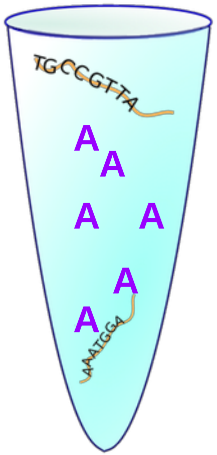
Genotype likelihoods, allele frequencies, and SNP calling from NGS data

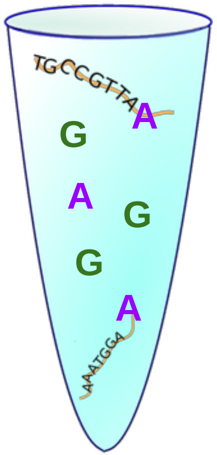
Tyler Linderoth
Physalia lcWGS course 2025

Why Probabilistic Methods?

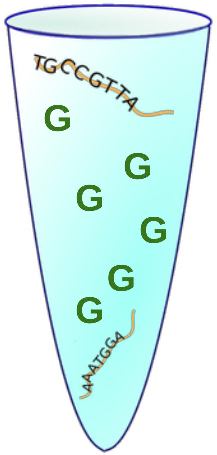
Why Probabilistic Methods?



The library for an individual homozygous for the A allele will consist only of As.

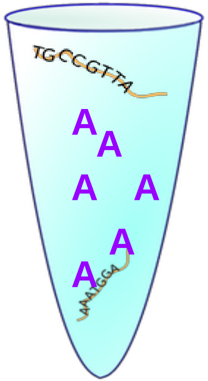


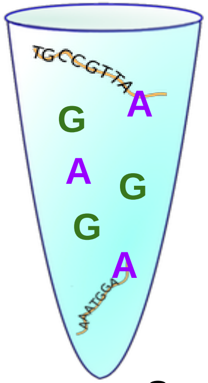
The library for a heterozygous individual at a site contains both As and Gs.

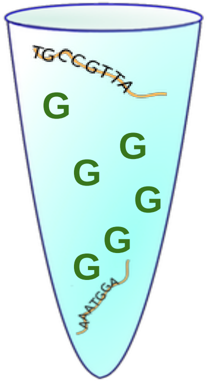


The library for an individual homozygous for the G allele will consist only of Gs.

Why Probabilistic Methods?







Sequence to average depth of 4x.

$$\text{Depth} \sim \text{Poisson}(\lambda=4)$$

$$E[\text{depth}] = \lambda$$

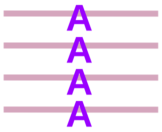
$$\text{Var}[\text{depth}] = \lambda$$

Sequencing (sampling) the two different alleles is just like flipping a coin.

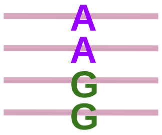
$$\#\,\text{A alleles} \sim \text{Binomial}(n\,reads, p=0.5)$$

$$E[\text{A depth}] = np = 0.5n$$

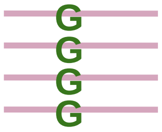
$$\text{Var}[\text{A depth}] = np(1-p) = 0.25n$$



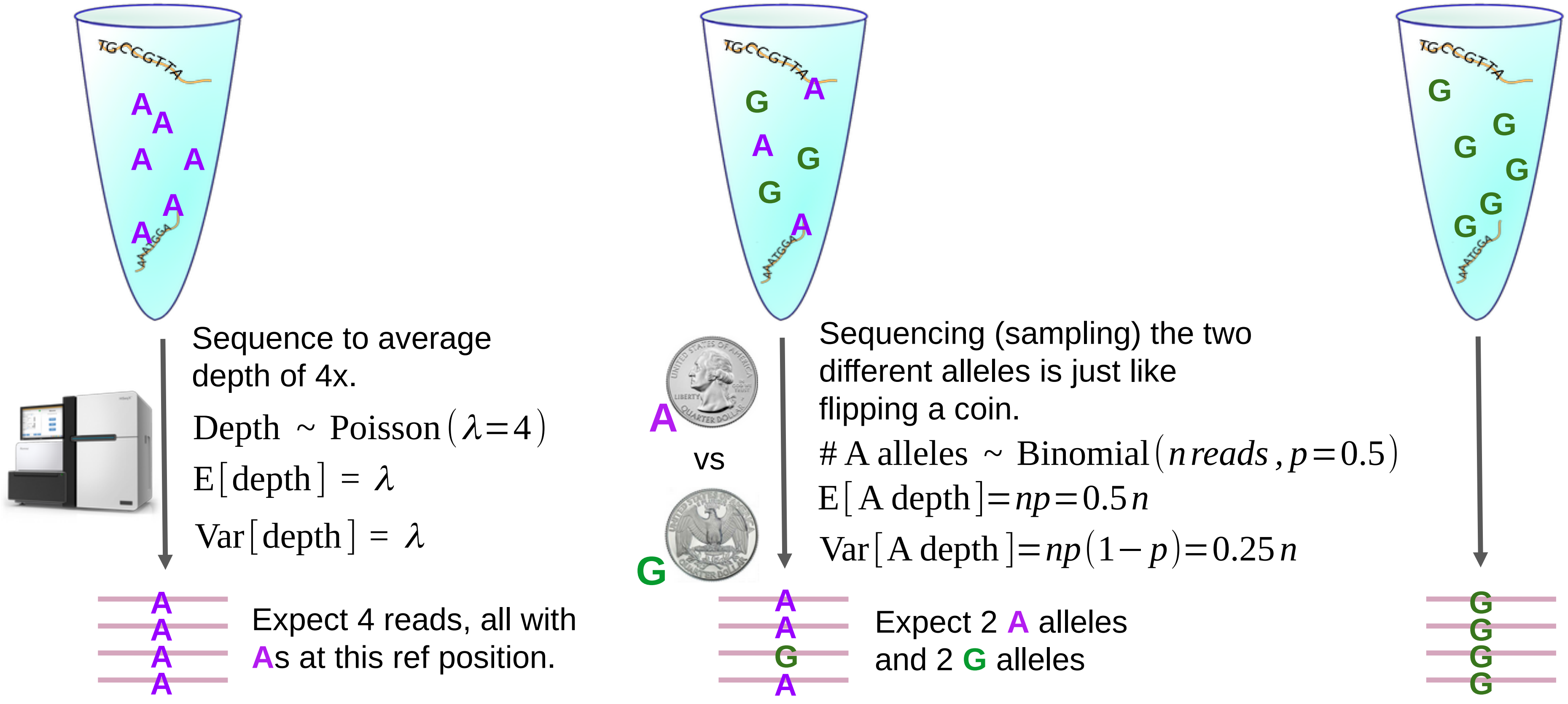
Expect 4 reads, all with As at this ref position.



Expect 2 A alleles and 2 G alleles



Why Probabilistic Methods?



Sequence to average depth of 4x.

$$\text{Depth} \sim \text{Poisson}(\lambda = 4)$$

$$E[\text{depth}] = \lambda$$

$$\text{Var}[\text{depth}] = \lambda$$

Expect 4 reads, all with As at this ref position.

Sequencing (sampling) the two different alleles is just like flipping a coin.

A vs G

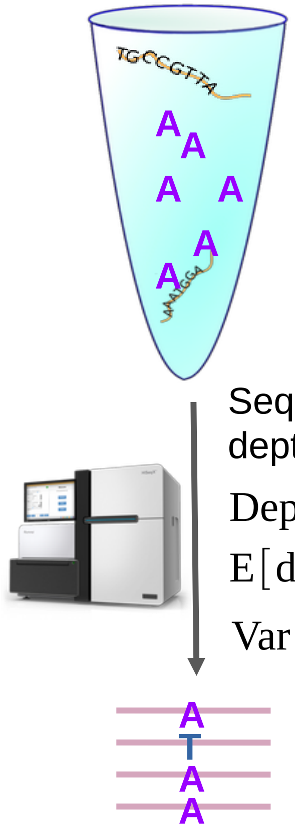
$$\#\,\text{A alleles} \sim \text{Binomial}(n\,reads, p = 0.5)$$

$$E[\text{A depth}] = np = 0.5n$$

$$\text{Var}[\text{A depth}] = np(1-p) = 0.25n$$

Expect 2 A alleles and 2 G alleles

Why Probabilistic Methods?



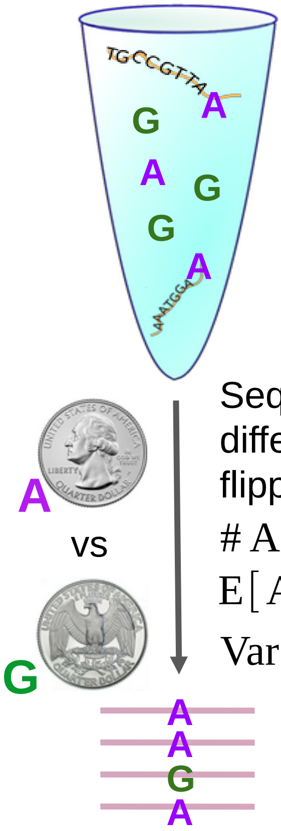
Sequence to average depth of 4x.

$$\text{Depth} \sim \text{Poisson}(\lambda=4)$$

$$E[\text{depth}] = \lambda$$

$$\text{Var}[\text{depth}] = \lambda$$

Sequencing error. Rates of ~0.1% for some Illumina platforms.



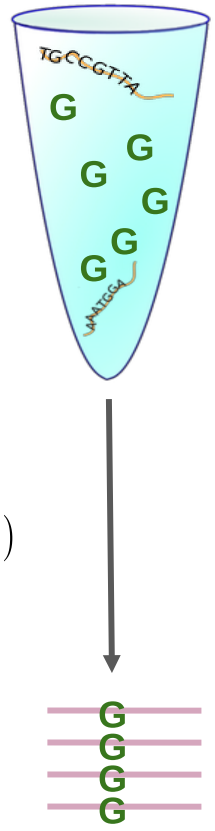
Sequencing (sampling) the two different alleles is just like flipping a coin.

$$\#\,\text{A alleles} \sim \text{Binomial}(n\,reads, p=0.5)$$

$$E[\text{A depth}]=np=0.5n$$

$$\text{Var}[\text{A depth}]=np(1-p)=0.25n$$

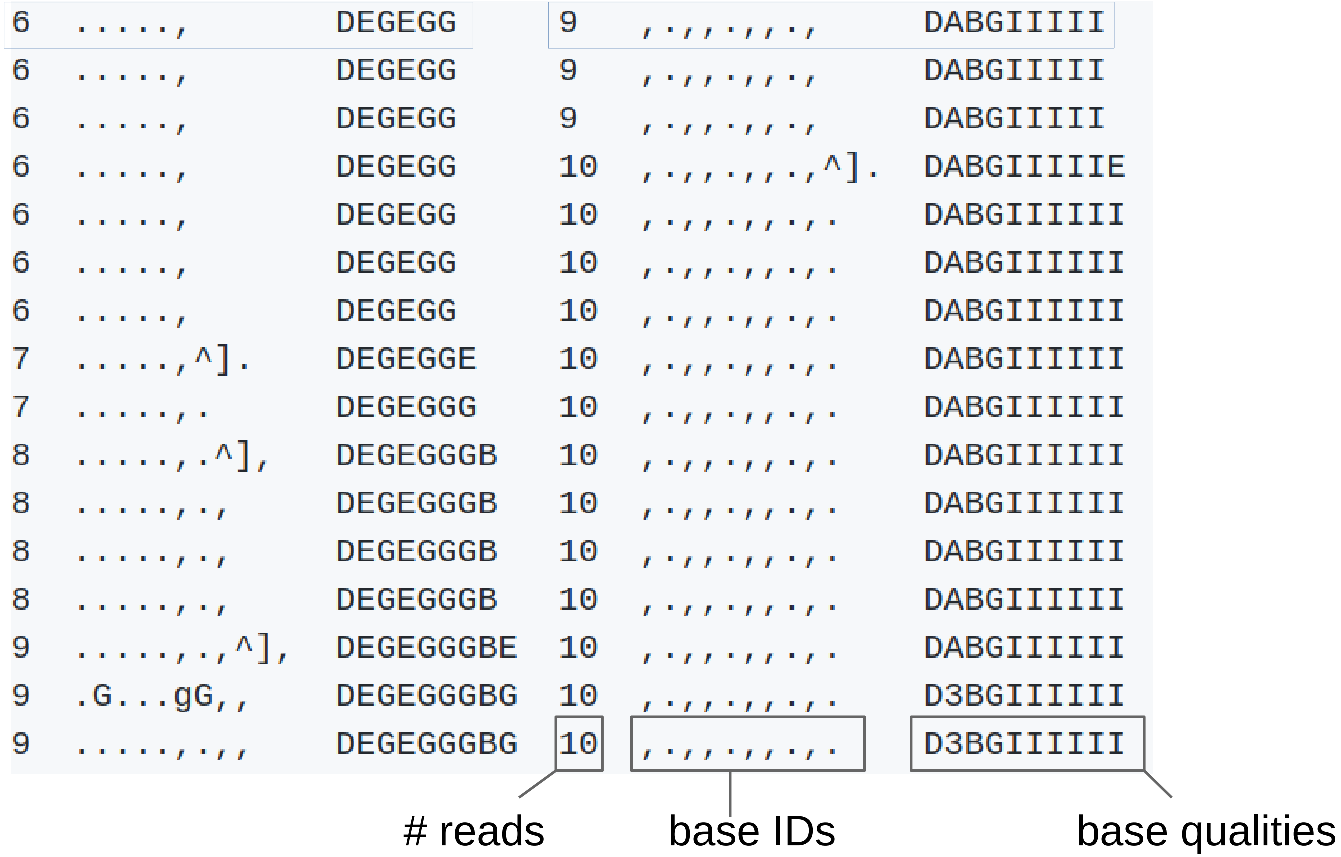
Expect 2 A alleles and 2 G alleles



SAMtools mpileup representation of sequencing data for two individuals

Individual 1 data

Individual 2 data



```
6  .....,       DEGEGG     9   ,.,,.,,.,     DABGIIIII
6  .....,       DEGEGG     9   ,.,,.,,.,     DABGIIIII
6  .....,       DEGEGG     9   ,.,,.,,.,     DABGIIIII
6  .....,       DEGEGG     10  ,.,,.,,.,^].  DABGIIIIE
6  .....,       DEGEGG     10  ,.,,.,,.,.    DABGIIIIII
6  .....,       DEGEGG     10  ,.,,.,,.,.    DABGIIIIII
6  .....,       DEGEGG     10  ,.,,.,,.,.    DABGIIIIII
7  .....,^].    DEGEGGE    10  ,.,,.,,.,.    DABGIIIIII
7  .....,.      DEGEGGG    10  ,.,,.,,.,.    DABGIIIIII
8  .....,.^],   DEGEGGGB   10  ,.,,.,,.,.    DABGIIIIII
8  .....,.,     DEGEGGGB   10  ,.,,.,,.,.    DABGIIIIII
8  .....,.,     DEGEGGGB   10  ,.,,.,,.,.    DABGIIIIII
8  .....,.,     DEGEGGGB   10  ,.,,.,,.,.    DABGIIIIII
9  .....,.,^],  DEGEGGGBE  10  ,.,,.,,.,.    DABGIIIIII
9  .G...gG,,    DEGEGGGBG  10  ,.,,.,,.,.    D3BGIIIIII
9  .....,.,,    DEGEGGGBG  10  ,.,,.,,.,.    D3BGIIIIII
```

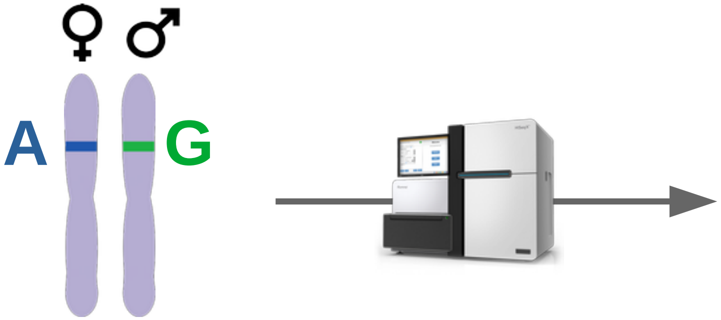
# reads

base IDs

base qualities

Each row is a different site in the reference genome

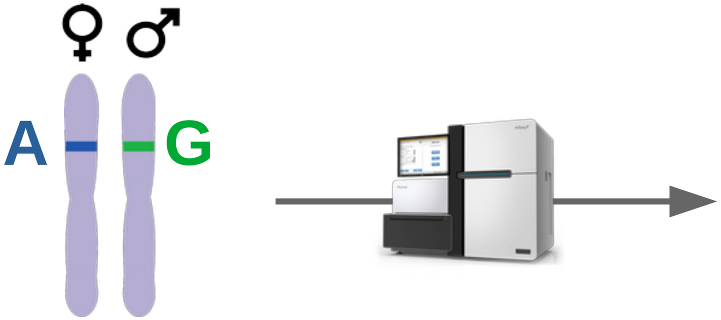
Example sequencing data for one individual at Chr1:472



Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

Maternally and paternally inherited chromosome 1 of a diploid individual.

Example sequencing data for one individual at Chr1:472



Maternally and paternally inherited chromosome 1 of a diploid individual.

Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

↓

ASCII character “D” = decimal value 68

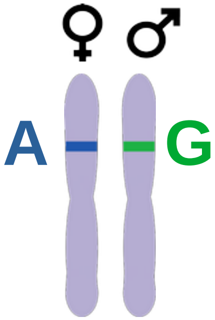
↓

68 – 33 = base quality of 35

↓

$$P(\text{error}) = \epsilon = 10^{\frac{-35}{10}} = 0.00032$$

Example sequencing data for one individual at Chr1:472



Maternally and paternally inherited chromosome 1 of a diploid individual.

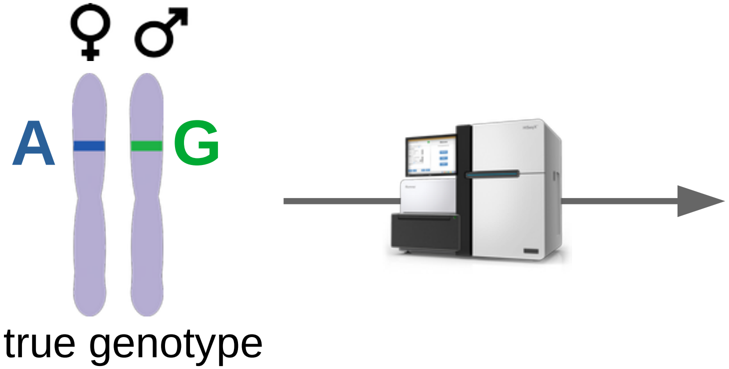
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

This individual could have any of the following 10 genotypes (we can only see the sequencing data):

AA, AC, AG, AT, CG, CC, CT, GG, GT, TT

How do we figure out which genotype they are most likely to have based on the observed sequence data?

Genotype likelihoods



Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

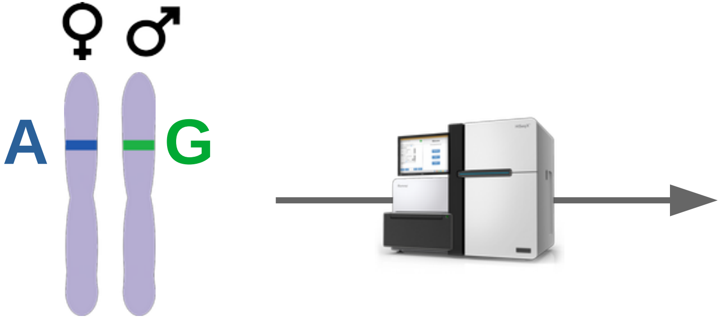
$$\mathrm{P}\left(\mathrm{Data}\middle|\mathrm{Genotype}=bh\right) = L\left(\mathrm{Genotype}\,bh\right) = \text{likelihood of genotype}\,bh$$

$$b,h \in \left\{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\right\}$$

Possible genotypes: AA, AC, AG, AT, CG, CC, CT, GG, GT, TT

What is the likelihood that the individual’s genotype is AC?

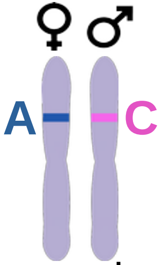
Possible genotypes: AA, **AC**, AG, AT, CG, CC, CT, GG, GT, TT



true genotype

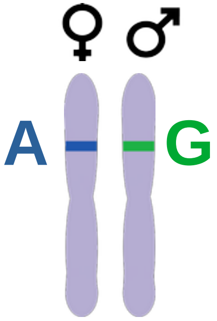
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)= \underbrace{\boxed{}}_{\substack{\text{Probability of sampling}\\ \text{maternal chromosome}\\ \text{(A allele)}}} \times \overbrace{\boxed{}}^{\text{P(observed } i\text{-th read | A allele)}} + \underbrace{\boxed{}}_{\substack{\text{Probability of sampling}\\ \text{paternal chromosome}\\ \text{(C allele)}}} \times \overbrace{\boxed{}}^{\text{P(observed } i\text{-th read | C allele)}}$$



Assumed true genotype

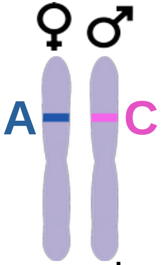
What is the likelihood that the individual’s genotype is AC?



true genotype

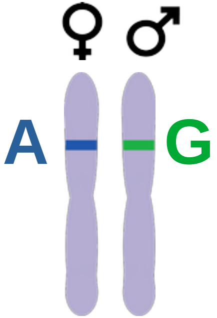
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)= \underbrace{\boxed{}}_{\textbf{Probability of sampling maternal chromosome (A allele)}} \times \overbrace{\boxed{}}^{\text{P(observed read A | A allele)}} + \underbrace{\boxed{}}_{\text{Probability of sampling paternal chromosome (C allele)}} \times \overbrace{\boxed{} \times \boxed{}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

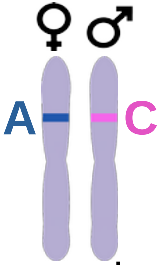
What is the likelihood that the individual's genotype is AC?



true genotype

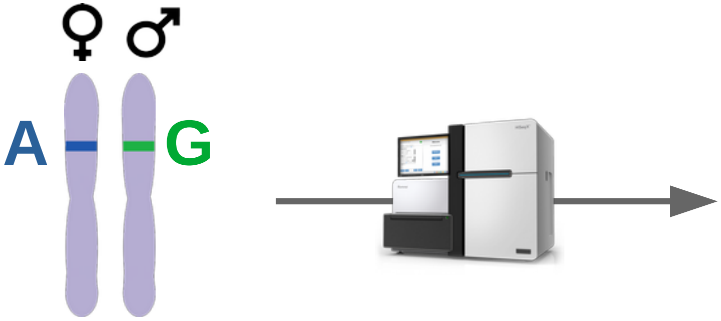
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\mathrm{Data}\middle|\mathrm{Genotype}=\mathrm{AC}\right)= \underbrace{\boxed{\frac{1}{2}}}_{\substack{\text{Probability of sampling}\\ \text{maternal chromosome}\\ \text{(A allele)}}} \times \overbrace{\boxed{\phantom{\frac{1}{2}}}}^{\text{P(observed read A | A allele)}} + \underbrace{\boxed{\phantom{\frac{1}{2}}}}_{\substack{\text{Probability of sampling}\\ \text{paternal chromosome}\\ \text{(C allele)}}} \times \overbrace{\boxed{\phantom{\frac{1}{2}}} \times \boxed{\phantom{\frac{1}{2}}}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

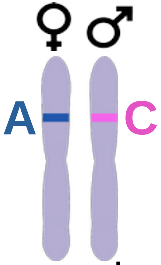
What is the likelihood that the individual's genotype is AC?



true genotype

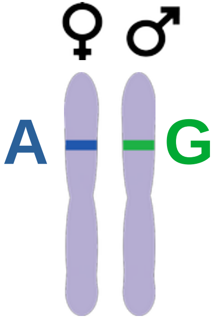
Scaffold	position	read bases	base qualities
Chr1	472	**A**AG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling maternal chromosome (A allele)}}\times\overbrace{\boxed{\phantom{\frac{1}{2}}}}^{\textbf{P(observed read A | A allele)}}+\underbrace{\boxed{\phantom{\frac{1}{2}}}}_{\text{Probability of sampling paternal chromosome (C allele)}}\times\overbrace{\boxed{\phantom{\frac{1}{2}}}\times\boxed{\phantom{\frac{1}{2}}}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

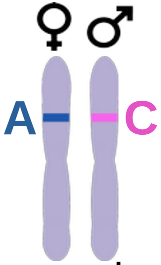
What is the likelihood that the individual’s genotype is AC?



true genotype

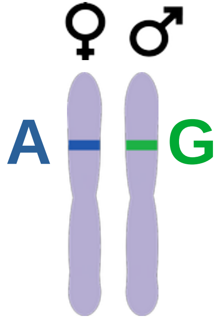
Scaffold	position	read bases	base qualities
Chr1	472	**A**AG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)= \underbrace{\boxed{\frac{1}{2}}}_{\substack{\text{Probability of sampling}\\ \text{maternal chromosome}\\ \text{(A allele)}}} \times \overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}} + \underbrace{\boxed{\phantom{\frac{1}{2}}}}_{\substack{\text{Probability of sampling}\\ \text{paternal chromosome}\\ \text{(C allele)}}} \times \overbrace{\boxed{\phantom{\frac{1}{2}}} \times \boxed{\phantom{\frac{1}{2}}}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

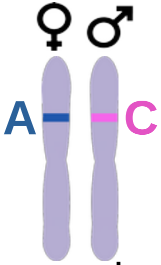
What is the likelihood that the individual’s genotype is AC?



true genotype

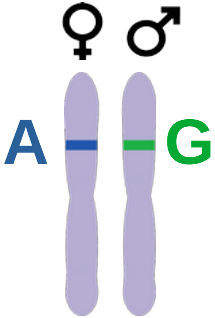
Scaffold	position	read bases	base qualities
Chr1	472	**A**AG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling maternal chromosome (A allele)}}\times\overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}}+\underbrace{\boxed{\phantom{\frac{1}{2}}}}_{\textbf{Probability of sampling paternal chromosome (C allele)}}\times\overbrace{\boxed{\phantom{\frac{1}{2}}}\times\boxed{\phantom{\frac{1}{2}}}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

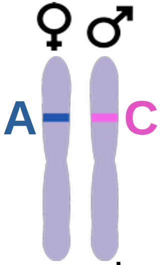
What is the likelihood that the individual’s genotype is AC?



true genotype

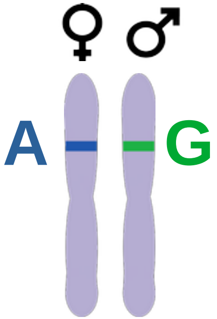
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\underbrace{\boxed{\frac{1}{2}}}_{\substack{\text{Probability of sampling}\\\text{maternal chromosome}\\\text{(A allele)}}}\times\overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}}+\underbrace{\boxed{\frac{1}{2}}}_{\substack{\text{Probability of sampling}\\\text{paternal chromosome}\\\text{(C allele)}}}\times\overbrace{\boxed{}\times\boxed{}}^{\text{P(observed read A | C allele)}}$$



Assumed true genotype

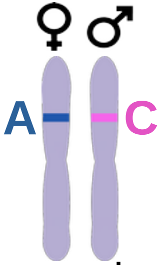
What is the likelihood that the individual’s genotype is AC?



true genotype

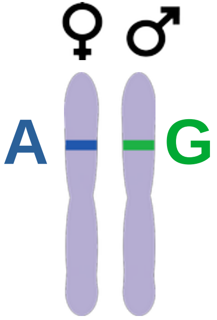
Scaffold	position	read bases	base qualities
Chr1	472	**A**AG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling maternal chromosome (A allele)}}\times\overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}}+\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling paternal chromosome (C allele)}}\times\overbrace{\boxed{}\times\boxed{}}^{\textbf{P(observed read A | C allele)}}$$



Assumed true genotype

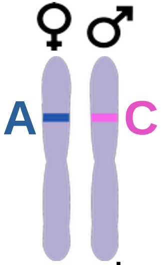
What is the likelihood that the individual’s genotype is AC?



true genotype

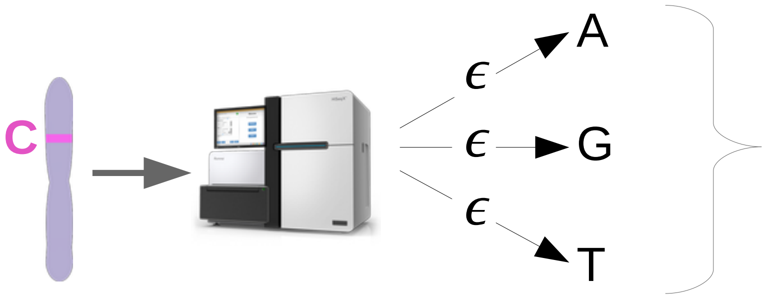
Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling maternal chromosome (A allele)}}\times\overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}}+\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling paternal chromosome (C allele)}}\times\overbrace{\boxed{\epsilon_1}\times\boxed{\frac{1}{3}}}^{\text{P(observed read A | C allele)}}$$



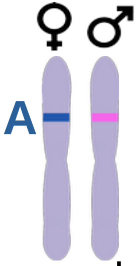
Assumed true genotype

What is the likelihood that the individuals genotype is AC?

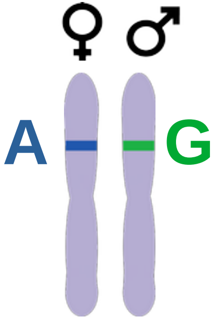


Assuming equal probability of changing to any of the possible 3 erroneous bases.

$$P\left(\text{Data}\middle|\text{Genotype}=AC\right)=\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling maternal chromosome (A allele)}}\times\overbrace{\boxed{1-\epsilon_1}}^{\text{P(observed read A | A allele)}}+\underbrace{\boxed{\frac{1}{2}}}_{\text{Probability of sampling paternal chromosome (C allele)}}\times\overbrace{\boxed{\epsilon_1}\times\boxed{\frac{1}{3}}}^{\text{P(observed read A | C allele)}}$$



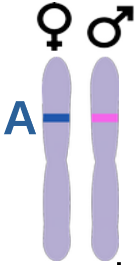
Assumed true genotype



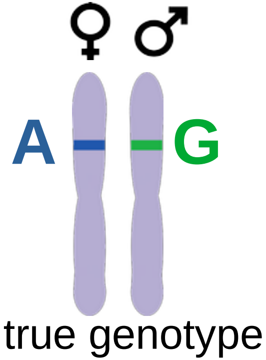
true genotype

Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=AC\right)=\left[\boxed{\frac{1}{2}}\times\boxed{1-\epsilon_1}+\boxed{\frac{1}{2}}\times\boxed{\epsilon_1}\times\boxed{\frac{1}{3}}\right]\times\left[\boxed{\frac{1}{2}}\times\boxed{1-\epsilon_2}+\boxed{\frac{1}{2}}\times\boxed{\epsilon_2}\times\boxed{\frac{1}{3}}\right]$$



Assumed true genotype



Scaffold	position	read bases	base qualities
Chr1	472	AAG	DEG

$$P\left(\text{Data}\middle|\text{Genotype}=\text{AC}\right)=\left[\boxed{\frac{1}{2}}\times\boxed{1-\epsilon_1}\quad\boxed{\frac{1}{2}}\times\boxed{\epsilon_1}\times\boxed{\frac{1}{3}}\right]\times$$

$$\left[\boxed{\frac{1}{2}}\times\boxed{1-\epsilon_2}+\boxed{\frac{1}{2}}\times\boxed{\epsilon_2}\times\boxed{\frac{1}{3}}\right]\times$$

$$\left[\boxed{\frac{1}{2}}\times\boxed{\epsilon_3}\times\boxed{\frac{1}{3}}+\boxed{\frac{1}{2}}\times\boxed{\epsilon_3}\times\boxed{\frac{1}{3}}\right]$$

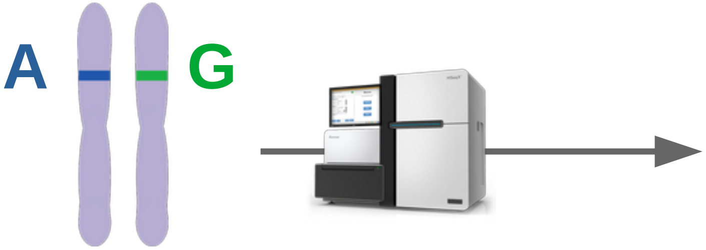
General genotype likelihood expression

$$\mathrm{P}\left(\text{Data} \mid \text{Genotype} = bh\right) = \prod_{i=1}^{r} \left(\frac{L_b^{(i)}}{2} + \frac{L_h^{(i)}}{2} \right)$$

$$b, h \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$$

$$\boxed{\begin{aligned} L_b^{(i)} &= \mathrm{P}\left(\text{observed read} = x_i \middle| \text{assumed true allele} = b\right) \\ L_h^{(i)} &= \mathrm{P}\left(\text{observed read} = x_i \middle| \text{assumed true allele} = h\right) \end{aligned}} \left.\vphantom{\begin{aligned} L \\ L \end{aligned}}\right\} \begin{array}{r} \dfrac{\epsilon_i}{3} \text{ if } b, h \neq x_i \\ 1 - \epsilon_i \text{ if } b, h = x_i \end{array}$$

Representation of genotype likelihoods in ANGSD



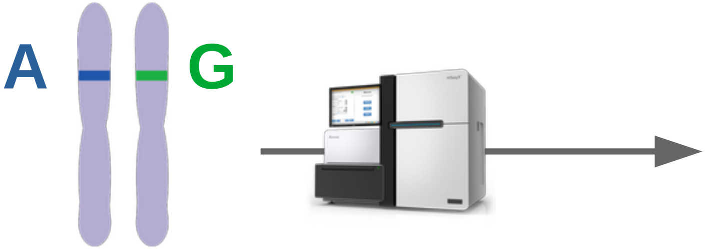
Chr1 472 AAG 555

ASCII character “5” = decimal value 53

53 – 33 = base quality of 20

$$\epsilon = 10^{\frac{-20}{10}} = 0.01$$

Representation of genotype likelihoods in ANGSD



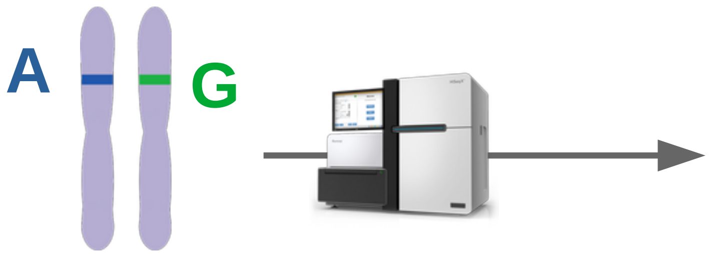
Chr1   472   AAG   555

Instead of a single genotype, AG, we have a distribution over all possible genotypes:

Genotype	AA	AC	AG	AT	CC	CG	CT	GG	GT	TT
Log$_{10}$ likelihood	-2.49	-3.08	-0.91	-3.08	-7.43	-5.26	-7.43	-4.96	-5.26	-7.43

$$\epsilon = 10^{\frac{-20}{10}} = 0.01$$

Representation of genotype likelihoods in ANGSD



Chr1 472 AAG 555

Instead of a single genotype, AG, we have a distribution over all possible genotypes:

Genotype	AA	AC	**AG**	AT	CC	CG	CT	GG	GT	TT
Log$_{10}$ likelihood	-2.49	-3.08	**-0.91**	-3.08	-7.43	-5.26	-7.43	-4.96	-5.26	-7.43

Maximum likelihood estimate of the genotype

$$\epsilon = 10^{\frac{-20}{10}} = 0.01$$

Exercise. Calculate genotype likelihoods with ANGSD.

Estimating allele frequencies

When genotypes are known, allele frequencies can be calculated by simply counting alleles.

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	Ind10	Minor allele frequency (MAF)
Site 1	0	0	1	0	0	0	1	1	0	0	0.3
Site 2	0	2	0	0	1	0	0	0	1	0	0.4

Genotype notation

0 = zero minor alleles
1 = one minor alleles
2 = two minor alleles

Estimating allele frequencies

When genotypes are known, allele frequencies can be calculated by simply counting alleles.

	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6	Ind7	Ind8	Ind9	Ind10		Minor allele frequency (MAF)
Site 1	0	0	1	0	0	0	1	1	0	0	=	0.3
Site 2	0	2	0	0	1	0	0	0	1	0	=	0.4

But how do you estimate allele frequencies when you have a distribution of genotype likelihoods?

Maximum likelihood estimation of allele frequencies

$$\mathrm{P}(\mathrm{Data}|f)=\prod_{i=1}^{n}\sum_{g\in\{0,1,2\}}\mathrm{P}(D_i|\mathrm{Genotype}_i=g)\,\mathrm{P}(\mathrm{Genotype}_i=g|f)$$

$D_i=$ sequencing data for individual i

$f=$ population minor allele frequency

<u>Genotype notation</u>

0 = zero minor alleles

1 = one minor alleles

2 = two minor alleles

Maximum likelihood estimation of allele frequencies

$$\mathrm{P}\left(\mathrm{Data}\middle|f\right)=\prod_{i=1}^{n}\sum_{g\in\{0,1,2\}}\underbrace{\mathrm{P}\left(D_i\middle|\mathrm{Genotype}_i=g\right)}\mathrm{P}\left(\mathrm{Genotype}_i=g\middle|f\right)$$

This the likelihood of genotype *g* for individual *i* calculated as shown previously.

$D_i=$ sequencing data for individual i

$f=$ population minor allele frequency

Genotype notation

0 = zero minor alleles
1 = one minor alleles
2 = two minor alleles

Maximum likelihood estimation of allele frequencies

$$\mathrm{P}\left(\text{Data} \middle| f\right) = \prod_{i=1}^{n} \sum_{g \in \{0,1,2\}} \underbrace{\mathrm{P}\left(D_i \middle| \text{Genotype}_i = g\right)}_{\text{This the likelihood of genotype } g \text{ for individual } i \text{ calculated as shown previously.}} \underbrace{\mathrm{P}\left(\text{Genotype}_i = g \middle| f\right)}_{\text{Hardy-Weinberg frequency of genotypes.}}$$

$D_i =$ sequencing data for individual i

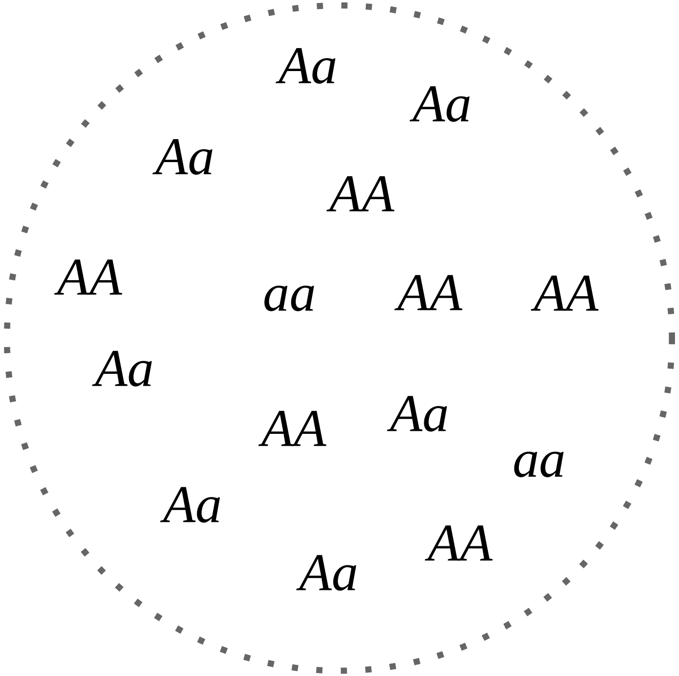
$f =$ population minor allele frequency

<u>Genotype notation</u>

0 = zero minor alleles
1 = one minor alleles
2 = two minor alleles

Genotype frequencies under Hardy-Weinberg Equilibrium (HWE)

$$\mathrm{P}\left(\mathrm{Data}|f\right)=\prod_{i=1}^{n}\sum_{g\in\{0,1,2\}}\mathrm{P}\left(D_i|\mathrm{Genotype}_i=g\right)\mathrm{P}\left(\mathrm{Genotype}_i=g|f\right)$$

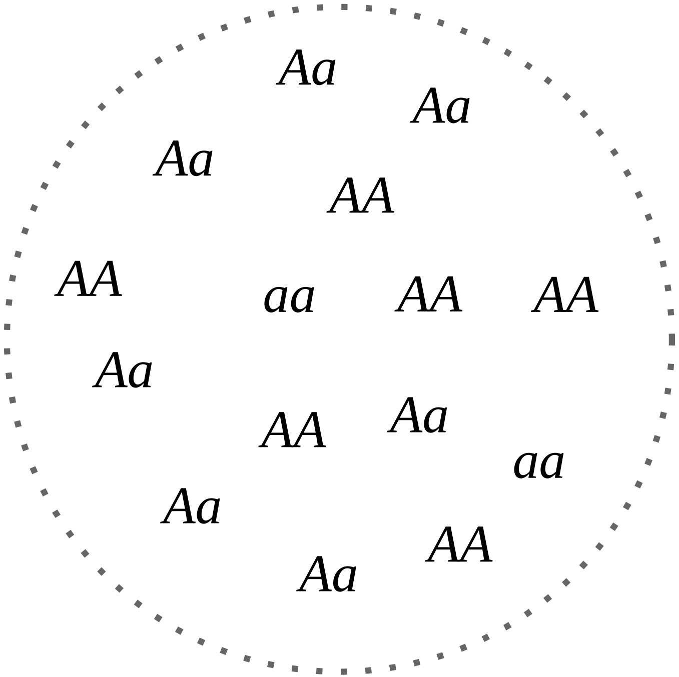


An “infinitely” large population of sexually reproducing diploid organisms segregating for alleles *A* and *a*, and for which

- Mating is random.
- Generations are nonoverlapping.
- Allele frequencies are the same in males and females.
- No migration, mutation, or selection.

Genotype frequencies under Hardy-Weinberg Equilibrium (HWE)

$$\mathrm{P}(\mathrm{Data}|f)=\prod_{i=1}^{n}\sum_{g\in\{0,1,2\}}\mathrm{P}(D_i|\mathrm{Genotype}_i=g)\,\mathrm{P}(\mathrm{Genotype}_i=g|f)$$



Offspring are formed through independent draws of gametes from this population, expressed as $(f_A + f_a)^2$, which expanded yields:

$$f_{AA} = f_A^2 = (1-f_a)^2 = \mathrm{P}(\mathrm{Genotype}=AA|f_a)$$

$$f_{Aa} = 2f_A f_a = 2(1-f_a)f_a = \mathrm{P}(\mathrm{Genotype}=1|f_a)$$

$$f_{aa} = f_a^2 = \mathrm{P}(\mathrm{Genotype}=2|f_a)$$

Maximum likelihood estimation of allele frequencies

$$\mathrm{P}\left(\mathrm{Data}\middle|f\right)=\prod_{i=1}^{n}\underbrace{\sum_{g\in\{0,1,2\}}}_{\substack{\text{Summing over the possible}\\ \text{genotypes accounts for}\\ \text{genotyping uncertainty.}}}\mathrm{P}\left(D_i\middle|\mathrm{Genotype}_i=g\right)\mathrm{P}\left(\mathrm{Genotype}_i=g\middle|f\right)$$

Marginal probabilities are used to account for uncertainty in various quantities associated with low coverage sequencing. In general, for random variables *X* and *Y*

$$\mathrm{P}\left(X=x\right)=\sum_{y}\mathrm{P}\left(X=x\middle|Y=y\right)\mathrm{P}\left(Y=y\right)$$

Maximum likelihood estimation of allele frequencies

$$\mathrm{P}(\mathrm{Data}|f)=\prod_{i=1}^{n}\sum_{g\in\{0,1,2\}}\mathrm{P}(D_i|\mathrm{Genotype}_i=g)\,\mathrm{P}(\mathrm{Genotype}_i=g|f)$$

The value of *f* that maximizes the likelihood function above yields a **maximum likelihood estimate of *f*** :

$$\hat{f}=\mathrm{argmax}_f\,\mathrm{P}(Data|f)$$

Using the ML allele frequency estimate to identify polymorphic sites

$$\mathrm{P}(\mathrm{Data}|f) = \prod_{i=1}^{n} \sum_{g \in \{0,1,2\}} \mathrm{P}(D_i|\mathrm{Genotype}_i = g)\,\mathrm{P}(\mathrm{Genotype}_i = g|f)$$

$$\hat{f} = \mathrm{argmax}_f\, \mathrm{P}(\mathit{Data}|f)$$

← **maximum likelihood estimate of *f***

Probability of the sequencing data when *f* = 0, i.e., the site is monomorphic (null case).

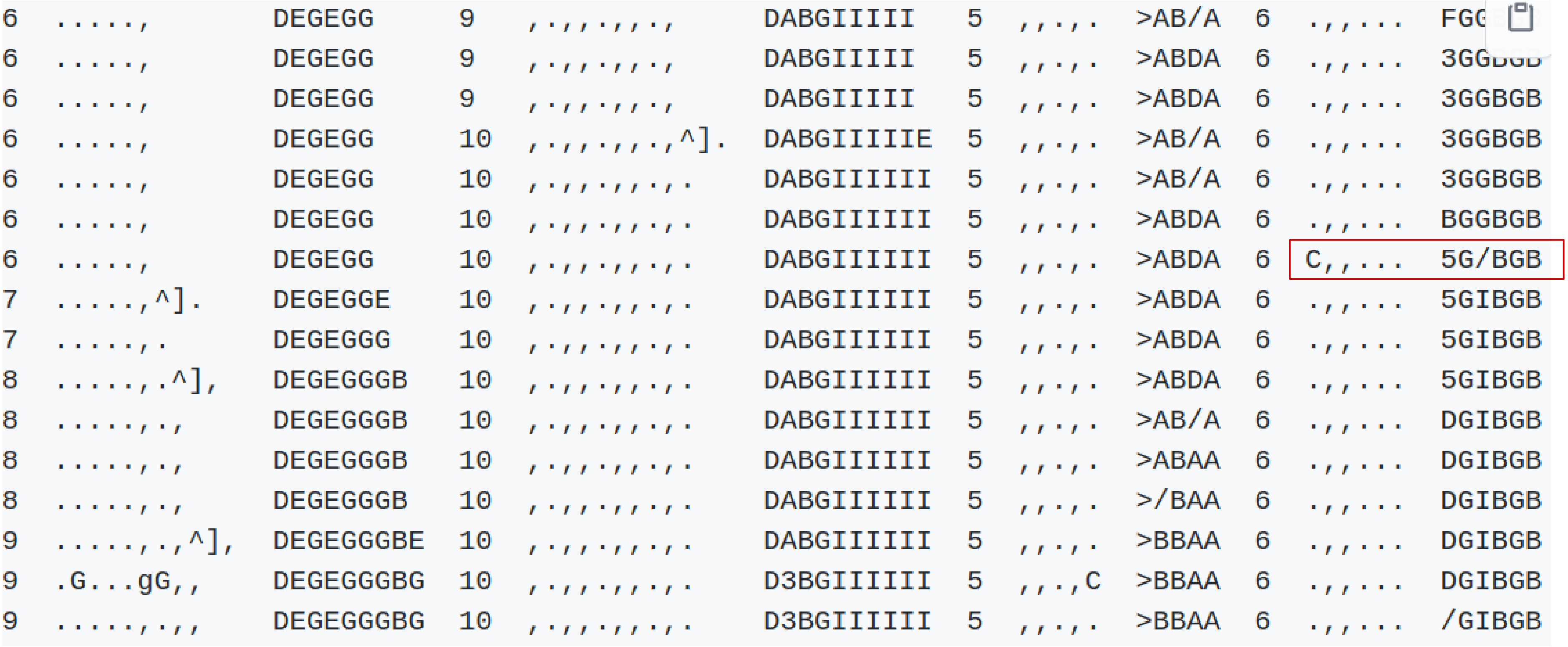
$$\lambda = -2\ln\left(\frac{\mathrm{P}(\mathrm{Data}|f_0)}{\mathrm{P}(\mathrm{Data}|\hat{f})}\right) = -2\left[\ln\left(\overbrace{\mathrm{P}(\mathrm{Data}|f_0)}\right) - \ln\left(\mathrm{P}(\mathrm{Data}|\hat{f})\right)\right]$$

$$\lambda \sim \chi^2(1\ \text{degree of freedom}) \longrightarrow$$

Call SNPs at a given level of statistical confidence.

Exercise. Estimate allele frequencies with ANGSD.

Can we use other information from our data to further increase our genotyping accuracy?

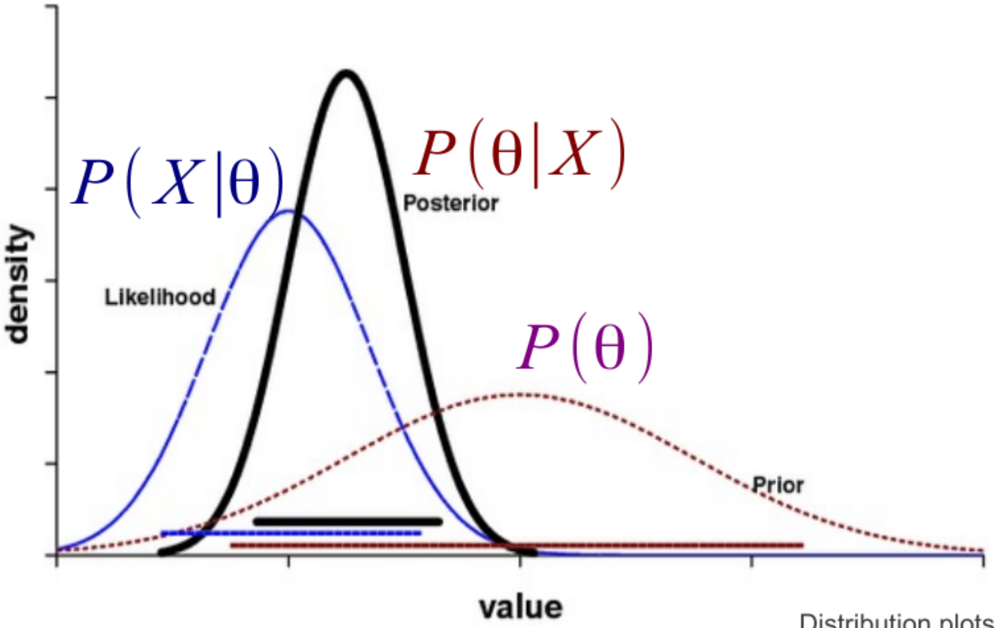


```
6  .....,      DEGEGG     9   ,.,,.,,.,      DABGIIIII   5  ,,.,.  >AB/A  6  .,,...  FGG
6  .....,      DEGEGG     9   ,.,,.,,.,      DABGIIIII   5  ,,.,.  >ABDA  6  .,,...  3GGBGB
6  .....,      DEGEGG     9   ,.,,.,,.,      DABGIIIII   5  ,,.,.  >ABDA  6  .,,...  3GGBGB
6  .....,      DEGEGG     10  ,.,,.,,.,^].   DABGIIIIIE  5  ,,.,.  >AB/A  6  .,,...  3GGBGB
6  .....,      DEGEGG     10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >AB/A  6  .,,...  3GGBGB
6  .....,      DEGEGG     10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABDA  6  .,,...  BGGBGB
6  .....,      DEGEGG     10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABDA  6  C,,...  5G/BGB
7  .....,^].   DEGEGGE    10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABDA  6  .,,...  5GIBGB
7  .....,.     DEGEGGG    10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABDA  6  .,,...  5GIBGB
8  .....,.^],  DEGEGGGB   10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABDA  6  .,,...  5GIBGB
8  .....,.,    DEGEGGGB   10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >AB/A  6  .,,...  DGIBGB
8  .....,.,    DEGEGGGB   10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >ABAA  6  .,,...  DGIBGB
8  .....,.,    DEGEGGGB   10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >/BAA  6  .,,...  DGIBGB
9  .....,.,^], DEGEGGGBE  10  ,.,,.,,.,.     DABGIIIIII  5  ,,.,.  >BBAA  6  .,,...  DGIBGB
9  .G...gG,,   DEGEGGGBG  10  ,.,,.,,.,.     D3BGIIIIII  5  ,,.,C  >BBAA  6  .,,...  DGIBGB
9  .....,.,,   DEGEGGGBG  10  ,.,,.,,.,.     D3BGIIIIII  5  ,,.,.  >BBAA  6  .,,...  /GIBGB
```

Wouldn’t it be awesome if you knew what the frequency of C was in the population.

Bayesian Inference

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



X = Data

θ = Parameter

Distribution plots from Bink 2008

Genotype posterior probabilities to improve genotyping accuracy

Using Bayes' Theorem, the posterior probability of genotype *g* is

$$\mathrm{P}\left(\mathrm{Genotype}=g \middle| \mathrm{Data}\right) = \frac{\mathrm{P}\left(\mathrm{Data} \middle| \mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}{\sum_{g\in\{0,1,2\}} \mathrm{P}\left(\mathrm{Data} \middle| \mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}$$

Genotype posterior probabilities to improve genotyping accuracy

Using Bayes' Theorem, the posterior probability of genotype *g* is

Likelihood of genotype *g* calculated as shown previously.

$$\mathrm{P}\left(\mathrm{Genotype}=g\,|\,\mathrm{Data}\right)=\frac{\mathrm{P}\left(\mathrm{Data}\,|\,\mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}{\sum_{g\in\{0,1,2\}}\mathrm{P}\left(\mathrm{Data}\,|\,\mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}$$

Genotype posterior probabilities to improve genotyping accuracy

Using Bayes' Theorem, the posterior probability of genotype *g* is

Given an estimate of the population minor allele frequency, *f*, under HWE

$$\mathrm{P}(\text{Genotype}=0|f)=(1-f)^2$$

$$\mathrm{P}(\text{Genotype}=1|f)=2f(1-f)$$

$$\mathrm{P}(\text{Genotype}=2|f)=f^2$$

Likelihood of genotype *g* calculated as shown previously.

$$\mathrm{P}(\text{Genotype}=g|\text{Data}) = \frac{\mathrm{P}(\text{Data}|\text{Genotype}=g)\,\mathrm{P}(\text{Genotype}=g)}{\sum_{g\in\{0,1,2\}} \mathrm{P}(\text{Data}|\text{Genotype}=g)\,\mathrm{P}(\text{Genotype}=g)}$$

Genotype posterior probabilities to improve genotyping accuracy

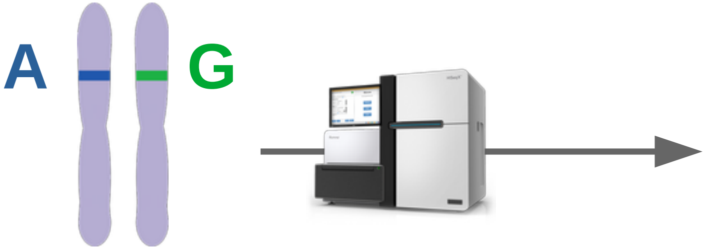
Using Bayes’ Theorem, the posterior probability of genotype *g* is

Note: factors like inbreeding can easily be incorporated into the genotype posterior probabilities by conditioning on the allele frequency and inbreeding coefficient.

Likelihood of genotype *g* calculated as shown previously.

$$\mathrm{P}\left(\mathrm{Genotype}=g \middle| \mathrm{Data}\right) = \frac{\mathrm{P}\left(\mathrm{Data} \middle| \mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}{\sum\limits_{g\in\{0,1,2\}} \mathrm{P}\left(\mathrm{Data} \middle| \mathrm{Genotype}=g\right)\mathrm{P}\left(\mathrm{Genotype}=g\right)}$$

Example genotype posterior probability distribution

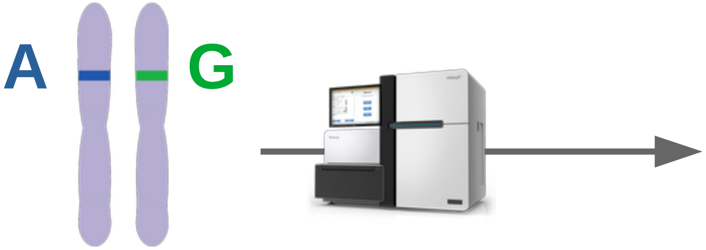


Chr1 472 AAG 555

Assume we estimate f(A) = 0.7, f(G) = 0.3

Genotype	Log_{10} likelihood	Prior	Posterior probability
AA	-2.49	$P(\text{Genotype} = AA) = 0.7^2 = 0.49$	0.03
AG	-0.91	$P(\text{Genotype} = AG) = 2 \times 0.7 \times 0.3 = 0.42$	0.97
GG	-4.96	$P(\text{Genotype} = GG) = 0.3^2 = 0.09$	0.00

Example genotype posterior probability distribution



Chr1 472 AAG 555

We could call most probable genotype, AG, and have an associated degree of confidence (prob = 0.97).

Assume we estimate f(A) = 0.7, f(G) = 0.3

Genotype	Log_{10} likelihood	Prior	Posterior probability
AA	-2.49	$P(\text{Genotype} = \text{AA}) = 0.7^2 = 0.49$	0.03
AG	-0.91	$P(\text{Genotype} = \text{AG}) = 2 \times 0.7 \times 0.3 = 0.42$	0.97
GG	-4.96	$P(\text{Genotype} = \text{GG}) = 0.3^2 = 0.09$	0.00

Exercise. Calculate genotype posterior probabilities and call genotypes with ANGSD.