Imperial College
London

# *Introduction to allele frequency estimation and SNP calling*

Matteo Fumagalli

# Intended Learning Outcomes

By the end of this session you will be able to

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

# 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?

# 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 $P(G=g|f)$?

Imperial College London

# Estimating allele frequencies

### Maximum Likelihood estimator

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

$P(D|G=g)$ 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 (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

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: $f \neq 0$

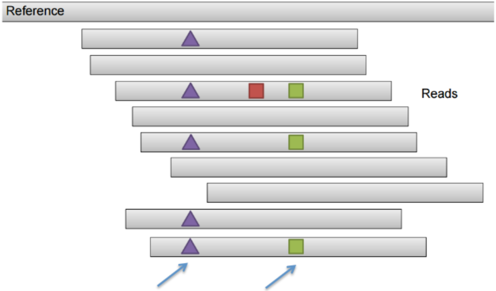
$$T = -2 \log \frac{L(f = 0)}{L(f = \hat{f}_{MLE})}$$

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

Practical: allele frequency estimation and SNP calling in ANGSD

# 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)

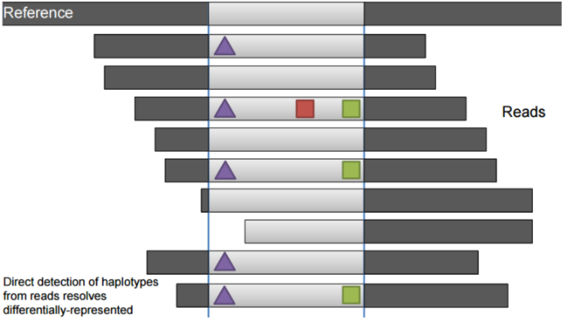
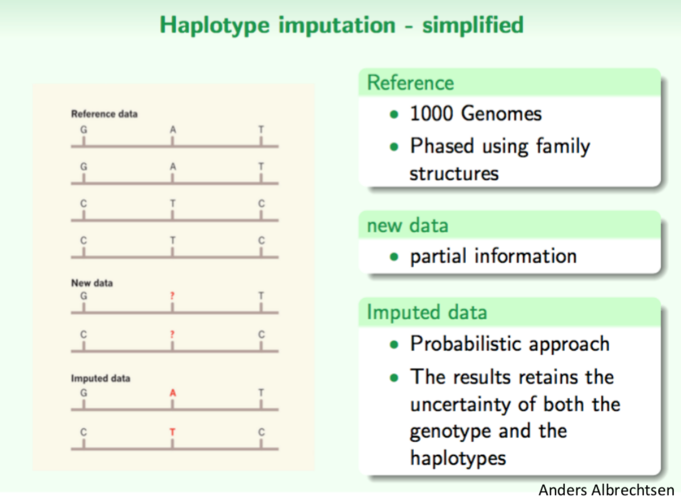


Figure from Erik Garrison

# Haplotype imputation



## Haplotype imputation - simplified

**Reference**

- 1000 Genomes
- Phased using family structures

**new data**

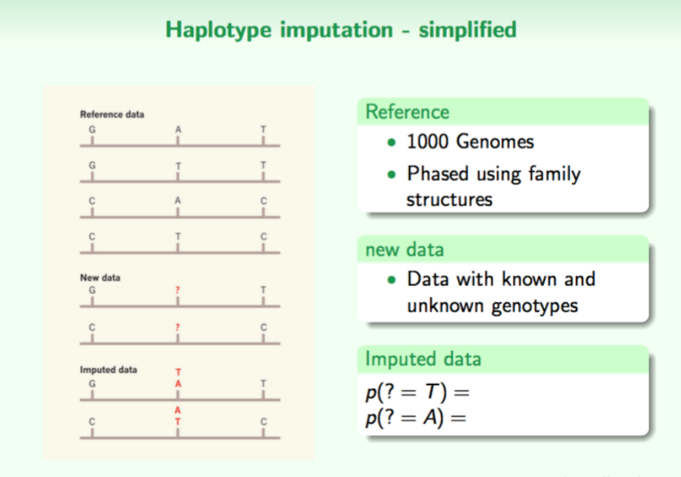
- partial information

**Imputed data**

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

Anders Albrechtsen

# Haplotype imputation



## Haplotype imputation - simplified

**Reference**

- 1000 Genomes
- Phased using family structures

**new data**

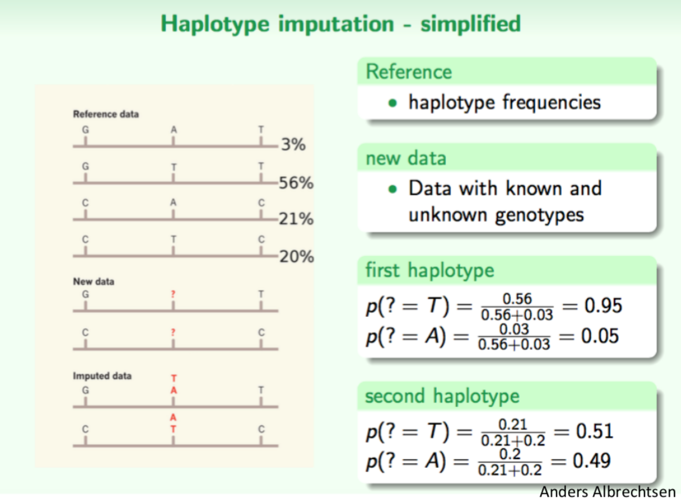
- Data with known and unknown genotypes

**Imputed data**

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

Anders Albrechtsen

# Haplotype imputation



## Haplotype imputation - simplified

**Reference**

- haplotype frequencies

**new data**

- Data with known and unknown genotypes

**first haplotype**

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

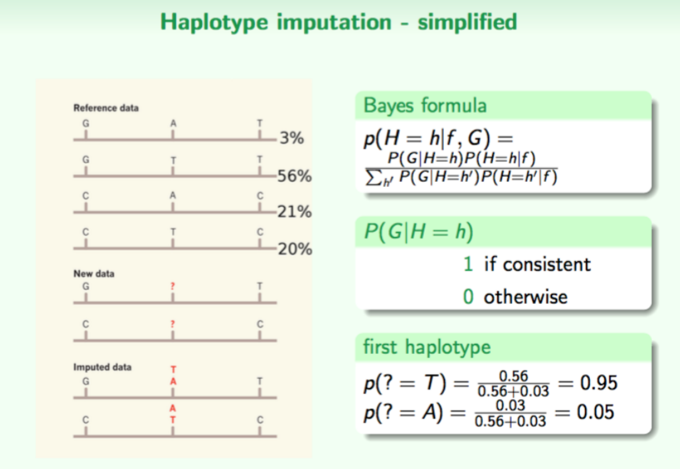
**second haplotype**

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

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

# Intended Learning Outcomes

At the end of this session you are now be able to

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