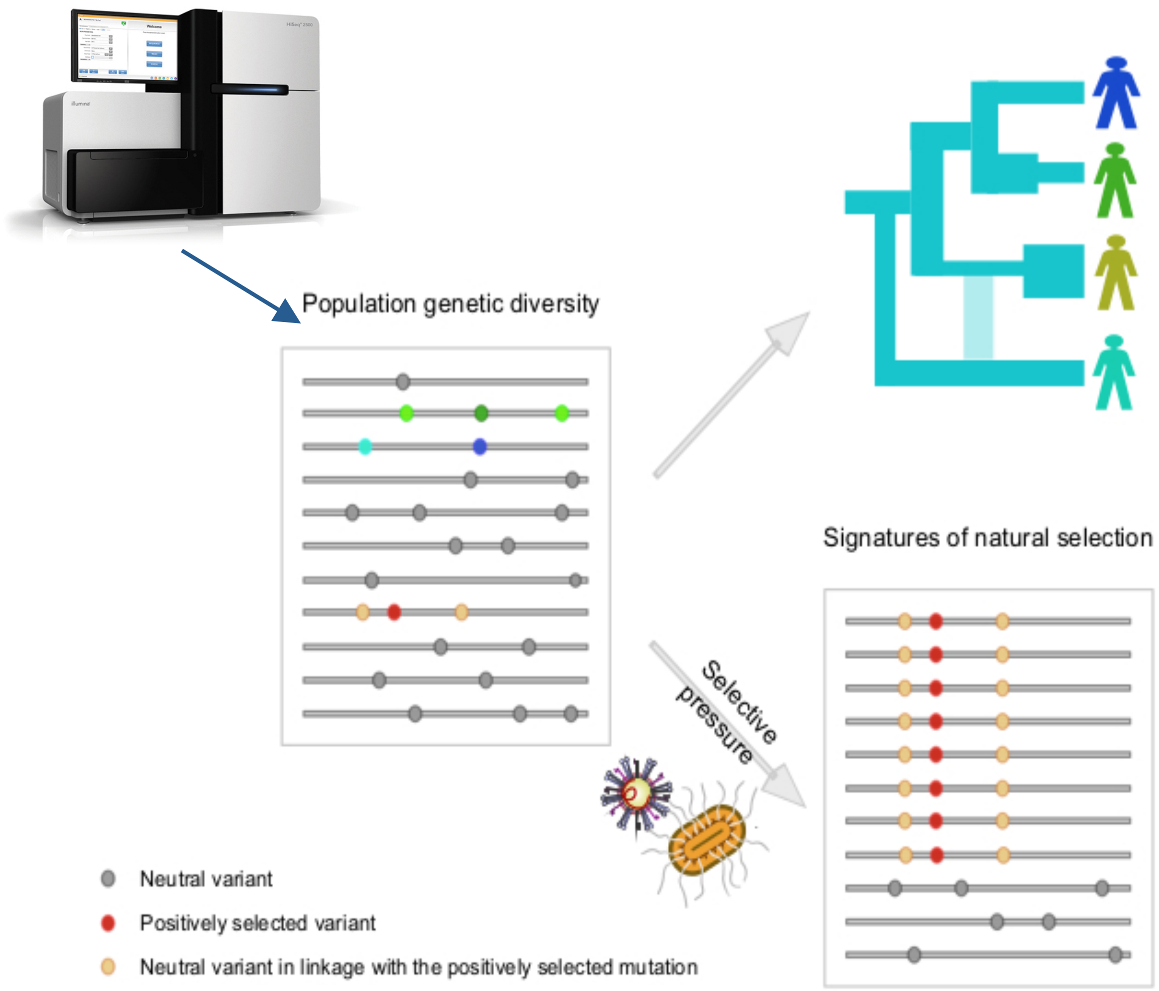
Introduction to NGS data: Genotype and SNP calling

Matteo Fumagalli

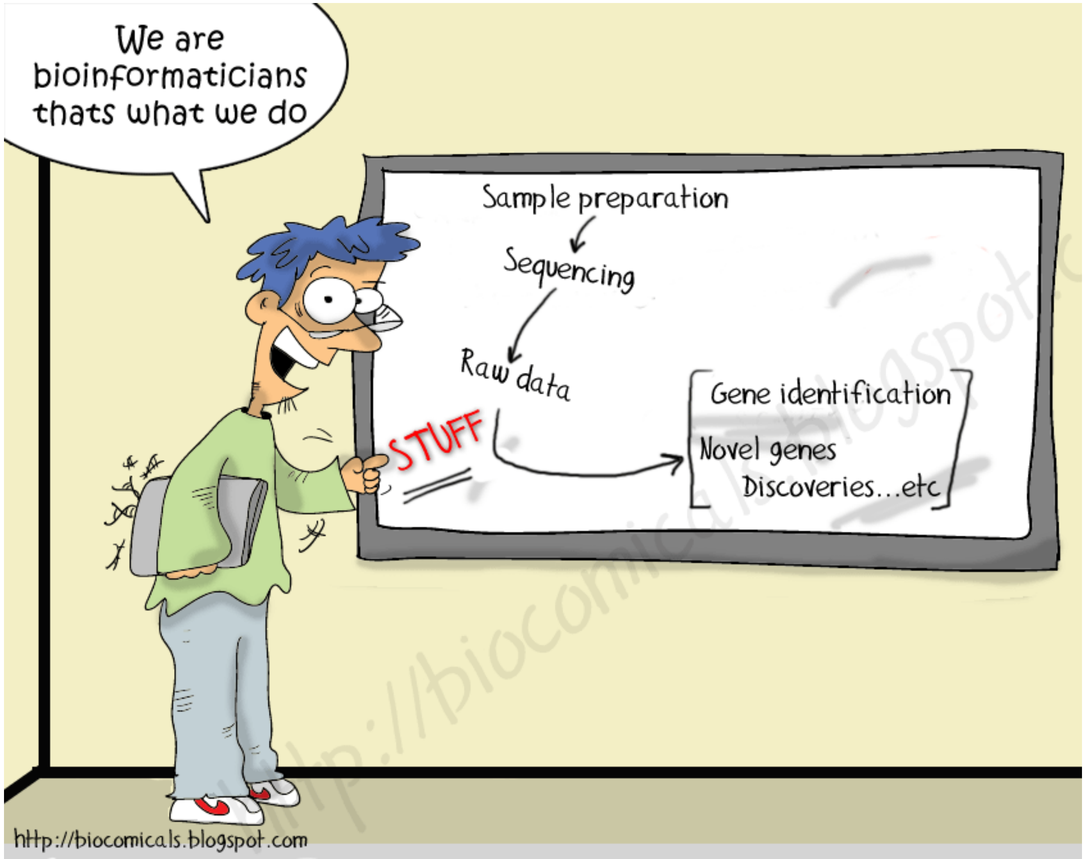
Population genetics


Population genetic diversity
Signatures of natural selection
Selective pressure
● Neutral variant
● Positively selected variant
● Neutral variant in linkage with the positively selected mutation

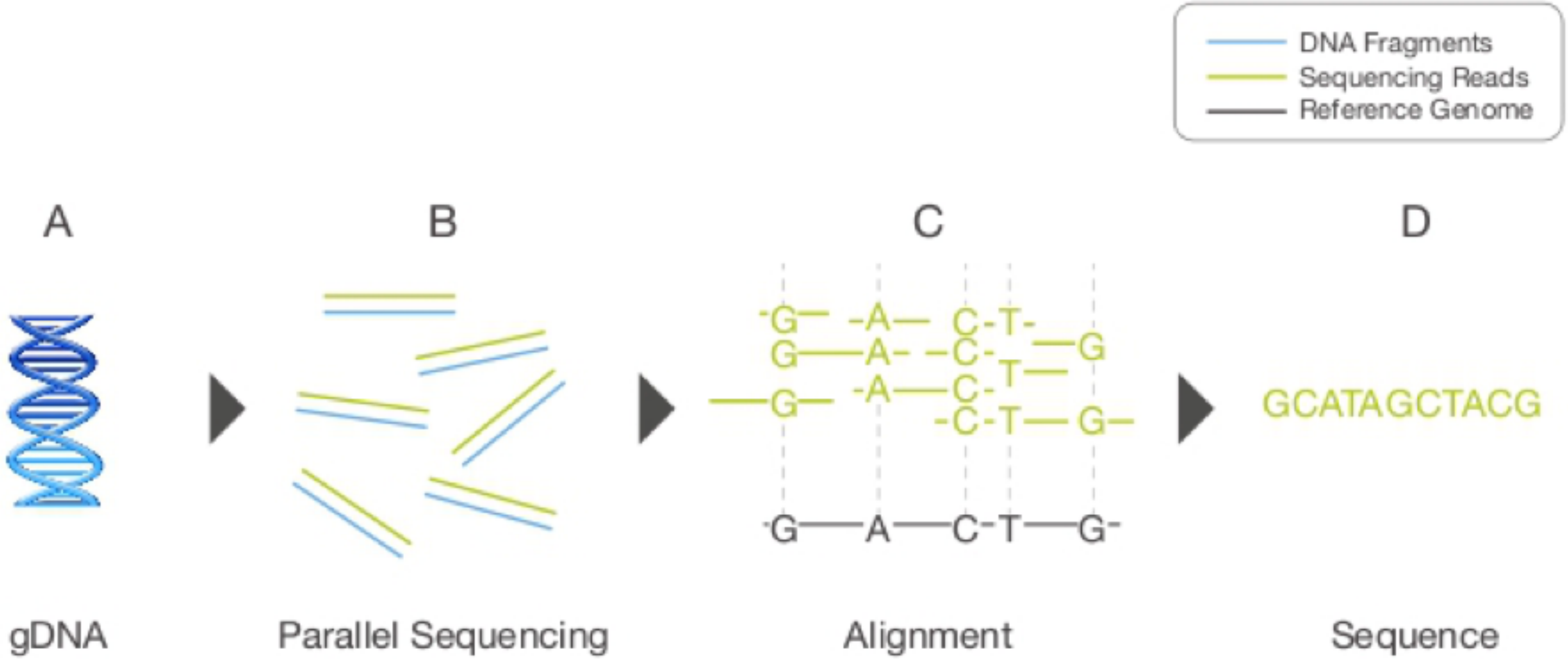
Demographic inference

Adaptive evolution

Find the bioinformatician inside yourself



Next-Generation Sequencing



A. Extracted gDNA

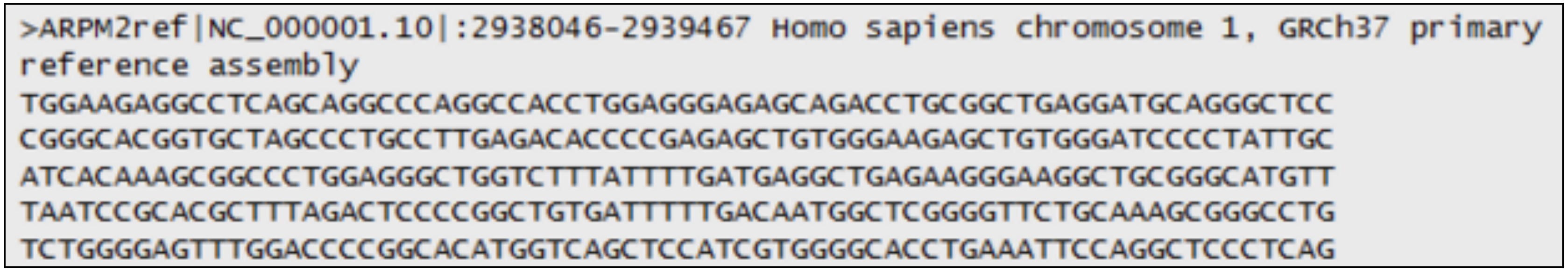
B. gDNA is fragmented into a library of small segments that are each sequenced in parallel.

C. Individual sequence reads are reassembled by aligning to a reference genome

D. The whole-genome sequence is derived from the consensus of aligned reads.

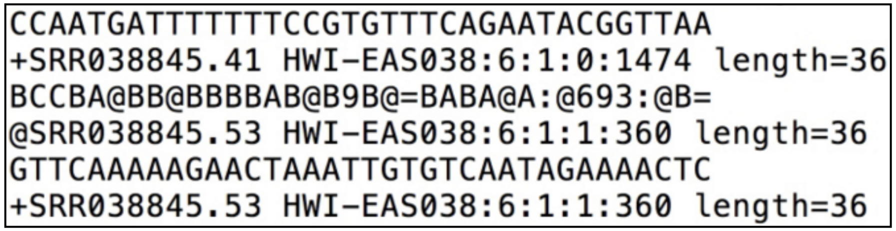
From genome to variants

Genome (FASTA)



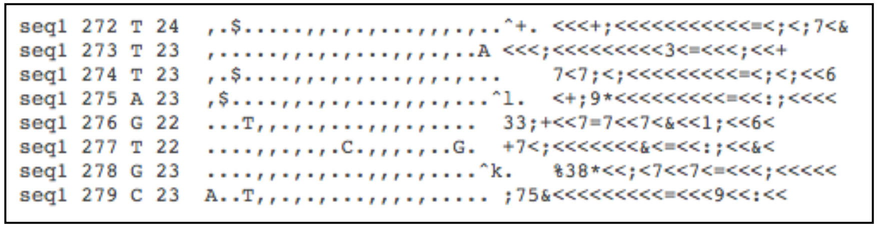
```
>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary
reference assembly
TGGAAGAGGCCTCAGCAGGCCCAGGCCACCTGGAGGGAGAGCAGACCTGCGGCTGAGGATGCAGGGCTCC
CGGGCACGGTGCTAGCCCTGCCTTGAGACACCCCGAGAGCTGTGGGAAGAGCTGTGGGATCCCCTATTGC
ATCACAAAGCGGCCCTGGAGGGCTGGTCTTTATTTTGATGAGGCTGAGAAGGGAAGGCTGCGGGCATGTT
TAATCCGCACGCTTTAGACTCCCCGGCTGTGATTTTTGACAATGGCTCGGGGTTCTGCAAAGCGGGCCTG
TCTGGGGAGTTTGGACCCCGGCACATGGTCAGCTCCATCGTGGGGCACCTGAAATTCCAGGCTCCCTCAG
```

Reads (FASTQ)



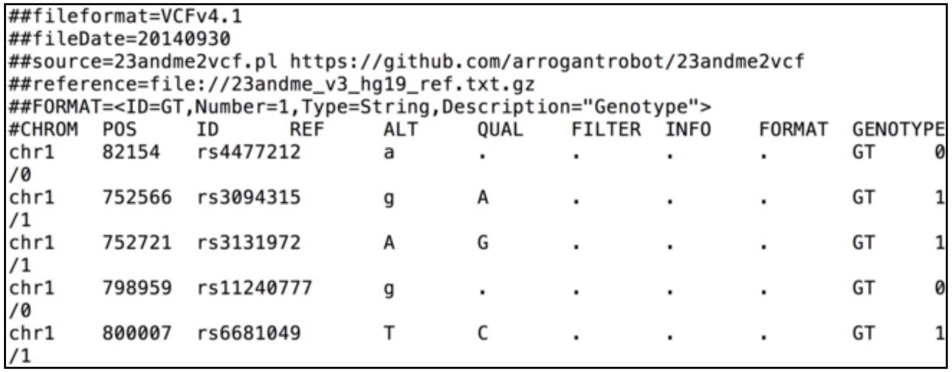
```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

Mapped Reads (mpileup, BAM)



```
seq1 272 T 24  ,.$.....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,....^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<=<<<9<<:<<
```

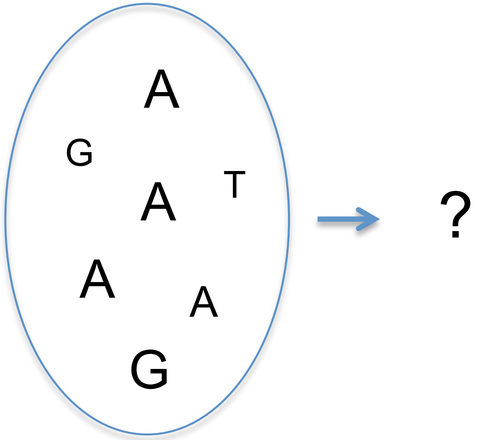
Variants (VCF)



```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23andme2vcf
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID          REF     ALT     QUAL    FILTER  INFO    FORMAT  GENOTYPE
chr1    82154   rs4477212   a       .       .       .       .       GT      0
/0
chr1    752566  rs3094315   g       A       .       .       .       GT      1
/1
chr1    752721  rs3131972   A       G       .       .       .       GT      1
/1
chr1    798959  rs11240777  g       .       .       .       .       GT      0
/0
chr1    800007  rs6681049   T       C       .       .       .       GT      1
/1
```

Statistical inference (1)

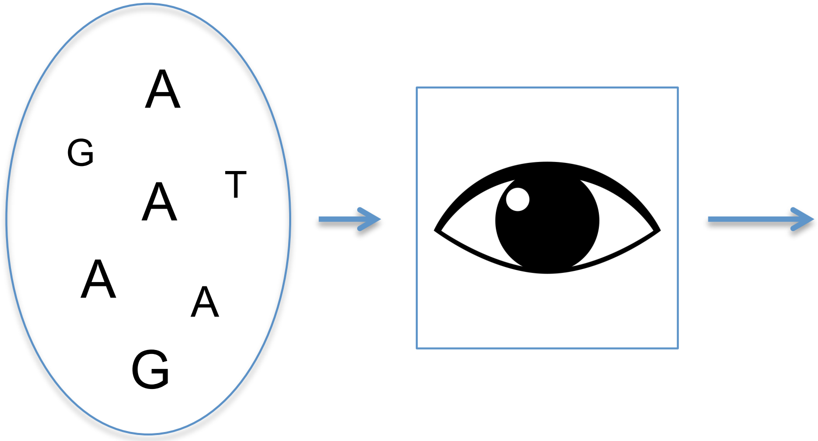
Data (D)



Parameter *f* is frequency of G

Statistical inference (1)

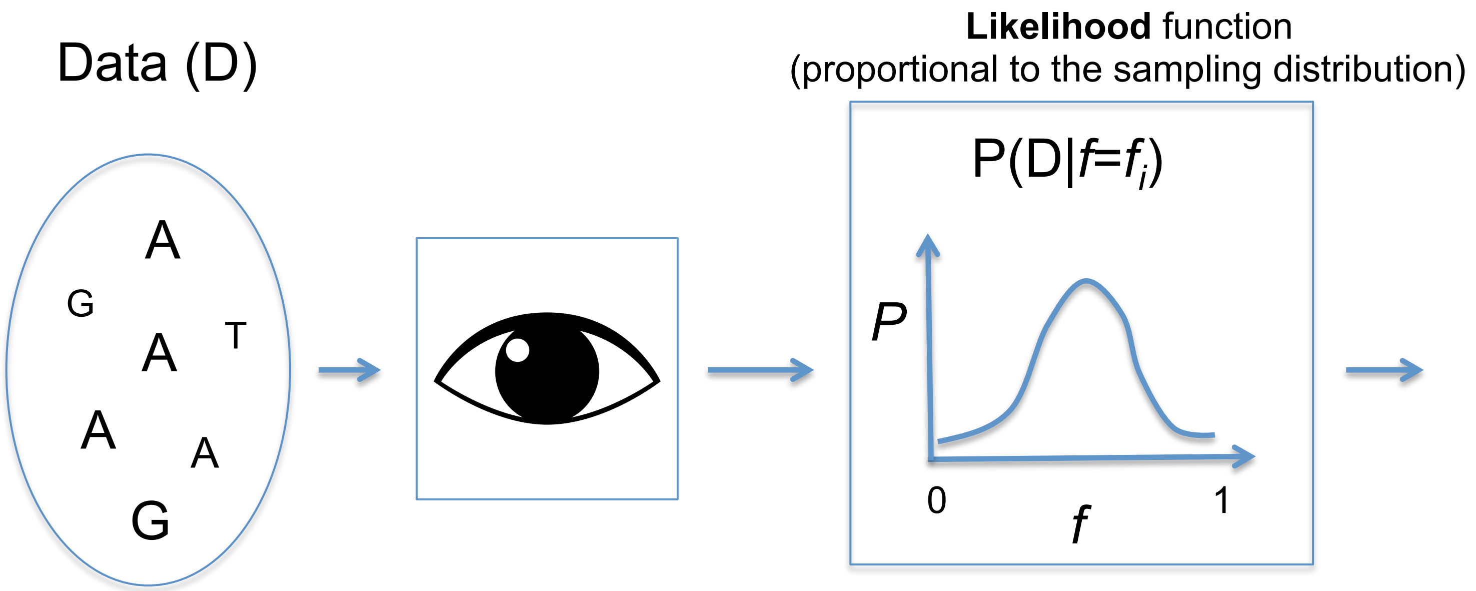
Data (D)



A
G
A
T
A
A
G

Parameter *f* is frequency of G

Statistical inference (1)



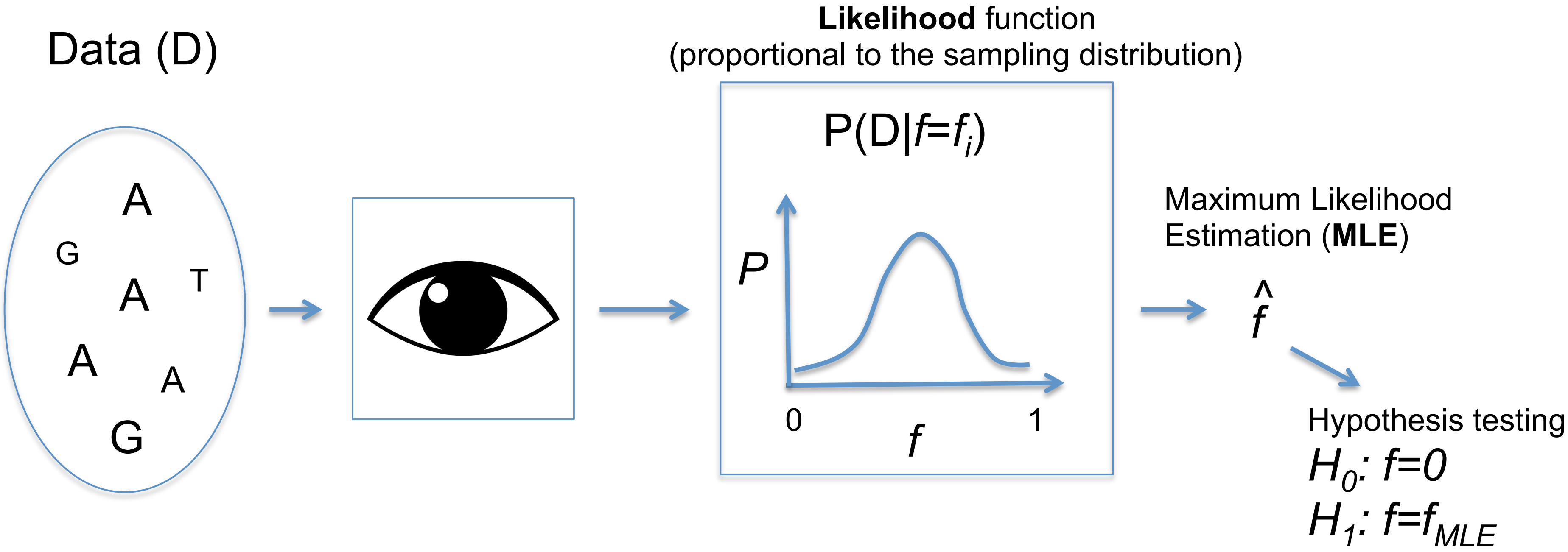
Data (D)

Likelihood function
(proportional to the sampling distribution)

$P(D|f=f_i)$

Parameter *f* is
frequency of G

Statistical inference (1)



Data (D)

Likelihood function
(proportional to the sampling distribution)

$P(D|f=f_i)$

Maximum Likelihood Estimation (**MLE**)

$\hat{f}$

Hypothesis testing

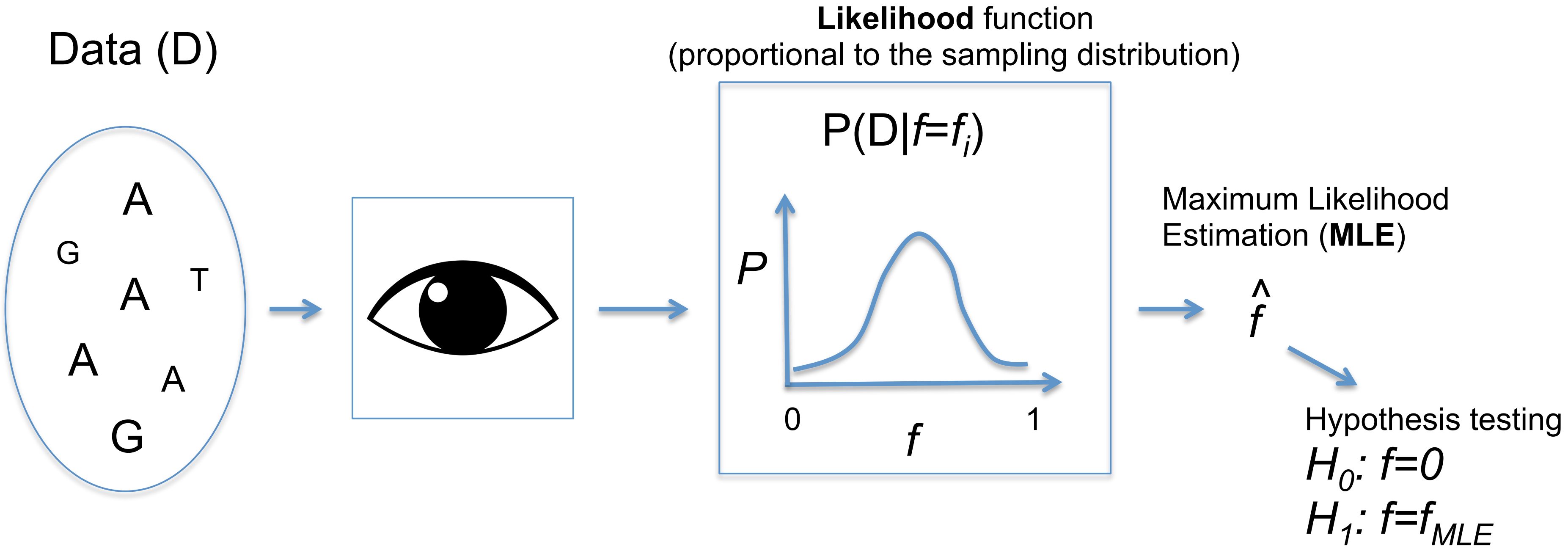
H_0: $f=0$

H_1: $f=f_{MLE}$

Likelihood approach:

- All the information on the parameter is in the likelihood function (we use all the data!).
- More data leads to less bias and less variance.
- Suitable for hypothesis testing.

Statistical inference (1)



Data (D)

Likelihood function
(proportional to the sampling distribution)

$P(D|f=f_i)$

Maximum Likelihood
Estimation (**MLE**)

$\hat{f}$

Hypothesis testing

H_0: $f=0$

H_1: $f=f_{MLE}$

Likelihood approach:

- All the information on the parameter is in the likelihood function (we use all the data!).
- More data leads to less bias and less variance.
- Suitable for hypothesis testing.

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

How many genotype likelihoods do we have for each individual at each site?

Genotype likelihoods

$$L(Data \mid G = \{A_1, A_2\})$$

$$A_i \in \{A, C, G, T\}$$

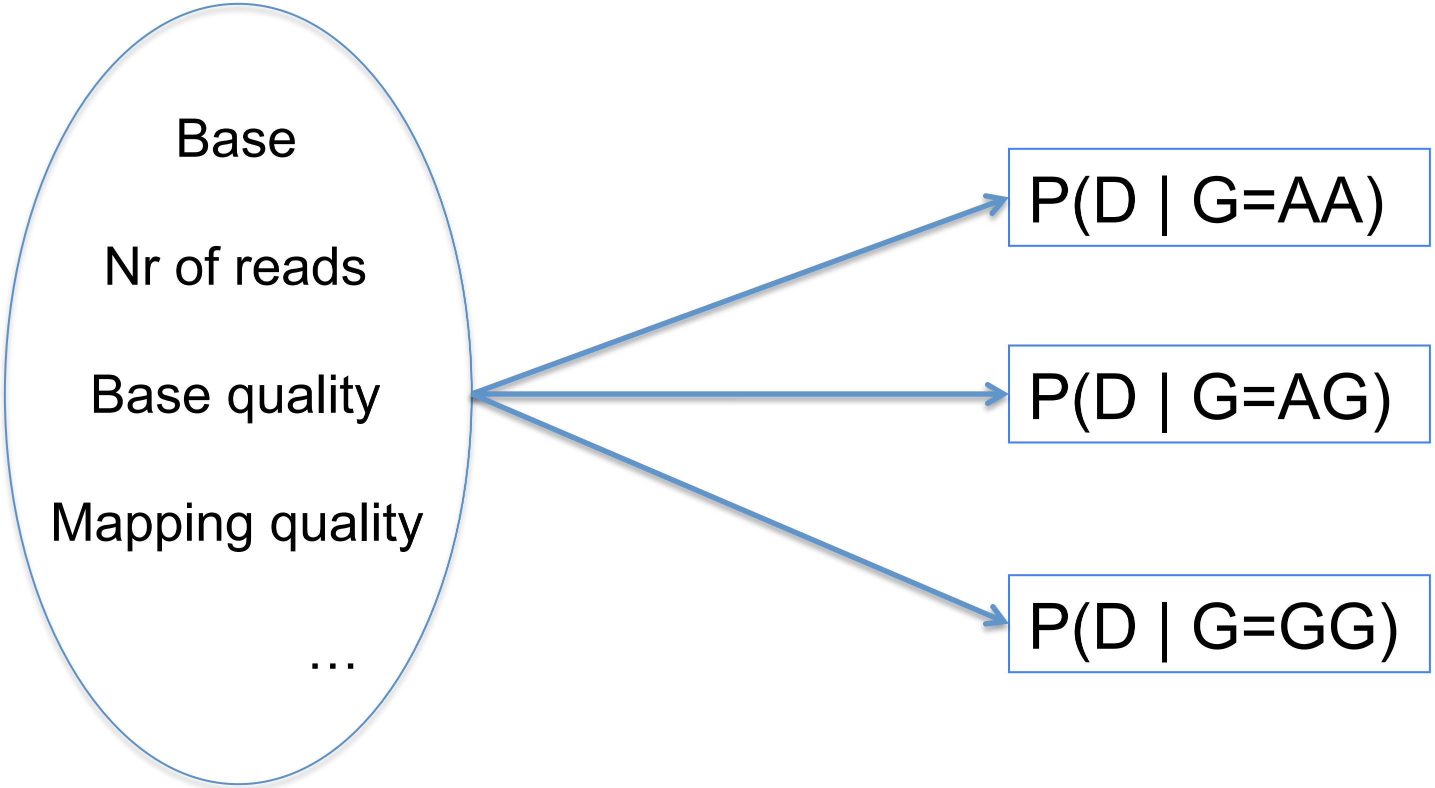
How many genotype likelihoods do we have
for each individual at each site?

3 if both alleles are known
10 if not

Genotype likelihoods

```
Chrom1    272      A        24      AAAAAGGAGAGGTAAG      <<<+;<<<<<<<<<<<=<;<;7<&
```

Base quality in Phred scale



Base

Nr of reads

Base quality

Mapping quality

…

P(D | G=AA)

P(D | G=AG)

P(D | G=GG)

Genotype likelihoods

- **SAMtools** (H Li et al., 2008): quality scores, quality dependency
- **soapSNP** (R Li et al., 2009): quality scores, quality dependency
- **GATK** (McKenna et al, 2010): quality scores
- Kim et al. (2011): type specific errors
- …

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342

A
T
T
T

Individual 1

T
T

Individual 2

A
A
T
T

Individual 3

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h \in \{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

Iterate through every read for every genotypic configuration...

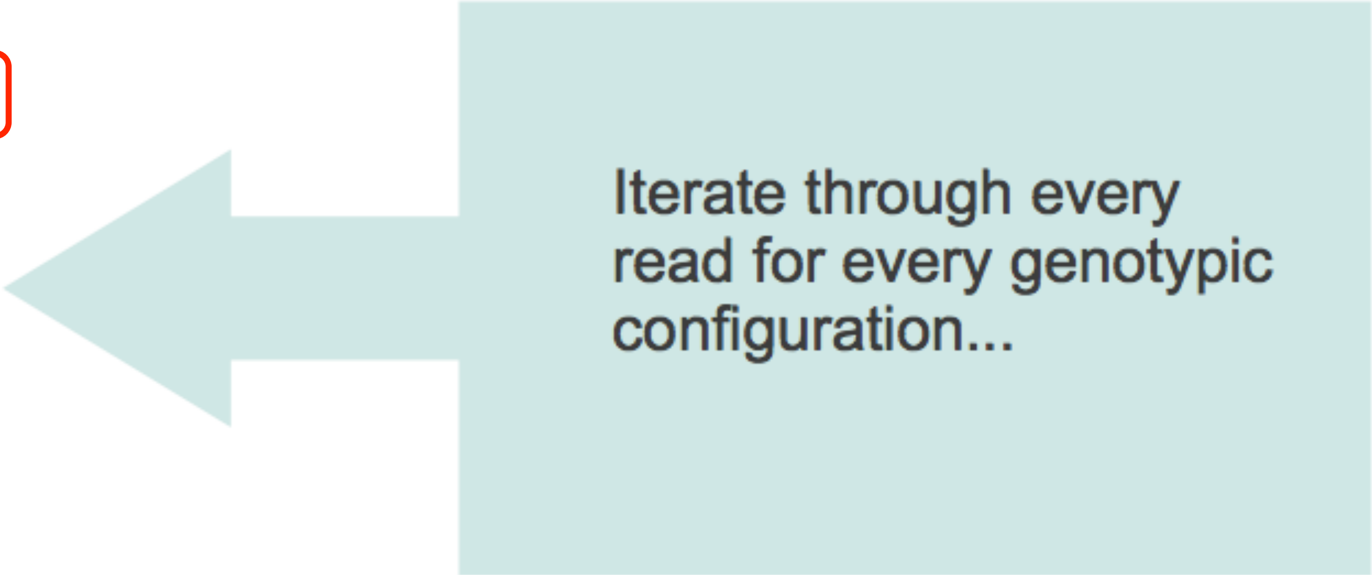
Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT



Iterate through every read for every genotypic configuration...

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$ T T T

$$P(X|G=AC)=$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$ T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

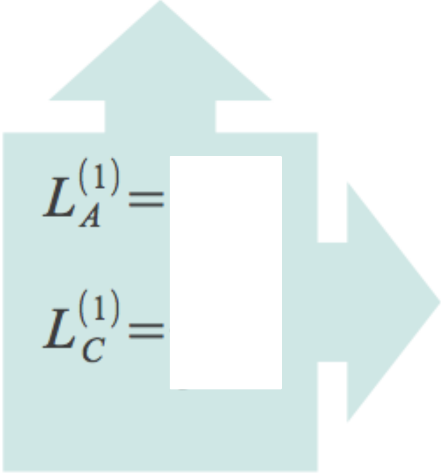
Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$ T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$



$L_A^{(1)}=$

$L_C^{(1)}=$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$ T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$$L_A^{(1)}=$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 $\boxed{\text{A}}$ T T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*$$

$$L_A^{(1)}=1-\varepsilon$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

$$P(X=A|G=AC)=\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \qquad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A $\boxed{\text{T}}$ T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*$$

↓

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A $\boxed{\text{T}}$ T T

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*$$

$$\Big\downarrow$$

$$\frac{\varepsilon}{3}$$

Calculating genotype likelihoods

$$P(X|G=bh)=\prod_{i=1}^{r}(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}) \quad b,h\in\{A,C,G,T\}$$

Example:

Chrom6 342 A T T T

$$P(X|G=AC)=(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2})*(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2})*(\frac{L_A^{(3)}}{2}+\frac{L_C^{(3)}}{2})*(\frac{L_A^{(4)}}{2}+\frac{L_C^{(4)}}{2})$$

$$=(\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6})*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}*\frac{\varepsilon}{3}$$

Genotype likelihoods

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

$ATTT$

$\varepsilon = 0.01$

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

$$ATTT$$

$$\varepsilon = 0.01$$

What is the genotype here?

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	**-1.22**
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

Simple genotype caller:
Maximum Likelihood

→ AT

Choose the genotype with the largest likelihood

Genotype calling

Genotype	Likelihood (log10)
AA	-7.44
AC	-7.74
AG	-7.74
AT	-1.22
CC	-9.91
CG	-9.91
CT	-3.38
GG	-9.91
GT	-3.38
TT	-2.49

Simple genotype caller: **Maximum Likelihood**

But **only** call the genotype if the largest likelihood is **much better** than the second best

Genotype calling

- Likelihood Ratio:

$$\log_{10} \frac{L_{G(1)}}{L_{G(2)}} > t$$

$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

(in our example *t*=1.27)

Genotype calling

- Likelihood Ratio:

$$\log_{10}\left(\frac{L_{G(1)}}{L_{G(2)}}\right) > t$$

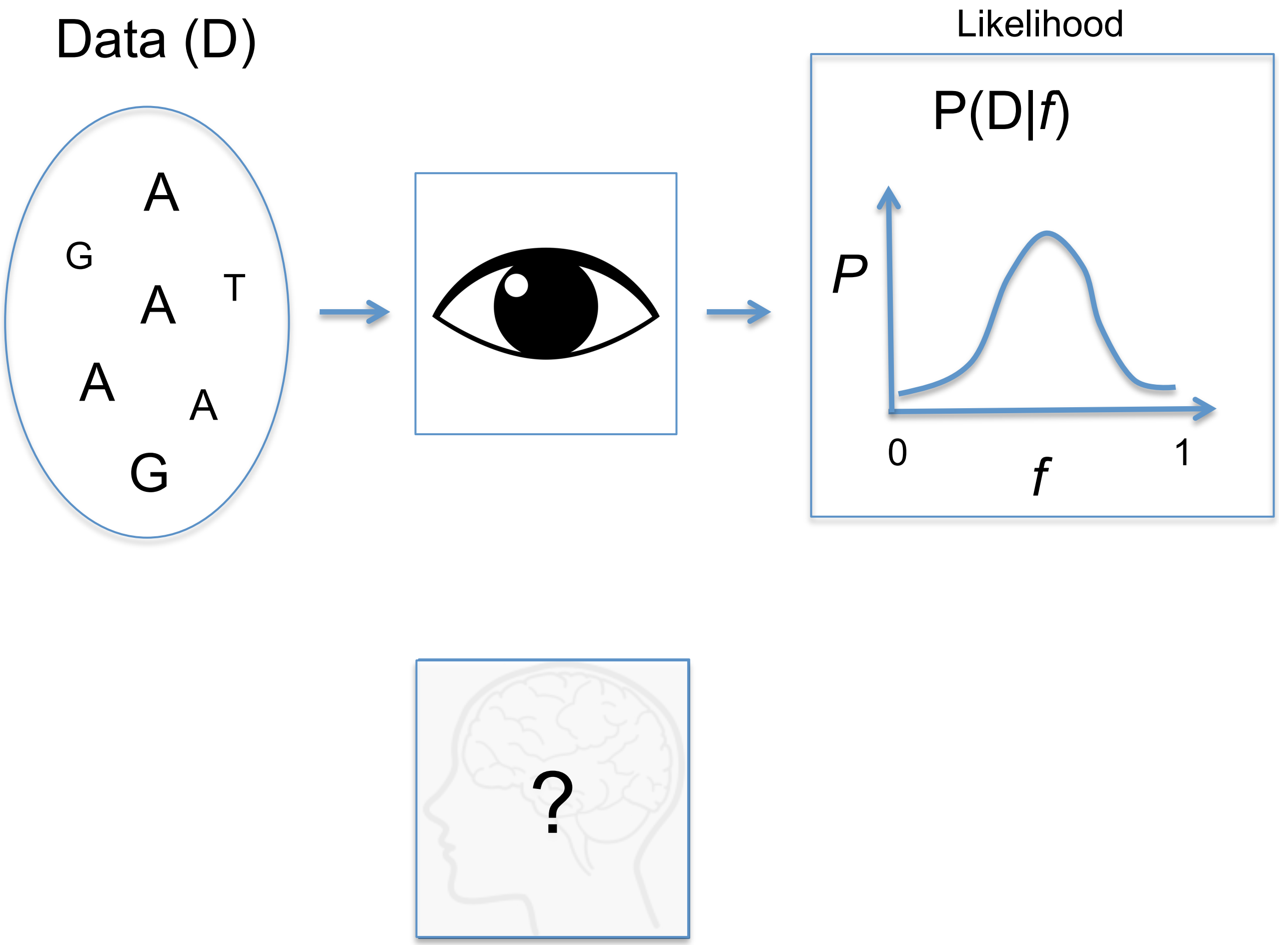
$$t = 1$$

The most likely genotype is at least **10 times** more likely than the second most likely one

↓

- Higher **confidence** of called genotypes
- More **missing** data

Statistical inference (2)

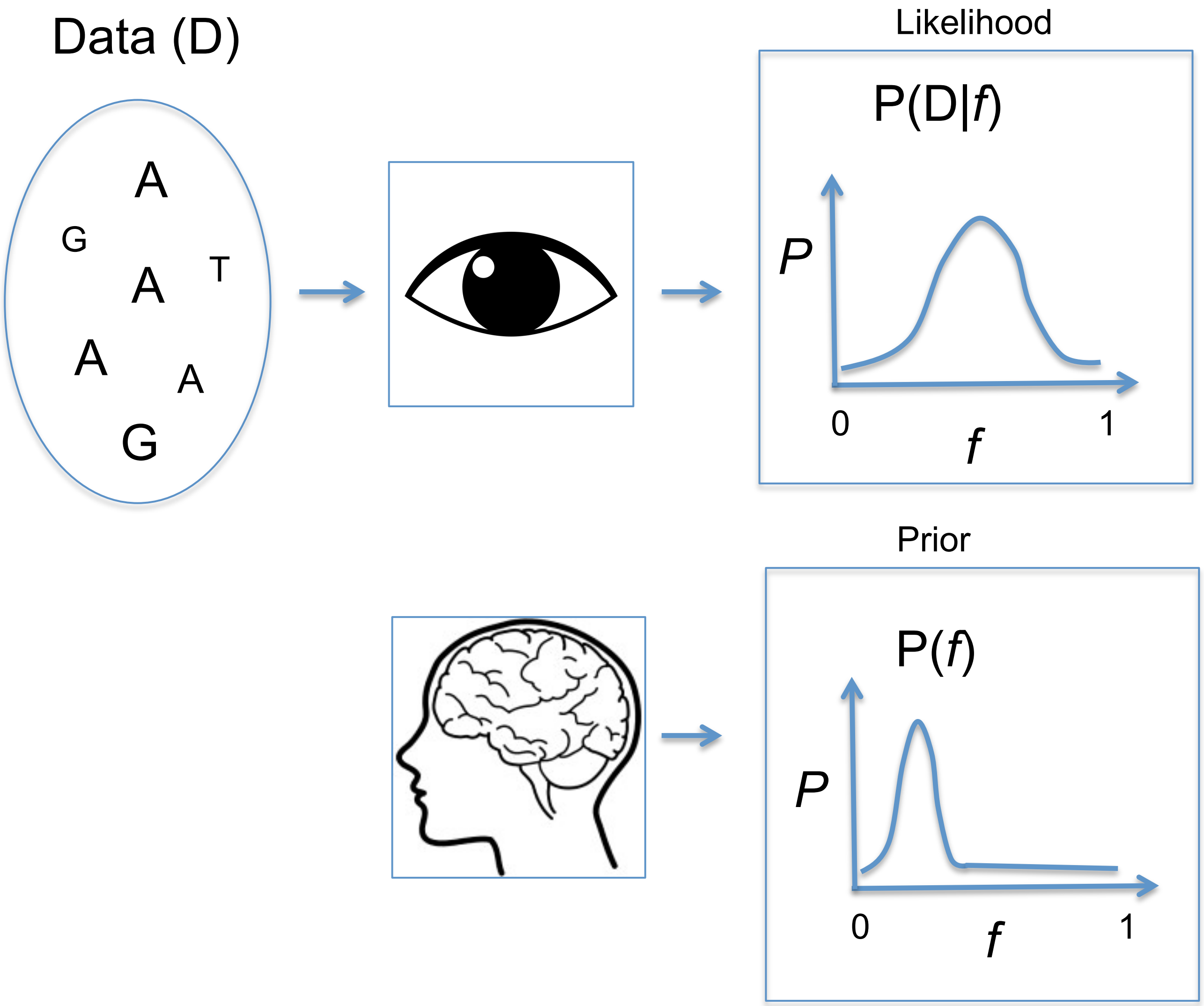


Data (D)

Likelihood

P(D|*f*)

Statistical inference (2)

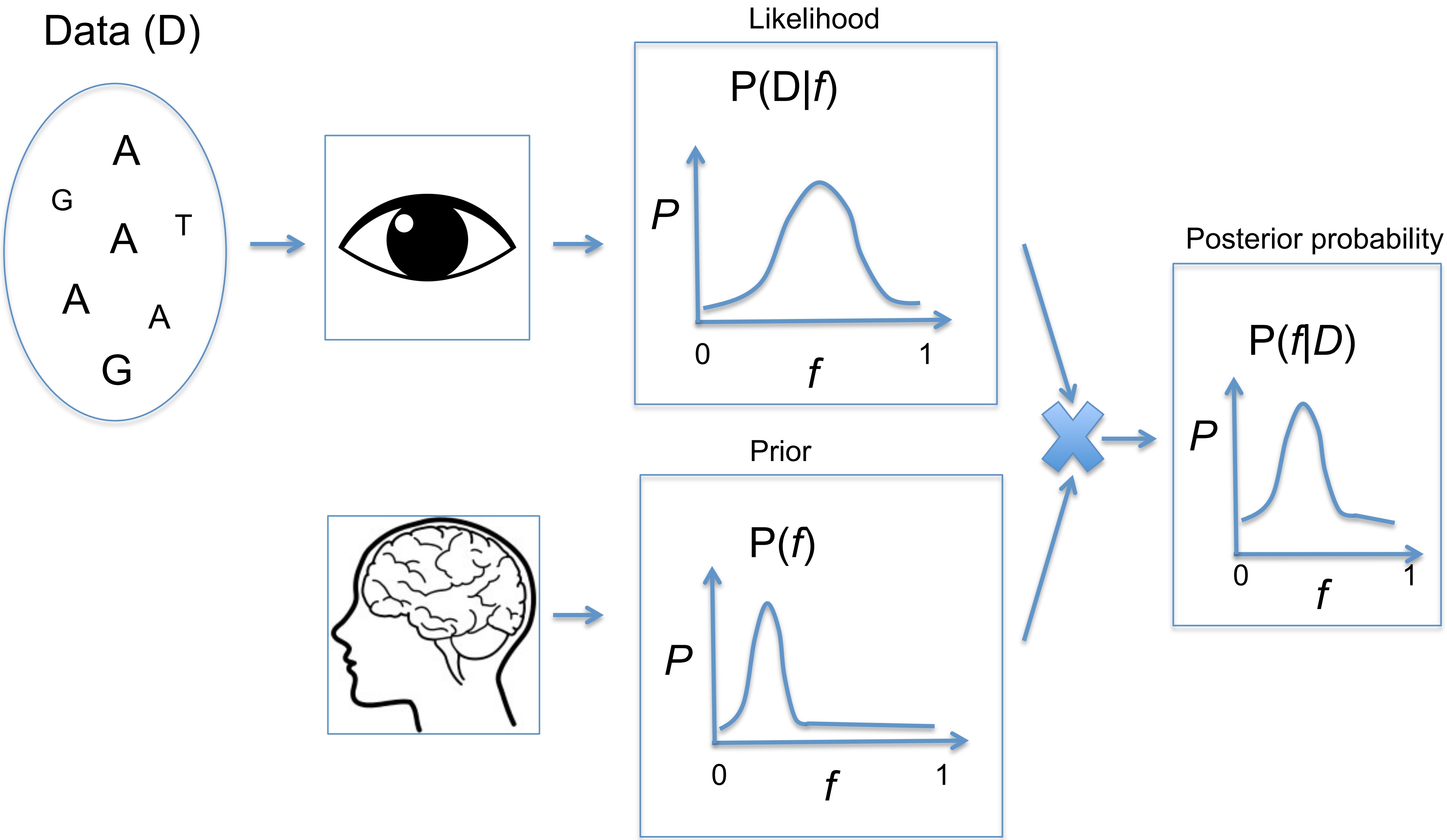


Data (D)

Likelihood

Prior

Statistical inference (2)



Bayesian inference

$$P(\theta|X)=\frac{P(X|\theta)P(\theta)}{P(X)}=\frac{P(X|\theta)P(\theta)}{\sum_{\theta}P(X|\theta)P(\theta)}$$

$P(X|\theta)$ ← Likelihood of θ

$P(\theta)$ ← Prior probability distribution of θ

$P(\theta|X)$ ← Posterior probability distribution of θ

- Parameter is not fixed (point estimate) but rather has a probability distribution
- We update our “belief” on the parameter after performing the experiment
- As P(f|D) is a proper probability distribution, we can easily derive credible intervals

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49		

Simple genotype caller:
Bayesian

→ ?

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	0.94
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller:
Bayesian

?

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller:
Bayesian

→ AT

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	1/10	~ 0
AC	-7.74	1/10	~ 0
AG	-7.74	1/10	~ 0
AT	-1.22	1/10	**0.94**
CC	-9.91	1/10	~ 0
CG	-9.91	1/10	~ 0
CT	-3.38	1/10	0.006
GG	-9.91	1/10	~ 0
GT	-3.38	1/10	0.006
TT	-2.49	1/10	0.05

Simple genotype caller: **Bayesian**

But **only** call the genotype if the largest probability is above a threshold (e.g. > 0.95)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49		

Simple genotype caller:
Bayesian

Example: reference is T

→ AT (?)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.01	~ 0
AC	-7.74	0.01	~ 0
AG	-7.74	0.01	~ 0
AT	-1.22	0.09	**0.67**
CC	-9.91	0.01	~ 0
CG	-9.91	0.01	~ 0
CT	-3.38	0.09	0.005
GG	-9.91	0.01	~ 0
GT	-3.38	0.09	0.0005
TT	-2.49	0.81	**0.32**

Simple genotype caller:
Bayesian

P(A) = 0.9 if A is the **reference** allele;
P(A) = 0.1 otherwise

→ AT (?)

Example: reference is T

$P(TT) = P(A)^2$

e.g. Illumina Casava

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49		

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

P(TT) = …
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22		
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller:
Bayesian

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
P(AT) = …
P(AA) = …

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44		
AC	-7.74		
AG	-7.74		
AT	-1.22	0.38	
CC	-9.91		
CG	-9.91		
CT	-3.38		
GG	-9.91		
GT	-3.38		
TT	-2.49	0.56	

Better genotype caller: **Bayesian**

P(A) = f

Where f (=0.75) is the **allele frequency** from a reference panel

Example: reference is T

$P(TT) = f^2$
$P(AT) = 2f(1-f)$
$P(AA) = \ldots$

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.06	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.38	**0.93**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.56	0.07

Better genotype caller: **Bayesian**

P(A) = f

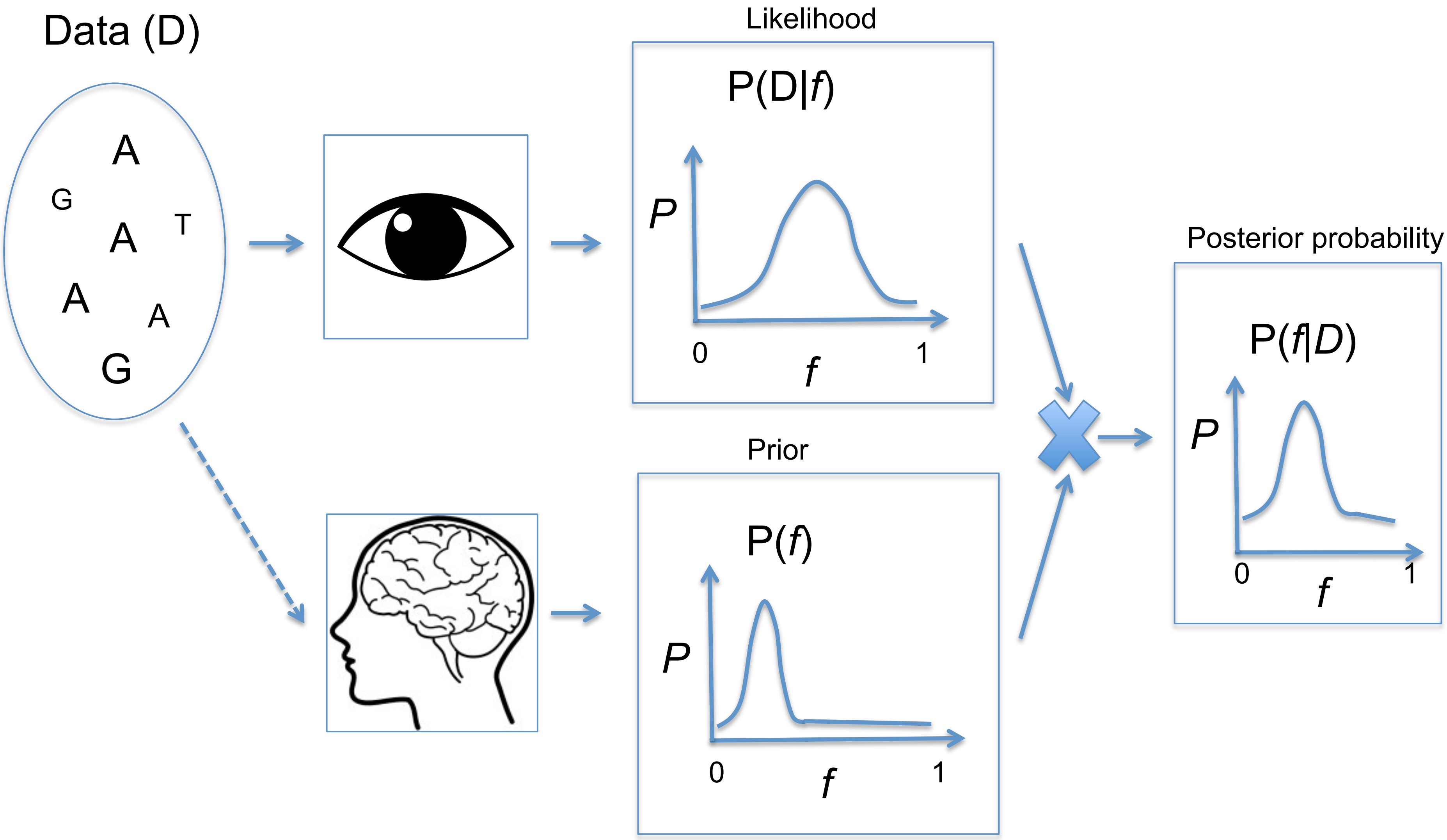
Where f is the **allele frequency** from a reference panel

Example: reference is T

P(TT) = f^2
P(AT) = 2f(1-f)
P(AA) = $(1-f)^2$

Assuming **f=0.75** and only **A and T** alleles

Statistical inference (3)



Data (D)

Likelihood

P(D|*f*)

Prior

P(*f*)

Posterior probability

P(*f*|*D*)

Genotype posterior probabilities

Genotype	Likelihood (log10)	Prior	Posterior probability
AA	-7.44	0.16	~ 0
AC	-7.74	0	0
AG	-7.74	0	0
AT	-1.22	0.48	**0.96**
CC	-9.91	0	0
CG	-9.91	0	0
CT	-3.38	0	0
GG	-9.91	0	0
GT	-3.38	0	0
TT	-2.49	0.36	0.38

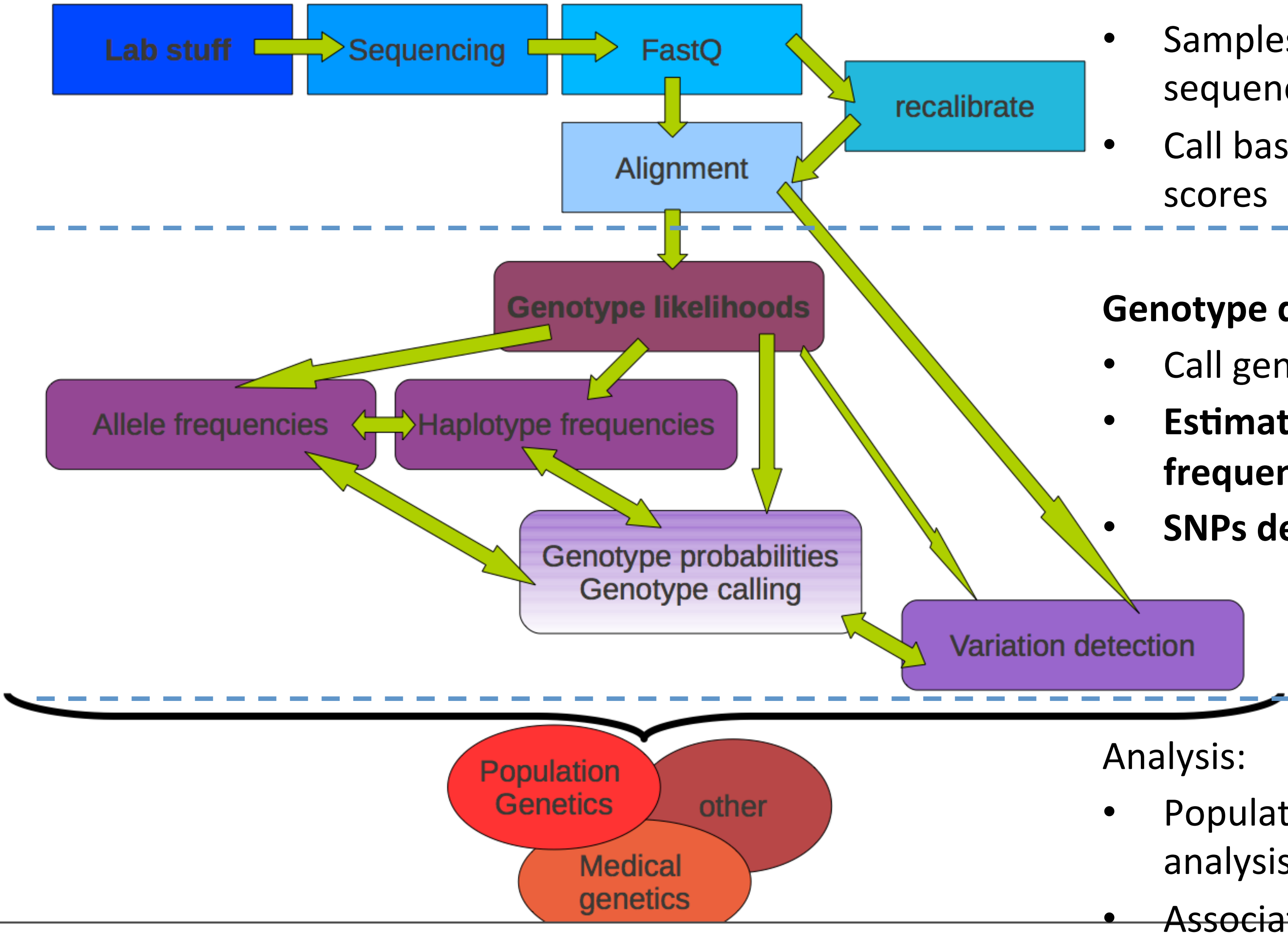
Better genotype caller: **Empirical Bayesian**

P(A) = f

Where f is the **allele frequency** estimated from the data itself

With **f=0.6**

Workflow



Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

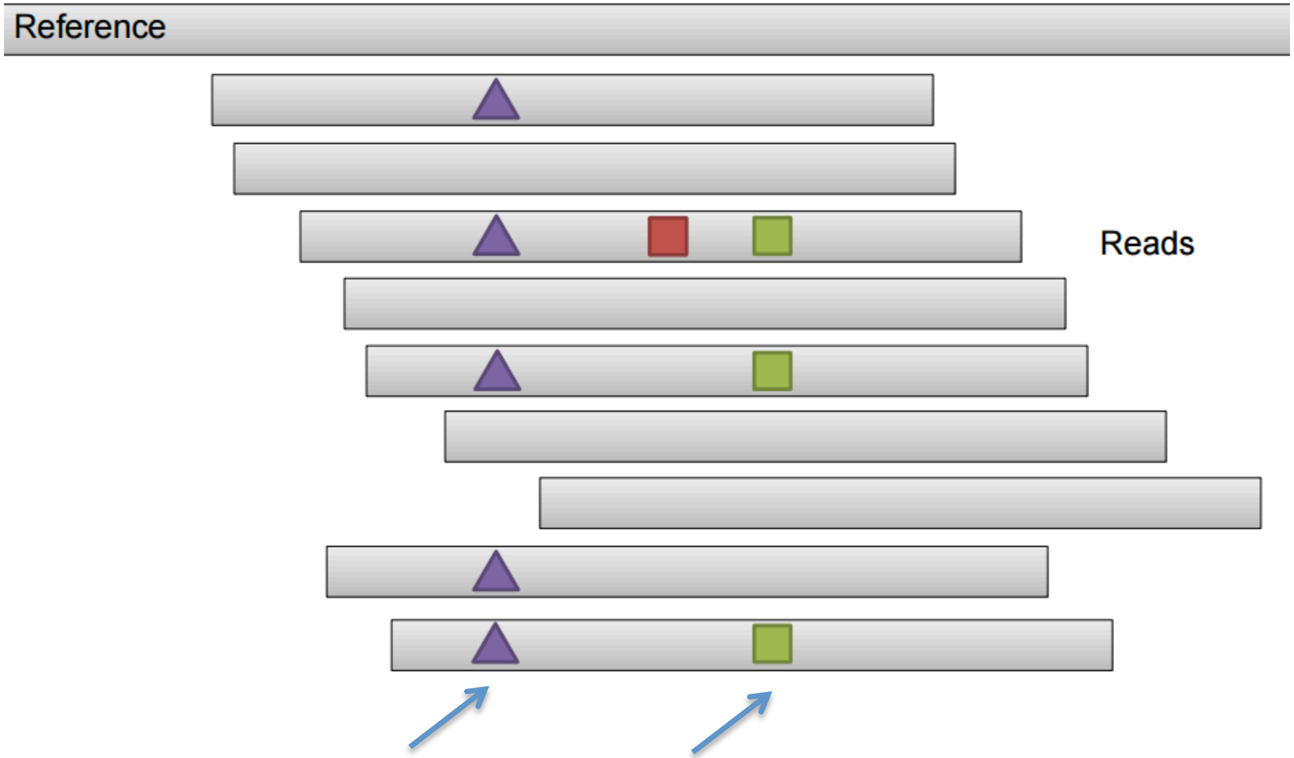
- Call genotypes
- **Estimate allele frequencies**
- **SNPs detection**

Analysis:

- Population genetics analysis
- Association studies

SNP calling procedures

- Alignment-based caller



We completely rely on how reads have been mapped

Figure from Erik Garrison

SNP calling procedures

- Assembly-based caller (as in GATK)

Local re-alignment around putative variants; better resolution for INDELs detection.

- Haplotype-based caller (as in freebayes)

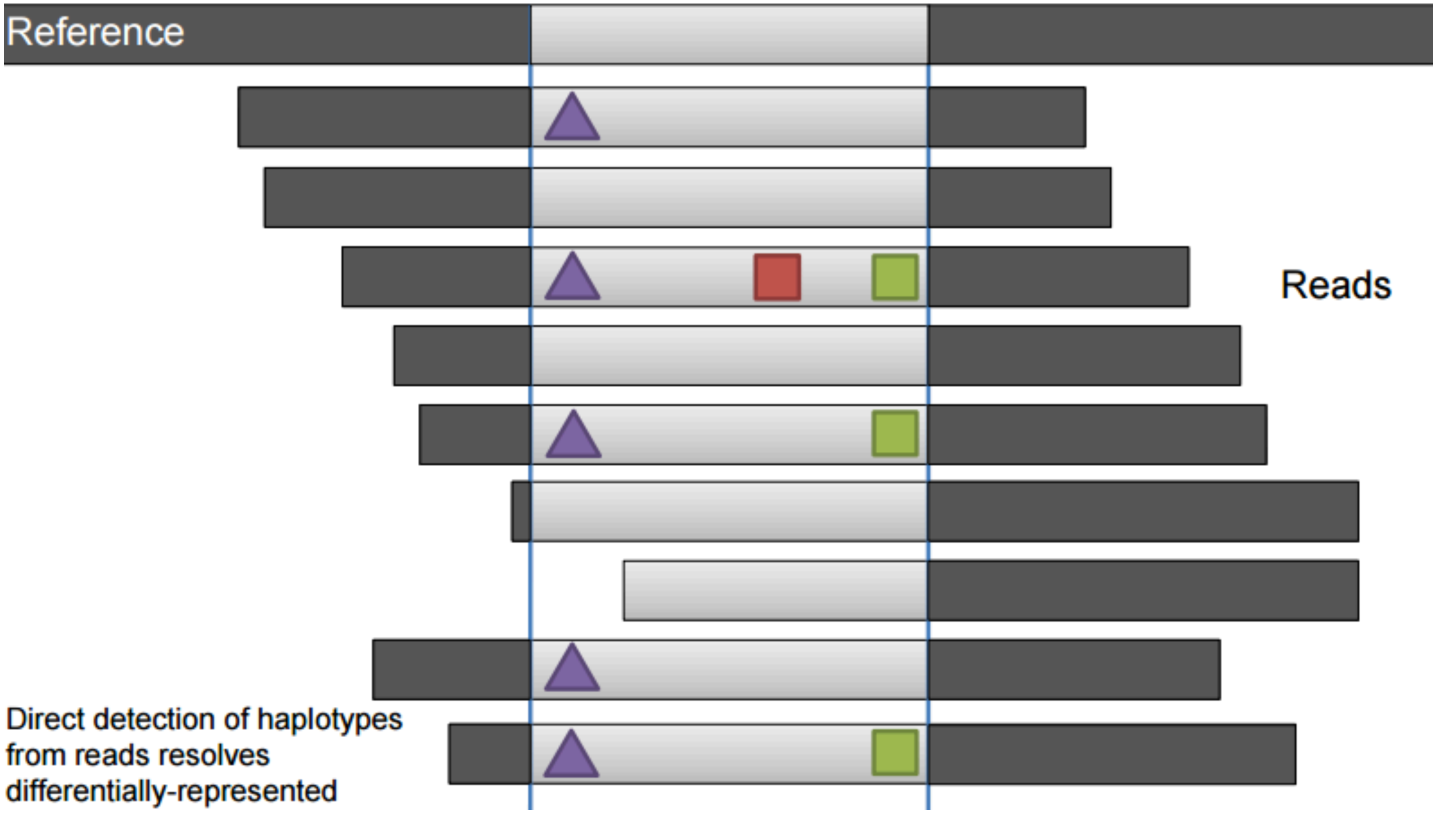

Reference
Reads
Direct detection of haplotypes from reads resolves differentially-represented

Figure from Erik Garrison

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA		
2	AA		
3	AG		
4	AG		
5	GG		
6	GG		
Tot.			

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individua l	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Assume only 2 allelic types

True allele frequency is 0.50

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})}$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})} = 0.75$$

Estimating allele frequencies

Individual	**True genotype**	**Reads allele A**	**Reads allele G**
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts**

$$\hat{f} = \frac{\sum_{i=1}^{N} n_{(A,i)}}{\sum_{i=1}^{N}\left(n_{(A,i)} + n_{(G,i)}\right)} = 0.75$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N} (n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})(1 - 2\varepsilon)}$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N} (n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N} (n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator:

from **reads counts with error**

$$\hat{f} = \frac{\sum_{i=1}^{N}(n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)}))}{\sum_{i=1}^{N}(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)} = 0.77$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Simple allele frequency estimator: from **reads counts with error and weights** (Y Li et al. 2010)

$$p_i = \frac{n_{(A,i)} - \varepsilon(n_{(A,i)} + n_{(G,i)})}{(n_{(A,i)} + n_{(G,i)})(1-2\varepsilon)}$$

Weighting function

$$\hat{f} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} p_i w_i = 0.57$$

Estimating allele frequencies

Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

Estimating allele frequencies

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$L = \prod_{i=1}^{N} p(D_i \mid f)$$

Genotype likelihoods

$$p(D_i \mid f) = \sum_{g \in \{0,1,2\}} p(D \mid G = g) p(G = g \mid f)$$

If we assume HWE: $p(G = AA \mid f) = f^2$

$$p(G = AG \mid f) = 2f(1-f)$$

$$p(G = GG \mid f) = (1-f)^2$$

Estimating allele frequencies

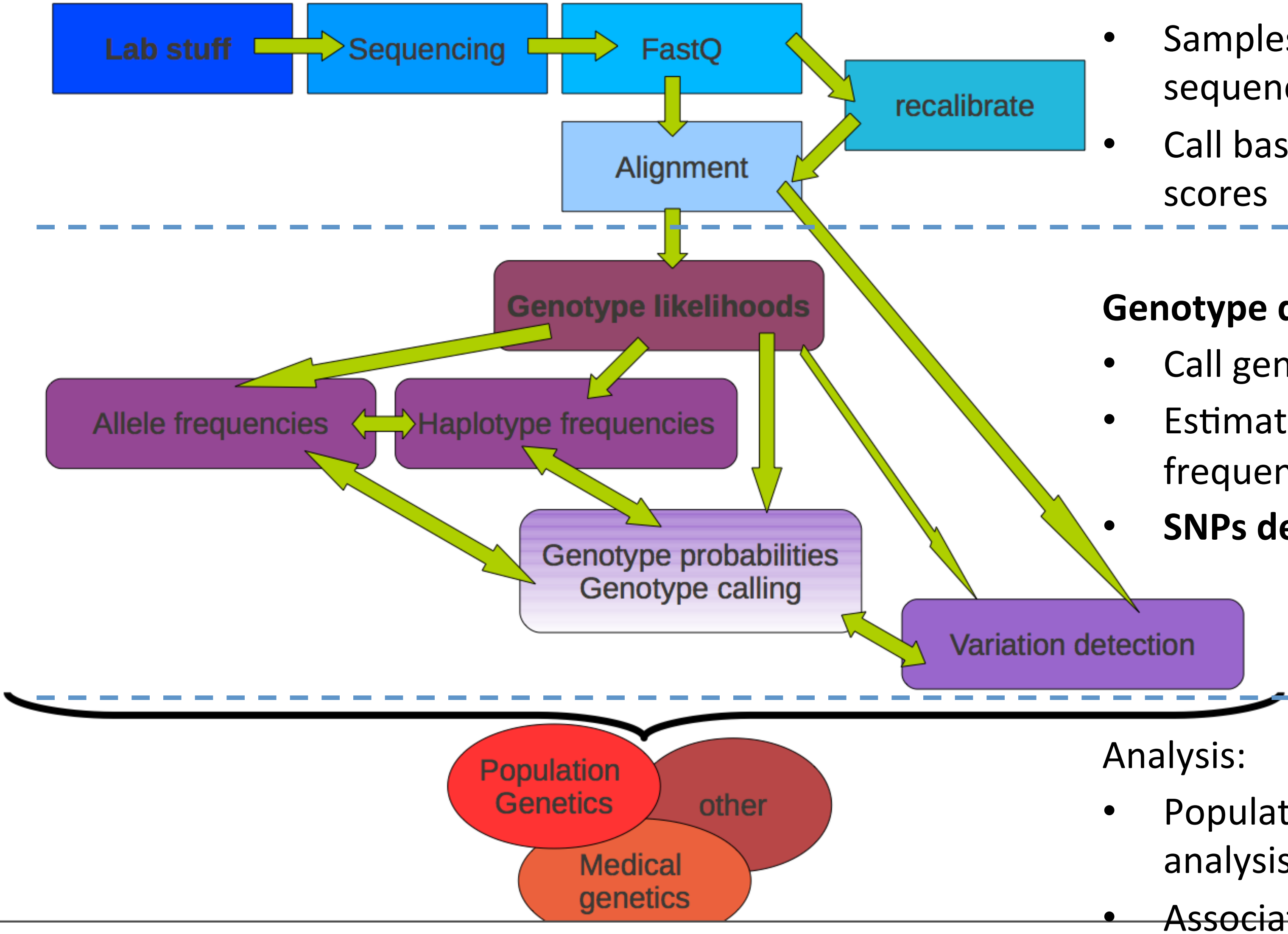
Individual	True genotype	Reads allele A	Reads allele G
1	AA	7	0
2	AA	25	1
3	AG	5	3
4	AG	4	4
5	GG	0	2
6	GG	0	4
Tot.		41	14

Maximum Likelihood (ML) estimator (Kim et al. 2011)

$$\hat{f} = \arg\max_p \prod_{i=1}^{N} p(D_i \mid f)$$

$$\hat{f} = 0.46$$

Workflow



Low-level data:

- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- **SNPs detection**

Analysis:

- Population genetics analysis
- Association studies

SNP calling

- A lot of missing data if calling genotypes at low depth (heterozygotes can be lost!)

- Rare variants are hard to detect

- Trade-off between False Positives and False Negatives

SNP calling

- What is the most straightforward method for SNP calling?

SNP calling

- What is the most straightforward method for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - …

SNP calling

- What is the most straightforward method for SNP calling?
  - Assign as SNPs sites where at least one heterozygote has been called
  - Assign as SNPs sites where the estimated allele frequency is above a certain threshold (e.g. ?)

SNP calling

- MLE of allele frequency at each site:

Call a SNP if

$$\hat{f}_{MLE} > t$$

Where *t* can be defined as the minimum sample allele frequency detectable (e.g. with 10 samples *t* can be set to 0.05)

Likelihood Ratio Test

- Compare the goodness of fit of the null and alternative model

- Null Model: frequency=0
- Alternative Model: frequency>0

The model with more parameters “tends” to fit better.

Whether or not this fit is “significantly better” is assessed by the comparison of the two likelihoods.

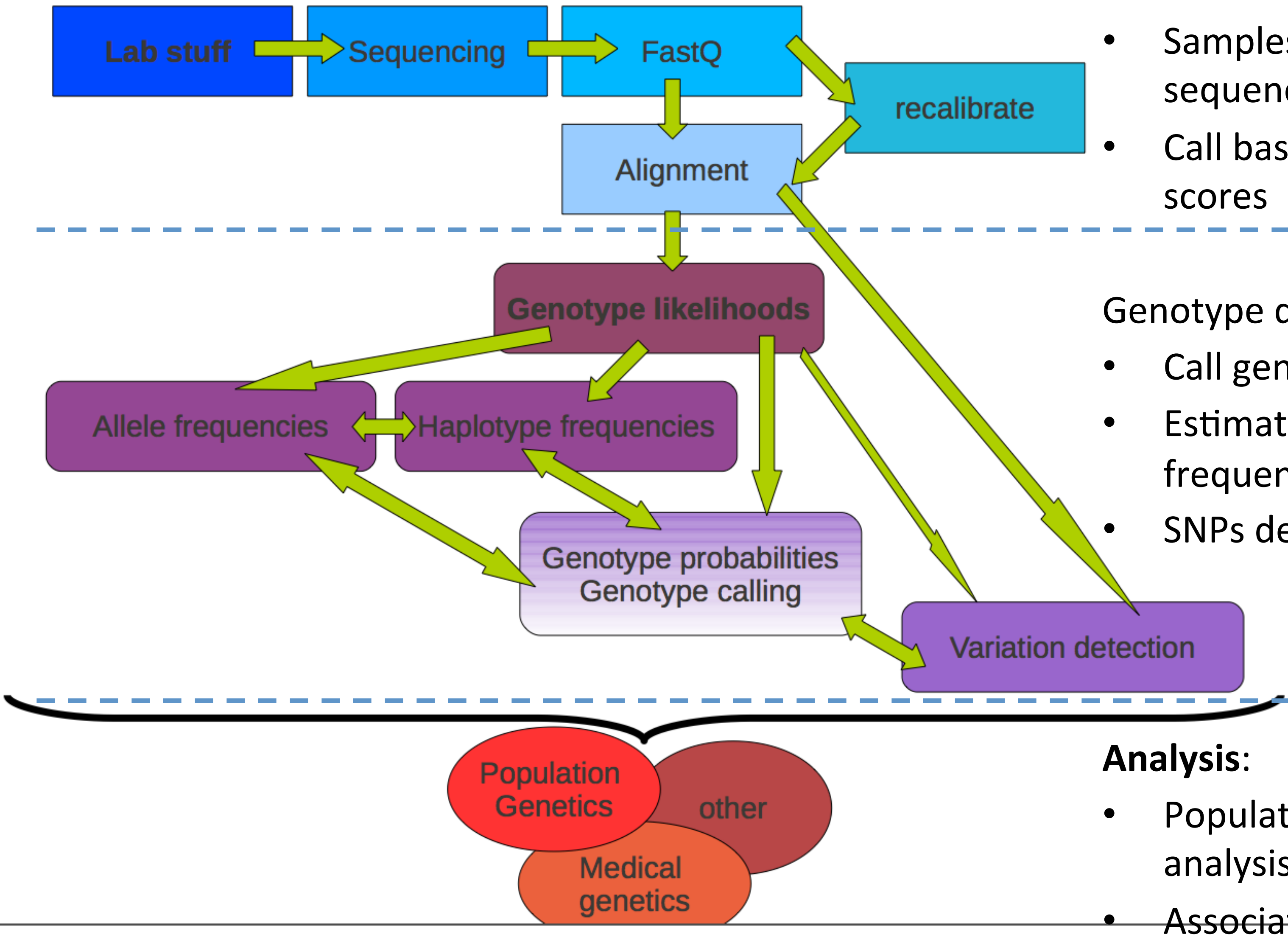
SNP calling

- Likelihood Ratio Test (**LRT**): test statistical hypotheses based on comparing the maximum likelihood under 2 different models.

$$T = -2\ln\left(\frac{L(f = 0)}{L(f \neq 0)}\right)$$

T is chi-squared distributed with 1 degree of freedom -> assign a *p*-value

Workflow



Low-level data:

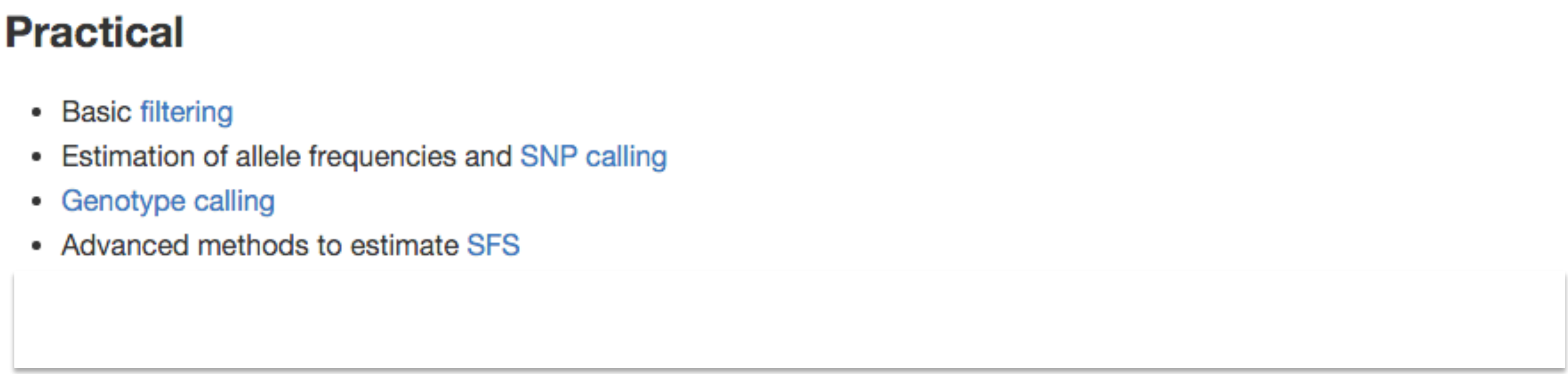
- Samples preparation + sequencing
- Call bases and quality scores

Genotype data:

- Call genotypes
- Estimate allele frequencies
- SNPs detection

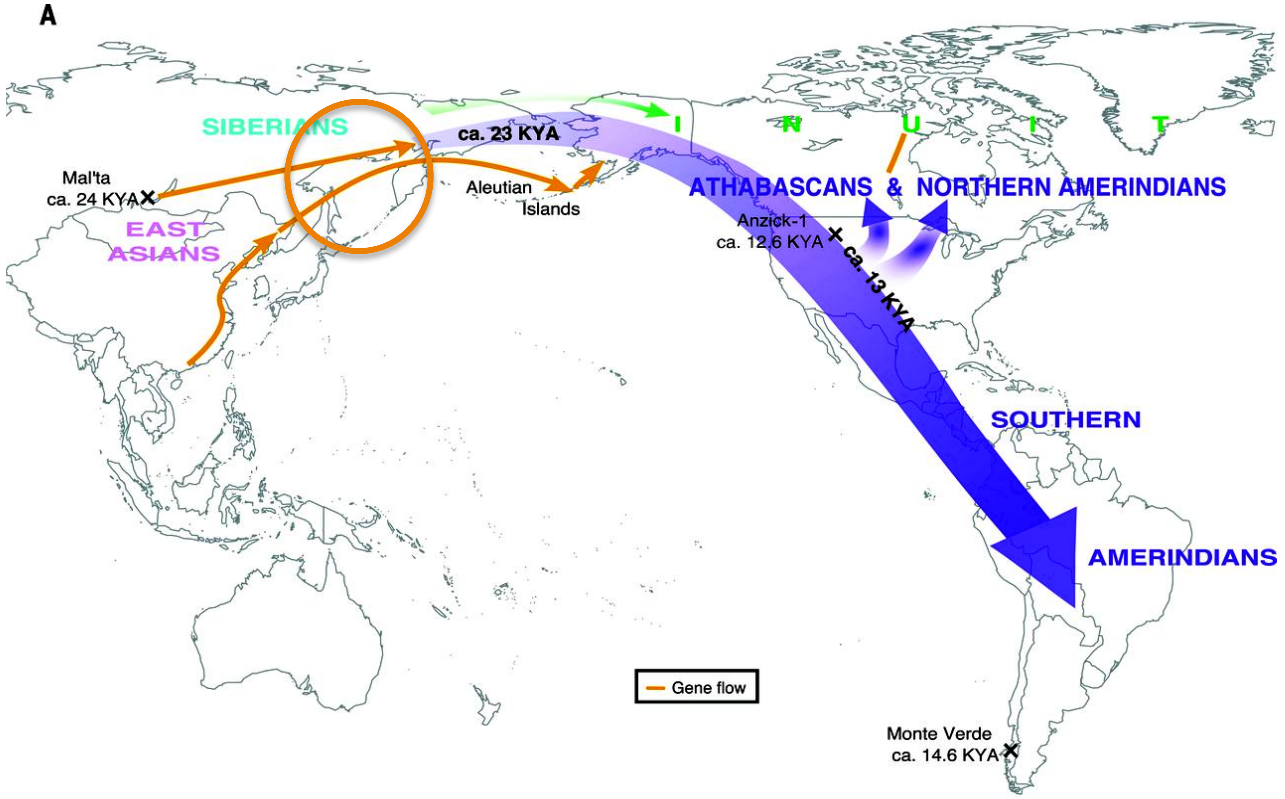
Analysis:

- Population genetics analysis
- Association studies



Practical

- Basic filtering
- Estimation of allele frequencies and SNP calling
- Genotype calling
- Advanced methods to estimate SFS



Raghavan et al. 2015 Science