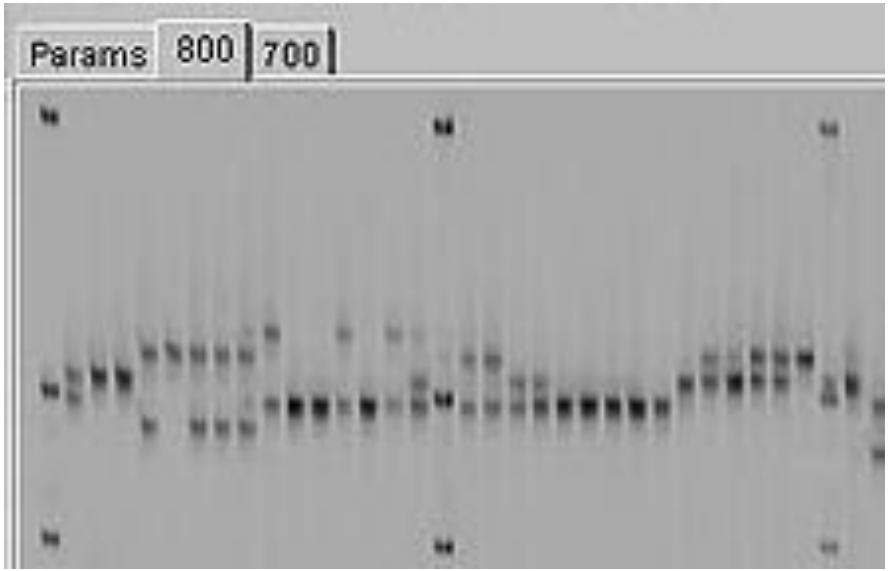
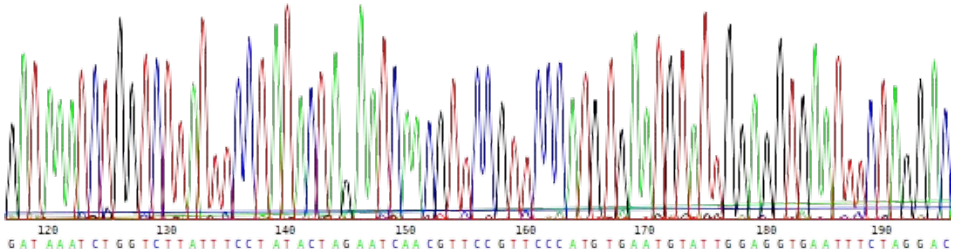
Estimation of allele frequencies, SNP calling, and genotype calling from NGS data

Tyler Linderoth
Copenhagen Popgen Summer Course 2021

Thanks Matteo Fumagalli for lending materials

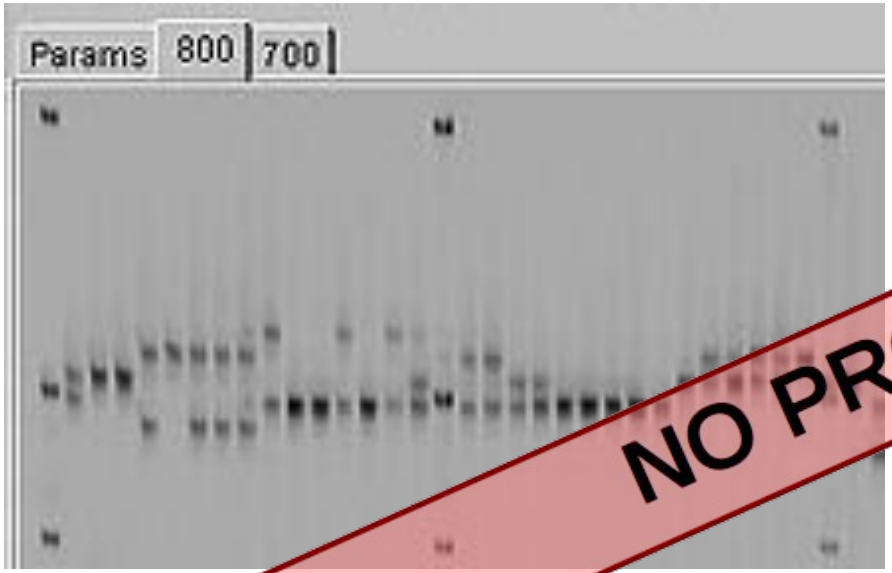
Why Probabilistic Methods?

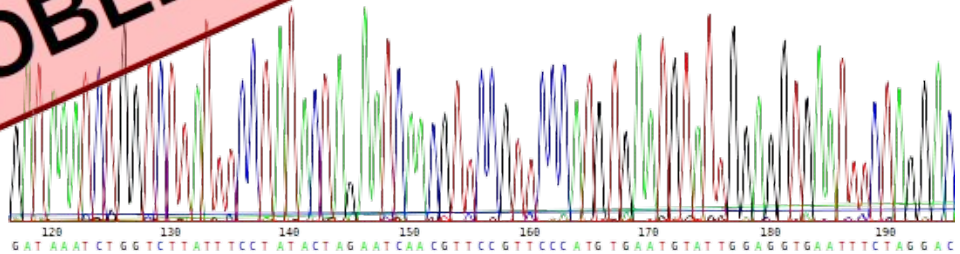




http://www.licor.com/bio/products/software/saga_gt/details.html

Why Probabilistic Methods?

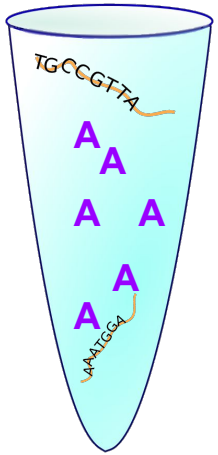




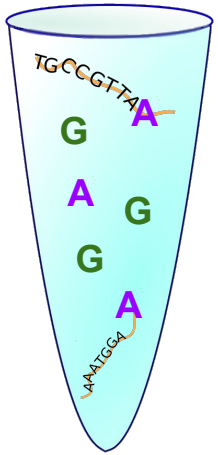
NO PROBLEM!

http://www.licor.com/bio/products/software/saga_gt/details.html

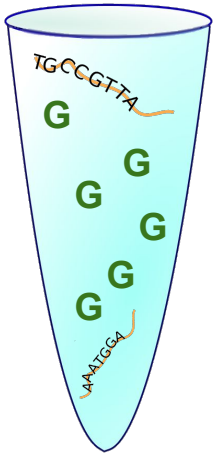
Why Probabilistic Methods?



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

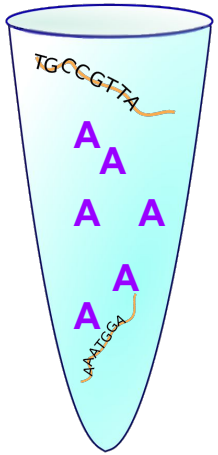


The library for a heterozygous individual at a site contains both **A**s and **G**s.



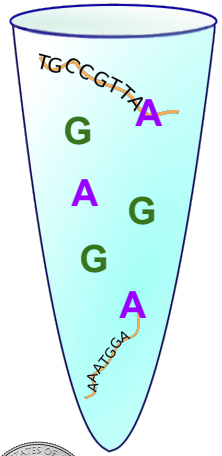
The library for an individual homozygous for the **G** allele w consist only of **G**s.

Why Probabilistic Methods?



Sequence to average depth of 4x.

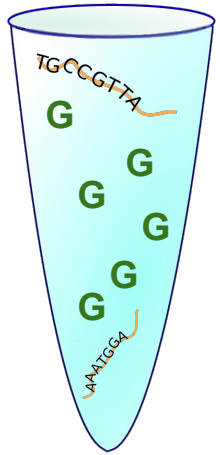
Expect 4 reads, all with **A**s at this ref position.
Depth ~ Poisson(λ = 4).



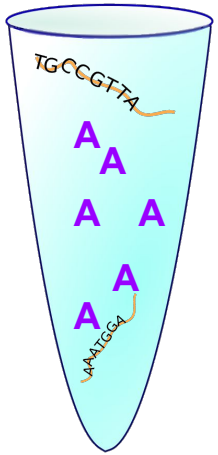
vs

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

P(k A alleles)
~Binom(n reads, p=0.5)

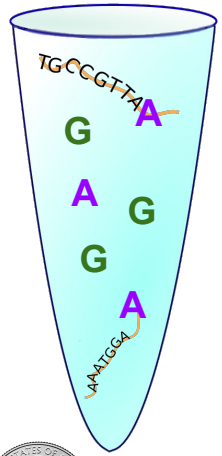


Why Probabilistic Methods?



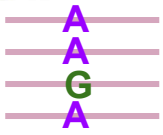
Sequence to average depth of 4x.

Expect 4 reads, all with **A**s at this ref position.
Depth ~ Poisson(λ = 4).

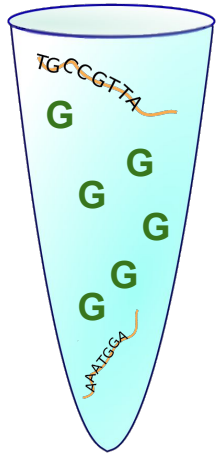


vs

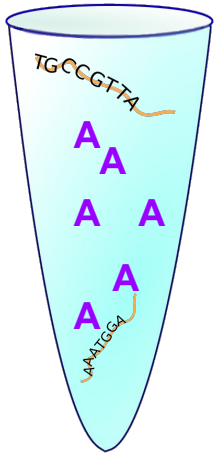
Sequencing (sampling) the two different alleles is just like flipping a coin.



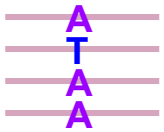
P(k A alleles)
~Binom(n reads, p=0.5)



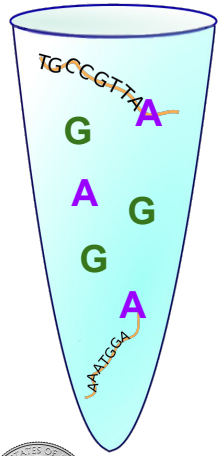
Why Probabilistic Methods?



A sequencing error occurs. This can occur at rates of around 0.1% in Illumina data.

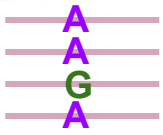


Expect 4 reads, all with **A**s at this ref position.
Depth ~ Poisson(λ = 4).

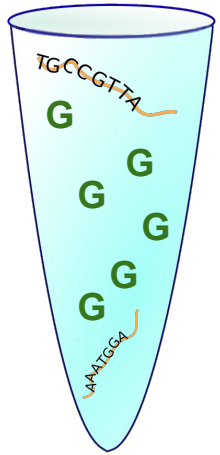


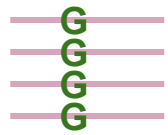
vs

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



P(k A alleles)
~Binom(n reads, p=0.5)





A basic model for a diploid individual's genotype

paternal allele

maternal allele

Product over all reads covering a site for an individual

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right)$$

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

Example for an individual with observed sequencing reads **AAAG** at a site.

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right)$$

10 potential genotypes for a diploid individual

AA
AC
AG
AT
CC
CG
CT
GG
GT
TT

For every genotype, we could calculate it's likelihood by iterating over all of the observed reads: AAAG.

Let's calculate the likelihood for the AC genotype.

Example for an individual with observed sequencing reads **AAAG** at a site.

$$P(X|G=bh)=\prod_{i=1}^{r}\left(\frac{L_b^{(i)}}{2}+\frac{L_h^{(i)}}{2}\right)$$

A,A,A,G

$$P(X|G=AC)=\left(\frac{L_A^{(1)}}{2}+\frac{L_C^{(1)}}{2}\right)*\left(\frac{L_A^{(2)}}{2}+\frac{L_C^{(2)}}{2}\right)*\left(\frac{L_A^{(3)}}{2}+\frac{L_C^{(3)}}{2}\right)*\left(\frac{L_A^{(4)}}{2}+\frac{L_C^{(4)}}{2}\right)$$

II

A

$$L_A^{(1)}=1-\varepsilon$$

$$L_C^{(1)}=\frac{\varepsilon}{3}$$

ε = Probability of an error

There are 3 potential erroneous reads for a error to turn into, hence the 1/3 * ε

$$P(X=A|G=AC)=\frac{1-\varepsilon}{2}+\frac{\varepsilon}{6}$$

Example for an individual with observed sequencing reads **AAAG** at a site.

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

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.

This is essentially what you are doing in ANGSD wh
you choose a uniform prior probability distribution fo
the genotypes.

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).

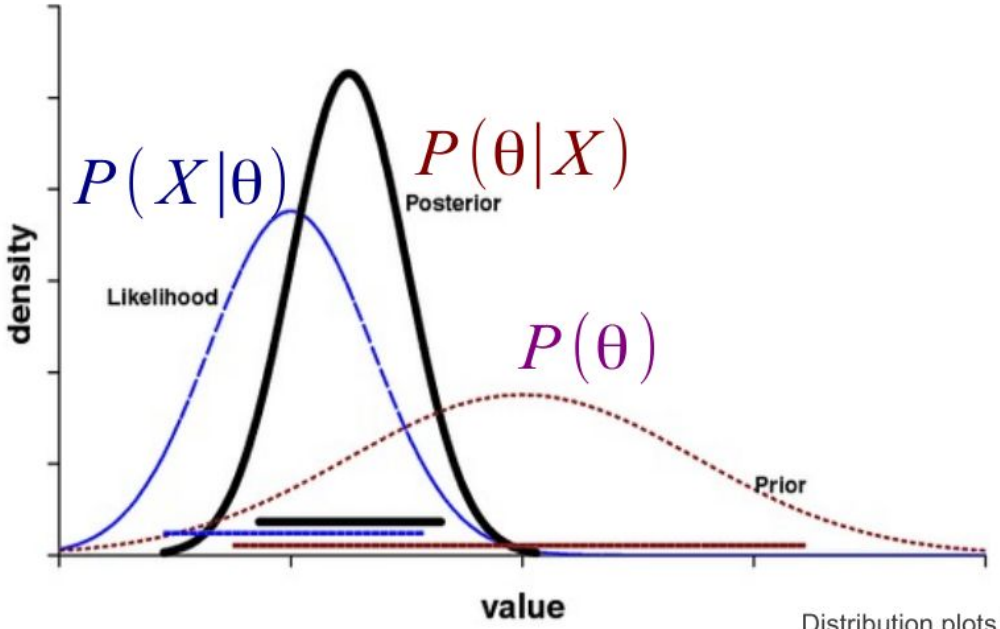
Can we somehow use other information present in our data to further increase our genotype calling accuracy?

6	.....,	DEGEGG	9	,.,,.,,.,	DABGIIIII	5	,,.,.	>AB/A	6	.,,...	FG(
6	.....,	DEGEGG	9	,.,,.,,.,	DABGIIIII	5	,,.,.	>ABDA	6	.,,...	3GG
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 rest of the sample or population.

Bayesian Inference

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



Distribution plots from Bink 2008

From genotype likelihoods to posterior probabilities

Having an estimate of the allele frequency in the population would enable us to have prior knowledge on the probabilities of observing a particular genotype, using for instance principles like Hardy-Weinberg Equilibrium (HWE):

$P(\text{Genotype = 0 minor alleles}) = (1-f_{minor})^2$

$P(\text{Genotype = 1 minor allele}) = 2 * f_{minor} * (1-f_{minor})$

$P(\text{Genotype = 2 minor alleles}) = (f_{minor})^2$

Things like inbreeding can easily be incorporated into these genotype probabilities. So, now we have to know how to estimate allele frequencies.

A simple model to estimate the population minor allele frequency, *f*, is given by

$$p(D_i|f) = \sum_{\mathbf{g}\in\{0,1,2\}} p(D_i|G_i = \mathbf{g})\,p(G_i = \mathbf{g}|f)$$

These are the genotype likelihoods (D_i is the sequencing data for the ith individual) that we now know how to calculate.

And here is just the probability of the genotype given the minor allele frequency, which we can get through HWE.

$$\hat{f} = \arg\max_f \prod_i p(D_i|f)$$

Figure out what minor allele frequency maximizes the above likelihood across all individuals in the sample, and you have a ML estimate of the minor allele frequency.

A simple model to estimate the population minor allele frequency, *f*, is given by

$$p(D_i|f) = \sum_{\mathbf{g}\in\{0,1,2\}} p(D_i|G_i = \mathbf{g})p(G_i = \mathbf{g}|f)$$

$$\hat{f} = \arg\max_f \prod_i p(D_i|f)$$

One thing to note here is that you can compare the likelihood under the ML minor allele frequency to the likelihood calculated from above with f set to zero:

λ = -2*log(ℒ(*f*=0|D) - ℒ(*f*=ML *f*|D))
λ ~ Chi-square(1 d.f.)

Now you have a way to test whether the ML MAF is statistically nonzero, i.e. whether the site is a SNP. Cool!

Now, getting back to this individual with sequencing data AAAG at a site. If we estimate f(A) = 0.7 and we consider only the two most likely alleles (A and G) and ε is always 0.01 (Phred quality of 20, remember that?), then the genotype likelihoods are

Genotype	Likelihood
AA	-5.73
AG	-2.80
GG	-17.12

Apply Bayes' Theorem

Prior probability using *f*(A) = 0.7 and HWE

$$P(G|D) = \frac{P(D|G)\pi(G)}{\sum_{G \in \{0,1,2\}} P(D|G)\pi(G)}$$

And you get genotype posterior probabilities!

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

Now, we can call the genotype as AG based on the max posterior probability (and we also have an estimate of how reliable this call is).

Allele frequency likelihoods

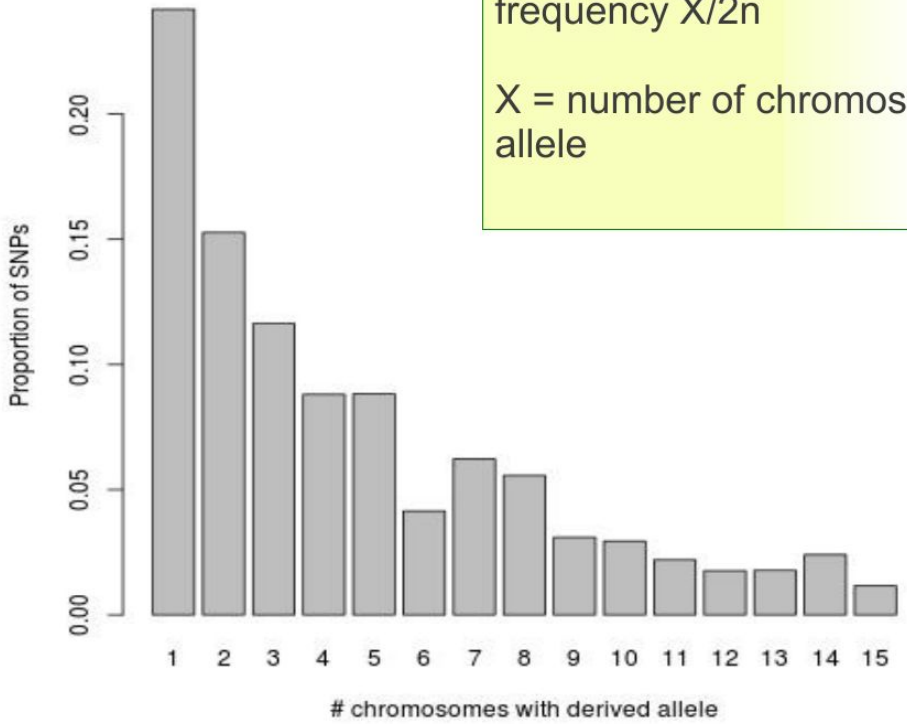
-doSaf 1:

	0	1	2	3	4	. . .	2n
Site1	0.00	-2.24	-4.53	-6.99	-9.63		-232.69
Site2	0.00	-2.24	-4.53	-6.99	-9.63		-232.69
Site3	-76.63	-37.87	-10.42	0.00	-9.59		-467.13
Site4	0.00	-2.24	-5.53	-6.99	-9.63		-237.55
.							
.							
.							
.							
.							
.							
Sitek	0.00	-8.62	-19.22	-30.67	-43.27		-626.78

$$P(X|\theta)$$

Site Frequency Spectrum

-realSFS



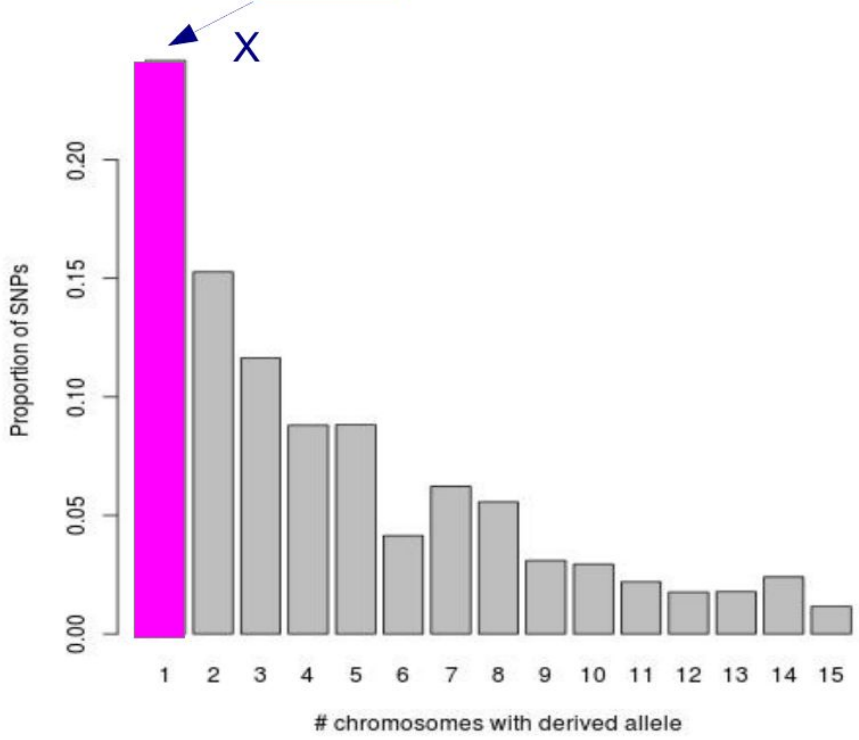
Probability of sampling a site with allele frequency X/2n

X = number of chromosomes with alternative allele

$$P(\theta)$$

	0	1	2	3	4	.	.	.	2n
Site1	0.00	-2.24	-4.53	-6.99	-9.63				-232.69
Site2	0.00	-2.24	-4.53	-6.99	-9.63				-232.69

$P(X|\theta)$

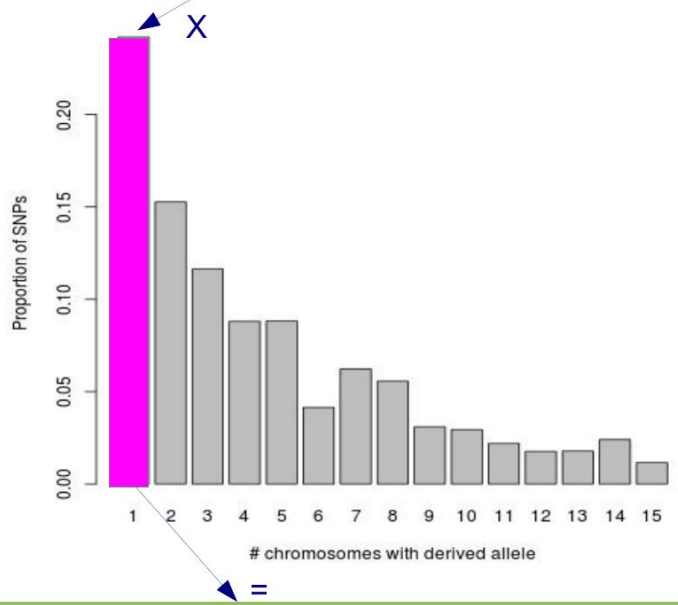


$P(\theta)$ Sample SFS

	0	1	2	3	4	.	.	.	2n
Site1	0.00	-2.24	-4.53	-6.99	-9.63				-232.69
Site2	0.00	-2.24	-4.53	-6.99	-9.63				-232.69

$P(X|\theta)$

X



=

$P(\theta)$ Sample SFS

	0	1	2	3	4	.	.	.	2n
Site1	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000
Site2	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000

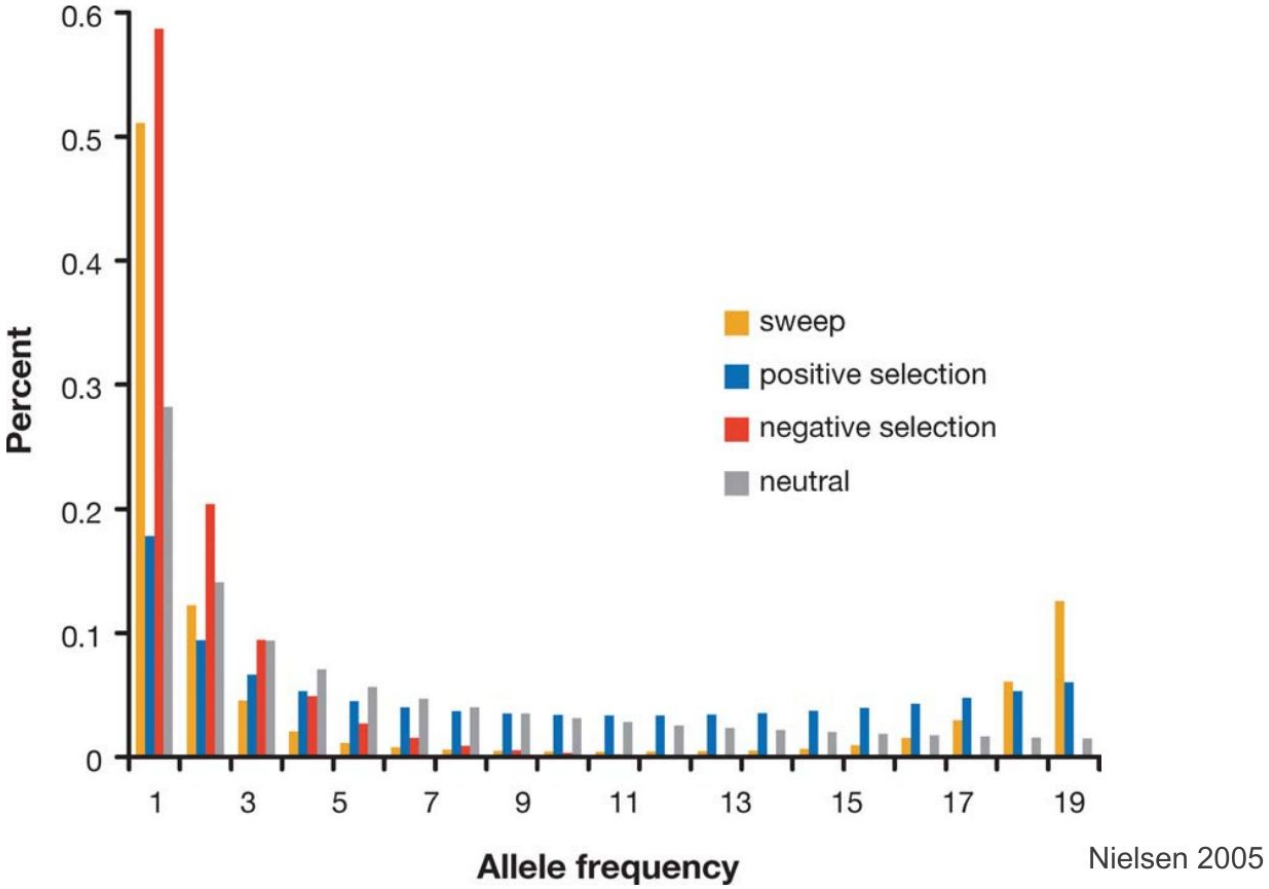
$P(\theta|X)$

Call SNPs

	0	1	2	3	4	.	.	.	2n
Site1	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000
Site2	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000
Site3	0.0000	0.0000	0.0000	0.9999	0.0000				0.0000
Site4	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000
.									
.									
.									
.									
.									
.									
Sitek	0.9999	0.0000	0.0000	0.0000	0.0000				0.0000

99.99% probability of allele frequency 3/2n

What about allele frequency posterior probabilities?



Nielsen 2005