

# **Describing Variation & Patterns of Diversity**

BIOL 1435

January 31, 2023

## Icebreaker

---

Name, Year, Major, and what was the last song you listened to today?

# Overview

---

1. ATGC's of life & encoding DNA
2. Measures of sequence diversity
3. In class coding exercise

# Overview

---

**1. ATGC's of life & encoding DNA**

2. Measures of sequence diversity

3. In class coding exercise

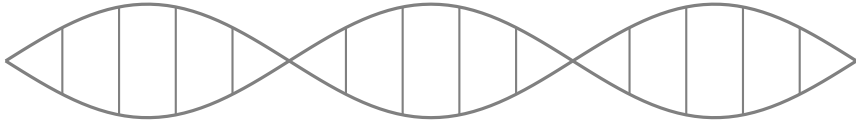
# DNA consists of four nucleotides

---

ATGC

# DNA is organized onto chromosomes

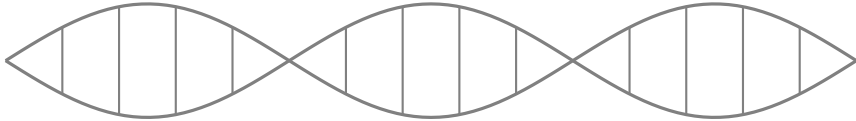
---



# Ploidy (#N): number of sets of chromosomes

---

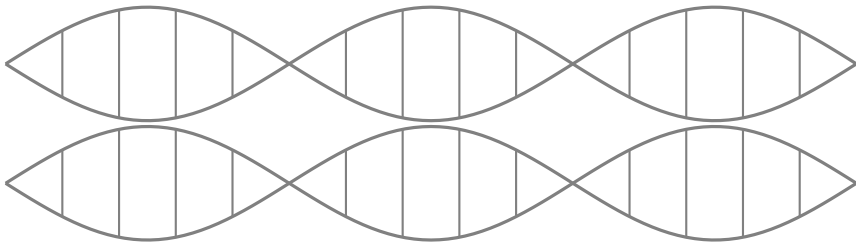
$N = \text{haploid}$



## Ploidy (#N): number of sets of chromosomes

---

$2N = \text{diploid}$

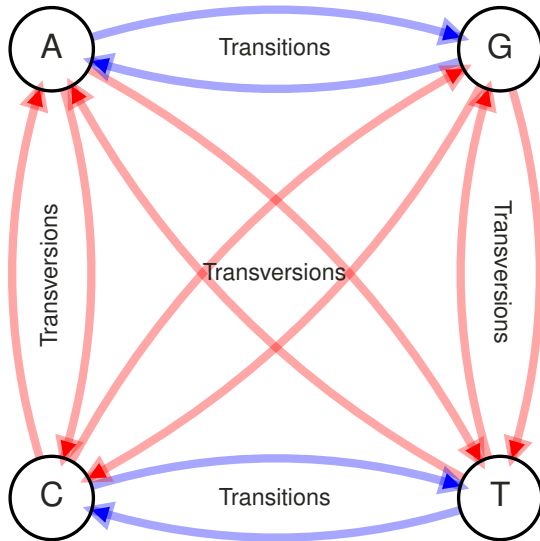




**Q: How does genetic variation  
arise?**

# A: Mutations

---



## How do we encode DNA?

---

$$m \text{ (sites)} \times n \text{ (chromosomes)}$$

# How do we encode DNA?

---

	<i>ind<sub>1</sub></i>	<i>ind<sub>2</sub></i>	<i>ind<sub>3</sub></i>	<i>ind<sub>4</sub></i>	<i>ind<sub>5</sub></i>
<i>pos<sub>1</sub></i>	T	T	T	T	T
<i>pos<sub>2</sub></i>	C	G	G	C	G
<i>pos<sub>3</sub></i>	A	T	A	T	T
<i>pos<sub>4</sub></i>	G	G	G	G	C
<i>pos<sub>5</sub></i>	T	A	A	A	A

# Genotype matrices

---

$$\begin{array}{c} \text{pos}_1 \\ \text{pos}_2 \\ \text{pos}_3 \\ \text{pos}_4 \\ \text{pos}_5 \end{array} \begin{array}{c} \text{ind}_1 \\ \text{ind}_2 \\ \text{ind}_3 \\ \text{ind}_4 \\ \text{ind}_5 \end{array} \begin{bmatrix} T & T & T & T & T \\ C & G & G & C & G \\ A & T & A & T & T \\ G & G & G & G & C \\ T & A & A & A & A \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

0 = reference or ancestral allele

1 = alternative or derived allele

# Genotype matrices

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

# Genotype matrices

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

# Some terminology...

---

- Single nucleotide polymorphism (SNP)
- Single nucleotide variant (SNV)
- Variant site
- Segregating site



# Overview

---

1. ATGC's of life & encoding DNA

**2. Measures of sequence diversity**

3. In class coding exercise

## How would you summarize this genotype matrix?

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

# Measurements of genetic variation

---

- Segregating sites ( $S$ )
- Site frequency spectrum (SFS)
- Gene diversity ( $h$  &  $H$ )
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity ( $\Pi$  &  $\pi$ )

# Measurements of genetic variation

---

- Segregating sites ( $S$ )
- Site frequency spectrum (SFS)
- Gene diversity ( $h$  &  $H$ )
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity ( $\Pi$  &  $\pi$ )

# Segregating sites ( $S$ )

---

## Definition

A segregating site is a site that is polymorphic in the data—i.e., there are multiple alleles observed.

## Segregating sites ( $S$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

## Segregating sites ( $S$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

## Segregating sites ( $S$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \rightarrow S = 4$$



# Measurements of genetic variation

---

- Segregating sites ( $S$ )
- Site frequency spectrum (SFS)
- Gene diversity ( $h$  &  $H$ )
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity ( $\Pi$  &  $\pi$ )

# Site frequency spectrum (SFS)

---

## Definition

Minor allele frequency spectrum: Histogram of the frequency of the less common allele which range from  $1/n$  to 0.5 where  $n$  is the total number of chromosomes.

# Site frequency spectrum (SFS)

---

## Definition

Minor allele frequency spectrum: Histogram of the frequency of the less common allele which range from  $1/n$  to 0.5 where  $n$  is the total number of chromosomes.

Derived allele frequency spectrum: Histogram of the frequency of the derived allele—normally determined by the use of an outgroup—which range from  $1/n$  to  $(n-1)/n$ .

# Site frequency spectrum (SFS)

---

## Definition

Minor allele frequency spectrum: Histogram of the frequency of the less common allele which range from  $1/n$  to 0.5 where  $n$  is the total number of chromosomes.

Derived allele frequency spectrum: Histogram of the frequency of the derived allele—normally determined by the use of an outgroup—which range from  $1/n$  to  $(n-1)/n$ .

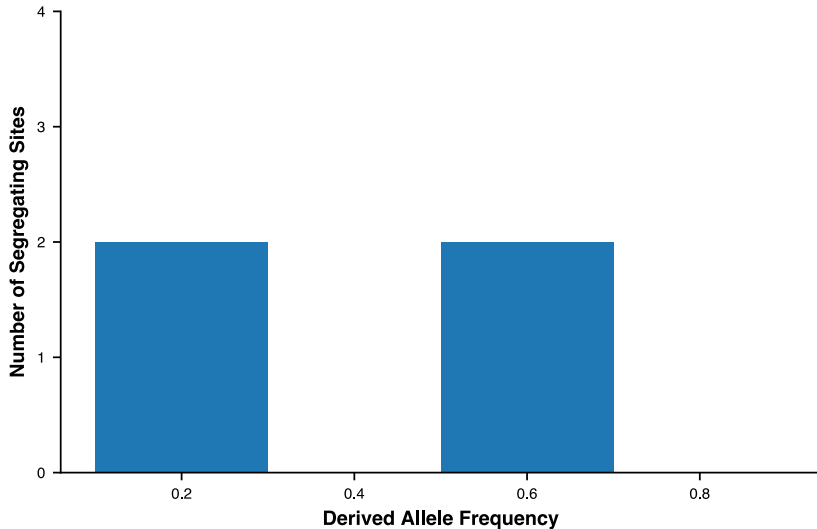
## Note

Minor allele frequency spectrum = Folded SFS

Derived allele frequency spectrum = Unfolded SFS

# Site frequency spectrum (SFS)

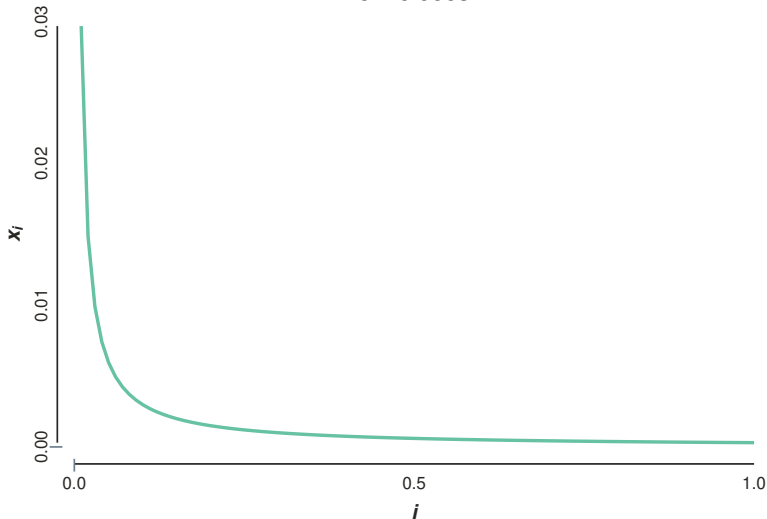
---



# Site frequency spectrum has the shape $x_i = \frac{\theta}{i}$

---

$\theta = 0.0003$



# Measurements of genetic variation

---

- Segregating sites ( $S$ )
- Site frequency spectrum (SFS)
- Gene diversity ( $h$  &  $H$ )
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity ( $\Pi$  &  $\pi$ )

# Gene diversity ( $h$ & $H$ )

---

## Definition

Gene diversity is the probability that two random DNA sequences are different.



# Gene diversity ( $h$ & $H$ )

---

## Definition

Gene diversity is the probability that two random DNA sequences are different.

## Equation

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

Where  $p_i$  is the frequency of the  $i^{th}$  allele out of  $m$  observed alleles.

$$H = \frac{1}{L} \sum_{j=1}^L h_j \quad (2)$$

Where  $h_j$  is the gene diversity for site  $j$  and  $L$  is to the total number of sites.

# Dave's tips and tricks

---

## Note

$$h = 1 - (p^2 + q^2)$$

Where  $p$  is the frequency of the derived/alternate allele and  $q = (1 - p)$

## Gene diversity ( $h$ & $H$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

## Gene diversity ( $h$ & $H$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (0^2 + 1^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \end{bmatrix}$$

## Gene diversity ( $h$ & $H$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (0^2 + 1^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \end{bmatrix} \rightarrow h_j = \begin{bmatrix} 0 \\ 12/25 \\ 12/25 \\ 8/25 \\ 8/25 \end{bmatrix}$$

## Gene diversity ( $h$ & $H$ )

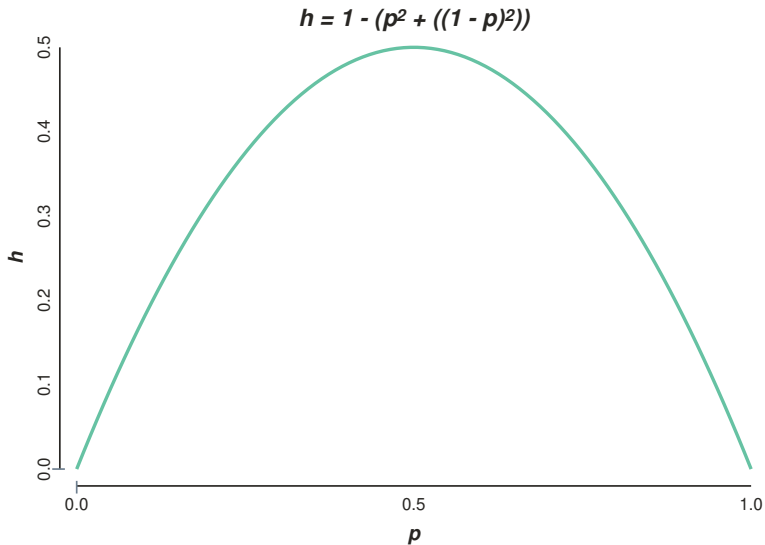
---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - (0^2 + 1^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (3^2/5^2 + 2^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \\ 1 - (1^2/5^2 + 4^2/5^2) \end{bmatrix} \rightarrow h_j = \begin{bmatrix} 0 \\ 12/25 \\ 12/25 \\ 8/25 \\ 8/25 \end{bmatrix}$$

$$H = 40/25 \times 1/5 = 8/25$$

# Behavior of $h$

---



# Measurements of genetic variation

---

- Segregating sites ( $S$ )
- Site frequency spectrum (SFS)
- Gene diversity ( $h$  &  $H$ )
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity ( $\Pi$  &  $\pi$ )



# Nucleotide diversity ( $\Pi$ & $\pi$ )

---

## Definition

Nucleotide diversity is the average number of pairwise differences between genotypes drawn from the same population.

# Nucleotide diversity ( $\Pi$ & $\pi$ )

---

## Equation

$$\Pi = \sum_{i < j} k_{ij} \quad (3)$$

Where  $k_{ij}$  is the number of nucleotide differences between the  $i^{th}$  and  $j^{th}$  sample at a given site.

$$\pi = \frac{1}{\binom{n}{2}} \sum_{j=1}^L \Pi_j \quad (4)$$

Where  $\Pi_j$  is the nucleotide diversity for site  $j$ ,  $L$  is the total number of sites in the genotype matrix, and  $\binom{n}{2}$  is number of unique comparisons between  $n$  samples in the genotype matrix.

## Dave's tips and tricks

---

Note

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

## Nucleotide diversity ( $\Pi$ & $\pi$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

## Nucleotide diversity ( $\Pi$ & $\pi$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 + 0 + 0 + 0 \\ 3 + 1 + 1 + 1 \\ 3 + 1 + 2 + 0 \\ 1 + 1 + 1 + 1 \\ 4 + 0 + 0 + 0 \end{bmatrix}$$

## Nucleotide diversity ( $\Pi$ & $\pi$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 + 0 + 0 + 0 \\ 3 + 1 + 1 + 1 \\ 3 + 1 + 2 + 0 \\ 1 + 1 + 1 + 1 \\ 4 + 0 + 0 + 0 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 0 \\ 6 \\ 6 \\ 4 \\ 4 \end{bmatrix}$$

## Nucleotide diversity ( $\Pi$ & $\pi$ )

---

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 + 0 + 0 + 0 \\ 3 + 1 + 1 + 1 \\ 3 + 1 + 2 + 0 \\ 1 + 1 + 1 + 1 \\ 4 + 0 + 0 + 0 \end{bmatrix} \rightarrow \Pi_j = \begin{bmatrix} 0 \\ 6 \\ 6 \\ 4 \\ 4 \end{bmatrix} \rightarrow \Pi = 20$$

$$\pi = \binom{5}{2}^{-1} \sum_{j=1}^5 \Pi_j = \frac{2}{5(5-1)} \times 20 = 2$$

# Overview

---

1. ATGC's of life & encoding DNA
2. Measures of sequence diversity
- 3. In class coding exercise**