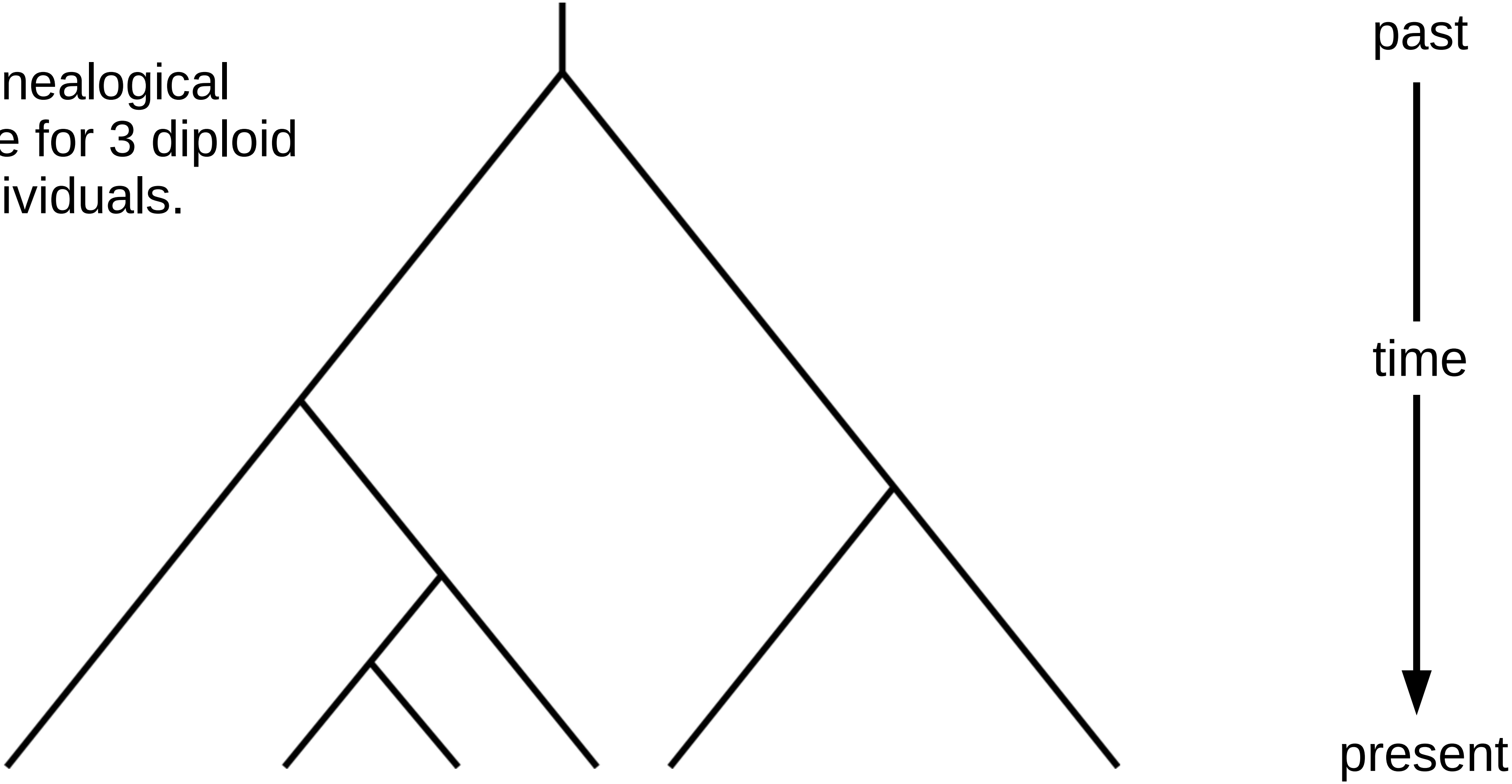
Site Frequency Spectrum (SFS)

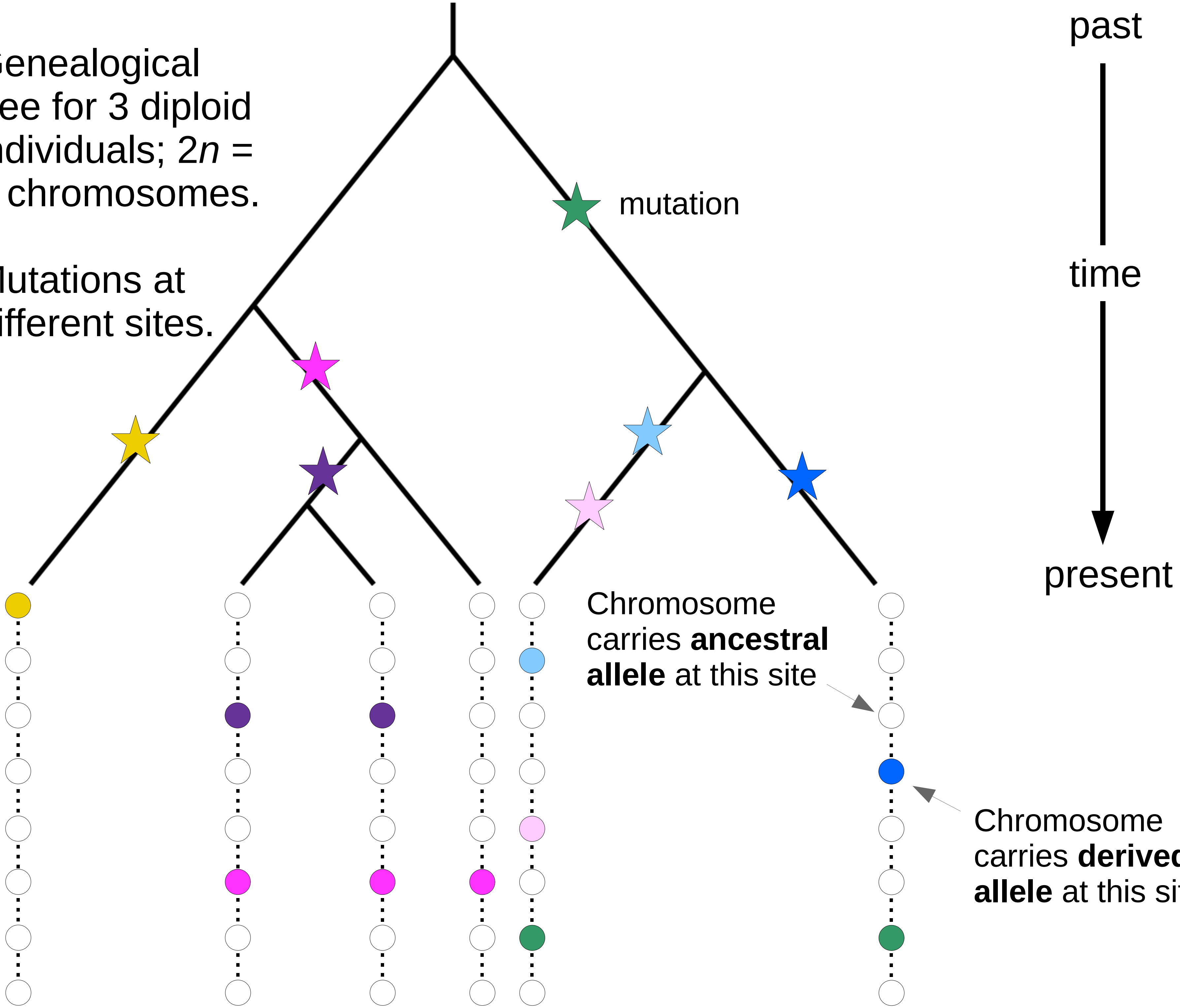
Tyler Linderoth
Physalia lcWGS course 2025

Genealogical tree for 3 diploid individuals.



Genealogical tree for 3 diploid individuals; 2*n* = 6 chromosomes.

Mutations at different sites.



mutation

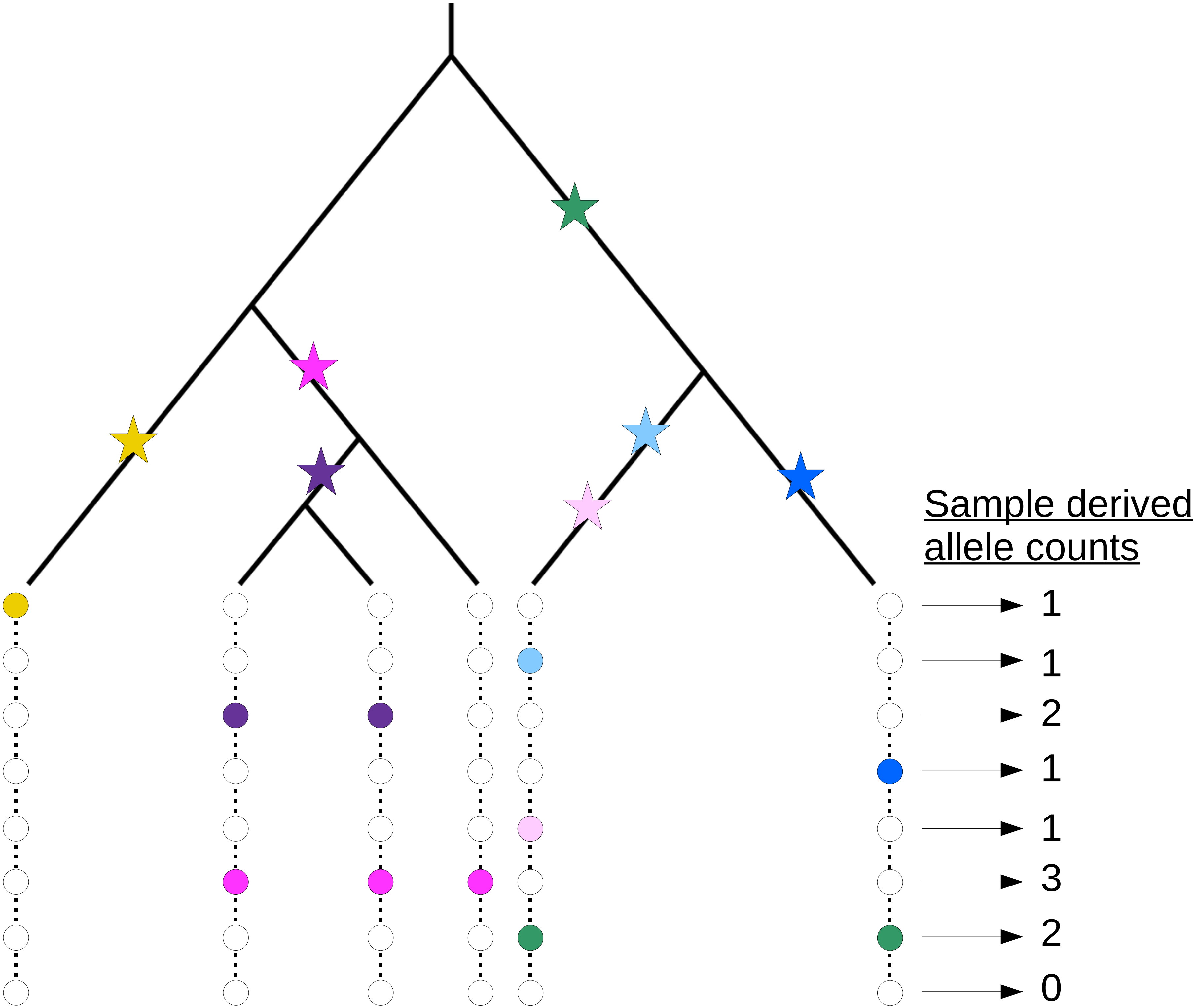
past

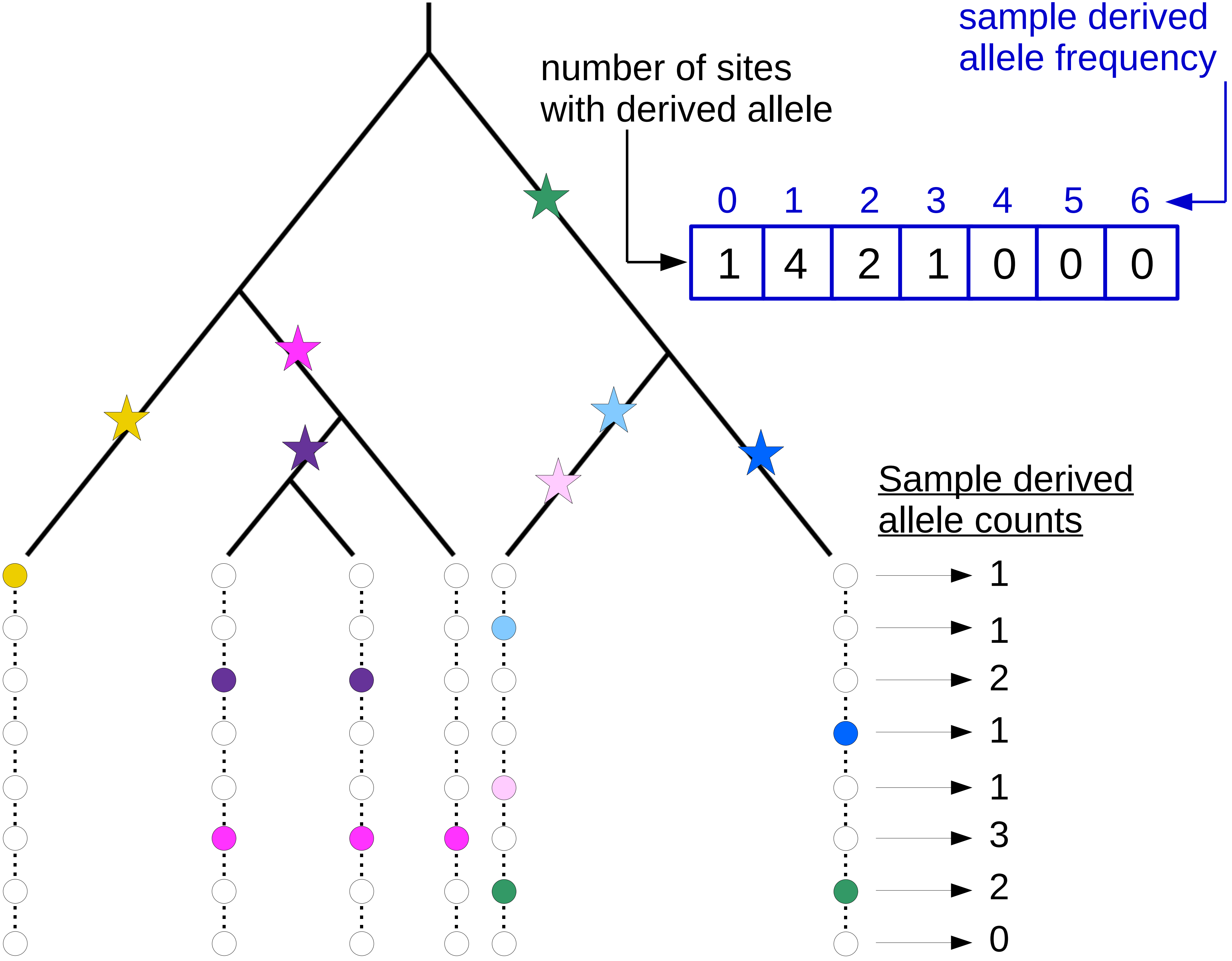
time

present

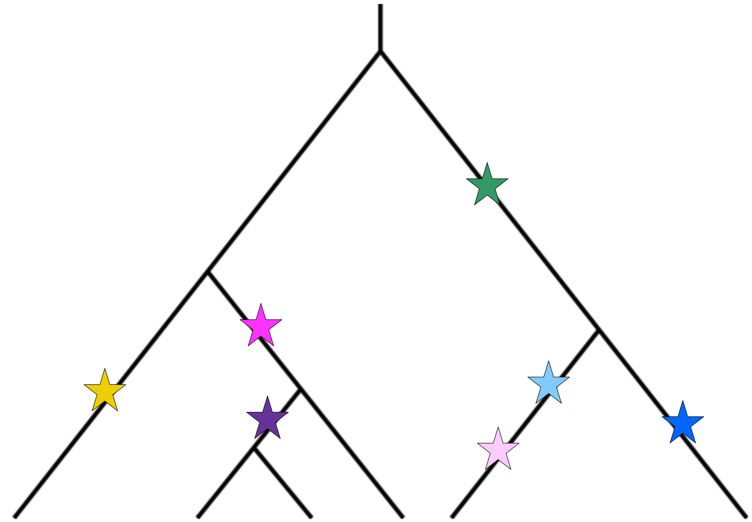
Chromosome carries **ancestral allele** at this site

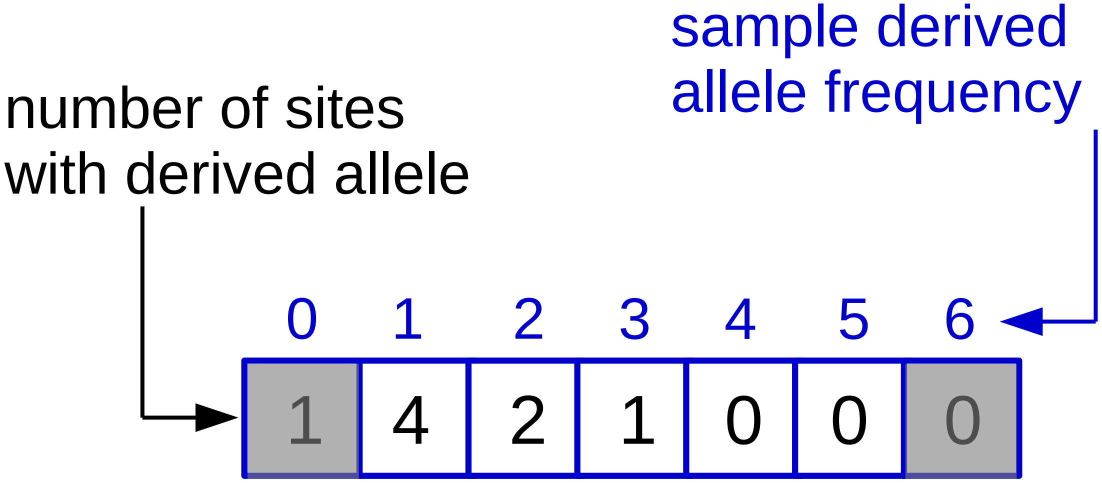
Chromosome carries **derived allele** at this site


Sample derived allele counts
1
1
2
1
1
3
2
0



0	1	2	3	4	5	6
1	4	2	1	0	0	0

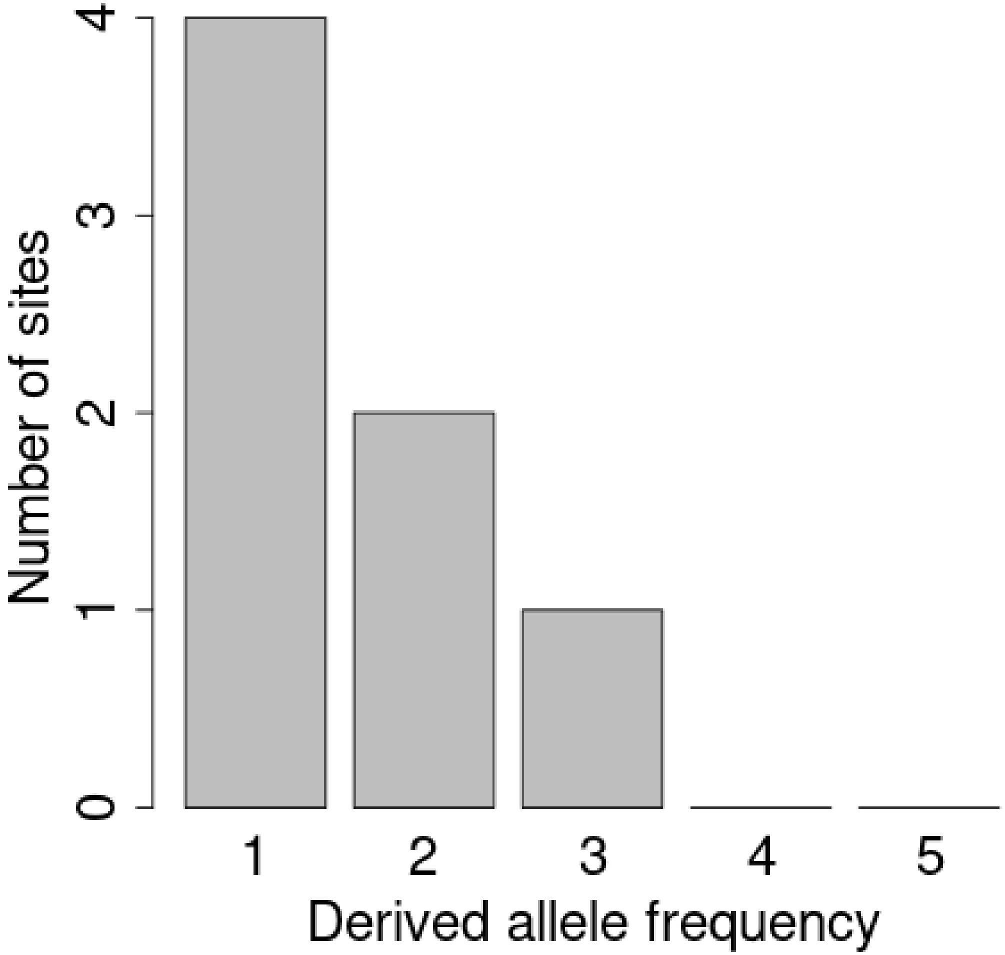




number of sites
with derived allele

sample derived
allele frequency

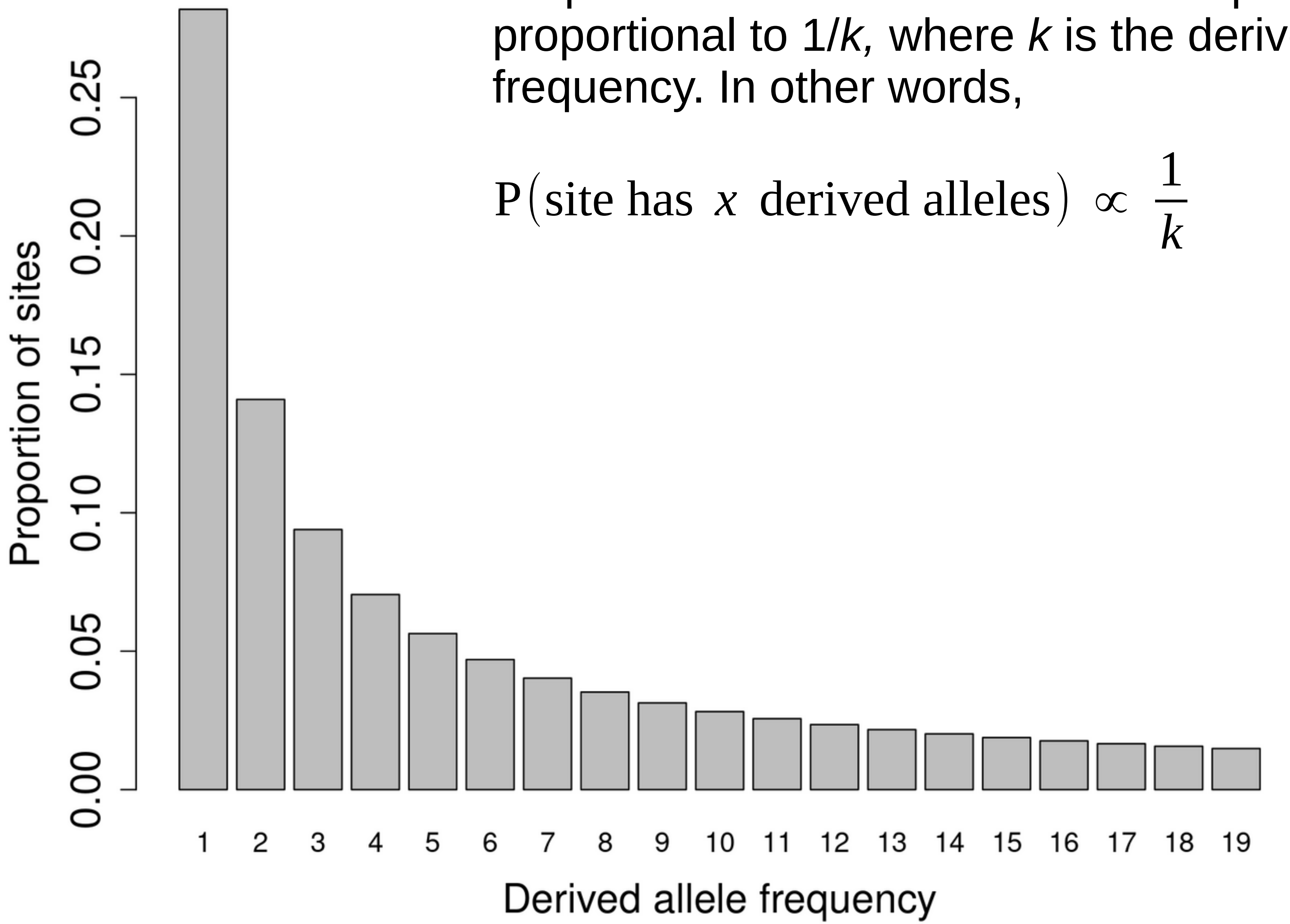
0	1	2	3	4	5	6
1	4	2	1	0	0	0



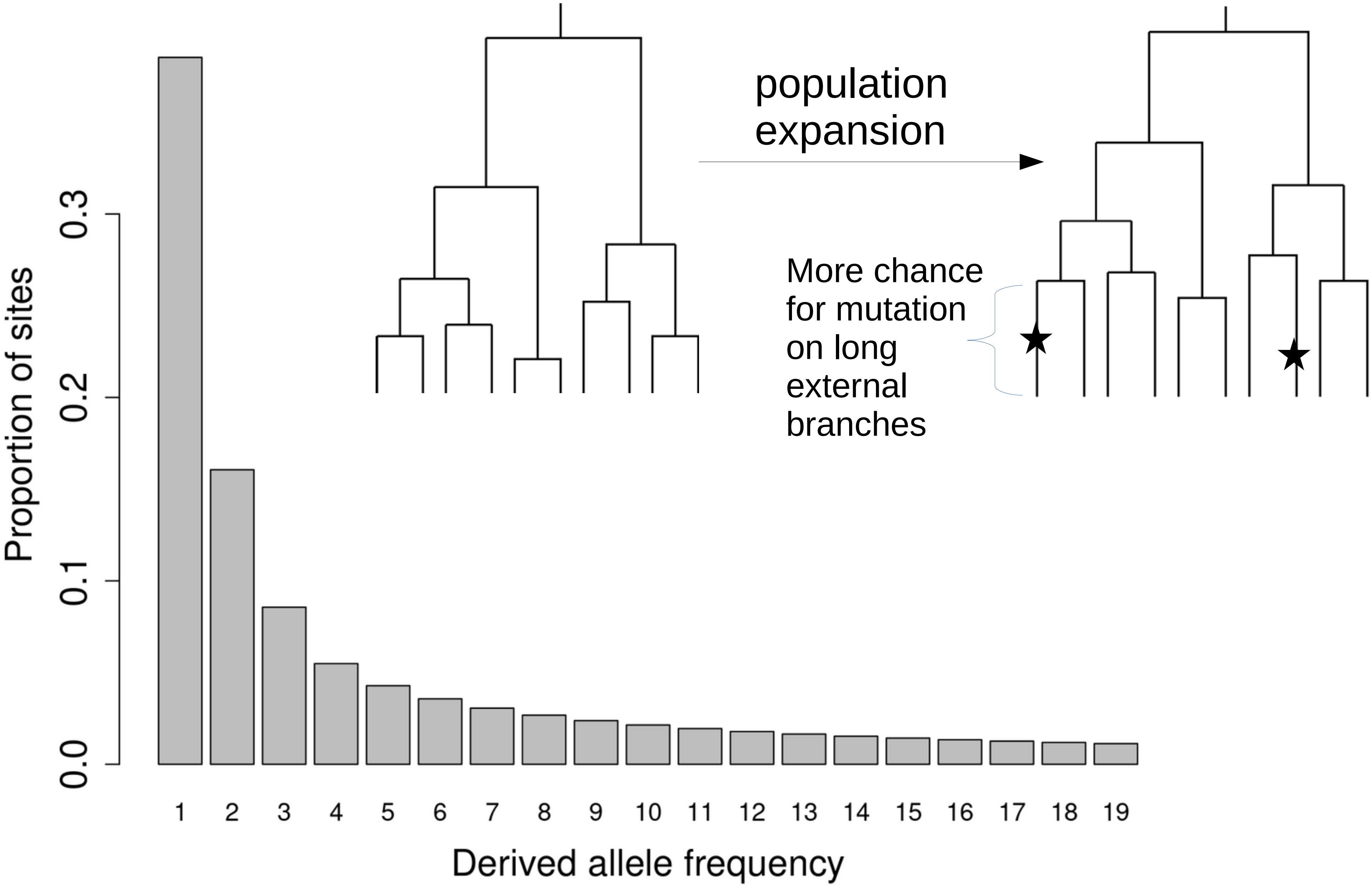
Neutral, constant-size population SFS

Proportion of sites in each allele frequency class proportional to 1/*k,* where *k* is the derived allele frequency. In other words,

$$\mathrm{P}\left(\text{site has } x \text{ derived alleles}\right) \propto \frac{1}{k}$$



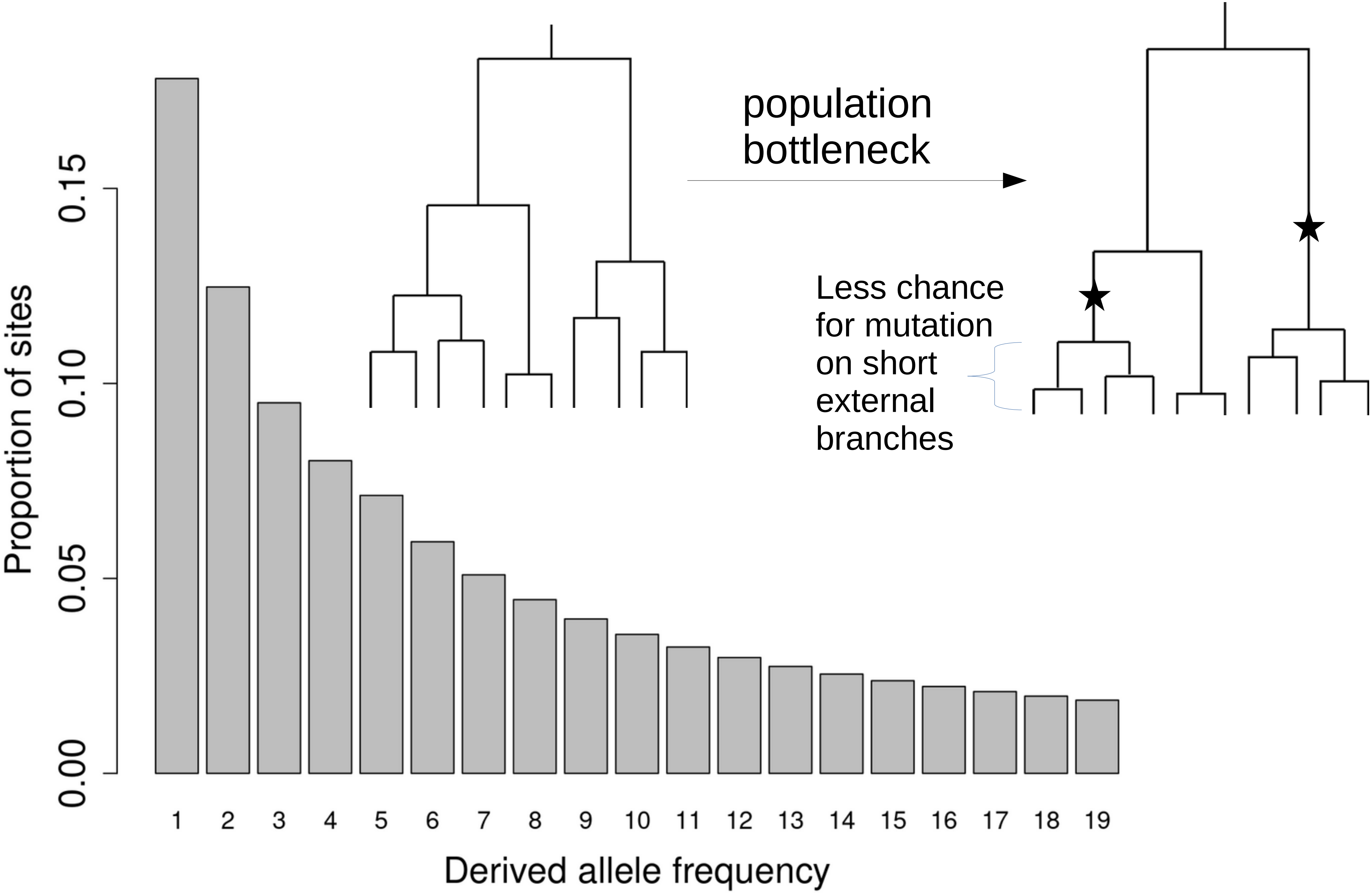
Lineages coalesce at rate 1/*N*. *N* is the effective population size.



population expansion

More chance for mutation on long external branches

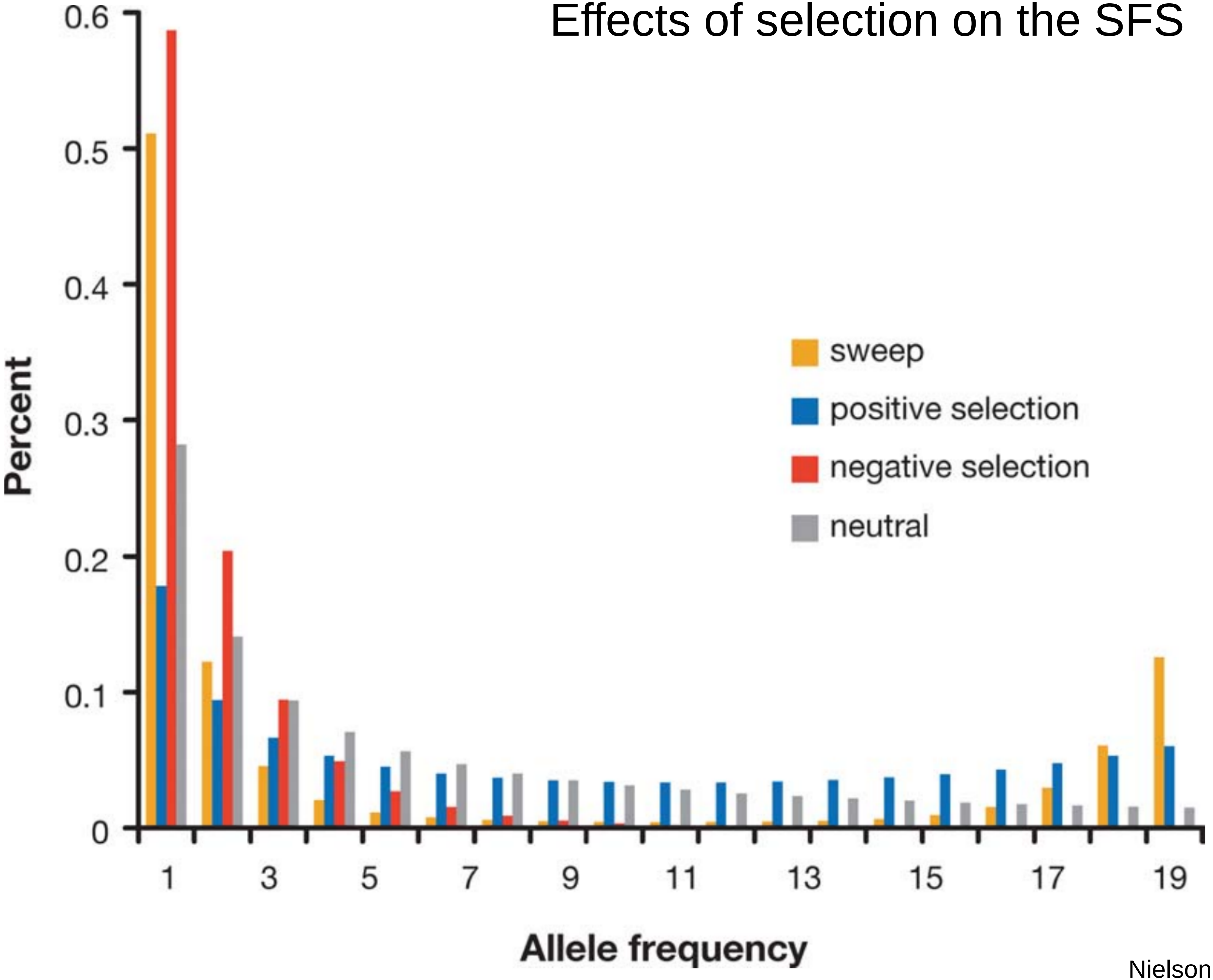
Lineages coalesce at rate 1/*N. N* is the effective population size.



population bottleneck

Less chance for mutation on short external branches

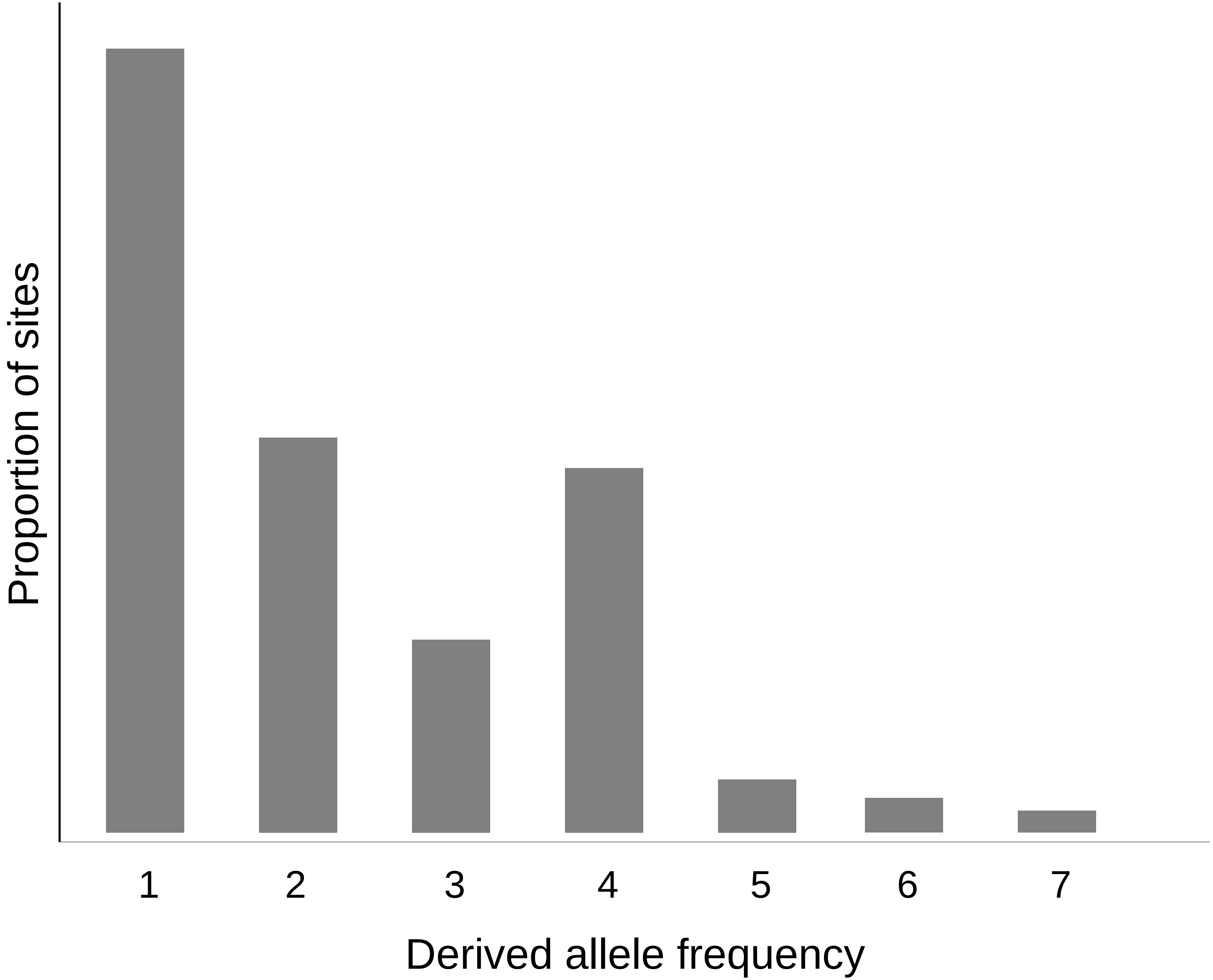
Effects of selection on the SFS



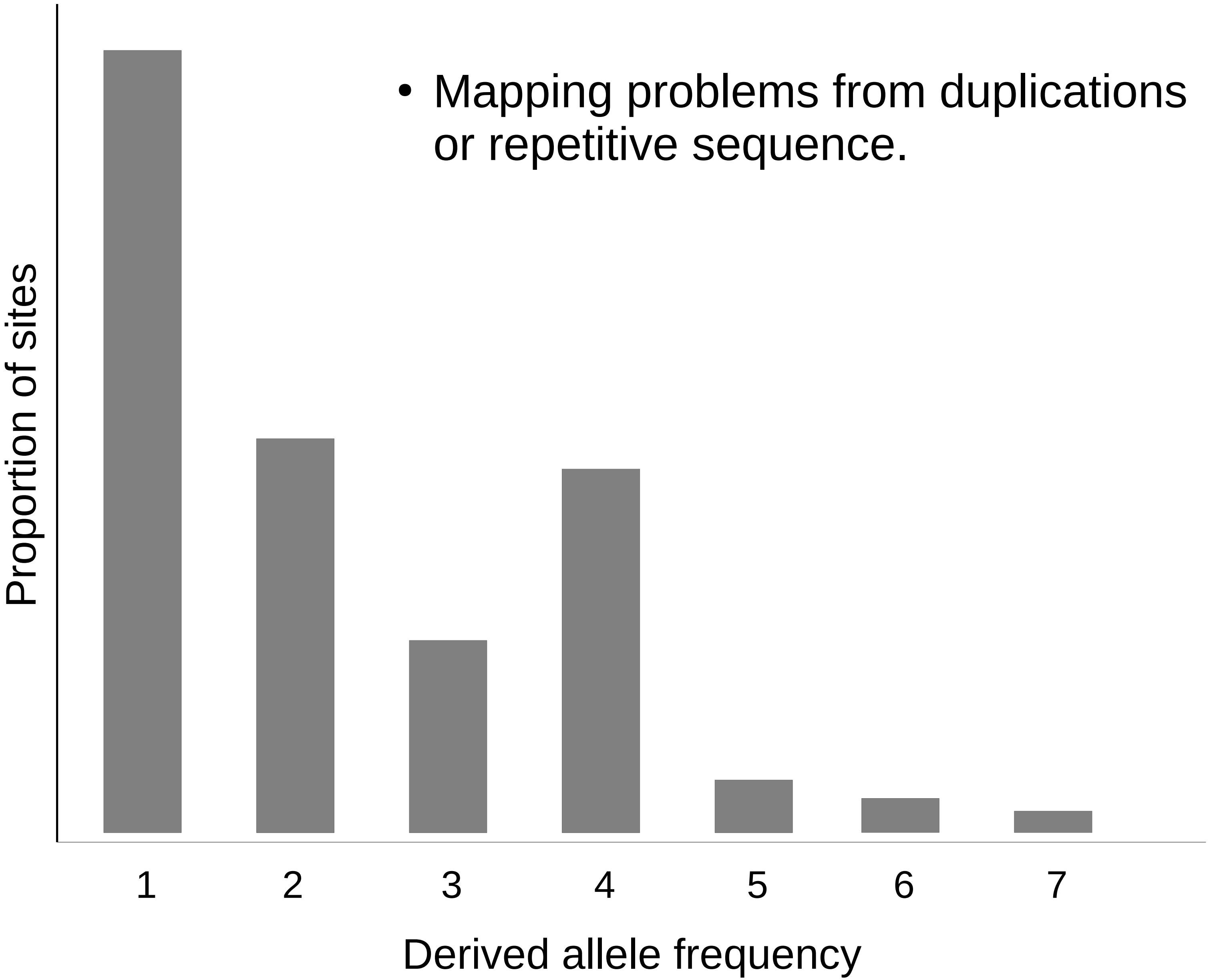
Nielson 2005

Let’s play the guess the bioinformatic problem game

SFS for a sample of 4 diploid individuals
Case 1

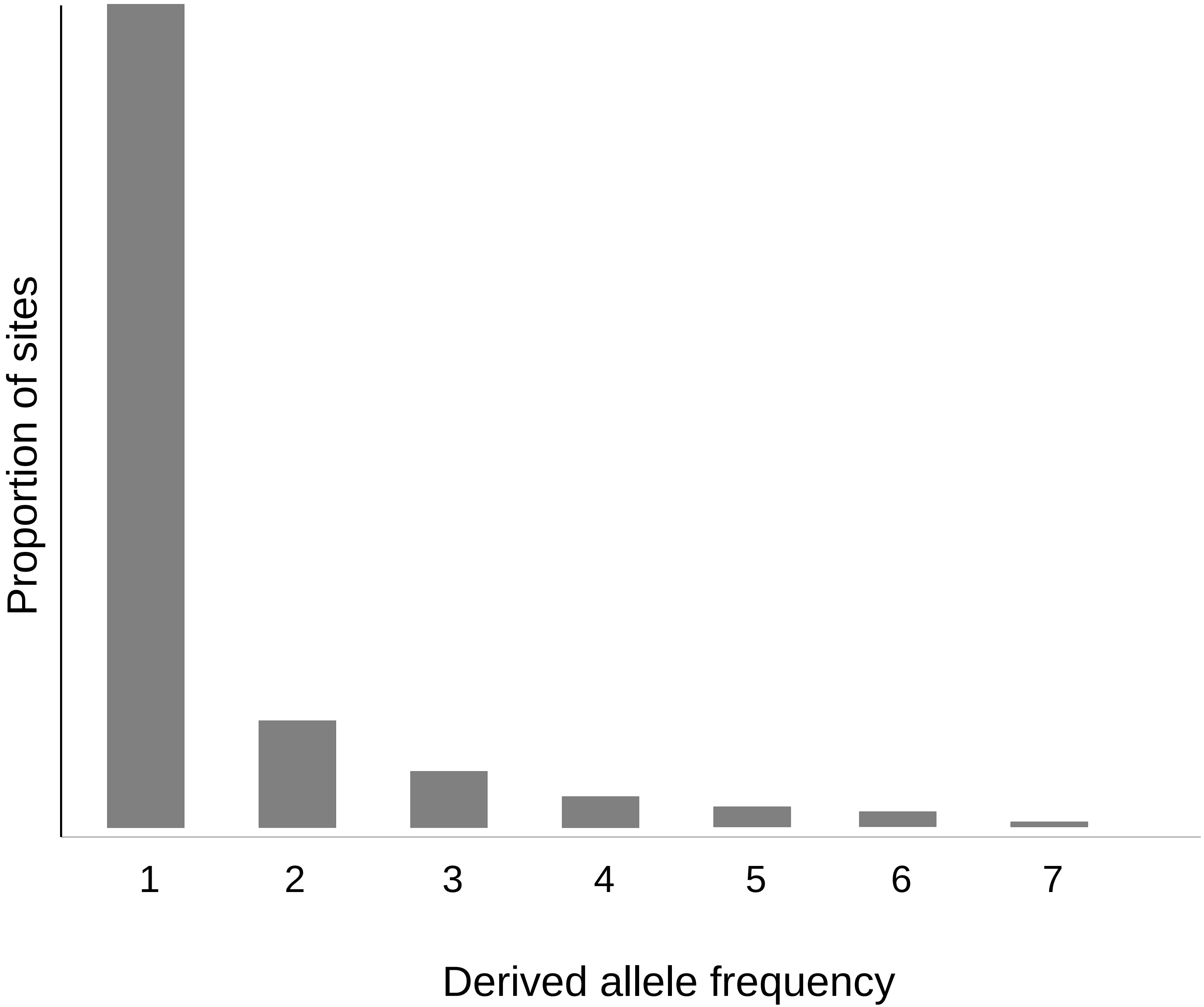


SFS for a sample of 4 diploid individuals
Case 1

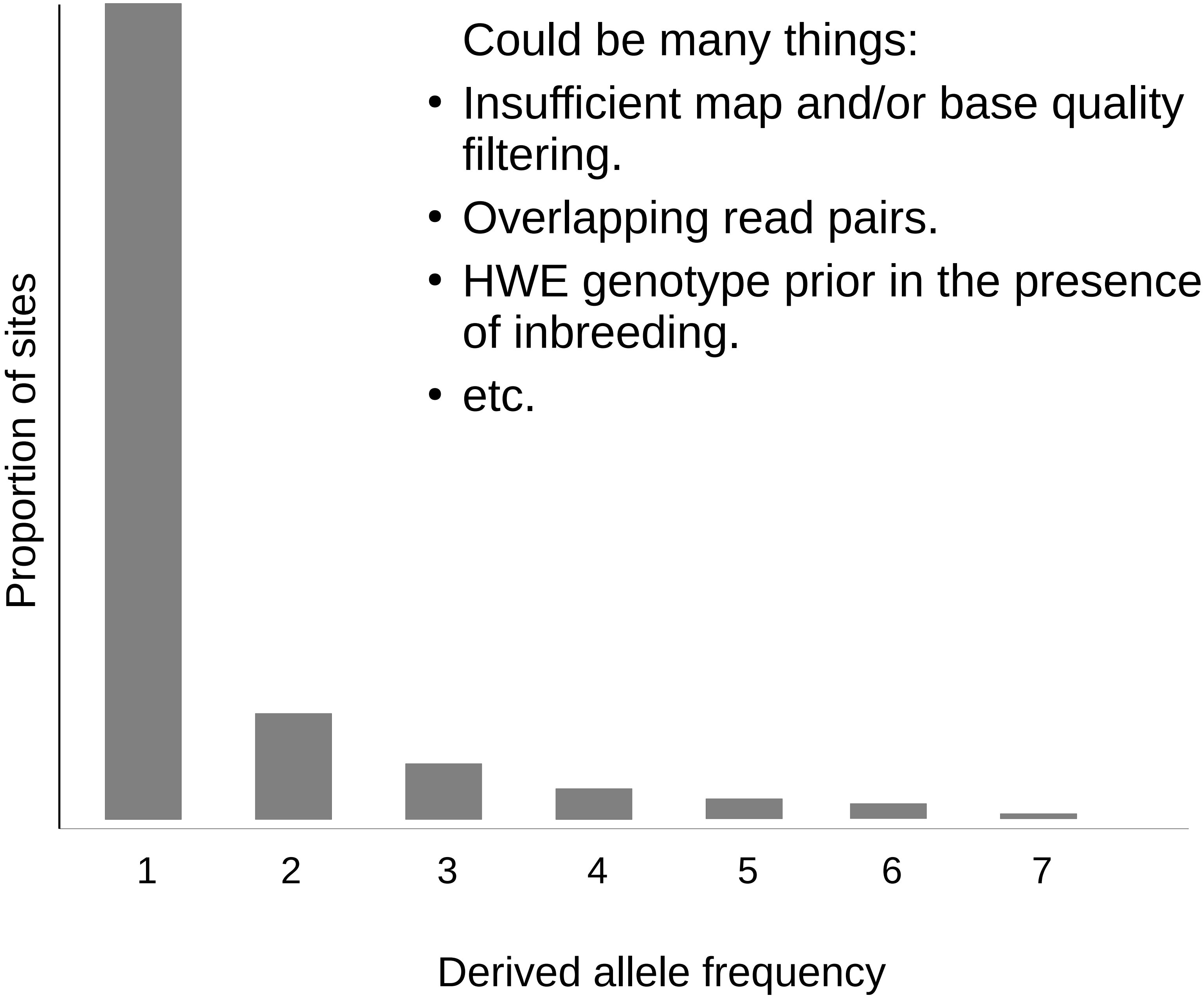


- Mapping problems from duplications or repetitive sequence.

SFS for a sample of 4 diploid individuals
Case 2



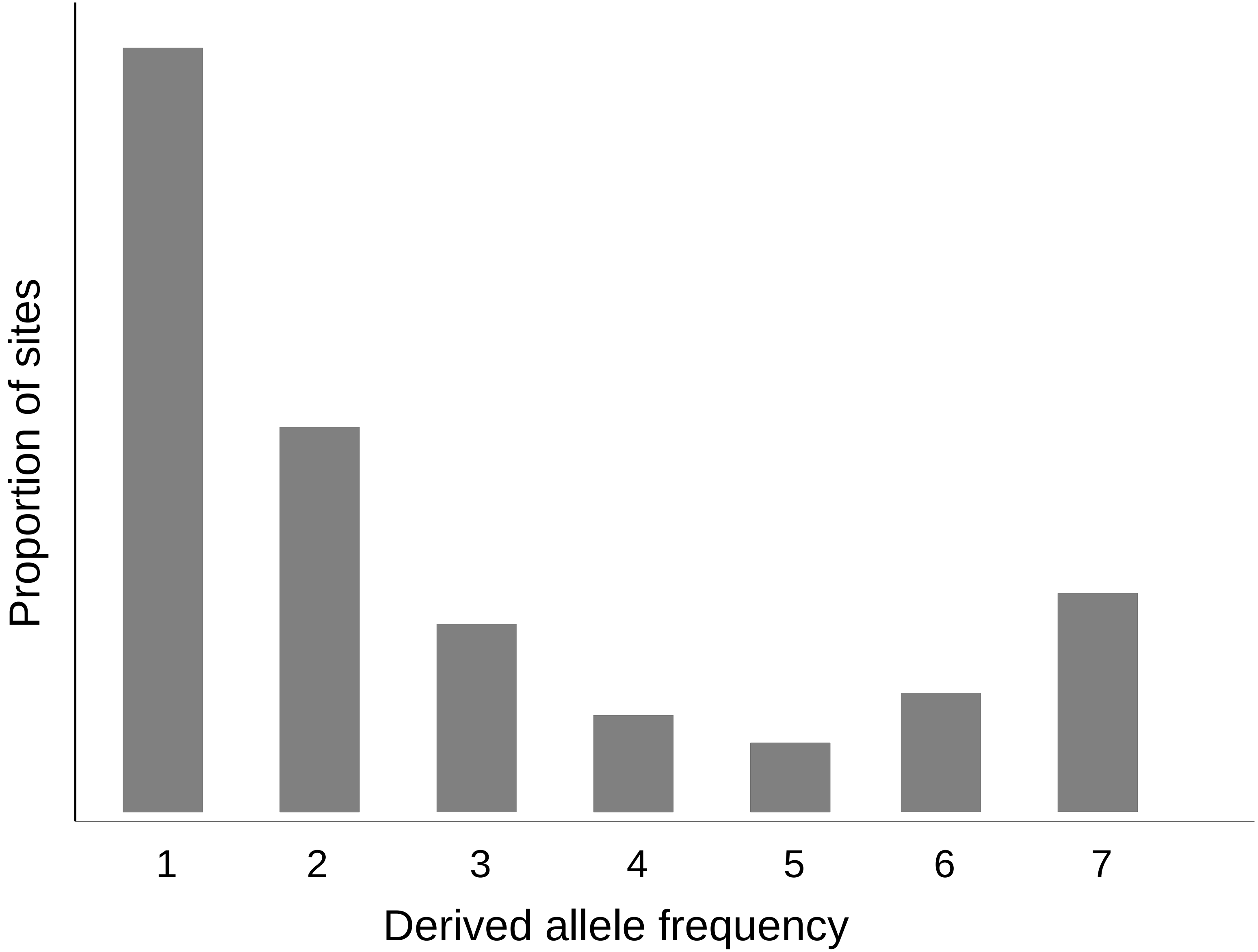
SFS for a sample of 4 diploid individuals
Case 2



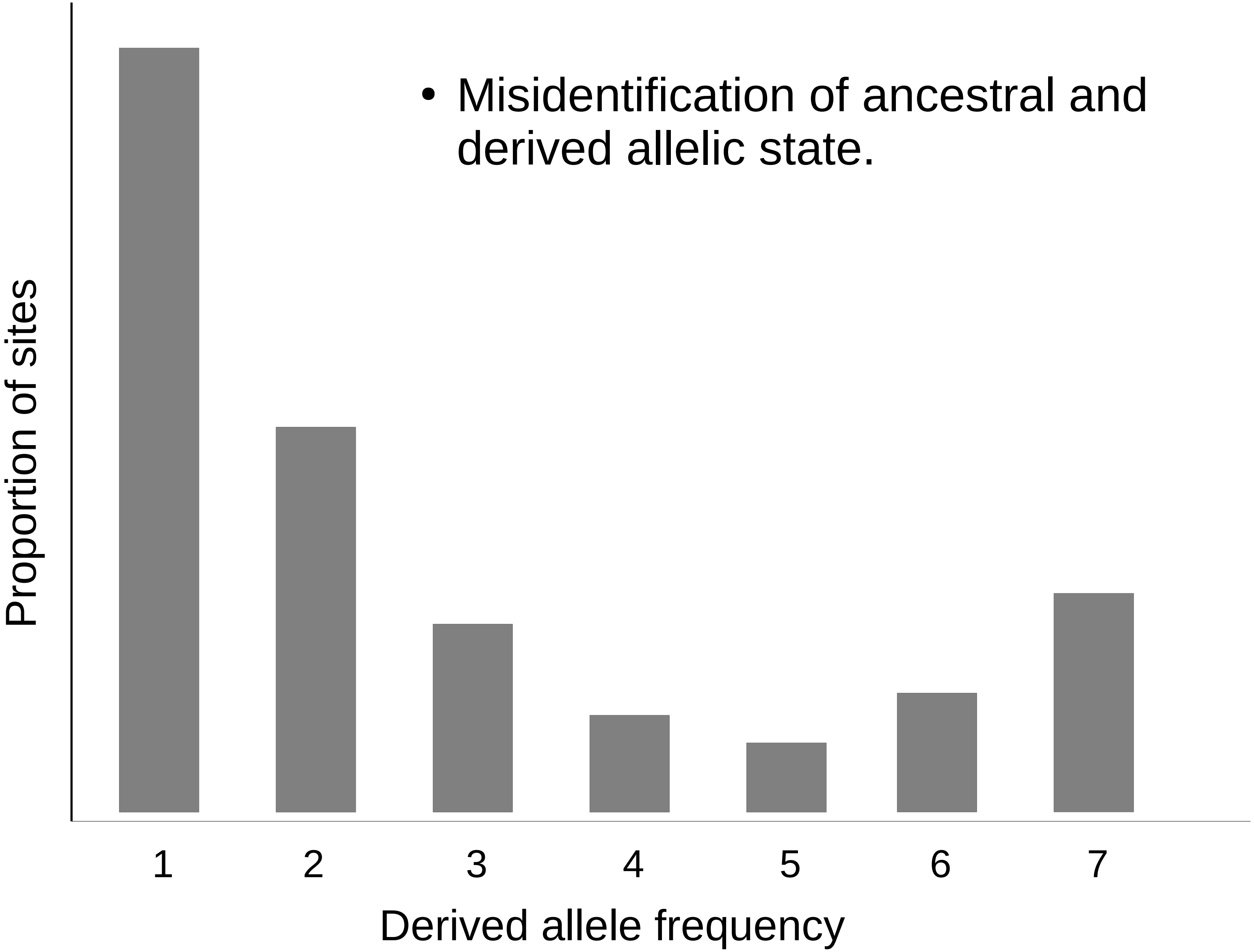
Could be many things:

- Insufficient map and/or base quality filtering.
- Overlapping read pairs.
- HWE genotype prior in the presence of inbreeding.
- etc.

SFS for a sample of 4 diploid individuals
Case 3

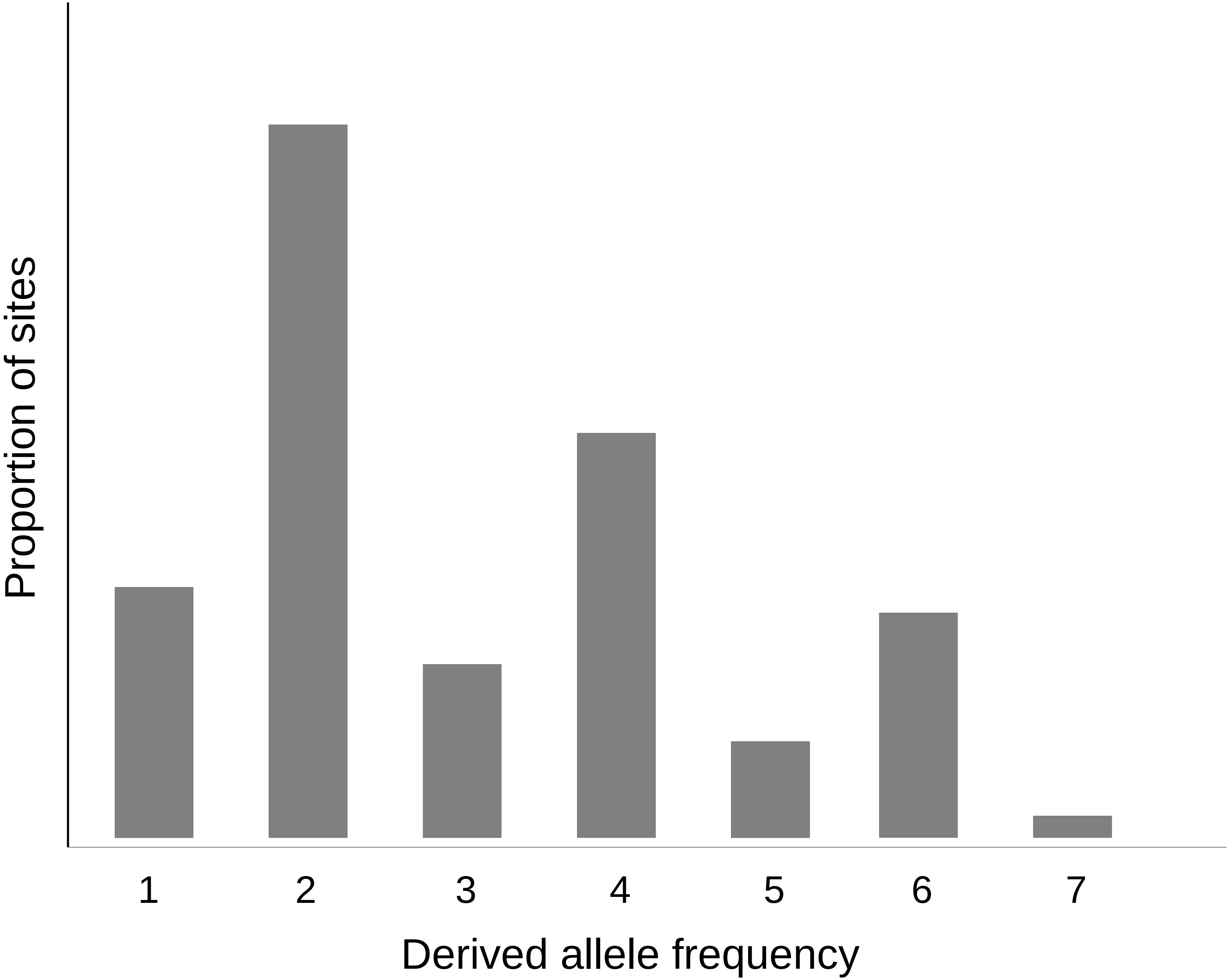


SFS for a sample of 4 diploid individuals
Case 3

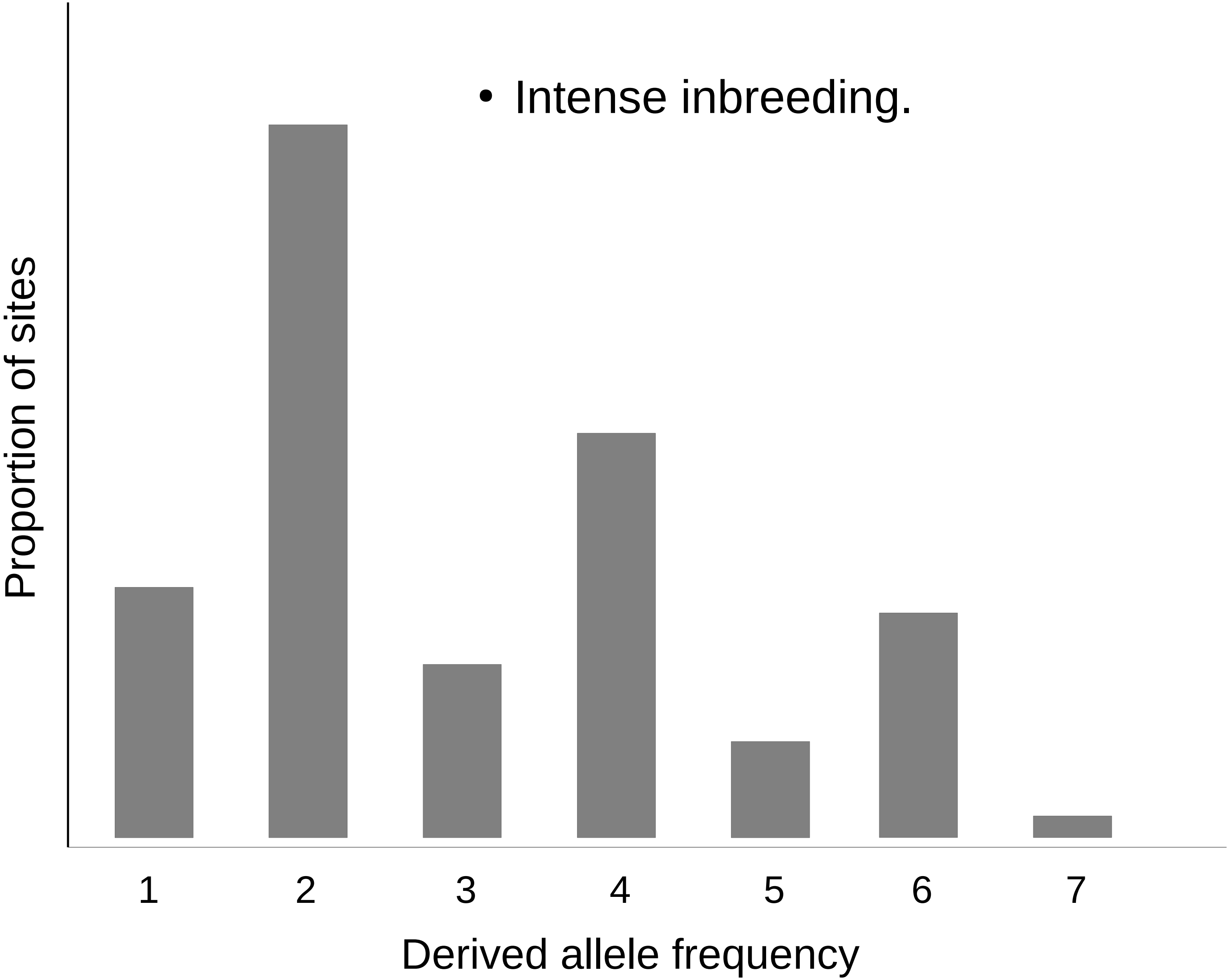


- Misidentification of ancestral and derived allelic state.

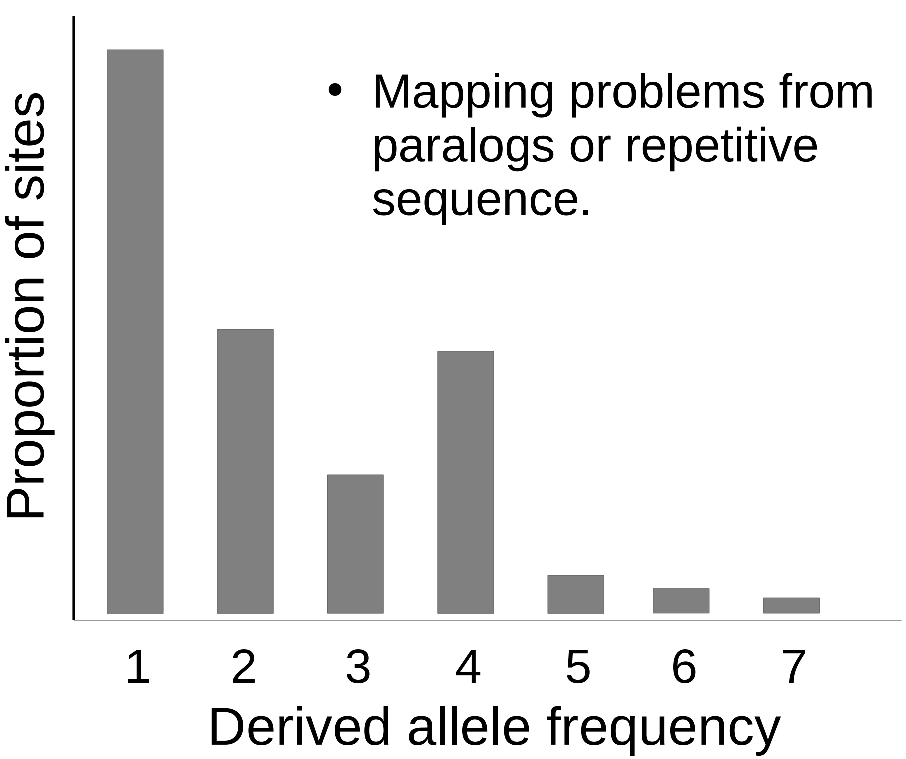
SFS for a sample of 4 diploid individuals
Case 4



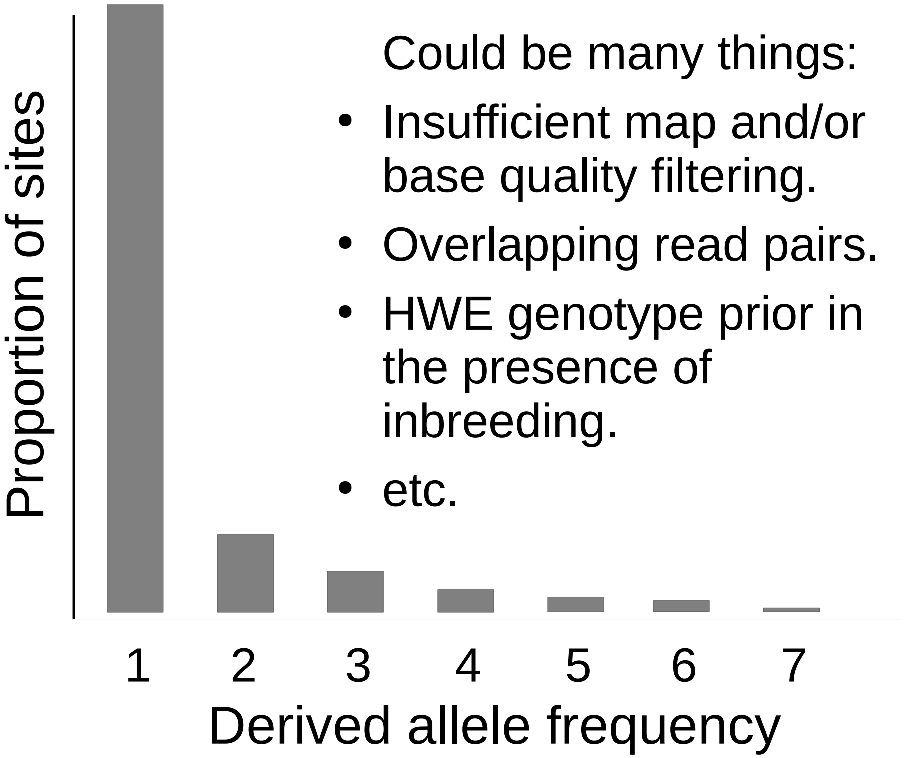
SFS for a sample of 4 diploid individuals
Case 4



- Intense inbreeding.

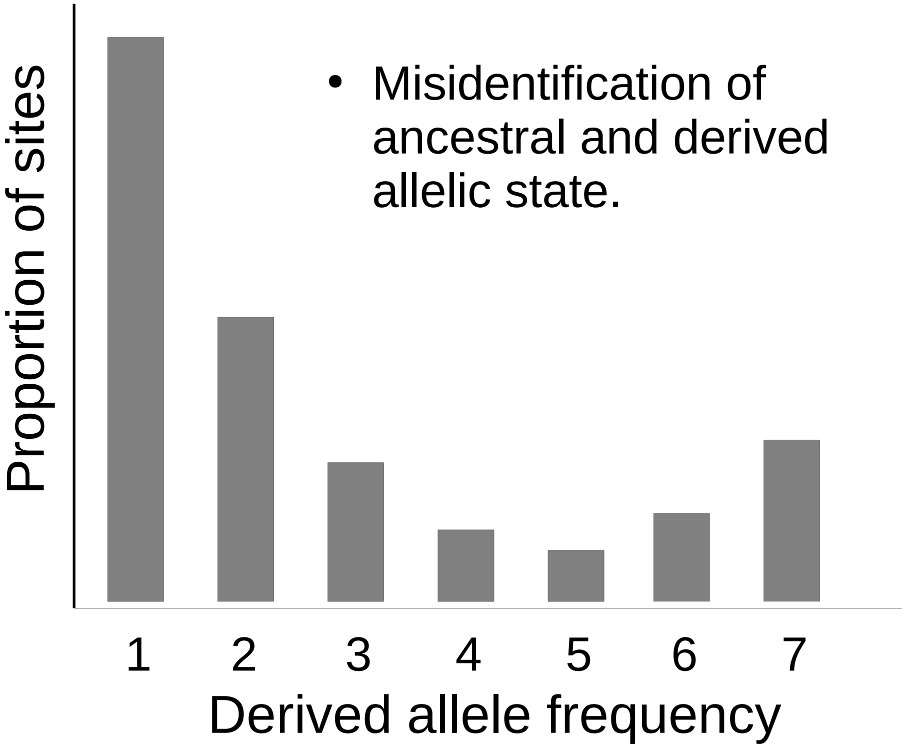


- Mapping problems from paralogs or repetitive sequence.

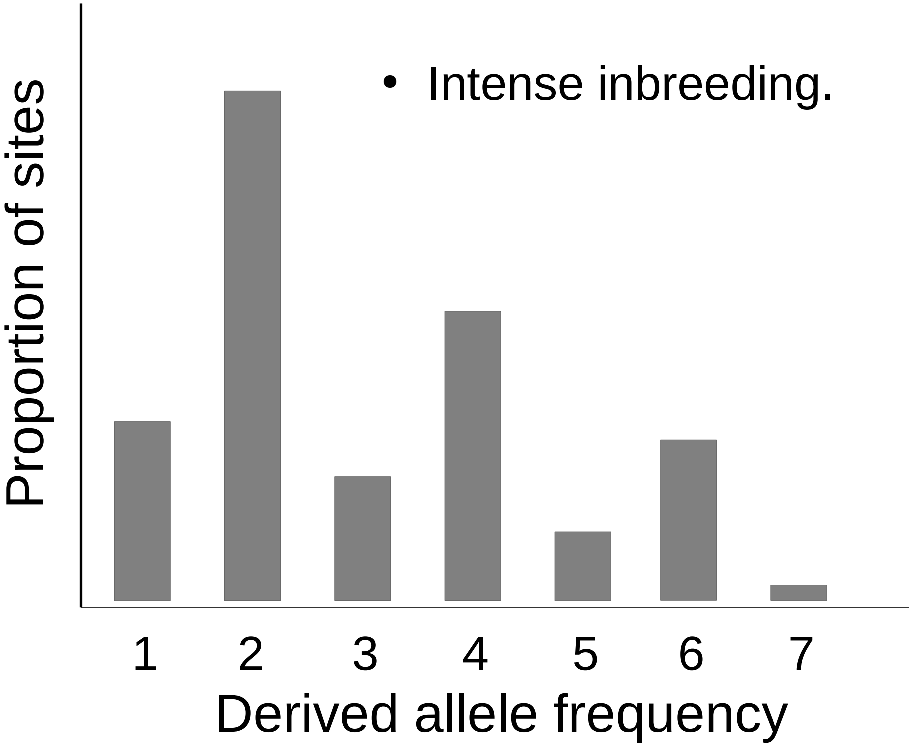


Could be many things:

- Insufficient map and/or base quality filtering.
- Overlapping read pairs.
- HWE genotype prior in the presence of inbreeding.
- etc.



- Misidentification of ancestral and derived allelic state.



- Intense inbreeding.

Likelihood function for the SFS

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

p_k: probability that a random site in the genome has *k* derived alleles
n = diploid sample size

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

Likelihood of the SFS with *known* genotypes considering a single site:

X: observed data
(sequencing reads)

$$\mathrm{P}\left(X, \boldsymbol{G} \middle| \boldsymbol{P}\right) = \sum_{k=0}^{2n} \mathrm{P}\left(X, G \middle| K = k\right) \mathrm{P}\left(K = k \middle| \boldsymbol{P}\right)$$

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

K: number of derived alleles in genotype vector ***G***

Likelihood of the SFS with *known* genotypes considering a single site:

$$\mathrm{P}(X,\boldsymbol{G}|\boldsymbol{P})=\sum_{k=0}^{2n}\boxed{\mathrm{P}(X,\boldsymbol{G}|K=k)}\underbrace{\boxed{\mathrm{P}(K=k|\boldsymbol{P})}}_{=p_k}$$

$$\Big\downarrow$$

$$\mathrm{P}(a,b|c)=\mathrm{P}(a|b)\,\mathrm{P}(b|c)$$

$$\Big\downarrow$$

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

$$=\sum_{k=0}^{2n}\boxed{\mathrm{P}(X|\boldsymbol{G})\,\mathrm{P}(\boldsymbol{G}|K=k)}\,\mathrm{P}(K=k|\boldsymbol{P})$$

Likelihood of the SFS with *known* genotypes considering a single site:

$$\mathrm{P}(X,\boldsymbol{G}|\boldsymbol{P})=\sum_{k=0}^{2n}\boxed{\mathrm{P}(X,\boldsymbol{G}|K=k)}\underbrace{\boxed{\mathrm{P}(K=k|\boldsymbol{P})}}_{=p_k}$$

$$\downarrow\ \mathrm{P}(a,b|c)=\mathrm{P}(a|b)\mathrm{P}(b|c)$$

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

$$=\sum_{k=0}^{2n}\boxed{\mathrm{P}(X|\boldsymbol{G})\mathrm{P}(\boldsymbol{G}|K=k)}\mathrm{P}(K=k|\boldsymbol{P})$$

↓ Consider the each individual’s data and genotype independently.

$$=\sum_{k=0}^{2n}\boxed{\left(\prod_{i=1}^{n}\mathrm{P}(X_i|G_i)\right)\mathrm{P}(\boldsymbol{G}|K=k)}\mathrm{P}(K=k|\boldsymbol{P})$$

Likelihood of the SFS with *known* genotypes considering a single site:

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

Probability based on the number of ways to have *k* derived alleles in ***G*** out of the total number of ways to to arrange *k* derived alleles among 2*n* sampled chromosomes. → **Assumes HWE**

Example for 4 diploid individual with derived allele *A* and ancestral allele *a*.

k = 4

m = # heterozygotes = 2

AA	aA	aa	aA
AA	Aa	aa	aA
AA	aA	aa	Aa
AA	Aa	aa	Aa

$2^m = 2^2$

= 4 combinations

$$P(\mathbf{G}|K=k)=\frac{2^m}{\binom{2n}{k}}=\frac{2^2}{\binom{8}{4}}=\frac{4}{70}$$

$$\mathrm{P}(X,\boldsymbol{G}|\boldsymbol{P})=\sum_{k=0}^{2n}(\prod_{i=1}^{n}\mathrm{P}(X_i|G_i))\,p(\boldsymbol{G}|K=k)\,\mathrm{P}(K=k|\boldsymbol{P})$$

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

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

$$\sum_{k=0}^{2n}\mathrm{P}(K=k|\boldsymbol{P})\sum_{G_1\in\{0,1,2\}}\cdots\sum_{G_n\in\{0,1,2\}}\mathrm{P}(G|K=k)\prod_{i=1}^{n}\mathrm{P}(X_i|G_i)$$

Summing over the uncertainty of the identity of genotypes contained in **G**

Likelihood of **P** considering site *v*

$$\mathrm{P}\left(X^v|\boldsymbol{P}\right)=$$

$$\sum_{k=0}^{2n}\mathrm{P}\left(K=k|\boldsymbol{P}\right)\sum_{G_1\in\{0,1,2\}}\cdots\sum_{G_i\in\{0,1,2\}}\mathrm{P}\left(G^v|K=k\right)\prod_{i=1}^{n}\mathrm{P}\left(X_i^v|G_i^v\right)$$

Assume sites are independent

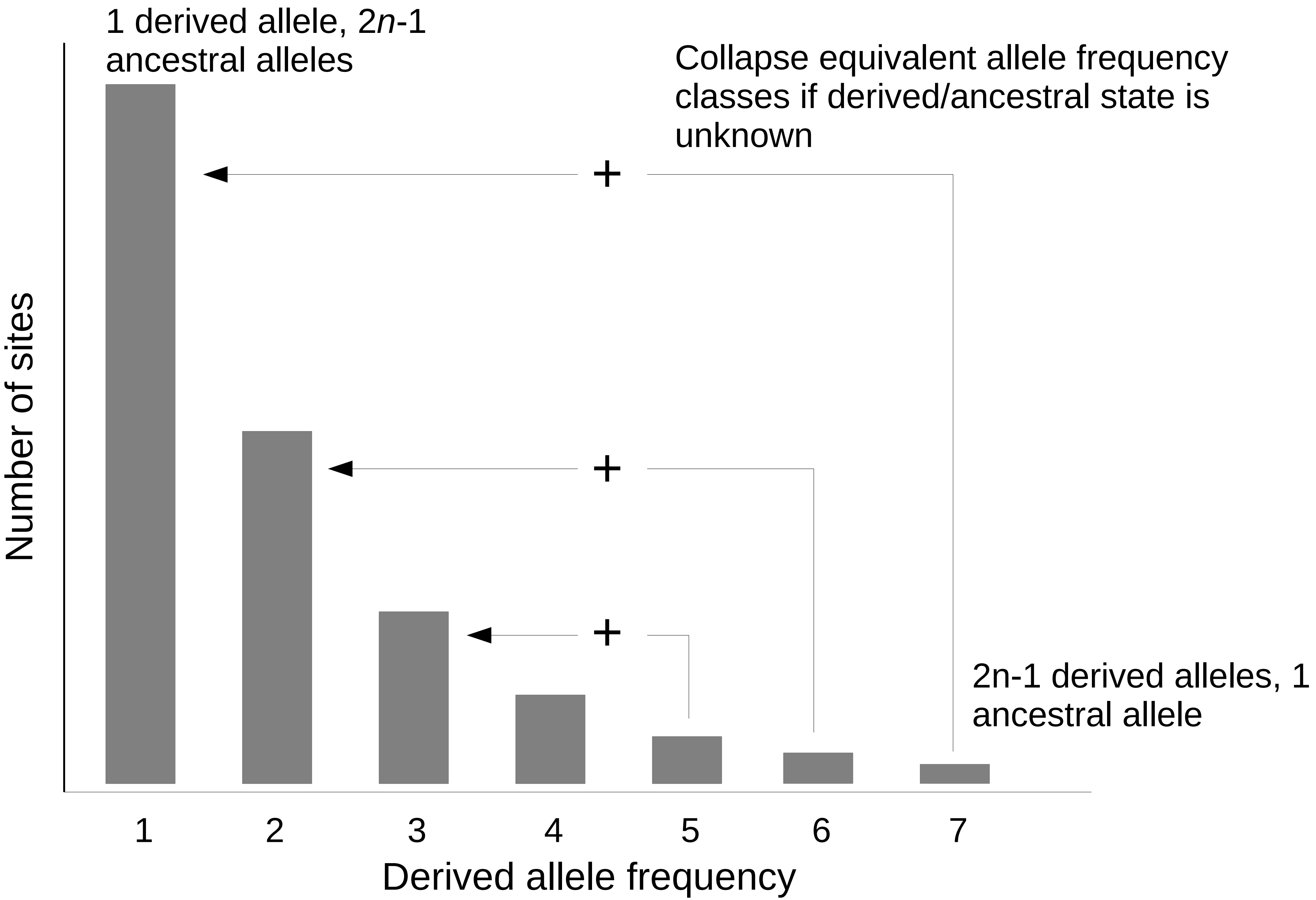
Likelihood of **P** considering all sites (the SFS)

$$\mathrm{P}\left(X|\boldsymbol{P}\right)=$$

$$\prod_{v=1}^{\text{all sites}}\sum_{k=0}^{2n}\mathrm{P}\left(K=k|\boldsymbol{P}\right)\sum_{G_1\in\{0,1,2\}}\cdots\sum_{G_d\in\{0,1,2\}}\mathrm{P}\left(G^v|K=k\right)\prod_{d=1}^{n}\mathrm{P}\left(X_i^v|G_i^v\right)$$

Product over all sites

Folding the SFS when ancestral state is unknown

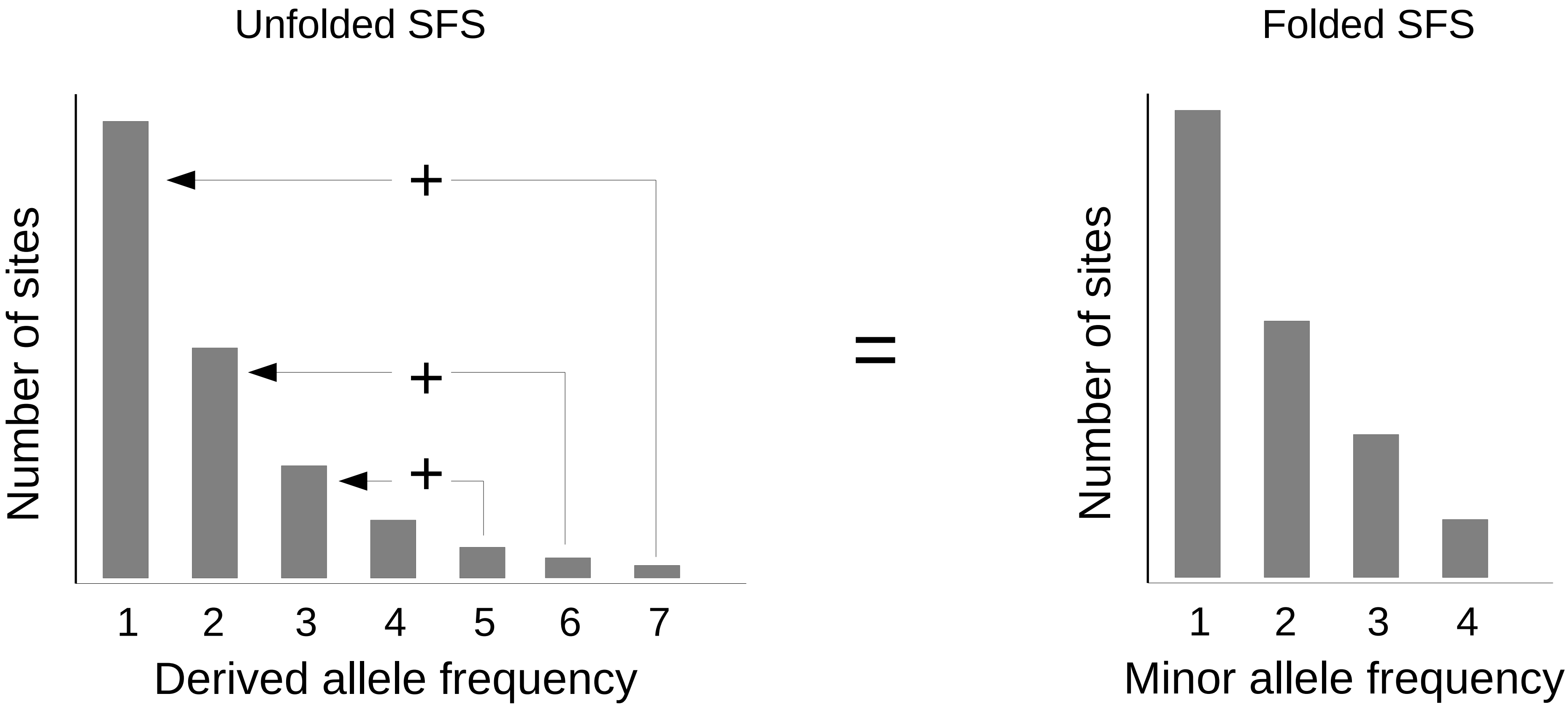


1 derived allele, 2*n*-1 ancestral alleles

Collapse equivalent allele frequency classes if derived/ancestral state is unknown

2n-1 derived alleles, 1 ancestral allele

Folding the SFS when ancestral state is unknown



Unfolded SFS

Folded SFS

How to estimate posterior probabilities of allele frequencies: allele frequency likelihoods

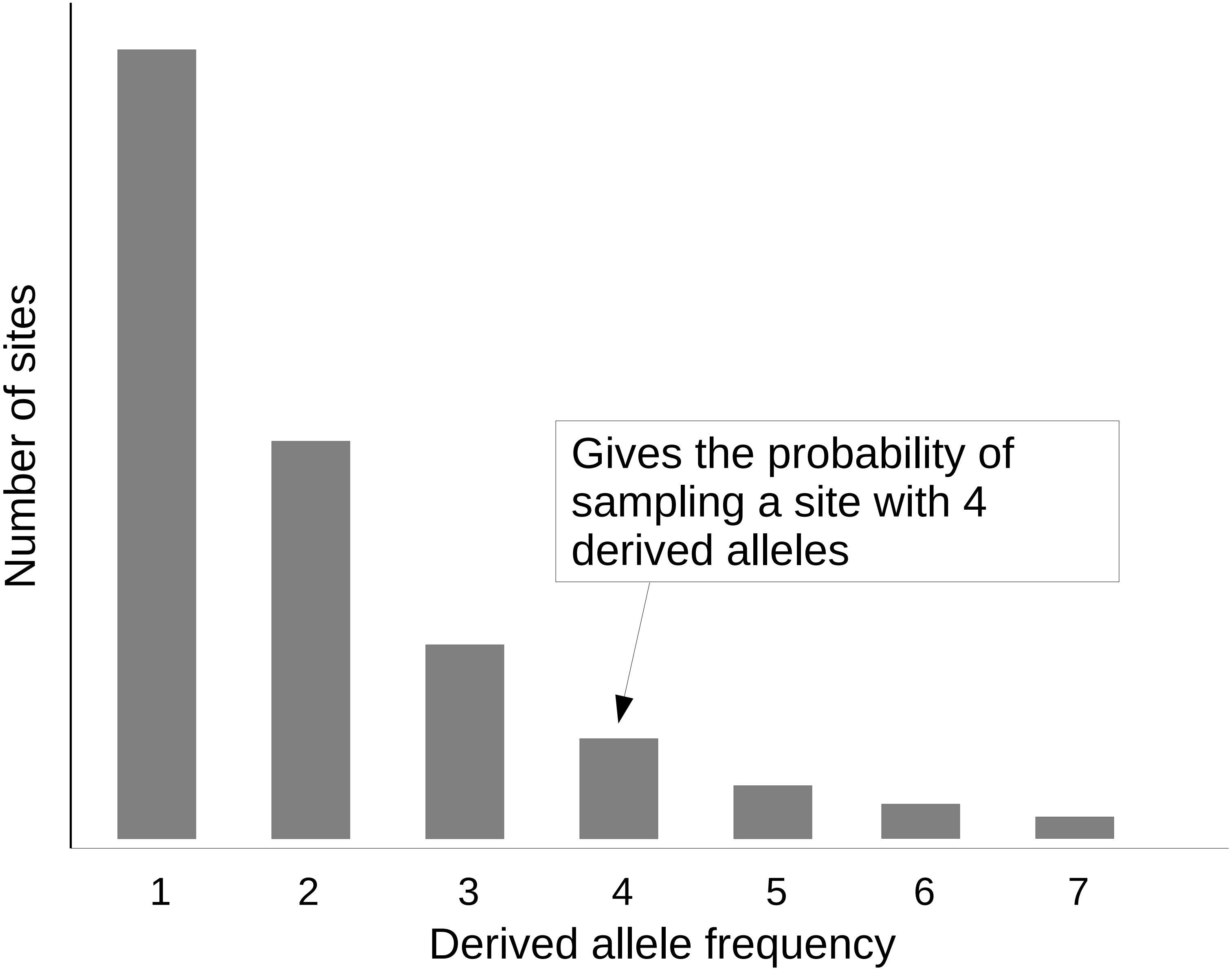
ANGSD -doSaf 1: Table of allele frequency likelihoods

	0	**1**	**2**	**3**	**4**	**.**	**.**	**.**	**2*n***
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
.									
.									
.									
.									
.									
.									
.									
site*m*	**0.00**	**-8.62**	**-19.22**	**-30.67**	**-43.27**				**-626.78**

← # derived alleles

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

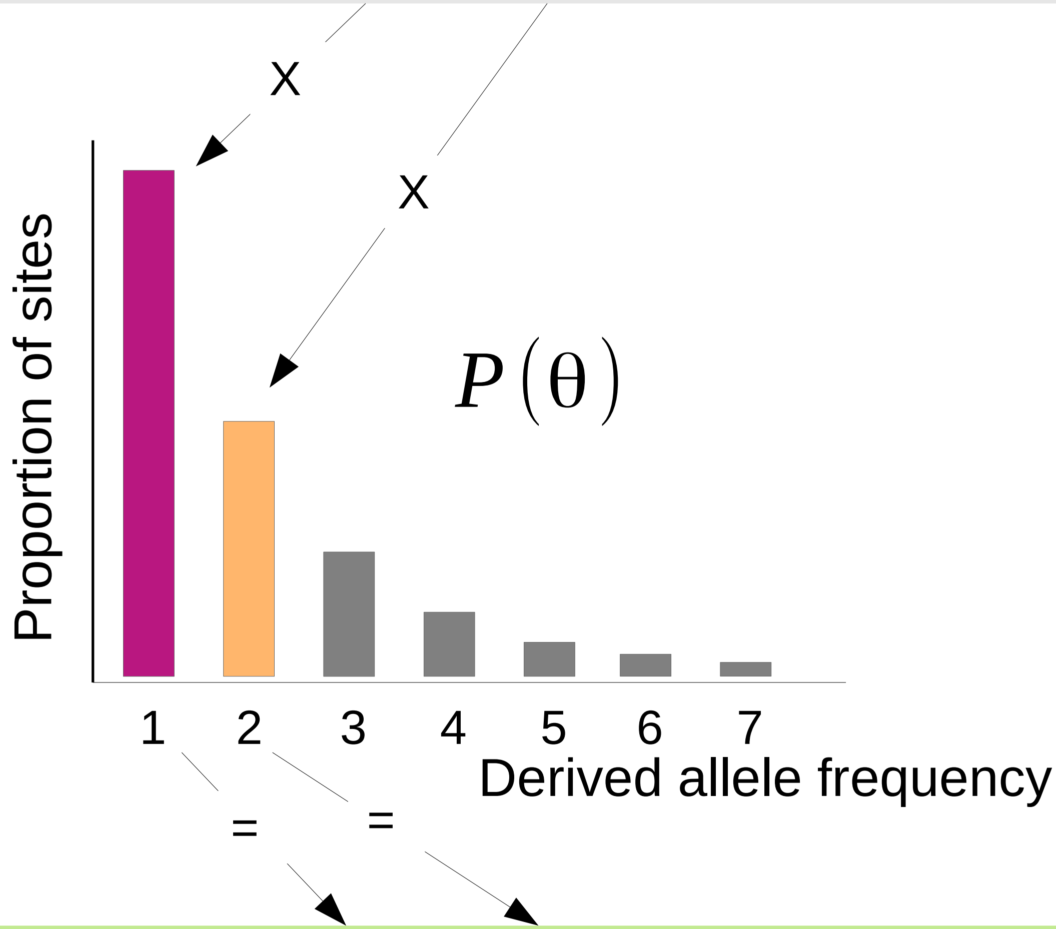
How to estimate posterior probabilities of allele frequencies: Allele frequency prior probabilities



Gives the probability of sampling a site with 4 derived alleles

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

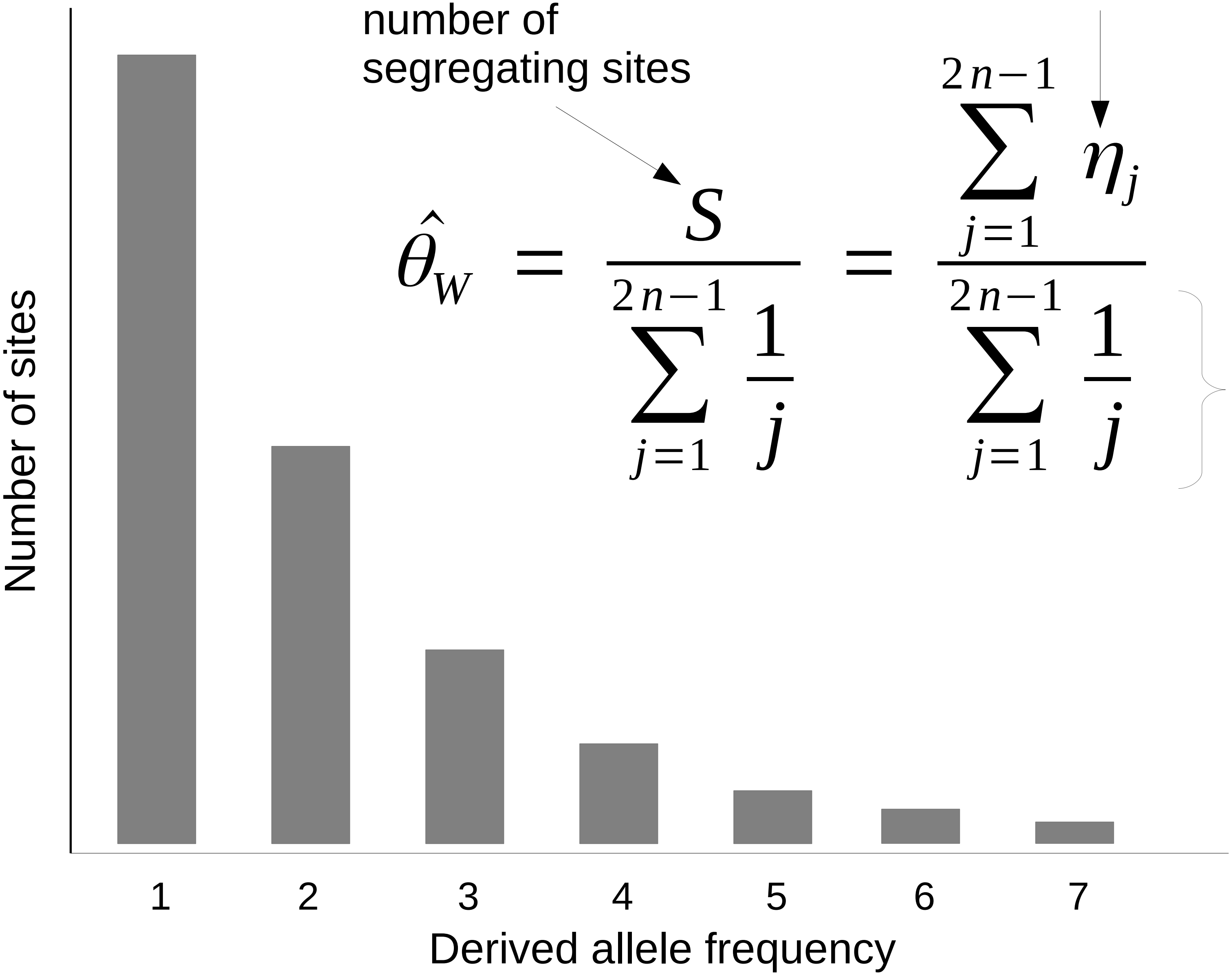
	0	1	2	3	4	. . .	2*n*	
site1	0.00	-2.24	-4.53	-6.99	-9.63		-232.69	$P(X\|\theta)$
site2	0.00	-2.24	-4.53	-6.99	-9.63		-232.69	



	0	1	2	3	4	. . . .	2*n*	
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	

Population genetic summary statistics derived from the SFS

Estimating diversity from the SFS



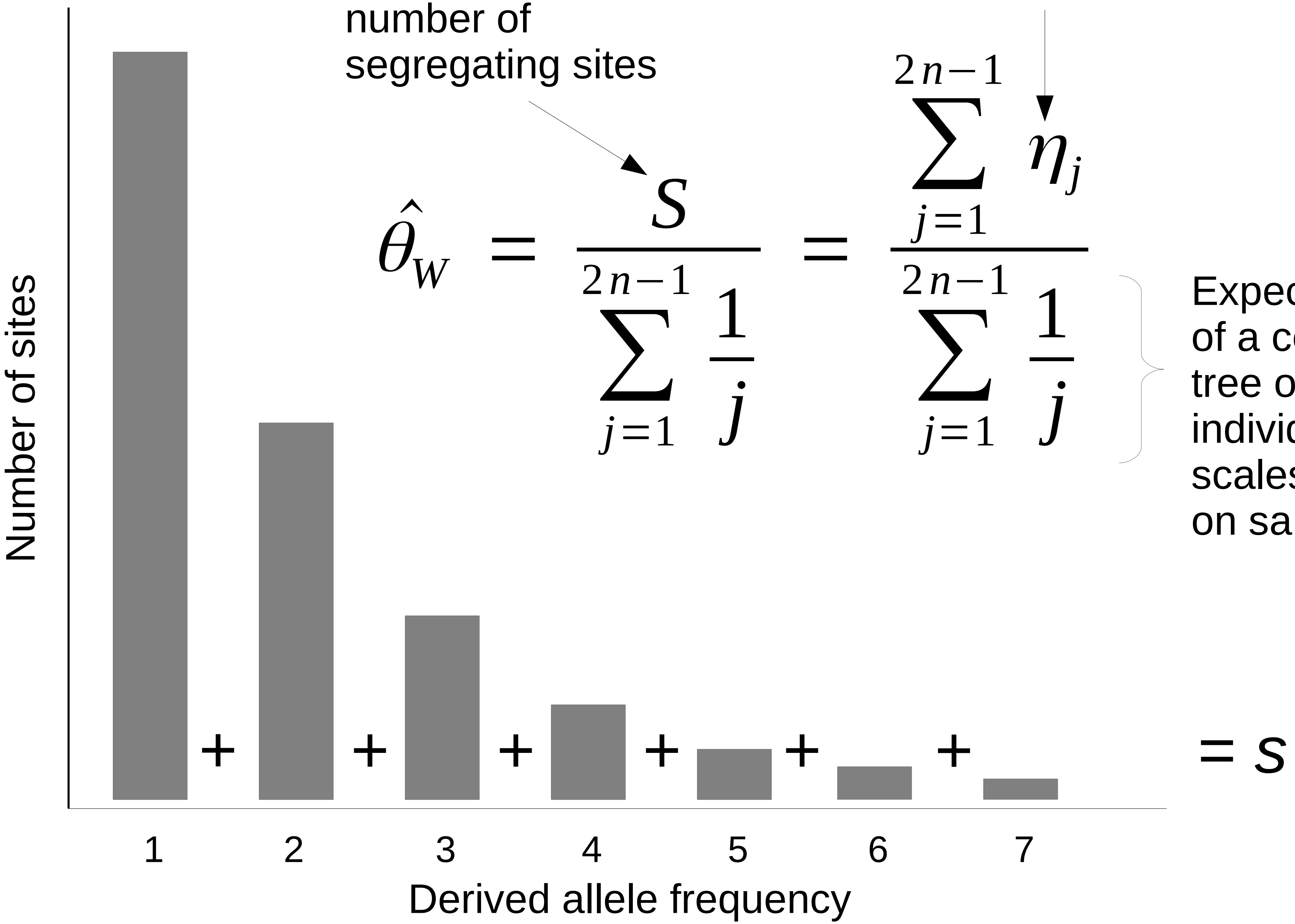
number of segregating sites

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* diploid individuals. This scales *S* based on sample size.

Estimating diversity from the SFS



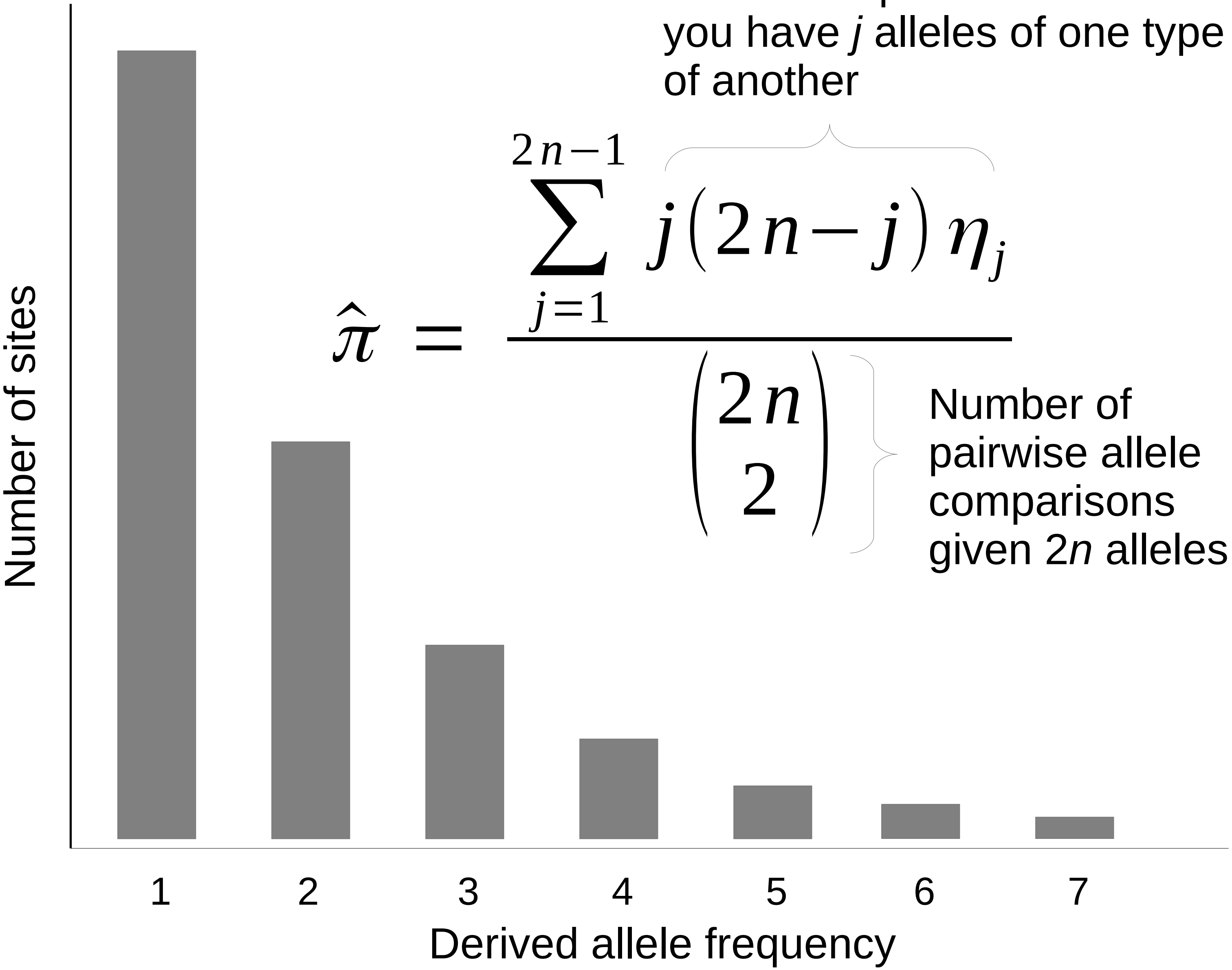
$$\hat{\theta}_W = \frac{S}{\sum_{j=1}^{2n-1} \frac{1}{j}} = \frac{\sum_{j=1}^{2n-1} \eta_j}{\sum_{j=1}^{2n-1} \frac{1}{j}}$$

number of segregating sites

number of sites with *j* derived alleles

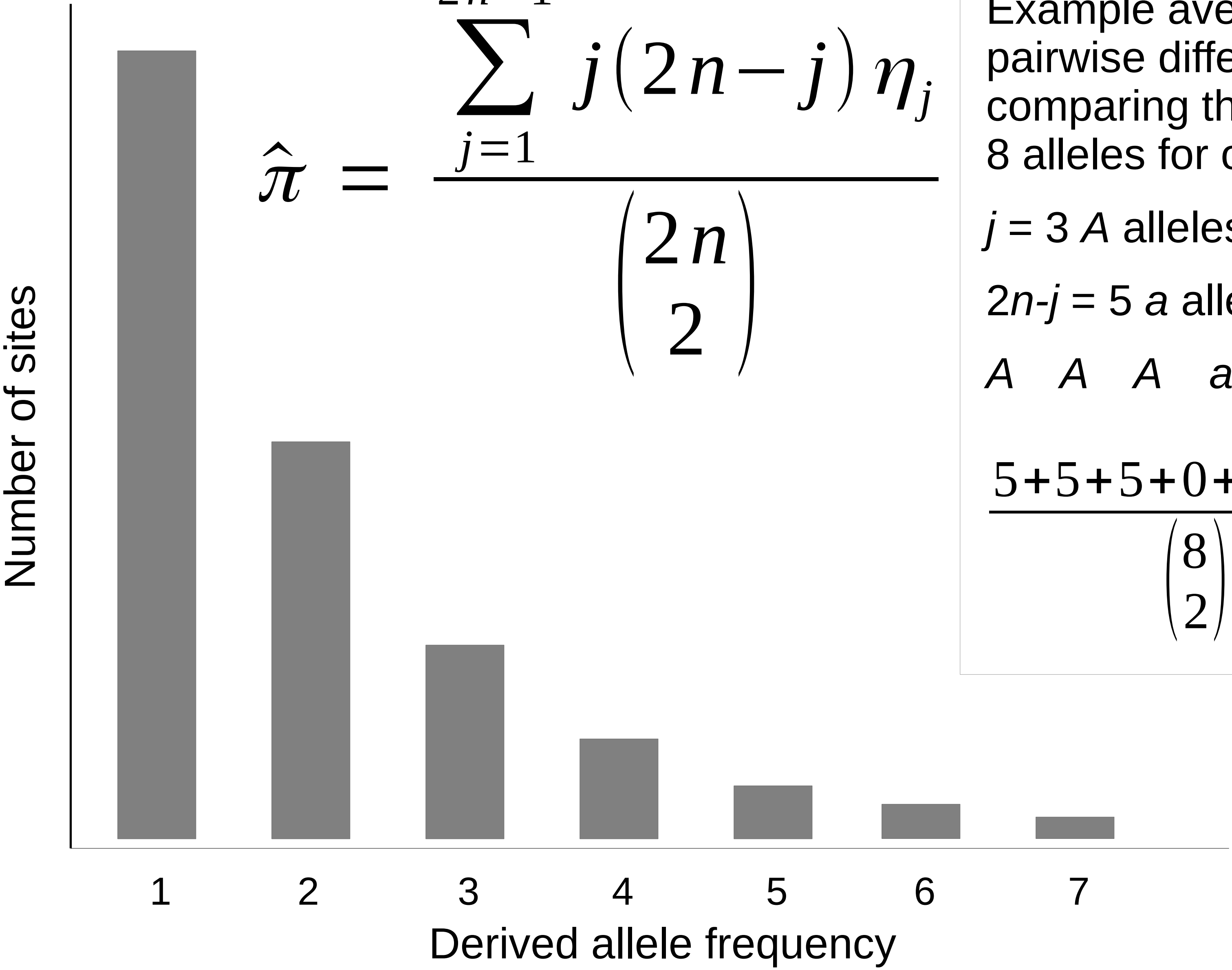
Expected length of a coalescent tree of *n* diploid individuals. This scales *S* based on sample size.

Estimating diversity from the SFS



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

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 for one site:

j = 3 *A* alleles

2*n*-*j* = 5 *a* alleles*:*

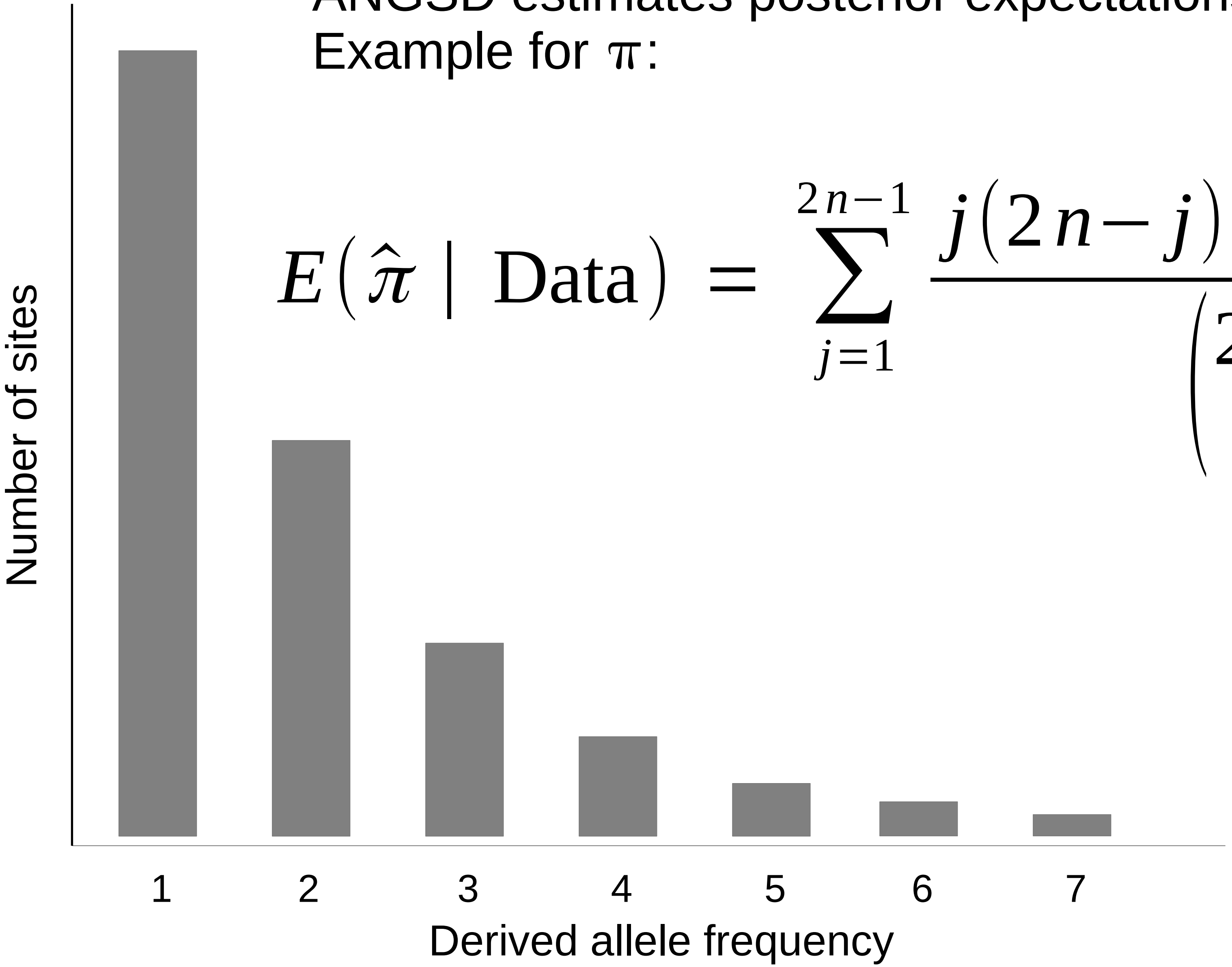
A *A* *A* *a* *a* *a* *a* *a*

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

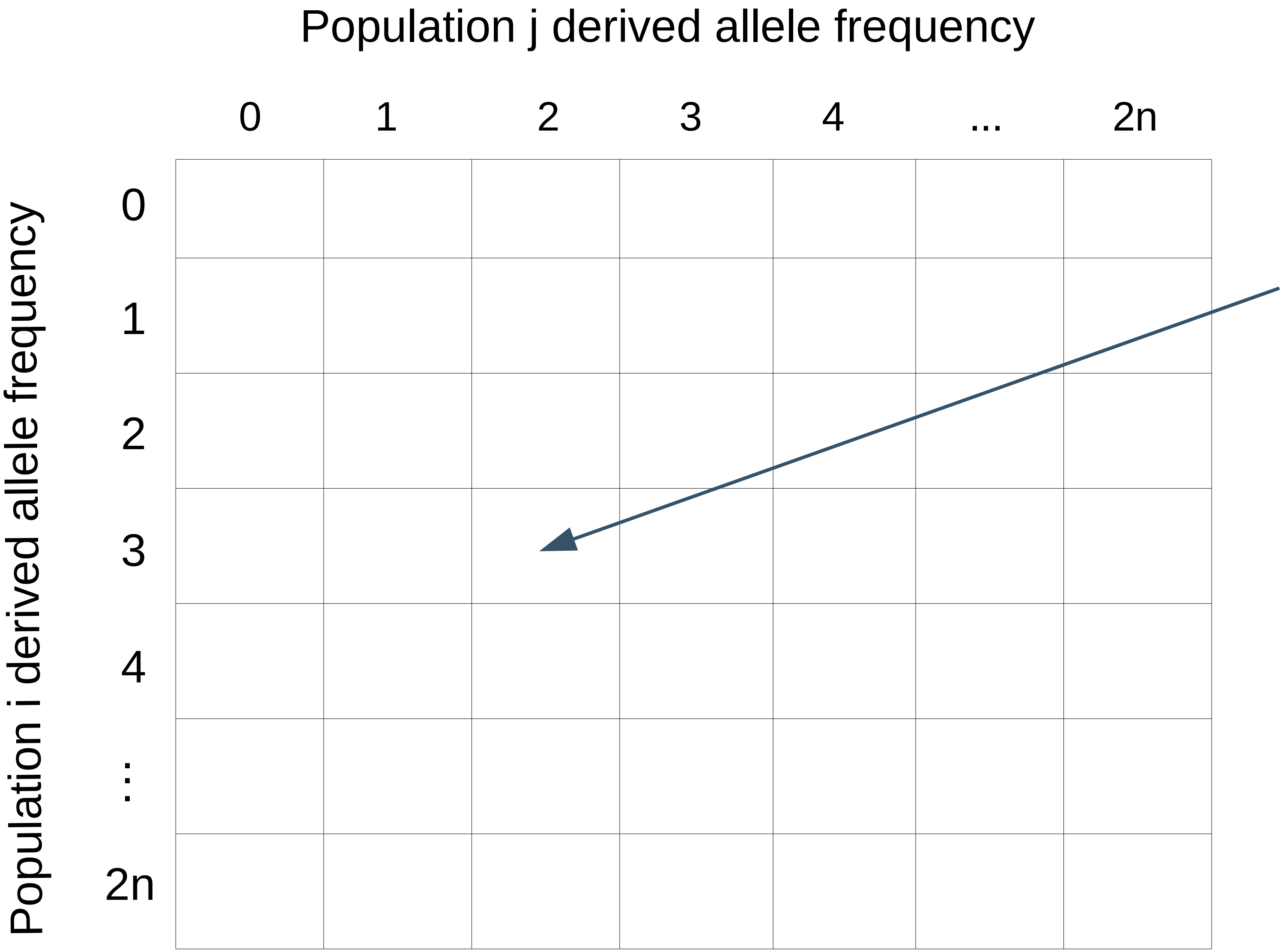
Estimating diversity from the SFS

ANGSD estimates posterior expectations of θ.
Example for π:

$$E\left(\hat{\pi} \mid \mathrm{Data}\right) = \sum_{j=1}^{2n-1} \frac{j\left(2n-j\right)\mathrm{E}\left(\eta_j \middle| \mathrm{Data}\right)}{\binom{2n}{2}}$$



2-dimensional SFS



Population j derived allele frequency

Population i derived allele frequency

	0	1	2	3	4	...	2n
0							
1							
2							
3							
4							
⋮							
2n							

Proportion of sites where:

pop *i* has 3 derived alleles

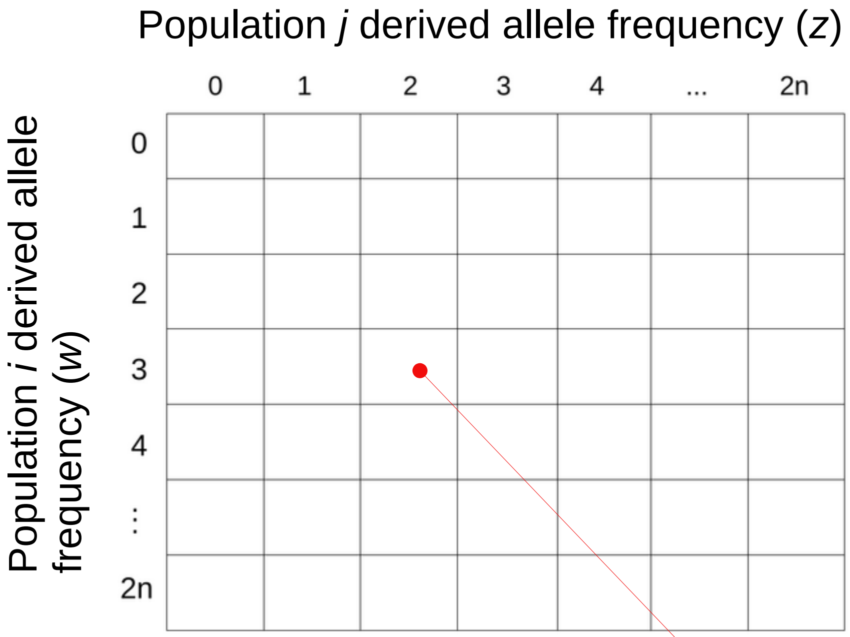
&

pop *j* has 2 derived alleles.

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

Genetic variance between populations

Genetic variance within populations

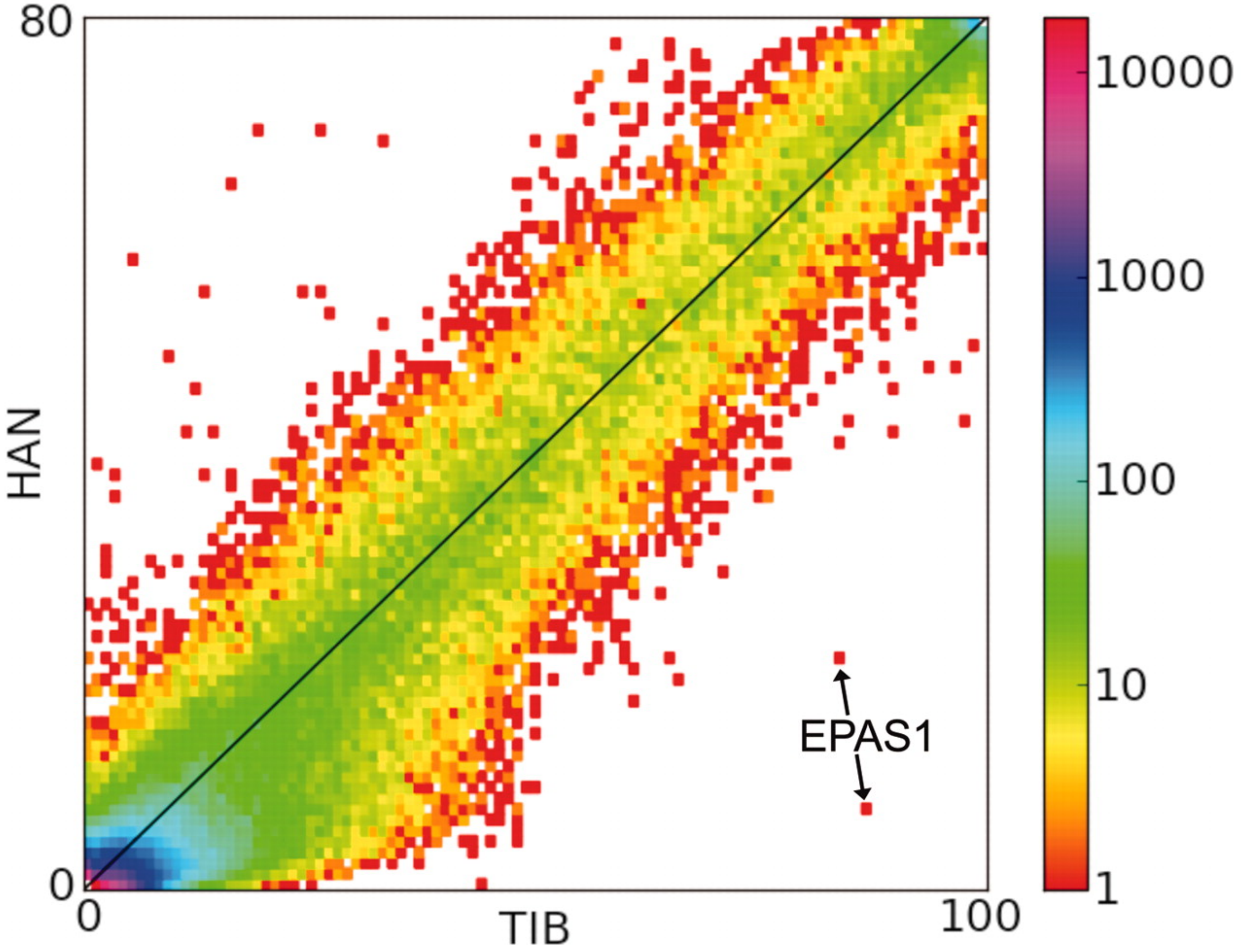

Population *j* derived allele frequency (*z*)
0
1
2
3
4
...
2n
Population *i* derived allele frequency (*w*)
0
1
2
3
4
⋮
2n

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

$$\mathrm{E}(a|X)=\sum_{k=0}^{2n}\sum_{z=0}^{2n} a^{w,z}_{pop\,i,\,pop\,j}\overbrace{\mathrm{P}(X_{i,v}|K_{i,v}=w)}^{\text{Likelihood of } w \text{ derived alleles in pop } i \text{ at site } v}\overbrace{\mathrm{P}(X_{j,v}|K_{j,v}=z)}^{\text{Likelihood of } z \text{ derived alleles in pop } j \text{ at site } v}\boxed{Q^{w,z}_{i,j}}$$

$$\mathrm{E}(b|X)=\sum_{k=0}^{2n}\sum_{z=0}^{2n} b^{w,z}_{pop\,i,\,pop\,j}\mathrm{P}(X_{i,v}|K_{i,v}=k)\mathrm{P}(X_{j,v}|K_{j,v}=z)Q^{w,z}_{i,j}$$

Identifying loci under selection



Yi *et al*. (2010)

Exercise. Estimating the SFS and summary statistics