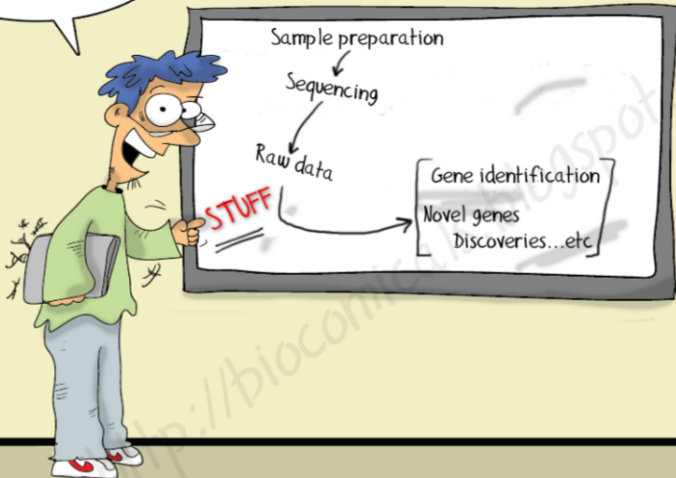


Genotype and SNP calling and estimation of allele frequencies

Matteo Fumagalli

We are
bioinformaticians
thats what we do



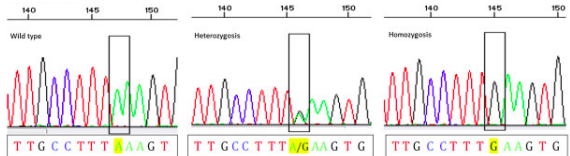
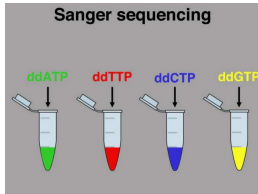
Presentation outline

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

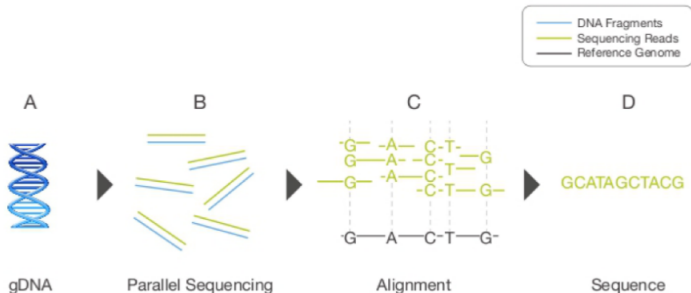
Motivation

Sanger sequencing

aka first/former generation sequencing



Next Generation Sequencing



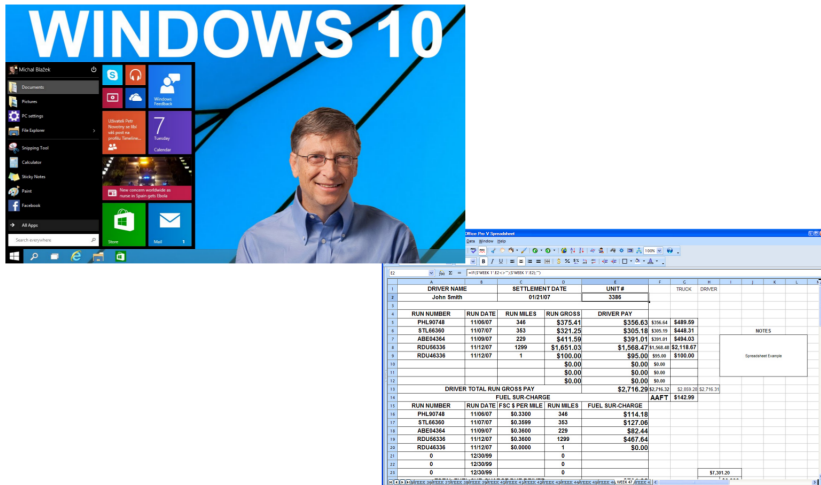
- Extracted gDNA
- gDNA is fragmented into a library of small segments that are each sequenced in parallel.
- Individual sequence reads are reassembled by aligning to a reference genome
- The whole-genome sequence is derived from the consensus of aligned reads.

>ARPM2ref|NC_000001.10|:2938046-2939467 Homo sapiens chromosome 1, GRCh37 primary reference assembly

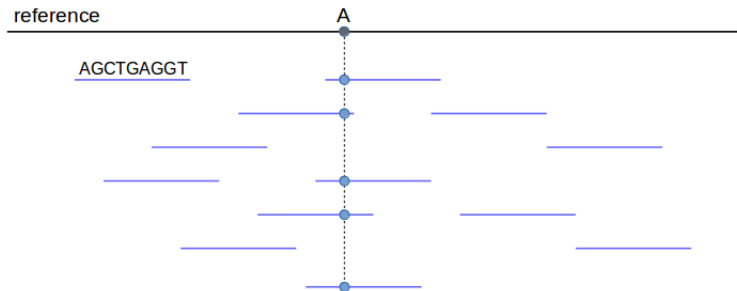
```
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B=-BABA@:@693:GB=
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGGAACATAAATTGTGTCAAATGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
```

[illegible]

```
##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 T . . . . GT
/r0
chr1 752566 rs3094315 G A . . . . GT
/r1
chr1 752721 rs3131972 A G . . . . GT
/r1
chr1 798959 rs11240777 G . . . . . GT
/r0
chr1 800007 rs6681049 T C . . . . GT
/r1
```



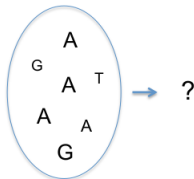
Why do we need statistics?



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

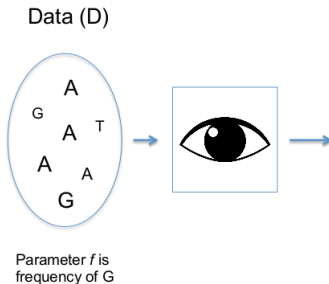
Statistical inference

Data (D)

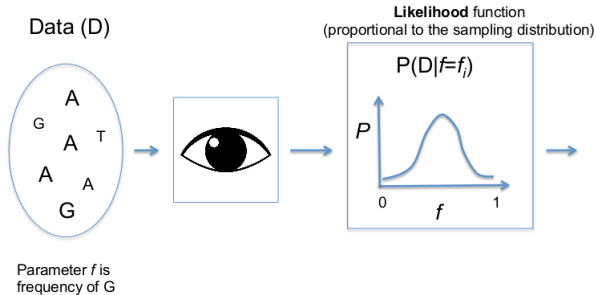


Parameter f is
frequency of G

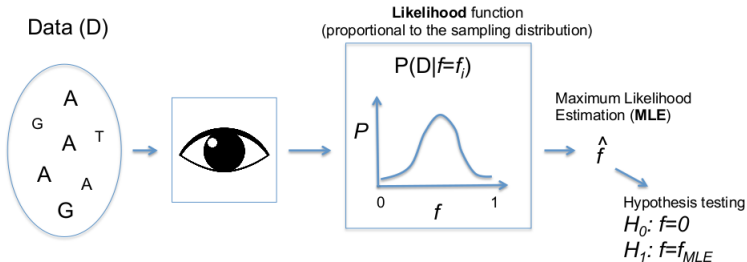
Statistical inference



Statistical inference



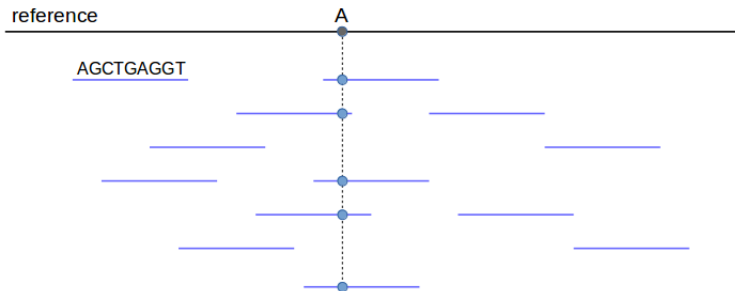
Statistical inference



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.

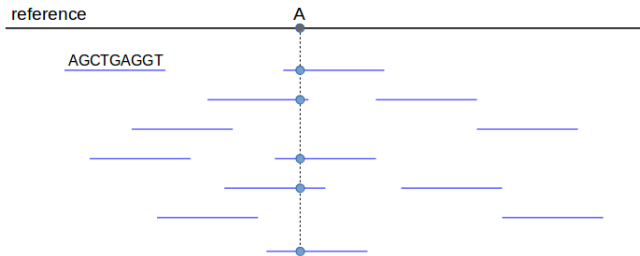
That is why we need statistics!



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

Genotype likelihoods

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

$$P(D|G = \{A_1, A_2, \dots, A_n\})$$

with

$A_i \in \{A, C, G, T\}$ and n being the ploidy level

Genotype likelihoods - equation

Likelihood

$$P(D|G = \{A_1, A_2, \dots, A_n\})$$

with

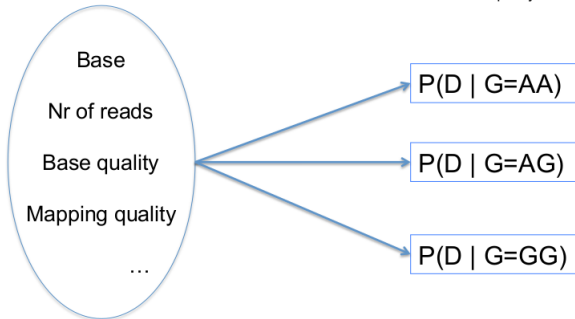
$A_i \in \{A, C, G, T\}$ and n being the ploidy level

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

Genotype likelihoods - rationale

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

Base quality in Phred scale



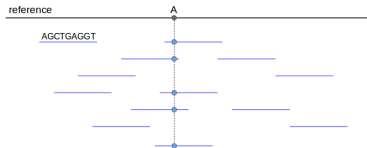
Genotype likelihoods - calculation

Likelihood function

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

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

Genotype likelihoods - example

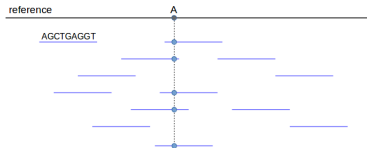


• is a **nucleotide/base/allele** with a certain **quality** score

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
A

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

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

Genotype likelihoods - example

Likelihood function

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

A
A
A
G
& Q=20

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

Genotype likelihoods - example

Likelihood function

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

A

A

A

G

& Q=20

$N = 2; i = 1; A_1 = A; A_2 = C$

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

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

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

$$L_{C,1} =$$

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

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

$$L_{A,1} =$$

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

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

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

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

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

$$L_{C,4} =$$

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

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

$$L_{A,4} =$$

Genotype likelihoods - example

Likelihood function

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

A
A
A
G

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

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

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

What is ϵ ?

Genotype likelihoods - example

Genotype	Likelihood (log10)	
AA	-2.49	
AC	-3.38	
AG	-1.22	A
AT	-3.38	A
CC	-9.91	A
CG	-7.74	G
CT	-9.91	$\epsilon = 0.01$
GG	-7.44	
GT	-7.74	
TT	-9.91	

Genotype calling

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.

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, 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**?

Open jupyter-notebook.

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 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:

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>

The monster dilemma



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

The monster dilemma - likelihood

Let's denote D (data) as the set of observations specifying whether I tell you that I saw Nessie ($D = 1$) or not ($D = 0$).

D is our sample space, the set of all possible outcomes of the experiment, and $D = \{0, 1\}$.

We want to make some inferences on the probability that Nessie exists, or that it is true that I saw it (her?). Let's denote this probability as N .

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

The monster dilemma - likelihood

Questions

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

The monster dilemma - likelihood

Let's assume that $p(D = 1|N = 0) = 0.01$ and $p(D = 1|N = 1) = 0.90$ are valid for each observer I , with $I = 3$.

Then the log-likelihood of $N = 0$ is given by

$\sum_{I=1}^3 \log(p(D = 1|N = 0)) = -6.91$ while the log-likelihood of $N = 1$ is given by $\sum_{I=1}^3 \log(p(D = 1|N = 1)) = -0.32$.

With 3 observations of $D = 1$ we obtained a likelihood ratio (LR, of $N = 1$ vs $N = 0$) of 6.59.

Does the Loch ness monster exist?

"Eyes" thinking

What's "wrong"?

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

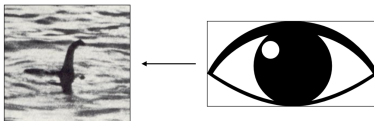


Figure 2: The eye: a "likelihood" organ.

"Blind Brain" thinking

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

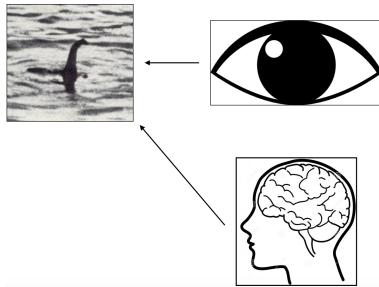


Figure 3: The brain: a "non-likelihood" organ.

* unfortunately in many cases

Eyes+Brain thinking

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

"Eyes + Blind Brain" thinking

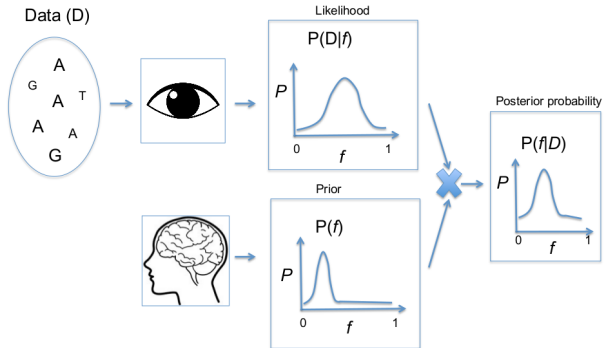
The "belief" function $p(N)$ is called **prior probability** and the joint product of the likelihood $p(D|N)$ and the prior is proportional to the **posterior probability** $p(N|D)$.

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

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

$$p(G|D) = \frac{f(D|G)\pi(G)}{\int f(D|G)\pi(G)dG}$$

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

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

$$P(AA) = ?$$

$$P(AG) = ?$$

$$P(GG) = ?$$

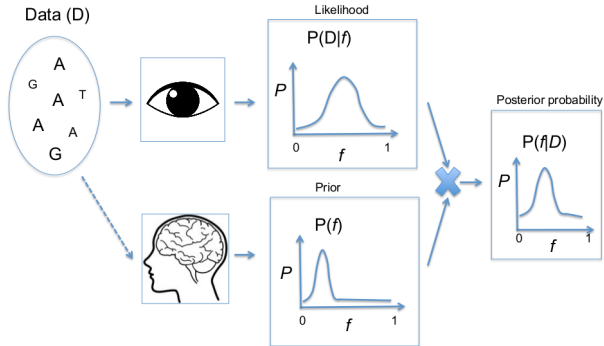
Genotype	Likelihood (log)	Prior	Posterior
AA	-5.73	0.49	0.06
AG	-2.80	0.42	0.94
GG	-17.12	0.09	0

If the assumption of Hardy Weinberg Equilibrium can be reasonably met.

What happens if that's not the case?

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

"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

<https://github.com/mfumagalli/Copenhagen>

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?

SNP calling

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?

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

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

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

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?

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

$$\hat{n}_A = \sum_{i=1}^N (1 - \epsilon)n_{A,i} + \epsilon n_{G,i} - \epsilon n_{A,i} - (1 - \epsilon)n_{G,i} = 0.77$$

Estimating allele frequencies

Maximum Likelihood estimator

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

with N samples.

What are $P(D|G = g)$ and $P(G = g|f)$?

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.

SNP calling (for low-coverage NGS data)

Challenges

SNP calling (for low-coverage NGS data)

Challenges

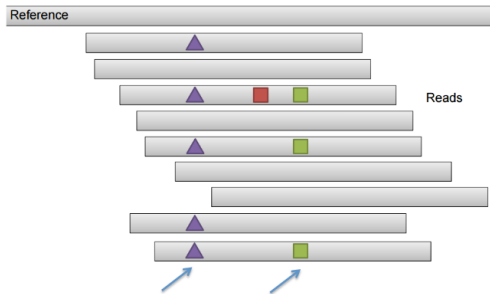
- If high levels of missing data, then genotypes can be lost.
- Rare variants are hard to detect.
- Trade off between false positive and false negative rates.

How to call SNPs (traditionally)?

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

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)

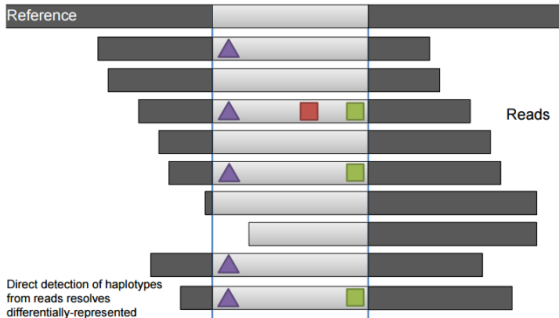


Figure from Erik Garrison

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).

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: $f \neq 0$

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 χ^2 distributed with 1 degree of freedom.

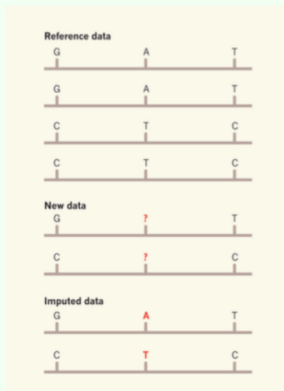
Practical: allele frequencies and SNP calling

<https://github.com/mfumagalli/Copenhagen>

Imputation

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

Haplotype imputation

Haplotype imputation - simplified

Reference data



New data



Imputed data



Reference

- 1000 Genomes
- Phased using family structures

new data

- Data with known and unknown genotypes

Imputed data

$$p(? = T) =$$

$$p(? = A) =$$

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

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

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

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