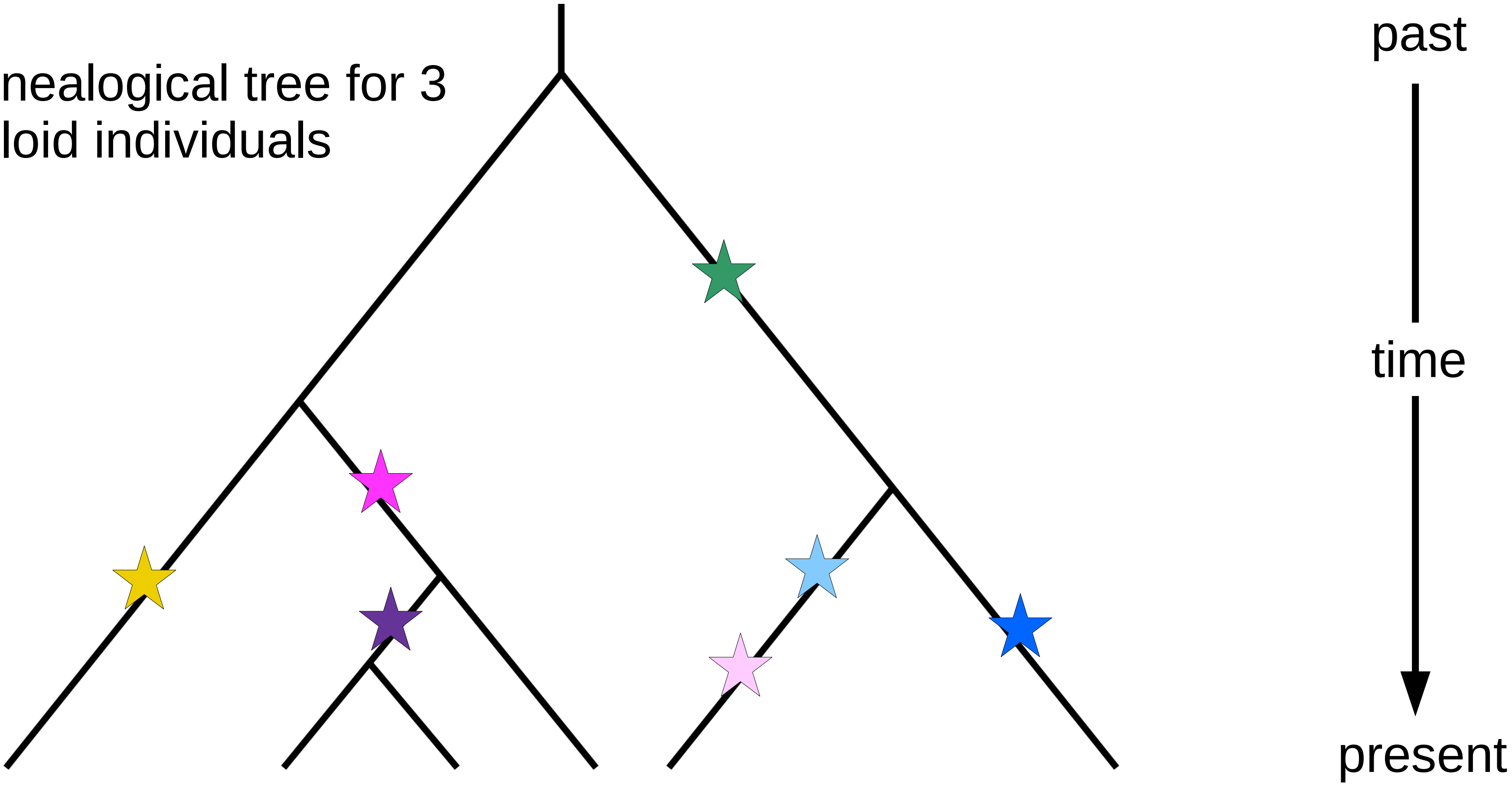
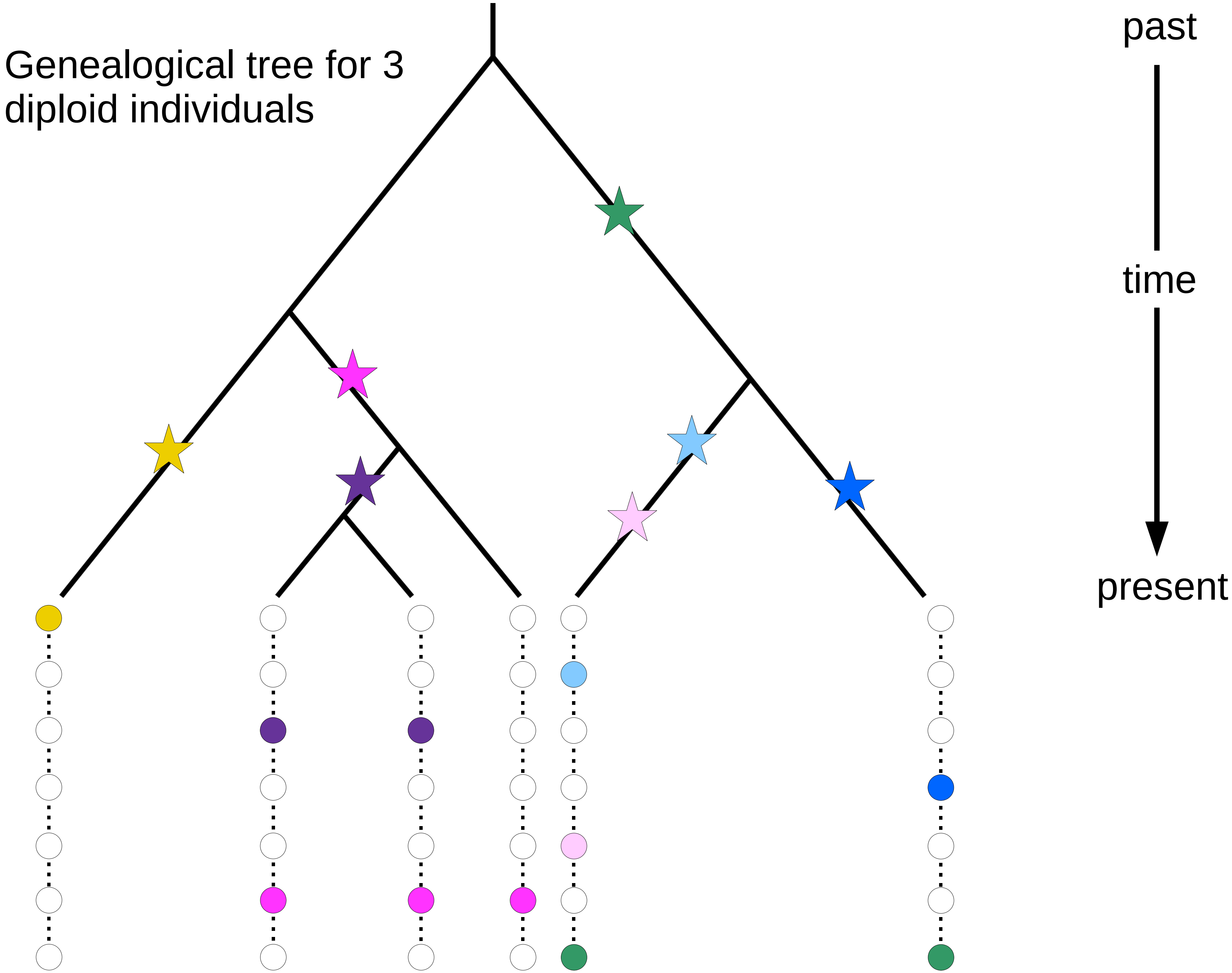
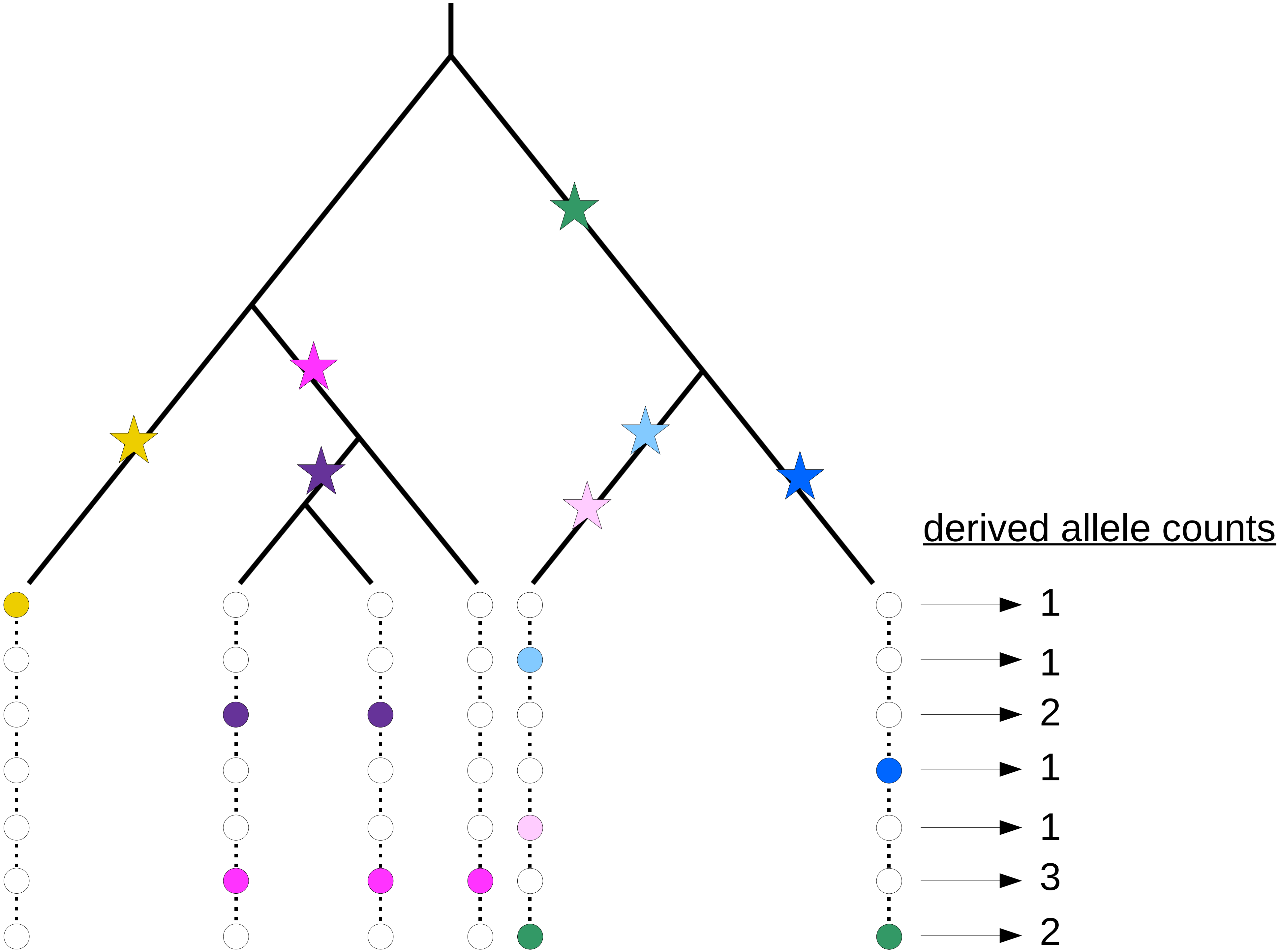
Site Frequency Spectrum (SFS)

Tyler Linderoth
Physalia lcWGS course 2024

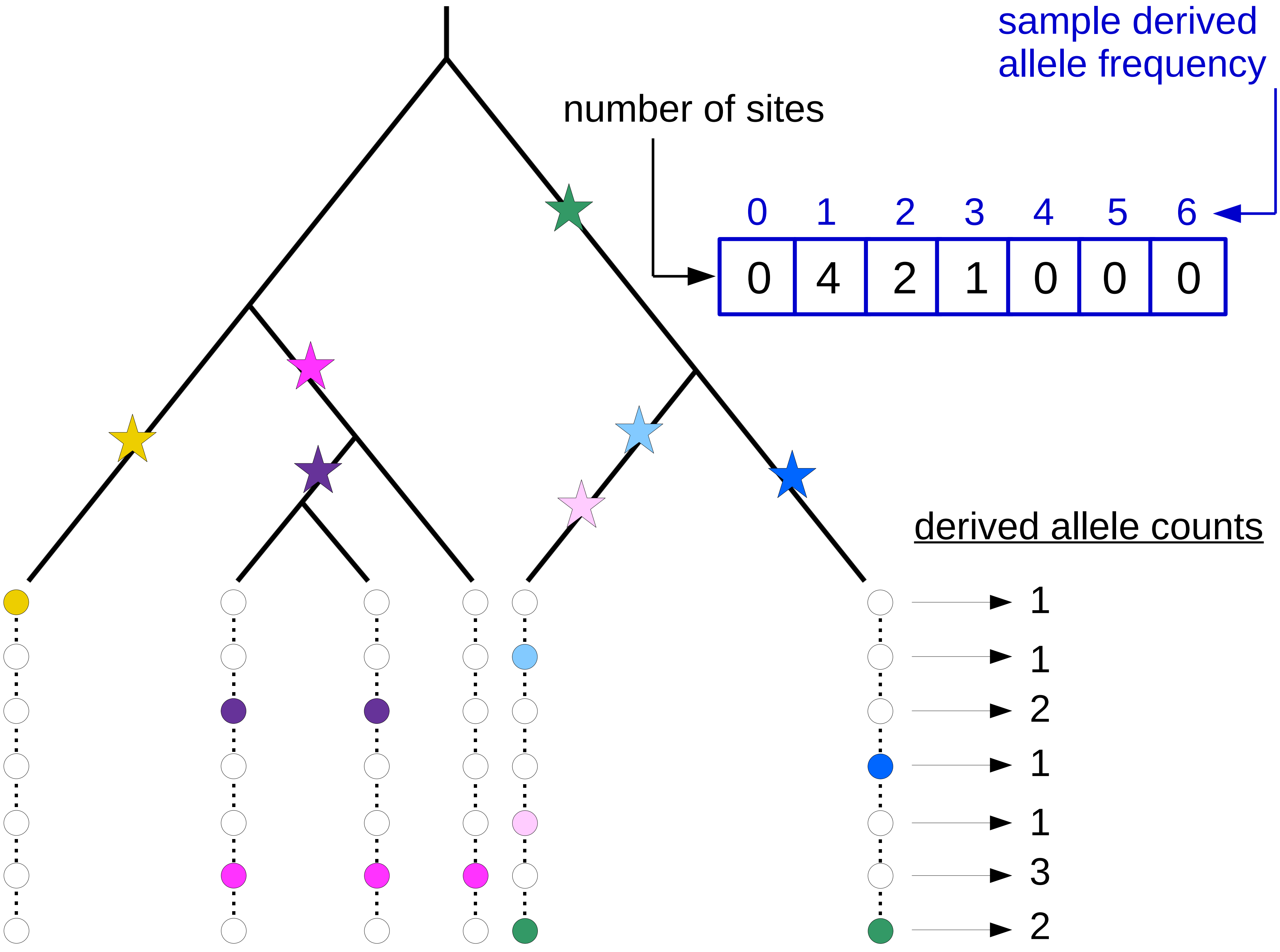
Genealogical tree for 3 diploid individuals



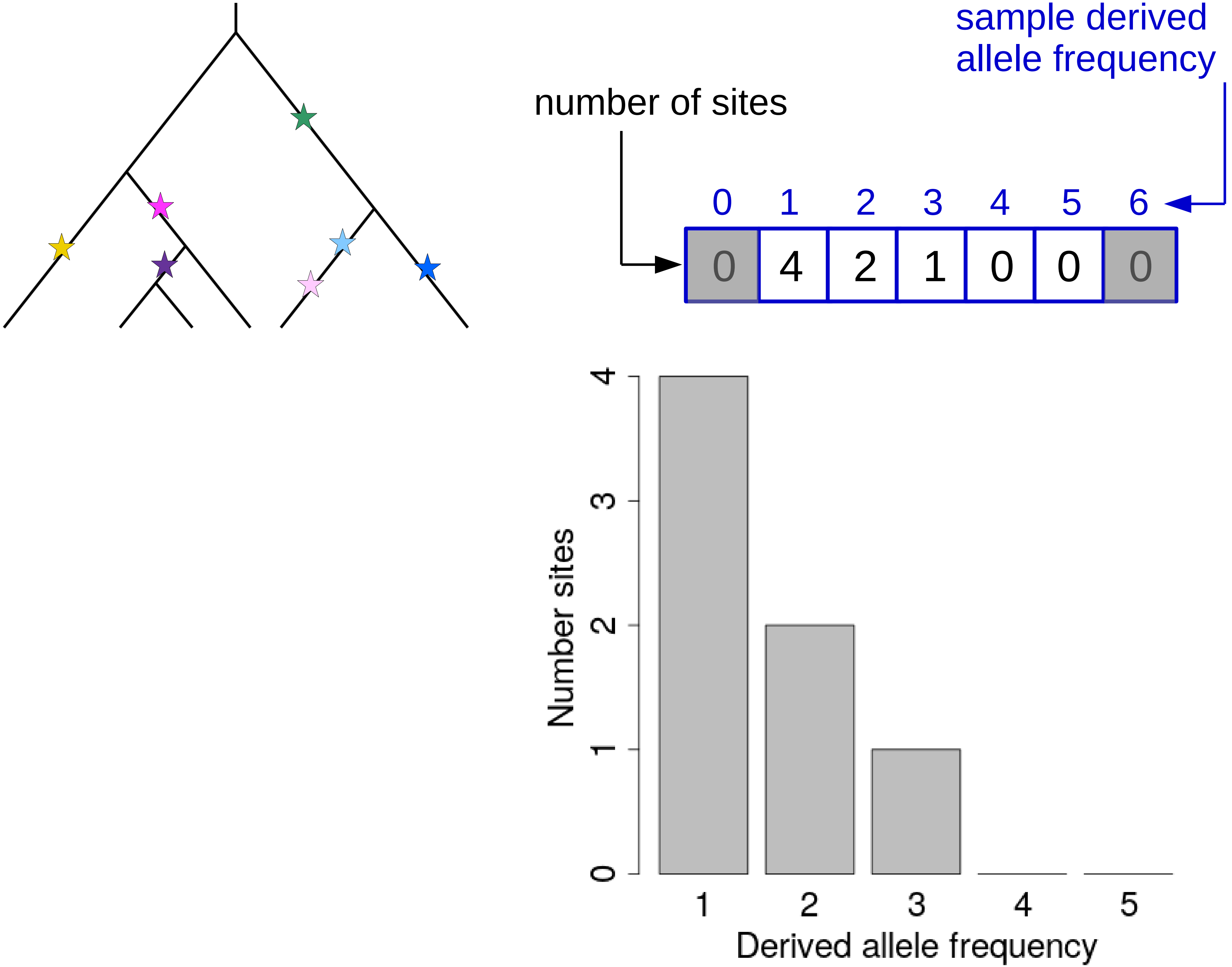

Genealogical tree for 3 diploid individuals
past
time
present



derived allele counts
1
1
2
1
1
3
2



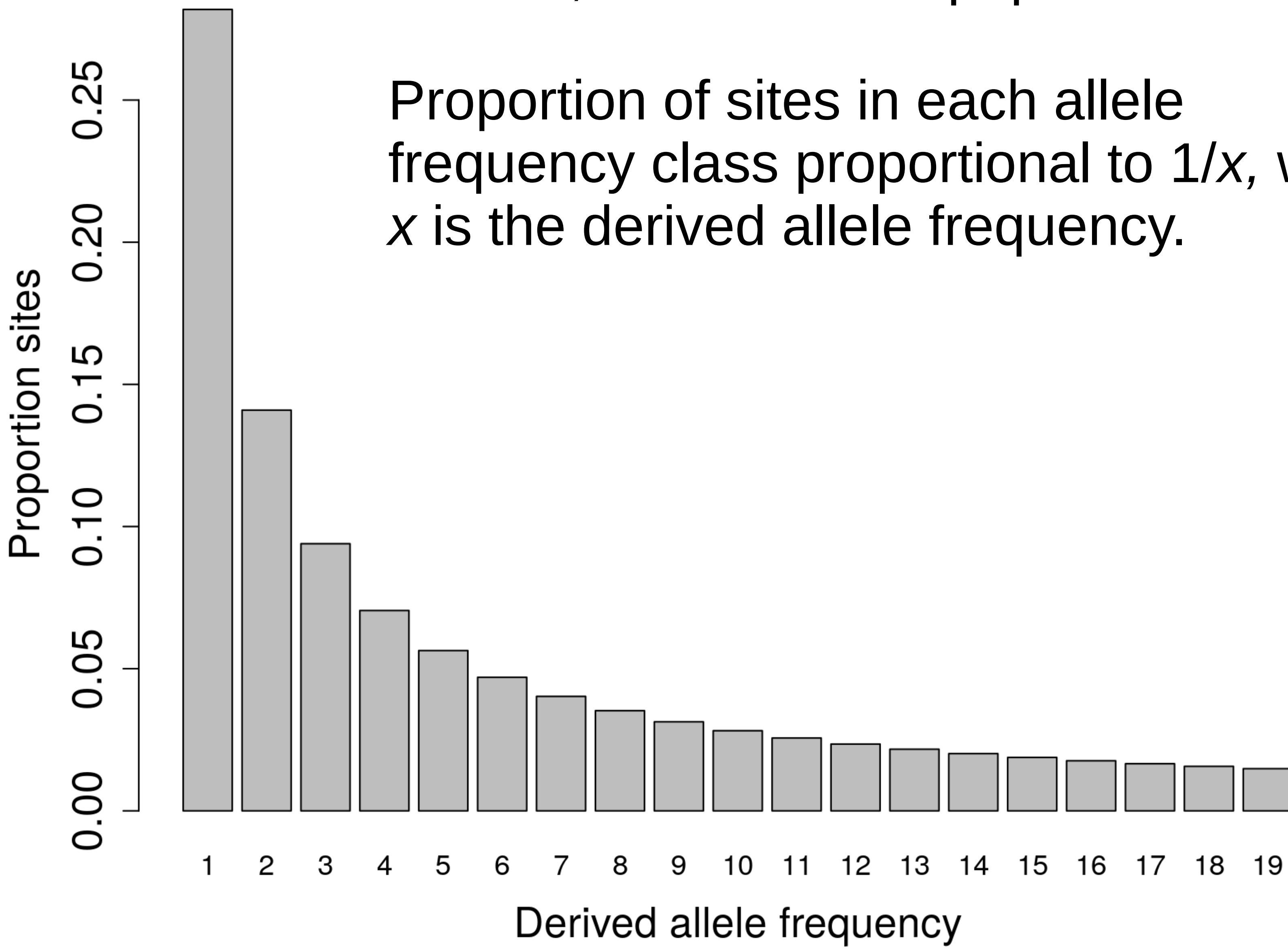
sample derived
allele frequency
number of sites
0 1 2 3 4 5 6
0 4 2 1 0 0 0
derived allele counts
1
1
2
1
1
3
2



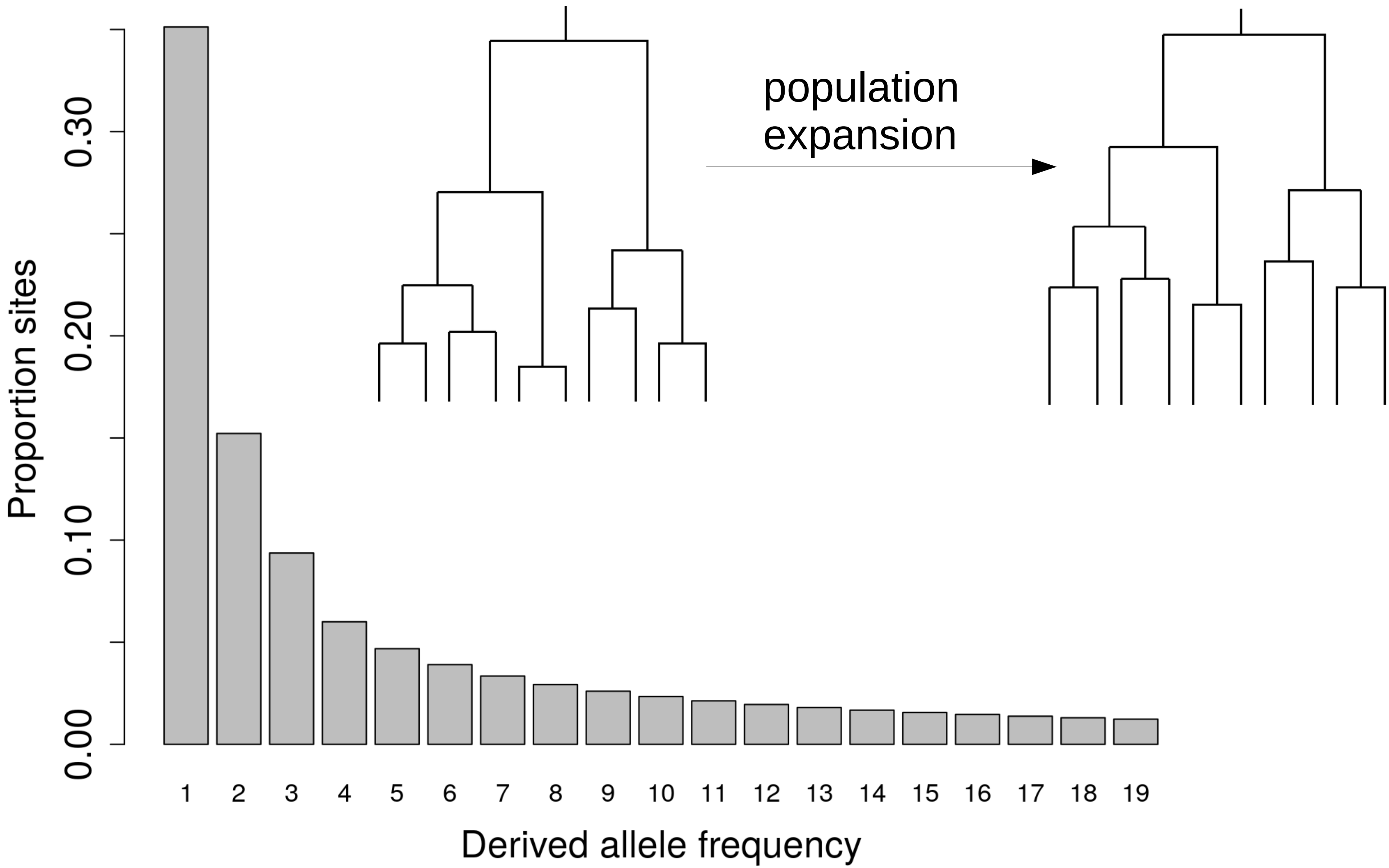
sample derived
allele frequency
number of sites
0 1 2 3 4 5 6
0 4 2 1 0 0 0
Number sites
0 1 2 3 4
1 2 3 4 5
Derived allele frequency

Neutral, constant-size population SFS

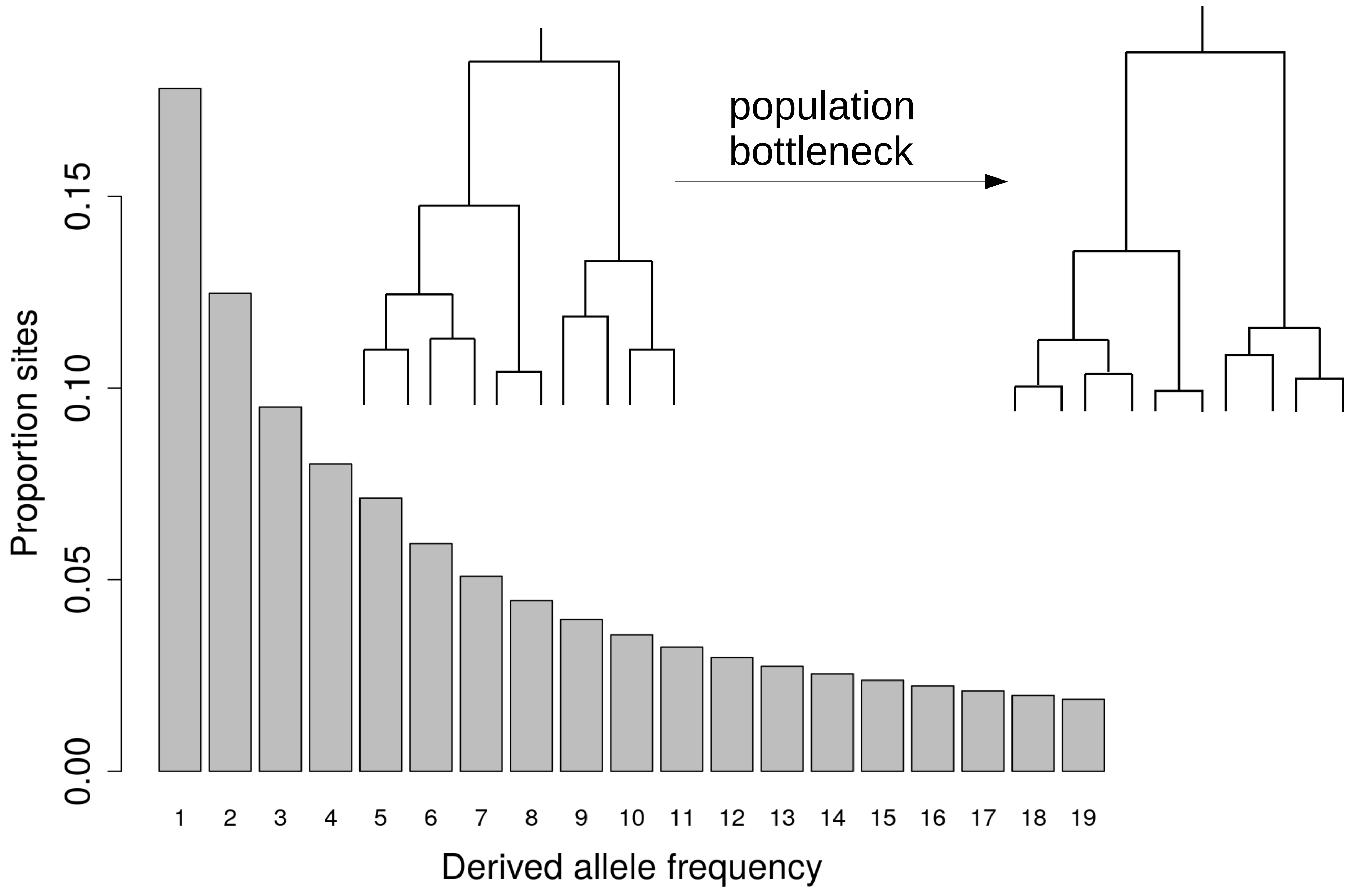
Proportion of sites in each allele frequency class proportional to 1/*x*, where *x* is the derived allele frequency.



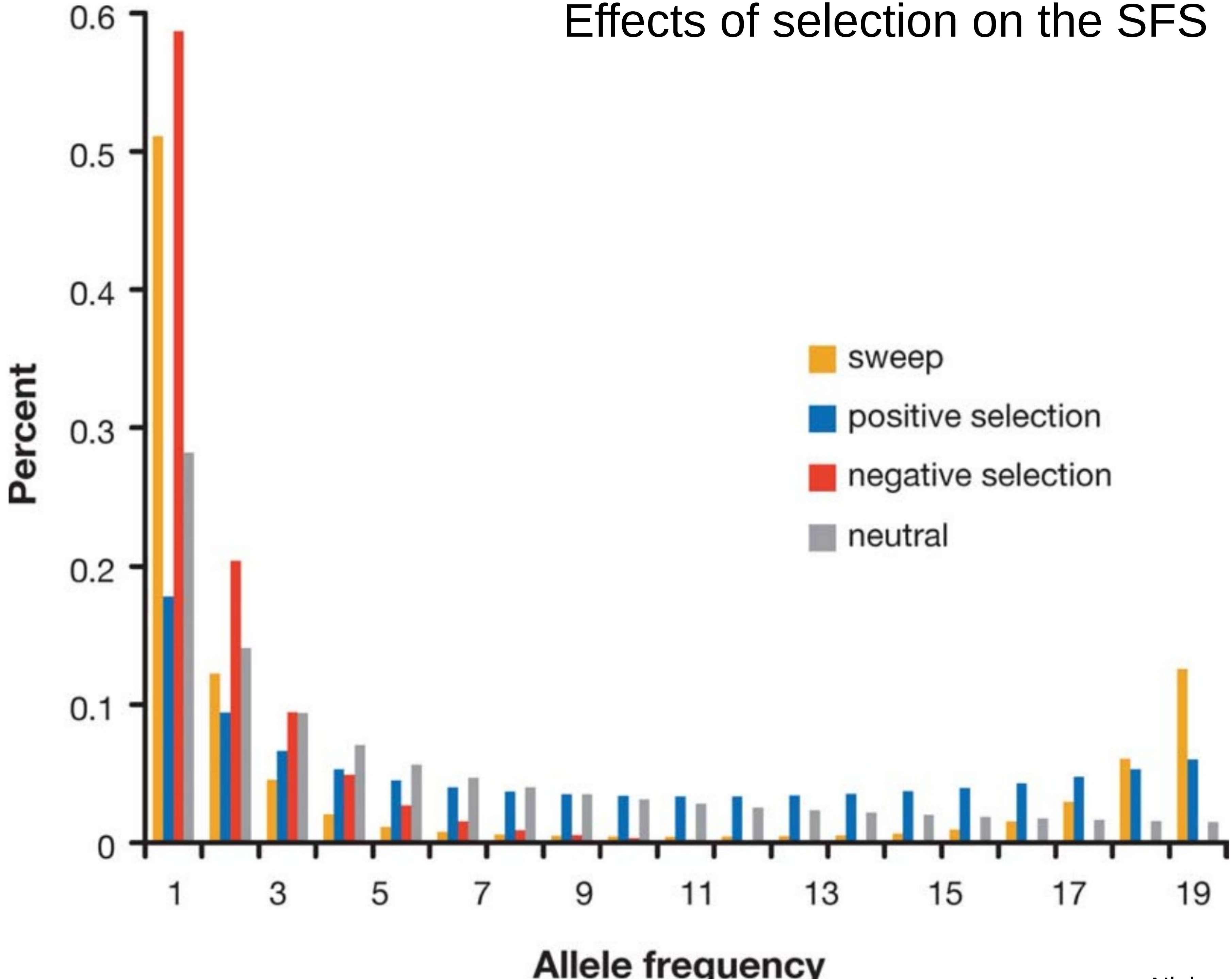
Lineages coalesce at rate 1/N, where N is the effective population size



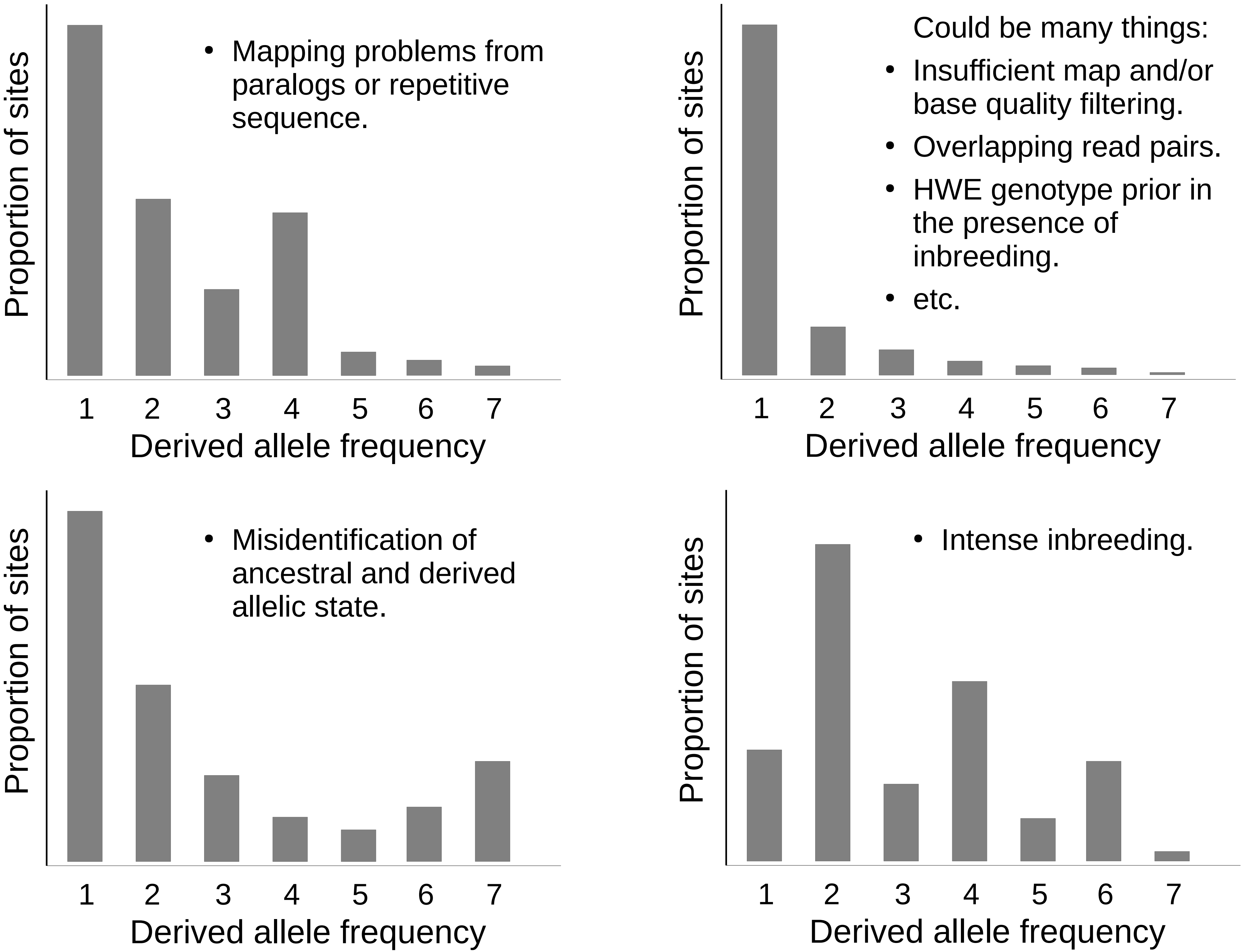
Lineages coalesce at rate 1/N, where N is the effective population size



Effects of selection on the SFS



Nielson 2005



Proportion of sites
• Mapping problems from paralogs or repetitive sequence.
1
2
3
4
5
6
7
Derived allele frequency
Proportion of sites
Could be many things:
• Insufficient map and/or base quality filtering.
• Overlapping read pairs.
• HWE genotype prior in the presence of inbreeding.
• etc.
1
2
3
4
5
6
7
Derived allele frequency
Proportion of sites
• Misidentification of ancestral and derived allelic state.
1
2
3
4
5
6
7
Derived allele frequency
Proportion of sites
• Intense inbreeding.
1
2
3
4
5
6
7
Derived allele frequency

$$\mathrm{SFS} = \boldsymbol{P} = \underbrace{\left(p_0, p_1, p_2, p_3, p_4, \ldots, p_{2n}\right)}$$

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

Likelihood of the SFS (***P***) assuming we know the genotypes at a single site:

X: observed data (sequencing reads)

G: genotype vector = (G_1, G_2, G_3, …, G_n)

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

s_d: number of derived alleles in genotype vector, ***G***

Likelihood of the SFS (***P***) assuming we know the genotypes at a single site:

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

$$p(a,b|c)=p(a|b)\,p(b|c)$$

Recall that SFS = ***P*** = (p_0, p_1, …, 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 (***P***) assuming we know the genotypes at a single site:

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

$$p(a,b|c)=p(a|b)\,p(b|c)$$

Recall that SFS = ***P*** = (p_0, p_1, …, p_{2n})

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

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

Likelihood of the SFS (***P***) assuming we know the genotypes at a single site:

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

Probability based on the number of ways to have *j* derived alleles in ***G*** out of the total number of ways to to arrange *j* derived alleles in 2*n* chromosomes (i.e. choose(2n, j)).

Assumes HWE

Example with 4 diploid individual with derived allele T.

j = 4, *k* = # heterozygotes = 2

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

$$p(\mathbf{G}|s_d=j)=\frac{2^k}{\binom{2n}{j}}=\frac{2^2}{\binom{8}{4}}=\frac{4}{70}$$

$2^k = 2^2 = 4$ combinations

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

Allow for unknown genotypes (consider the likelihood for all possible genotypes)

$$p(X|\boldsymbol{P})=$$

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

Likelihood of **P** for site *v* (-doSaf 1):

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

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

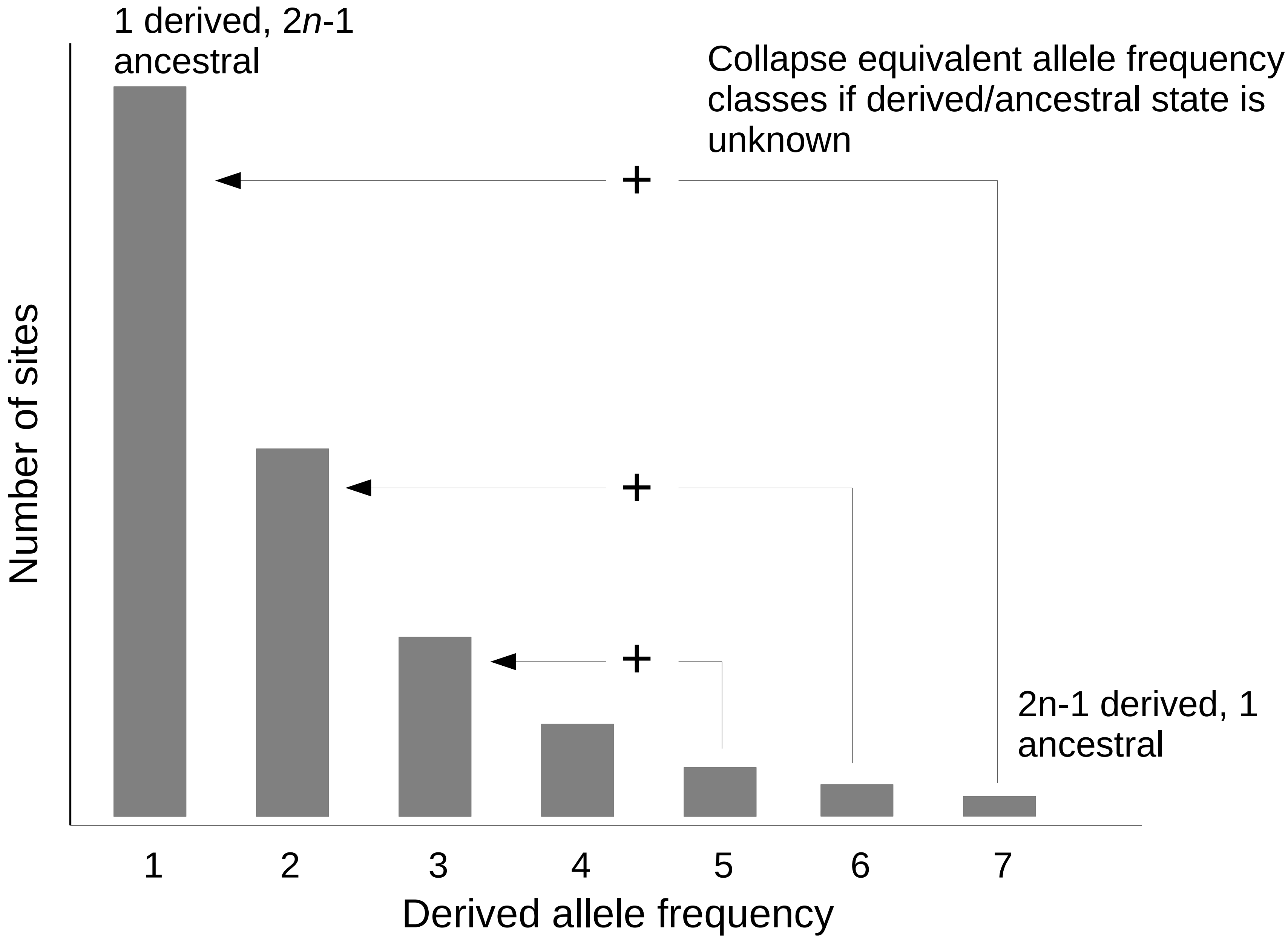
Assume sites are independent

Likelihood of **P**:

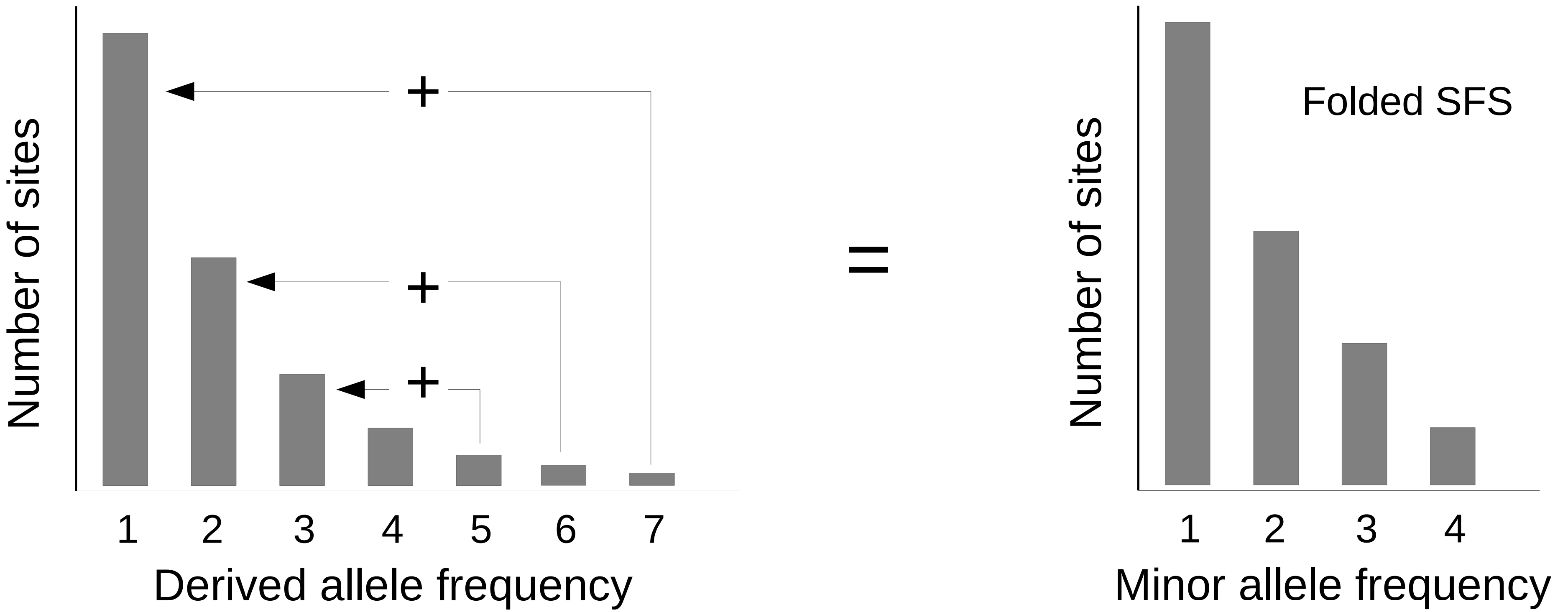
$$p(X|\boldsymbol{P}) =$$

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

Folding the SFS



Folding the SFS



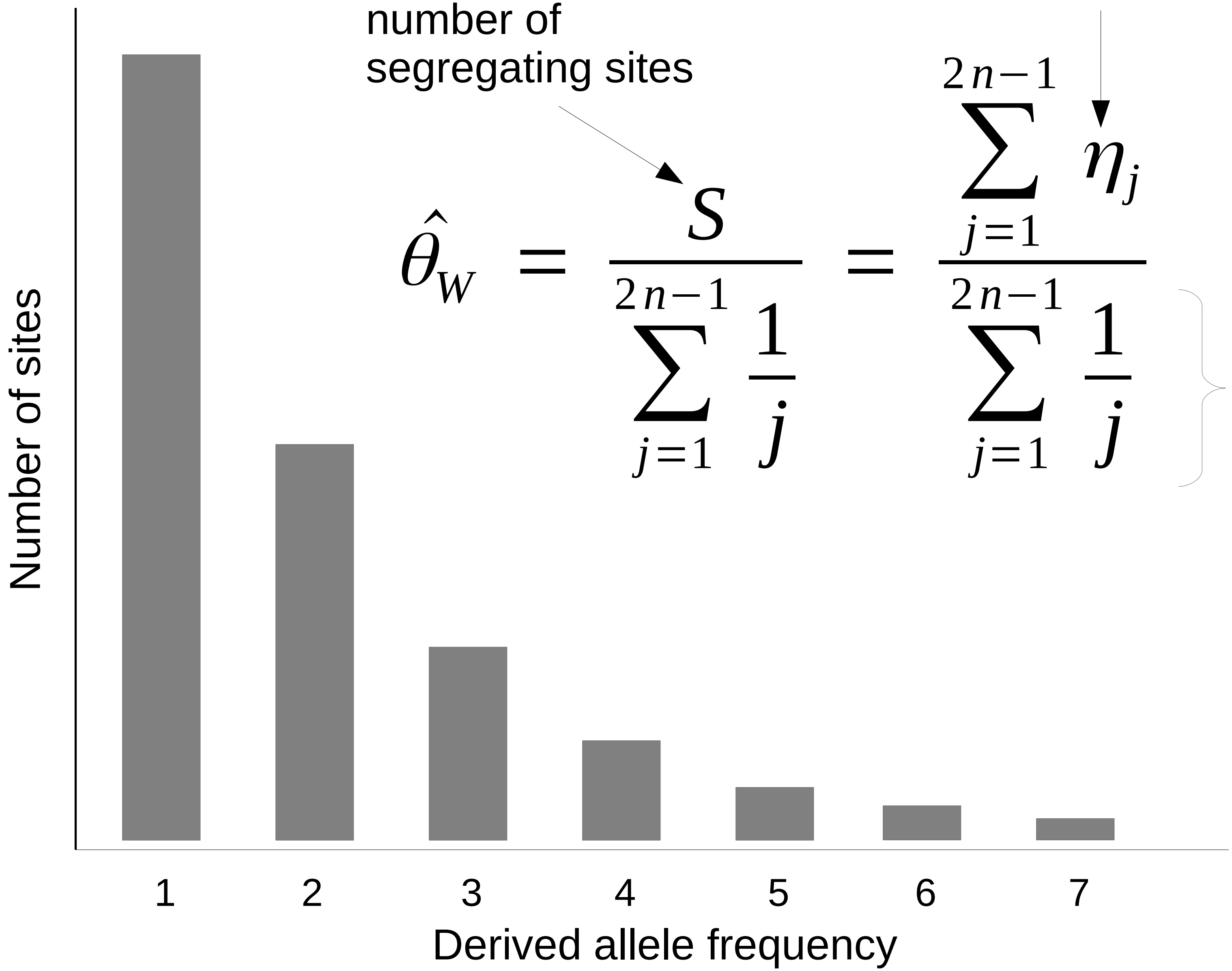
Estimating diversity from the SFS

number of segregating sites

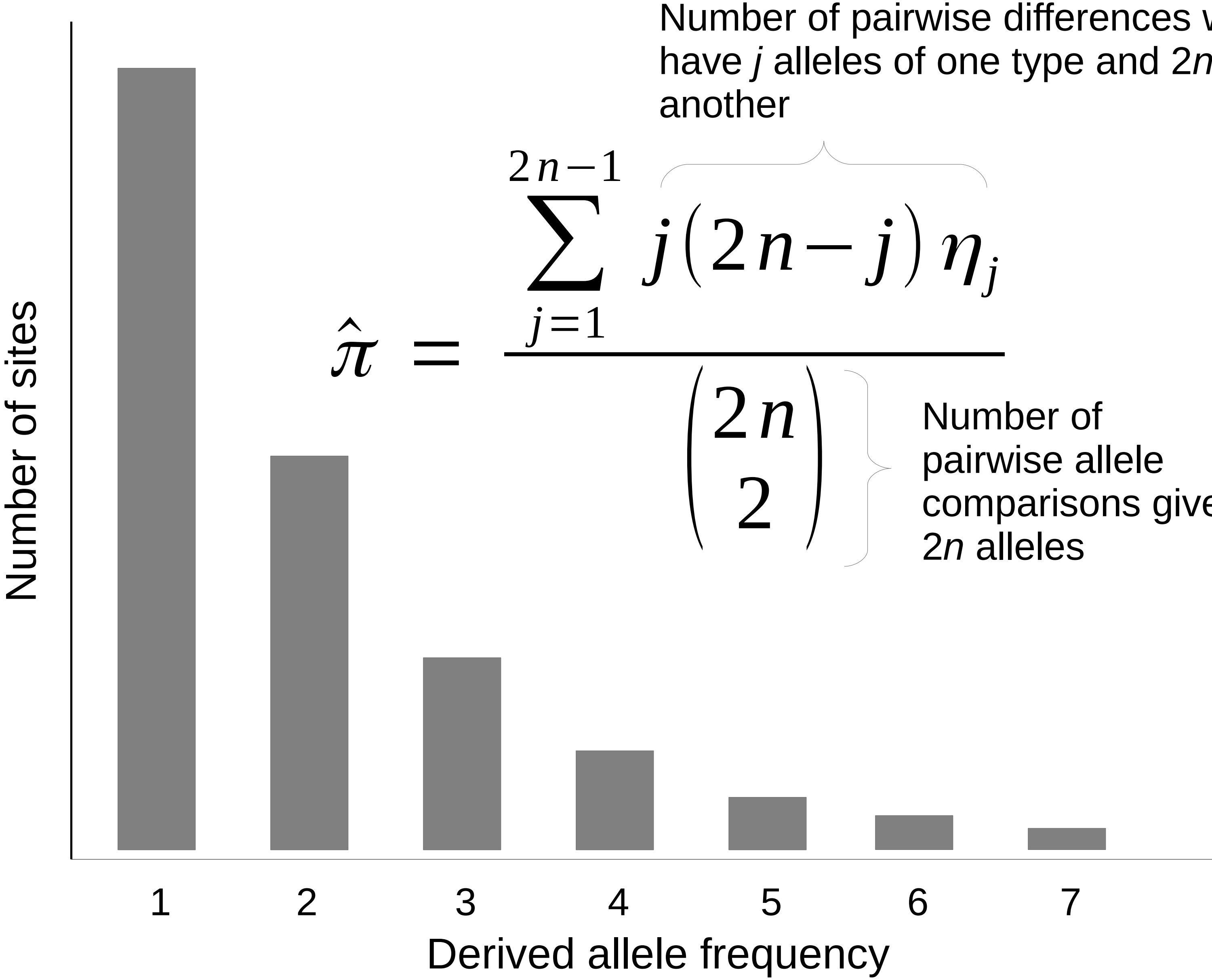
number of sites with j derived alleles

$$\hat{\theta}_W = \frac{S}{\sum_{j=1}^{2n-1} \frac{1}{j}} = \frac{\sum_{j=1}^{2n-1} \eta_j}{\sum_{j=1}^{2n-1} \frac{1}{j}}$$

Expected length of a coalescent tree of n individuals (corrects for sample size)



Estimating diversity from the SFS

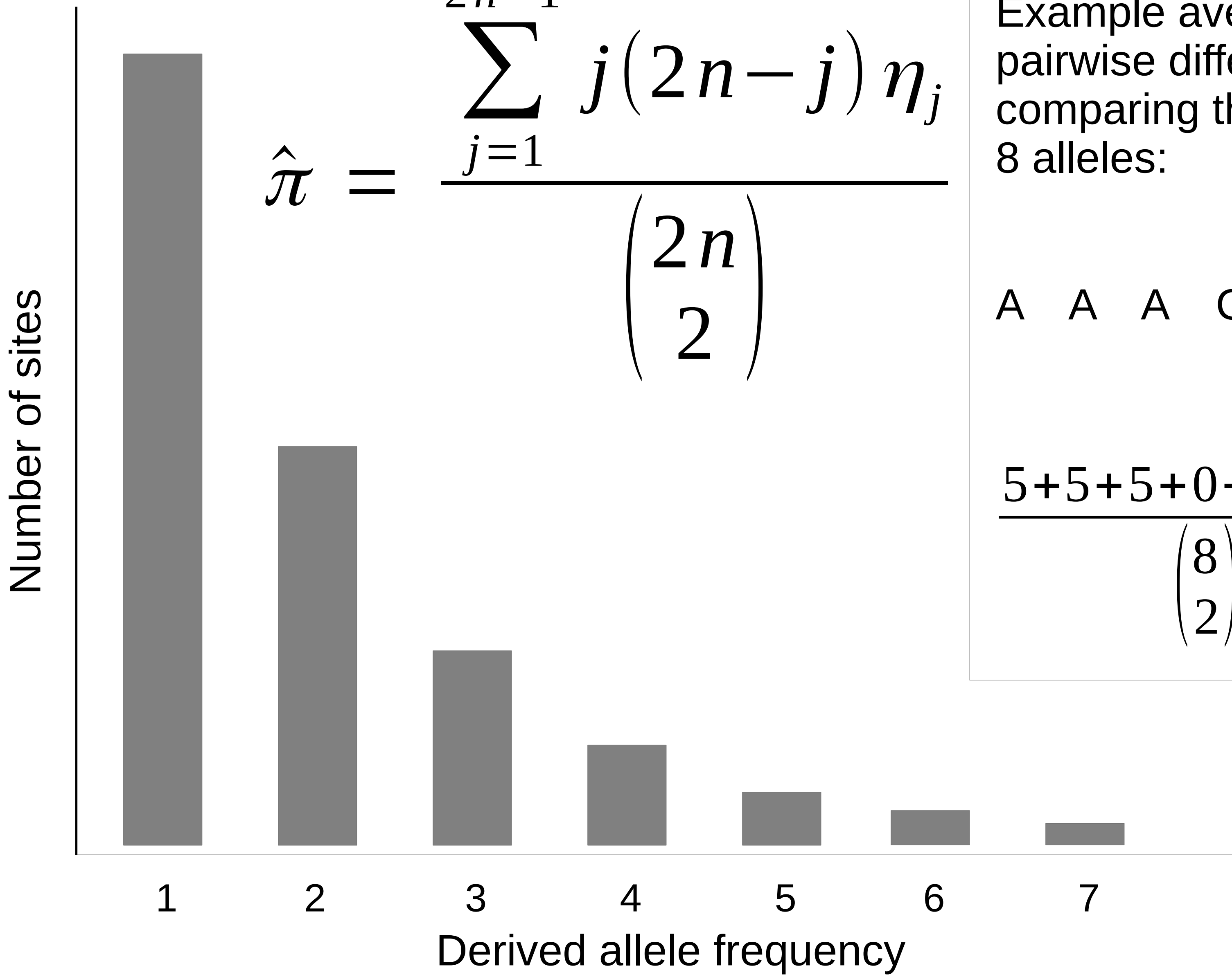


$$\hat{\pi} = \frac{\sum_{j=1}^{2n-1} \overbrace{j(2n-j)\,\eta_j}^{\text{Number of pairwise differences when you have } j \text{ alleles of one type and } 2n-j \text{ of another}}}{\binom{2n}{2}}$$

Number of pairwise allele comparisons given $2n$ alleles

Estimating diversity from the SFS

$$\hat{\pi} = \frac{\sum_{j=1}^{2n-1} j(2n-j)\,\eta_j}{\binom{2n}{2}}$$



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

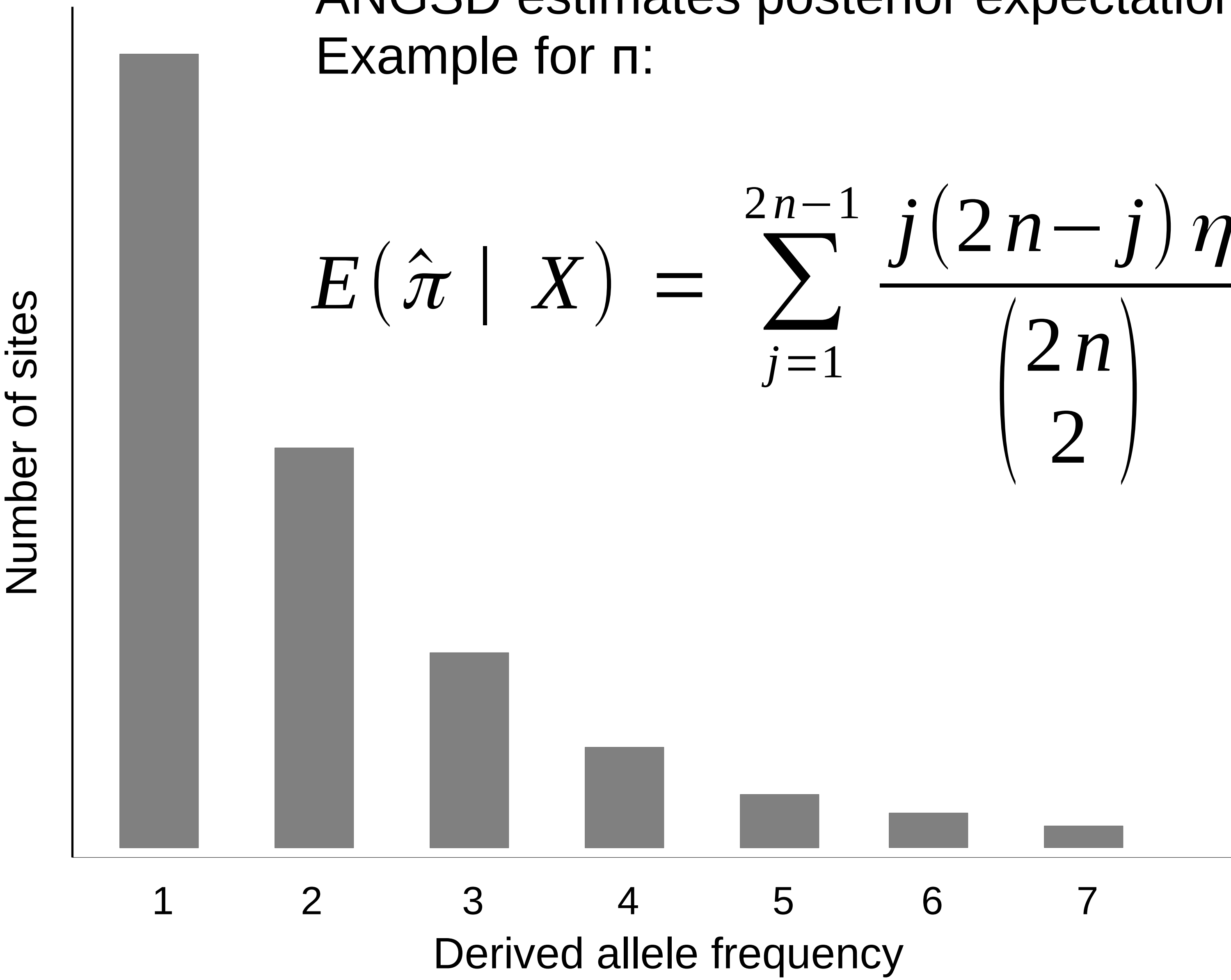
A A A C C C C C

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

Estimating diversity from the SFS

ANGSD estimates posterior expectation of θ.
Example for π:

$$E(\hat{\pi} \mid X) = \sum_{j=1}^{2n-1} \frac{j(2n-j)\eta_j}{\binom{2n}{2}} \underbrace{E(\eta_j | X)}_{\text{Given by genome-wide SFS}}$$



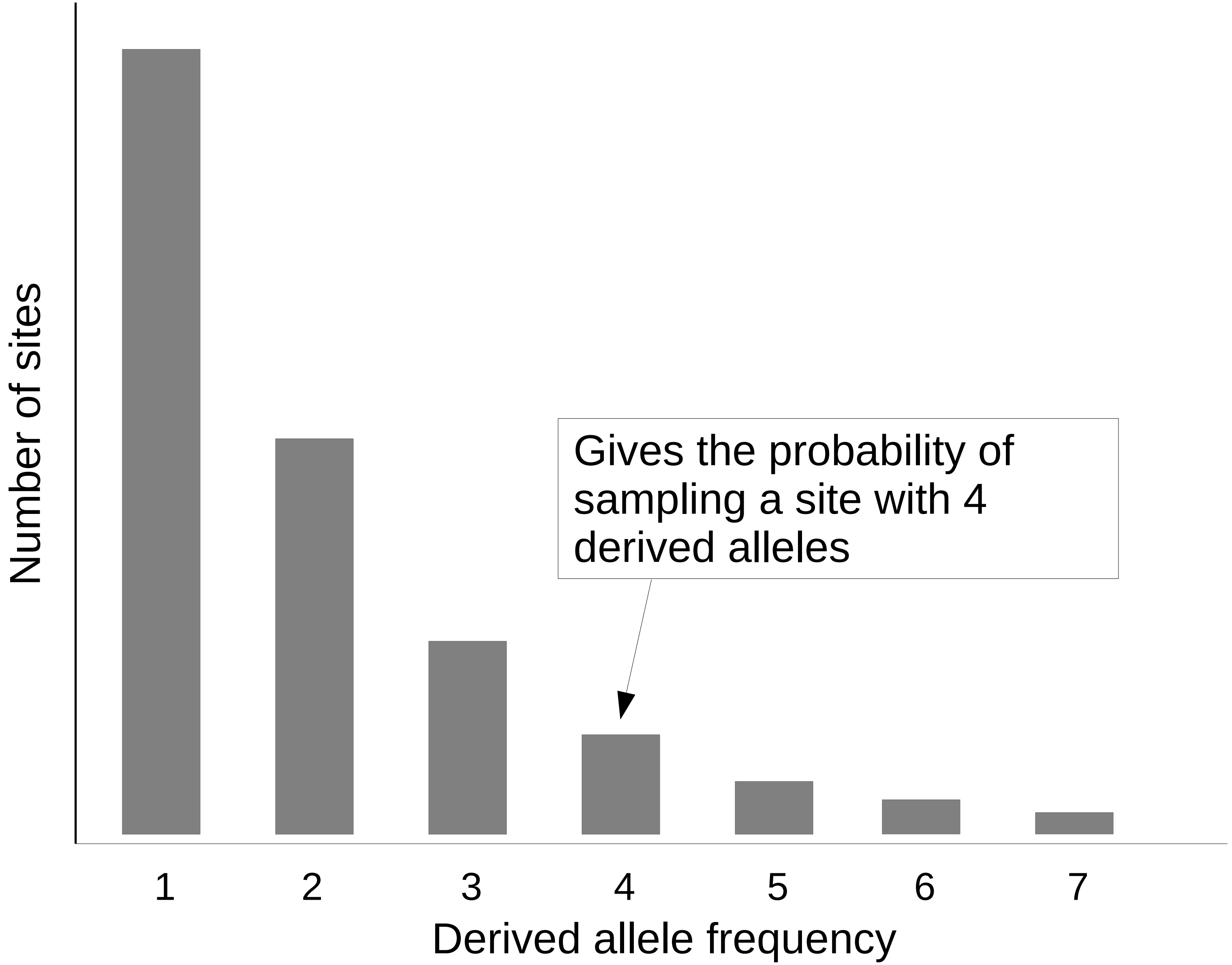
How to estimate posterior probabilities of allele frequencies

-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**

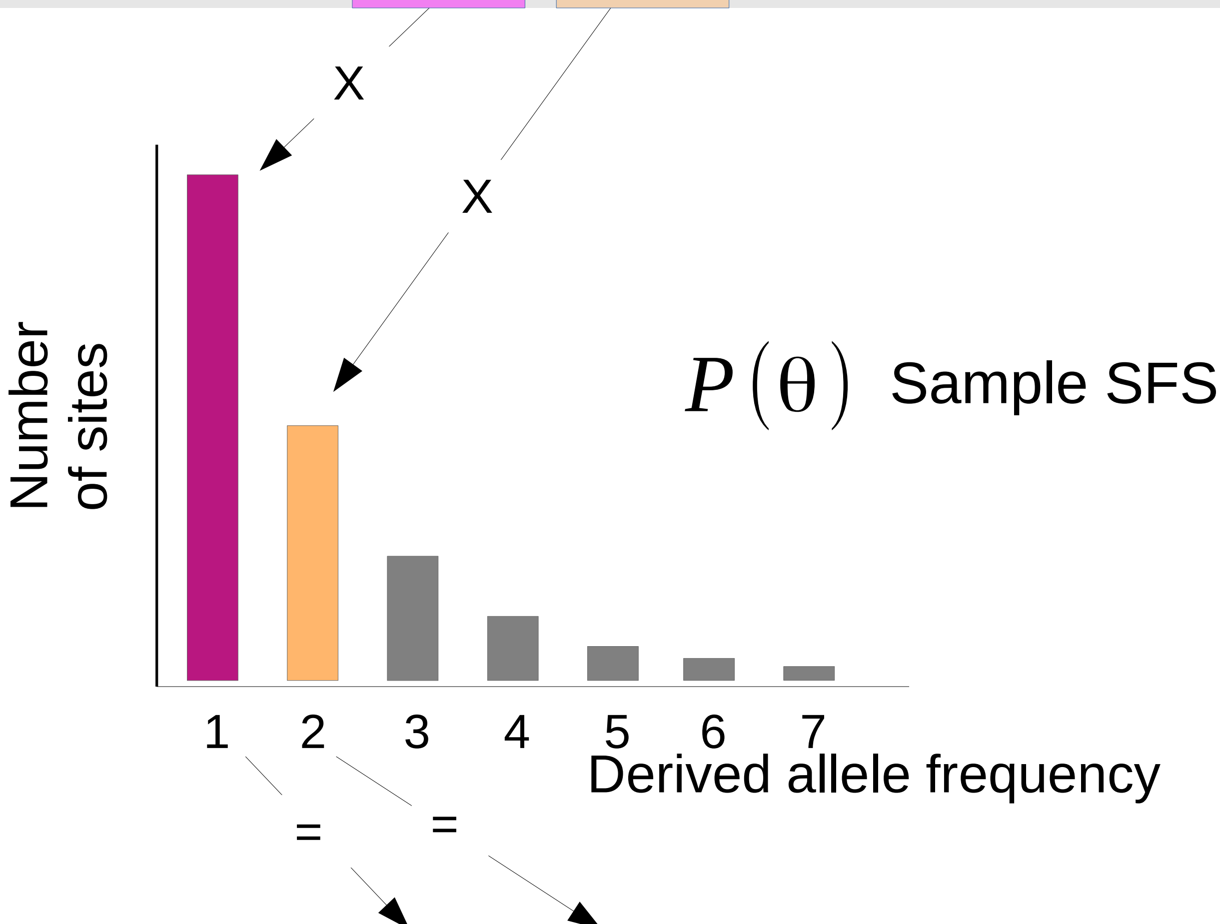
$$P(\theta|X) = \color{green}{P(X|\theta)} P(\theta)$$

How to estimate posterior probabilities of allele frequencies



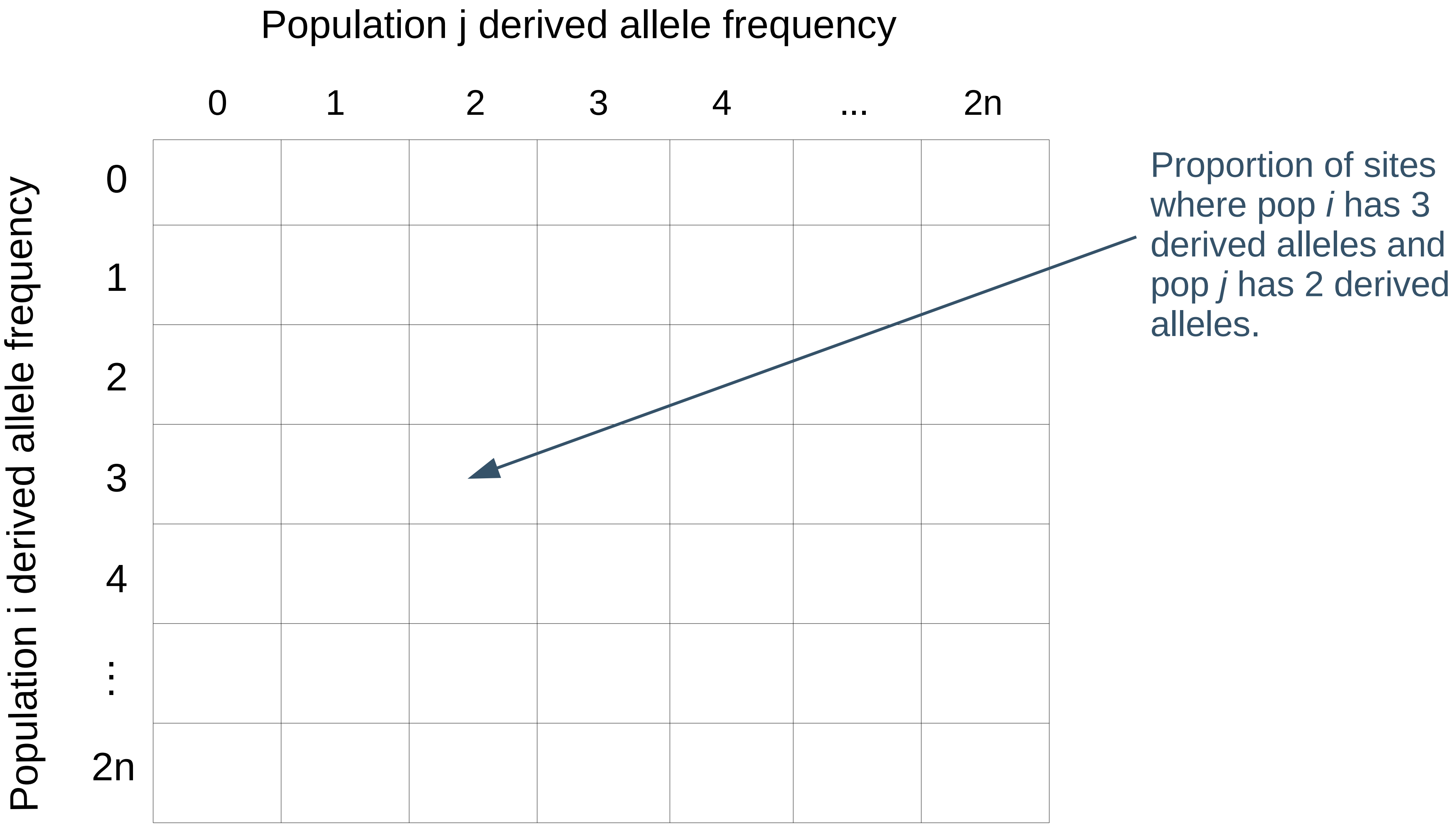
$$P(\theta|X) = P(X|\theta)\,P(\theta)$$

	0	1	2	3	4	. . .	2n	
Site1	**0.00**	**-2.24**	**-4.53**	**-6.99**	**-9.63**		**-232.69**	$P(X \mid \theta)$
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**	$P(\theta \mid X)$
Site2	**0.9892**	**0.0101**	**0.0006**	**0.0000**	**0.0000**		**0.0000**	

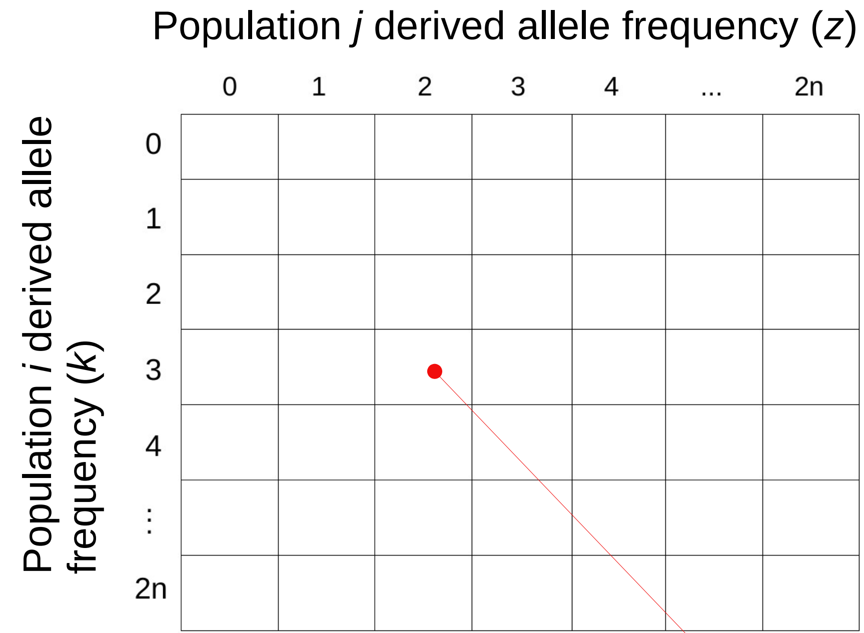
2-dimensional SFS



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

Genetic variance between populations

Genetic variance within populations



Probability of *k* derived alleles in pop *i* and *z* derived alleles in pop *j*

Likelihood of k derived alleles in pop *i* at site *v*

Likelihood of *z* derived alleles in pop *j* at site *v*

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

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