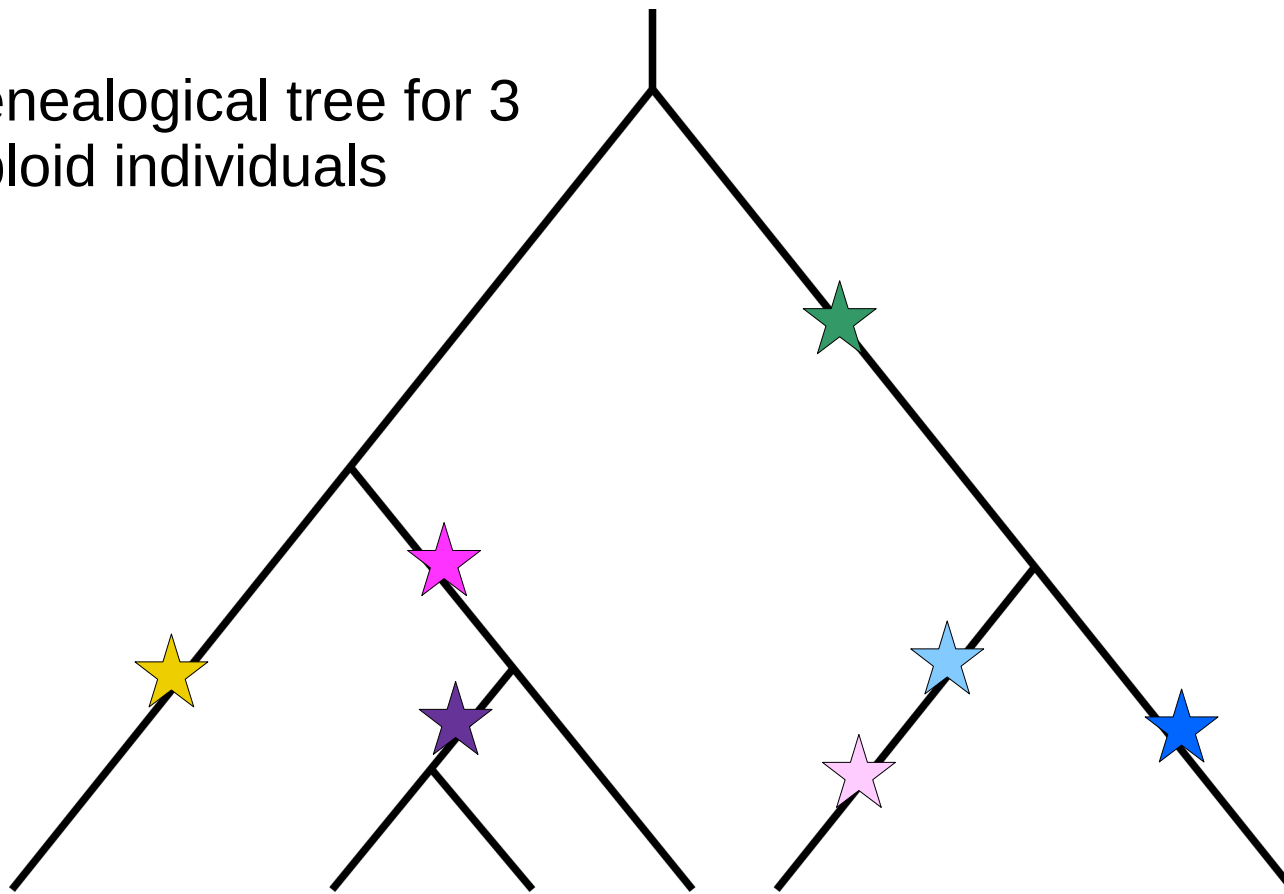


Site frequency spectrum (SFS)

Genealogical tree for 3
diploid individuals

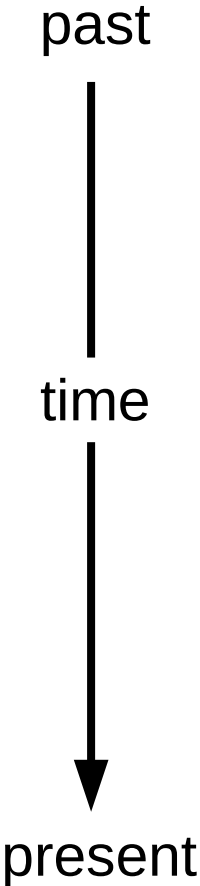
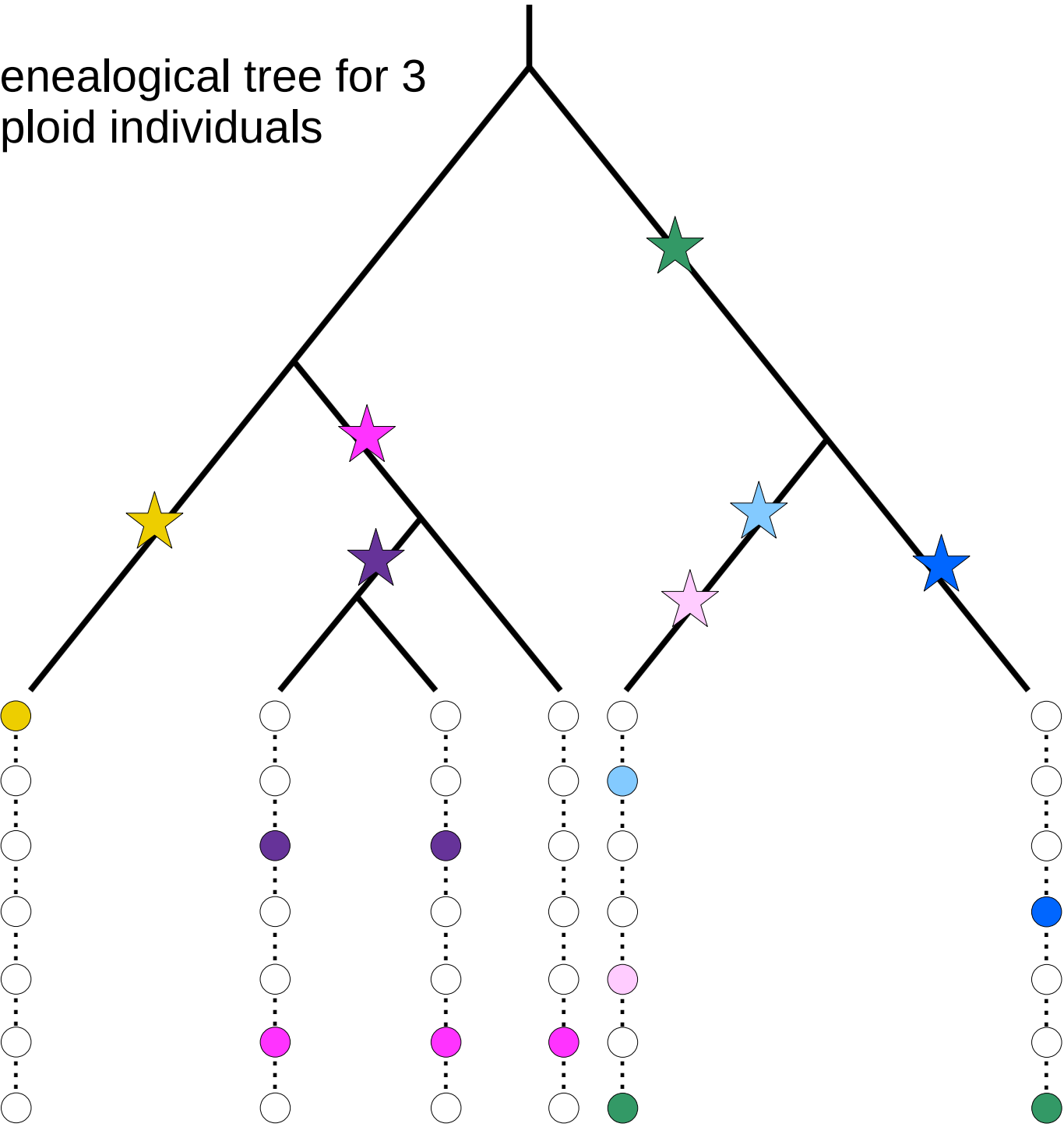


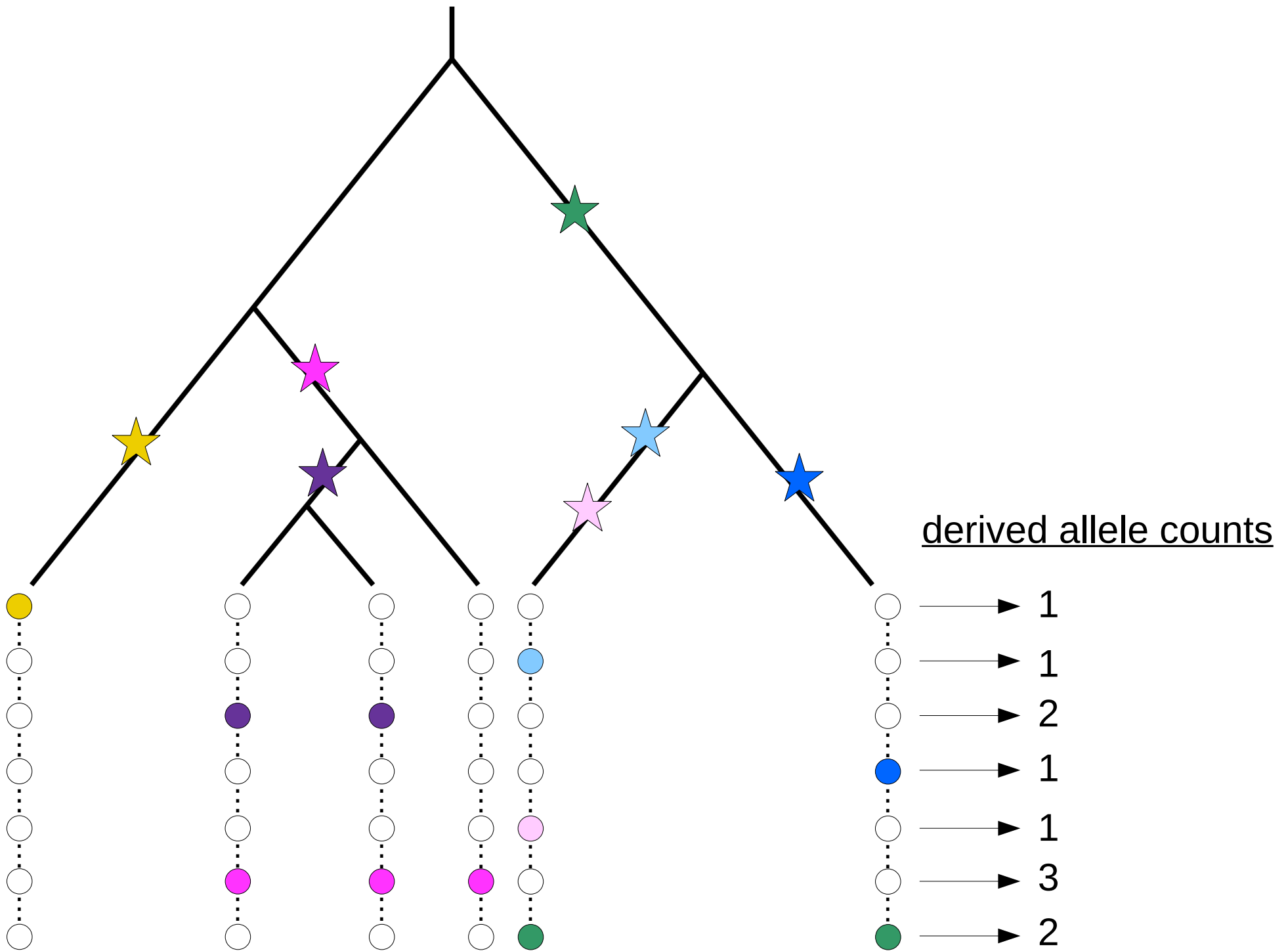
past

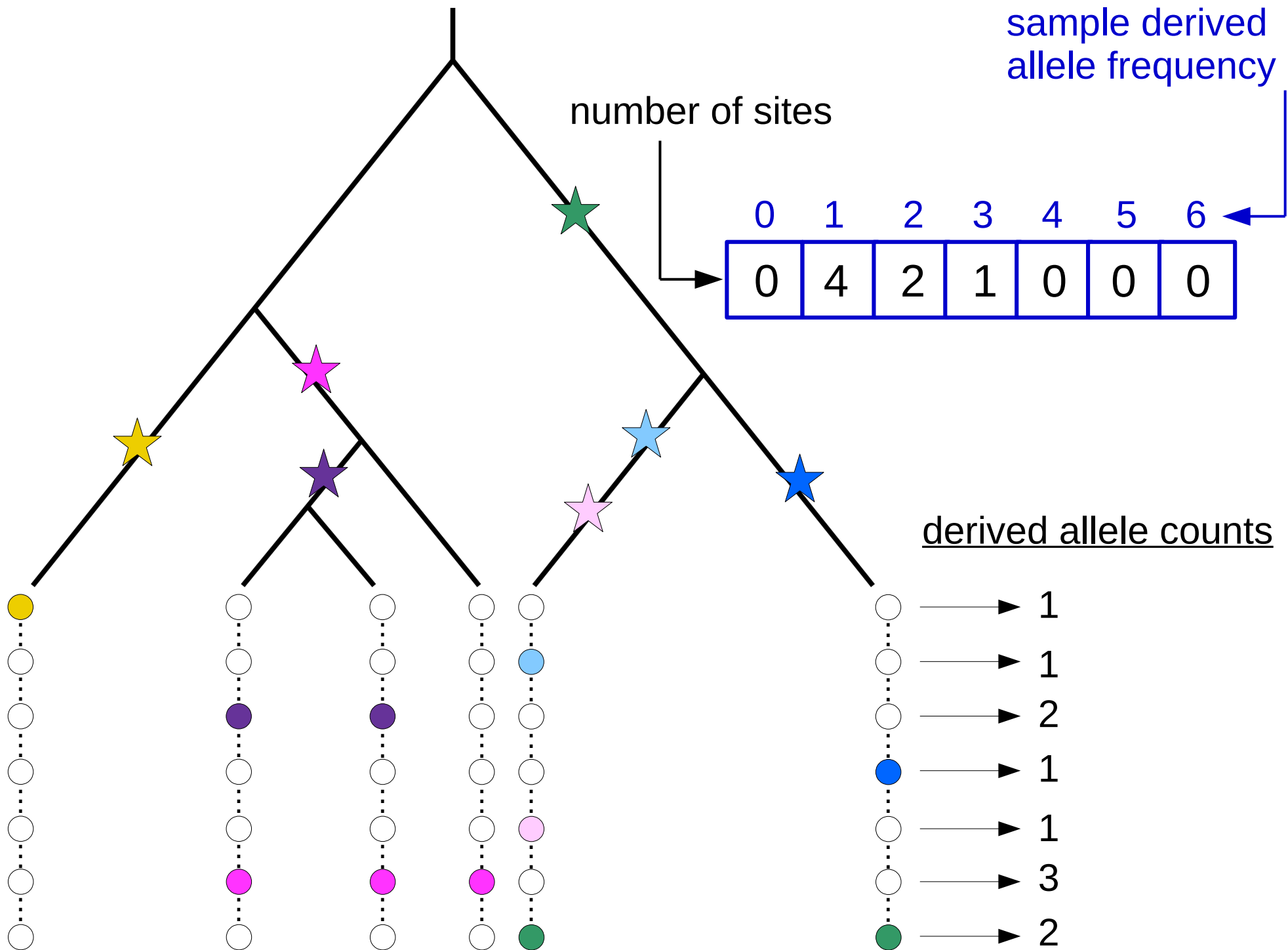
time

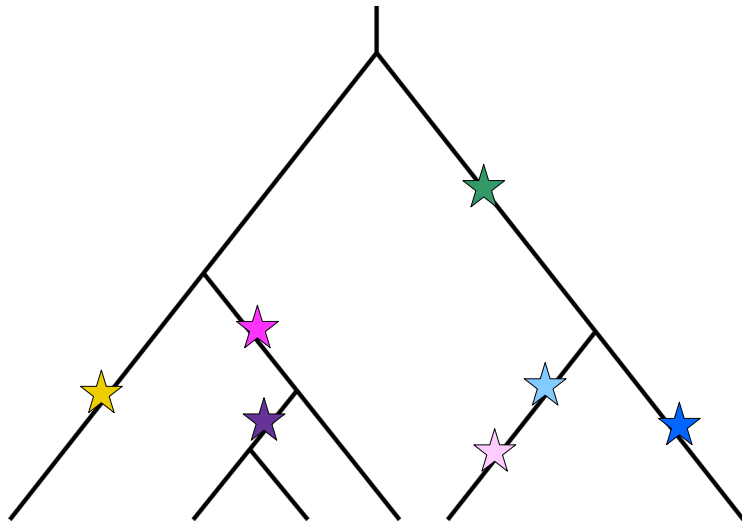
present

Genealogical tree for 3
diploid individuals



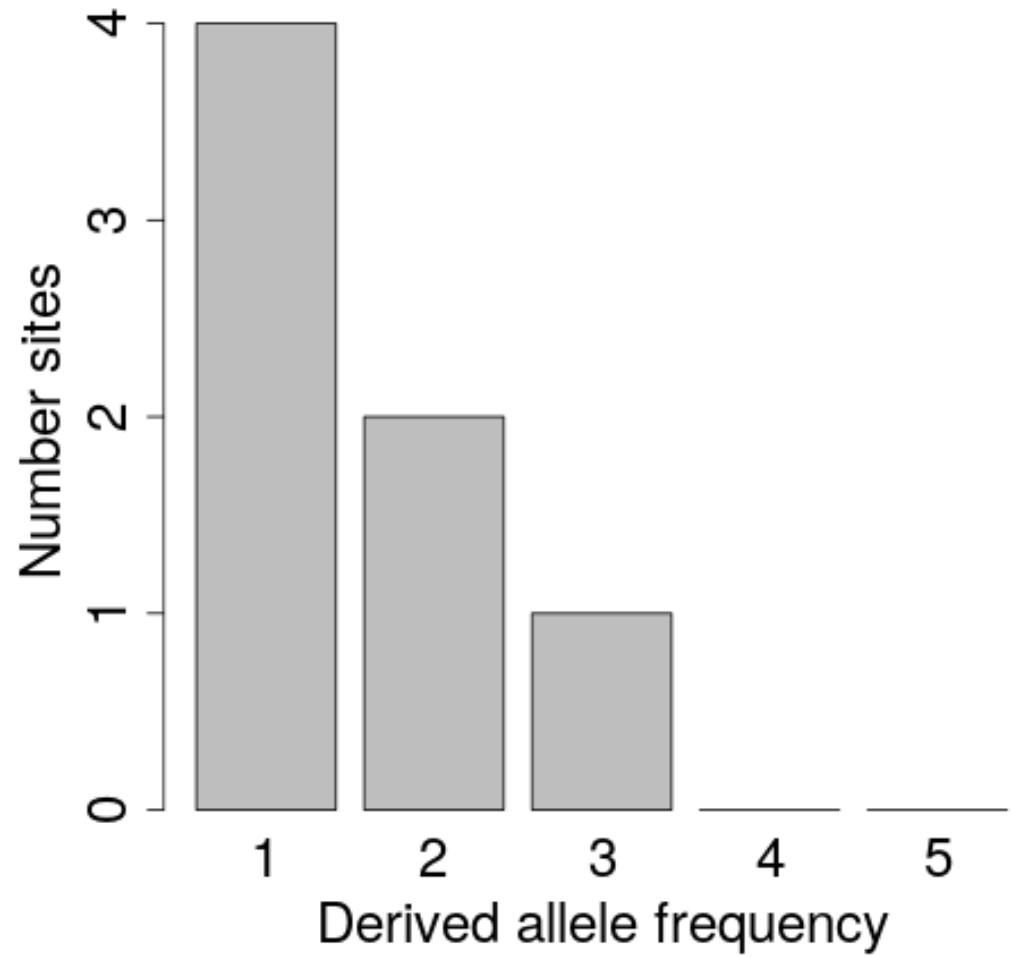
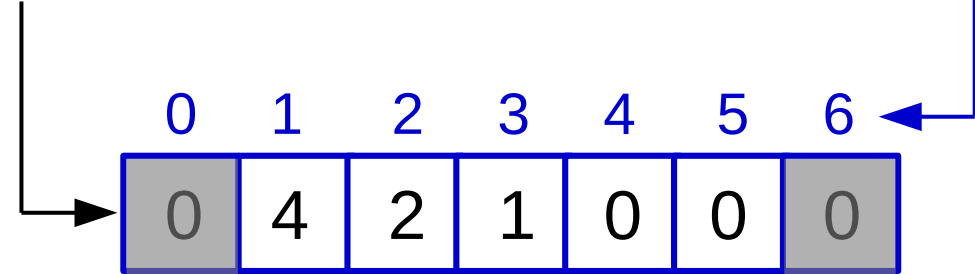




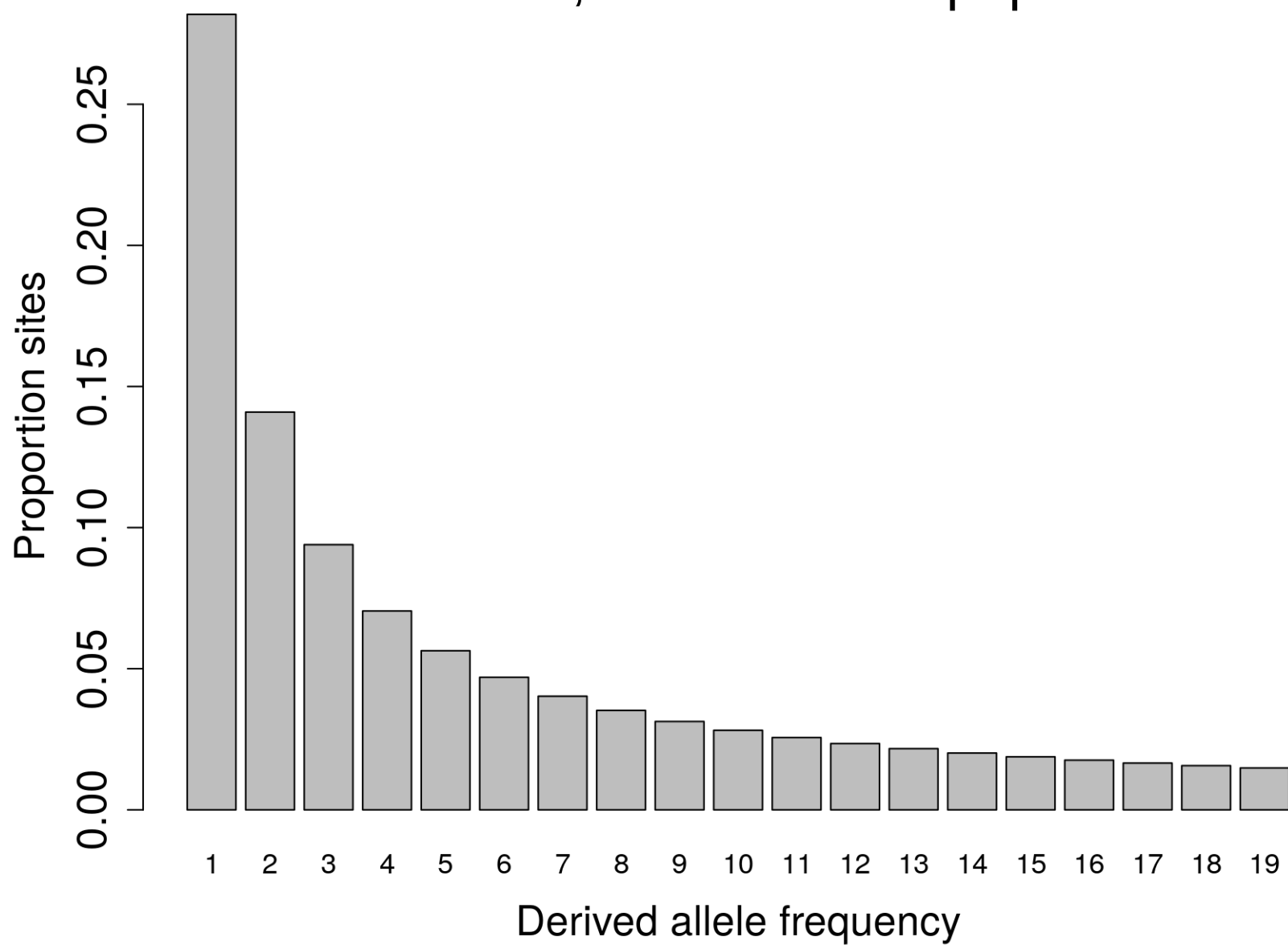


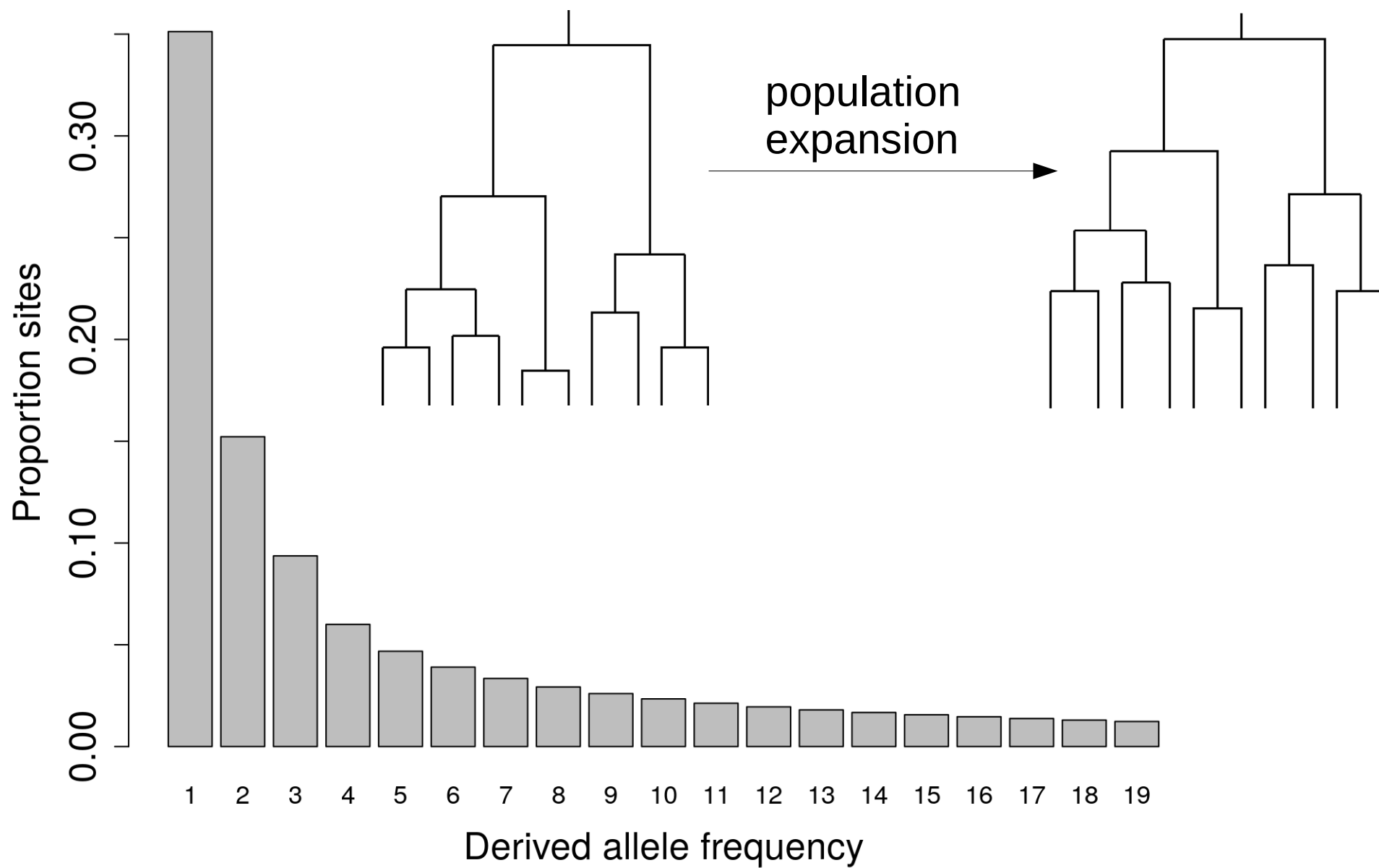
number of sites

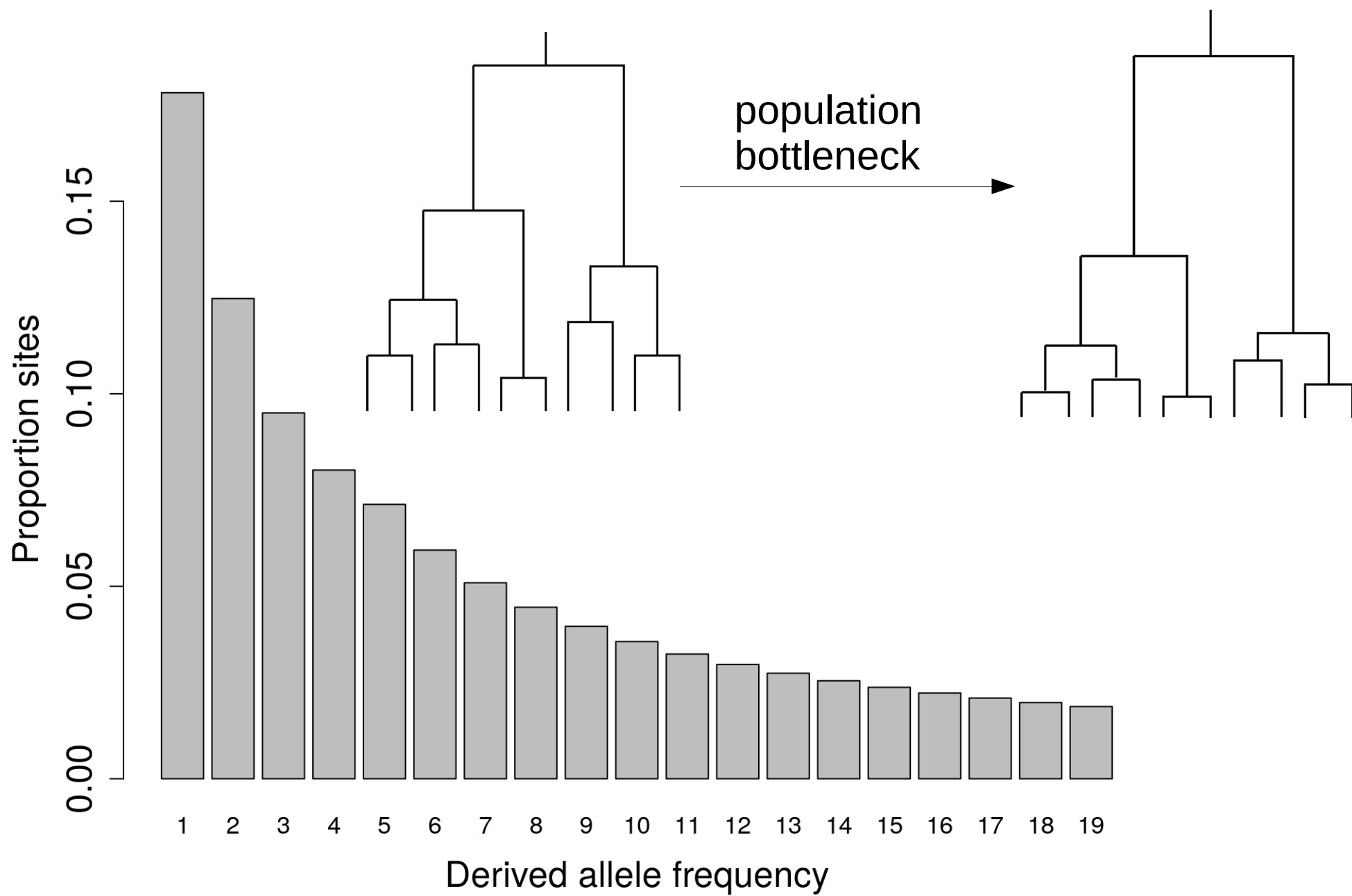
sample derived
allele frequency



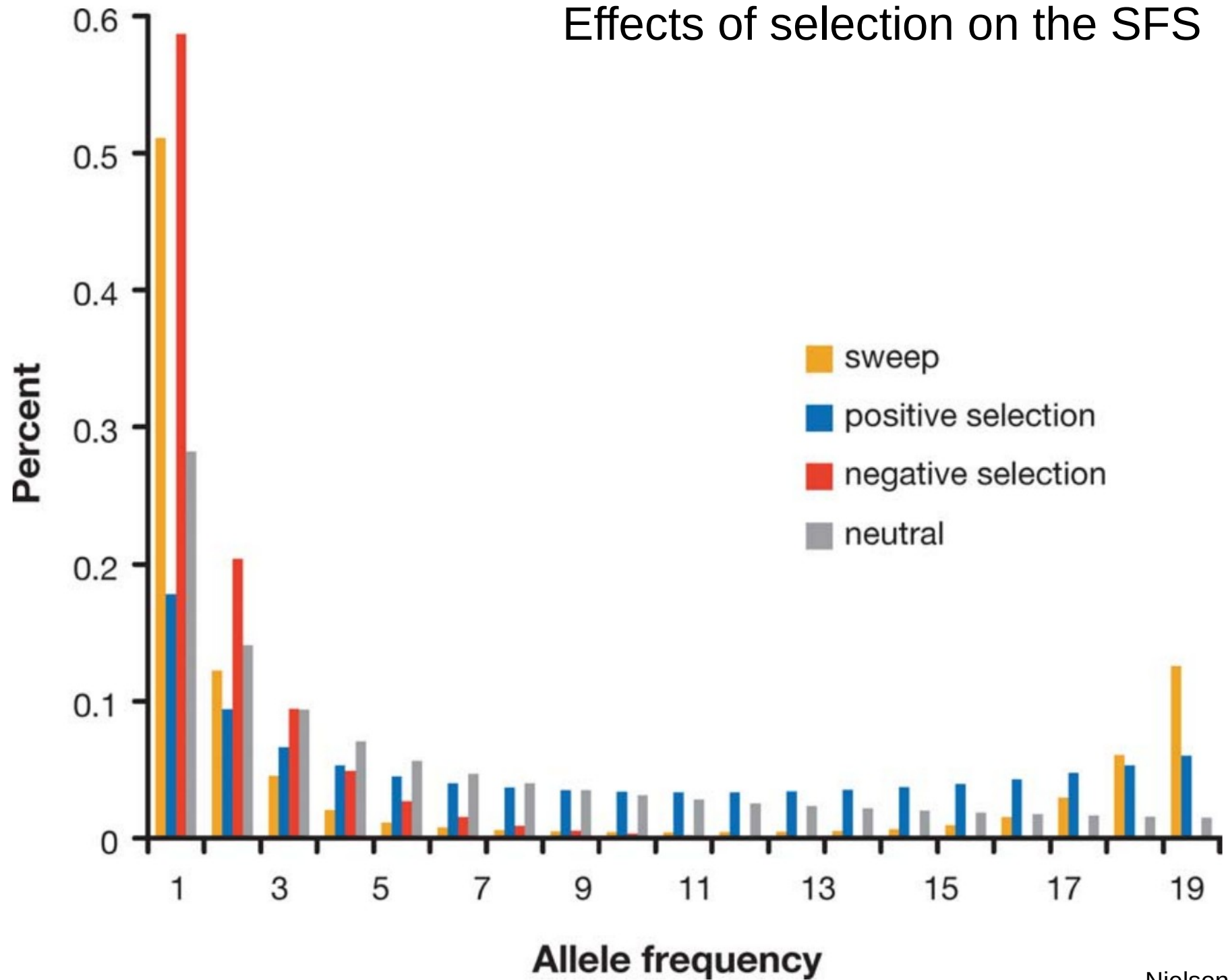
Neutral, constant-size population SFS








Effects of selection on the SFS



SFS for quality control game

$$\text{SFS} = \mathbf{P} = (p_0, p_1, p_2, p_3, p_4, \dots, p_{2n})$$


p_j : proportion of sites in the genome with j derived alleles
 n = diploid sample size

Likelihood of the SFS (\mathbf{P}) assuming we know the genotypes:

X : observed data
 (sequencing reads)

G : genotype vector = $(G_1, G_2, G_3, \dots, G_n)$

$$P(X, \mathbf{G} | \mathbf{P}) = \sum_{j=0}^{2n} p(X, G | s_d = j) p(s_d = j | \mathbf{P})$$

s_d : number of derived alleles in genotype vector, G

Likelihood of the SFS (\mathbf{P}) assuming we know the genotypes:

$$P(X, \mathbf{G} | \mathbf{P}) = \sum_{j=0}^{2n} \boxed{p(X, \mathbf{G} | s_d = j)} \underbrace{\boxed{p(s_d = j | \mathbf{P})}}_{= p_j}$$

$p(a, b | c) = p(a | b) p(b | c)$ Recall that SFS = $\mathbf{P} = (p_0, p_1, \dots, p_{2n})$

$$= \sum_{j=0}^{2n} \boxed{p(X | \mathbf{G}) p(\mathbf{G} | s_d = j)} p(s_d = j | \mathbf{P})$$

Likelihood of the SFS (\mathbf{P}) assuming we know the genotypes:

$$\begin{aligned}
 P(X, \mathbf{G} | \mathbf{P}) &= \sum_{j=0}^{2n} \boxed{p(X, \mathbf{G} | s_d = j)} \underbrace{\boxed{p(s_d = j | \mathbf{P})}}_{= p_j} \\
 &\quad p(a, b | c) = p(a | b) p(b | c) \quad \text{Recall that SFS} = \mathbf{P} = (p_0, p_1, \dots, p_{2n}) \\
 &= \sum_{j=0}^{2n} \boxed{p(X | \mathbf{G}) p(\mathbf{G} | s_d = j)} p(s_d = j | \mathbf{P}) \\
 &= \sum_{j=0}^{2n} \boxed{\left(\prod_{i=1}^n p(X | G_i) \right) p(\mathbf{G} | s_d = j)} p(s_d = j | \mathbf{P})
 \end{aligned}$$

Likelihood of the SFS (\mathbf{P}) assuming we know the genotypes:

$$P(\mathbf{X}, \mathbf{G} | \mathbf{P}) = \sum_{j=0}^{2n} \left(\prod_{i=1}^n p(\mathbf{X} | G_i) \right) p(\mathbf{G} | s_d = j) p(s_d = j | \mathbf{P})$$

Probability based on the number of ways to have j derived alleles in \mathbf{G} out of the total number of ways to arrange j derived alleles in $2n$ chromosomes (i.e. $\text{choose}(2n, j)$).

Assumes HWE

Example with 4 diploid individual with derived allele T.

$j = 2$, $k = \# \text{ heterozygotes} = 2$

CC	CT	TT	CT
CC	TC	TT	CT
CC	CT	TT	TC
CC	TC	TT	TC

$$p(\mathbf{G} | s_d = j) = \frac{2^k}{\binom{2n}{j}} = \frac{2^2}{\binom{8}{2}} = \frac{4}{28}$$

$2^j = 2^2 = 4$ combinations

$$P(X, \mathbf{G} | \mathbf{P}) = \sum_{j=0}^{2n} p(X, \mathbf{G} | s_d = j) p(s_d = j | \mathbf{P})$$

Allow for unknown genotypes
(need to use genotype likelihoods)



$$P(X | \mathbf{P}) =$$

$$\sum_{j=0}^{2n} p(s_d = j | \mathbf{P}) \sum_{G_1 \in \{0,1,2,\}} \cdots \sum_{G_d \in \{0,1,2,\}} \prod_{d=1}^n p(X_d | G_d)$$

Likelihood of \mathbf{P} for a site \rightarrow these are the allele frequency likelihoods at a site (-doSaf 1):

$$P(X^v | \mathbf{P}) =$$

$$\sum_{j=0}^{2n} p(s_d = j | \mathbf{P}) \sum_{G_1 \in \{0,1,2,\}} \cdots \sum_{G_d \in \{0,1,2,\}} \prod_{d=1}^n p(X_d | G_d)$$

Assume sites are independent

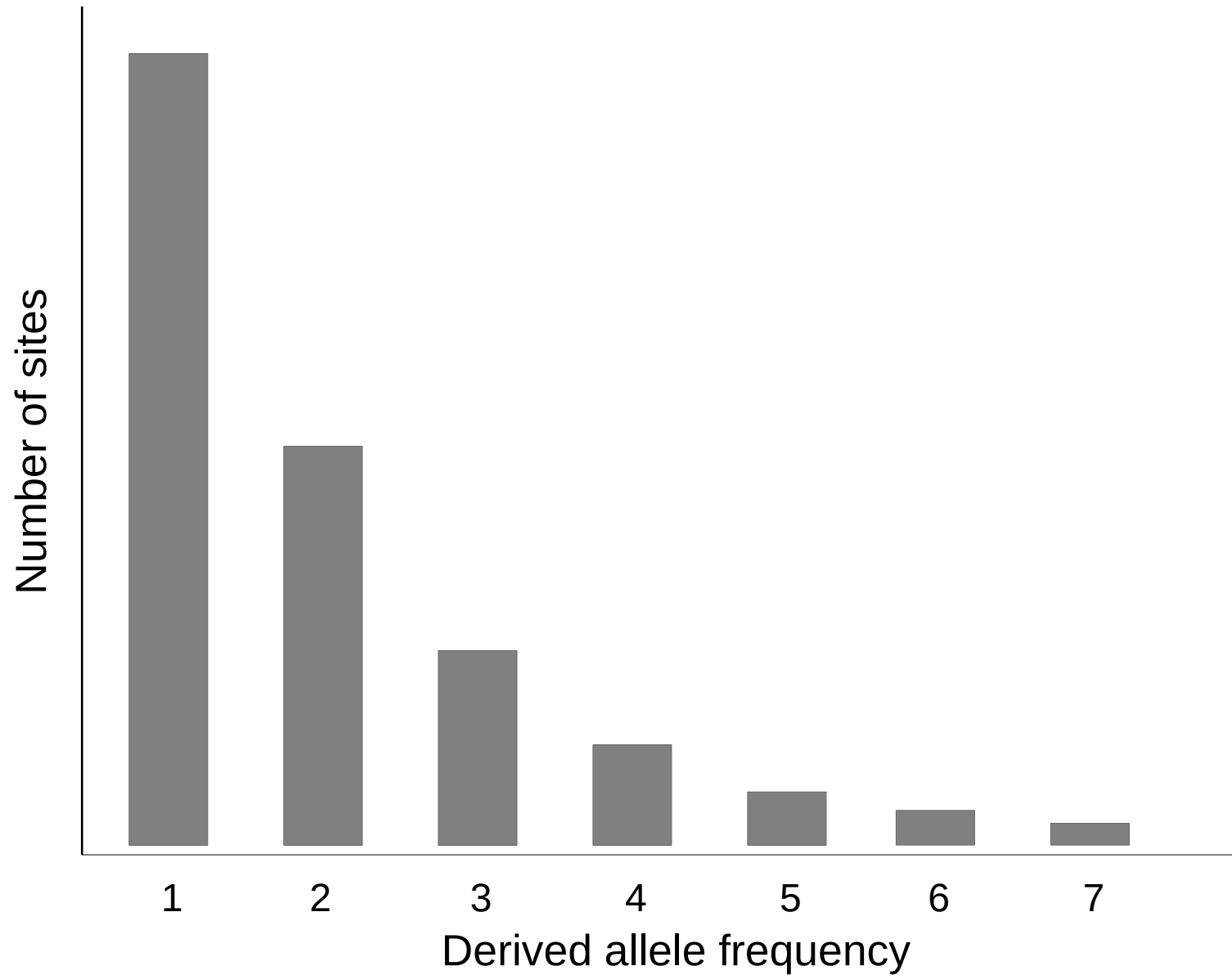


Likelihood of genome-wide \mathbf{P} :

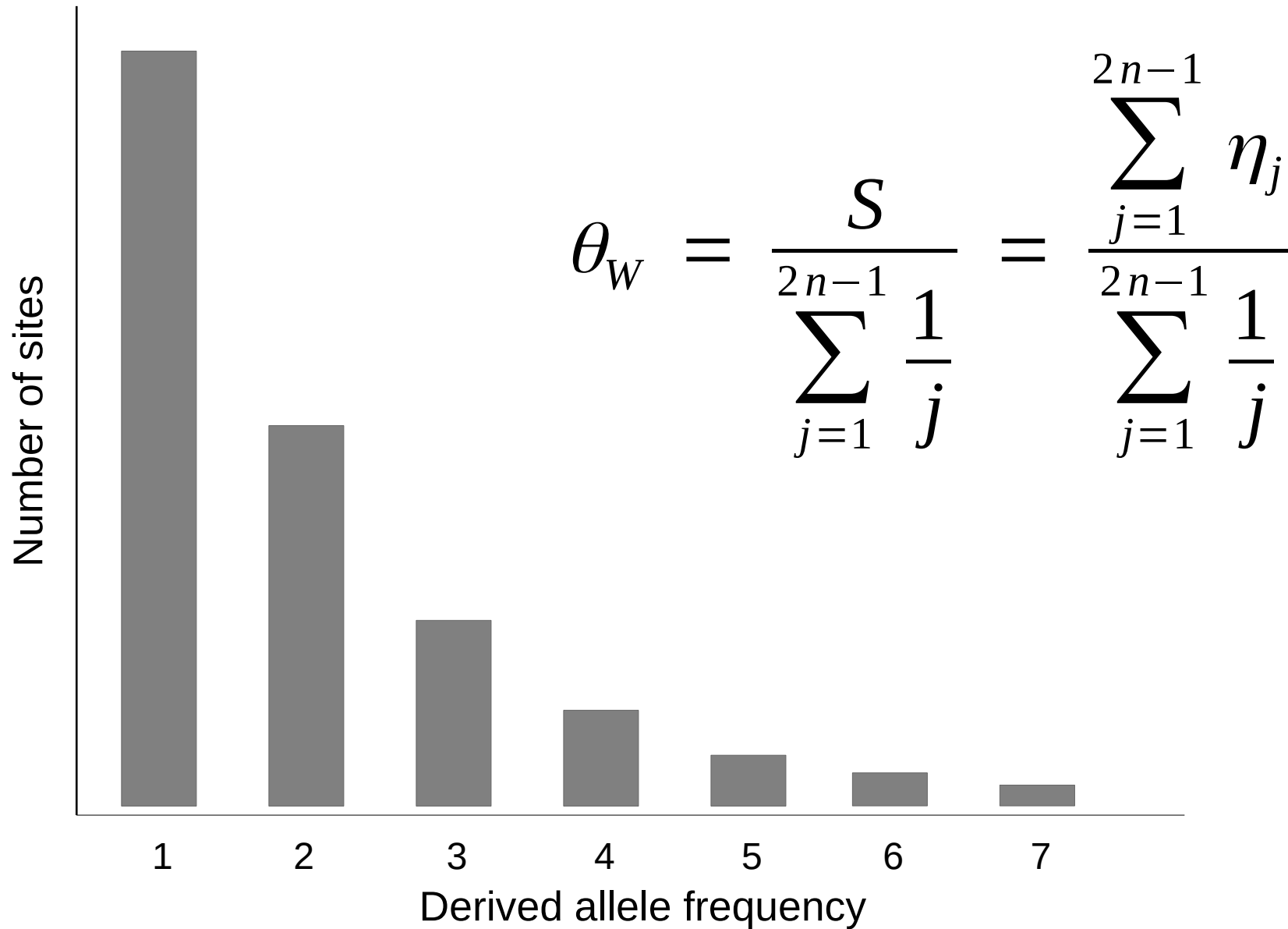
$$P(X | \mathbf{P}) =$$

$$\prod_{v=1}^{\text{all sites}} \sum_{j=0}^{2n} p(s_d = j | \mathbf{P}) \sum_{G_1 \in \{0,1,2,\}} \cdots \sum_{G_d \in \{0,1,2,\}} \prod_{d=1}^n p(X_d^v | G_d^v)$$

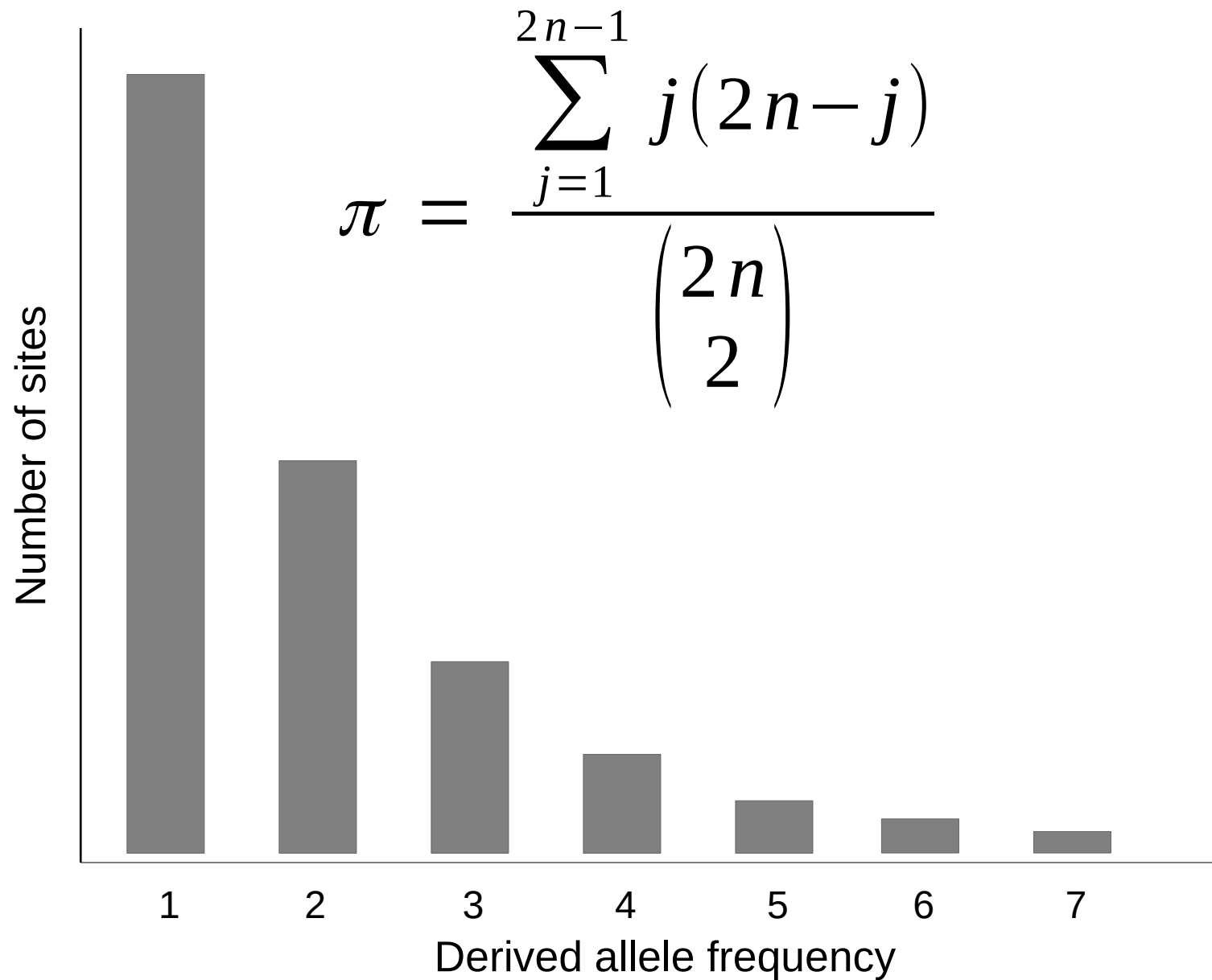
Folding the SFS



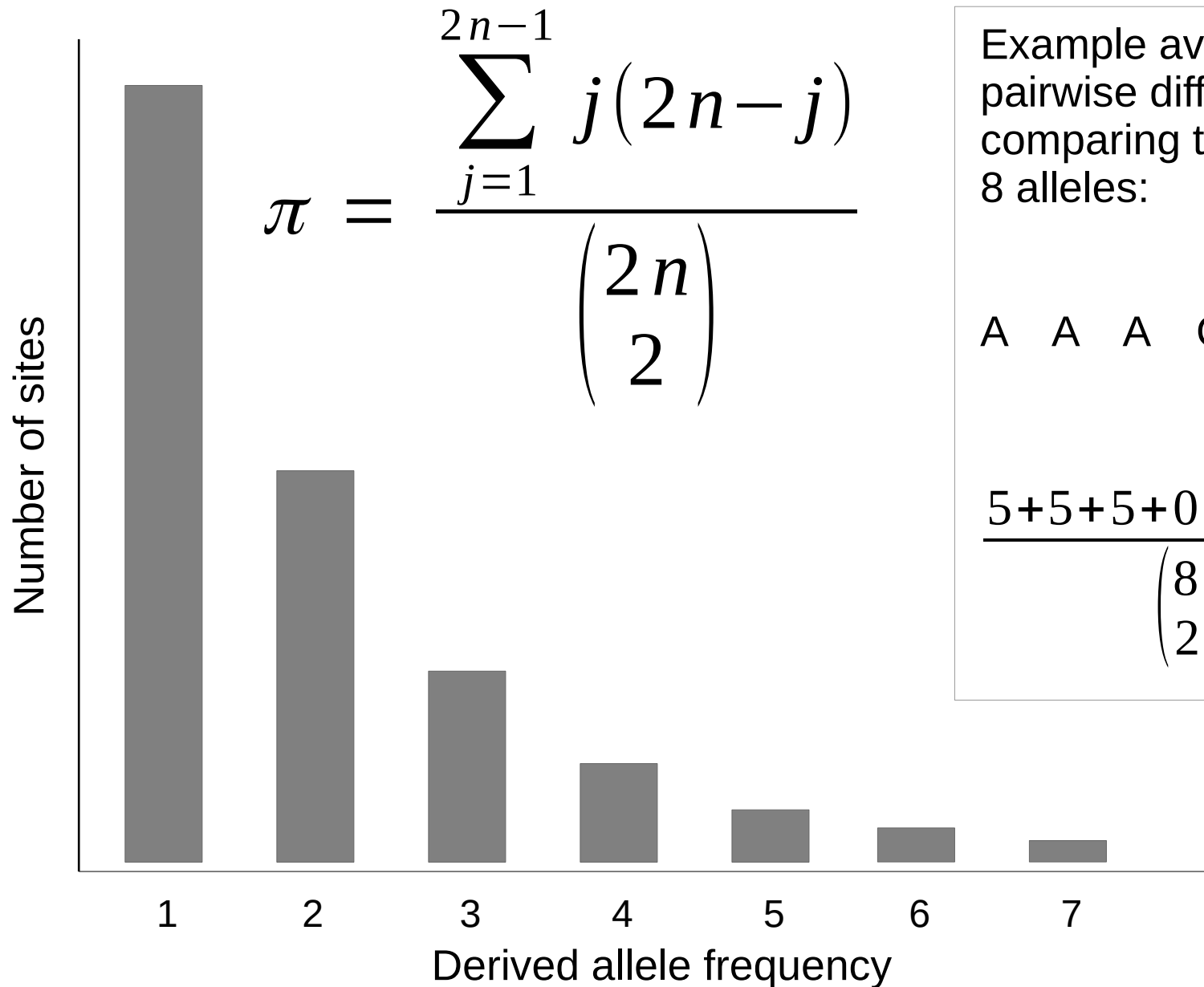
Estimating diversity from the SFS



Estimating diversity from the SFS



Estimating diversity from the SFS



Example average number of pairwise differences comparing the following set of 8 alleles:

A A A C C C C C

$$\frac{5+5+5+0+0+0+0}{\binom{8}{2}} = \frac{15}{28}$$

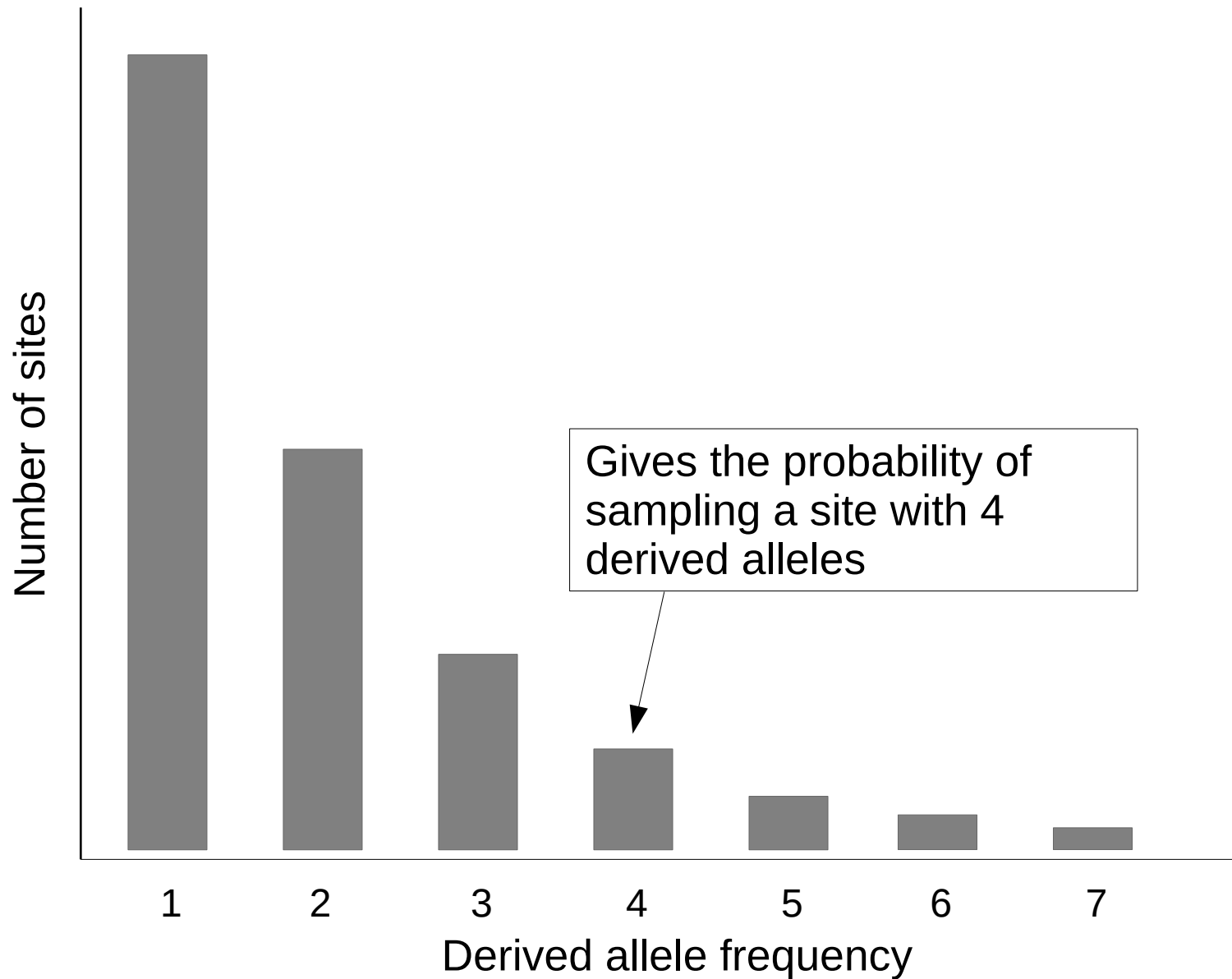
How to estimate posterior probabilities of allele frequencies

-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

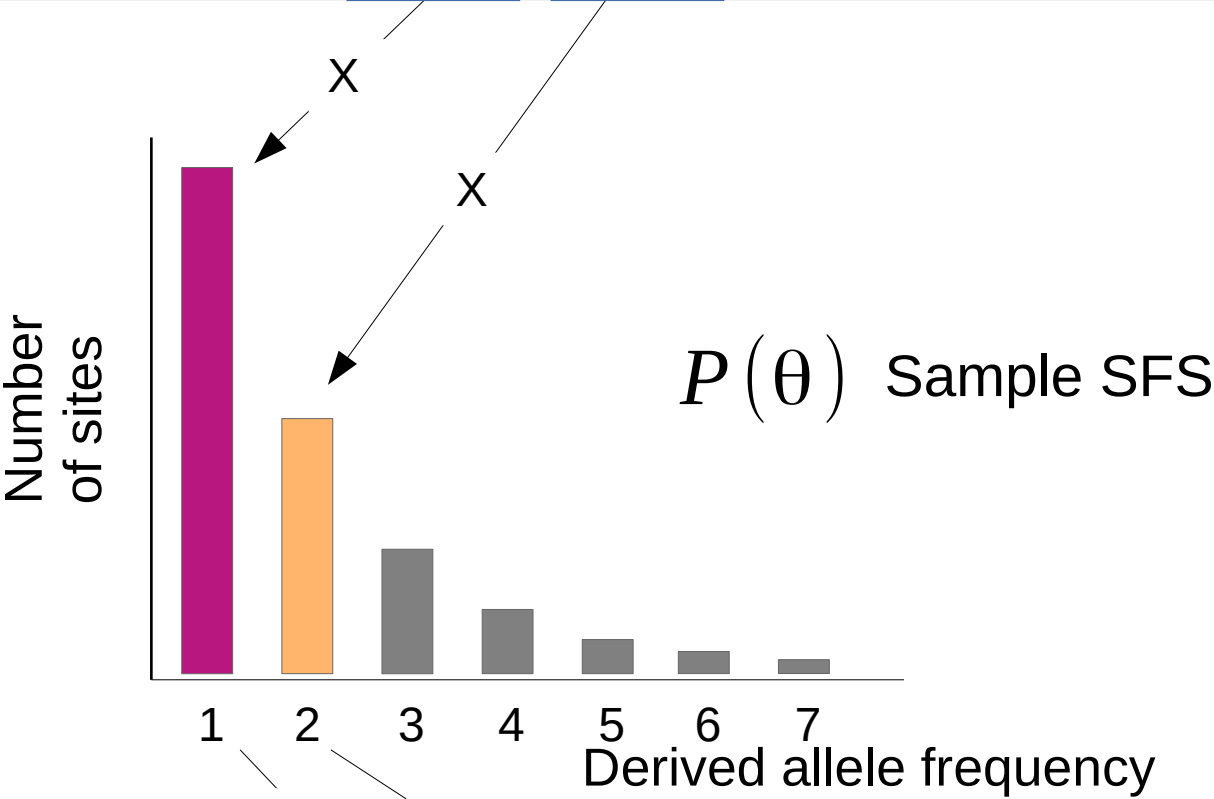
$$P(\theta|X) = P(X|\theta)P(\theta)$$

How to estimate posterior probabilities of allele frequencies



$$P(\theta|X) = P(X|\theta)P(\theta)$$

	0	1	2	3	4	.	.	.	2n	
Site1	0.00	-2.24	-4.53	-6.99	-9.63				-232.69	$P(X \theta)$
Site2	0.00	-2.24	-4.53	-6.99	-9.63				-232.69	



	0	1	2	3	4	.	.	.	2n	
Site1	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000	$P(\theta X)$
Site2	0.9892	0.0101	0.0006	0.0000	0.0000				0.0000	

2-dimensional SFS

		Population j derived allele frequency						
		0	1	2	3	4	...	2n
Population i derived allele frequency	0							
	1							
	2							
	3							
	4							
	⋮							
	2n							

$$F_{ST} = \frac{E(a|X)}{E(a|X)+E(b|X)}$$

		Population j derived allele frequency						
		0	1	2	3	4	...	2n
Population i derived allele frequency	0							
	1							
	2							
	3							
	4							
	⋮							
	2n							

$$E(a|X)=\sum_{k=0}^{2n}\sum_{z=0}^{2n}a_{pop\,i,pop\,j}^{k,z}P(X_{i,j}|s_d=k)P(X_{i,j}|s_d=z)Q_{i,j}^{k,z}$$

$$E(b|X)=\sum_{k=0}^{2n}\sum_{z=0}^{2n}b_{pop\,i,pop\,j}^{k,z}P(X_{i,j}|s_d=k)P(X_{i,j}|s_d=z)Q_{i,j}^{k,z}$$