Imperial College London

Introduction to genotype likelihoods and calling

Matteo Fumagalli

Intended Learning Outcomes

By the end of this session you will be able to

- understand the theory underpinning genotype calling
- calculate genotype likelihoods
- appreciate the need to avoid genotype calling for low-depth data
- implement a pipeline in ANGSD to perform the aforementioned analyses

From raw data to 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  ,.$.....,,,.,.,,.,,,.,..."+. <<<+;<<<<<<<<<<<<<<<<<<<j?<&
seq1 273 T 23  ,.....,,.,,,.,.,,.,..A <<<;<<<<<<<<<<<3<<<<<;<<<
seq1 274 T 23  ,.$....,,.,.,,...,,,.,..... 7<7;<;<<<<<<<<<<<;<;7<<6
seq1 275 A 23  ,$.....,,.,.,,...,,,.,.....^l. <+;9*<<<<<<<<<<<<;;<<<<
seq1 276 G 22  ...T,,,.,.,,...,,,.,.... 33;+<<7*<7<&<<1;<<6<
seq1 277 T 22  ......,,.,.,,.C,,,.,..G. +7<;<<<<<<&<<=<;<<:<<
seq1 278 G 23  .....,,.,.,,...,,,.,..."k. %38*<<;<7<<7<=<<<j;<<<<<
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
```
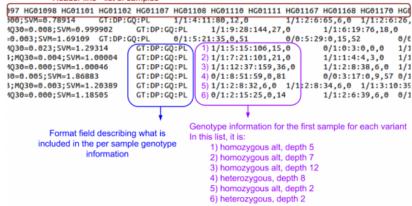
How about?

Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

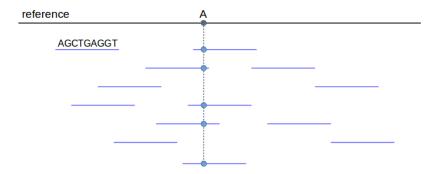perhaps not the best idea ever... but is file size the only issue?

Header line - list of samples

997 HG01098 HG01101 HG01102 HG01107 HG01108 HG01110 HG01111 HG01167 HG01168 HG01170 HG0

```
900;SVM=0.78914        GT:DP:GQ:PL     1/1:4:11:80,12,0        1/1:2:6:65,6,0  1/1:2:6:26,
4Q30=0.008;SVM=0.999902      GT:DP:GQ:PL     1/1:9:28:144,27,0      1/1:6:19:76,18,0
=0.003;SVM=1.69109  GT:DP:GQ:PL      0/1:5:21:35,0,51      0/0:5:29:0,15,52       0/6
4Q30=0.023;SVM=1.29314   GT:DP:GQ:PL  1) 1/1:5:15:106,15,0      0/1:0:3:0,0,0     1/1
4;MQ30=0.004;SVM=1.00004  GT:DP:GQ:PL  2) 1/1:7:21:101,21,0      1/1:1:4:4,3,0     1/1
4Q30=0.000;SVM=1.00046   GT:DP:GQ:PL  3) 1/1:12:37:159,36,0     1/1:2:8:38,6,0    1/1
30=0.005;SVM=1.86883     GT:DP:GQ:PL  4) 0/1:8:51:59,0,81       0/0:3:17:0,9,57 0/1
3;MQ30=0.003;SVM=1.20389  GT:DP:GQ:PL  5) 1/1:2:8:32,6,0  1/1:2:8:34,6,0  1/1:3:10:39
4Q30=0.000;SVM=1.18505   GT:DP:GQ:PL  6) 0/1:2:15:25,0,14       1/1:2:6:39,6,0   0/1
```

Format field describing what is included in the per sample genotype information

Genotype information for the first sample for each variant
In this list, it is:
1) homozygous alt, depth 5
2) homozygous alt, depth 7
3) homozygous alt, depth 12
4) heterozygous, depth 8
5) homozygous alt, depth 2
6) heterozygous, depth 2

genome.sph.umich.edu

what else do we need?

Why do we need statistics?

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

Low-level data

```
FASTQ
a'X_\Va\J'KaYJHG^]b\a^BBBBBBBBBBBB   <-- quality score
@FC42BF1AAXX:6:1:5:732#0/1           <-- read ID
TGATTCTCTCGATATCCAGTCCTTAGTGNCATAGN <-- read (bases)
+
a^_aaaa'aa'_aaa_aaa'__''_'VBBBBBBBB
@FC42BF1AAXX:6:1:5:492#0/1
AACAGTGGGAGGCTGCAGCAGGAGGATTNCTGAAN
+
ababb_abbbZbabaab^'aaTaabbaBBBBBBBB
@FC42BF1AAXX:6:1:5:480#0/1
ACCTCCTCAGAGTTCTCGAGCTCGAGAANTCTGGN
```

Quality scores

Qscore

- The ASCII values can be interpreted as a probability
- A Q20 (ASCII 'T') score is probability of 1%
- The score is the probability, P, that the base is incorrect

$$Q_{score} = -10log_{10}(P)$$

$$P = 10^{\frac{-Q}{10}}$$

Quality scores

The qscores are encoded as ASCII characters, and are shifted by +33 (now the standard) or +64.

Phred Quality Score	Probability of error	Base call accuracy
0 … 9	1 … 0.13	!"#$%&'()*
10 … 19	0.1 … 0.013	+,-./01234
20 … 29	0.01 … 0.0013	56789:;<=>
30 … 39	0.001 … 0.00013	?@ABCDEFGH
40 … 49	0.0001 … 0.000014	IJKLMNOPQR

Quality scores

> ### Example
>
> From a fastq file we observe an 'A' with a qscore encoded as '7'. What is the probability of being 'A'? What is the probability of geing 'G'?

1. We find that the corresponding ASCII value of '7' is ?
   (hint: `http://www.asciitable.com`)

Quality scores

Example

From a fastq file we observe an 'A' with a qscore encoded as '7'. What is the probability of being 'A'? What is the probability of geing 'G'?
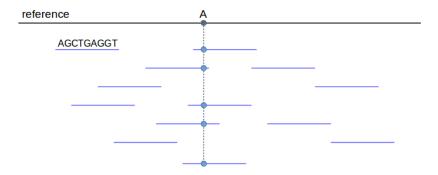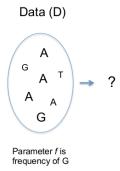
1. We find that the corresponding ASCII value of '7' is ? (hint: `http://www.asciitable.com`)
2. We substract 33 to get a value of ?. This is our qscore.
3. The probability of 'A' being incorrect is ? (hint: $p = 10^{\frac{-Q}{10}}$)
4. The probability of 'A' being correct is ?
5. The probability of being 'G' (or 'C' or 'T') is ?

Why do we need statistics?

reference   A

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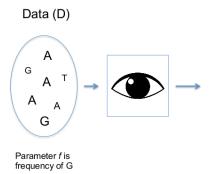
- is a **nucleotide**/base/allele with a certain **quality** score

Statistics to the rescue!
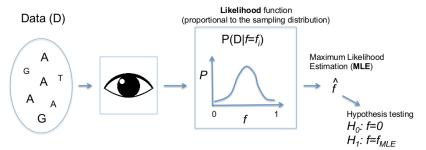
Data (D)

A

G

A   T

A

A

G

Parameter *f* is
frequency of G

$\rightarrow$?

Statistics to the rescue!

Data (D)

Parameter *f* is
frequency of G

Statistics to the rescue!

Data (D)

Likelihood function
(proportional to the sampling distribution)

$P(D|f=f_i)$

A
G
A T
A
A A
G

P

0   f 1

Parameter f is
frequency of G

Statistics to the rescue!

Data (D)

Likelihood function
(proportional to the sampling distribution)

$P(D|f=f_i)$

Maximum Likelihood
Estimation (**MLE**)

$\hat{f}$

Hypothesis testing
$H_0: f=0$
$H_1: f=f_{MLE}$
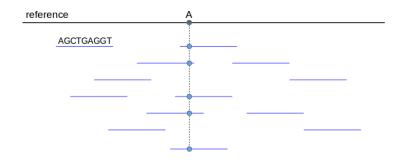
Likelihood approach:

- 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.

Let's move to the whiteboard!

Introduction to likelihoods: the case of GC-content.

That is why we need statistics and we should not be afraid of it*.

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

* (but statisticians are scary, with all their formulas and notations and expectations and variances... and who cares about p-values anyway?)

Let's move to the whiteboard!

Diagram: what is the relationship between data and parameter?
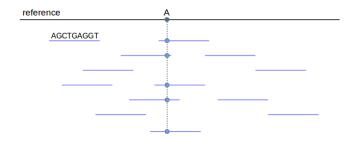
Genotype likelihoods

what are these genotype "quality" and what do they mean?

Genotype likelihoods


```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      Sample1        Sample2        Sample3
2      4370    rs6057    G   A      29   .      NS=2;DP=13;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.
2      7330    .         T   A      3    q10    NS=5;DP=12;AF=0.017               GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3   0/0:41:3
2      110696  rs6055    A   G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
2      130237  .         T   .      47   .      NS=2;DP=16;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2      134567  microsat1 GTCT G,GTACT 50  PASS   NS=2;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
chr1   45796269 .         .   G      .
chr1   45797505 .         .   C      G
chr1   45798555 .         .   T      C
chr1   45798901 .         .   C      T
chr1   45805566 .         .   G      C
```

what are these genotype "quality" and what do they mean?
Challenge: by the end of this session you will be able to calculate
these genotype likelihoods by hand!

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

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

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

Genotype likelihoods - equation

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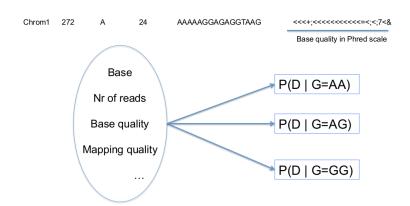
$P(D|G = \{A_1, A_2, ..., A_n\})$
with
$A_i \in \{A, C, G, T\}$ and n being the ploidy level

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

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)

Genotype likelihoods - example

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

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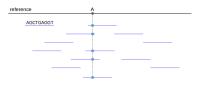
with all with quality scores equal to 20 (in phred score)

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

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

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

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 $L_{C,1}$?

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

$$L_{C,1} =$$

Genotype likelihoods - example

Likelihood function

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

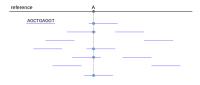A
A
A
G

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

$$L_{A,1} =$$

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

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

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

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

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

$$L_{C,4} =$$

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

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

$$L_{A,4} =$$

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 ϵ?

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
$\epsilon = 0.01$

why in log-scale?

Practical

Introduction to ANGSD and genotype likelihoods
`https://github.com/nt246/physalia-lcwgs`

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?

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.

PS: You did it by hand!

Major and minor alleles

Likelihood function

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

Where is the product? Why the sum now?
AAAG & $\epsilon = 0.01$

Major and minor alleles

Likelihood function

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

Where is the product? Why the sum now?

AAAG & $\epsilon = 0.01$

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

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

Genotype calling

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

Genotype	Likelihood
AA	-5.73
AG	-2.80
GG	-17.12

At what extent is the data affecting the called genotype and its **confidence**?

Let me open a jupyter notebook to illustrate the effect of data uncertainty on genotype calling.

Genotype likelihood ratio for calling

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

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 likelihood ratio for calling

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

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:

Genotype likelihood ratio for calling

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

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: gentle introduction to ANGSD for genotype likelihoods and (basic) genotype calling.

The monster dilemma

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

Let's move to the whiteboard.

"Eyes" thinking

What's "wrong"?
Our inference on θ, our parameter, is driven solely by our observations, given by our likelihood function.
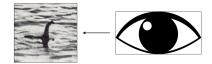
Figure 2: The eye: a "likelihood" organ.

"Blind Brain" thinking

In real life we take many decisions based not only on what we observe but also on some believes of ours*.
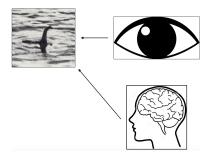
Figure 3: The brain: a "non-likelihood" organ.

* unfortunately in many cases

Eyes+Brain thinking

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

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

Question

How can we define $p(\theta)$?

"Eyes + Blind Brain" thinking
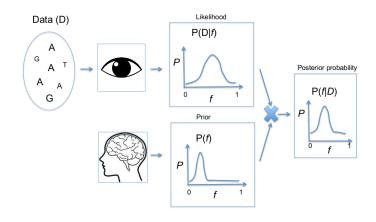
The "belief" function $p(\theta)$ is called **prior probability** and the joint product of the likelihood $p(D|\theta)$ and the prior is proportional to the **posterior probability** $p(N|\theta)$.

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

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.

Bayesian inference

Bayes' Theorem

$$p(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}$$

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

Imperial College London

Genotype posterior probability

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

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73		

Genotype posterior probability

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

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	1/3	

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		

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		

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

What is the called genotype? What's its confidence?

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

What is the called genotype? What's its confidence?
Only call genotypes if the largest probability is above a certain
threshold (e.g. 0.95).

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		

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		

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		

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

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

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		

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		

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		

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 Hardy Weinberg Equilibrium can be reasonably met.
What happens if that's not the case?

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from a reference panel
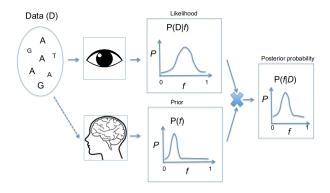
$P(AA) =?$
$P(AG) =?$
$P(GG) =?$

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 Hardy Weinberg Equilibrium can be reasonably met.

What happens if that's not the case?

Inbreeding can be incorporated: $f_{AA} = (1-f)^2 + (1-f)fF$ ...

"Eyes + non-Blind Brain" inference

Empirical Bayesian

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ from the data itself

Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73		

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ 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(+-F) can be met (no population structure)
- if enough samples to have a robust estimate of the allele frequencies

Practical: (advanced) genotype calling with ANGSD

Genotype posterior probability

AAAG & $\epsilon = 0.01$ & A,G alleles & $f(A) = 0.7$ 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

How can we estimate allele frequencies from NGS data?

Intended Learning Outcomes

At the end of this session you are now able to

- understand the theory underpinning genotype calling
- calculate genotype likelihoods
- appreciate the need to avoid genotype calling for low-depth data
- implement a pipeline in ANGSD to perform the aforementioned analyses