Computer Simulation #5: F Statistics of Population Structure

We have used the **F** statistic to describe several kinds of population structures, based on a difference between the *observed* and *expected* numbers of heterozygous genotypes within or among populations. When examination of a population for expected Hardy-Weinberg proportions shows a deficiency of heterozygotes, that deficiency can be quantified as $\mathbf{F} = (\mathbf{H_{obs}} - \mathbf{H_{exp}}) / \mathbf{H_{exp}}$. We called this the **Wahlund Effect**, where the analysis of two (or more) populations, each of which follows Hardy-Weinberg expectation, as a *single* population results in a deficiency of heterozygotes, and as such led to faulty rejection of the null hypothesis of *no* population structure.

We also used **F** as the **inbreeding coefficient**, where inbreeding is defined as the expectation of drawing two alleles from a population that are identical by descent, that is, genetic copies of an allele existing in a common ancestor. This expectation applies to both (1) **gamete formation** by two alleles **I by D**, and (2) **random '*draw*'** of two alleles **I by D** from different individuals in a population because of common ancestry. We saw that inbreeding necessarily occurs in any *closed finite* population, because over time all individuals become more or less closely related. **F** can then be used to estimate the effective size (N_e) of different populations, and also to detect mating patterns where mating between more or less closely related individuals is variously avoided, encouraged, or occurs randomly.

F as a measure of breeding structure *within* populations can be extended to **measures of structure *among* sub-populations *within* more inclusive population units**, which may be hypothesized geographic or ecological units. **F** becomes a *hierarchal* statistic, including as components the structure of individuals *within* sub-populations ($\mathbf{F_{IS}}$), and individuals *within* the *total* population ($\mathbf{F_{IT}}$). The most informative structure is often that of *sub-populations* with respect to the *total* population ($\mathbf{F_{ST}}$). $\mathbf{F_{ST}}$ is one of the most widely used tools in evolutionary population genetics.

The **Excel Workbook** for **F-Statistics** provides spreadsheets for calculation of **F statistics** in the special case of a *single* locus with *two* alleles **A** & **a**, among *three* sub-populations, where sample size **N** in all sub-populations is *identical*. This avoids the computational complication of weighting the contributions of sub-populations of different sizes to the various calculations, which obscures the logic of what is really a simple hierarchal procedure.

This lab exercise will demonstrate the calculation of F statistics, and develop a sense for analysis of population structures as inferred F statistics. The data are (1) Hypothetical data for population structures, including hybrid zones between two isolated populations, and (2) Real data for **MN** blood types from an investigation of population structure among Philippine islanders (Arcellana *et al*., 2011; Carr, 2021).

The Philippine archipelago comprises more than 7,600 islands and more than 180 ethnic groups, many of which are to some degree isolated by distance, ethnicity, and (or) language. The data allow tests of (a) *intra*-island differentiation among populations on the islands of Luzon in the north, and on Mindanao in the south, (b) *inter*-island differentiation, including isolation by distance and two alternative samplings of Luzon, Mindanao, the islands of Cebu between them, and Palawan in the west.

References:

AES Arcellana, RMS Guzman, and IKC Fontanilla (2011). Distribution of MN blood group types in local populations in Philippines. *Journal of Genetics* **90**, e90–e93.

SM Carr (2021). Corrigendum to Arcellana et al. 2011. *Journal of Genetics* **100**, 76-77.

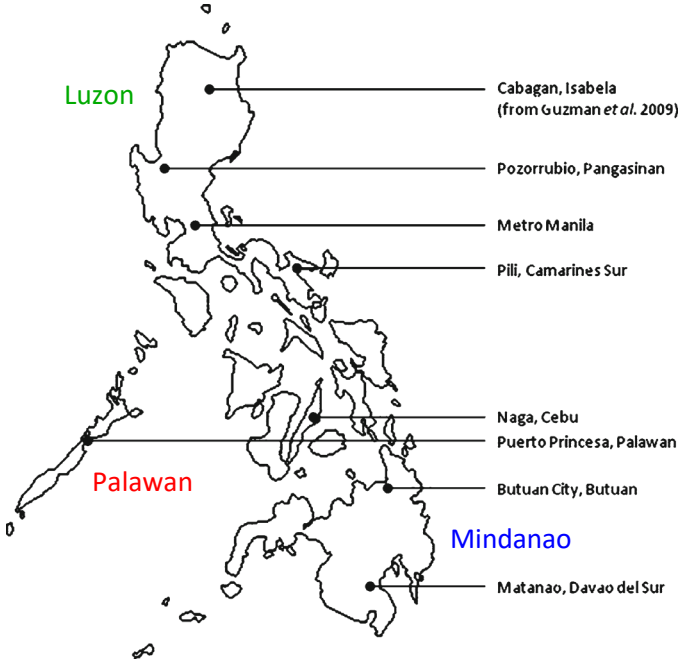
Instructions:

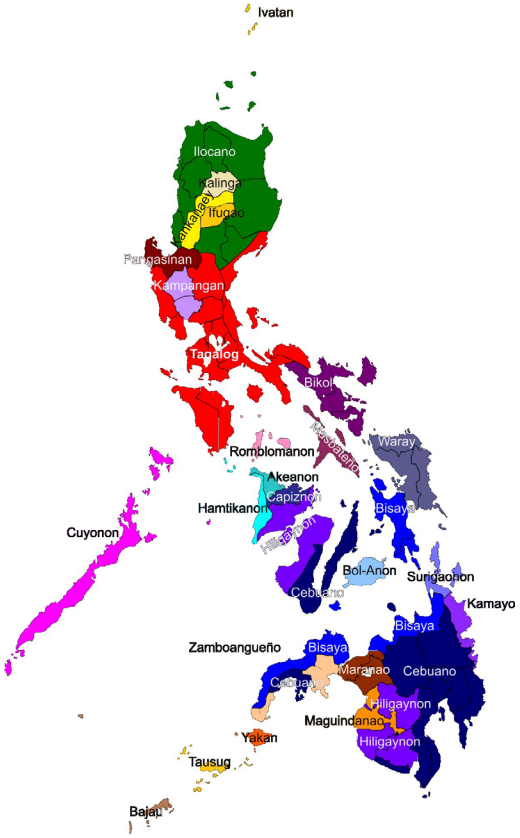
1. The data are **3x3** blocks of **counts** of **AA**, **Aa**, & **aa** for each of **three** sub-populations selected from the tests of hypotheses below. The worksheet will then calculate (1) **Local F** for each of the three sub-populations, and (2) F_{is}, F_{st}, & F_{it} for the combination of sub-populations entered.
    a. **Arrange data blocks on the spreadsheet; Copy & paste *Values Only*** into the Grey Block
    b. For all exercises below, **record** data blocks (1) and (2) for the sets of sub-populations analyzed.

2. Examine the series of simple models of population structure, including
    a. All sub-populations identical for allele frequencies in Hardy-Weinberg proportions (HWP)
    b. All sub-populations identical for allele frequencies ***not*** in HWP
    c. Sub-populations with *different* allele frequencies, in HWP
    d. Sub-populations with *different* allele frequencies, ***not*** in HWP
    e. Any other simple models that occur to you

3. The data below are **MN blood group counts** from populations in the **Philippine Islands** (Arcellana *et al.*, 2011; corrected by Carr, 2021). Counts have been normalized to 100 @.
Examine the population structure of the Philippine archipelago, with the following *a priori* hypotheses:
    a. Structure occurs according to *distance*
    b. Structure occurs *within* individual islands
    c. Structure occurs *among* different islands

    Calculate Chi-Square tests of whether the **F-statistics** show significant departure from HWE.





Group	n	MM	MN	NN
Isabela	100	73	12	15
Pangasinan	100	28	54	18
Manila	100	25	53	22
Camarines Sur	100	20	67	13
Palawan	100	39	46	15
Cebu	100	34	59	7
Butuan	100	74	16	10
Davao del Sur	100	33	45	22