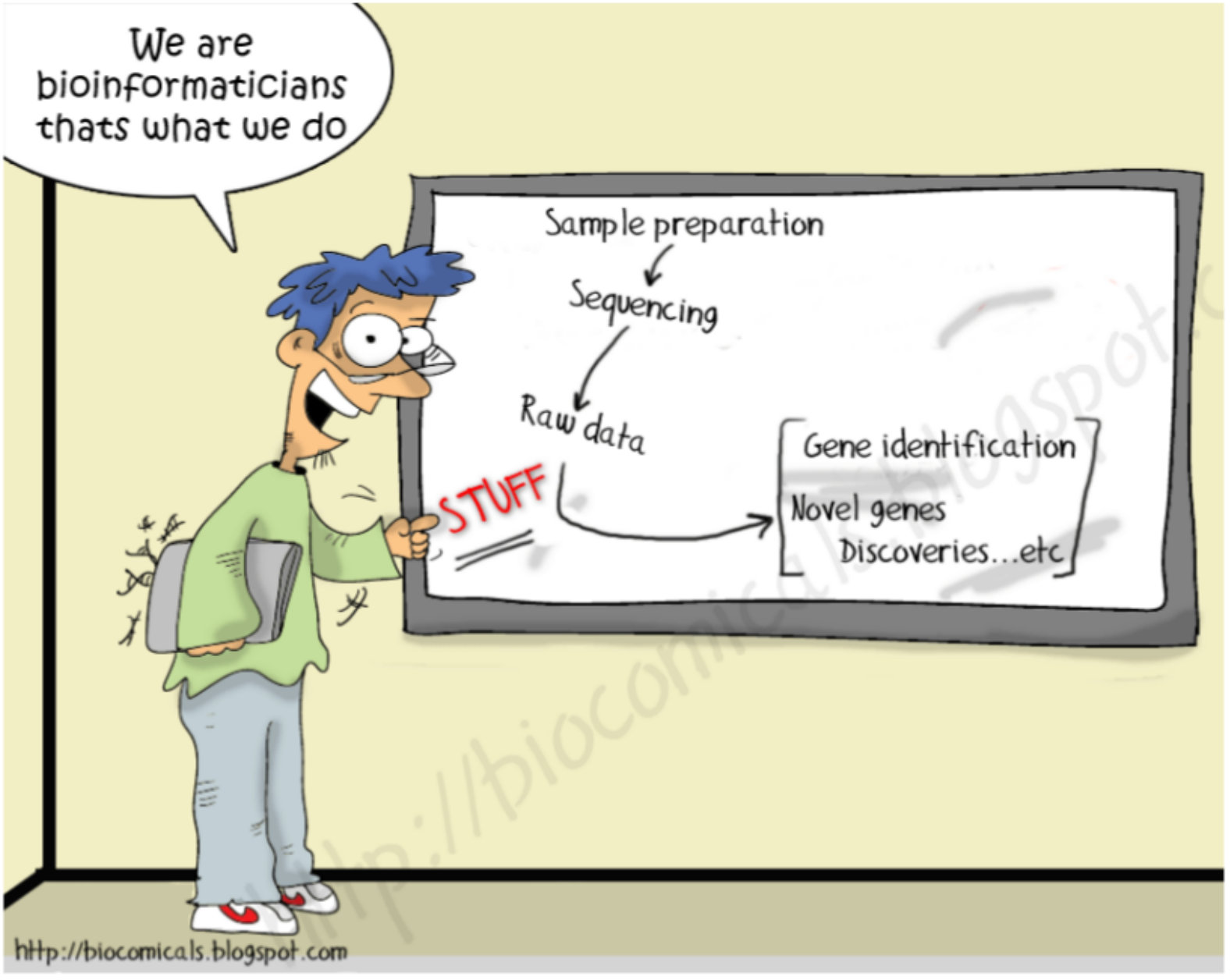
Imperial College
London

15th July 2019

# *Genotype and SNP calling and estimation of allele frequencies*

Matteo Fumagalli


We are bioinformaticians thats what we do
Sample preparation
Sequencing
Raw data
STUFF
Gene identification
Novel genes
Discoveries...etc
http://biocomicals.blogspot.com

Imperial College London

# Presentation outline

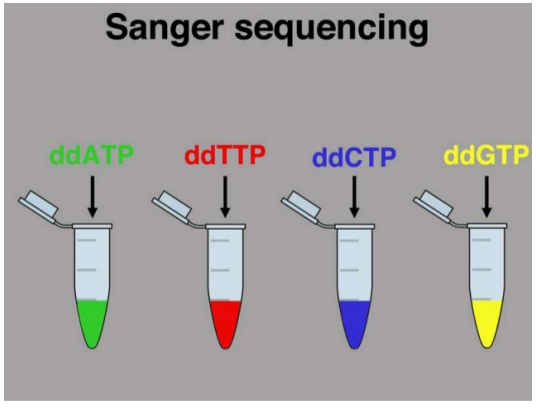
1. Motivation
2. Genotype likelihoods
3. Genotype calling
4. SNP calling
5. Imputation

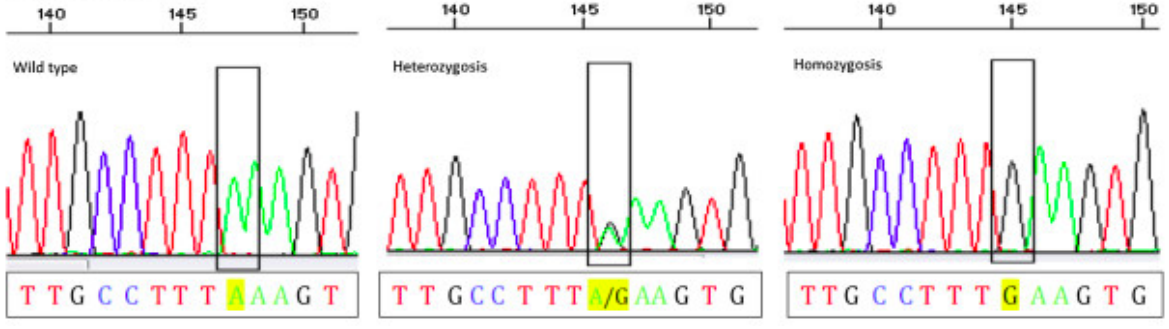
# Motivation

Imperial College London

# Sanger sequencing

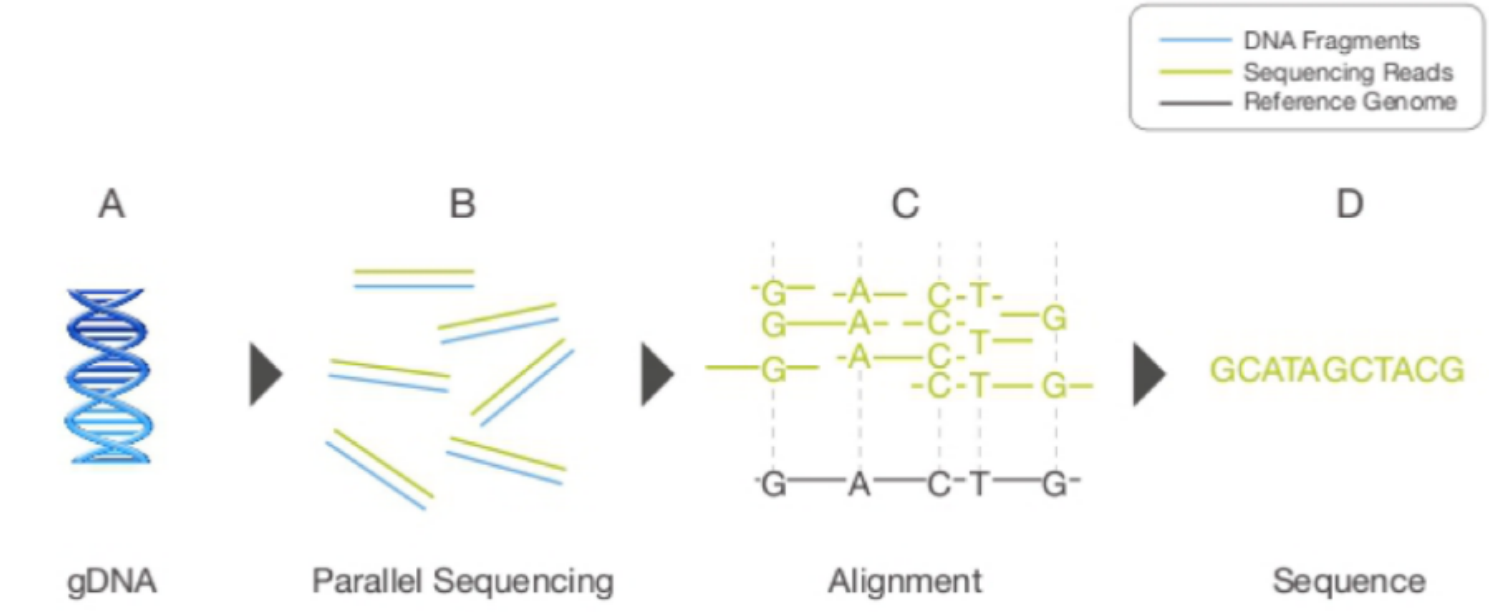
aka first/former generation sequencing



Imperial College London

# 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.

Imperial College London

# From genomes 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)

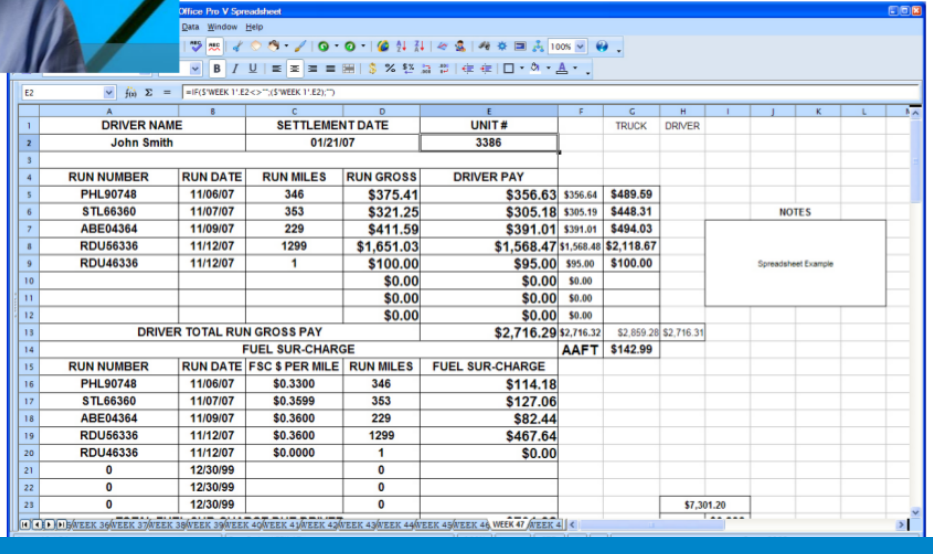
```
seq1 272 T 24  ,.$.....,,.,.,...,,,,,,,,^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,,,,,,A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$.....,,.,.,...,,,,,,,,    7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$.....,,.,.,...,,,,,,,,^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,.,...,,,,,,,,  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
```

Imperial College London

# Forget about

| DRIVER NAME | SETTLEMENT DATE | UNIT # | | TRUCK | DRIVER |
|---|---|---|---|---|---|
| John Smith | 01/21/07 | 3386 | | | |

| RUN NUMBER | RUN DATE | RUN MILES | RUN GROSS | DRIVER PAY | | | |
|---|---|---|---|---|---|---|---|
| PHL90748 | 11/06/07 | 346 | $375.41 | $356.63 | $356.64 | $489.59 | |
| STL66360 | 11/07/07 | 353 | $321.25 | $305.18 | $305.19 | $448.31 | |
| ABE04364 | 11/09/07 | 229 | $411.59 | $391.01 | $391.01 | $494.03 | |
| RDU56336 | 11/12/07 | 1299 | $1,651.03 | $1,568.47 | $1,568.48 | $2,118.67 | |
| RDU46336 | 11/12/07 | 1 | $100.00 | $95.00 | $95.00 | $100.00 | |
| | | | $0.00 | $0.00 | $0.00 | | |
| | | | $0.00 | $0.00 | $0.00 | | |
| | | | $0.00 | $0.00 | $0.00 | | |
| DRIVER TOTAL RUN GROSS PAY | | | | $2,716.29 | $2,716.32 | $2,859.28 | $2,716.31 |

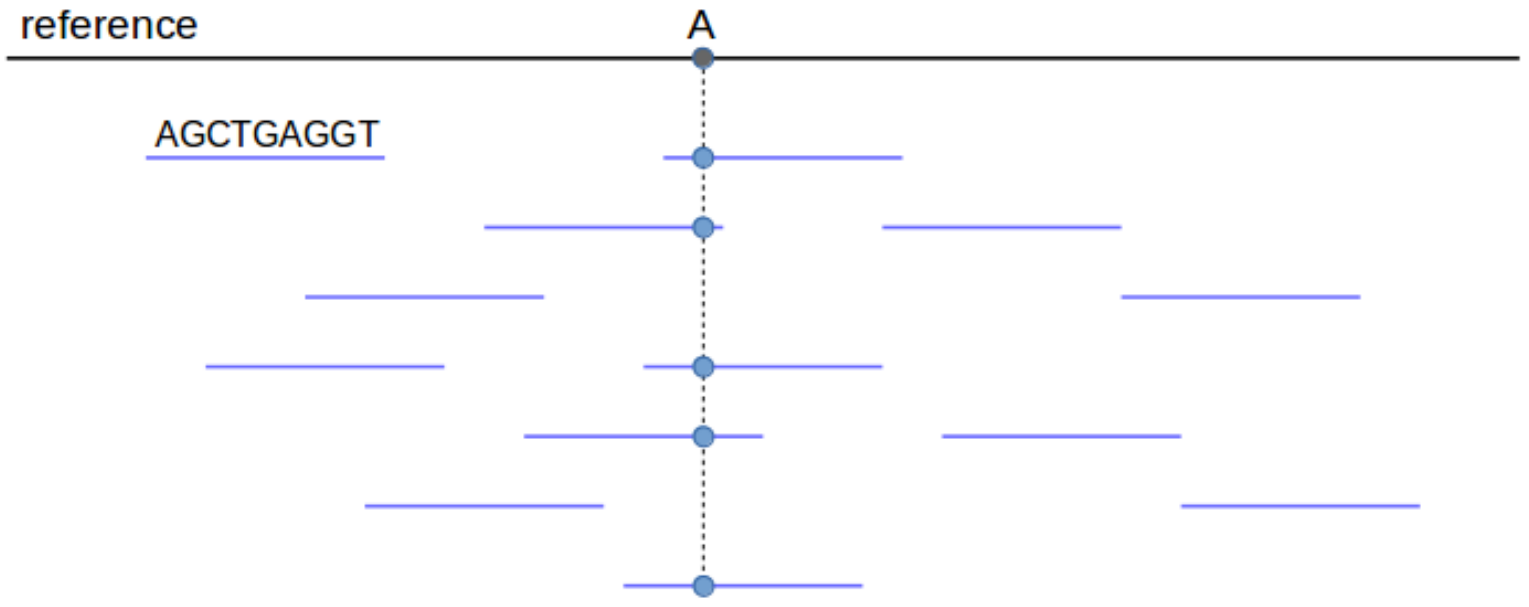
NOTES

Spreadsheet Example

| FUEL SUR-CHARGE | | | | | AAFT | $142.99 | |
|---|---|---|---|---|---|---|---|
| RUN NUMBER | RUN DATE | FSC $ PER MILE | RUN MILES | FUEL SURCHARGE | | | |
| PHL90748 | 11/06/07 | $0.3300 | 346 | $114.18 | | | |
| STL66360 | 11/07/07 | $0.3599 | 353 | $127.06 | | | |
| ABE04364 | 11/09/07 | $0.3600 | 229 | $82.44 | | | |
| RDU56336 | 11/12/07 | $0.3600 | 1299 | $467.64 | | | |
| RDU46336 | 11/12/07 | $0.0000 | 1 | $0.00 | | | |
| 0 | 12/30/99 | | 0 | | | | |
| 0 | 12/30/99 | | 0 | | | | |
| 0 | 12/30/99 | | 0 | | | | $7,301.20 |

Imperial College London

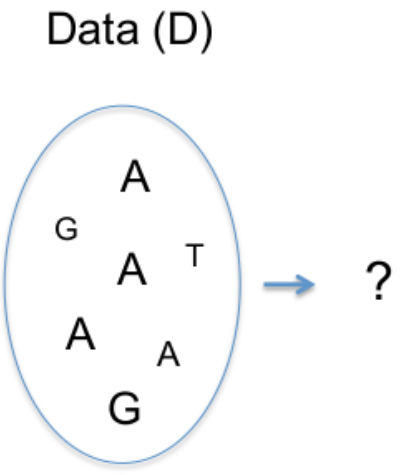
# Why do we need statistics?




- is a **nucleotide**/base/allele with a certain **quality** score

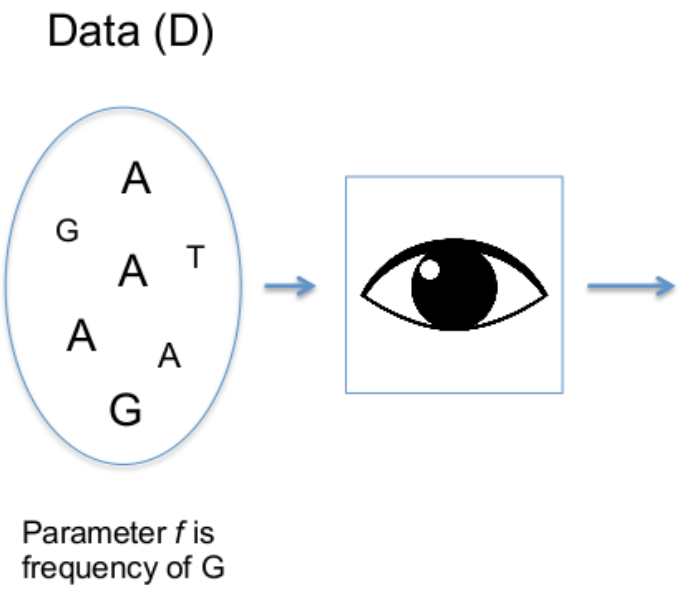
Imperial College London

# Statistical inference



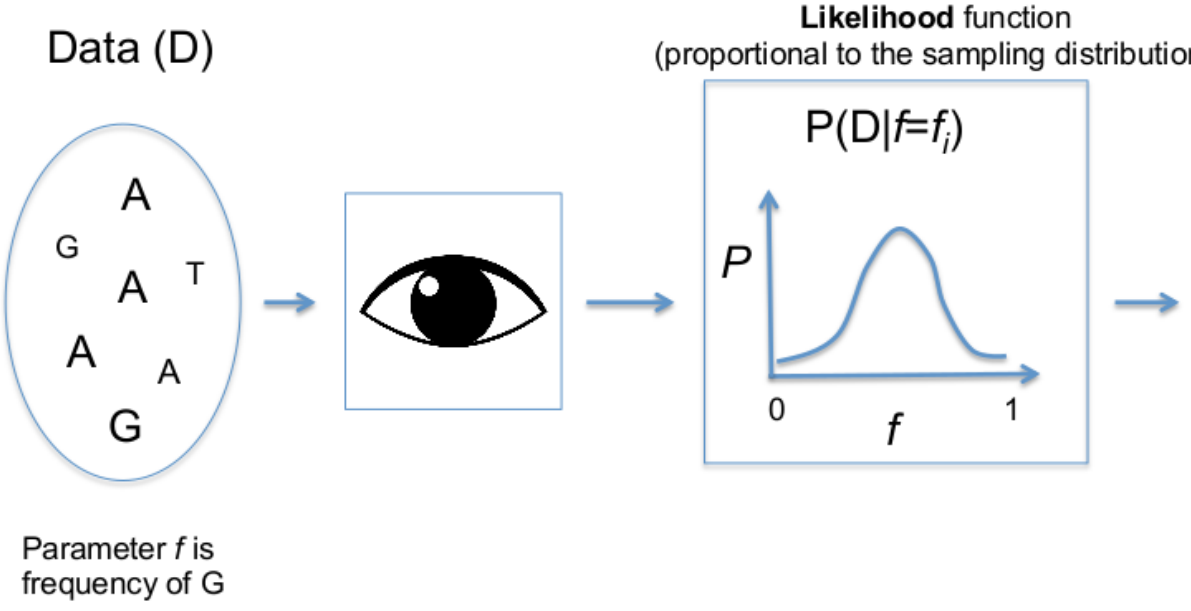
Imperial College London

# Statistical inference



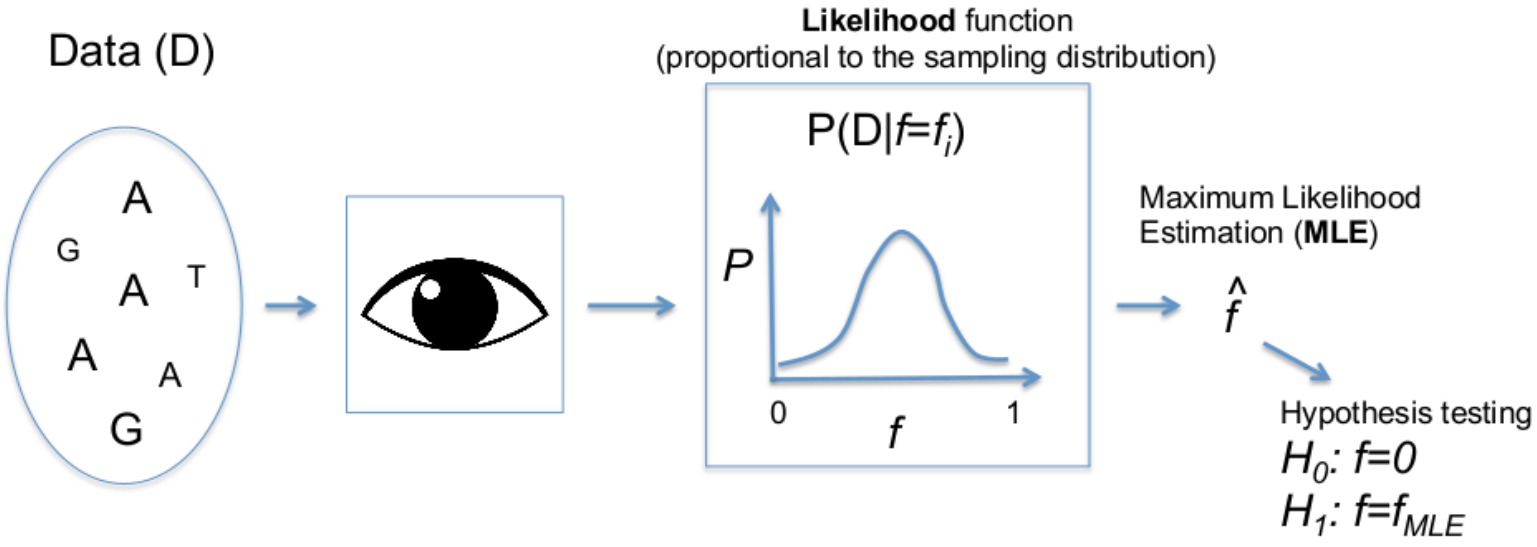
Imperial College London

# Statistical inference



Imperial College London

# Statistical inference

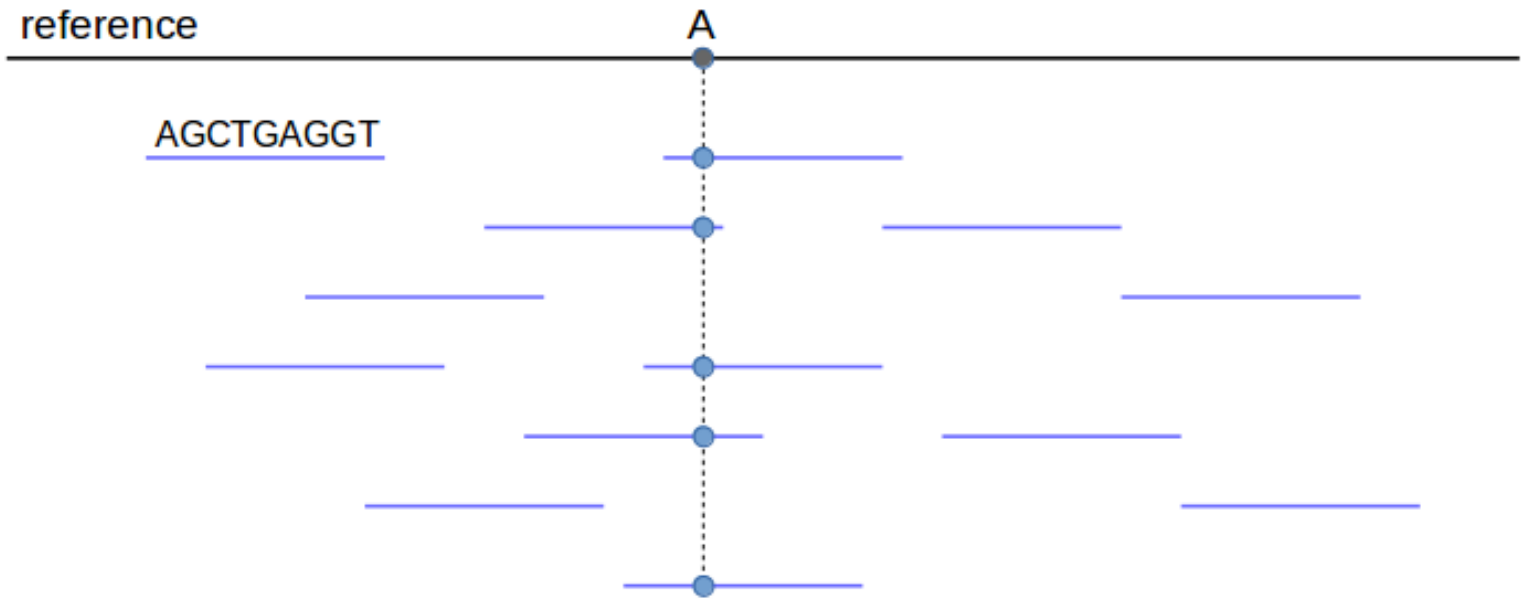



**Likelihood** approach:

- 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.

Imperial College London

# That is why we need statistics!

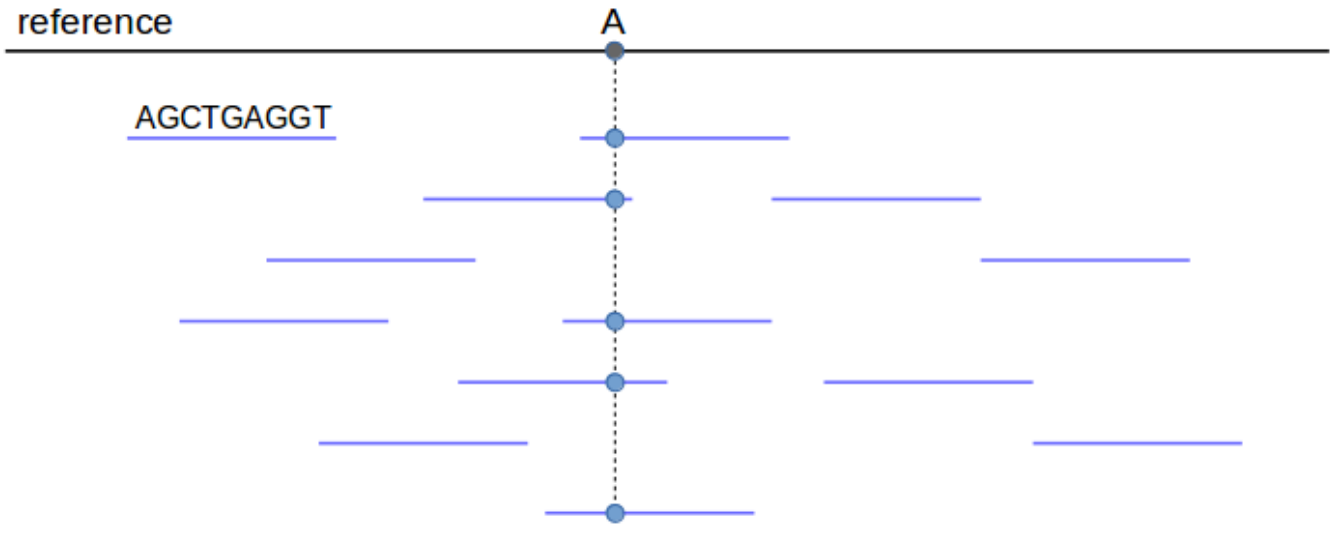



- is a **nucleotide**/base/allele with a certain **quality** score

# Genotype likelihoods

Imperial College London

Given a possible genotype, what is the probability of observing this NGS data?




- is a **nucleotide**/base/allele with a certain **quality** score

Imperial College London

# Genotype likelihoods - equation

## Likelihood

$P(D|G = \{A_1, A_2, ..., A_n\})$
with
$A_i \in \{A, C, G, T\}$ and $n$ being the ploidy level

Imperial College London

# Genotype likelihoods - equation

## Likelihood

$P(D|G = \{A_1, A_2, ..., A_n\})$
with
$A_i \in \{A, C, G, T\}$ and $n$ being the ploidy level

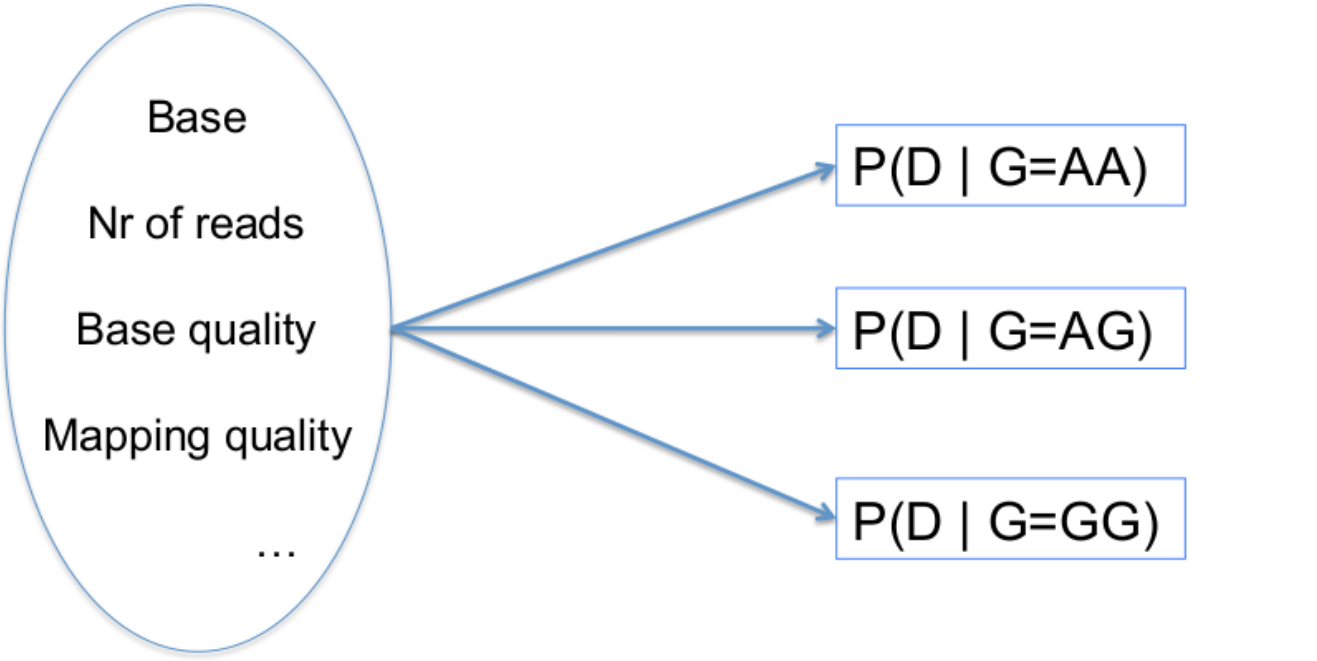
How many genotypes likelihoods do we need to calculate for each each diploid individual at each site?

Imperial College London

# Genotype likelihoods - rationale

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

Base quality in Phred scale



Imperial College London

# Genotype likelihoods - calculation

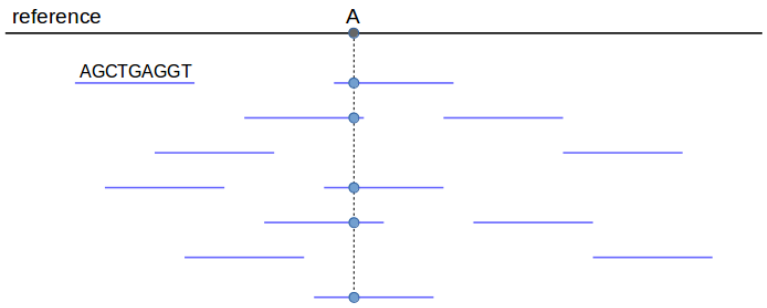
## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

- $L_{A_j,i} = P(D|A_G = A_j)$
- $A_i \in \{A, C, G, T\}$
- $R$ is the depth (nr. of reads)
- $N$ is the ploidy level (nr. of chromosomal copies)

Imperial College London

# Genotype likelihoods - example



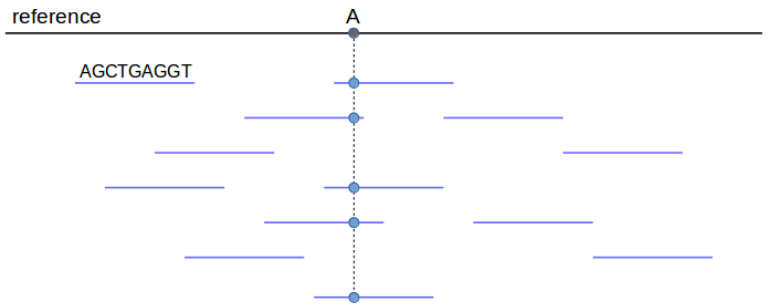

- is a **nucleotide**/base/allele with a certain **quality** score

A
A
A
G

with all with quality scores equal to 20 (in phred score)

Imperial College London

# Genotype likelihoods - example




- is a **nucleotide**/base/allele with a certain **quality** score

A
A
A
G

with all with quality scores equal to 20 (in phred score)

What is $P(D|G = AC) =?$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20

$$P(D|G = \{A, C\}) = ...$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G
& Q=20
$N = 2; i = 1; A_1 = A; A_2 = C$

$$P(D|G = \{A, C\}) = (\frac{L_{A,1}}{2} + \frac{L_{C,1}}{2}) \times ...$$

What are $L_{A,1}$ and $L_{C,1}$?

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} =$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} = \frac{\epsilon}{3}$$

$$L_{A,1} =$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,1} = \frac{\epsilon}{3}$$

$$L_{A,1} = 1 - \epsilon$$

$$P(D|G = \{A, C\}) = (\frac{1-\epsilon}{2} + \frac{\epsilon}{6}) \times ...$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} =$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} = \frac{\epsilon}{3}$$

$$L_{A,4} =$$

Imperial College London

# Genotype likelihoods - example

## Likelihood function

$$P(D|G = \{A_1, A_2, ..., A_N\}) = \prod_{i=1}^{R} \sum_{j=1}^{N} \frac{L_{A_j,i}}{N}$$

A
A
A
G

$$L_{C,4} = \frac{\epsilon}{3}$$

$$L_{A,4} = \frac{\epsilon}{3}$$

$$P(D|G = \{A, C\}) = (\frac{1-\epsilon}{2} + \frac{\epsilon}{6})^3 \times \frac{\epsilon}{3}$$

What is $\epsilon$?

Imperial College London

# Genotype likelihoods - example

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

A

A

A

G

$\epsilon = 0.01$

# Genotype calling

Imperial College London

# Genotype calling

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

AAAG & $\epsilon = 0.01$

What is the genotype here?

Imperial College London

# Genotype calling

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

AAAG & $\epsilon = 0.01$
What is the genotype?
AG.

## Maximum Likelihood

The simplest genotype caller: choose the genotype with the highest likelihood.

# Major and minor alleles

## Likelihood function

$$\log P(D|G = A) = \sum_{i=1}^{R} \log L_{A_j,i}$$

AAAG & $\epsilon = 0.01$

| Allele | Likelihood |
|---|---|
| **A** | **-2.49** |
| C | -3.38 |
| **G** | **-1.22** |
| T | -3.38 |

We can reduce the genotype space to 3 entries (from 10).

Imperial College London

# Genotype calling

AAAG & $\epsilon = 0.01$ & A,G alleles

| Genotype | Likelihood |
|---|---|
| AA | -5.73 |
| AG | -2.80 |
| GG | -17.12 |

Examples varying qualities and reads... open Julia script.

# Genotype likelihood ratio

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

i.e. $t = 1$ meaning that the most likely genotype is 10 times more likely than the second most likely one

Pros and cons?

- Yes:
- No:

# Genotype likelihood ratio

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

i.e. $t = 1$ meaning that the most likely genotype is 10 times more likely than the second most likely one

Pros and cons?

- Yes: genotype are called with higher **confidence**
- No: more **missing** data

Practical: genotype likelihoods and (basic) genotype calling
`https://github.com/mfumagalli/Copenhagen`

Imperial College London

# Statistical thinking

Figure 1: Nessie, the Loch Ness Monster. True or fake?

Imperial College London

# Statistical thinking

- $D = \{0, 1\}$, whether I tell you I saw Nessie or not.
- $N = \{0, 1\}$, whether Nessie exists or not.

## Questions

- What are $p(D = 1|N = 1)$ and $p(D = 1|N = 0)$?
- What is a Maximum Likelihood Estimate of $N$?

# Statistical thinking

Our inference on $N$, our parameter, is driven solely by our observations, given by our likelihood function.

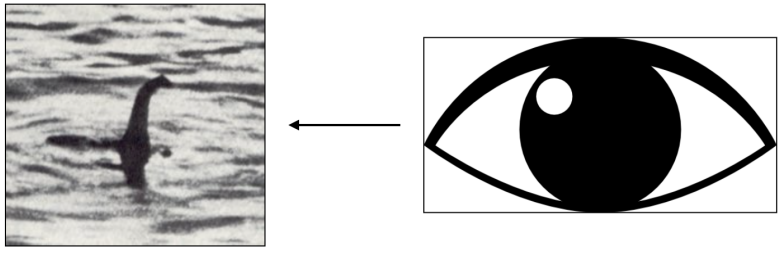


Figure 2: The eye: a "likelihood" organ.

# Statistical thinking

In real life we take many decisions based not only on what we observe but also on some believes of ours.

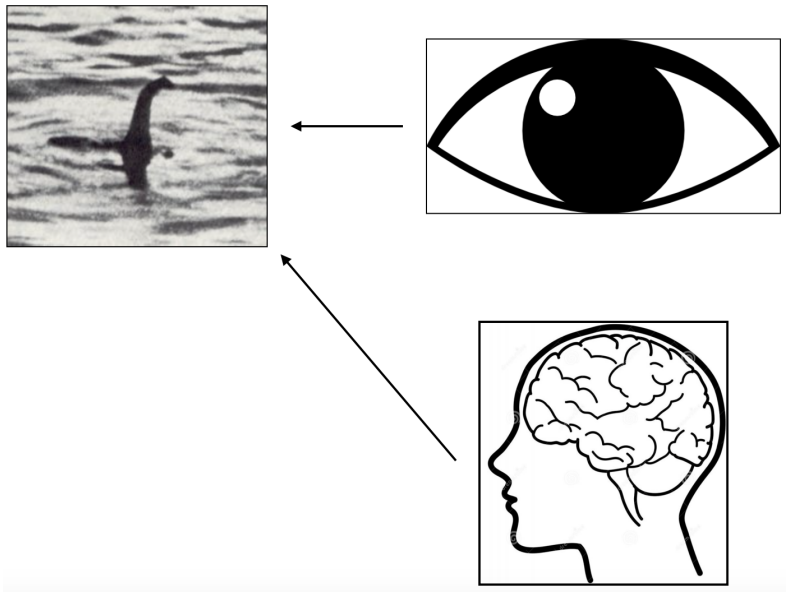


Figure 3: The brain: a "non-likelihood" organ.

# Bayesian thinking

- with "eyes only" our intuition is that $p(N|D) \approx p(D|N)$
- with "the brain" our intuition is that $p(N|D) \approx p(D|N)p(N)$

Our "belief" expresses the probability $p(N)$ **unconditional** of the data.

**Question**

How can we define $p(N)$?

# Bayesian thinking

The "belief" function $p(N)$ is called **prior probability** and the joint product of the likelihood $p(D|N)$ and the prior is proportional to the **posterior probability** $p(N|D)$.

The use of posterior probabilities for inferences is called Bayesian statistics.

# Statistical inference

If $D$ is the data and $\theta$ is your unknown parameter, then

- the frequentist conditions on parameters and integrates over the data, $p(D|\theta)$,
- the Bayesian conditions on the data and integrates over the parameters, $p(\theta|D)$.

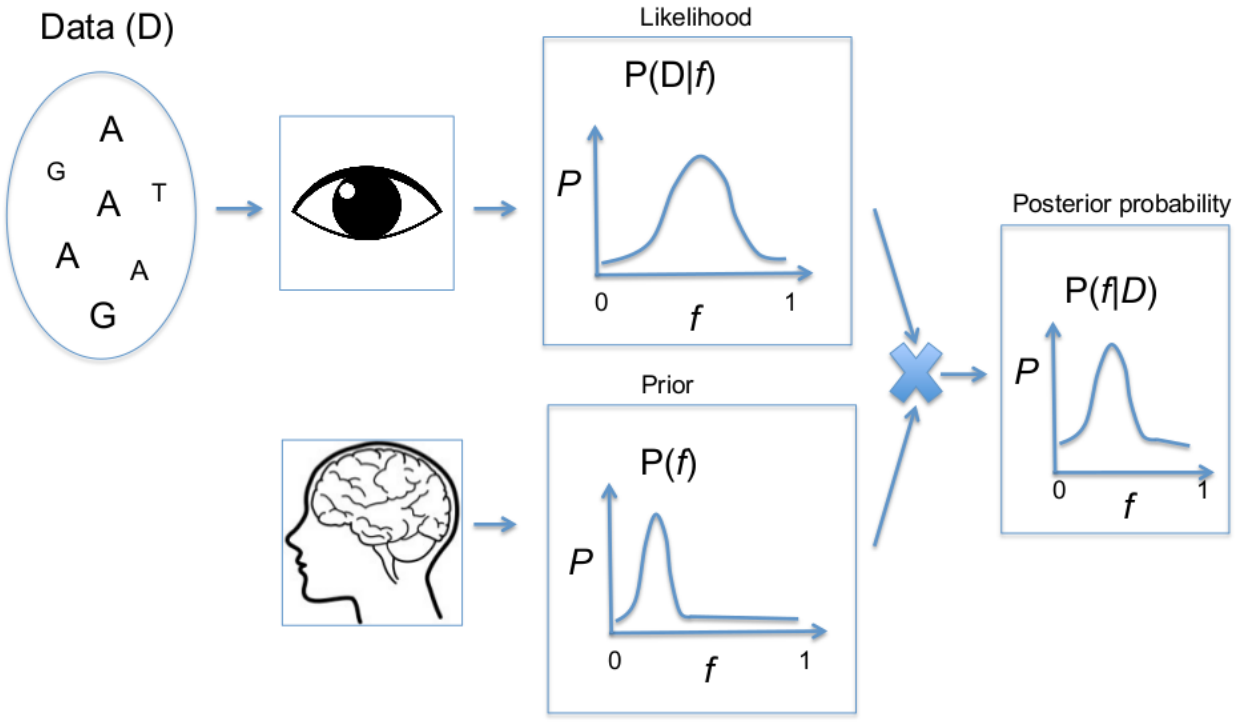
# Statistical inference

## Bayesian *vs.* Likelihoodist

- we derive "proper" probability distributions of our parameters rather than deriving a point estimate;
- a probability is assigned to a hypothesis rather than a hypothesis is tested;
- we can "accept" the null hypothesis rather than "fail to reject" it;
- parsimony imposed in model choice rather than correcting for multiple tests.

Imperial College London

# Bayesian inference



# Bayesian concepts

## Bayes' Theorem

$$p(\vec{\theta}|\vec{y}) = \frac{f(\vec{y}|\vec{\theta})\pi(\vec{\theta})}{m(\vec{y})} = \frac{f(\vec{y}|\vec{\theta})\pi(\vec{\theta})}{\int f(\vec{y}|\vec{\theta})\pi(\vec{\theta})d\vec{\theta}} \quad (1)$$

- $\vec{\theta}$ is not a fixed parameter but a random quantity with prior distribution $\pi(\vec{\theta})$
- $p(\vec{\theta}|\vec{y})$ is the posterior probability distribution of $\vec{\theta}$
- $\int p(\vec{\theta}|\vec{y})d\vec{\theta} = 1$

Imperial College London

# Genotype posterior probability

A

A

A

G

$\epsilon = 0.01$

A,G alleles

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | | |
| AG | -2.80 | | |
| GG | -17.12 | | |

Imperial College London

# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | 1/3 | 0.05 |
| AG | -2.80 | 1/3 | 0.95 |
| GG | -17.12 | 1/3 | 0 |

Only call genotypes if the largest probability is above a certain threshold (e.g. 0.95).

Imperial College London

# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & **A is the reference allele**
$P(AA) > P(AG) > P(GG)$

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | 0.80 | 0.22 |
| AG | -2.80 | 0.15 | 0.78 |
| GG | -17.12 | 0.05 | 0 |

The reference allele is just one of the possible alleles, often chosen arbitrarily: why give it so much weight?

Imperial College London

# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from a reference panel

$P(AA) =?$; $P(AG) =?$; $P(GG) =?$

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | | |
| AG | -2.80 | | |
| GG | -17.12 | | |

Imperial College London

# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from a reference panel

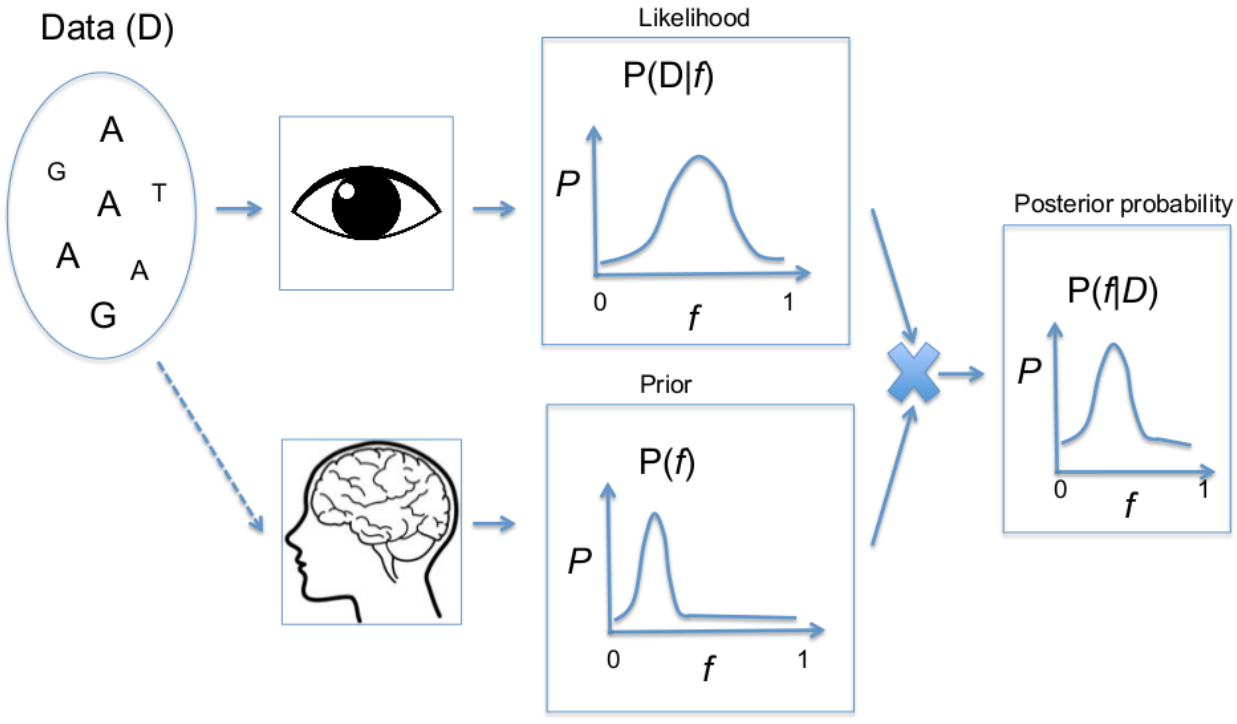
$P(AA) = ?$; $P(AG) = ?$; $P(GG) = ?$

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | 0.49 | 0.06 |
| AG | -2.80 | 0.42 | 0.94 |
| GG | -17.12 | 0.09 | 0 |

If the assumption of HWE can be reasonably met.

Imperial College London

# Empirical Bayesian inference



# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.6$ from the data itself

$P(AA) = ?$; $P(AG) = ?$; $P(GG) = ?$

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | 0.49 | 0.04 |
| AG | -2.80 | 0.42 | 0.96 |
| GG | -17.12 | 0.09 | 0 |

- if the assumption of HWE can be reasonably met
- if you have enough samples to have a robust estimate of the allele frequencies

Practical: genotype calling

`https://github.com/mfumagalli/Copenhagen`

Imperial College London

# Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.6$ from the data itself

| Genotype | Likelihood (log) | Prior | Posterior |
|---|---|---|---|
| AA | -5.73 | 0.49 | 0.04 |
| AG | -2.80 | 0.42 | 0.96 |
| GG | -17.12 | 0.09 | 0 |

- if the assumption of HWE can be reasonably met
- if you have enough samples to have a robust estimate of the allele frequencies

How can we estimate allele frequencies?

# SNP calling

Imperial College London

# Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

| Sample | True genotype | Reads allele A | Read 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 |

What is the simplest estimator of allele frequencies?

Imperial College London

# Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

| Sample | True genotype | Reads allele A | Read 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 |
| Total | | 41 | 14 |

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

$\hat{f} = 0.75$

What is wrong with this estimator?

Imperial College London

# Estimating allele frequencies

Assuming 2 alleles (A,G) with true allele frequency of 0.50

| Sample | True genotype | Reads allele A | Read 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 |
| Total | | 41 | 14 |

$$\hat{n_A} = \sum_{i=1}^{N} (1-\epsilon) n_{A,i} + \epsilon n_{G,i} - \epsilon n_{A,i} - (1-\epsilon) n_{G,i}$$

$\hat{f} = 0.77$

Imperial College London

# Estimating allele frequencies

## Maximum Likelihood estimator

$$P(D|f) = \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} P(D|G=g)P(G=g|f)$$

# Estimating allele frequencies

## Maximum Likelihood estimator

$$P(D|f) = \prod_{i=1}^{N} \sum_{g \in \{0,1,2\}} P(D|G=g)P(G=g|f)$$

$P(D|G=g)$ is the genotype likelihood and $P(G=g|f)$ is given by HWE (for instance).

In our previous example, $\hat{f} = 0.46$ which is much closer to the true value than previous estimators.

Imperial College London

# SNP calling

## Challenges

- If high levels of missing data, then genotypes can be lost.
- Rare variants are hard to detect.
- Trade off between false positive and false negative rates.

## How to call SNPs?

- If at least one heterozygous genotype has been called.
- If the estimated allele frequency is above a certain threshold.

Imperial College London

# SNP calling

Call a SNP if

$$\hat{f} \geq t$$

where $t$ can be the minimum sample allele frequency detectable (e.g. $t = 1/2N$ with $N$ diploids).

Imperial College London

# Likelihood Ratio Test

A Likelihood Ratio Test (LRT) compares the goodness of fit between the null and the alternative model:

- Null model: $f = 0$
- Alternative model: $f \neq 0$

$$T = -2 \log \frac{L(f = 0)}{L(f = \hat{f}_{MLE})}$$

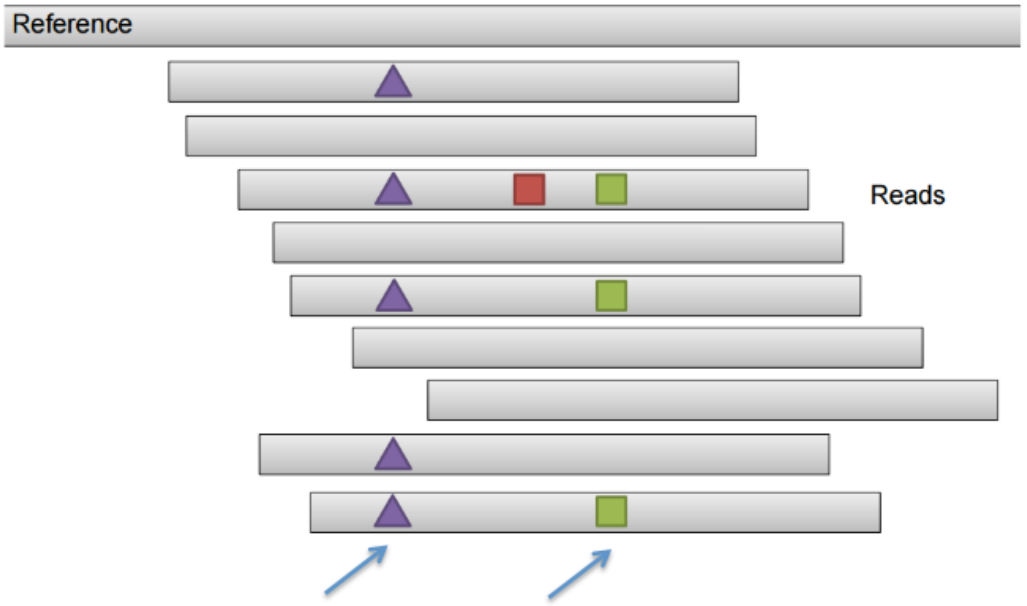
where $T$ is $\chi^2$ distributed with 1 degree of freedom.

Practical: allele frequencies and SNP calling
`https://github.com/mfumagalli/Copenhagen`

Imperial College London

# SNP calling procedures

- Alignment-based caller




We completely rely on how reads have been mapped

Figure from Erik Garrison

Imperial College London

# 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)

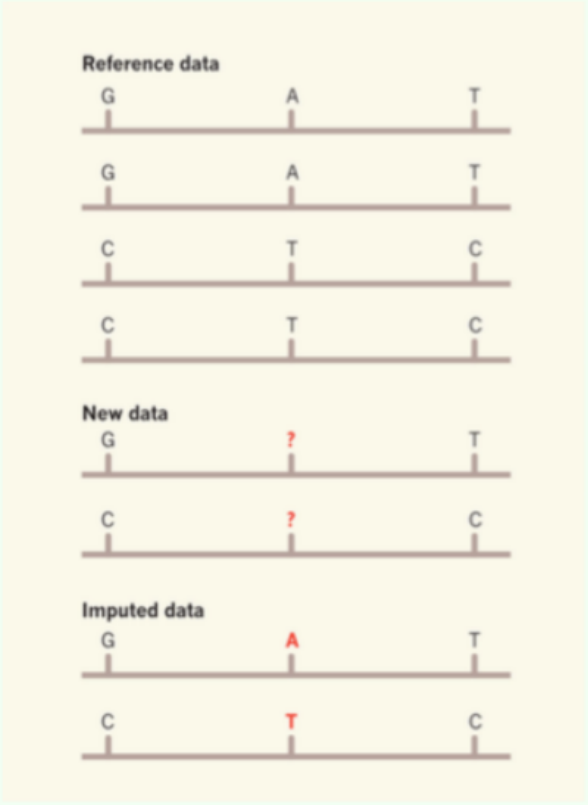

Figure from Erik Garrison

# Imputation

Imperial College London

# Haplotype imputation

## Haplotype imputation - simplified




**Reference**

- 1000 Genomes
- Phased using family structures

**new data**

- partial information

**Imputed data**

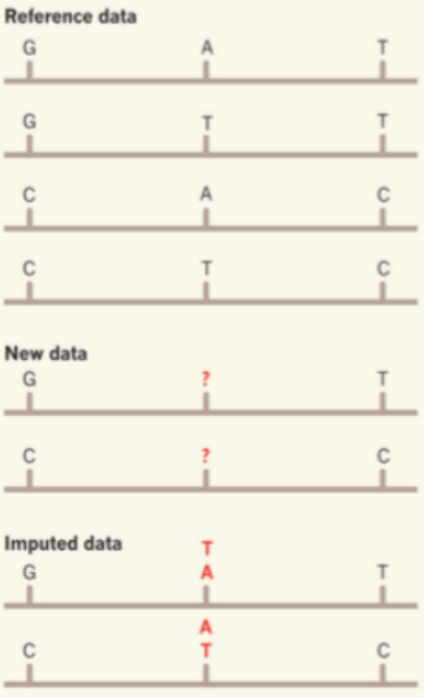
- Probabilistic approach
- The results retains the uncertainty of both the genotype and the haplotypes

Anders Albrechtsen

Imperial College London

# Haplotype imputation

## Haplotype imputation - simplified




**Reference**

- 1000 Genomes
- Phased using family structures

**new data**

- Data with known and unknown genotypes

**Imputed data**

$p(? = T) =$
$p(? = A) =$

Anders Albrechtsen

Imperial College London

# Haplotype imputation

## Haplotype imputation - simplified


**Reference**

- haplotype frequencies

**new data**

- Data with known and unknown genotypes

**first haplotype**

$$p(? = T) = \frac{0.56}{0.56+0.03} = 0.95$$
$$p(? = A) = \frac{0.03}{0.56+0.03} = 0.05$$

**second haplotype**

$$p(? = T) = \frac{0.21}{0.21+0.2} = 0.51$$
$$p(? = A) = \frac{0.2}{0.21+0.2} = 0.49$$

Anders Albrechtsen

Imperial College London

# Haplotype imputation

## Haplotype imputation - simplified


**Bayes formula**

$$p(H = h|f, G) = \frac{P(G|H=h)P(H=h|f)}{\sum_{h'} P(G|H=h')P(H=h'|f)}$$

**$P(G|H = h)$**

1 if consistent

0 otherwise

**first haplotype**

$$p(? = T) = \frac{0.56}{0.56+0.03} = 0.95$$
$$p(? = A) = \frac{0.03}{0.56+0.03} = 0.05$$

Anders Albrechtsen