# Using genetic markers in analyses of natural populations

Mark P. Miller
U.S. Geological Survey
Forest and Rangeland Ecosystem Science Center
3200 SW Jefferson Way
Corvallis, OR 97331

# MIGRATE Student Experience and Interest Survey

| Experience | Interest |
|---|---|
| 1=no experience | 1=little interest |
| 2=moderate experience | 2=moderate interest |
| 3=very experienced | 3=very interested |

| | Molecular Markers | |
|---|---|---|
| | **Experience** | **Interest** |
| | 1 | 3 |
| | 1 | 3 |
| | 1 | 3 |
| | 1 | 3 |
| | 1 | 2 |
| | 2 | 3 |
| | 1 | 2 |
| | 1 | 3 |
| | 1 | 1 |
| | 1 | 2 |
| | 1 | 2 |
| | 2 | 3 |
| | 1 | 3 |
| | 1 | 3 |
| | 1 | 3 |
| | 1 | 2 |
| | 1 | 3 |
| **Mean** | **1.11764706** | **2.58823529** |

# Basic terminology (Genetics 101)

- *Genome*: all of the genetic information contained in an organism. In general, each cell in an organism has a complete copy of the entire genome
  - *Genomes* are comprised of *chromosomes*

- *Chromosomes*: large physical aggregations of DNA found within cells.
  - Chromosome numbers vary among different species (ex. humans have 23 pairs of chromosomes)

## Chromosome numbers in some plants

| Plant Species | # |
|---|---|
| Arabidopsis | 10 |
| Rye | 14 |
| Maize | 20 |
| Einkorn wheat | 14 |
| Pollard wheat | 28 |
| Bread wheat | 42 |
| Wild tobacco | 24 |
| Cultivated tobacco | 48 |
| Fern | 1200 |

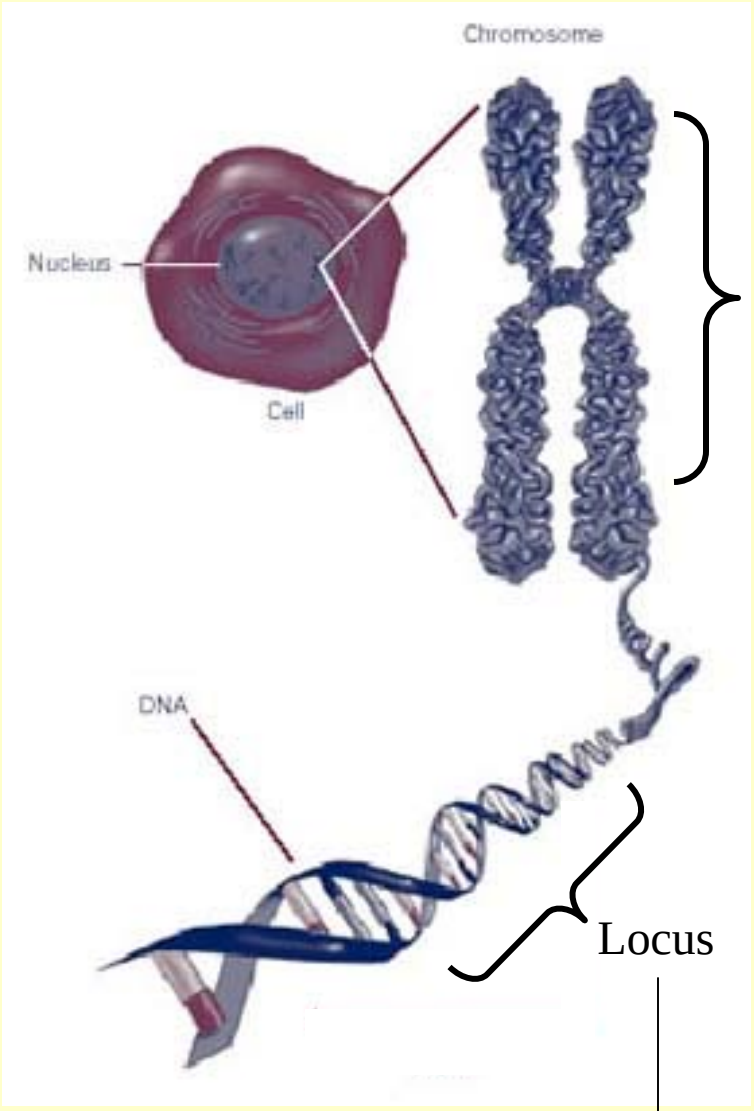
## Chromosome numbers in some animals

| Species | # | Species | # |
|---|---|---|---|
| Fruit fly | 8 | Guinea Pig | 16 |
| Dove | 16 | Snail | 24 |
| Worm | 36 | Fox | 36 |
| Cat | 38 | Pig | 38 |
| Mouse | 40 | Rat | 42 |
| Rabbit | 44 | Hamster | 44 |
| Hare | 46 | Human | 46 |
| Ape | 48 | Sheep | 54 |
| Elephant | 56 | Cow | 60 |
| Donkey | 62 | Horse | 64 |
| Dog | 78 | Chicken | 78 |
| Carp | 104 | Butterflies | 380 |

# Basic terminology (Genetics 101 cont.)

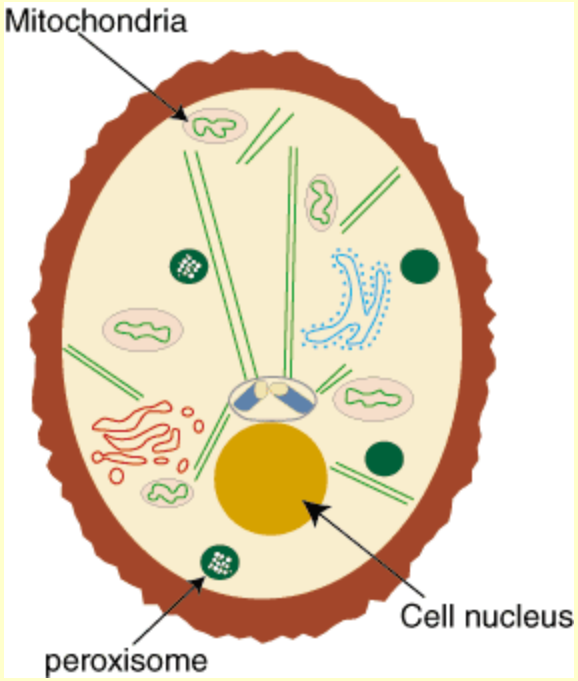
- *Locus*: specific physical location on a chromosome
  - The term *Locus* (or *loci*, plural) is often used to refer to the individual genes that play specific roles for organismal function
- *Alleles*: the specific DNA sequence variants observed at loci
  - Some *loci* are extremely diverse. i.e., there are lots of different *alleles* that may be observed in a sample of individuals from a population
  - The combination of *alleles* observed at a *locus* within a particular individual is called its *genotype*

Chromosomes in cell




Variants at a locus = different Alleles

# Animals actually have TWO different genomes:




**Nuclear Genome:**

1) **Contains the very large majority of genes**
2) **Variable numbers of chromosomes depending on species**
3) **In most animals, organisms are DIPLOID, meaning that there are two copies of each chromosome (and the possibility for two different alleles at each locus)**

**Mitochondrial Genome:**

1) **Contains only a very small number of genes**
2) **Contains a single chromosome**
3) **The mitochondrial genome is HAPLOID (only one chromosome copy and one allele at each locus)**
4) **Mitochondrial alleles are also referred to as *haplotypes***
5) **Almost always maternally inherited**

*****If you are dealing with plants, then you can also examine the chloroplast genome**

# Some additional terms used when dealing with diploid organisms

- Genotype:  A general reference to the combination of alleles at a locus for an individual
- Homozygote:  An individual that has 2 identical copies of an allele at a locus.  The individual is *homozygous* at that locus.
- Heterozygote:  An individual that has 2 different copies of an allele at a locus.  The individual is *heterozygous* at that locus

# Ultimately, we need to:

Characterize variation among

alleles

at multiple different

loci

across multiple different

chromosomes

so that we can make inferences about

genomes

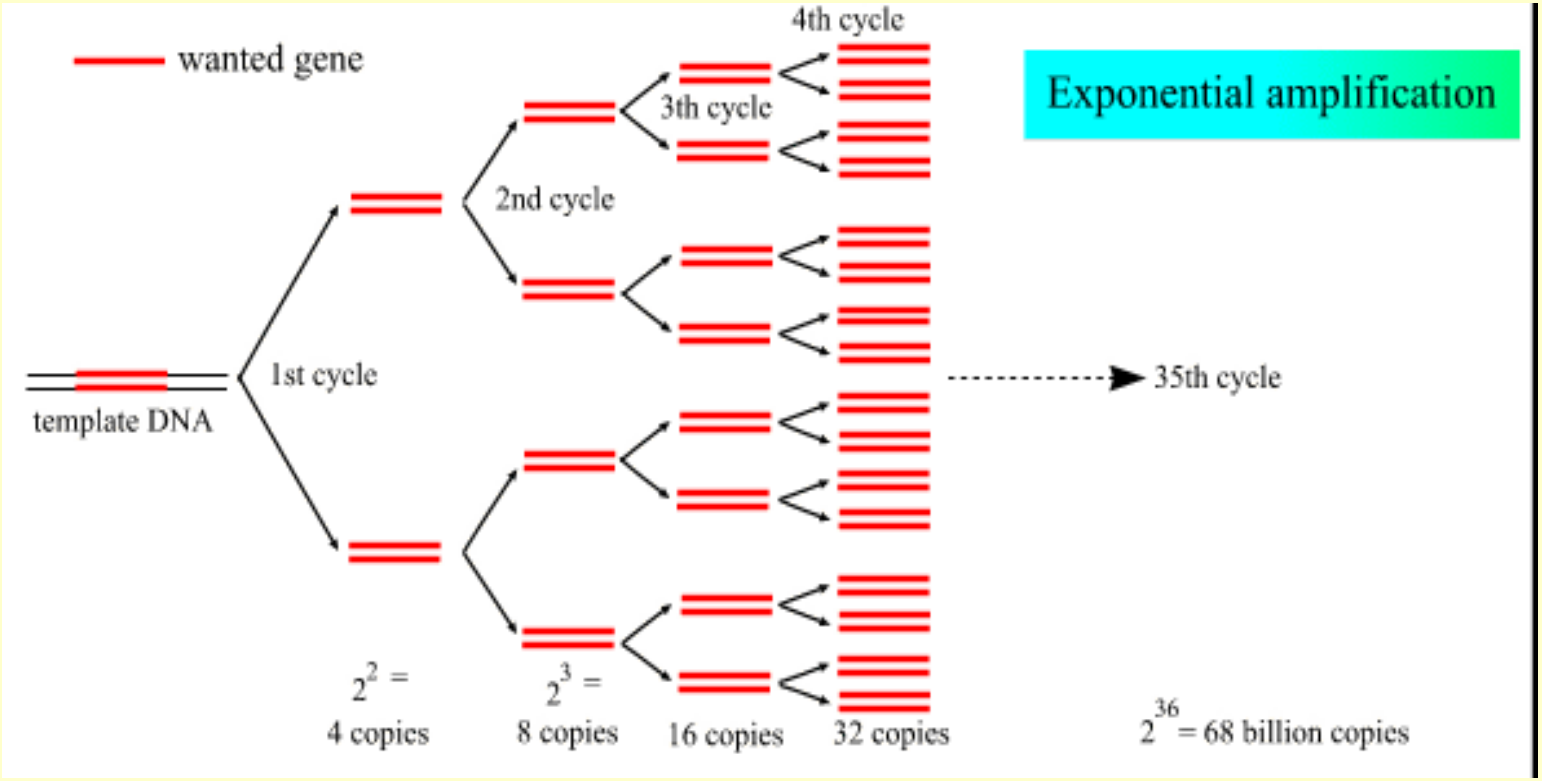
# Characterizing allelic variation at a locus

**Some common techniques for generating genetic data:**

- DNA sequences
- Microsatellites
- Amplified Fragment Length Polymorphisms (AFLP markers)

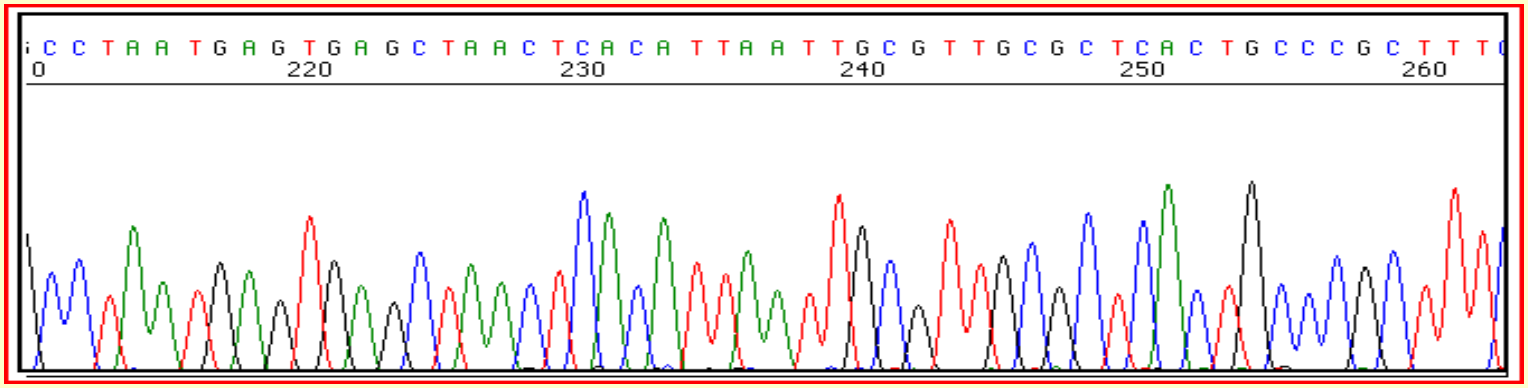
There are many more types too!

# All of these techniques rely on the Polymerase Chain Reaction (PCR)




PCR results in the exponential amplification of a region of target DNA (i.e., a locus)

# DNA sequences

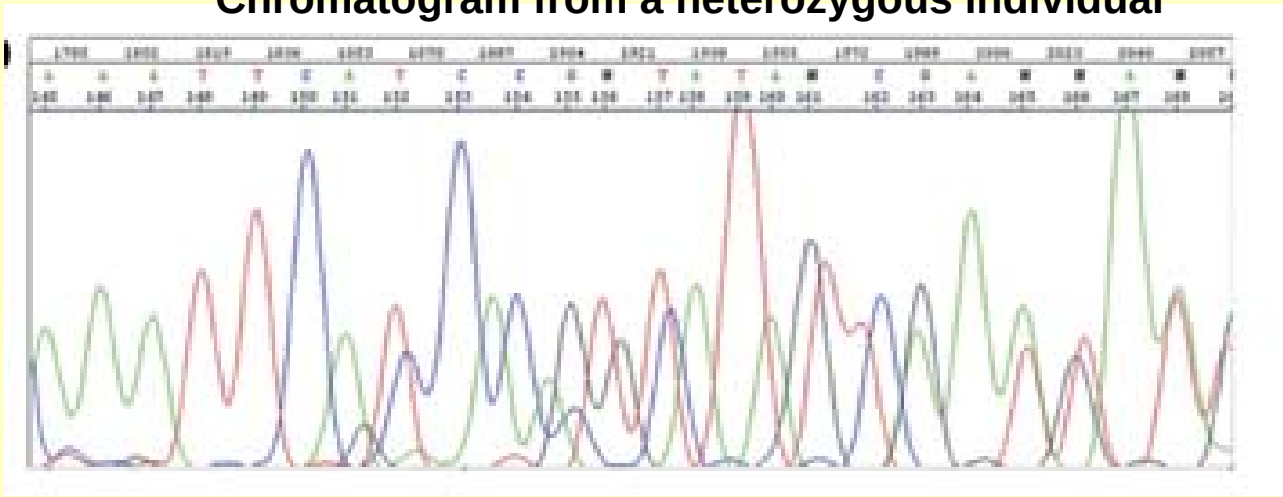



- The ‘ultimate’ approach for characterizing genetic variation
- Most usually, DNA sequences are obtained for organelle genomes (haploid genomes)
- Data come in the form of *chromatograms* (see above), which provide information on the order of nucleotides (A,C,G,T) in a given allele.

# DNA sequences

- More labor intensive and expensive to generate relative to other types of genetic markers
- Additional expensive and time consuming steps are required when generating sequences for diploid genomes
  - If there are 2 alleles at a locus (if the individual is heterozygous), additional steps are required to separate the 2 alleles from the PCR mixture prior to sequencing

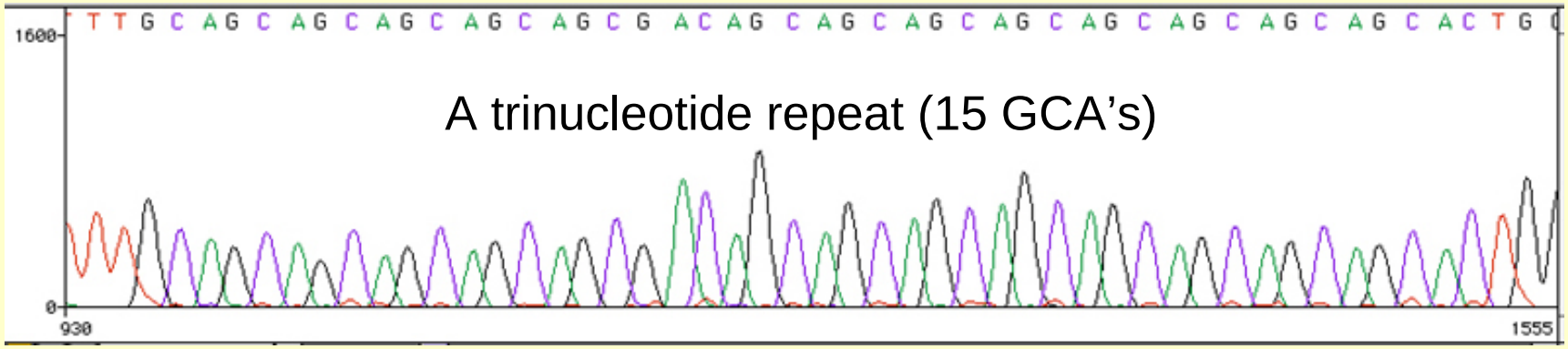
**Chromatogram from a heterozygous individual**



- Used for phylogeographic analysis, discerning population structure, phylogenetics

# Microsatellites

- Tandomly-repeated stretches of DNA
- May be dinucleotide, trinucleotide, tetranucleotide, etc.

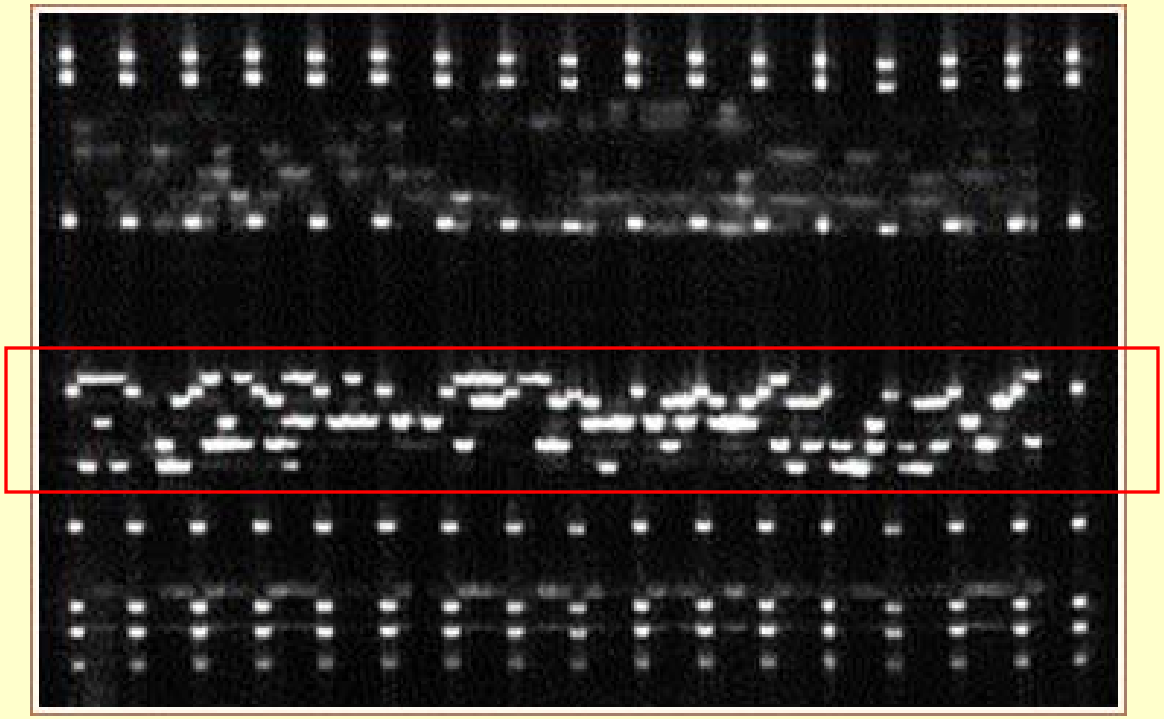



A trinucleotide repeat (15 GCA's)

- Mutations tend to cause the gain or loss of repeat units, which changes the length (size) of fragments
- Microsatellite loci have high mutation rates and are highly variable

# Microsatellites

- Easy to distinguish between homozygous and heterozygous genotypes




- Requires time and money to develop these loci before they can be used

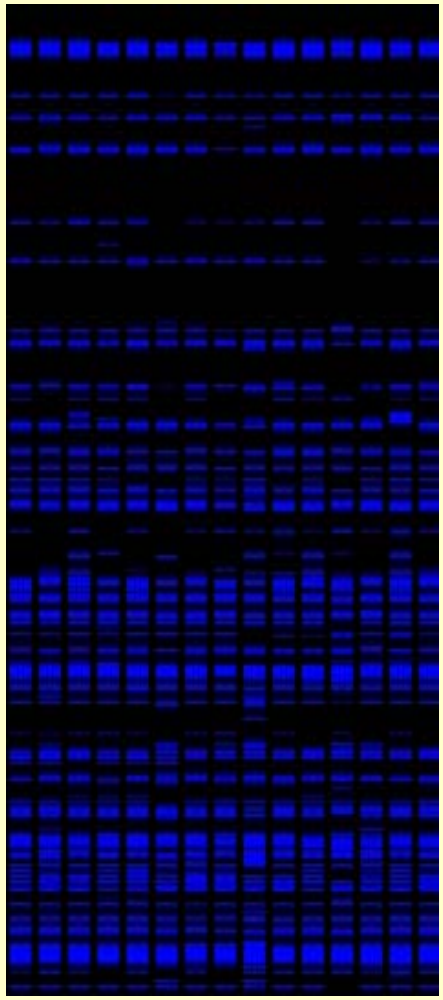
Used for population structure, phylogeography, pedigrees, forensics

# Amplified Fragment Length Polymorphism (AFLP) Markers

- A fast, efficient, and relatively inexpensive approach for generating multilocus DNA fingerprints
- Large numbers of markers can be generated with a single PCR reaction
- Data are represented as “Presence/Absence” information
- Downside:  AFLP markers are “Dominant” markers, meaning that heterozygous individuals can not be discerned.

If allele ‘A’ at a locus produces a band and allele ‘a’ does not, we do not know if an individual has the genotype ‘AA’ or ‘Aa’

Larger size



Smaller size

# Why should Ecologists and Organismal Biologists care about Genetics??

- With genetic data in hand, diverse inferences can be made about the ecology, demography, behavior, and history of a given species
- May be able to make inferences that would otherwise not be feasible (or at the very least, not possible for a given investigation)

# Specific types of questions that you can ask with genetic data

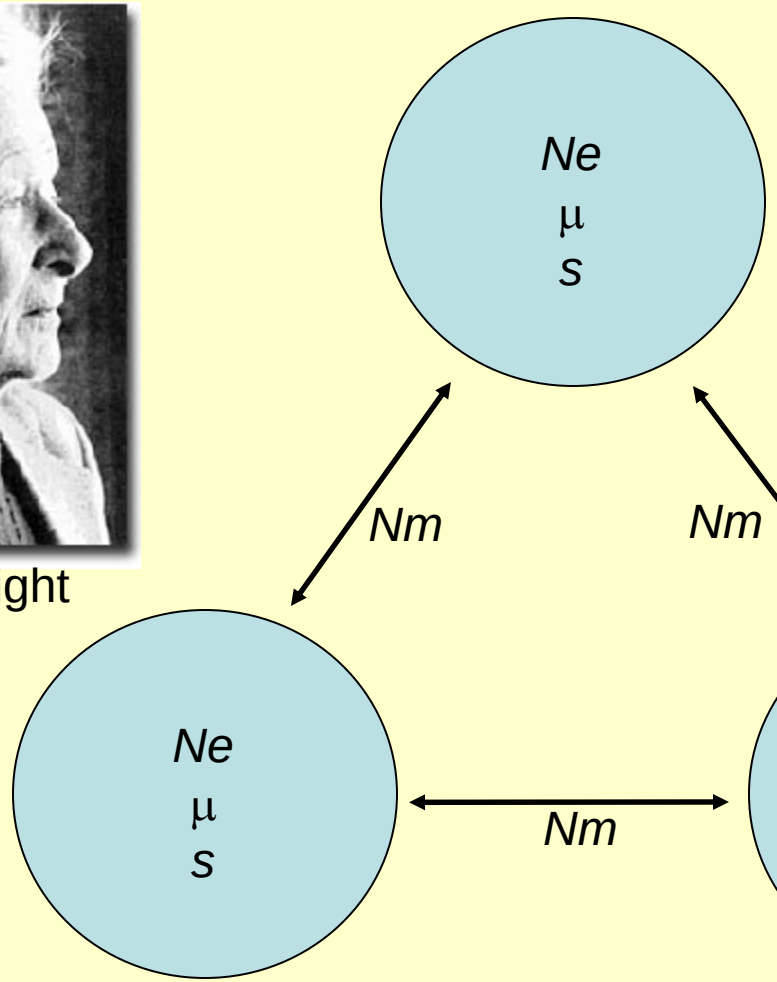
- Are populations genetically different from one another? Does the level of genetic differentiation vary among populations?
- How many populations are there?
- How much genetic diversity is in each population?
- Have there been demographic changes to a population? Is the population increasing or decreasing in size? Have there been genetic bottlenecks?
- Is there evidence for inbreeding within a population?
- From what geographical region or population did an individual originate?
- What is the effective size of a population?
- Are different species/subspecies hybridizing?

Many, many more too!

But this is the MIGRATE course: we will focus on general overviews of inferring movement patterns using genetic data

# Population Genetics: The Foundations

Sewall Wright




R.A. Fisher

*Ne* = effective population size
*s* = selection coefficient

μ = mutation rate
***Nm* = migrants per generation**

# Links between organismal movement and population genetic structure

- Gene flow effectively transfers alleles among different populations Under high gene flow, populations become more similar to one another
- Under low (or no) gene flow, populations become completely genetically differentiated
- Under intermediate gene flow, populations become slightly differentiated from one another

An example: Genetic structure differences in migratory vs. sedentary house wrens (*Troglodytes aedon* vs. *T. musculus*)


FIGURE 1. Sampling sites for migratory and sedentary House Wrens (*Troglodytes* spp.). The shaded area shows the breeding distribution of House Wrens on the American continents.

From Arguedas and Parker (2000)
The Condor 102: 517-528

How genetically differentiated are populations from each species?

Quantified allelic variation at 5 microsatellite loci using $F_{ST}$, a measure of genetic differentiation (more information about $F_{ST}$ later).

Migratory:
$F_{ST}$ = 0.0033

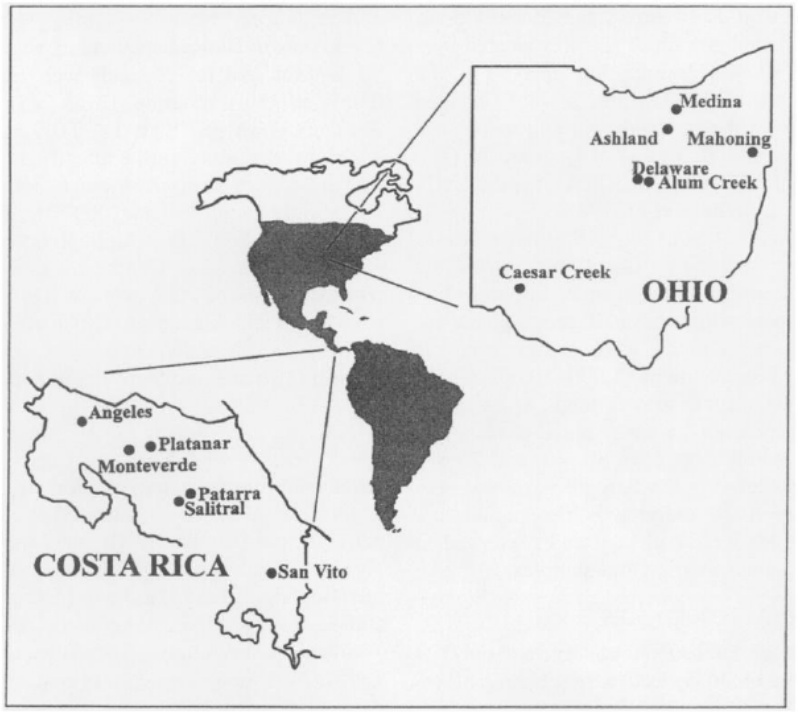



FIGURE 1. Sampling sites for migratory and sedentary House Wrens (*Troglodytes* spp.). The shaded area shows the breeding distribution of House Wrens on the American continents.

Sedentary:
$F_{ST}$ = 0.012

Migratory bird populations show less genetic differentiation than their sedentary counterparts, reflecting their greater tendency to move among locations

What if you know little about the dispersal behavior of a given species??

Very likely to have genetic structure!

Will genetic structure exist???

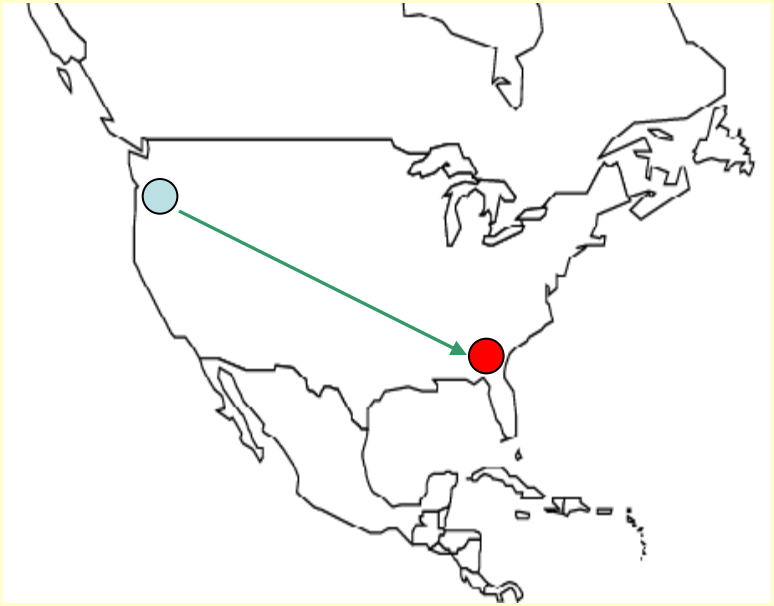
Genetic analyses can start to provide insights to organismal movement

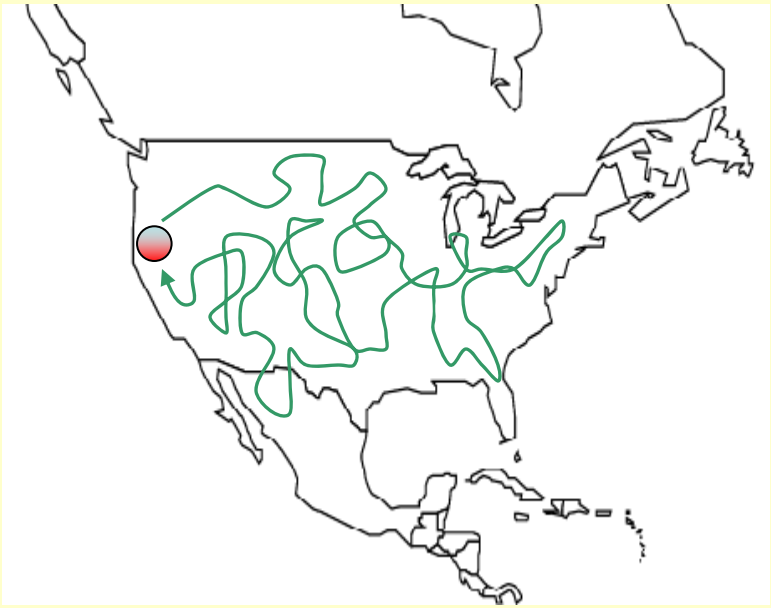
# What type of dispersal is important from a population genetics perspective??

- The only important type of movement from a population genetics perspective involves differences between the location of BIRTH and BREEDING sites
- Over time, the degree of breeding site fidelity and natal philopatry are important
  - *Breeding site fidelity*: returning to same location every year to breed
  - *Natal philopatry*: returning to birth site for breeding purposes
- All other movements do not matter!

Birth Site

Breeding Site



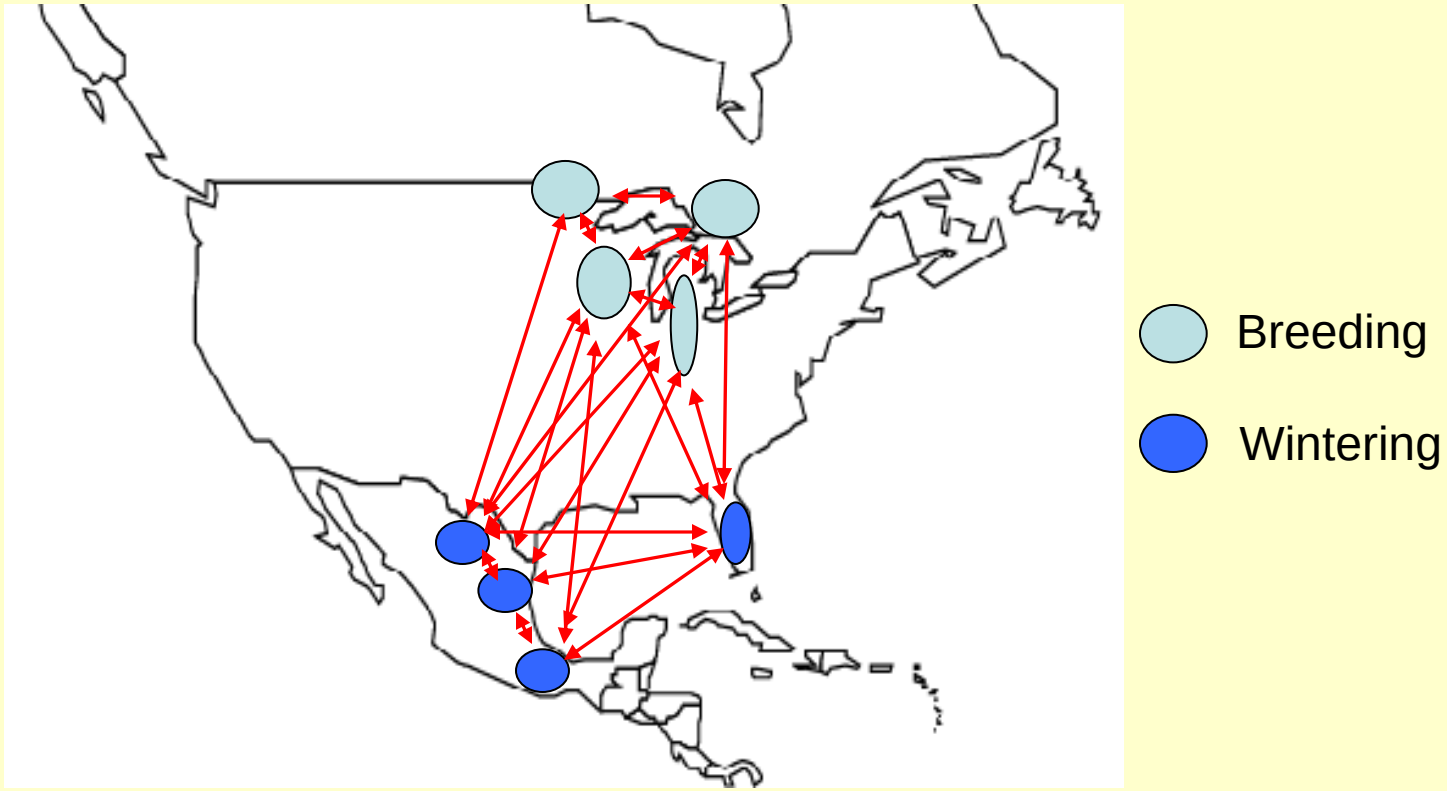


- Birth and breeding sites are different
- Gene flow has occurred
- Genetic structure is reduced

- Birth and breeding sites are identical
- No gene flow has occurred
- Genetic structure becomes greater

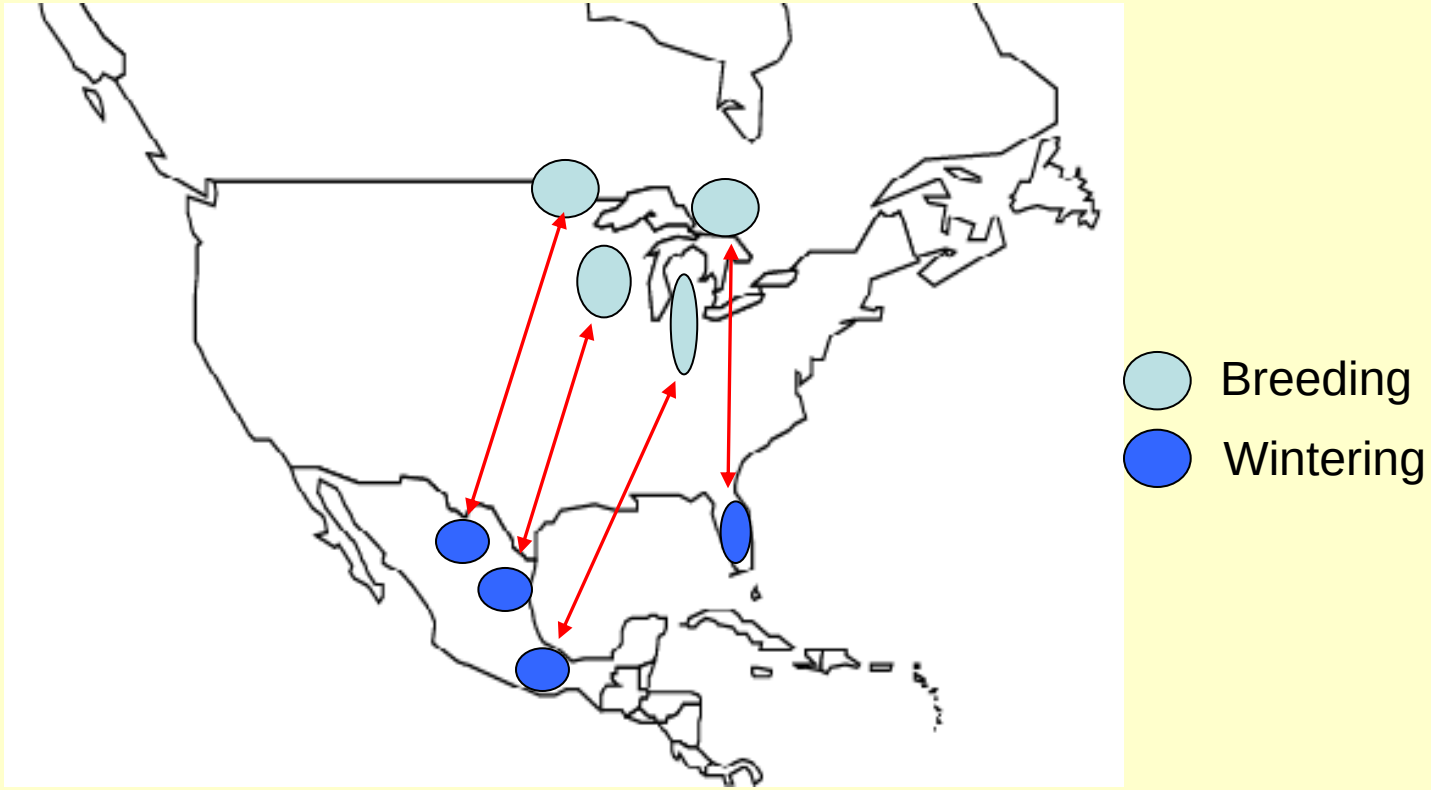
Based on this idea, what are some potential types of patterns that could be observed in migratory birds?

Case 1: High gene flow among breeding sites, with high movement among wintering areas




- No genetic structure exists
- There is a single “population” of organisms that inhabits a large geographical area

Case 2:
Extreme breeding site and wintering site fidelity
First year breeders return to birth sites (natal philopatry)




- Genetic structure will arise
- There are multiple “populations”

of organisms that inhabit different geographical areas

A short quiz….

Does wintering site fidelity matter????

# Once you have a genetic data set, how can you determine if population genetic structure exists?

Populations sampled for genetic analysis

# Data analysis

- Conventional approaches
  - Contingency table analyses
  - $F_{st}$, $\Phi_{st}$, $R_{st}$, $\theta$
- Isolation by distance analyses (Mantel tests)
  - Simple spatial approach
- Bayesian Clustering algorithms

# Conventional approaches: Ask the simple question: "Are populations genetically differentiated from one another?"

| | Allele | | |
|---|---|---|---|
| "population" | 1 | 2 | Total |
| 1 | 10 | 10 | 20 |
| 2 | 10 | 10 | 20 |
| Total | 20 | 20 | 40 |

| | Allele | | |
|---|---|---|---|
| "population" | 1 | 2 | Total |
| 1 | 20 | 0 | 20 |
| 2 | 0 | 20 | 20 |
| Total | 20 | 20 | 40 |

- Treat "populations" (or groups of individuals) as independent/causative variables and allele frequencies as dependent/response variables
- Perform analysis using $X^2$ or Fisher's exact tests
- Contingency table analysis where there are *R* rows (populations) and *C* columns (alleles)
- HIGH genetic differentiation is most generally interpreted as being associated with LOW gene flow between populations

# Other common approaches for determining the degree of differentiation of populations

- $F_{ST}$ measures:
  - Uses an ANOVA-like framework to quantify the among-population genetic variation
  - How much of the total genetic variation is found among populations relative to within populations?
  - Some different FST analogs that you may see in the literature:
    - $\theta$, $R_{ST}$, $\Phi_{ST}$
  - See http://www.uwyo.edu/dbmcd/popecol/Maylects/FST.html for an explanation of how simple values of Fst can be calculated
- In all of the above, values = 0 indicate no differentiation, whereas values = 1 indicate very high differentiation.

Genetic structure differences in migratory vs. sedentary house wrens (*Troglodytes aedon* vs. *T. musculus*)

Migratory:
$F_{ST}$ = 0.0033

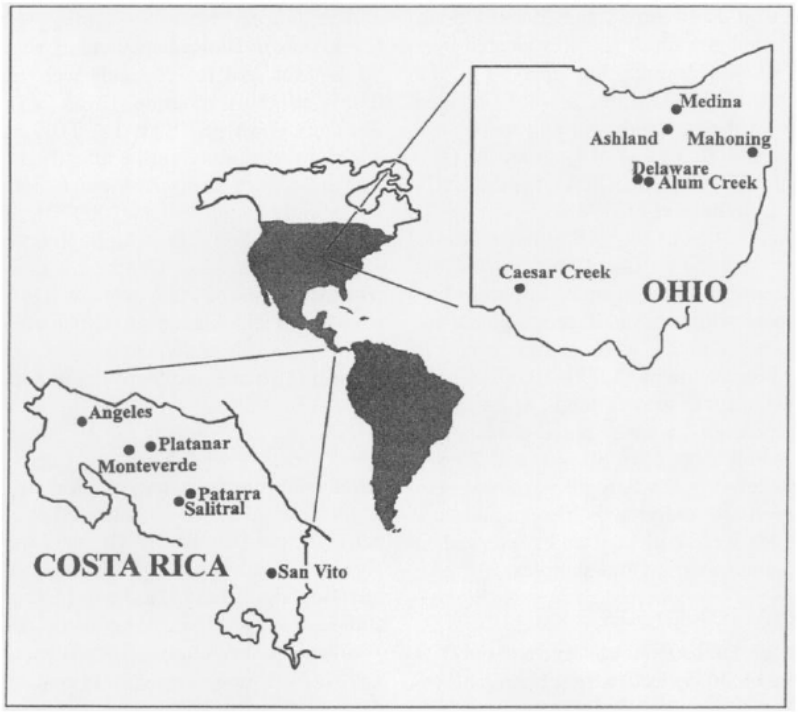



FIGURE 1. Sampling sites for migratory and sedentary House Wrens (*Troglodytes* spp.). The shaded area shows the breeding distribution of House Wrens on the American continents.

Sedentary:
$F_{ST}$ = 0.012

Migratory bird populations show less genetic differentiation than their sedentary counterparts, reflecting their greater tendency to move among locations

Genetic structure in the Southwestern Willow Flycatcher.
Busch et al. (2000), The Auk 117: 586-595

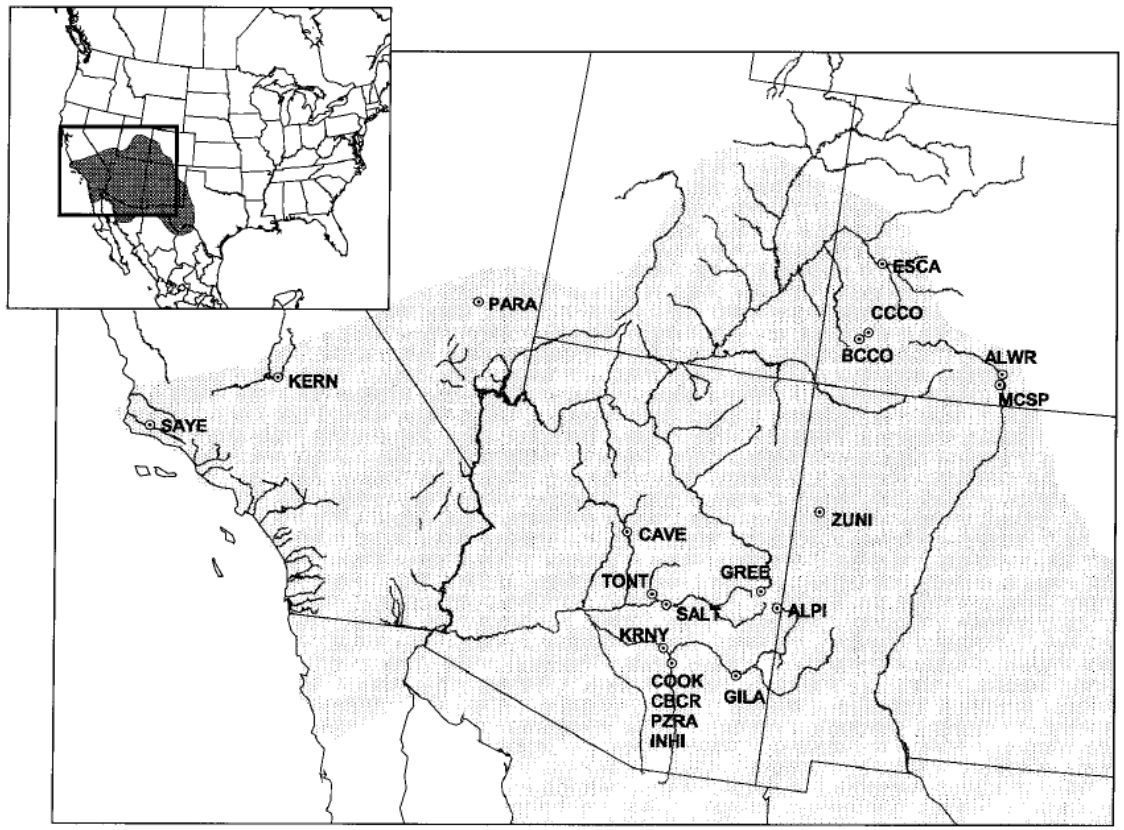



FIG. 1. Breeding range of the Southwestern Willow Flycatcher and collection locales (circles). Shading indicates the approximate boundary (Unitt 1987:figure 1, Browning 1993). Site abbreviations are in Table 1. COOK, CBCR, PZRA, and INHI are represented by a single circle but are distinct breeding sites in close proximity to each other.

$\theta = 0.0816$
95% CL = 0.061 to 0.103

$\Phi ST = 0.0458$, $P = 0.001$

**Suggests fairly high breeding site fidelity!**

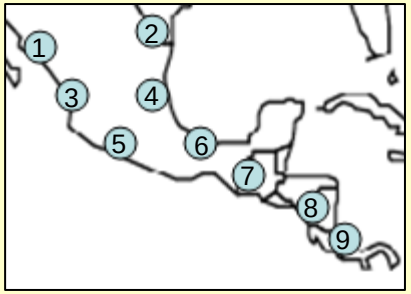
# A simple spatial analysis approach: Mantel tests

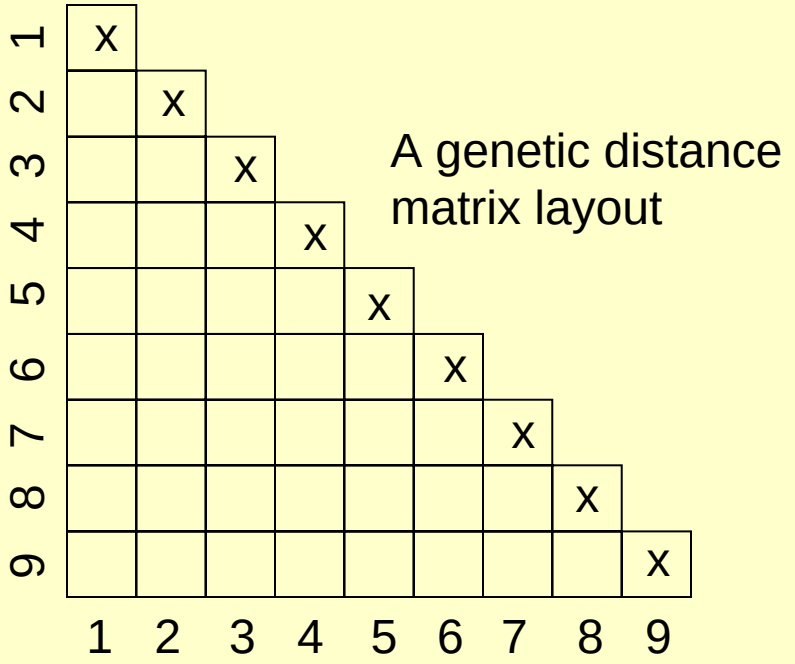
Populations sampled for genetic analysis

What if gene flow is higher between adjacent populations relative to populations that are far apart?

# Mantel tests:

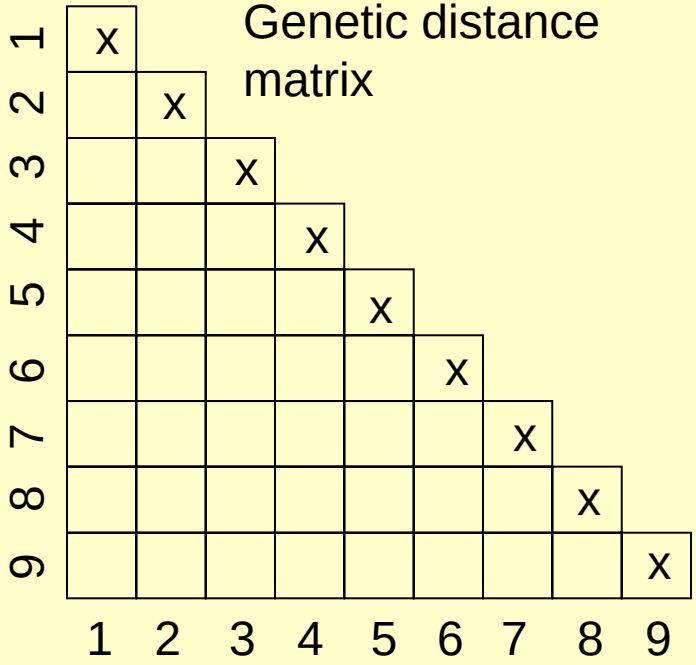
- Rather than calculating one overall statistic to reflect the level of genetic differentiation, genetic distances are calculated for all PAIRWISE combinations of populations.
- If you have *n* populations, there will be (*n**(*n*-1))/2 pairwise genetic distances

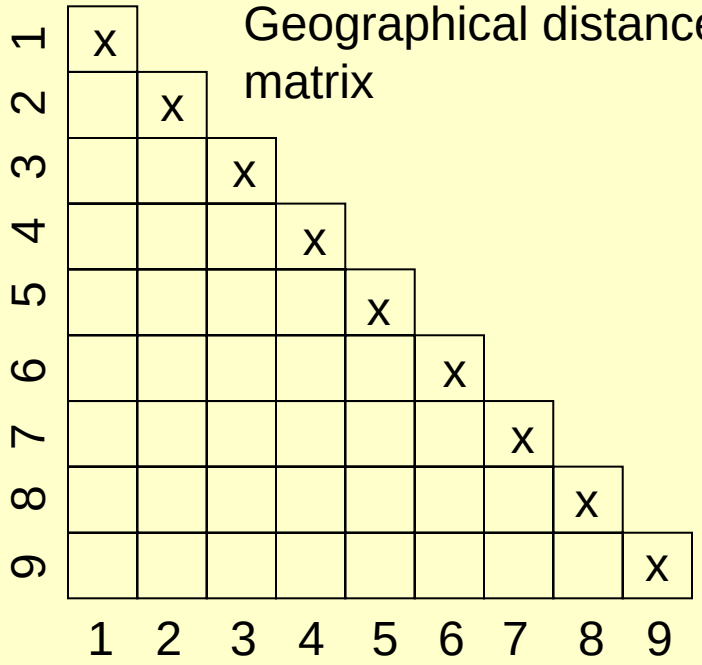



A genetic distance matrix layout

- Once you have a genetic distance matrix calculated, you can also generate a congruent geographical distance matrix.




- Can then use Mantel tests to determine if genetic distances are correlated with geographical distances
- Does genetic distance increase as geographical distance increases?.

# POPULATION GENETICS OF THE GALÁPAGOS HAWK (*BUTEO GALAPAGOENSIS*): GENETIC MONOMORPHISM WITHIN ISOLATED POPULATIONS. Bollmer et al. (2005), The Auk 122: 1210-1224

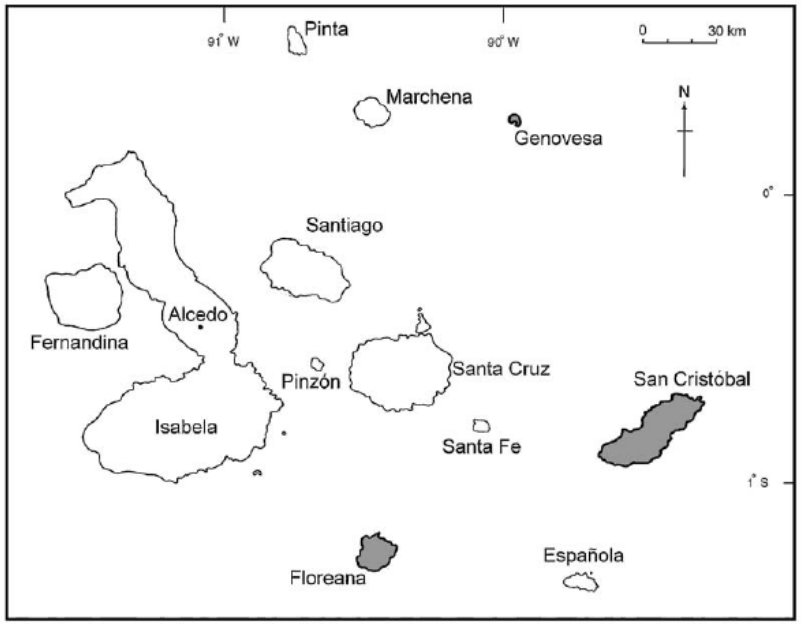



FIG. 1. Distribution of the Galápagos Hawk on the Galápagos Islands. All labeled islands currently have Galápagos Hawk populations, except for the three islands that are shaded. Genovesa has never supported a Galápagos Hawk population, and the populations on San Cristóbal and Floreana have been extirpated by humans.

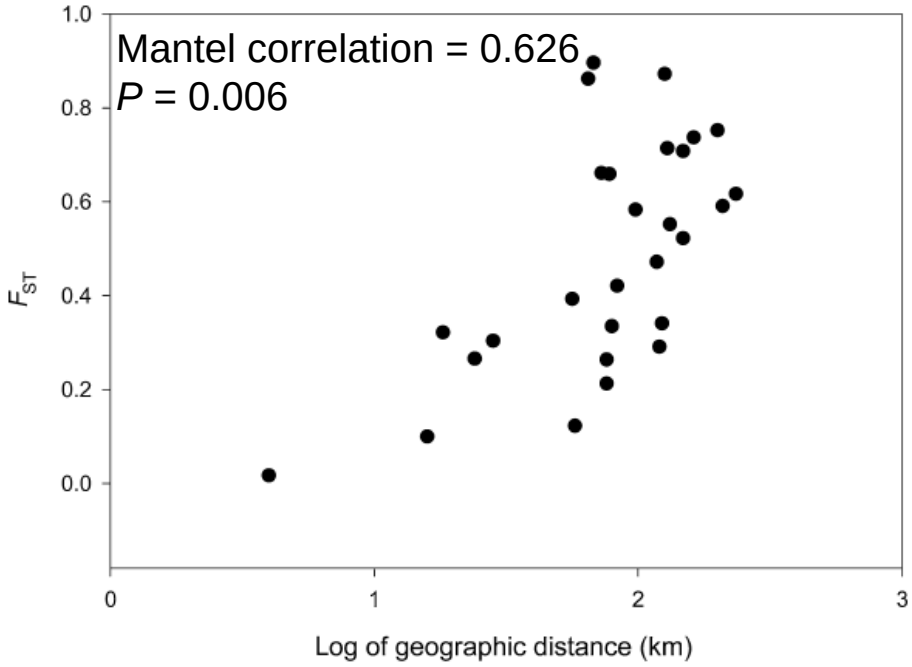
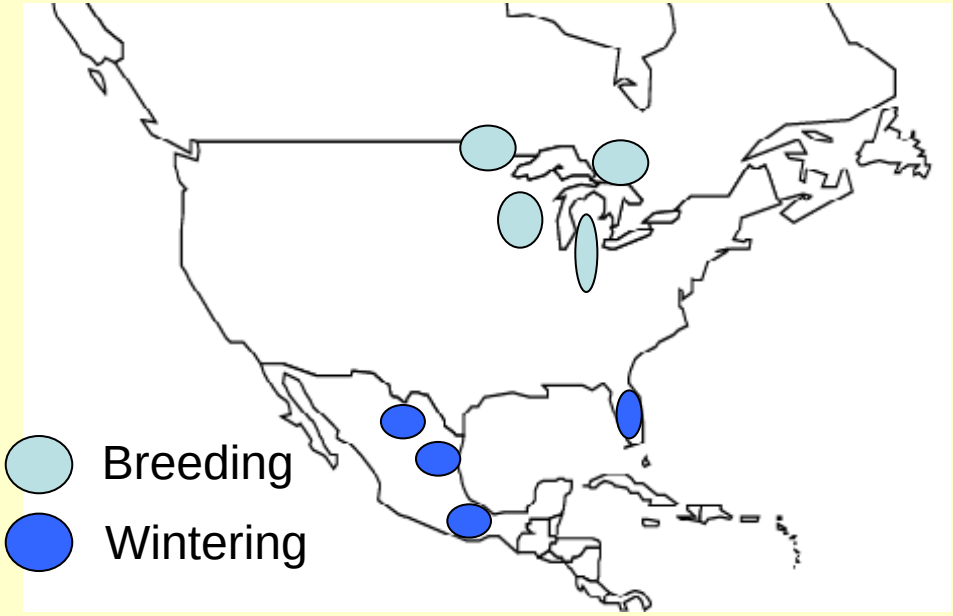



FIG. 4. Plot of pairwise interisland $F_{ST}$ values against the log of geographic distances (km) between islands for Galápagos Hawks. Degree of genetic differentiation between populations increases with increasing geographic distance.

# One more approach: Bayesian clustering procedures

- Computer program called “STRUCTURE”
  - Pritchard et al. (2000), Genetics 155: 945-959
  - http://pritch.bsd.uchicago.edu/software.html
- What if you have no ‘population’ information for your samples?
- What if you only have access to samples collected on the wintering grounds?




- We know that wintering ground fidelity does not affect genetic structure
- Birds collected throughout the wintering grounds may still be from separate breeding populations
- It would still be useful to “sort” birds into different groups to see if there is evidence for multiple populations!

- STRUCTURE is a *very* complex program based on a sophisticated model
- The analysis tries to:
  1) Identify how many 'clusters' of individuals exist based on the genetic data
  2) Assign each individual from your data set to one of the clusters

- An interesting note: when Pritchard et al. first described this approach, they used data from the Taita thrush (*Turdus helleri*) as an example (Data from Galbusera et al. 2000)

- Had data from 155 birds genotypes at 7 microsatellite loci

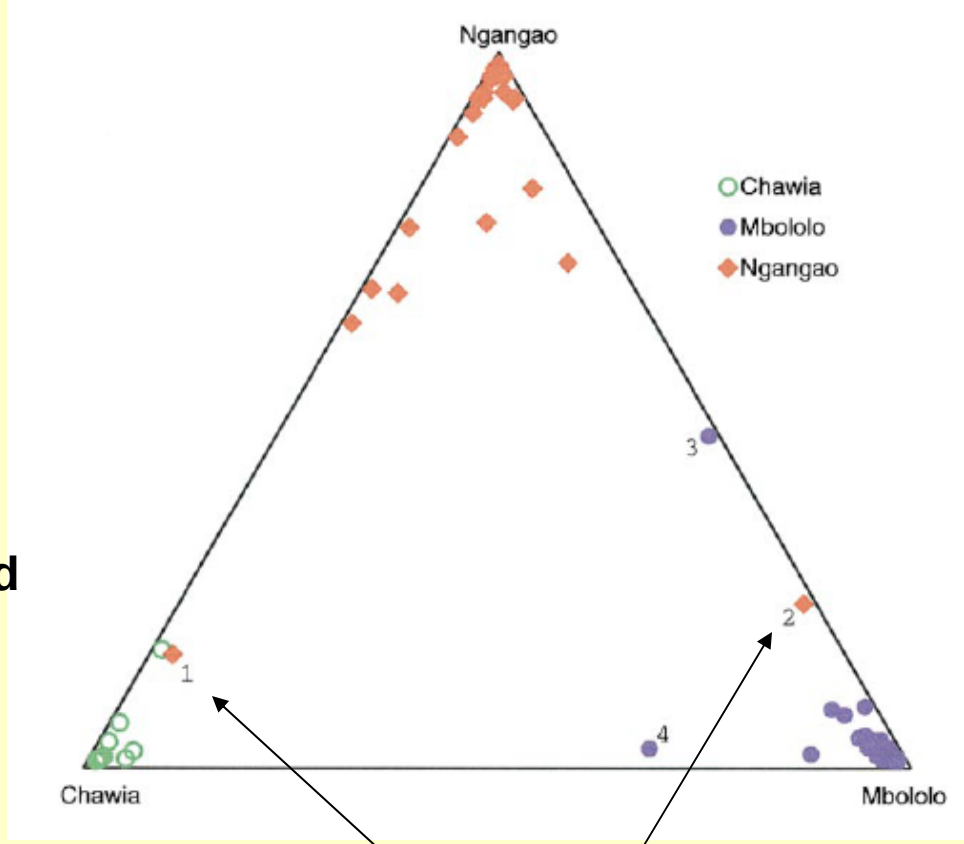
- When Pritchard et al. analyzed the data, they had no knowledge of where each individual was originally collected

- When Pritchard et al. analyzed the data, they had no knowledge of where each individual was originally collected

- Each dot represents an individual

-Dots closer to vertices of the triangle have greater probabilities of correct assignment.

- **Based on the analysis, 3 primary clusters of individuals were identified that largely corresponded to the *true* collection locations of individuals**




Migrants??

# How can I get started?

1) Read papers dealing with topics that are closely related to your specific research question
   - Pay attention to sampling designs from different studies
2) Learn more about laboratory techniques used to generate genetic data
   - Web searches!
3) Learn more about statistical procedures that are used to analyze genetic data
   - Use published papers and web searches!

# Genetic data analysis software

- Mark’s genetic software page
  - www.marksgeneticsoftware.net
- Arlequin
  - http://cmpg.unibe.ch/software/arlequin3/
- Genepop
  - http://genepop.curtin.edu.au/
- STRUCTURE
  - http://pritch.bsd.uchicago.edu/software.html
- Some additional lists of different software packages
  - http://courses.washington.edu/fish543/Software.htm
  - http://www.biology.lsu.edu/general/software.html