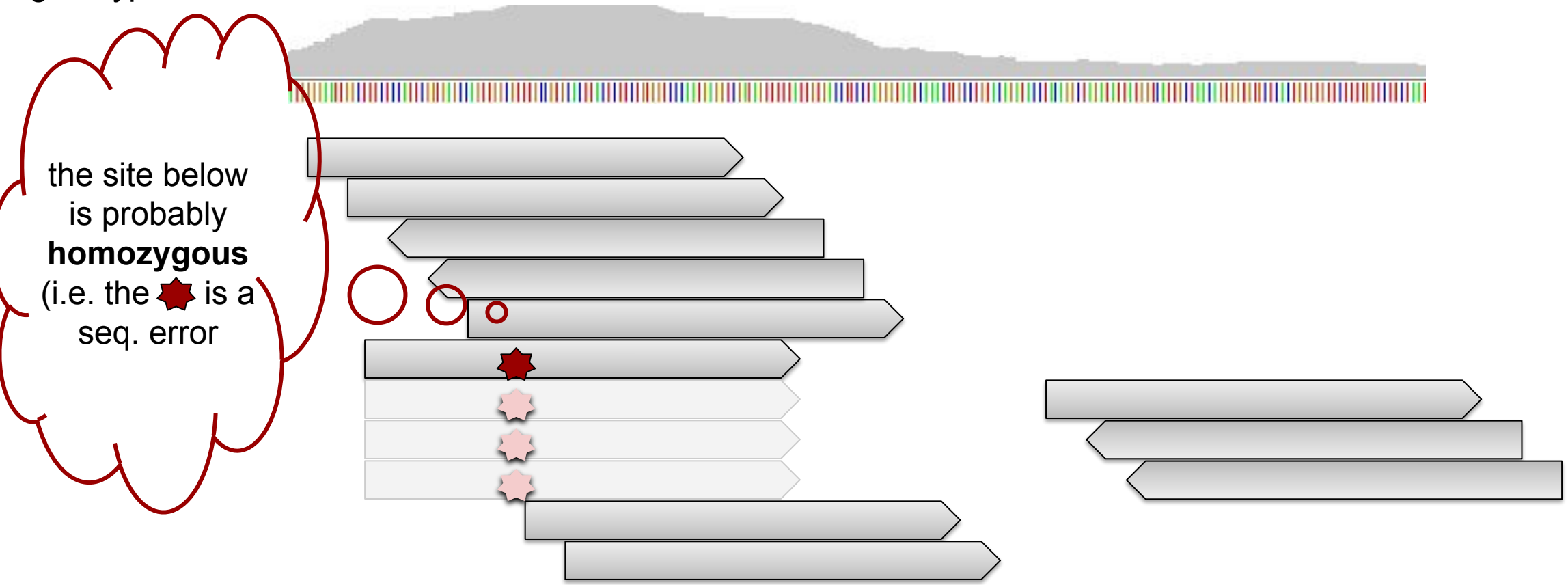
We want: remove or mark them to avoid false calls

genotyper:



Genotypers will ignore reads marked as duplicates

genotyper:



Duplicate/marking removal

Basic concepts of duplicate marking algorithm:

- Identify genomic position and strand for 5’-most bases.
- Mark reads that are duplicates of each other.
- Within a group of duplicate reads, the read with the highest sum of base quality scores is retained.

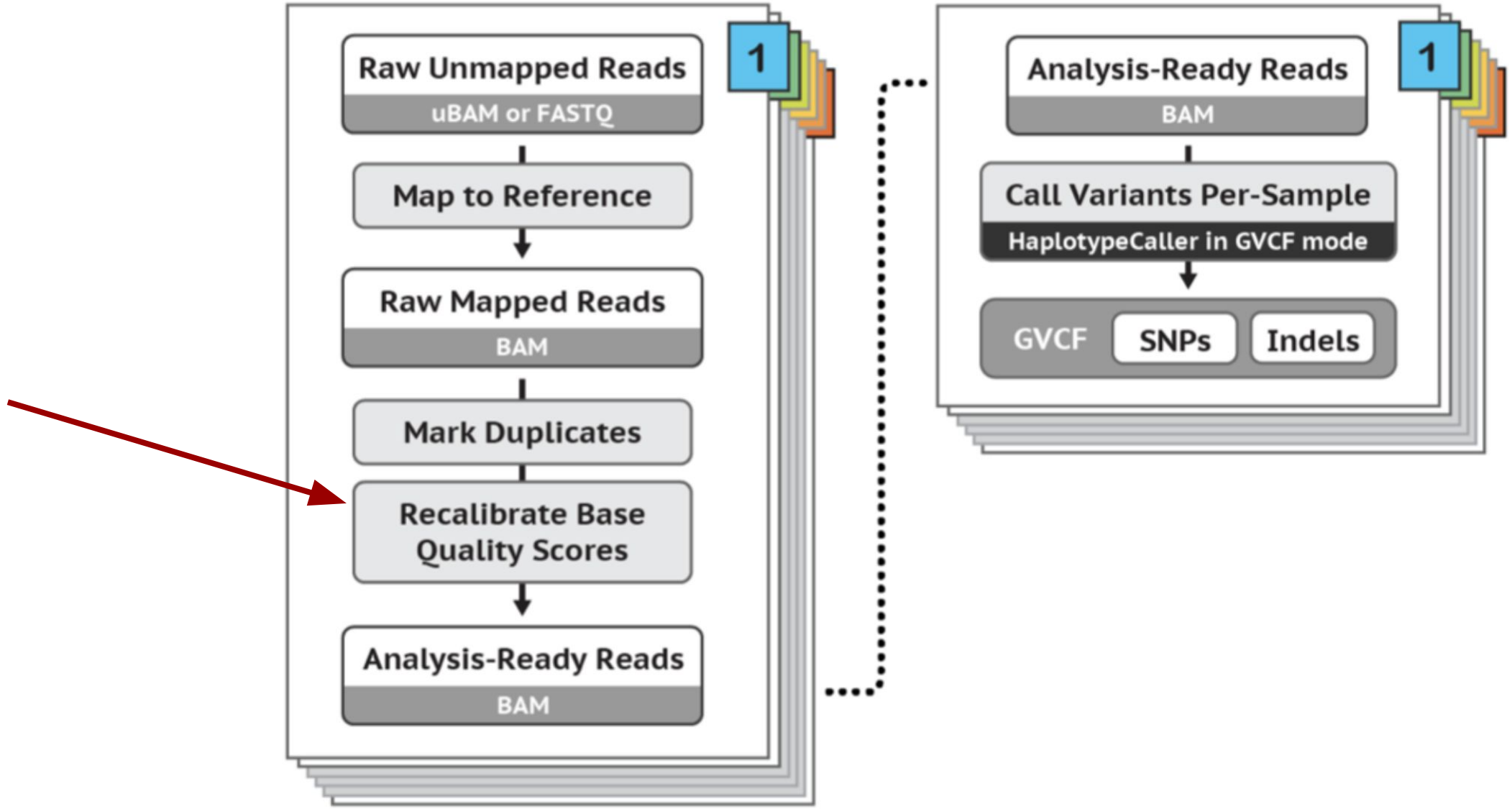
http://picard.sourceforge.net/

Duplicate/marking removal

Problems:

- Does not account for sequencing errors.
- Does not account for natural duplicates.
- Does not account for duplicate reads with different mapping locations.

GATK's recommended workflow



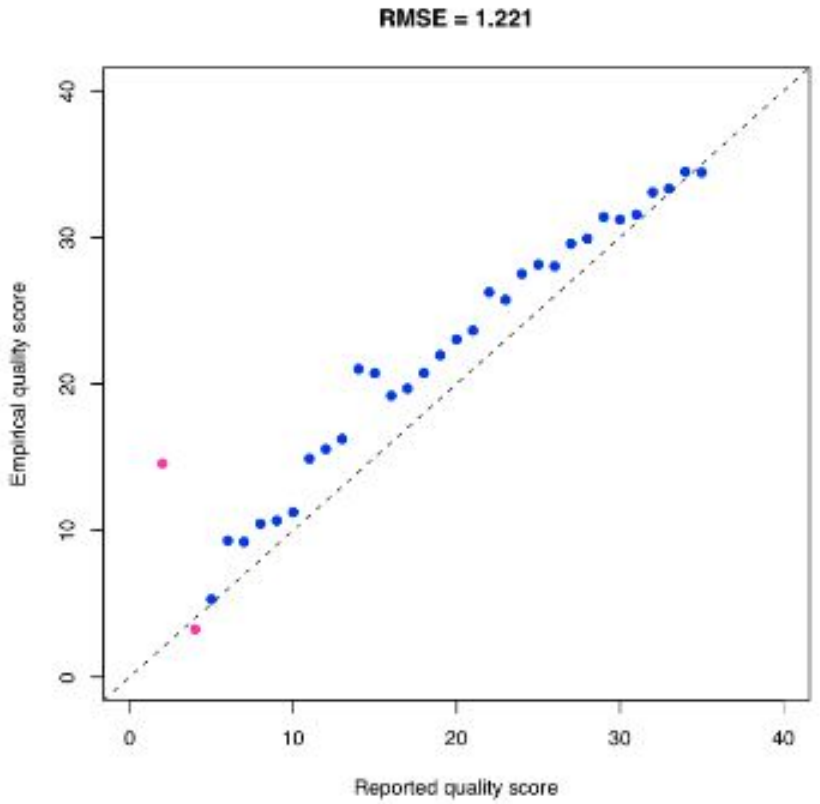
https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Base quality score recalibration?

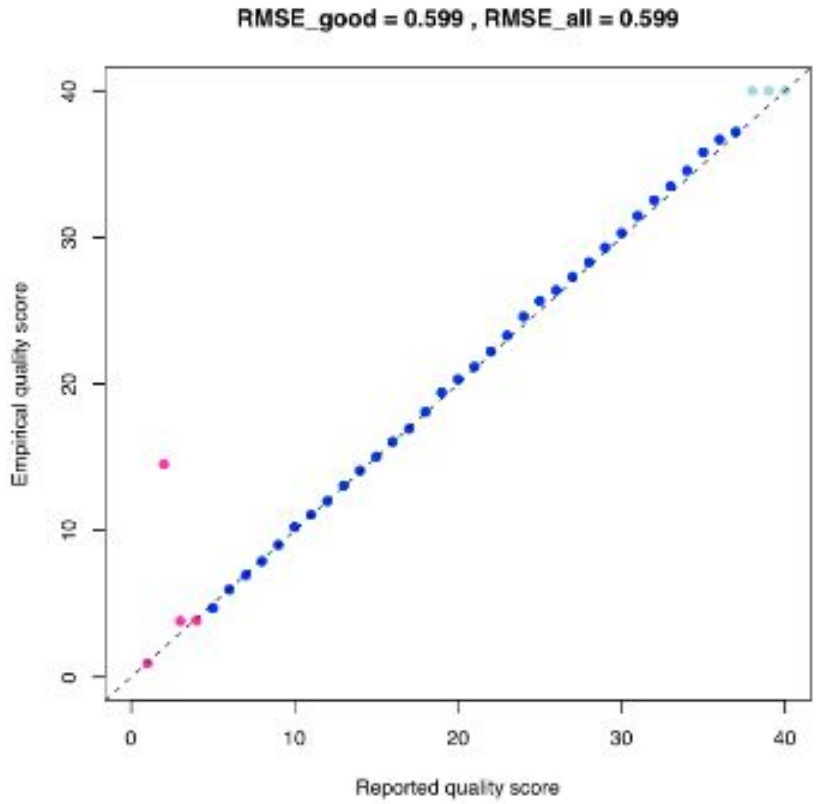
- remember those?
- There are supposed to reflect P(error)
- They are not always accurate: problem for genotyping

Reported Quality vs. Empirical Quality

Idea: use documented variants in the genome



Original Data



After GATK Recalibration

COMMENTARY | VOLUME 177, ISSUE 1, P26-31, MARCH 21, 2019

The Missing Diversity in Human Genetic Studies

Giorgio Sirugo [6] • Scott M. Williams [6] • Sarah A. Tishkoff [6] • Show footnotes

DOI: https://doi.org/10.1016/j.cell.2019.02.048 • Check for updates

The majority of studies of genetic association with disease have been performed in Europeans. This European bias has important implications for risk prediction of diseases across global populations. In this commentary, we justify the need to study more diverse populations using both empirical examples and theoretical reasoning.

Base quality score recalibration

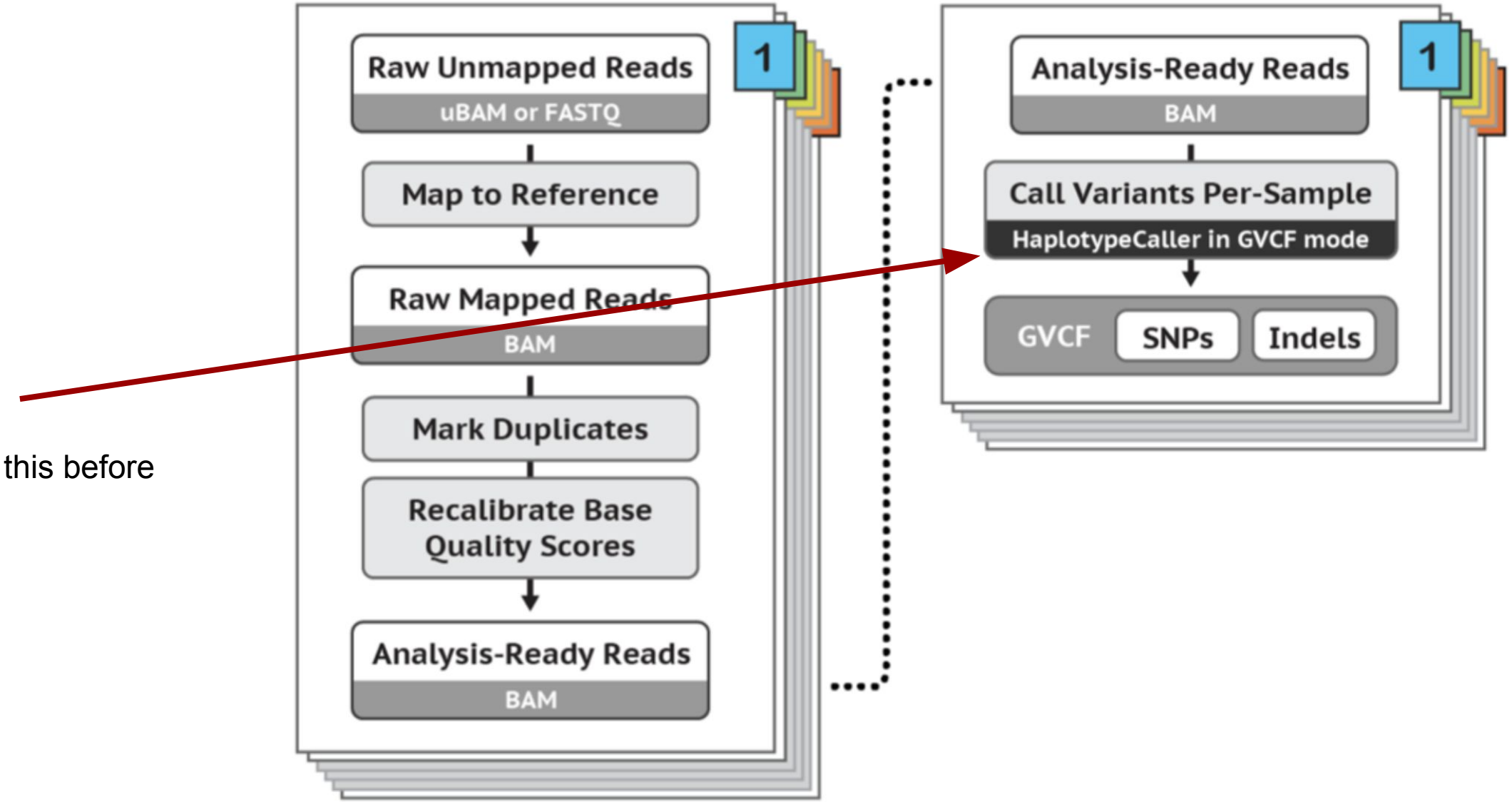
To work we need:

- East Asian or European (as in mostly West European) samples
- WGS
- Sufficient coverage

My biased opinion:

- Just don’t bother

GATK's recommended workflow



We covered this before

https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Variant call format (VCF)

- Details which variants have been called
- Can be bgzip (block gzip) and indexed using tabix
- Using tabix, queries can be made like:
  - return all variants in the region chr22:323,340-361,152

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

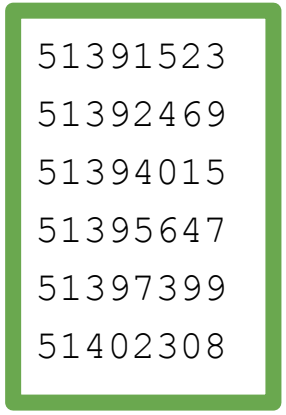
Variant call format (VCF)

```
20   61391523   .   A   G   173.96   .   AC=2;DP=5;MQ=52.03   GT:AD:DP:GQ:PL   1/1:0,5:5:15:188,15,0
20   61392469   .   C   T   146.14   .   AC=2;DP=4;MQ=60.00   GT:AD:DP:GQ:PL   1/1:0,4:4:12:160,12,0
20   61394015   .   T   C   97.64    .   AC=1;DP=6;MQ=60.00   GT:AD:DP:GQ:PL   0/1:3,3:6:66:105,0,66
20   61395647   .   A   C   89.64    .   AC=1;DP=7;MQ=57.28   GT:AD:DP:GQ:PL   0/1:4,3:7:97:97,0,100
20   61397399   .   C   T   93.64    .   AC=1;DP=7;MQ=60.00   GT:AD:DP:GQ:PL   0/1:4,3:7:99:101,0,120
20   61402308   .   C   T   161.64   .   AC=1;DP=9;MQ=60.00   GT:AD:DP:GQ:PL   0/1:3,6:9:63:169,0,63
```

name of chromosome (ex: chr1, chr2 ...

Variant call format (VCF)

```
20  51391523  .  A  G  173.96  .  AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469  .  C  T  146.14  .  AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015  .  T  C  97.64   .  AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647  .  A  C  89.64   .  AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399  .  C  T  93.64   .  AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308  .  C  T  161.64  .  AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```



coordinate on chromosome

Variant call format (VCF)

```
20  51391523  .  A  G  173.96  .  AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469  .  C  T  146.14  .  AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015  .  T  C  97.64   .  AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647  .  A  C  89.64   .  AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399  .  C  T  93.64   .  AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308  .  C  T  161.64  .  AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

ID (ex: rs23534)

Variant call format (VCF)

```
20  51391523  .  A  G  173.96  .  AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469  .  C  T  146.14  .  AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015  .  T  C  97.64   .  AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647  .  A  C  89.64   .  AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399  .  C  T  93.64   .  AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308  .  C  T  161.64  .  AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

reference base

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

alternative base

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

quality field

Variant call format (VCF)

```
20  51391523    .   A   G   173.96   .   AC=2;DP=5;MQ=52.03   GT:AD:DP:GQ:PL   1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14   .   AC=2;DP=4;MQ=60.00   GT:AD:DP:GQ:PL   1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64    .   AC=1;DP=6;MQ=60.00   GT:AD:DP:GQ:PL   0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64    .   AC=1;DP=7;MQ=57.28   GT:AD:DP:GQ:PL   0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64    .   AC=1;DP=7;MQ=60.00   GT:AD:DP:GQ:PL   0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64   .   AC=1;DP=9;MQ=60.00   GT:AD:DP:GQ:PL   0/1:3,6:9:63:169,0,63
```

Filter (ex: ‘LowQual’)

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Info field ex:

- AC= allele count
- DP = depth
- MQ = root mean square of the mapping quality

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Format field, what do the next fields mean?

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Most likely genotype

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Allele distribution

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Depth

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

Genotype quality

Variant call format (VCF)

```
20  51391523    .   A   G   173.96  .   AC=2;DP=5;MQ=52.03  GT:AD:DP:GQ:PL  1/1:0,5:5:15:188,15,0
20  51392469    .   C   T   146.14  .   AC=2;DP=4;MQ=60.00  GT:AD:DP:GQ:PL  1/1:0,4:4:12:160,12,0
20  51394015    .   T   C   97.64   .   AC=1;DP=6;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,3:6:66:105,0,66
20  51395647    .   A   C   89.64   .   AC=1;DP=7;MQ=57.28  GT:AD:DP:GQ:PL  0/1:4,3:7:97:97,0,100
20  51397399    .   C   T   93.64   .   AC=1;DP=7;MQ=60.00  GT:AD:DP:GQ:PL  0/1:4,3:7:99:101,0,120
20  51402308    .   C   T   161.64  .   AC=1;DP=9;MQ=60.00  GT:AD:DP:GQ:PL  0/1:3,6:9:63:169,0,63
```

PHRED-scaled likelihood

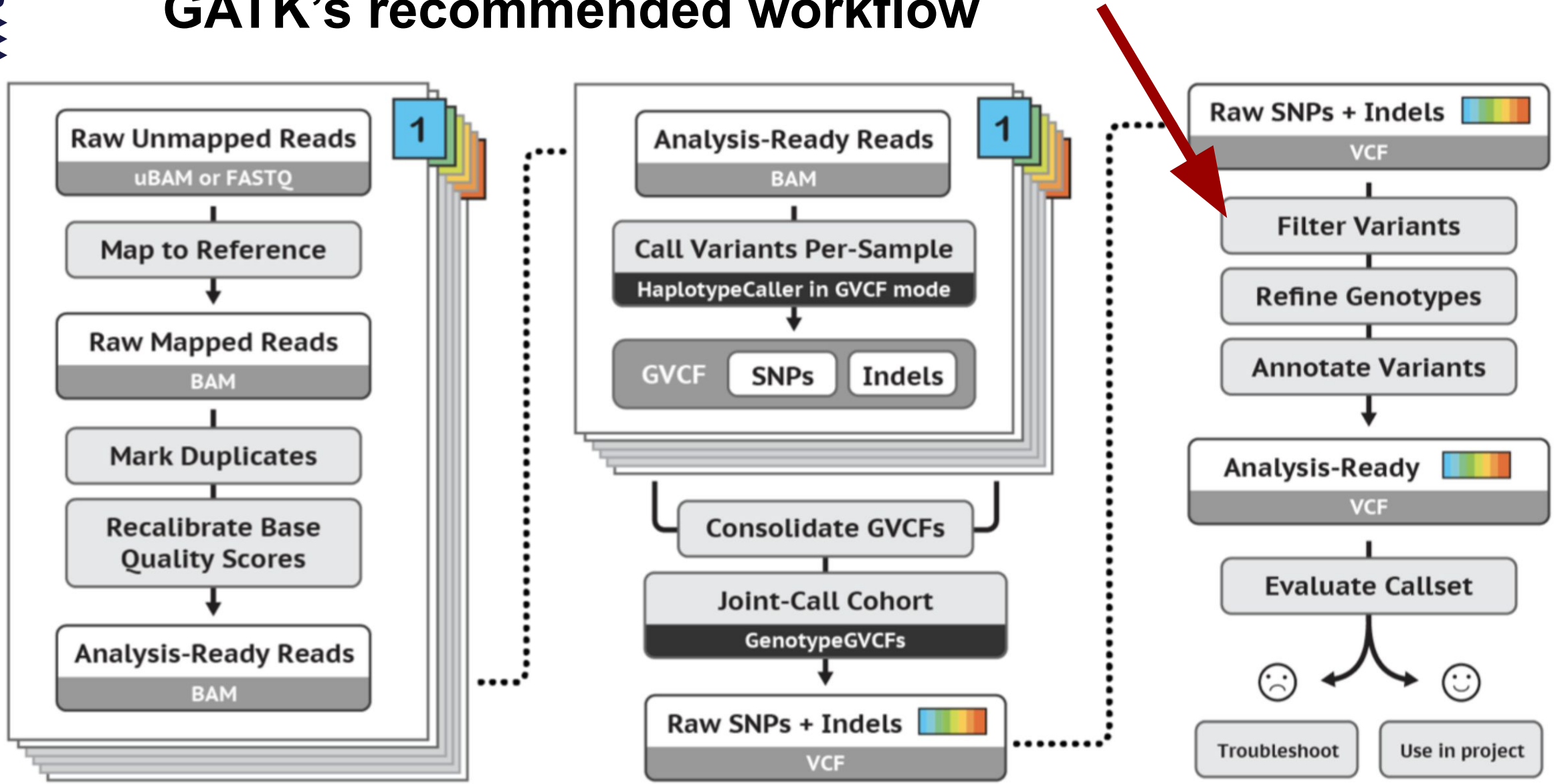
The likelihood $P(D|G)$

PL field

	PHRED	PHRED-scaled
$P(GG|D)$ =	6.7e-05	41.70	40.60
$P(GT|D)$ =	0.77888	1.09	0.00
$P(TT|D)$ =	0.22104	6.56	5.47

GQ field

GATK's recommended workflow



https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Variant filtration (soft)

- How do we remove false positive calls?
- Use known polymorphic sites to estimate what a real variant and a false variant “looks like”
- Learn how does the known sites (=truth set) look like in our data
- Evaluate on all our data, filter sites that look different!

Train a predictor & Test:

Known polymorphic sites

Pos 123
Pos 345
Pos 778

Our called sites

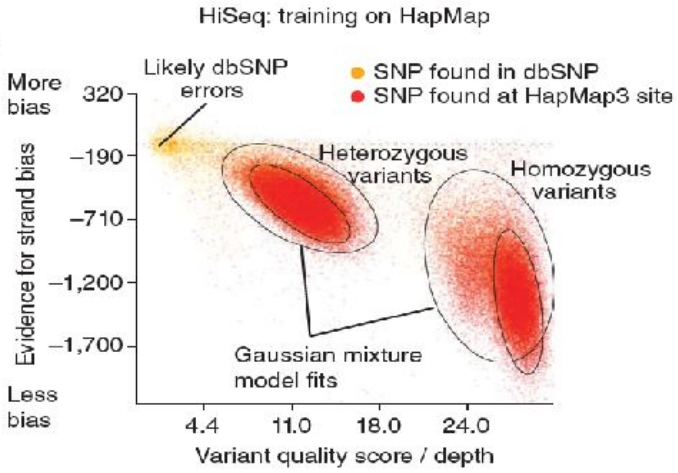
Pos 55
Pos 123
Pos 345
Pos 400
Pos 665
Pos 988

Take out overlap

Pos 123
Pos 345

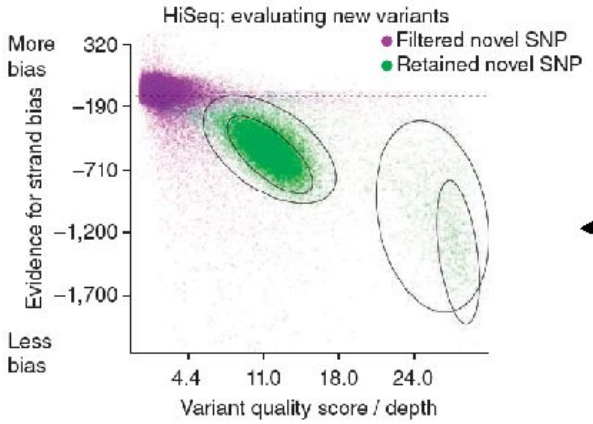
How does these sites look in our data?

Pos 123: GQ, Depth, Strand bias, …
Pos 345: GQ, Depth, Strand bias, …



Apply models to all our sites

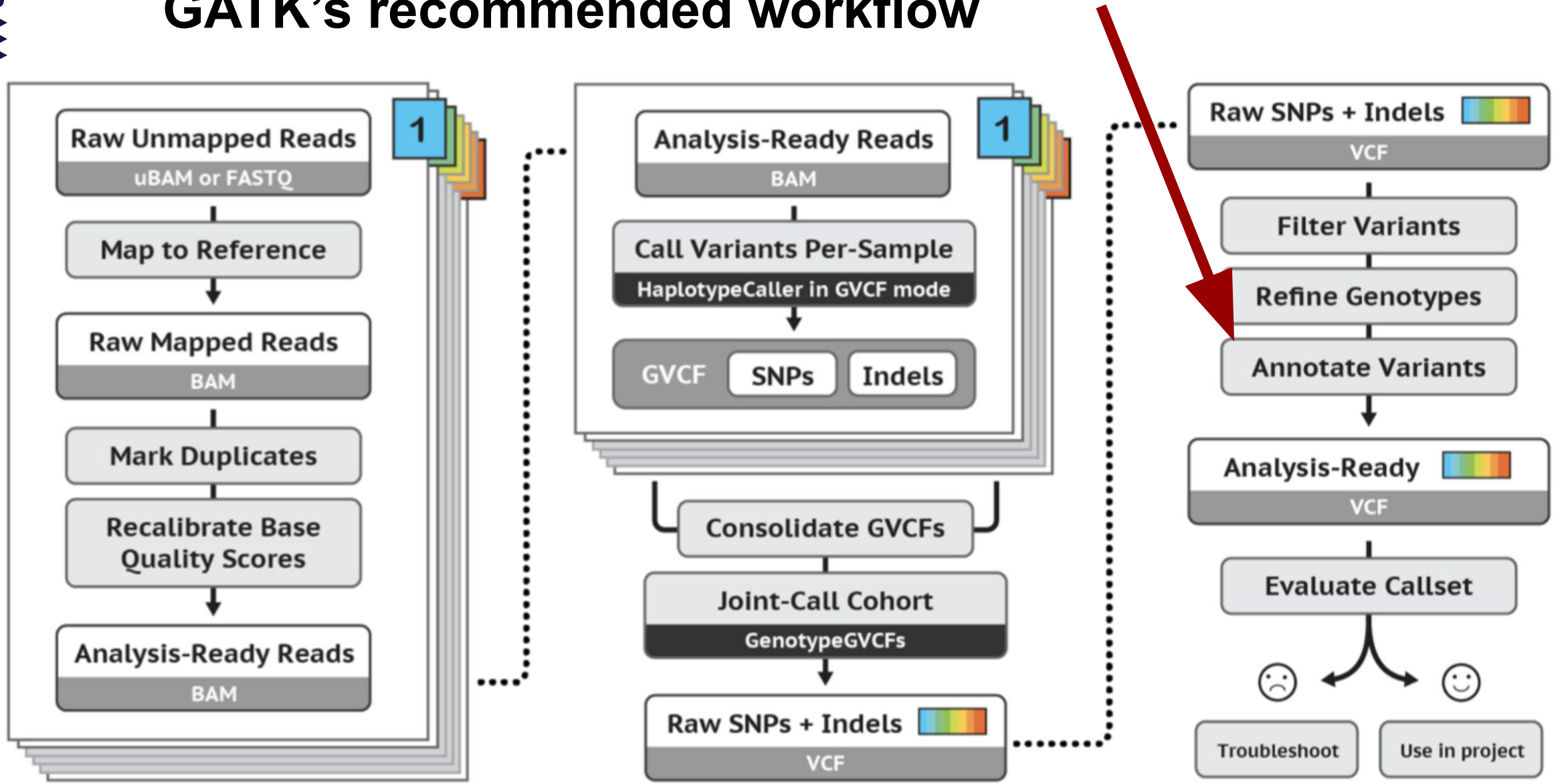
Pos 55
Pos 123
Pos 345
Pos 400
Pos 665
Pos 988



Variant filtration (hard)

- Hard filtering:
  - Variant quality score /depth
  - Mapping quality
  - Mappability
  - Strand bias (the variant being seen only on the forward strand or only on the reverse strand)
  - Depth
- BCFtools can perform this
- Depends on the project at hand
- Be careful of introducing a bias in favor of certain types of variants

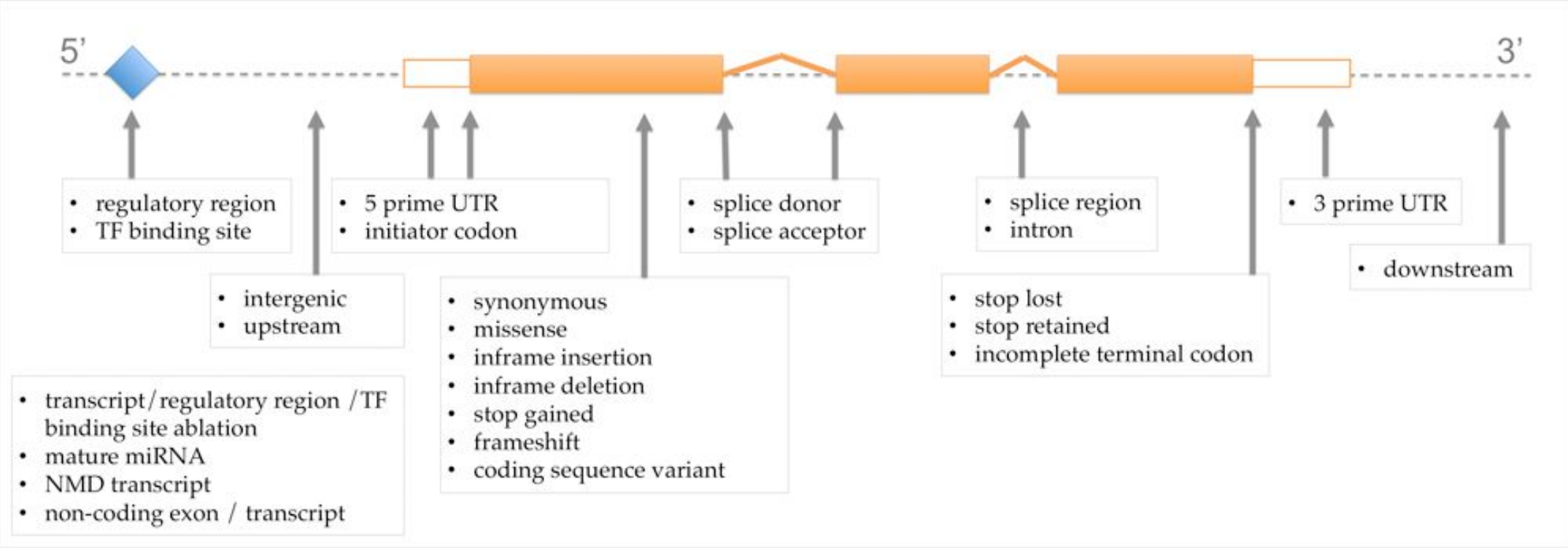
GATK's recommended workflow



https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

Variant annotation

What does the SNP do?



Variant annotation

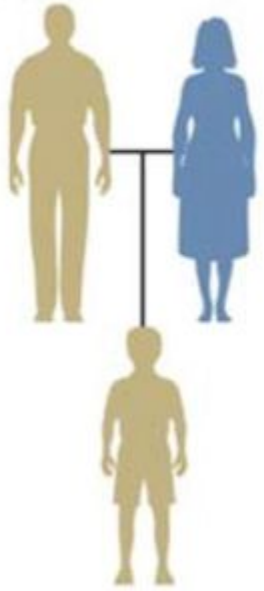
- Some example of tools:
  - Annovar
  - Ensembl Variant Effect Predictor (VEP)
  - SnpEff
- As good as annotations
- Beware of gene expression

we did not cover (in detail)...

Germline vs somatic

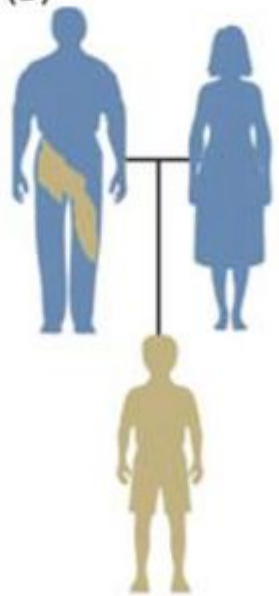
Inherited

(A)



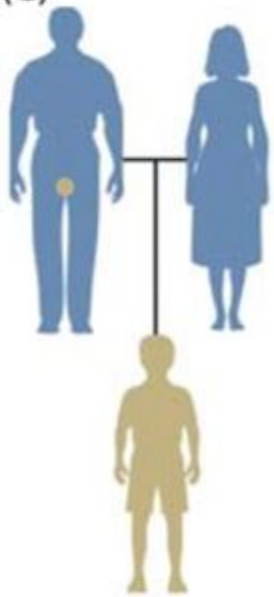
Father has mutation in all cells and transmits it on to his child. Child is heterozygous in every cell.

(B)



Father has mosaic mutation that affects germline and somatic cells. Child is heterozygous in every cell.

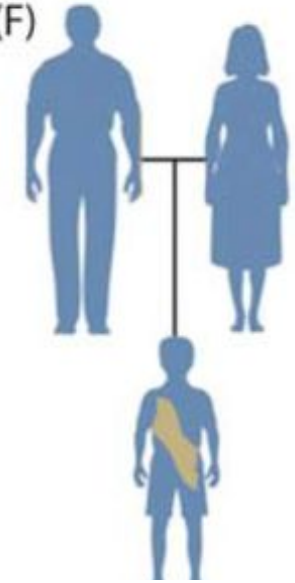
(C)



Father has germline mosaic mutation. Child is heterozygous in every cell.

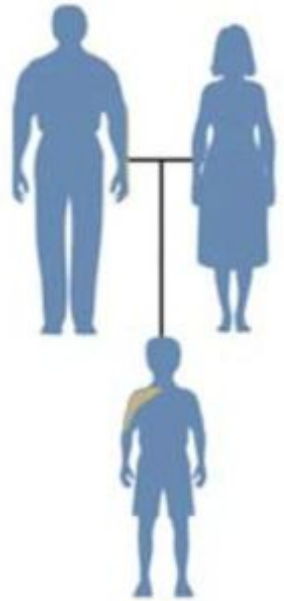
Somatic

(F)



Child has mosaic somatic mutation that occurrs early in postzygotic development and is present in a percentage of his cells.

(G)

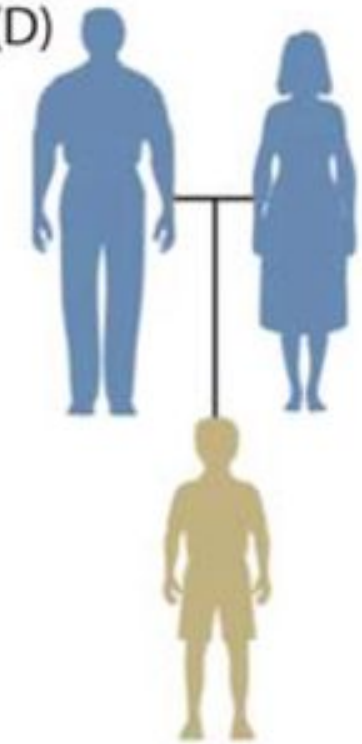


Child has mosaic mutation that occurrs later in development and affects fewer cells (e.g. skin cells)

de novo

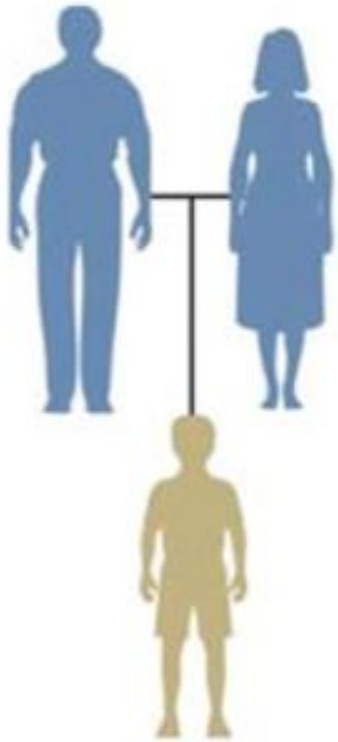
De novo

(D)



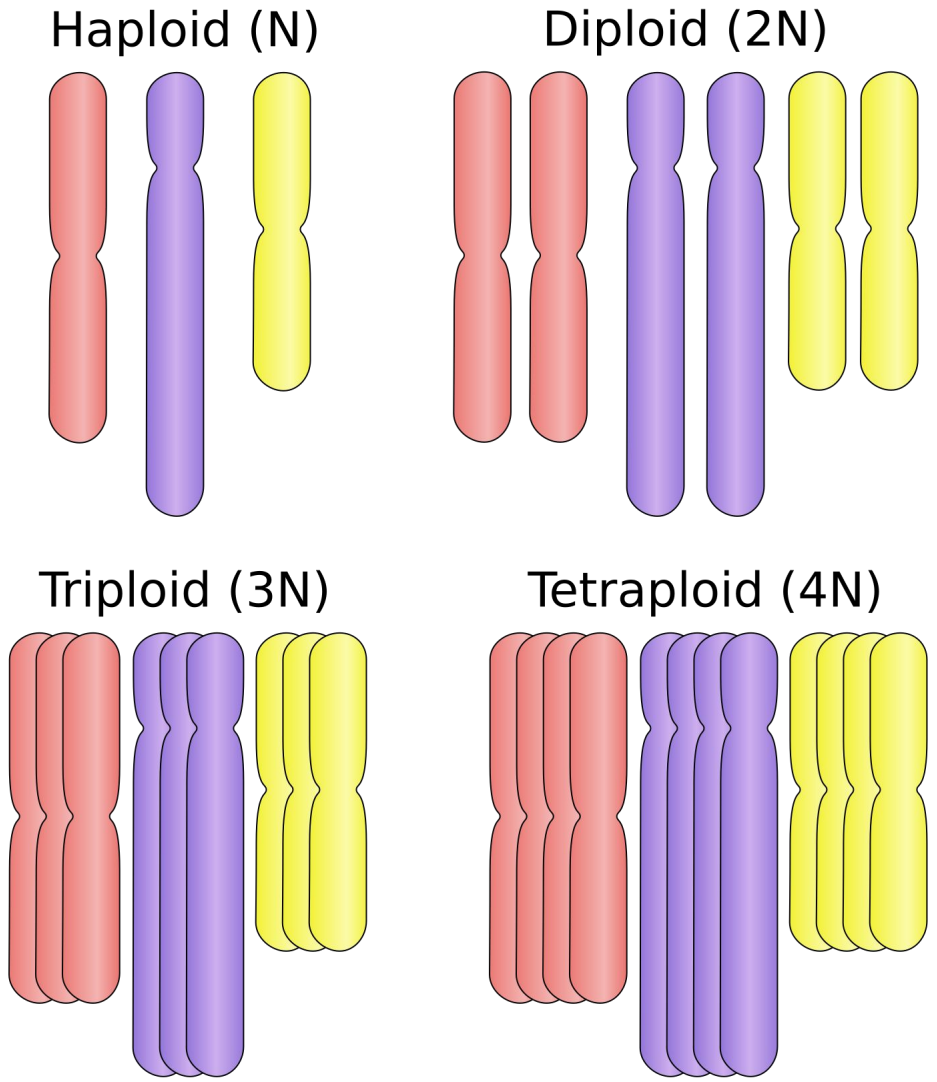
Father has mutation in a single sperm cell and transmits it to the child. Child is heterozygous in every cell.

(E)

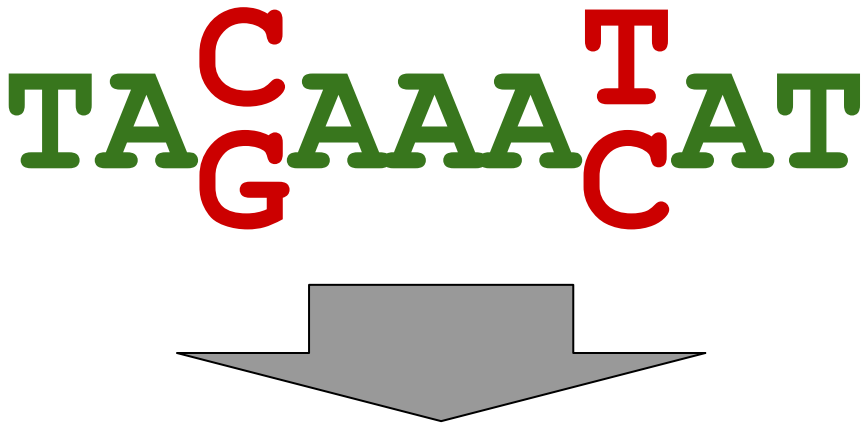


Mutation occurs in zygote within first few cell divisions. Child is heterozygous in every cell.

Polyploid



Phasing



TACAAATAT

vs

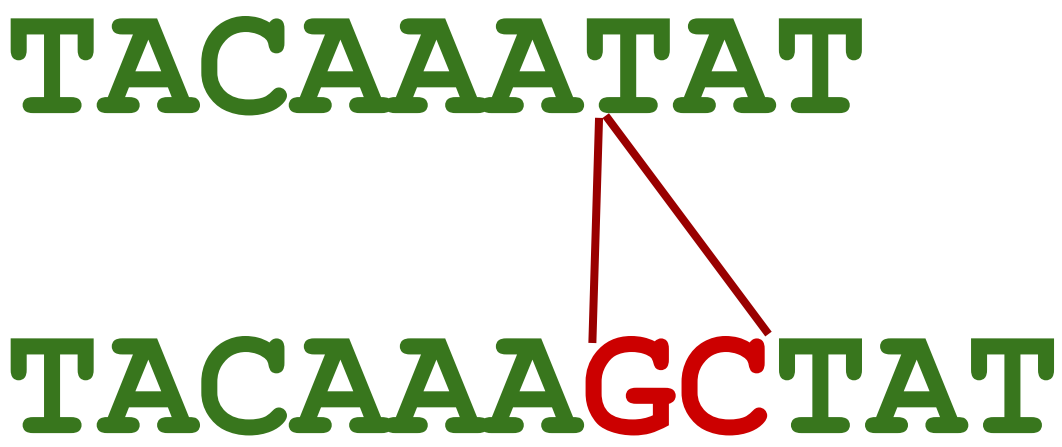
TAGAAACAT

TACAAACAT

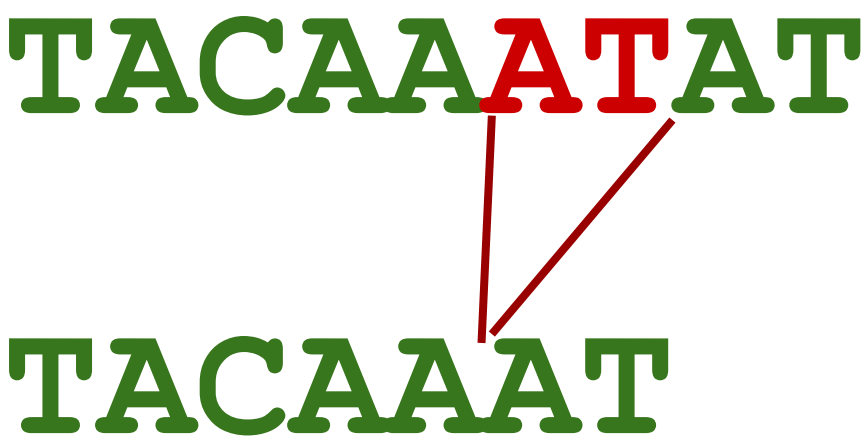
TAGAAATAT

INDELs

Insertions

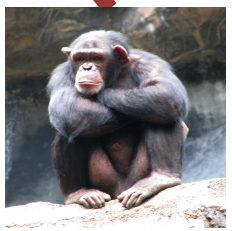


Deletion



INDELs

Caution:

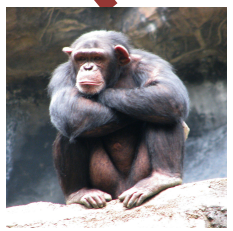


TACAAA--TAT **TACAAAGCTAT**

GC was inserted

INDELs

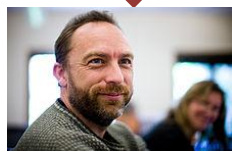
Caution:

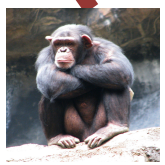


TACAAA--TAT

TACAAAGCTAT

GC was deleted



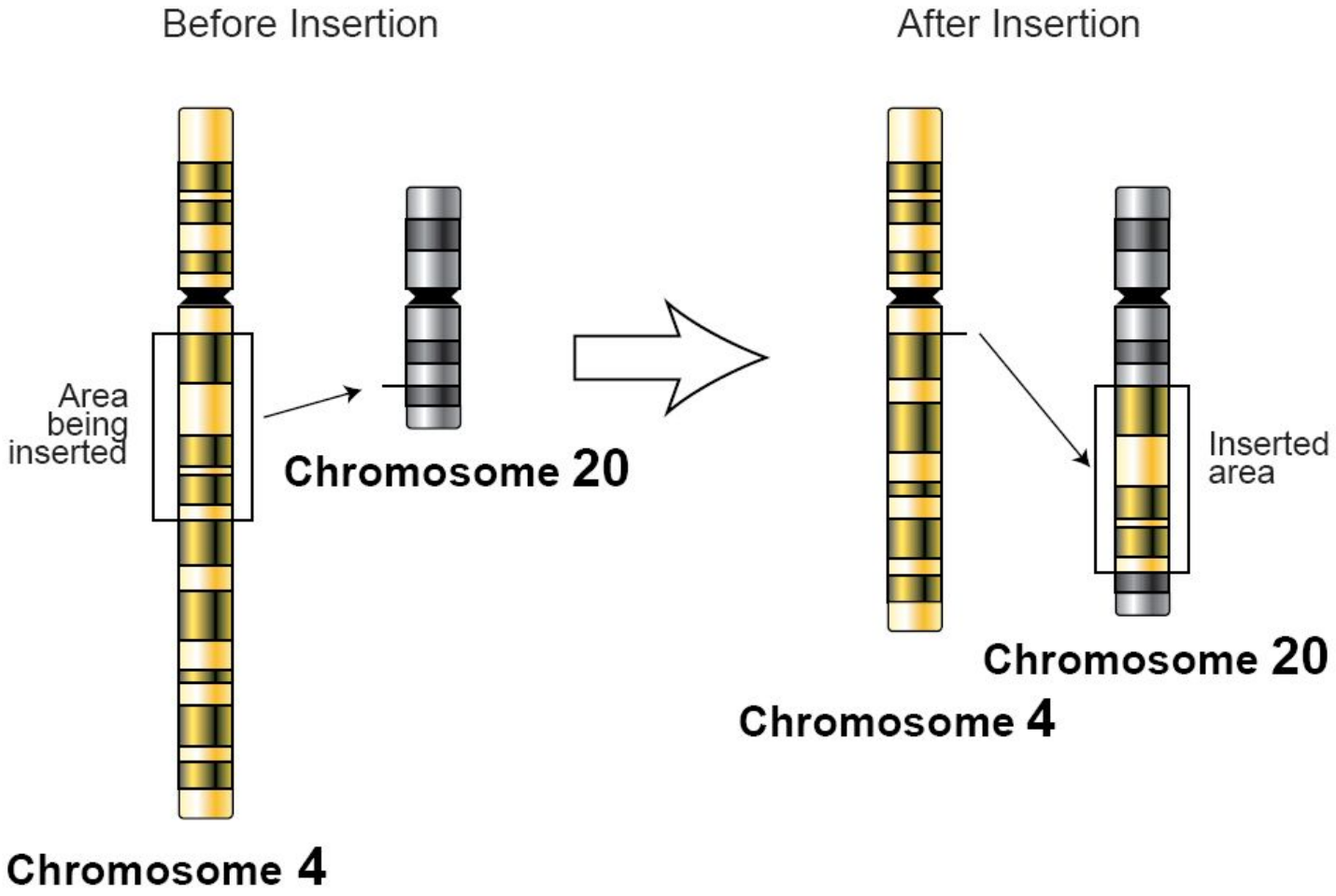


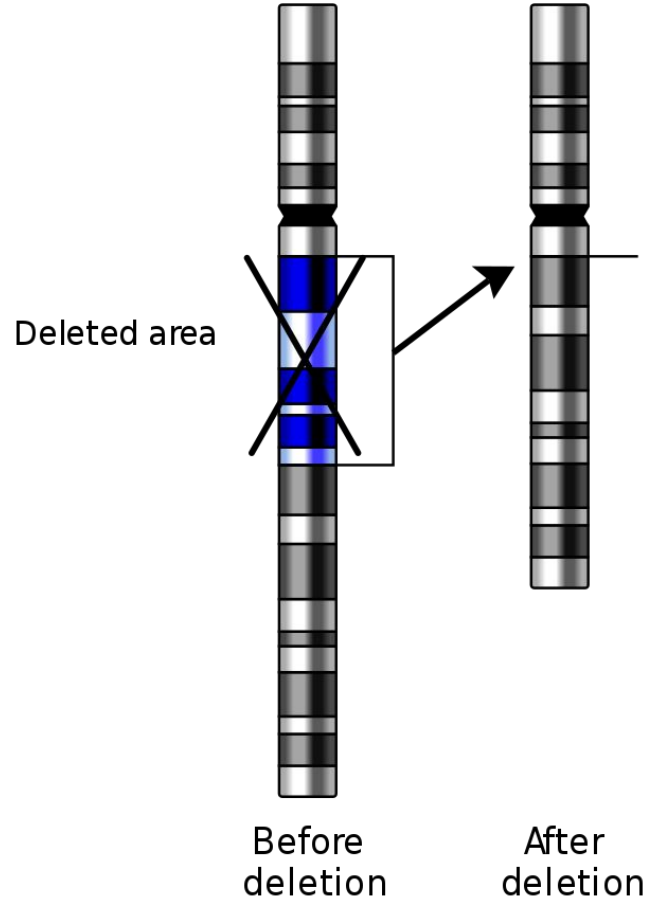
TACAAA--TAT TACAAAGCTAT TACAAAGCTAT

GC was deleted

more likely, not guaranteed!

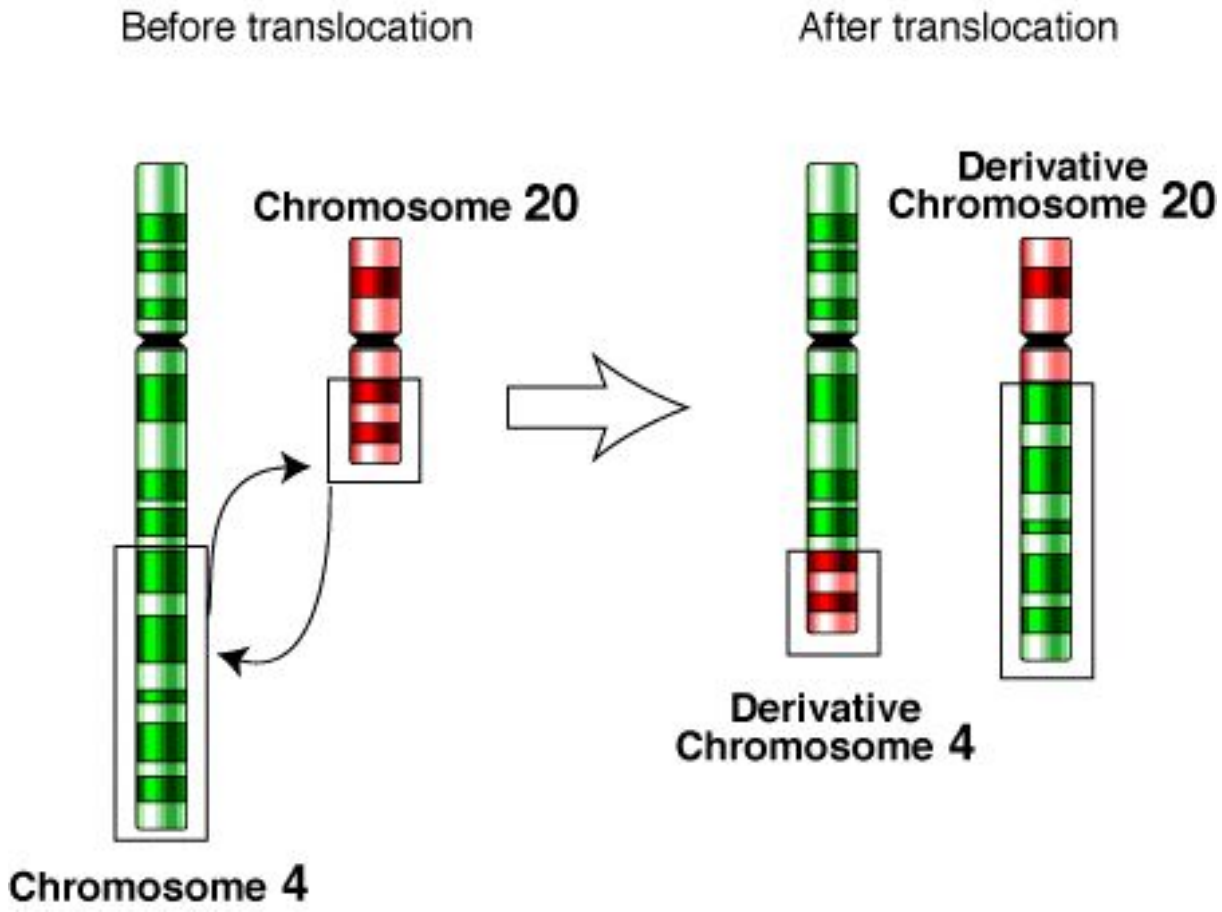
Structural variants





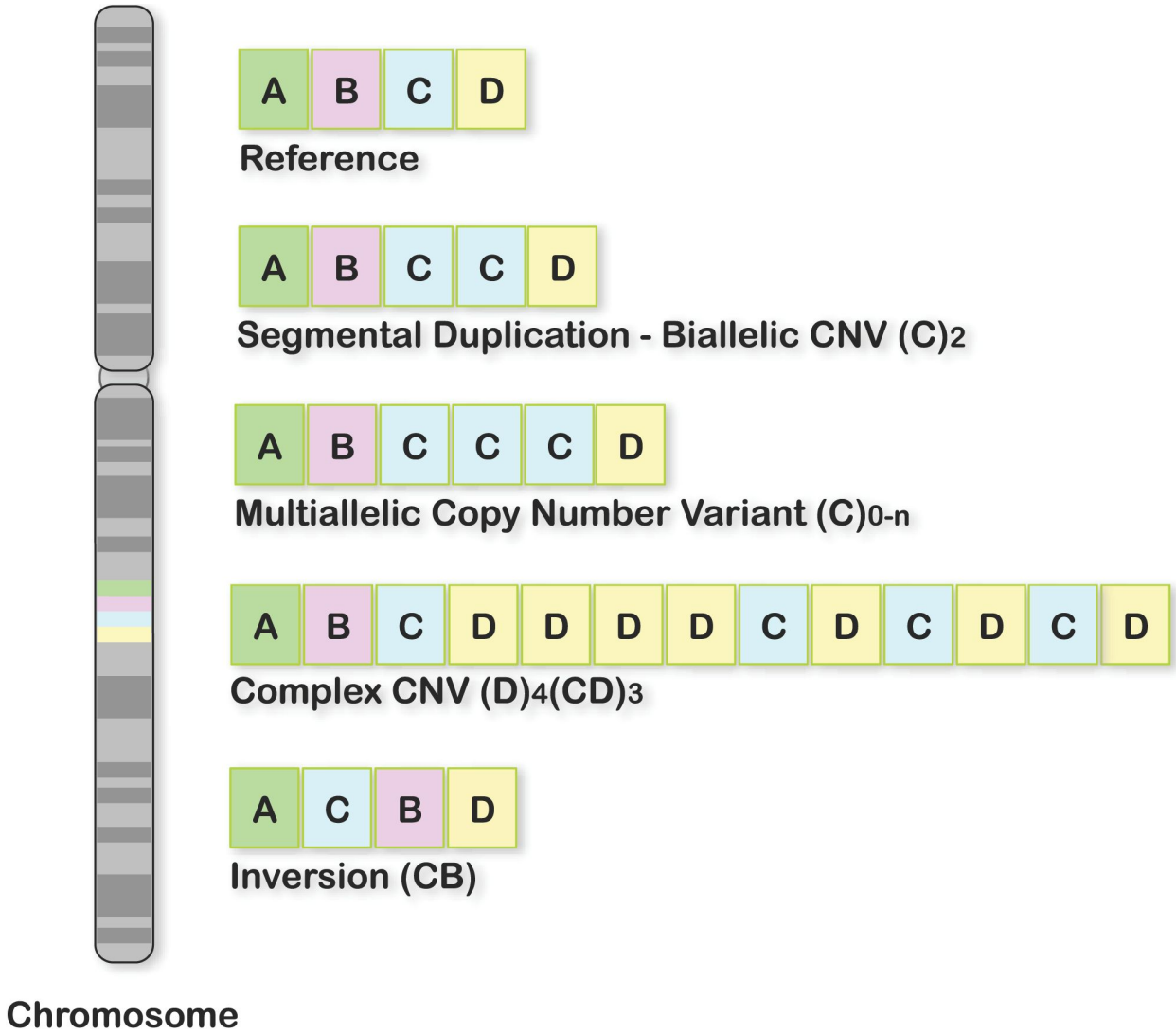
Structural variants

Translocation:



Structural variants

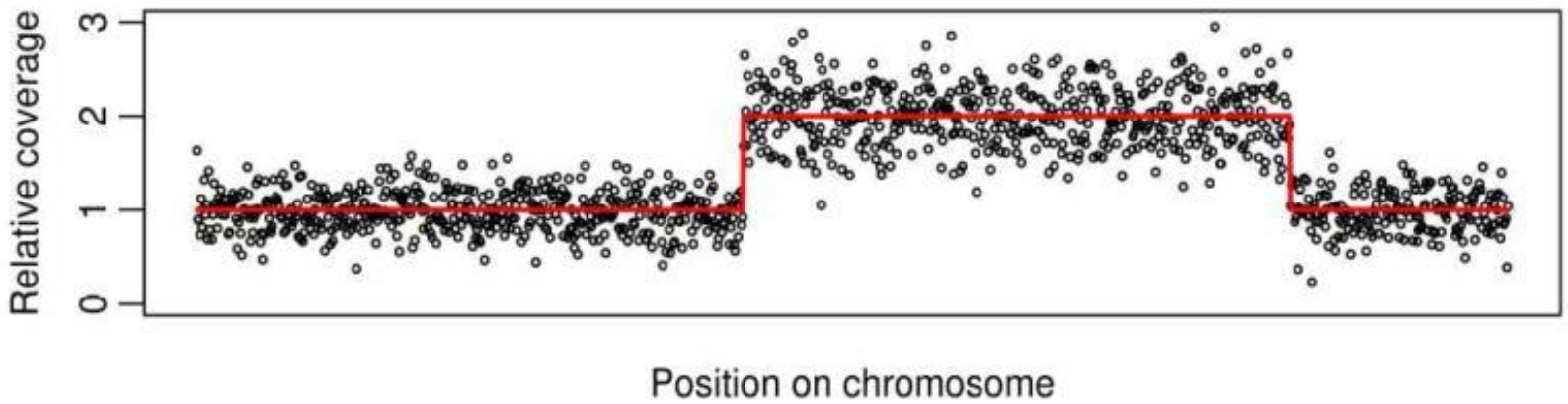
Copy number variations (CNV)



Estivill, Xavier, and Lluís Armengol. "Copy Number Variants and Common Disorders: Filling the Gaps and Exploring Complexity in Genome-Wide Association Studies." PLoS Genet 3.10 (2007): e190.

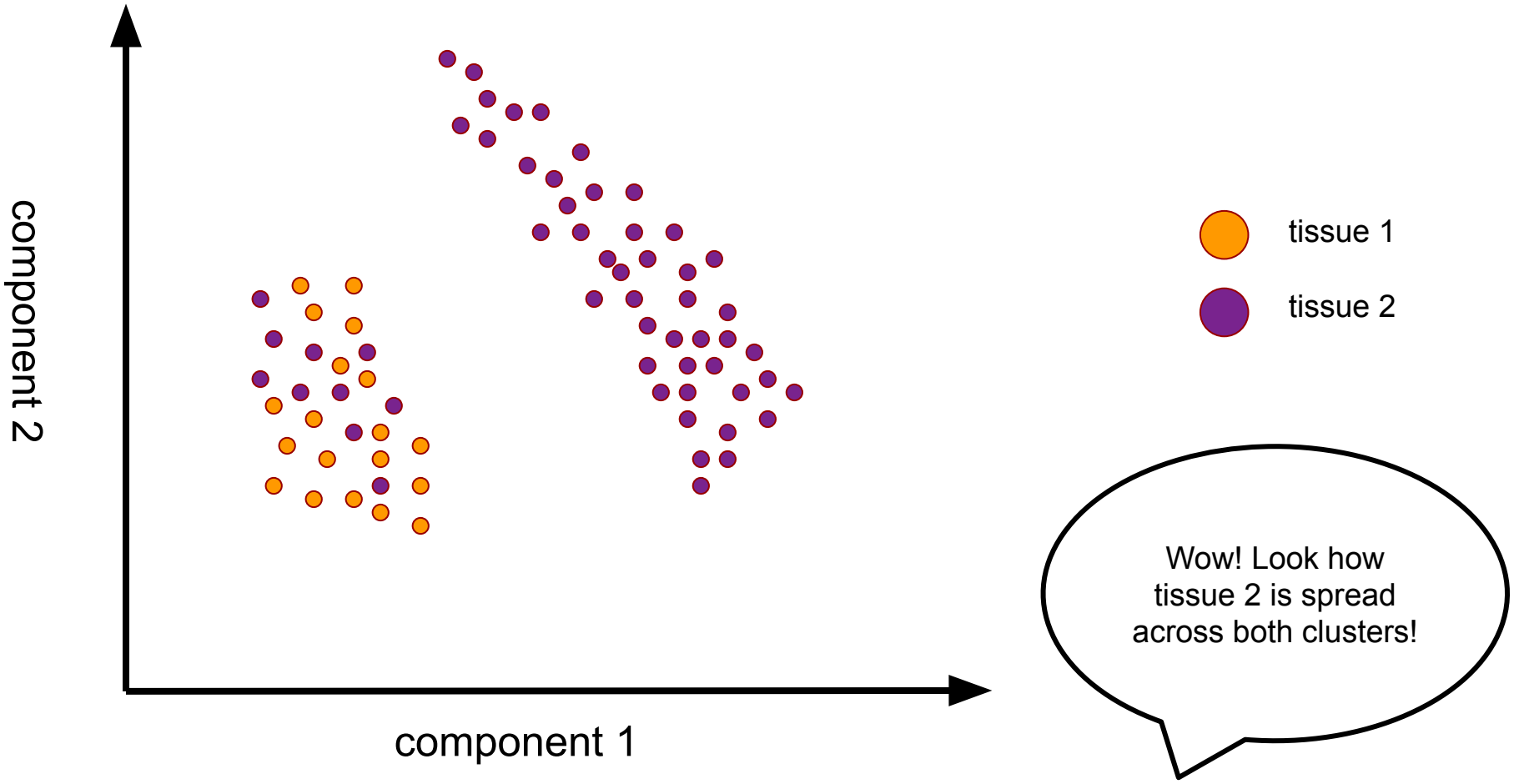
Structural variants

Copy number variations (CNV)
effect on coverage

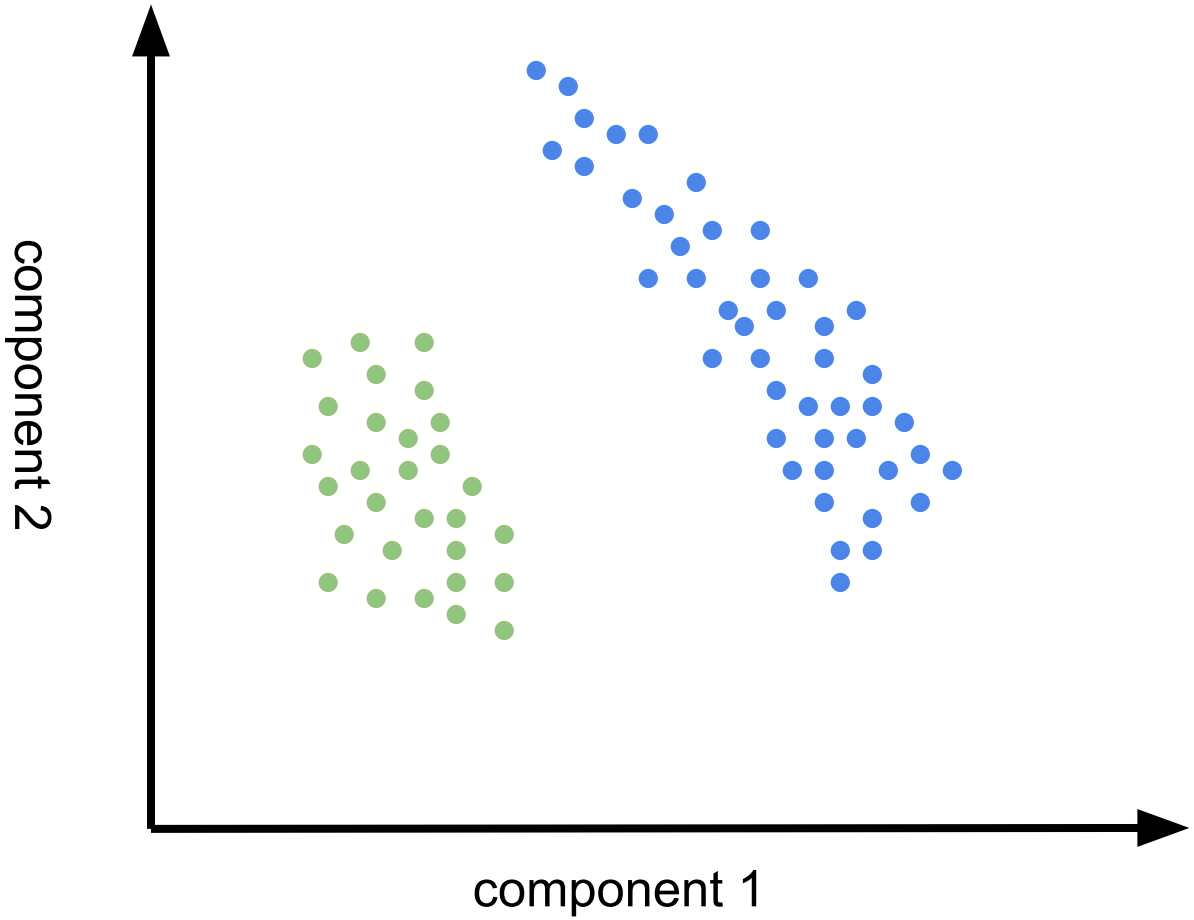


Weetman, David, Luc S. Djogbenou, and Eric Lucas. "Copy number variation (CNV) and insecticide resistance in mosquitoes: evolving knowledge or an evolving problem?." *Current Opinion in Insect Science* 27 (2018): 82-88.

Batch effects



Batch effects



sequencing center #1

sequencing center #2

Wow! Look how tissue 2 is spread across both clusters!

hmm yeah… about that...

Ethical concerns

privacy, justice, fairness etc..

Exercise time!

http://teaching.healthtech.dtu.dk/22126/index.php/Postprocess_exercise