Describing Variation: Definitions

Parameter: a property of population to be estimated (e.g., $\theta = 4N_e\mu$)

Estimator: an estimate of a population parameter typically derived from a sample of DNA sequences

cannot test all people

Heterozygosity

Heterozygosity is the probability that two alleles drawn randomly from a population will be different alleles

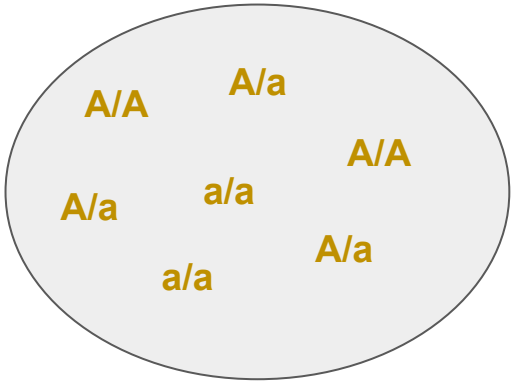
Consider a single bi-allelic locus with allele frequencies p_1 and p_2, where

$p_A + p_a = 1$

The probability of drawing two of the same allele from a population is:

$$\sum p_i^2$$

Where *i* is the ith allele



What is the frequency of the “A” allele (p_A)?

What is the frequency of the “a” allele (p_a)?

Heterozygosity

Heterozygosity is the probability that two alleles drawn randomly from a population will be different alleles

Consider a single bi-allelic locus with allele frequencies p_1 and p_2, where

$p_1 + p_2 = 1$

We calculate heterozygosity, *h*, as

$$h = \frac{n}{n-1}(1 - \sum_{i}^{m} p_i^2)$$

where *n*, is the number of sequences (i.e., chromosomes) in a sample, m is the number of alleles

*n/(n-1) is a correction for sampling bias

Measures of diversity from DNA sequences: π (“pi”)

π (=nucleotide diversity) is a measure of heterozygosity from DNA sequence data

looks for whole genome

Sometimes referred to as θ_π in reference to the parameter (θ; “theta”) which π is an estimator of

π can be calculated from the sum of site heterozygosities as:

$$\pi = \sum_{j=1}^{S} h_j$$

where, S is the number of segregating sites,

h is heterozygosity (defined above)

Measures of diversity from DNA sequences: π (“pi”)

Equivalently, π can calculated from the average number of pairwise differences

$$\pi = \frac{\sum_{i<j} k_{ij}}{n(n-1)/2}$$

differences between i and j

where,

n number of samples sequences,

k_{ij} is the number of differences between sequences i and j

Empirical estimates of nucleotide diversity π

Dividing π by the length (*L*) yields a per site measure of nucleotide diversity

Length includes all sites (both monomorphic and polymorphic)

Division by *L* allows meaningful comparisons of π between different regions of the genome or between different populations/species

Example: calculation of per site nucleotide diversity from the average number of pairwise differences

Sequence pair (ij)	Number of differences (*k*)
1,2	3
1,3	4
2,3	5
Numerator:	12
Denominator:	3
Length	15
π (per site):	0.2667

$$\pi = \frac{\sum_{i<j} k_{ij}}{n(n-1)/2}$$

```
1 TTACAATCCGATCGT
2 TTACGATGCGCTCGT
3 TCACAATGCGATGGA
```

Example: nucleotide diversity by continent

Zhao et al. (2000) sequenced a 10 kb region from many individuals on each continent

Table 3. Nucleotide diversity (%) in different populations and between populations

Population	African	Asian	European	Oceanian
African	0.085			
Asian	0.083	0.075		
European	0.108	0.091	0.077	
Oceanian	0.093	0.079	0.070	0.057

Average number of pairwise differences per site expressed as %, so divide by 100 to get per site estimate (=0.00057)

Zhao et al. (2000) PNAS

Example: nucleotide diversity in Hausa, Italian, Chinese

Voight et al. collected sequence data collected for many loci in three human populations

Hausa (Cameroon/Africa)
Italian (European)
Chinese (East Asian)

Average segregating sites per locus

Average number of pairwise differences

Table 1. Observed summary statistics

Population	$\bar{D}$	$\widehat{\mathrm{Var}[D]}$	$\bar{D}^*$	$\bar{S}$	$\bar{\pi}$, %	$\hat{\rho}$
Hausa	−0.20	0.55	−0.17	11.1	0.110	0.0006
Italian	0.28*	1.19**	0.18	7.1	0.085	0.0003
Chinese	0.18	1.08*	0.05	6.9	0.079	0.0002*

Voight et al. (2005) PNAS

How much nucleotide diversity (π) is there in humans

The typically cited number for π is 0.0001 in humans

That is, on average, a randomly drawn pair of chromosomes sampled from a population will differ at 1 in 1000 bp

Take home question: How many differences do you expect on average between a pair of haploid genomes?

*hint: the human genome is approximately 3 billion bp (in a single haploid set of 23 chromosomes)

Li and Sadler (1991) Genetics

Measures of diversity from DNA sequences: Watterson's θ

Watterson's θ (θ_W) is a measure of nucleotide diversity from the number of segregating sites

$$\theta_W = \frac{S}{a}$$

where, *S* is the number of segregating sites and a is defined as:

$$a = \sum_{i=1}^{n-1} \frac{1}{i}$$

Where, *n* is the number of chromosomes

The population mutation parameter θ

Both π and θ_W are estimators of the population parameter θ

Under a Wright-Fisher model at equilibrium between mutation and genetic drift, the following equality holds:

$E(\pi) = E(\theta_W) = \theta = 4N_e\mu$

where, E(π) is the expectation of π

$E(\theta_W)$ is the expectation of θ_W

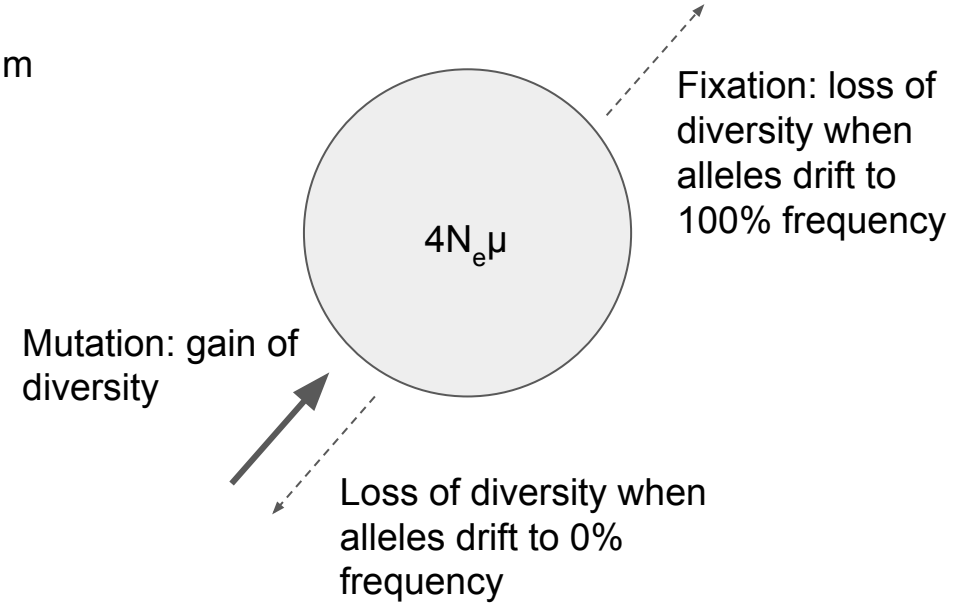
θ is the population mutation parameter

N_e is the effective population size

μ is the mutation rate per generation

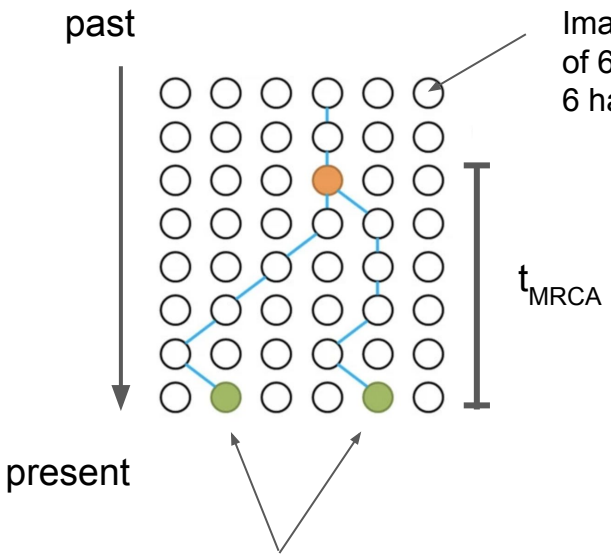
Mutation-drift equilibrium

A population with constant rate of genetic drift (i.e., constant Ne), no selection, and no migration is expected to reach an equilibrium level of nucleotide diversity



$4N_e\mu$ is the expected diversity in a Wright-Fisher population

How much genetic variation do we expect in a sample of DNA sequences from a population?



The two chromosomes sampled from a present-day population “coalesce” 5 generations in the past

Chromosomes in the direct line of ancestry of the two sampled sequences

Most Recent Common Ancestor (MRCA)

Chromosomes randomly sampled from present-day population

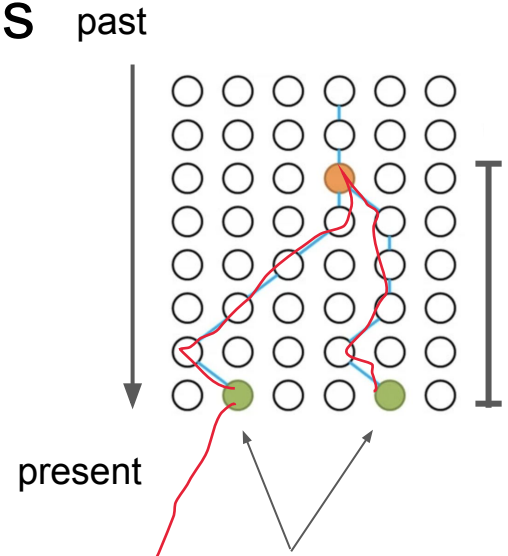
The expected number of nucleotide differences between a pair of sampled chromosomes

- The expected (i.e., mean) time to coalescence of two sequences drawn at random from a Fisher-Wright population is $2N_e$ generations

$$E[t_{MRCA}] = 2N_e$$

- If mutations occur at a rate of μ mutations per bp per generation, then we can calculate the expected number of mutations between a pair of sequences sampled in the present:

$$\theta = 2 * 2N_e * \mu = 4N_e\mu$$



The population mutation parameter θ (=”theta”)

How can we quantify the amount of genetic variation in a population?

The population mutation parameter, θ (=”theta”) is a theoretical value that quantifies diversity in a population

θ is the amount of genetic variation in a hypothetical Wright-Fisher population at mutation-drift equilibrium

Population geneticists estimate θ in real populations

Goal is to (1) have a measure of genetic diversity that both connects empirical observations to simple theoretical predictions (2) that can be compared among gene regions, among populations, or even among species

π and θ_W are sensitive to allele frequencies (but in different ways)

π is especially sensitive to intermediate frequency polymorphisms, but relatively insensitive to high or low frequency polymorphisms

θ_W is sensitive to all polymorphisms (irrespective of allele frequency)

Key point: understanding the different sensitivities of π and θ_W is key to gaining insight into Tajima's D and other summaries of the site frequency spectrum

The site frequency spectrum (SFS)

The SFS is a histogram of allele frequencies observed in a sample of sequences

The SFS represent:

(1) minor allele frequencies (“folded SFS”)

(2) derived allele frequencies (“unfolded SFS”)

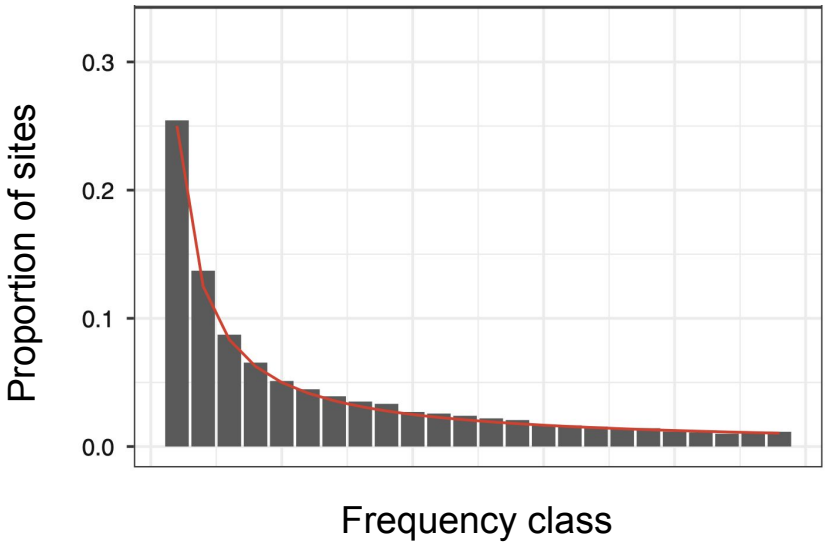
The site frequency spectrum (SFS)

The SFS is a histogram of allele frequencies observed in a sample of sequences

The SFS represent:

(1) minor allele frequencies (“folded SFS”)

(2) derived allele frequencies (“unfolded SFS”)



Example: calculating the folded SFS

(1) Calculate minor allele frequencies for all SNPs

(2) Count how many SNPs fall into each minor allele frequency class

*note: the folded, or minor allele frequency, spectrum will always have a max allele frequency of 0.5

```
1 TCAATCCCCGT
2 TCAAAGCCGGA
3 TCAATGCCGGA
4 TTAATGACCAA
5 TTTGTGCTCGA
6 ATAATGCTCGA
```

	Count of sites
Frequency class 1/6:	
Frequency class 2/6:	
Frequency class 3/6:	

Inferring ancestral and derived alleles

What is the ancestral state at position 2 of the multiple sequence alignment?

Example: use parsimony criterion (=accept ancestral state requiring fewest mutational steps)

```
1 TCAATCCCCGT
2 TCAAAGCCGGA
3 TCAATGCCGGA
4 TTAATGACCAA
5 TTTGTGCTCGA
6 ATAATGCTCGA
Outgroup TTACAGCTCAA
```

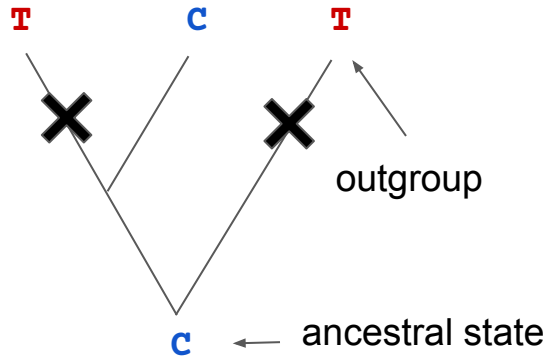
Inferring ancestral and derived alleles

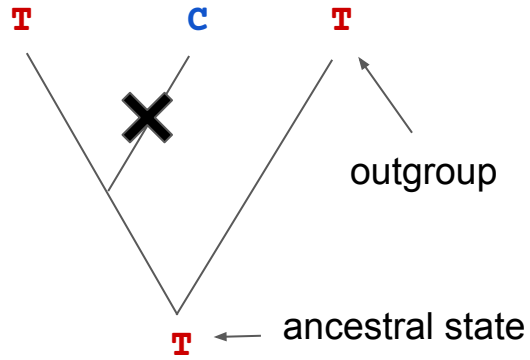
What is the ancestral state at position 2 of the multiple sequence alignment?

Example: use of a parsimony criterion to infer the ancestral and derived alleles

*parsimony accepts scenario with fewest number of mutational steps

```
1          TCAATCCCCGT
2          TCAAAGCCGGA
3          TCAATGCCGGA
4          TTAATGACCAA
5          TTTGTGCTCGA
6          ATAATGCTCGA
Outgroup   TTACAGCTCAA
```





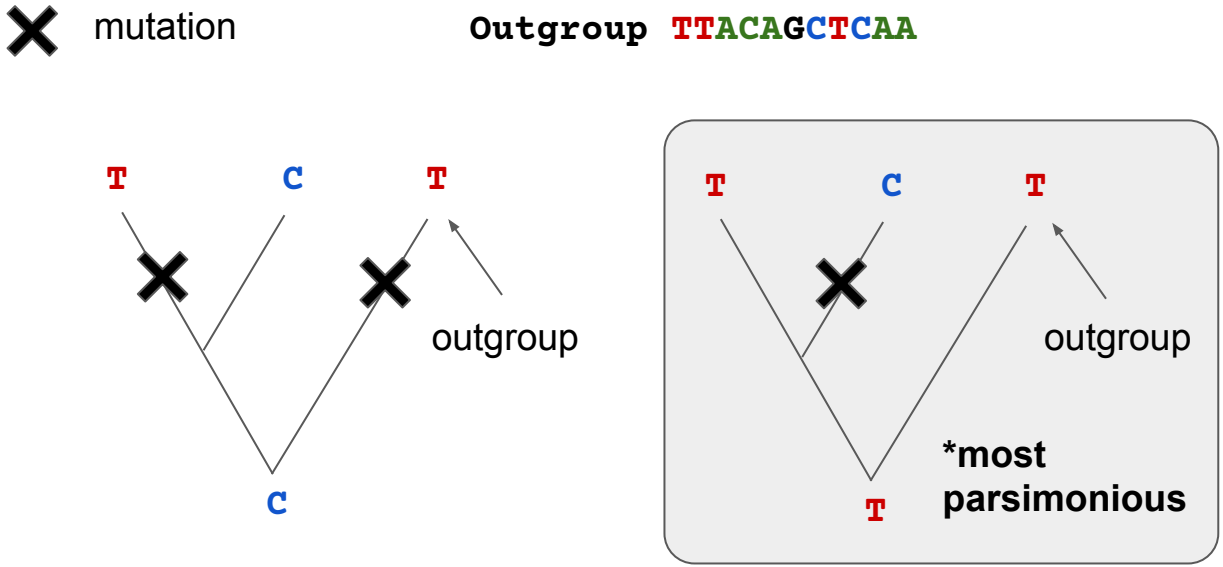
Inferring ancestral and derived alleles

What is the ancestral state at position 2 of the multiple sequence alignment?

Example: use of a parsimony criterion to infer the ancestral and derived alleles

*parsimony accepts scenario with fewest number of mutational steps

```
1          TCAATCCCCGT
2          TCAAAGCCGGA
3          TCAATGCCGGA
4          TTAATGACCAA
5          TTTGTGCTCGA
6          ATAATGCTCGA
Outgroup   TTACAGCTCAA
```



Example: Unfolded SFS

How many sites have derived allele frequencies in each frequency class?

```
1 TCAATCCCCGT
2 TCAAAGCCGGA
3 TCAATGCCGGA
4 TTAATGACCAA
5 TTTGTGCTCGA
6 ATAATGCTCGA
Outgroup TTACAGCTCAA
```

Count

Frequency class: 1/6

Frequency class: 2/6

Frequency class: 3/6

Frequency class: 4/6

Frequency class: 5/6

Other estimators of θ based on part of the SFS

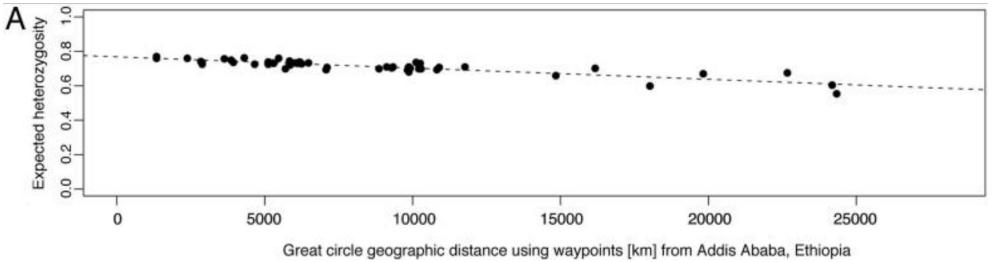
Example: Fay and Wu's H

Diversity in African and non-African human populations

Heterozygosity declines in human populations that are farther from East Africa

Geographical distances measured to Addis Ababa, Ethiopia

Consistent with a serial bottleneck model



Ramachandran et al. (2005) PNAS

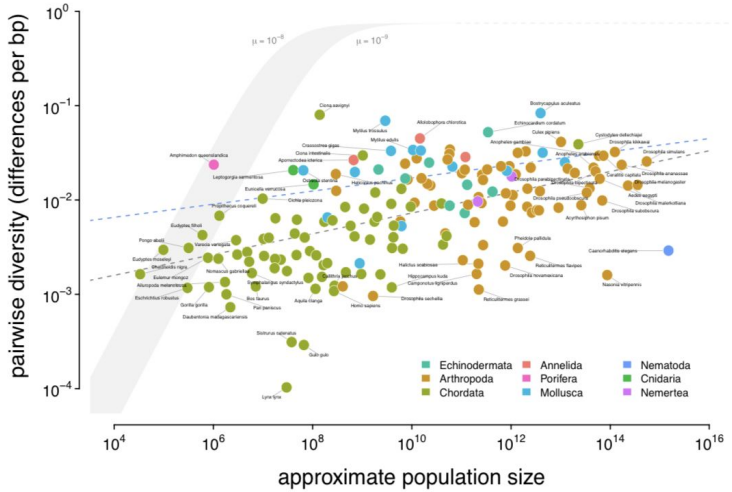
Lewontin’s Paradox

Species with large populations (e.g., species of marine phytoplankton) are not as genetically diverse as expected at mutation-drift equilibrium

Two explanations:

(1) Greater effects of linked selection in abundant species

(2) Abundant species more likely to experience non-equilibrium processes (e.g. fluctuations in population size)



Buffalo (2021) eLife