Population Structure

BIOL 1435

February 23, 2023

Overview

1. Quiz Review
2. Population Differentiation & Structure
3. Relative Differentiation
4. Absolute Differentiation
5. Population Structure

Overview

1. **Quiz Review**
2. Population Differentiation & Structure
3. Relative Differentiation
4. Absolute Differentiation
5. Population Structure

Overview

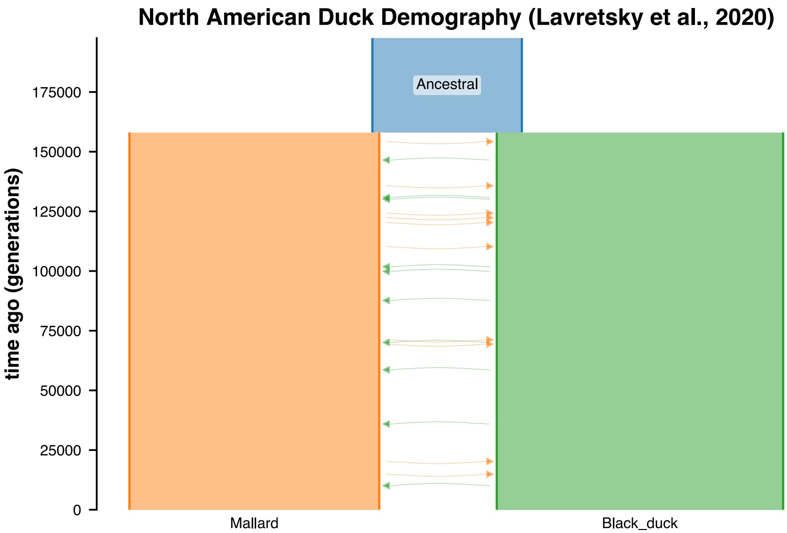
1. Quiz Review
2. **Population Differentiation & Structure**
3. Relative Differentiation
4. Absolute Differentiation
5. Population Structure

Q: What is a population?

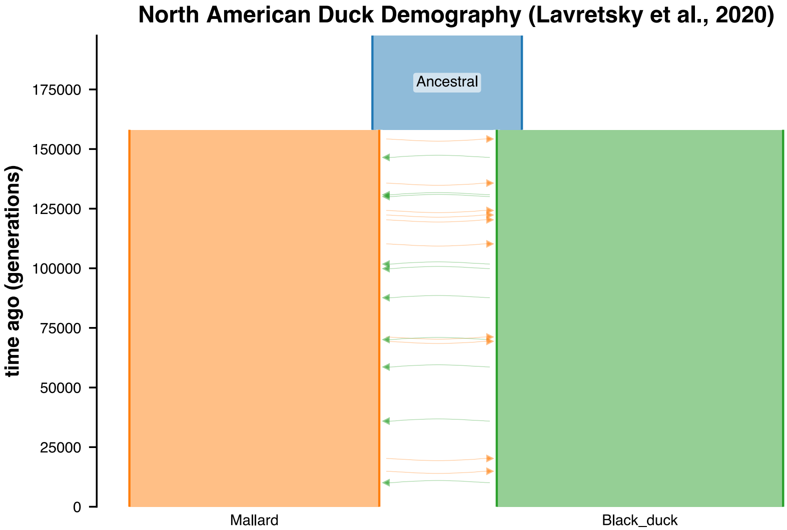
Q: What is a population?

A: A group of freely interbreeding individuals.

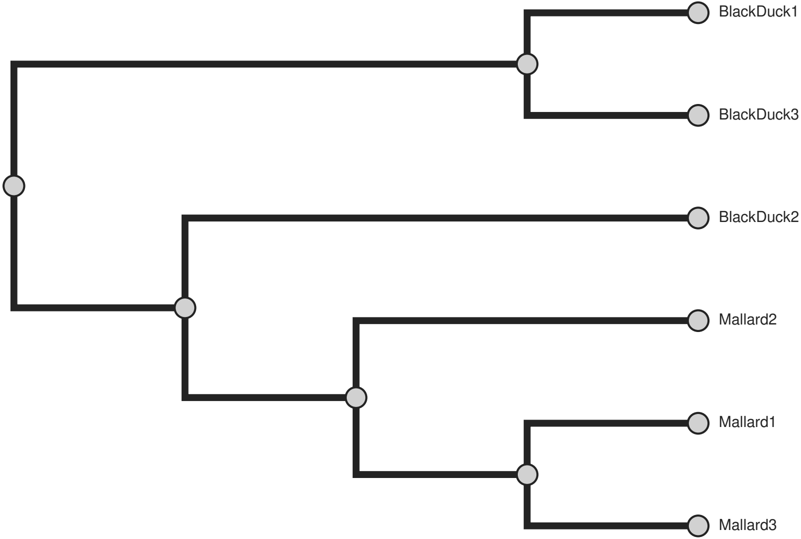
Populations



Differentiated Populations



Population Structure



Overview

1. Quiz Review
2. Population Differentiation & Structure
3. **Relative Differentiation**
4. Absolute Differentiation
5. Population Structure

F_{ST}

F_{ST}

Definition

F_{ST} is the proportion of the genetic variance contained in population relative to the total genetic variance in the entire population.

F_{ST}

Definition

F_{ST} is the proportion of the genetic variance contained in population relative to the total genetic variance in the entire population.

OR

F_{ST}

Definition

F_{ST} is the proportion of the genetic variance contained in population relative to the total genetic variance in the entire population.

OR

F_{ST} quantifies genetic drift between two populations relative to the average drift between the two populations.

Hudon's Estimator of F_{ST}

Hudon's Estimator of F_{ST}

Equation

$$F_{ST} = \frac{N}{D} \tag{1}$$

$$N = \left(p_1 - p_2\right)^2 \tag{2}$$

$$D = p_1\left(1 - p_2\right) + p_2\left(1 - p_1\right) \tag{3}$$

Where p_i is the frequency of the derived/alternative allele at a given site from the i^{th} population.

Hudon's Estimator of F_{ST}

Equation

$$F_{ST} = \frac{N}{D} \tag{1}$$

$$N = (p_1 - p_2)^2 \tag{2}$$

$$D = p_1 (1 - p_2) + p_2 (1 - p_1) \tag{3}$$

Where p_i is the frequency of the derived/alternative allele at a given site from the i^{th} population.

$$F_{ST} = \frac{\sum_{j=1}^{L} N_j}{\sum_{j=1}^{L} D_j} \tag{4}$$

Where L is to the total number of sites.

Understanding F_{ST}

Understanding F$_{ST}$

Interpretation

$$F_{ST} = 0 : \text{ no differentiation}$$

$$F_{ST} = 1 : \text{ maximum differentiation}$$

Limitations of F_{ST}

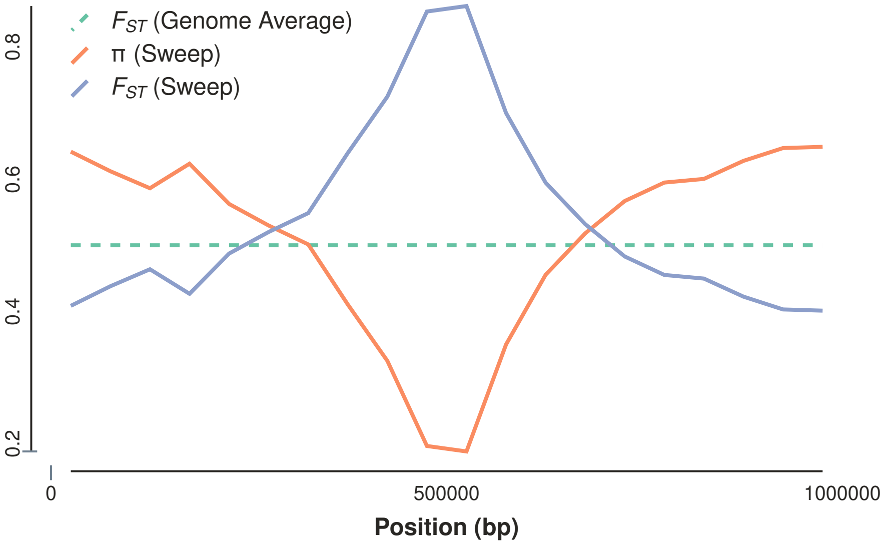
Q: Why is F_{ST} a relative measure of population differentiation?

Limitations of F_{ST}

Q: Why is F_{ST} a relative measure of population differentiation?

A: F_{ST} is strongly influenced by within-subpopulation levels of diversity!

Linked-Selection & F_{ST}



Overview

1. Quiz Review
2. Population Differentiation & Structure
3. Relative Differentiation
4. **Absolute Differentiation**
5. Population Structure

d_{XY}

$\mathbf{d_{XY}}$

Definition

d_{XY} average number of pairwise differences between chromosomes from two populations.

$\mathbf{d_{XY}}$

Definition

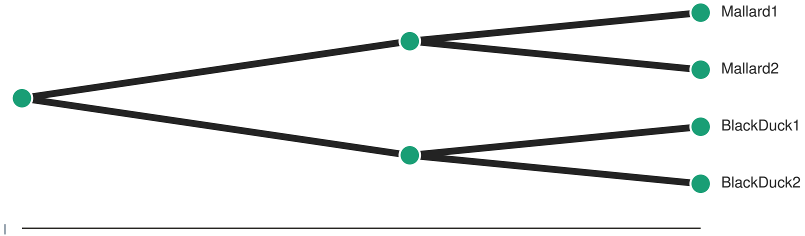
d_{XY} average number of pairwise differences between chromosomes from two populations.

Equation

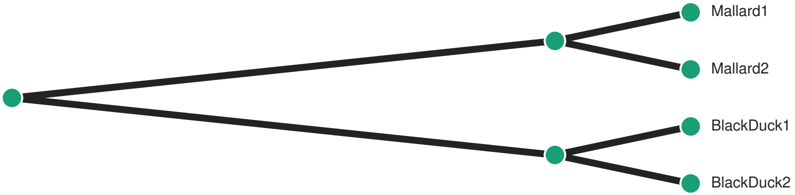
$$d_{XY} = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_y} k_{ij} \tag{5}$$

Where n_X and n_Y correspond to the number of chromosomes in populations X and Y, , and k_{ij} is the number of nucleotide differences between the i^{th} and j^{th} chromosome.

Advantages of d_{XY}



d_{XY}



Overview

1. Quiz Review
2. Population Differentiation & Structure
3. Relative Differentiation
4. Absolute Differentiation
5. **Population Structure**

Identifying Population Structure

Q: Given a set of samples how can we assess if there is population structure?

Identifying Population Structure

Q: Given a set of samples how can we assess if there is population structure?

A: Principle Component Analysis (PCA)

PCA Steps

1. Zero-center the allele count matrix
2. Calculate the covariance matrix
3. Perform eigendecomposition

The Allele Count Matrix (C)

$$\mathbf{C}\left(i,j\right)=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} \\ c_{21} & c_{22} & \cdots & c_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} \end{bmatrix}$$

Where $\mathbf{C}\left(i,j\right)$ consists of m (individuals) $\times$ n (sites), and $\mathbf{C}\left(i,j\right) \in \{\ 0,1,2\}$ alternative/derived allele count.

The Standardized Allele Count Matrix (M)

$$\mu(j) = \frac{\sum_{i=1}^{m} \mathbf{C}(i,j)}{m}$$

$$\mathbf{M}\,(i,j) = \frac{\mathbf{C}(i,j)-\mu(j)}{\sqrt{(m\times 2)p(j)(1-p(j))}}$$

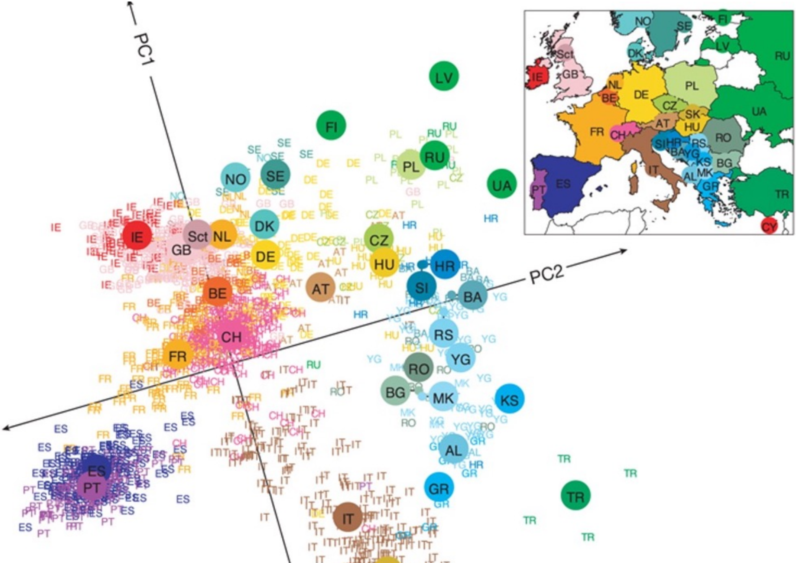
Where $\mu(j)$ is the column mean, $p(j) = \mu(j) \div 2$ is the allele frequency for site j, and $\sqrt{(m \times 2)\, p(j)\, (1 - p(j))}$ is the standard deviation of the binomial distribution and is proportional to the rate change in allele frequency per generation due to genetic drift.

Covariance Matrix (X)

$$\mathbf{X} = \frac{1}{m-1}\mathbf{M}\mathbf{M}^{\mathbf{T}}$$

Where X is a square matrix consisting of the the covariance between all individuals.

PCA In Action


PC1
PC2
LV
FI
RU
PL
UA
NO
SE
DK
IE
GB
Sct
NL
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
GR
TR
ES
PT
IT
CY
SK