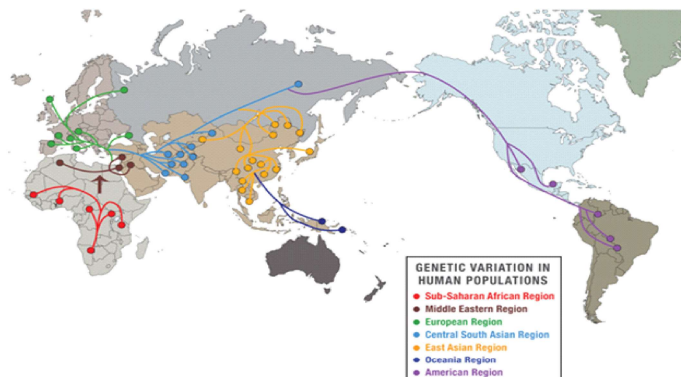


14 Population Substructure, Stratification, and Admixture

14.1: Learning Objectives

By the end of this chapter, the reader should be able to define population stratification, admixture, and cryptic relatedness as distinct forms of population substructure, explain the Wahlund effect and its consequence for Hardy–Weinberg equilibrium, describe how population stratification confounds population-based association tests, and detect and correct for stratification using genomic control and principal component analysis in R and Python.

14.2: What is Population Substructure?

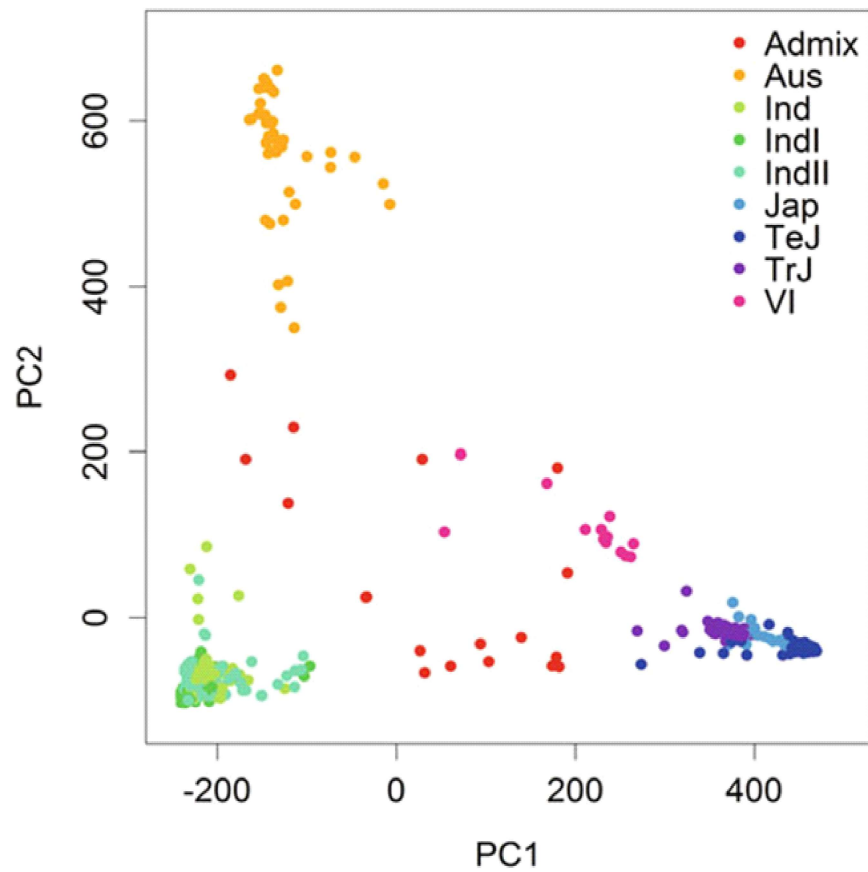


Global human genetic diversity and migration patterns out of Africa, the deep demographic history underlying present-day population substructure.

Population substructure refers to any feature of a sampled population that causes allele and/or genotype frequencies to vary systematically across individuals in ways that do not reflect random Mendelian sampling alone. Three related but distinct phenomena are usually distinguished:

- **Population stratification:** the sample is drawn from two or more genetically distinct subgroups (e.g., distinct ethnic or ancestral groups) mixed together, without the individuals themselves being recent-generation admixed individuals — for example, a Nigerian national sample combining Yoruba, Igbo, and Hausa/Fulani participants, each with somewhat different allele frequencies at many loci due to their distinct population histories.
- **Population admixture:** individuals themselves carry ancestry from two or more historically distinct source populations, arising from inter-population mating over recent generations (often occult, i.e. not evident from self-reported ethnicity) — a feature of many African-American, Afro-Brazilian, and other diaspora populations, and increasingly relevant within Nigeria itself given inter-ethnic marriage.
- **Cryptic relatedness:** some individuals in the sample are more closely related to one another (as in Chapter 6's kinship framework) than would be expected under simple random sampling from the population, without this relatedness being known or recorded by the investigators.

All three forms of substructure can distort standard statistical tests that assume individuals are independent draws from a single, homogeneous, randomly mating population.



Principal component analysis of genome-wide genotype data from multiple real human populations (e.g., admixed, Australian, several Indian and Japanese subgroups), showing that each population forms a visually distinct cluster in PC1–PC2 space — the standard diagnostic display of population substructure in a genetic study.

14.3: The Wahlund Effect: Substructure and Hardy–Weinberg Equilibrium

A foundational result, due to Sten Wahlund (1928), shows that combining two or more subpopulations that are each individually at Hardy–Weinberg equilibrium (Chapter 3), but which differ in allele frequency, produces a **combined sample that departs from Hardy–Weinberg equilibrium** — specifically, showing a deficit of heterozygotes relative to the HWE expectation computed from the pooled allele frequency. Suppose the population consists of two subpopulations in proportions w_1 and $w_2 = 1 - w_1$, with allele frequencies p_1 and p_2 ($p_1 \neq p_2$) at a bi-allelic locus, each subpopulation internally at HWE. The pooled allele frequency is $\bar{p} = w_1 p_1 + w_2 p_2$, but the pooled heterozygote frequency is:

$$\begin{aligned} Aa(\text{pooled}) &= w_1 \cdot 2p_1(1 - p_1) + w_2 \cdot 2p_2(1 - p_2) \\ &= 2\bar{p}(1 - \bar{p}) - 2w_1 w_2 (p_1 - p_2)^2 \end{aligned}$$

The term subtracted, $2w_1 w_2 (p_1 - p_2)^2$, is always non-negative, and is exactly zero only when $p_1 = p_2$ (no true subdivision) — so the pooled sample always shows *at least as many* homozygotes, and correspondingly fewer heterozygotes, as Hardy–Weinberg equilibrium computed at the pooled allele frequency would predict. This **Wahlund effect** is the population-genetic signature of undetected stratification, and is one of the classical diagnostic clues (alongside an excess of significant HWE tests genome-wide) that a sample thought to be drawn from a single homogeneous population is, in fact, structured.

Theorem 14.1: (The Wahlund Effect)

If a population is a mixture of K subpopulations in proportions $w_1, \dots, w_K$ ($\sum_k w_k = 1$), each internally at Hardy–Weinberg equilibrium at a bi-allelic locus with allele frequency p_k , then the pooled heterozygosity satisfies $H_{pooled} \leq 2\bar{p}(1 - \bar{p})$, where $\bar{p} = \sum_k w_k p_k$, with equality if and only if all p_k are equal.

Proof (two-subpopulation case; the general case follows by the same argument applied to the variance of p_k across subpopulations). The pooled heterozygote frequency is the weighted average of each subpopulation's own HWE heterozygosity: $H_{pooled} = \sum_k w_k \cdot 2p_k(1 - p_k)$. Write $2\bar{p}(1 - \bar{p}) - H_{pooled}$ and expand, using $\bar{p} = \sum_k w_k p_k$:

$$\begin{aligned} & 2\bar{p}(1 - \bar{p}) - \sum_k w_k \cdot 2p_k(1 - p_k) \\ &= 2 \left[\bar{p} - \bar{p}^2 - \sum_k w_k p_k + \sum_k w_k p_k^2 \right] \\ &= 2 \left[\sum_k w_k p_k^2 - \bar{p}^2 \right] = 2 \text{Var}_w(p_k), \end{aligned}$$

where $\text{Var}_w(p_k) = \sum_k w_k p_k^2 - \bar{p}^2 = \sum_k w_k (p_k - \bar{p})^2$ is the (weighted) variance of the allele frequency across subpopulations — itself always ≥ 0 as a sum of squares, with equality if and only if every $p_k = \bar{p}$ (no true differentiation). Hence $2\bar{p}(1 - \bar{p}) - H_{pooled} = 2 \text{Var}_w(p_k) \geq 0$, establishing $H_{pooled} \leq 2\bar{p}(1 - \bar{p})$ as claimed, with the deficit exactly equal to twice the cross-subpopulation allele-frequency variance. ■

This proof reveals something the raw formula alone does not make obvious: the Wahlund deficit is not an arbitrary

correction term but is *exactly* twice the variance of allele frequency across the mixed subpopulations — connecting the population-genetic phenomenon directly to a basic statistical quantity, and showing immediately why more genetically differentiated subpopulations (larger $\text{Var}_w(p_k)$) produce a larger apparent heterozygote deficit when pooled.

Worked Example 14.2 (Wahlund effect across three Nigerian ethno-linguistic groups)

A national biobank pools genotype data from three broad Nigerian ethno-linguistic groups without correcting for group membership: Yoruba ($w_1 = 0.35$, $p_1 = 0.62$ at a candidate marker), Igbo ($w_2 = 0.35$, $p_2 = 0.58$), and Hausa/Fulani ($w_3 = 0.30$, $p_3 = 0.45$). The pooled allele frequency is $\bar{p} = 0.35(0.62) + 0.35(0.58) + 0.30(0.45) = 0.217 + 0.203 + 0.135 = 0.555$. The weighted variance across groups is $\text{Var}_w(p_k) = 0.35(0.62 - 0.555)^2 + 0.35(0.58 - 0.555)^2 + 0.30(0.45 - 0.555)^2 = 0.35(0.00423) + 0.35(0.000625) + 0.30(0.011025) = 0.00148 + 0.000219 + 0.00331 = 0.00501$. By Theorem 14.1, the Wahlund deficit is $2 \times 0.00501 = 0.01002$, so naive pooled HWE would predict heterozygosity $2(0.555)(0.445) = 0.494$, while the true pooled heterozygosity is only $0.494 - 0.010 = 0.484$ — a modest but real deficit that a genome-wide analyst would need to distinguish from true inbreeding or selection before drawing any biological conclusion, exactly the caution raised throughout this chapter for multi-ethnic Nigerian genetic studies.

Worked Example 14.1.

Two Nigerian subpopulations each internally at HWE have allele frequencies $p_1 = 0.85$ and $p_2 = 0.55$ at a marker, mixed in equal proportions ($w_1 = w_2 = 0.5$).

$$\bar{p} = 0.5(0.85) + 0.5(0.55) = 0.70$$

$$\begin{aligned}\text{Expected } Aa \text{ under naive pooled HWE} &= 2(0.70)(0.30) \\ &= 0.420\end{aligned}$$

$$\begin{aligned}\text{Actual pooled } Aa &= 0.420 - 2(0.5)(0.5)(0.85 - 0.55)^2 \\ &= 0.420 - 0.045 = 0.375\end{aligned}$$

The pooled sample shows a heterozygote deficit of 0.045 (about 11% fewer heterozygotes than naive HWE would predict) purely as a mechanical consequence of mixing two differentiated subpopulations, with no biological non-random mating, inbreeding, or selection required at all — illustrating why an apparent departure from HWE in a genetic epidemiology dataset must always be checked against the possibility of hidden population structure before being attributed to genotyping error, selection, or inbreeding (Chapter 5).

14.4: Confounding of Population-Based Association Tests

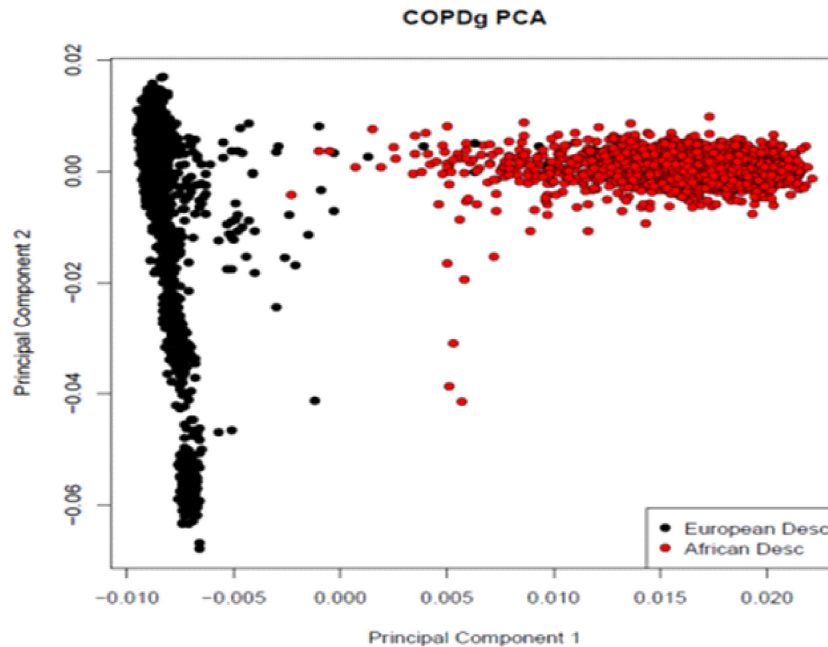
Population stratification becomes a serious statistical problem specifically when case and control samples (Chapter 11) differ, even modestly, in their proportions of underlying subpopulations, *and* the tested marker's allele frequency differs across those subpopulations — a form of **confounding** exactly analogous to confounding in classical epidemiology, with ancestry playing the role of the confounding variable. If a disease happens to be more common in one ancestral group for reasons entirely unrelated to the tested marker (e.g., differing environmental exposures, diet, or unrelated genetic risk factors more prevalent in that group), and that group also happens to carry a different frequency of the tested allele, a population-based case–control test (Chapter 11) will detect a spurious association at that marker — a false positive driven entirely by ancestry confounding rather than by any true biological effect of the marker on disease risk. This is precisely

the vulnerability that motivated the family-based designs of Chapter 13, and it is a serious practical concern for genetic studies conducted across Nigeria's genetically and ethnically diverse population, where case and control recruitment sites (e.g., different teaching hospitals) may inadvertently draw disproportionately from different ethnic-linguistic groups.

14.5: Detecting and Correcting for Stratification

Genomic control: As introduced in Chapter 18 (Section 18.5), genomic control compares the median of the observed test statistics across many genome-wide markers (most of which are not truly associated with the trait) to the median expected under the null distribution, yielding the inflation factor $\lambda_{GC} = \text{median}(\chi^2_{\text{obs}}) / \text{median}(\chi^2_{\text{null}})$. A value of λ_{GC} appreciably greater than 1 signals systematic inflation of test statistics — consistent with, though not uniquely diagnostic of, population stratification — and can be used to deflate all test statistics by dividing by λ_{GC} before assessing significance, a simple if somewhat crude correction.

Principal component analysis (PCA): The modern standard approach, introduced to genetics by Price and colleagues (2006), applies **principal component analysis** to a matrix of genome-wide genotypes to extract the major axes of ancestry variation within the sample — the leading principal components typically correspond closely to known continental or sub-continental ancestry gradients (for example, distinguishing Yoruba, Igbo, and Hausa/Fulani ancestry components within a Nigerian sample, or separating West African from other African or non-African ancestry in an admixed diaspora sample).



Principal component analysis separating two ancestry clusters (e.g. African vs. European descent) from genome-wide genotype data.

Including the leading principal components as covariates in the logistic (or linear) regression association model of Section 11.7 adjusts each marker's association test for the individual's estimated ancestry, removing most of the stratification-induced confounding while retaining the statistical power of the population-based design.

Mixed models: An alternative, increasingly standard in large modern GWAS, replaces PCA-adjusted regression with a **linear mixed model** that includes a random effect with covariance structure given by a genome-wide **genetic relatedness matrix** (a genome-wide analogue of the kinship matrix of Chapter 6),

simultaneously correcting for both distant population stratification and cryptic close relatedness within the sample in a single unified framework.

Theorem 14.2 (The Wahlund Inequality).

Let a population consist of K subpopulations with allele frequencies $p_1, \dots, p_K$ (each subpopulation in Hardy-Weinberg equilibrium internally) mixed in proportions $w_1, \dots, w_K$. Then the expected heterozygosity in the pooled population is less than or equal to the heterozygosity computed from the pooled (average) allele frequency, with equality if and only if all subpopulations have identical allele frequency.

Proof. Let $\bar{p} = \sum_k w_k p_k$ be the pooled allele frequency. The true expected heterozygosity in the pooled population is $H_{\text{pooled}} = \sum_k w_k [2 p_k (1-p_k)]$, since each subpopulation individually satisfies Hardy-Weinberg proportions. The heterozygosity that would be expected if the pooled population were itself in Hardy-Weinberg equilibrium at the pooled allele frequency is $H_{\text{HWE}} = 2 \bar{p} (1 - \bar{p})$. Consider the variance of p_k across subpopulations (weighted by w_k): $\text{Var}(p) = \sum_k w_k (p_k - \bar{p})^2 = \sum_k w_k p_k^2 - \bar{p}^2 \geq 0$. Rearranging, $\sum_k w_k p_k^2 \geq \bar{p}^2 + \text{Var}(p)$. Then $H_{\text{HWE}} - H_{\text{pooled}} = 2\bar{p}(1-\bar{p}) - 2 \sum_k w_k p_k(1-p_k) = 2[\bar{p} - \bar{p}^2 - \sum_k w_k p_k + \sum_k w_k p_k^2] = 2[\sum_k w_k p_k^2 - \bar{p}^2] = 2 \text{Var}(p) \geq 0$, with equality if and only if $\text{Var}(p) = 0$, i.e., all p_k are equal. This proves $H_{\text{pooled}} \leq H_{\text{HWE}}$, the formal statement of the Wahlund effect introduced descriptively in Section 14.3: population substructure always reduces observed heterozygosity relative to Hardy-Weinberg expectation, and strictly so whenever subpopulation allele frequencies genuinely differ.

14.6: Chapter Summary

- Population stratification (distinct subgroups), admixture (mixed ancestry within individuals), and cryptic relatedness are three related forms of population substructure that violate the assumption of a single, homogeneous, randomly mating sample.
- The Wahlund effect shows that pooling HWE subpopulations with different allele frequencies mechanically produces a heterozygote deficit in the combined sample, a classical signature of hidden structure.
- Population stratification confounds population-based case–control association tests whenever case/control ancestry proportions and tested-marker allele frequencies both differ across ancestral subgroups.
- Genomic control, principal component adjustment, and genome-wide mixed models are the standard tools for detecting and correcting stratification in modern association studies, including GWAS (Chapter 18).

14.7: Computational Corner: Simulating the Wahlund Effect and Detecting Structure by PCA

R.

Simulate the Wahlund effect: two subpopulations, unequal allele frequencies

```
set.seed(7)
```

```
n_per_pop <- 1000
```

```
p1 <- 0.85; p2 <- 0.55
```

```
geno_pop1 <- rbinom(n_per_pop, 2, p1)
```

```
geno_pop2 <- rbinom(n_per_pop, 2, p2)
```

```
pooled <- c(geno_pop1, geno_pop2)
```

```
p_bar <- mean(pooled) / 2
```

```

observed_het <-mean(pooled ==1)
expected_het_naive <-2* p_bar * (1- p_bar)
cat("Pooled allele freq:", p_bar, "\n")
cat("Observed heterozygosity:", observed_het, " Naive-HWE
expected:", expected_het_naive, "\n")

```

PCA-based stratification detection: simulate 200 markers
differentiated between two groups

```

n_snps <-200
group <-rep(c(1, 2), each = n_per_pop)
allele_freqs_pop1 <-runif(n_snps, 0.1, 0.9)
allele_freqs_pop2 <-pmin(pmax(allele_freqs_pop1
+rnorm(n_snps, 0, 0.15), 0.01), 0.99)

geno_matrix <-matrix(0, nrow =2* n_per_pop, ncol =
n_snps)
for (j in 1:n_snps) {
  geno_matrix[group ==1, j] <-rbinom(n_per_pop, 2,
allele_freqs_pop1[j])
  geno_matrix[group ==2, j] <-rbinom(n_per_pop, 2,
allele_freqs_pop2[j])
}
pca <-prcomp(geno_matrix, scale. =TRUE)
plot(pca$x[, 1], pca$x[, 2], col = group, pch =20,
xlab = "PC1", ylab = "PC2", main = "PCA reveals two
simulated subpopulations")

```

Python.

```

import numpy as np
from sklearn.decomposition import PCA

```

```

rng = np.random.default_rng(7)
n_per_pop =1000
p1, p2 =0.85, 0.55

```

```

geno_pop1 = rng.binomial(2, p1, n_per_pop)
geno_pop2 = rng.binomial(2, p2, n_per_pop)
pooled = np.concatenate([geno_pop1, geno_pop2])

p_bar = pooled.mean() / 2
observed_het = np.mean(pooled == 1)
expected_het_naive = 2 * p_bar * (1 - p_bar)
print("Pooled allele freq:", p_bar)
print("Observed heterozygosity:", observed_het, "Naive-
HWE expected:", expected_het_naive)

n_snps = 200
group = np.array([1] * n_per_pop + [2] * n_per_pop)
af1 = rng.uniform(0.1, 0.9, n_snps)
af2 = np.clip(af1 + rng.normal(0, 0.15, n_snps), 0.01, 0.99)

geno_matrix = np.zeros((2 * n_per_pop, n_snps))
for j in range(n_snps):
    geno_matrix[group == 1, j] = rng.binomial(2, af1[j],
n_per_pop)
    geno_matrix[group == 2, j] = rng.binomial(2, af2[j],
n_per_pop)

pca = PCA(n_components=2)
scores = pca.fit_transform((geno_matrix -
geno_matrix.mean(0)) / geno_matrix.std(0))
# scores[:, 0] vs scores[:, 1], coloured by 'group', will
separate the two simulated subpopulations.

```

The first simulation reproduces the Wahlund heterozygote deficit of Worked Example 14.1 directly from simulated genotype data rather than algebra, and the second shows the now-standard genome-wide PCA approach: even though no individual was told which subpopulation they belonged to, the first principal component of the genotype matrix cleanly separates the two

simulated groups, exactly the diagnostic plot used in practice (Section 14.5) to check a real GWAS sample for hidden ancestry structure before running any association test.

Mini-Project

Mini-Project 14: Detecting and Correcting for Population Stratification

Simulate a stratified population exhibiting a Wahlund effect and demonstrate both the problem and a standard correction.

Step 1: Simulate genotypes for two subpopulations with different allele frequencies at 200 markers, then pool them into a single sample.

Step 2: Demonstrate the Wahlund effect (Section 14.3) by comparing observed heterozygosity in the pooled sample to Hardy-Weinberg expectations.

Step 3: Run a case-control association test on a null marker (no true effect) using the pooled, unstratified sample and show it produces inflated false positives.

Step 4: Apply principal component analysis (Section 14.5/14.7) to the marker data and show the first principal component recovers the subpopulation structure.

Step 5: Re-run the association test adjusting for the top principal component and show the false-positive inflation is reduced.

Deliverable: *Simulation code, a PCA plot showing recovered structure, and before/after inflation statistics (e.g., genomic control lambda) for the association test.*

14.8: Practice Questions

- Q1. Two Nigerian subpopulations are internally at HWE with allele frequencies $p_1 = 0.70$ and $p_2 = 0.40$, mixed in proportions $w_1 = 0.6$, $w_2 = 0.4$. Compute the pooled allele frequency, the naive-HWE expected heterozygosity, and the actual pooled heterozygosity, and state the size of the Wahlund heterozygote deficit.
- Q2. Explain, in two or three sentences, the difference between population stratification and cryptic relatedness, giving a hypothetical Nigerian example of each.
- Q3. A case-control GWAS finds genomic control inflation factor $\lambda_{GC} = 1.35$. Explain what this suggests, and describe one further analysis (beyond simply dividing test statistics by 1.35) that a careful analyst should perform before concluding the inflation is due to stratification rather than genuine widespread polygenic association.
- Q4. Explain briefly why including the top principal components of genome-wide genotype data as covariates in a logistic regression association model (Section 11.7) can correct for stratification-induced confounding.
- Q5. Using the R or Python code of Section 14.7, increase the allele-frequency difference between the two simulated subpopulations (e.g., $p_1 = 0.9$, $p_2 = 0.3$) and re-run the Wahlund-effect simulation. How does the size of the heterozygote deficit change, and why?
- Q6. Looking at a real multi-population PCA plot such as the one in Section 14.2, in which several distinct populations each form a tight, separate cluster, explain what an *intermediate* individual — one who falls between two clusters rather than inside either — most plausibly represents biologically.

