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**

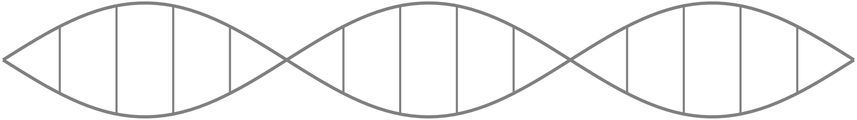
Overview

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

DNA consists of four nucleotides

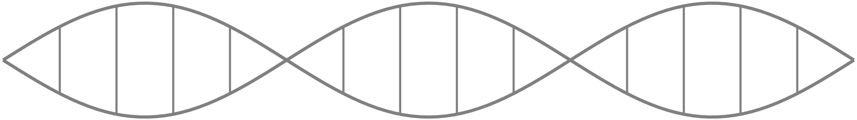
ATGC

DNA is organized onto chromosomes



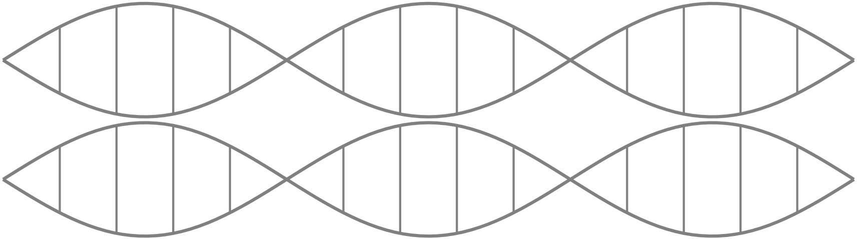
Ploidy (#N): number of sets of chromosomes

N = haploid



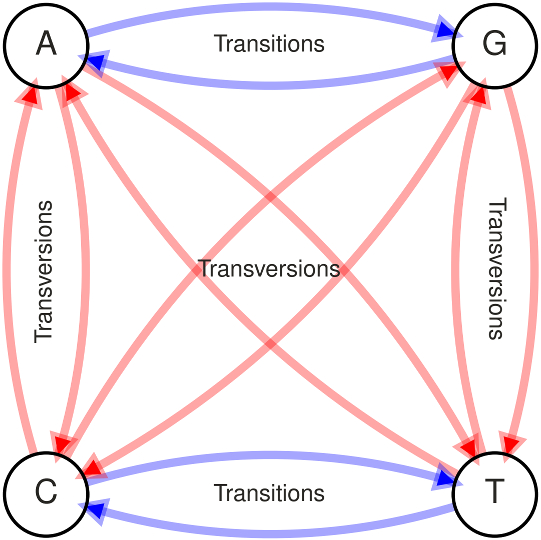
Ploidy (#N): number of sets of chromosomes

2N = diploid



Q: How does genetic variation arise?

A: Mutations


A
G
C
T
Transitions
Transversions
Transversions
Transversions
Transitions

How do we encode DNA?

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

How do we encode DNA?

$$\begin{array}{c c}
 & \begin{array}{ccccc} ind_1 & ind_2 & ind_3 & ind_4 & ind_5 \end{array} \\
\begin{array}{c} pos_1 \\ pos_2 \\ pos_3 \\ pos_4 \\ pos_5 \end{array} &
\begin{bmatrix}
T & T & T & T & T \\
C & G & G & C & G \\
A & T & A & T & T \\
G & G & G & G & C \\
T & A & A & A & A
\end{bmatrix}
\end{array}$$

Genotype matrices

$$\begin{array}{cc} & \begin{array}{ccccc} ind_1 & ind_2 & ind_3 & ind_4 & ind_5 \end{array} \\ \begin{array}{c} pos_1 \\ pos_2 \\ pos_3 \\ pos_4 \\ pos_5 \end{array} & \begin{bmatrix} T & T & T & T & T \\ C & G & G & C & G \\ A & T & A & T & T \\ G & G & G & G & C \\ T & A & A & A & A \end{bmatrix} \end{array} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$0 =$ 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

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

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 (Π & π)

Measurements of genetic variation

- Segregating sites (S)
- Site frequency spectrum (SFS)
- Gene diversity (h & H)
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity (Π & π)

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 (Π & π)

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.

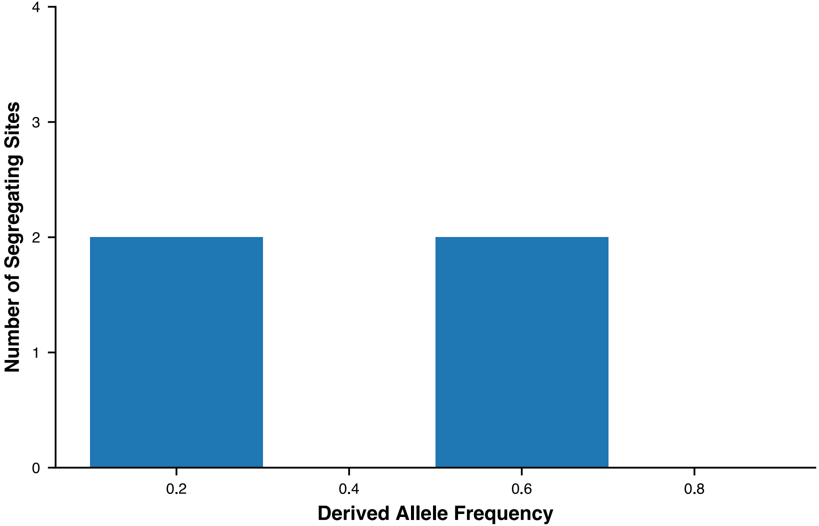
Derived allele frequency spectrum: Histogram of the frequency of the derived allele—normally determined by the use of an outgroup—which range from $^1/_n$ to $^{(n-1)}/_n$.

Note

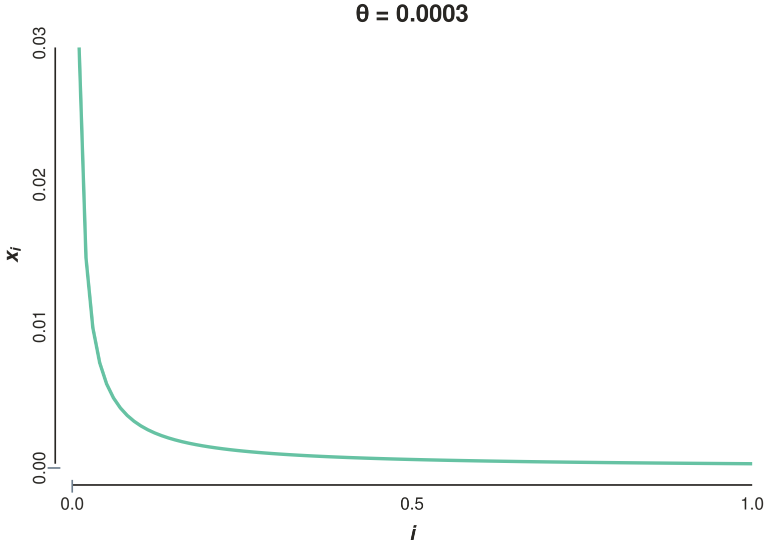
Minor allele frequency spectrum = Folded SFS

Derived allele frequency spectrum = Unfolded SFS

Site frequency spectrum (SFS)



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



Measurements of genetic variation

- Segregating sites (S)
- Site frequency spectrum (SFS)
- Gene diversity (h & H)
  - Also referred to as *expected heterozygosity*
- Nucleotide diversity (Π & π)

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

$$h = 1 - \sum_{i=1}^{m} p_i^2 \tag{1}$$

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

$$H = \frac{1}{L} \sum_{j=1}^{L} h_j \tag{2}$$

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)

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

Gene diversity $(h \;\&\; H)$

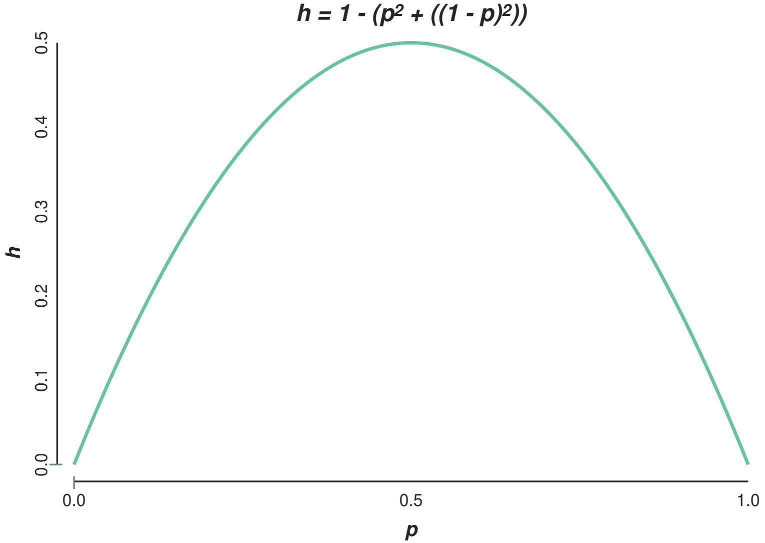
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 - \left(1^2 + 0^2\right) \\ 1 - \left({}^{2^2}\!/_{5^2} + {}^{3^2}\!/_{5^2}\right) \\ 1 - \left({}^{2^2}\!/_{5^2} + {}^{3^2}\!/_{5^2}\right)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

Gene diversity $(h \,\&\, H)$

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

$$H = \nicefrac{40}{25} \times \nicefrac{1}{5} = \nicefrac{8}{25}$$

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 (Π & π)

Nucleotide diversity (Π & π)

Definition

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

Nucleotide diversity (Π & π)

Definition

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

Equation

$$\Pi = \frac{\sum_{i<j} k_{ij}}{\binom{n}{2}} \tag{3}$$

Where k_{ij} is the number of nucleotide differences between the i^{th} and j^{th} sequence in the sample and the denominator represents the number of unique comparisons being made between n sequences.

$$\pi = \frac{\Pi}{L} \tag{4}$$

Where L is to the total number of sites.

Dave's tips and tricks

Note

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

Nucleotide diversity (Π & π)

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

Nucleotide diversity (Π & π)

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

Nucleotide diversity (Π & π)

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{array}{l} (0+0+0+0) = 0 \\ (3+1+1+1) = 6 \\ (3+1+2+0) = 6 \\ (1+1+1+1) = 4 \\ (4+0+0+0) = 4 \end{array} \rightarrow \Pi = 20 \div \frac{5(5-1)}{2} = 2$$

Nucleotide diversity (Π & π)

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{matrix} (0+0+0+0) = 0 \\ (3+1+1+1) = 6 \\ (3+1+2+0) = 6 \\ (1+1+1+1) = 4 \\ (4+0+0+0) = 4 \end{matrix} \rightarrow \Pi = 20 \div \frac{5(5-1)}{2} = 2$$

$$\pi = 2 \times {}^{1}\!/_{5} = {}^{2}\!/_{5}$$

Overview

1. ATGC's of life & encoding DNA

2. Measures of sequence diversity

3. **In class coding exercise**