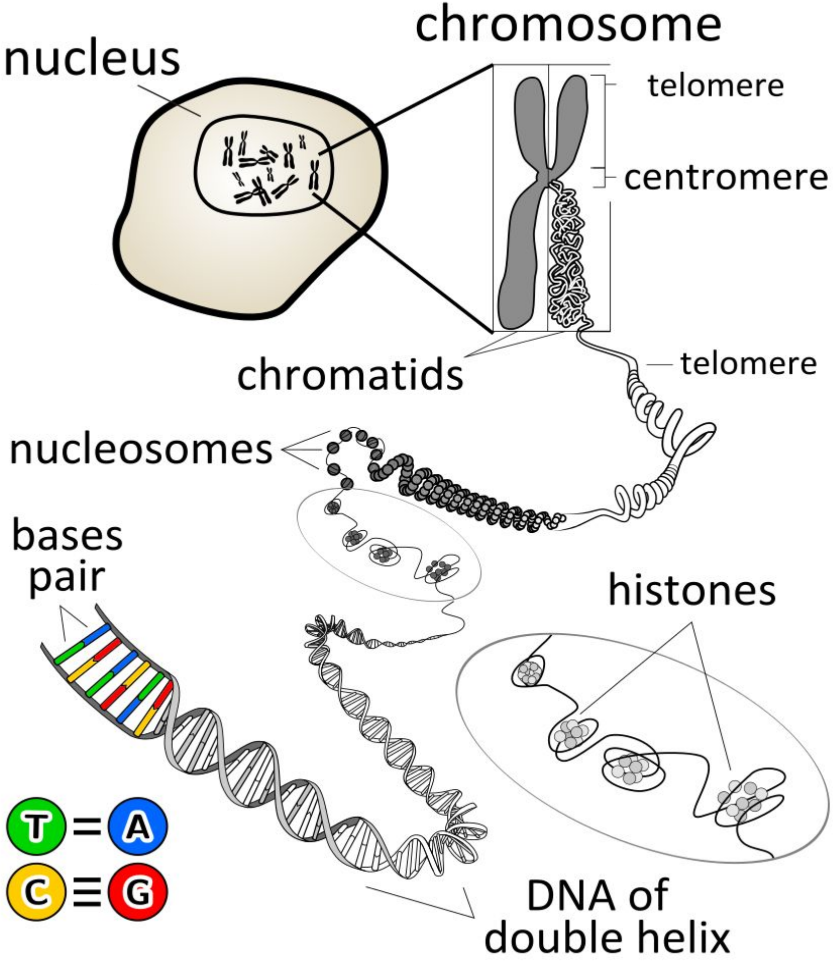
DTU

DTU Health Technology
Bioinformatics

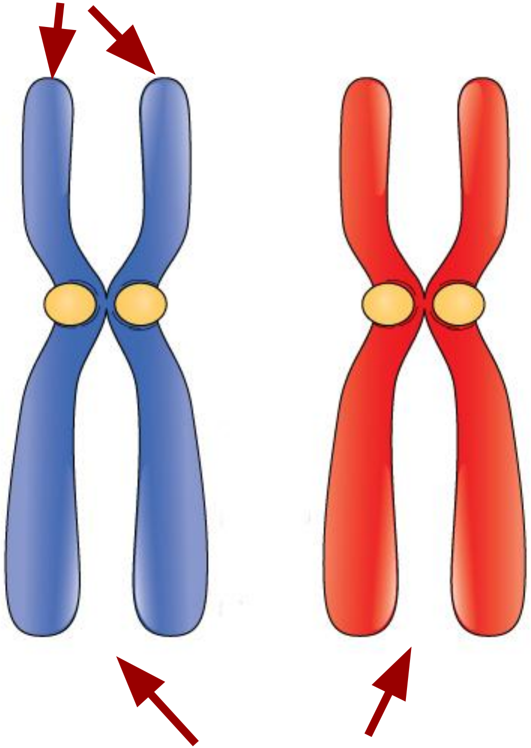
Alignment post-processing and variant calling part 1

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

A brief reminder



A brief reminder



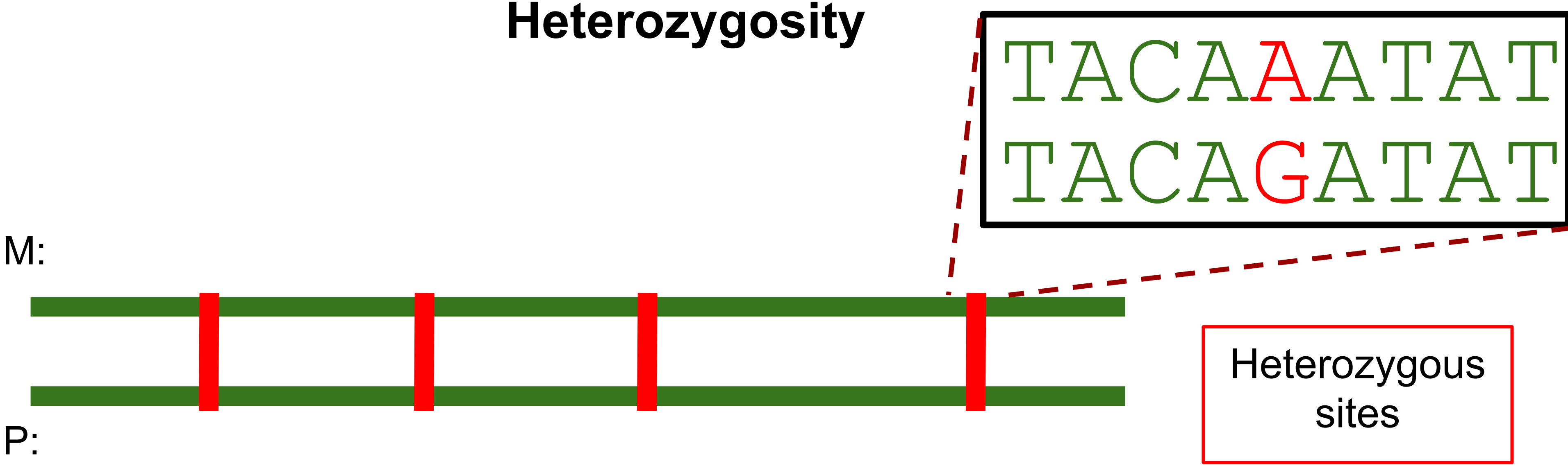
sister chromatids

Paternal

Maternal

homologous chromosomes

Heterozygosity



TACAAATAT

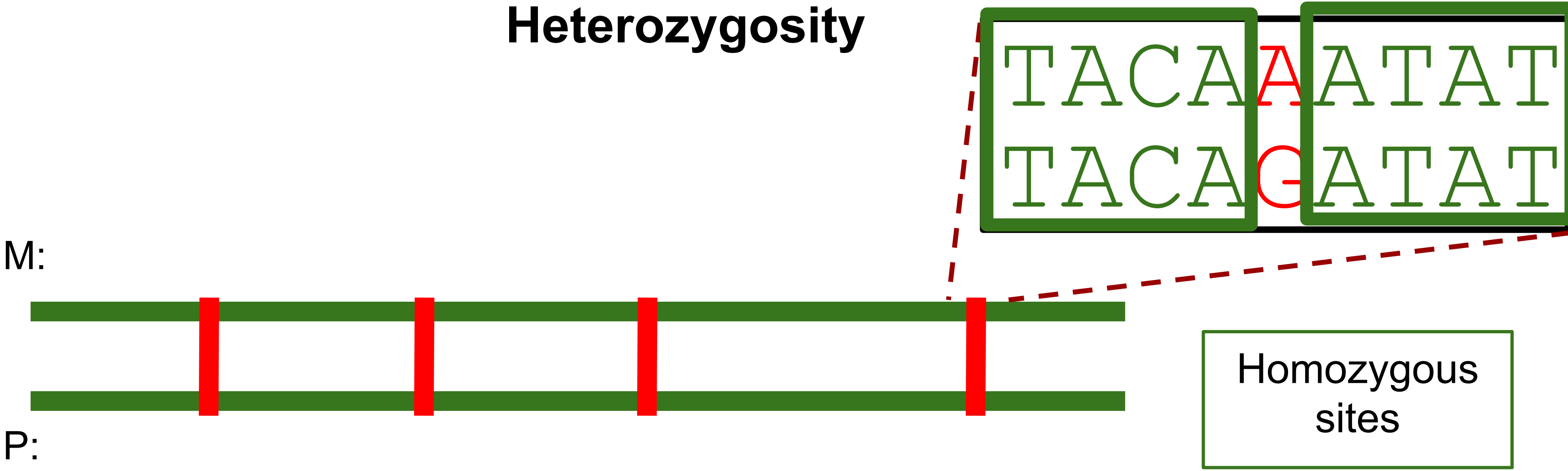
TACAGATAT

M:

P:

Heterozygous sites

Heterozygosity



M:

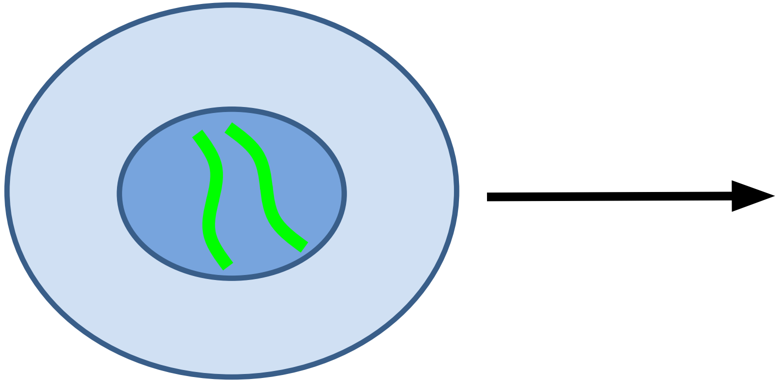
P:

TACAAATAT

TACAGATAT

Homozygous sites

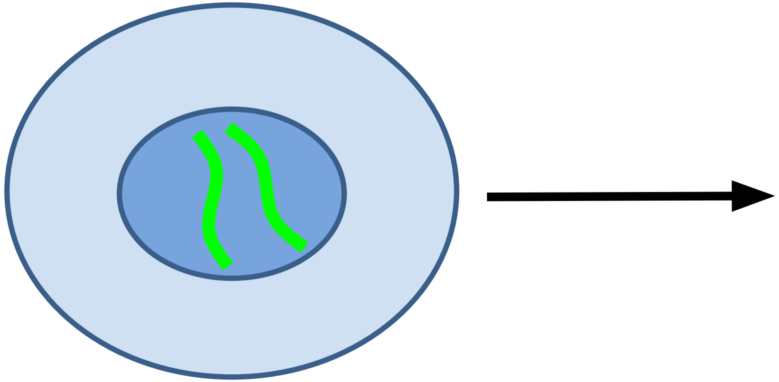
ind#A



M: TACAAATAT

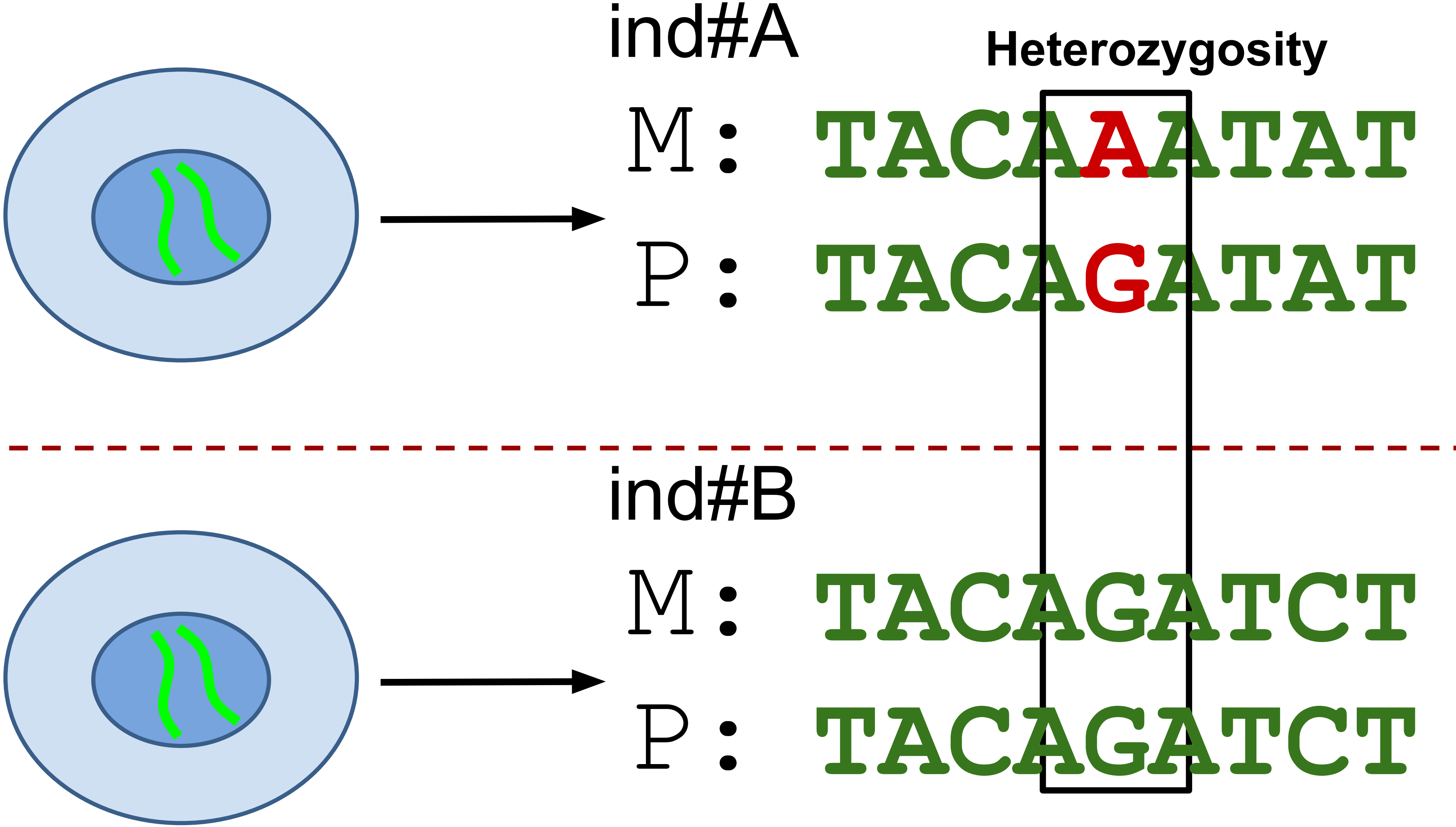
P: TACAGATAT

ind#B



M: TACAGATCT

P: TACAGATCT



ind#A

Heterozygosity

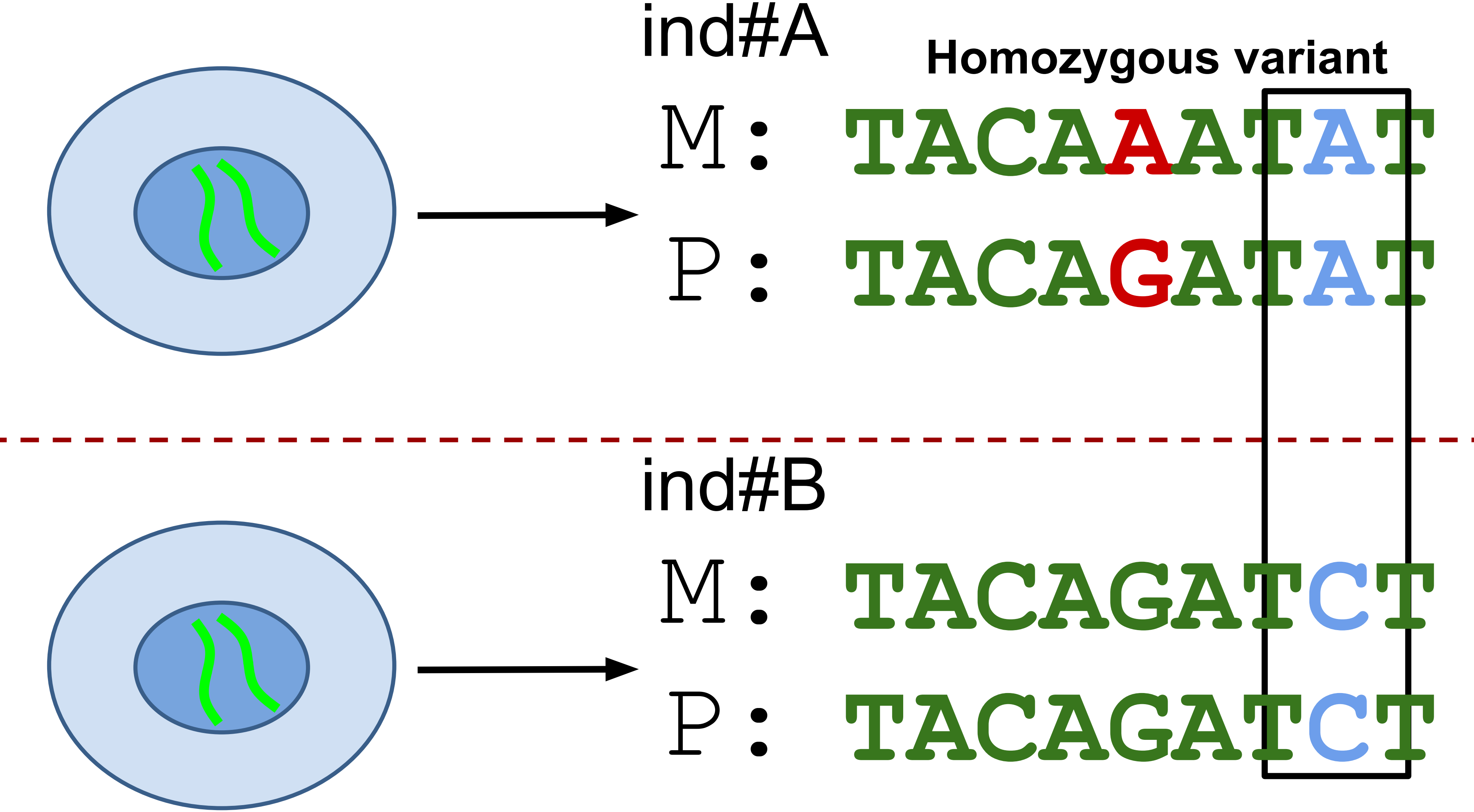
M: TACAAATAT

P: TACAGATAT

ind#B

M: TACAGATCT

P: TACAGATCT



ind#A

Homozygous variant

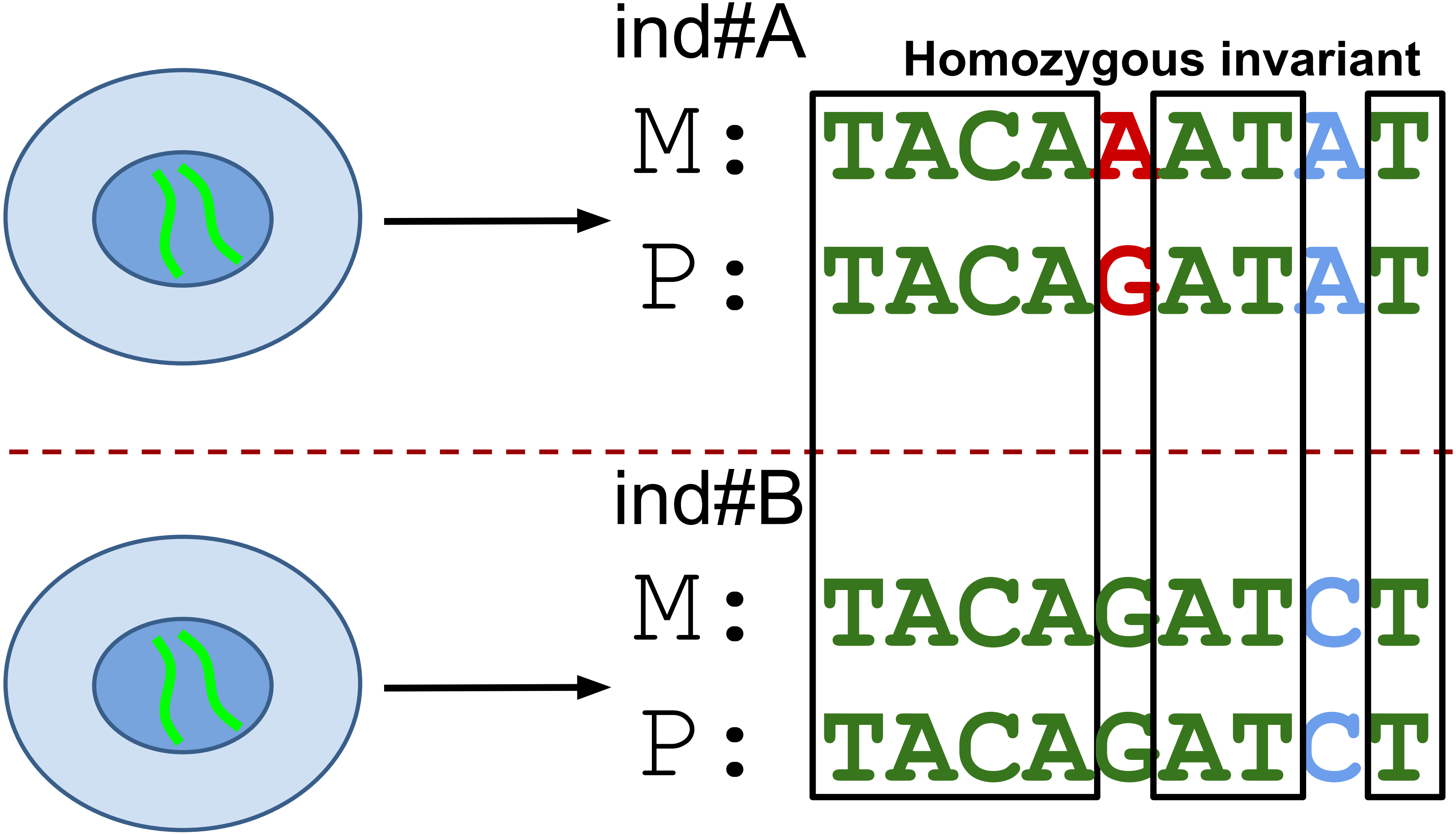
M: **TACAAATAT**

P: **TACAGATAT**

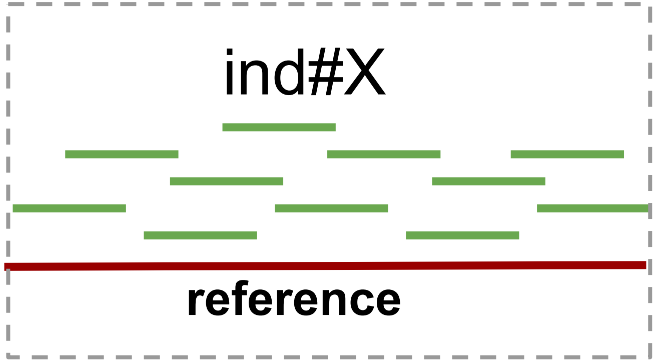
ind#B

M: **TACAGATCT**

P: **TACAGATCT**


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

Genotyping

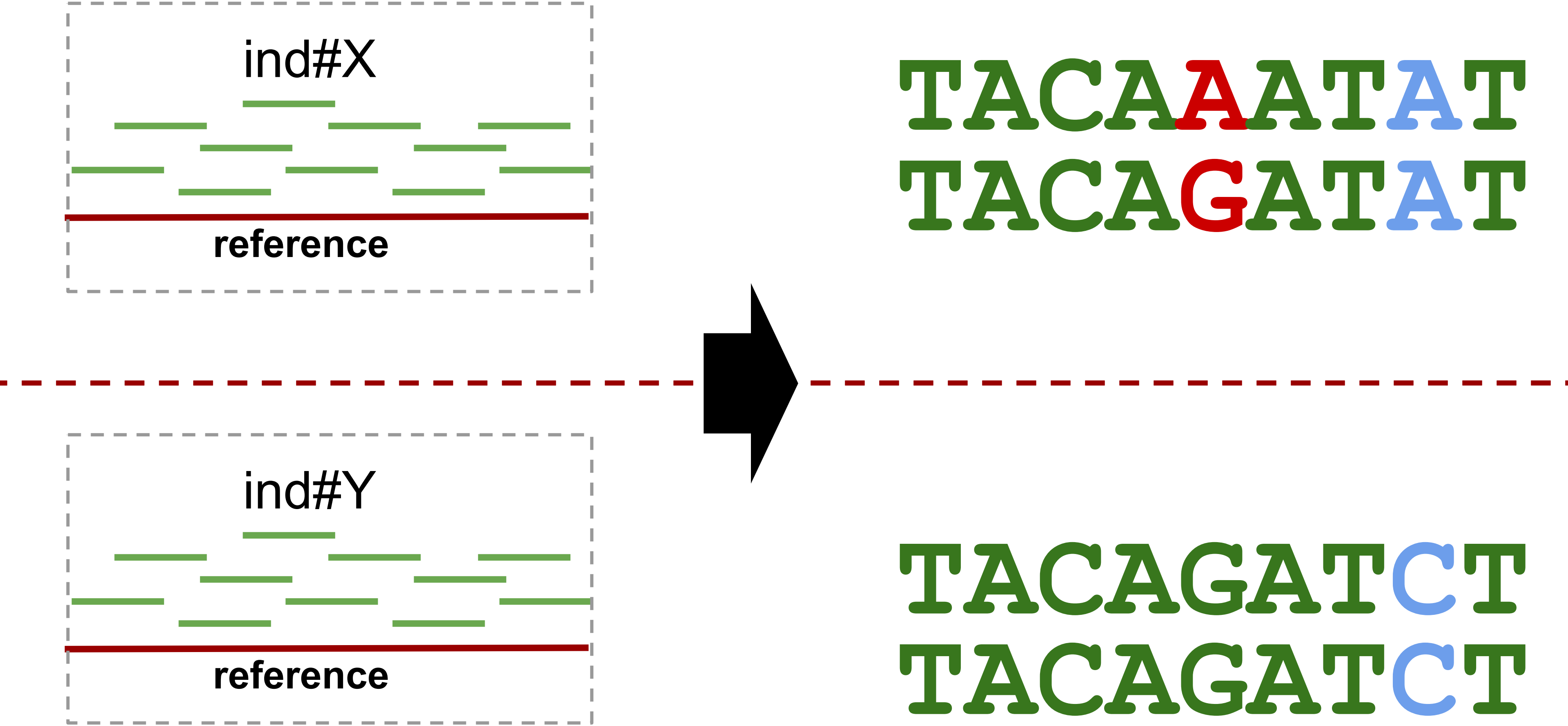


TACAAATAT
TACAGATAT

Which of the 10 possible genotypes is the most likely?

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

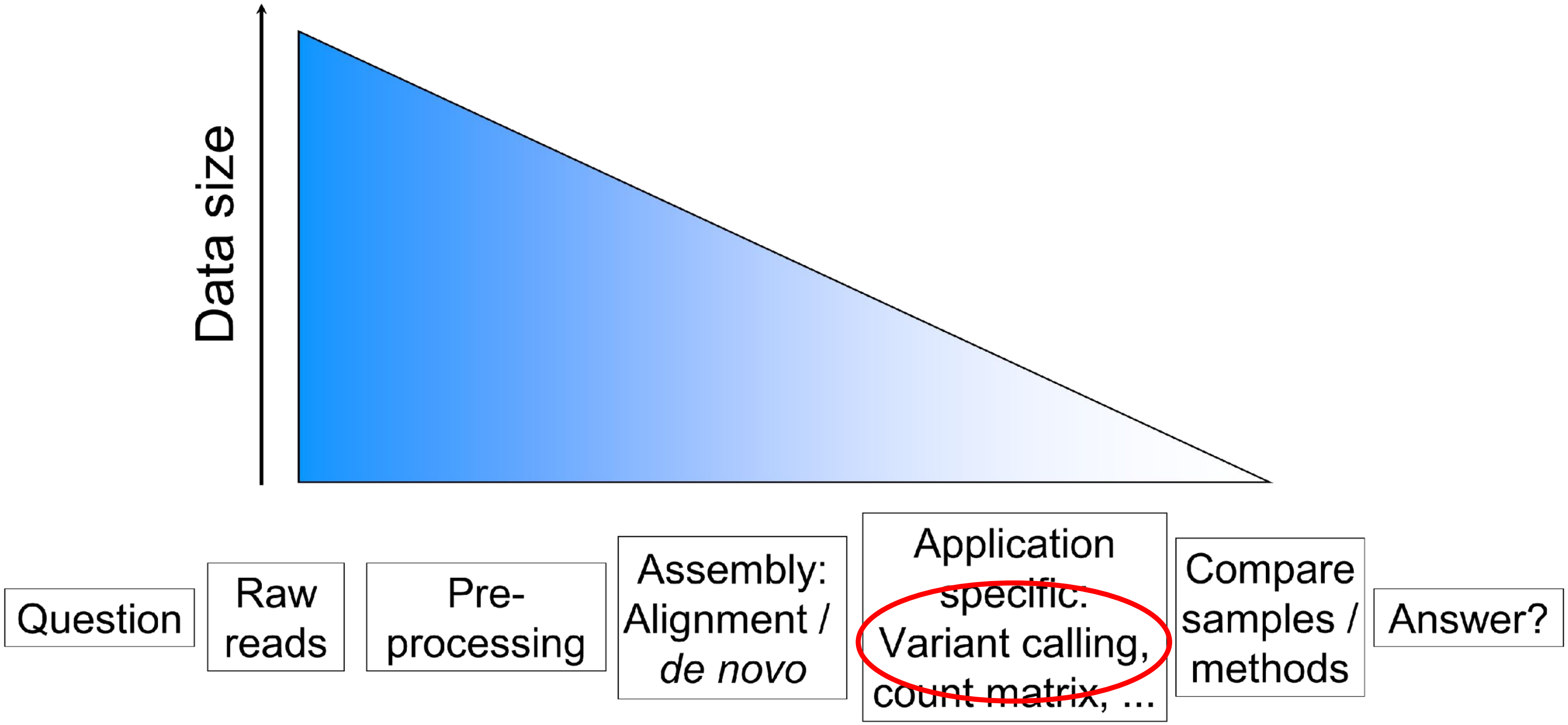
Joint Genotyping



Menu

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

Generalized NGS analysis



Question

Raw reads

Pre-processing

Assembly: Alignment / *de novo*

Application specific: Variant calling, count matrix, ...

Compare samples / methods

Answer?

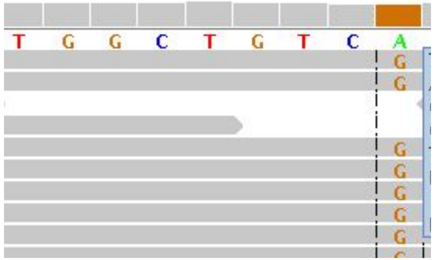
What is genotyping?

Genotyping is determining which genotype maximizes:

$$P(\underset{\substack{\nearrow \\ \text{genotype}}}{G} \mid \underset{\substack{\nwarrow \\ \text{data}}}{D})$$

What is genotyping?

Genotyping is determining which genotype maximizes:

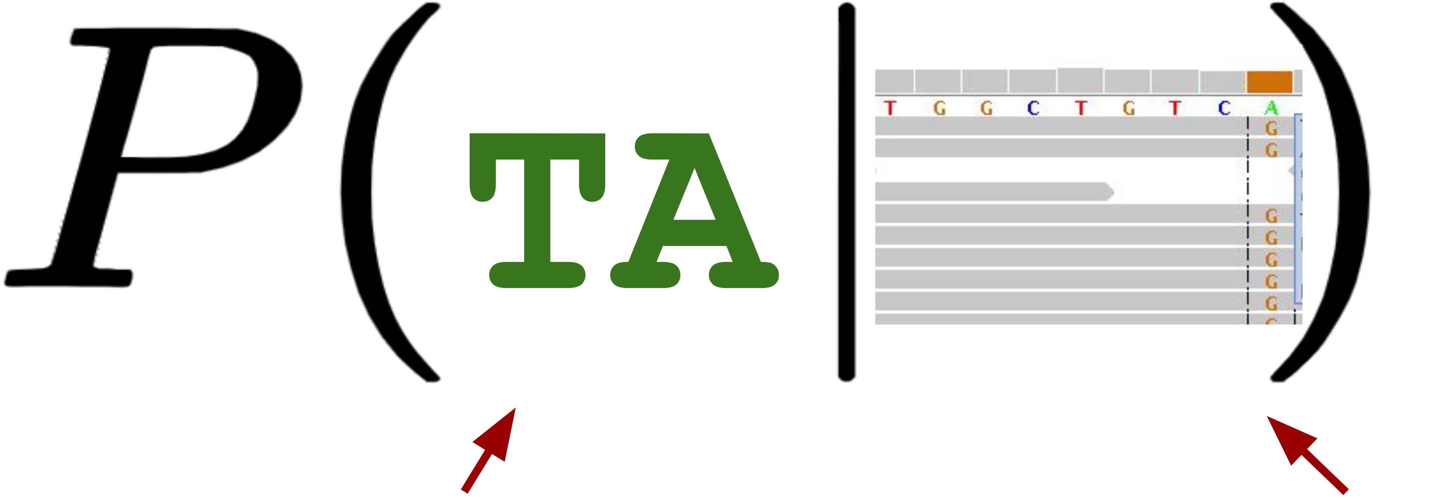
$P(G \mid$  $)$

genotype

data

What is genotyping?

Genotyping is determining which genotype maximizes:



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?

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

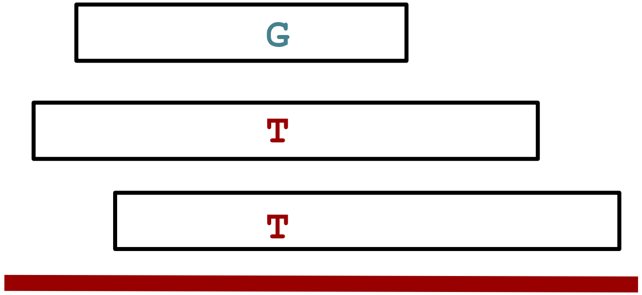
The likelihood

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

i.e. each reads is an independent observation

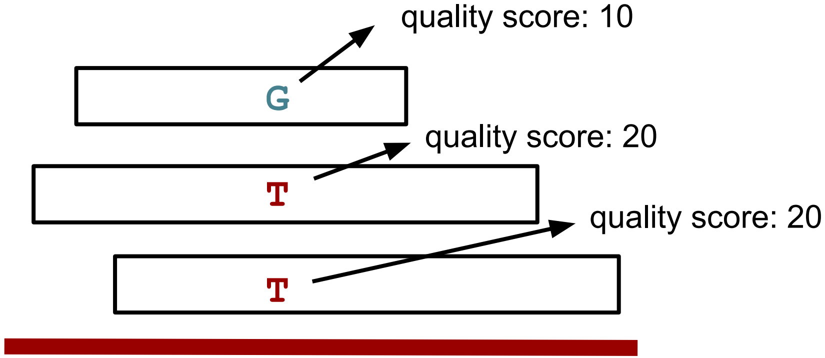
The likelihood $P(D|G)$

Toy example:



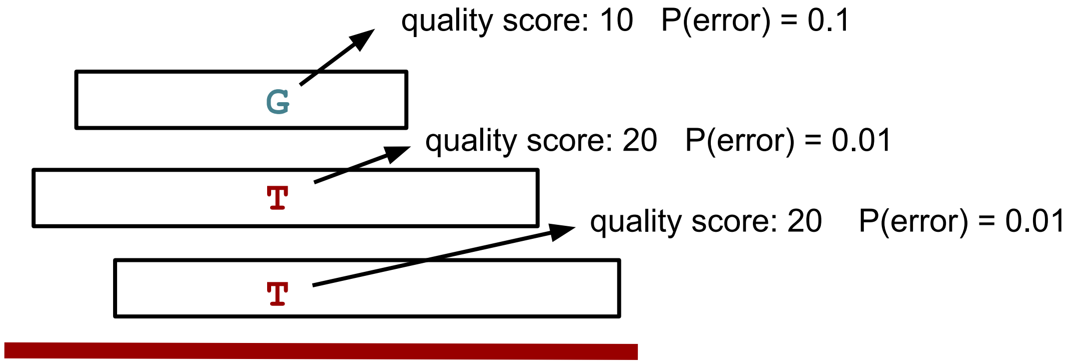
The likelihood $P(D|G)$

Toy example:



The likelihood $P(D|G)$

Toy example:



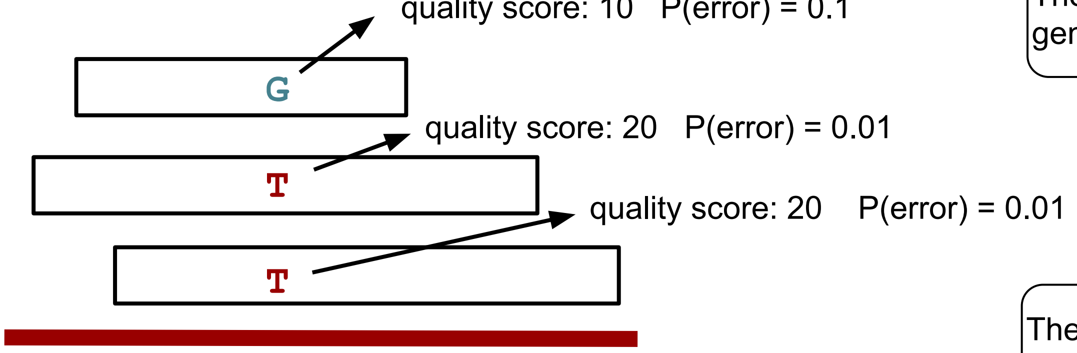
quality score: 10   P(error) = 0.1

quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01

The likelihood $P(D|G)$

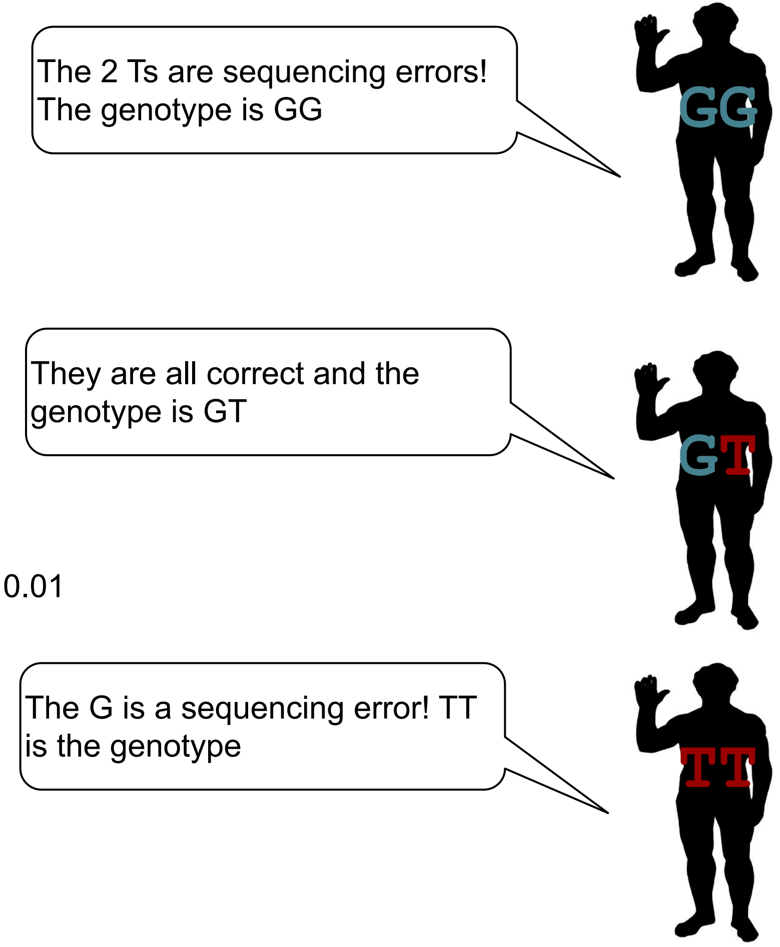
Toy example:



quality score: 10   P(error) = 0.1

quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01



The 2 Ts are sequencing errors! The genotype is GG

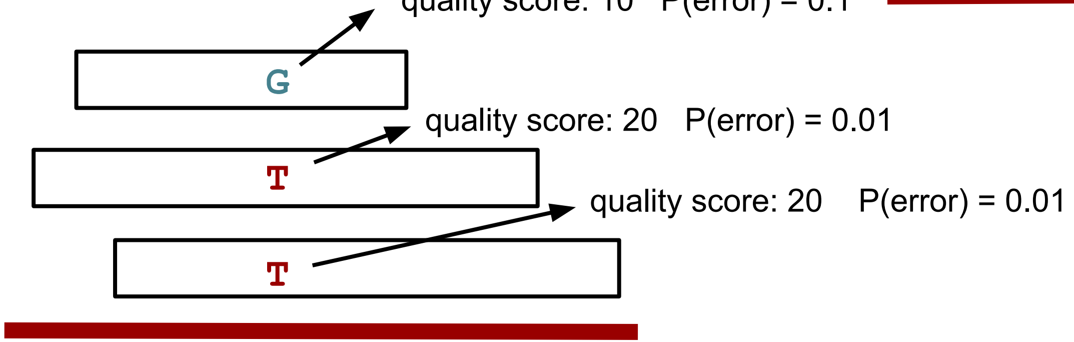
They are all correct and the genotype is GT

The G is a sequencing error! TT is the genotype

The likelihood $P(D|G)$

Error model

Toy example:



quality score: 10 P(error) = 0.1

quality score: 20 P(error) = 0.01

quality score: 20 P(error) = 0.01

What I think is the base

probability of the data given the base

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

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

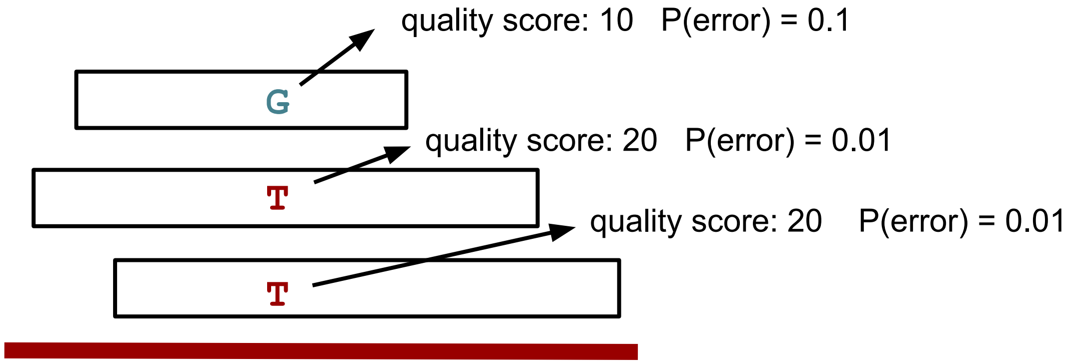
$$\mathbf{P}(G|G) = 0.9$$

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

Let’s evaluate 3 possible genotypes:

The likelihood $P(D|G)$

Toy example:



quality score: 10 P(error) = 0.1

quality score: 20 P(error) = 0.01

quality score: 20 P(error) = 0.01

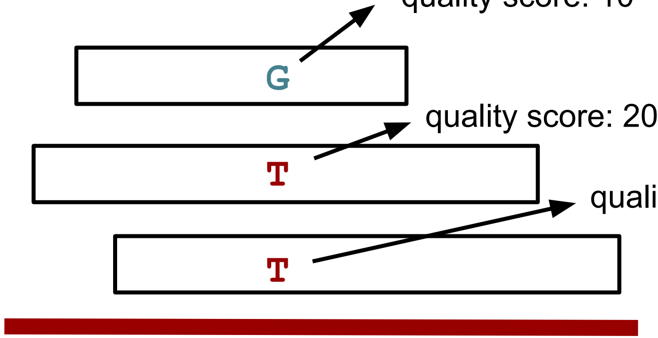
GG

GT

TT

$$P(D|GG)$$

The likelihood $P(D|G)$



quality score: 10   P(error) = 0.1

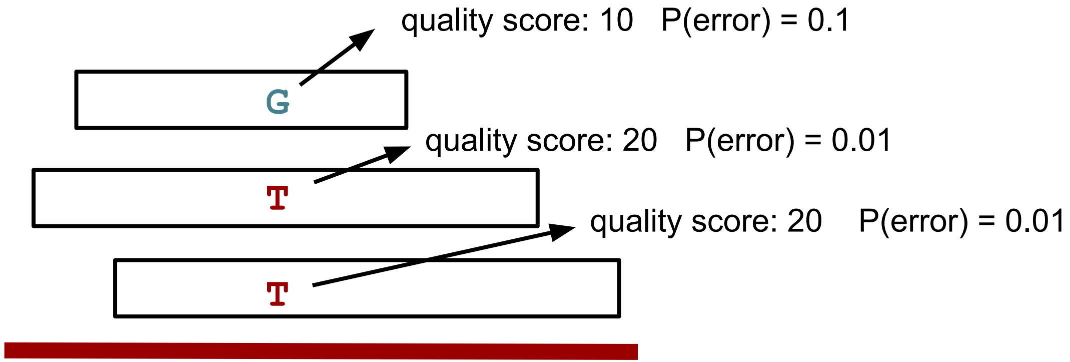
quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01

The likelihood $P(D|G)$

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

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



quality score: 10 P(error) = 0.1

quality score: 20 P(error) = 0.01

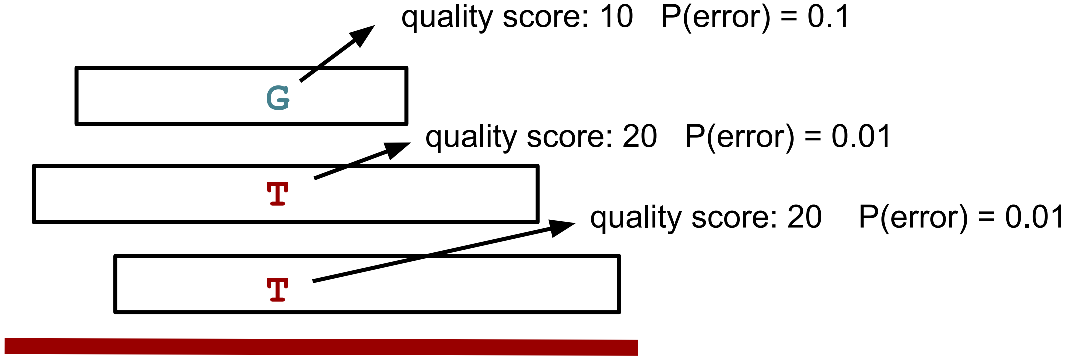
quality score: 20 P(error) = 0.01

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

The likelihood $P(D|G)$

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

✓ 0.9 ✓ 0.9



quality score: 10 P(error) = 0.1

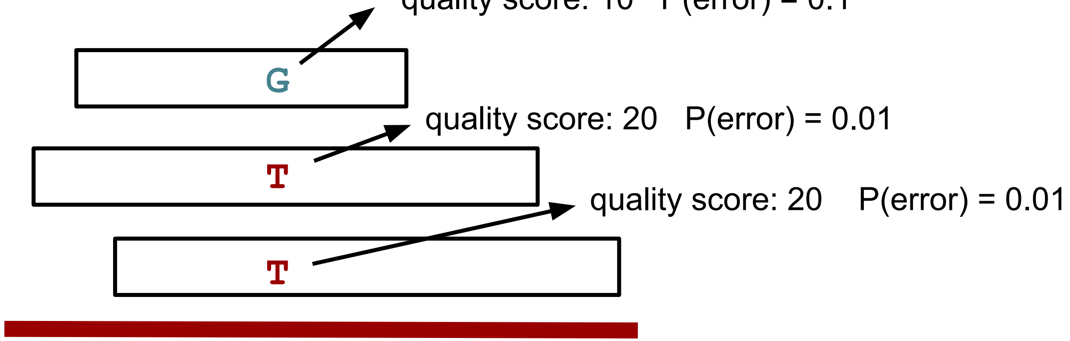
quality score: 20 P(error) = 0.01

quality score: 20 P(error) = 0.01

The likelihood $P(D|G)$

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

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

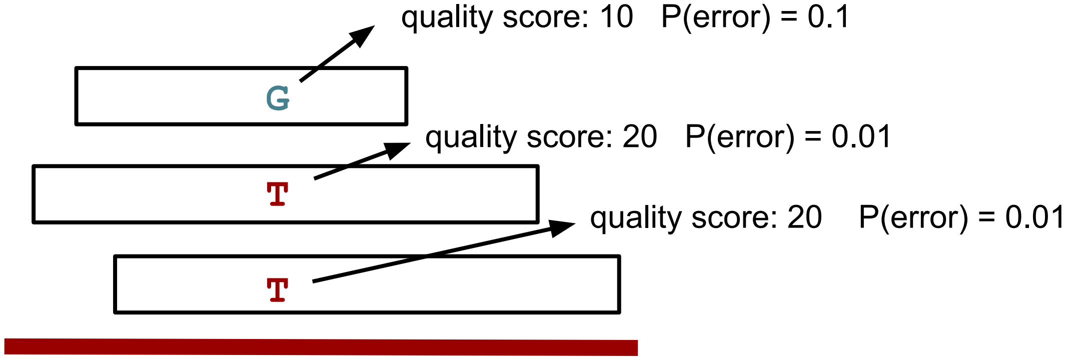


quality score: 10 P(error) = 0.1

quality score: 20 P(error) = 0.01

quality score: 20 P(error) = 0.01

The likelihood $P(D|G)$



quality score: 10   P(error) = 0.1

quality score: 20   P(error) = 0.01

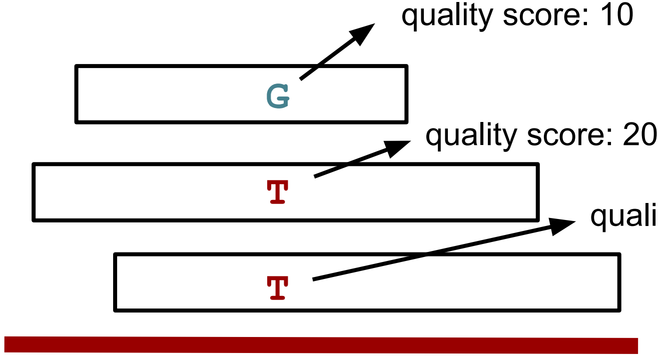
quality score: 20   P(error) = 0.01

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

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

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

The likelihood $P(D|G)$



quality score: 10   P(error) = 0.1

quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01

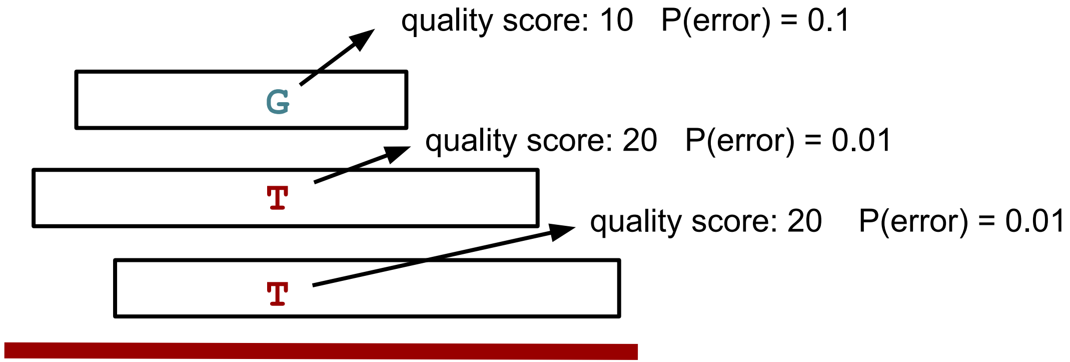
$\frac{1}{2}$	G	$\frac{1}{2}$	G
✓	0.9	✓	0.9
✗	$\frac{0.01}{3}$	✗	$\frac{0.01}{3}$
✗	$\frac{0.01}{3}$	✗	$\frac{0.01}{3}$

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

The likelihood $P(D|G)$

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

$\frac{1}{2}$ G		$\frac{1}{2}$ T	
✓	0.9	✗	$\frac{0.1}{3}$
✗	$\frac{0.01}{3}$	✓	0.99
✗	$\frac{0.01}{3}$	✓	0.99



quality score: 10 P(error) = 0.1

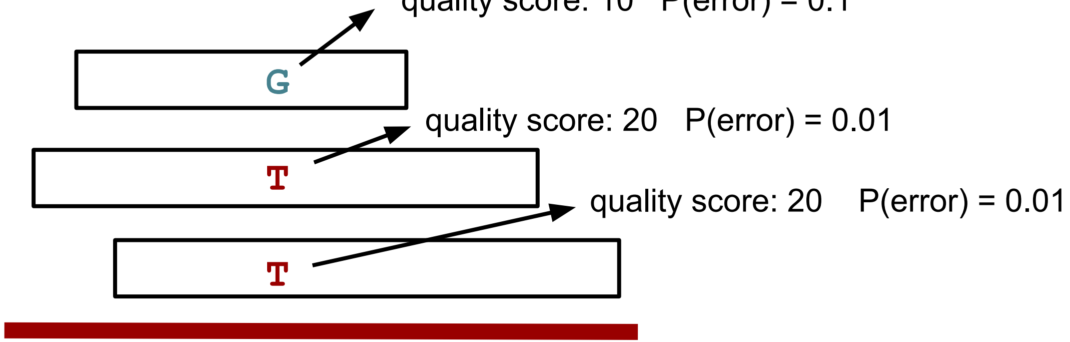
quality score: 20 P(error) = 0.01

quality score: 20 P(error) = 0.01

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

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

The likelihood $P(D|G)$



quality score: 10   P(error) = 0.1

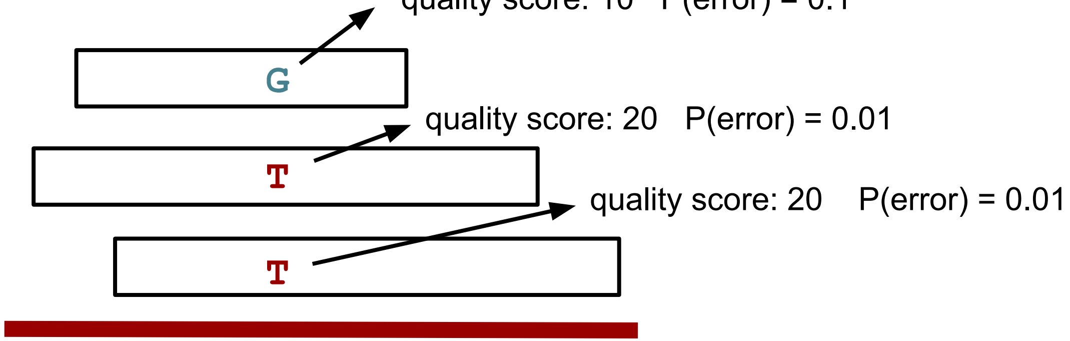
quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01

$\frac{1}{2}$	T	$\frac{1}{2}$	T
✗	$\frac{0.1}{3}$	✗	$\frac{0.1}{3}$
✓	0.99	✓	0.99
✓	0.99	✓	0.99

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

The likelihood $P(D|G)$



quality score: 10   P(error) = 0.1

quality score: 20   P(error) = 0.01

quality score: 20   P(error) = 0.01

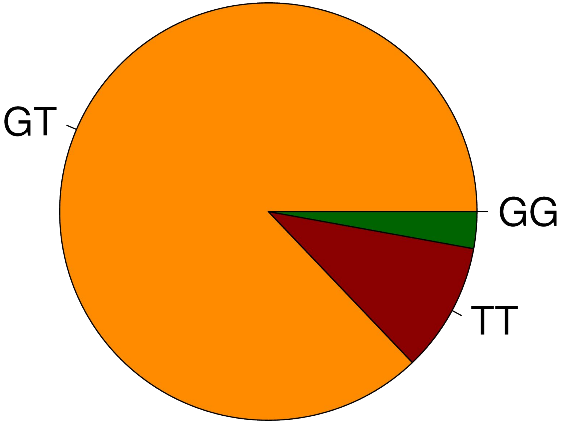
$$P(D|\mathrm{GG}) = 0.00001$$

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

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

The likelihood $P(D|G)$

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



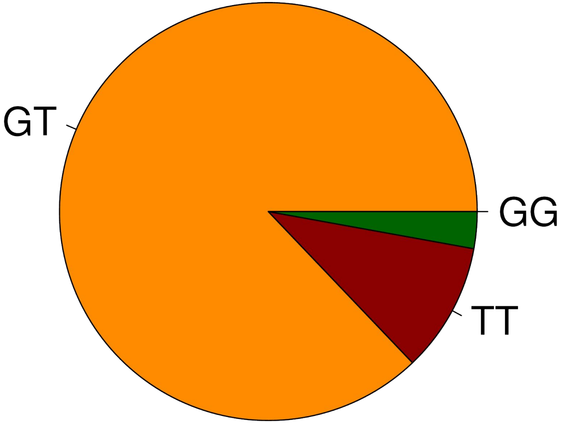
$$P(D|\text{GG}) = 0.00001$$

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

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

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

The likelihood $P(D|G)$



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

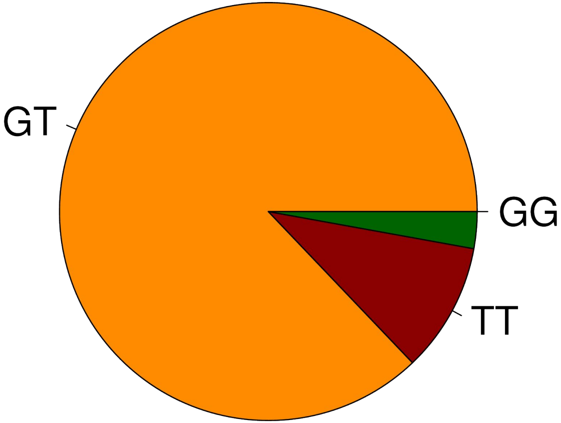
$$P(D|\text{GT}) = 0.11511$$

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

We will neglect the genotype prior this time

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

The likelihood $P(D|G)$

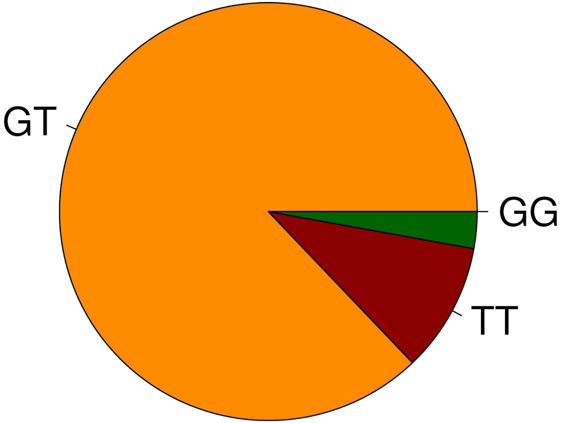


$P(\mathbf{GG}|D)$= 6.7e-05

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

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

The likelihood $P(D|G)$



$P(GG|D)$= 6.7e-05

$P(GT|D)$= 0.77888

$P(TT|D)$= 0.22104

Important point: More coverage ⟶More multiplications ⟶The relative difference between models become larger

The likelihood $P(D|G)$

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

What is the difference between the **most likely** and **second most likely**

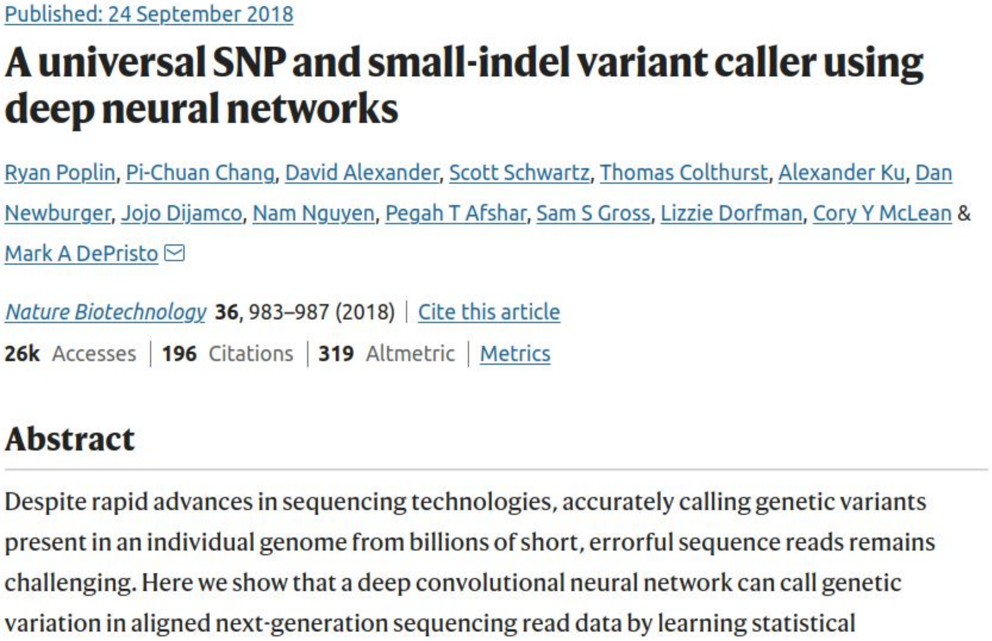
Details I did not cover

- 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

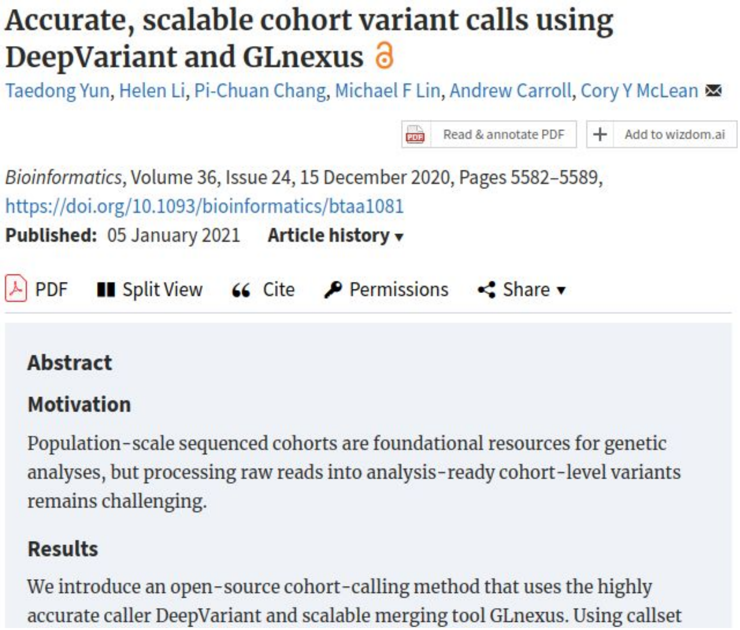
Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean & Mark A DePristo

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

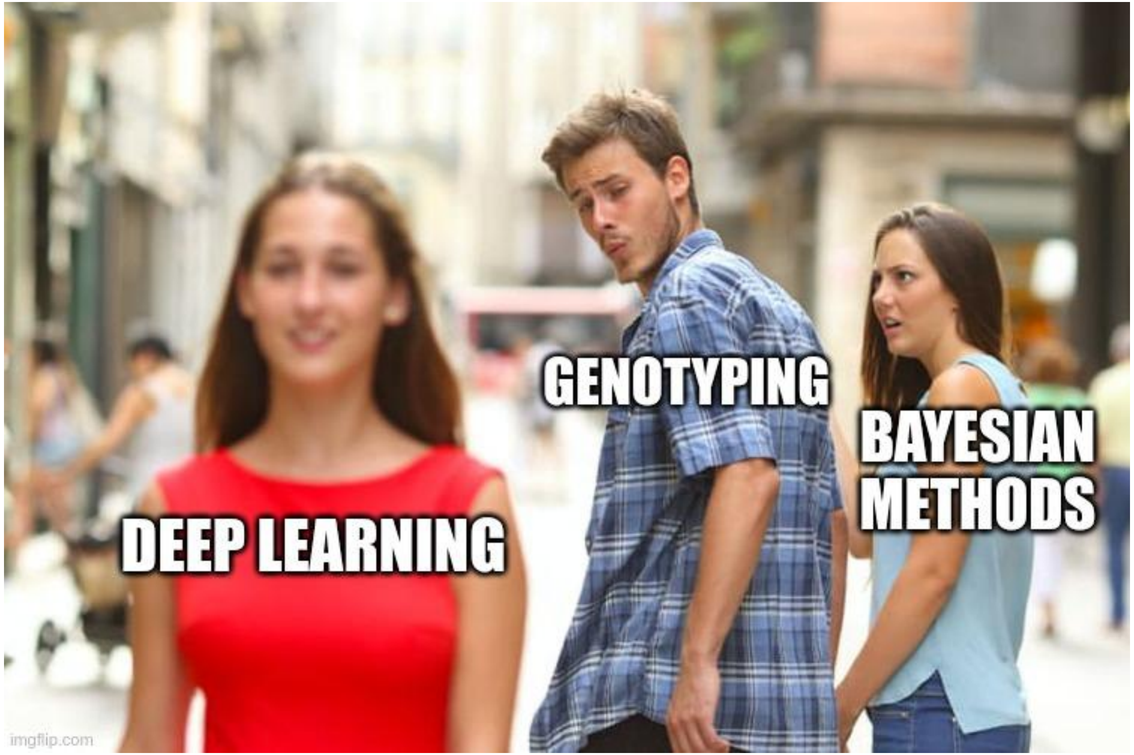
Motivation

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

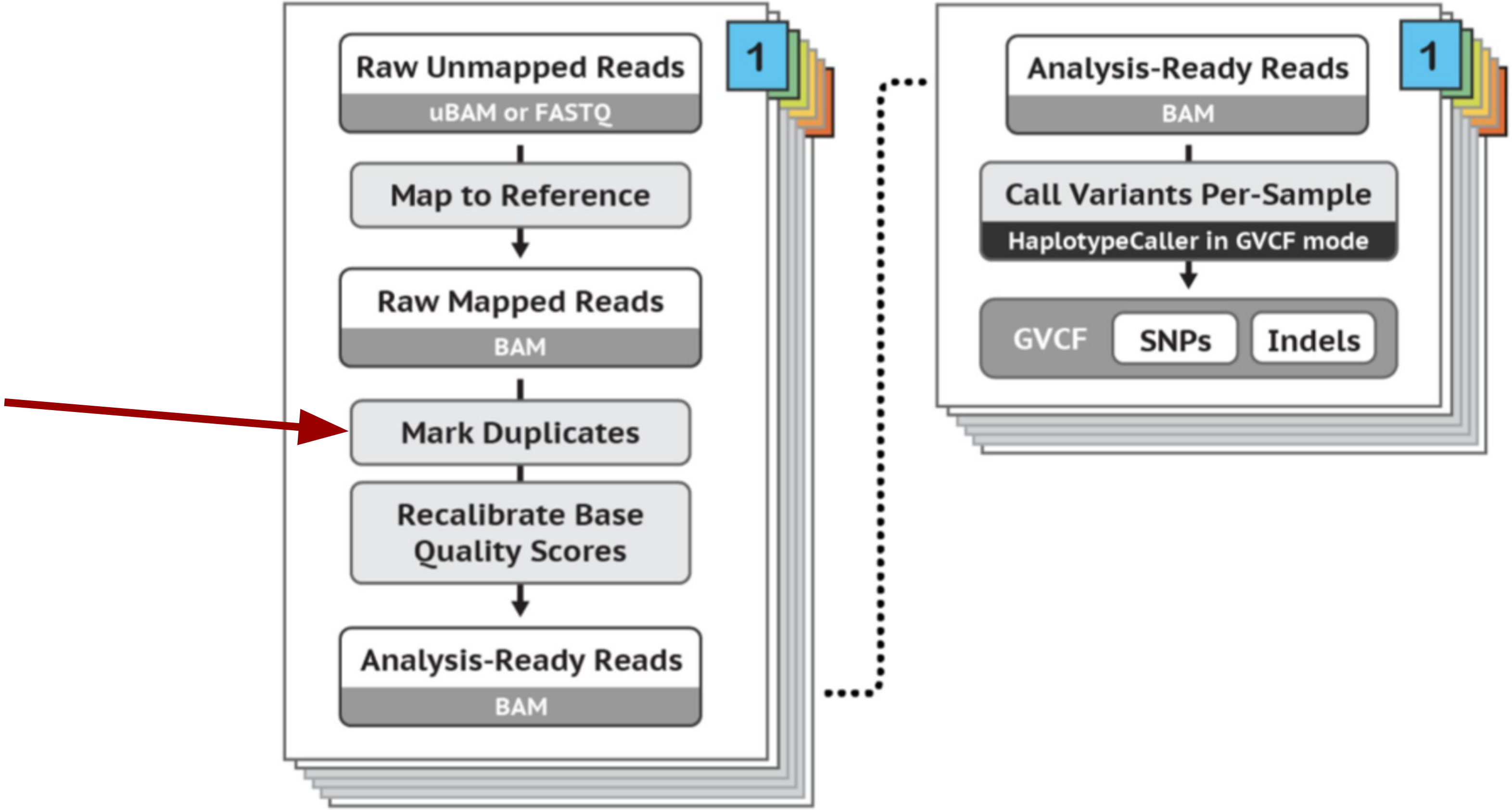
Results

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?

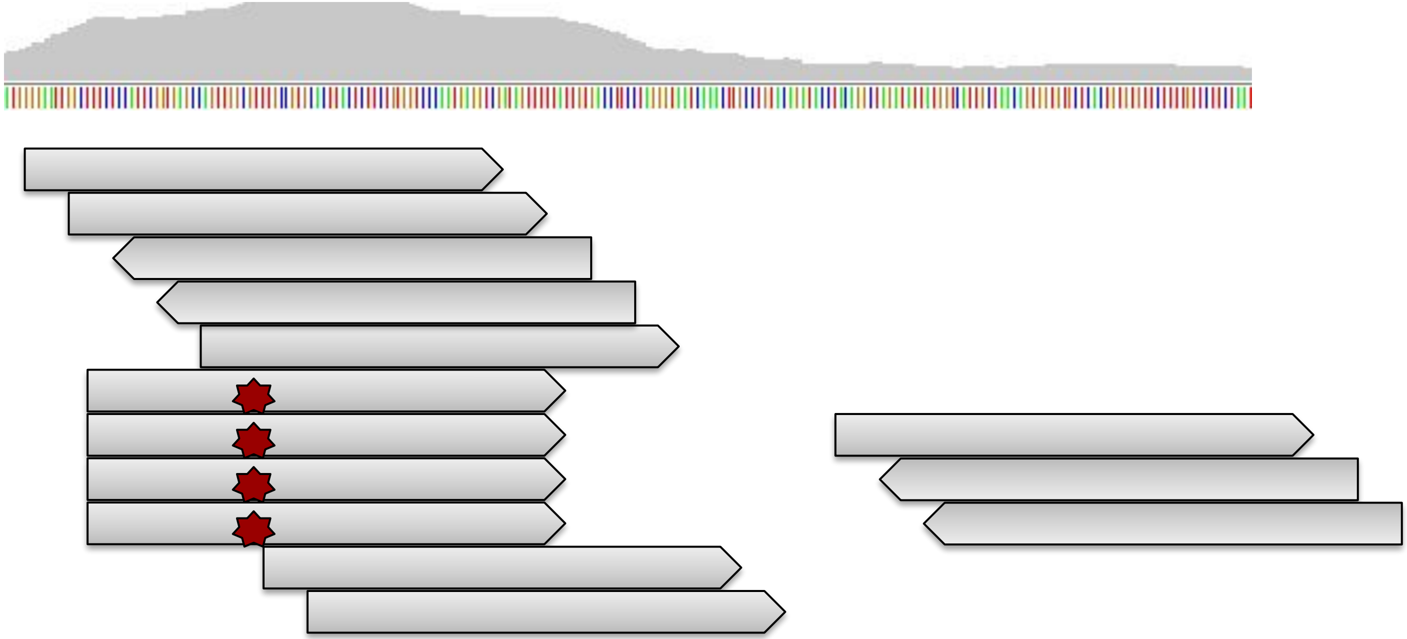


GATK’s recommended workflow

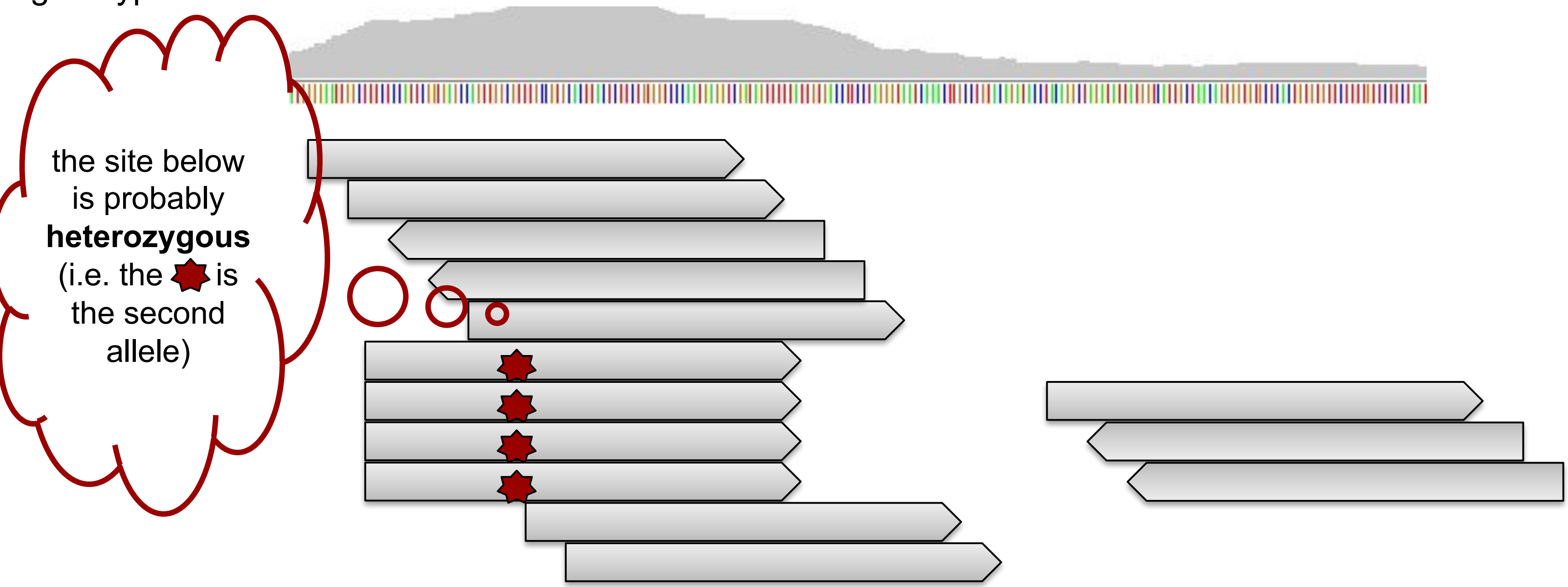


https://gatk.broadinstitute.org/hc/en-us/articles/360035535932-Germline-short-variant-discovery-SNPs-Indels-

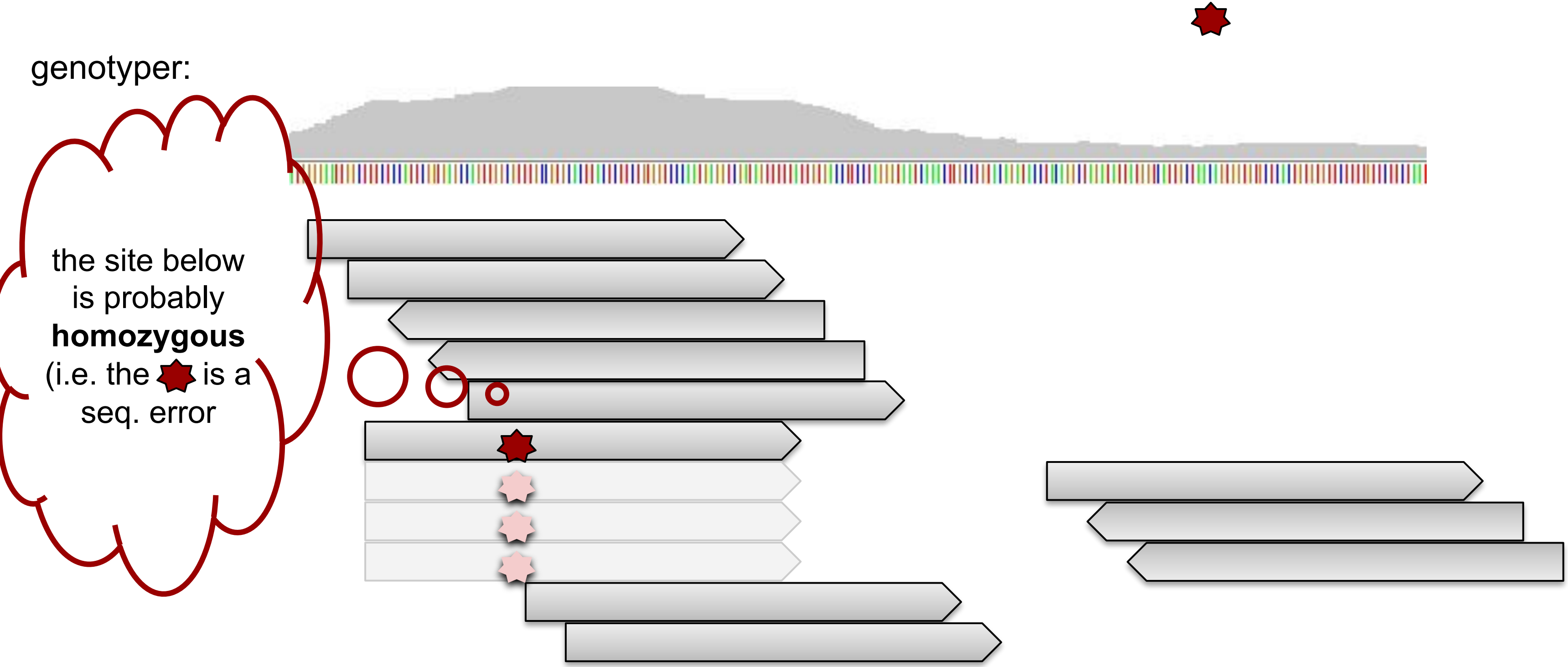
The PCR amplification step included in the majority of NGS library construction techniques can introduce duplicates in the data.



We want: remove or mark them to avoid false calls



Genotypers will ignore reads marked as duplicates



genotyper:

the site below is probably **homozygous** (i.e. the ✸ is a seq. error

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.

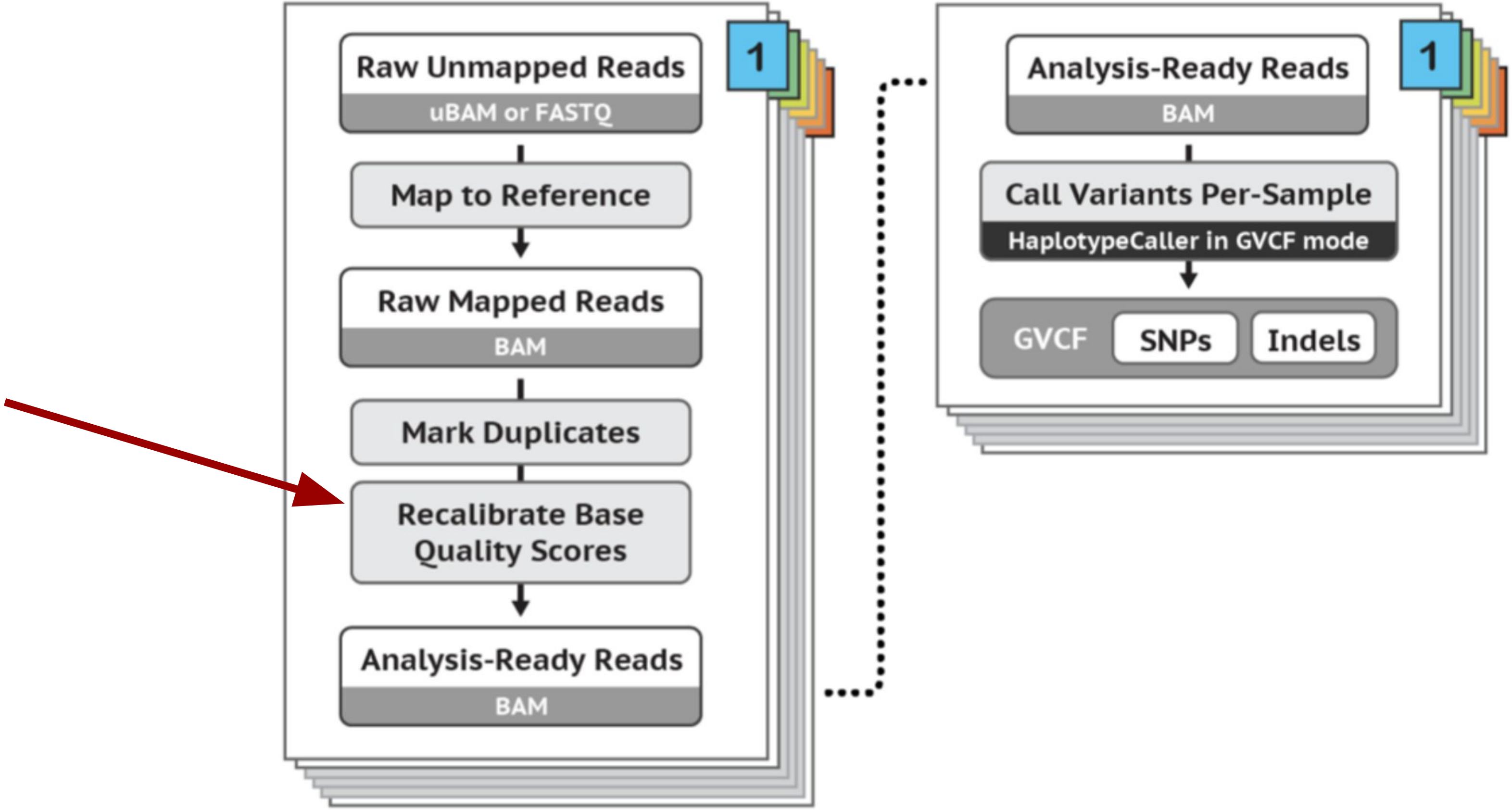
http://picard.sourceforge.net/

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?

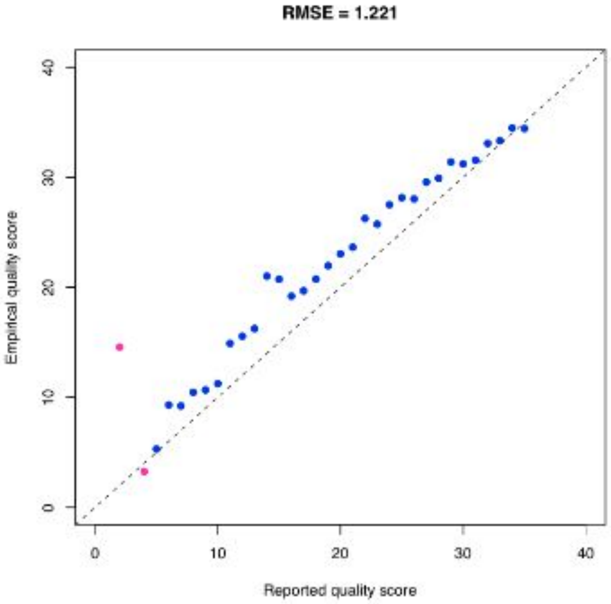
```
@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNNNN/1
ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC
+
BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?
```

- There are supposed to reflect P(error)

- They are not always accurate: problem for genotyping

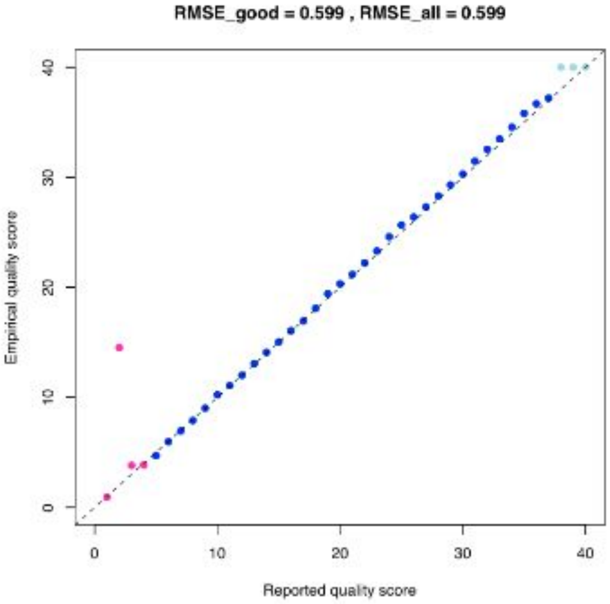
Reported Quality vs. Empirical Quality

Idea: use documented variants in the genome



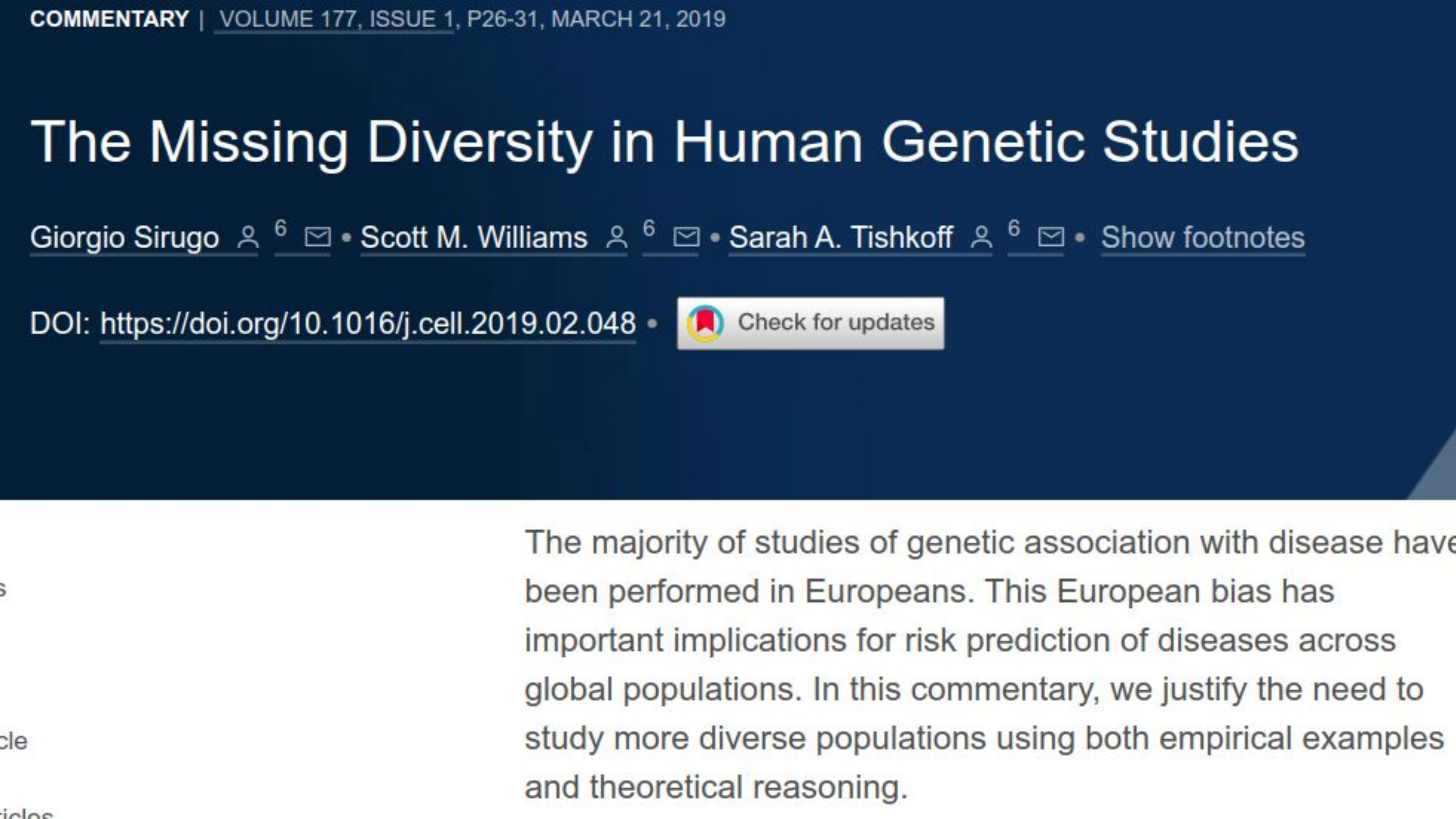
RMSE = 1.221

Original Data



RMSE_good = 0.599 , RMSE_all = 0.599

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

s

cle

icles

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

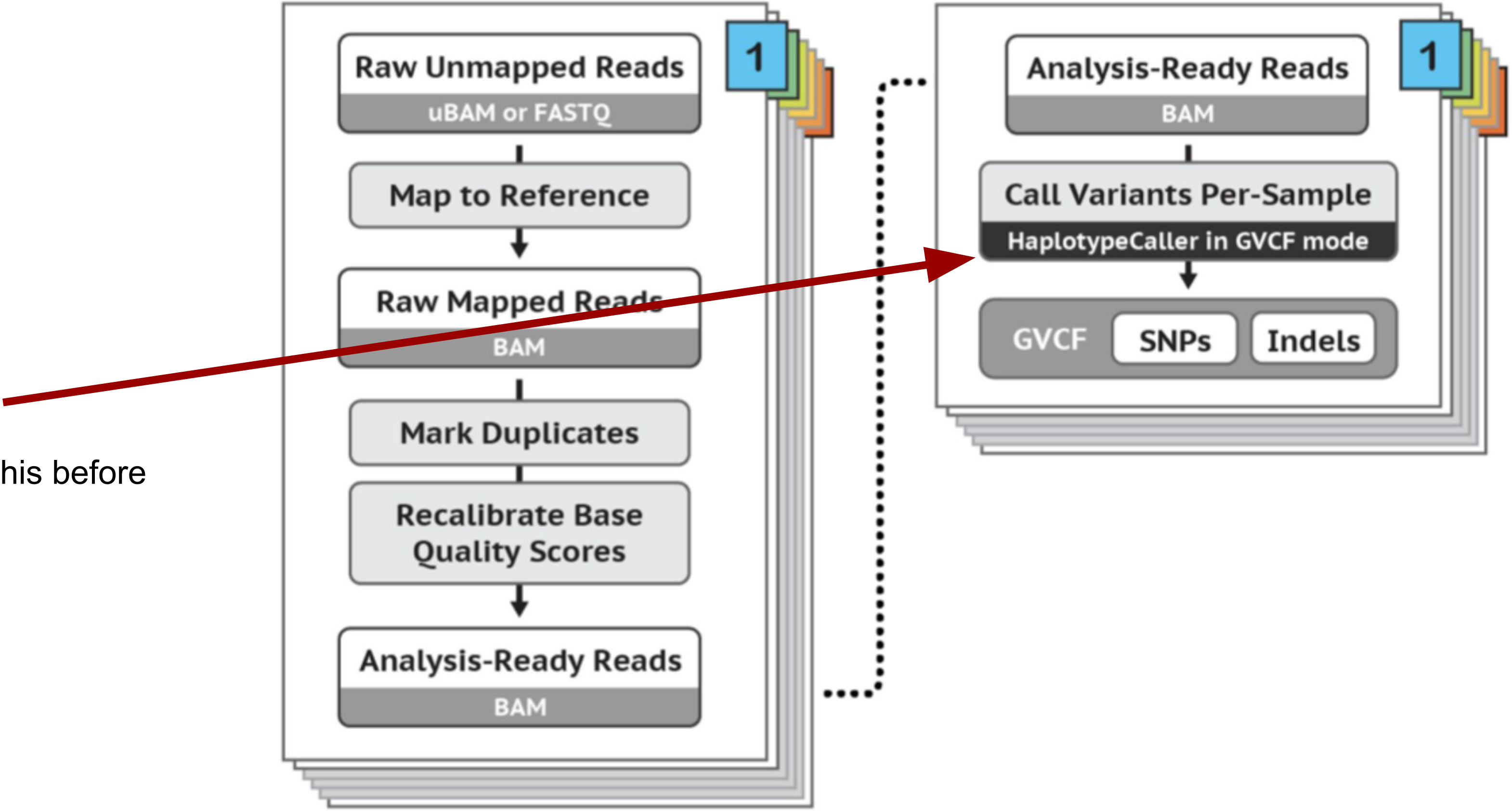
To work we need:

- 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    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
```

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

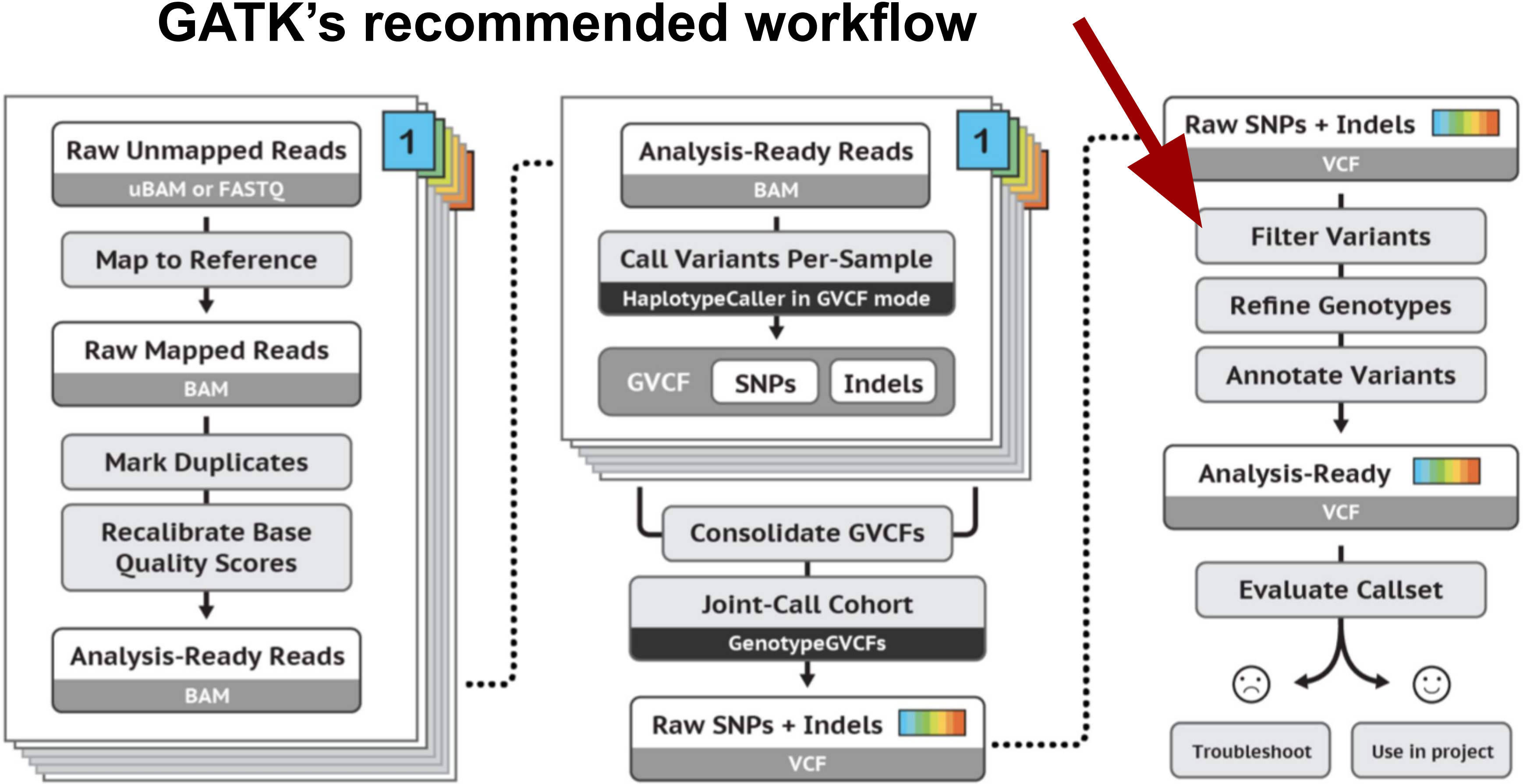
The likelihood $P(D|G)$

PL field

	PHRED	PHRED-scaled
$P(\mathrm{GG} \mid D)$= 6.7e-05	41.70	40.60
$P(\mathrm{GT} \mid D)$= 0.77888	1.09	0.00
$P(\mathrm{TT} \mid D)$= 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-

Later, we will...

- Filter variants
- Annotate the variants
- Other types of variants
- Final considerations about genomic variants

… but for now we have:

- removed duplicates to get independent observations
- Used them to call the most likely genotype
- Saw the VCF format

Exercise time!

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