Studying the Evolution of the Domesticated Dog through Population Genetics and Bioinformatics

Lab adapted from Village Dog Genetic Diversity Project, c/o Dr. Adam Boyko (Cornell University College of Veterinary Medicine). Adapted for the University of Miami by Allie M. Graham, M.S.

Primary Literature and Review Papers:

Boyko, Adam R., Ryan H. Boyko, Corin M. Boyko, Heidi G. Parker, Marta Castelhano, Liz Corey, Jeremiah D. Degenhardt et al. "Complex population structure in African village dogs and its implications for inferring dog domestication history." *Proceedings of the National Academy of Sciences* 106, no. 33 (2009): 13903-13908. **Required**

Pennisi, Elizabeth. "Old Dogs Teach a New Lesson About Canine Origins."*Science* 342.6160 (2013): 785-786. **Required**

Shearin AL and EA Ostrander. 2010. Leading the way: canine models of genomics and disease. *Disease Models and Mechanisms*. 3: 27-34. **Suggested**

Parker HG, Shearin AL and EA Ostrander. 2010. Man's Best Friend Becomes Biology's Best in Show: Genome Analyses in the Domestic Dog. *Annual Reviews in Genetics*. 44:309-336. **Required**

Part I: Introduction to Genetic Markers

What is a Microsatellite?

Microsatellites consist of short, tandem repeats of DNA sequence. The sequence CAGCAGCAGCAG is a tri-nucleotide microsatellite with the *motif* CAG repeated four times.

Variation in the number of repeats in a microsatellite arises due to DNA-polymerase errors during the DNA replication process. During DNA replication, DNA-polymerase moves along a DNA sequence and adds complementary bases to the template strand. When this template is highly repetitive, as in a microsatellite sequence, the DNA-polymerase may "hiccup" and move forward or backward one full repeat before continuing replication.

For example if we begin with the microsatellite CAGCAGCAGCAG and DNA-polymerase moves back three nucleotide positions before it continues replication, this would result in a microsatellite with five repeats CAGCAGCAGCAGCAG or $(CAG)_5$. If DNA-polymerase moved forward three nucleotide positions this would result in a microsatellite with three repeats instead of four.

This process of gaining or losing a single repeat of a motif is called the *stepwise mutation process*. DNA-polymerase replication errors of highly repetitive sequences are fairly common thus the mutation rate in microsatellites, particularly microsatellites with large numbers of repeats, is very high.

Differences in the number of motif repeats for a particular microsatellite are also called *alleles*. Unlike a *single nucleotide polymorphism*(SNP) allele which can only be one of four possible states (A, C, G or T) there are a large number of possible motif repeats for a particular microsatellite which means there are countless possible alleles.

Microsatellites are considered to be *neutral* markers because unlike other parts of the genome, they do not code for proteins and thus we can assume they are not under selection. High mutation rate and large number of alleles make microsatellites especially useful molecular markers for population genetic and parentage assignment studies.

Microsatellites are characterized in individuals using *polymerase-chain reactions* (PCR) to amplify the desired genome region from a DNA sample. The PCR makes thousands of copies of the microsatellite, which are labeled with fluorescent markers. These sequences can then be visualized by a DNA analyzing machine, which separates fragments according to length. The process is conceptually similar to an agarose gel where shorter DNA fragments move faster through the gel matrix than longer fragments. Since microsatellites with varying numbers of repeats vary in total sequence length the alleles are actually scored according to length polymorphism (Figure 1). When characterizing microsatellites, the DNA analyzing machine reports fluorescent peaks corresponding to a particular sequence length. Though the actual DNA sequence remains unknown we can infer the number of repeats, and thus the alleles present, based on the total sequence length.

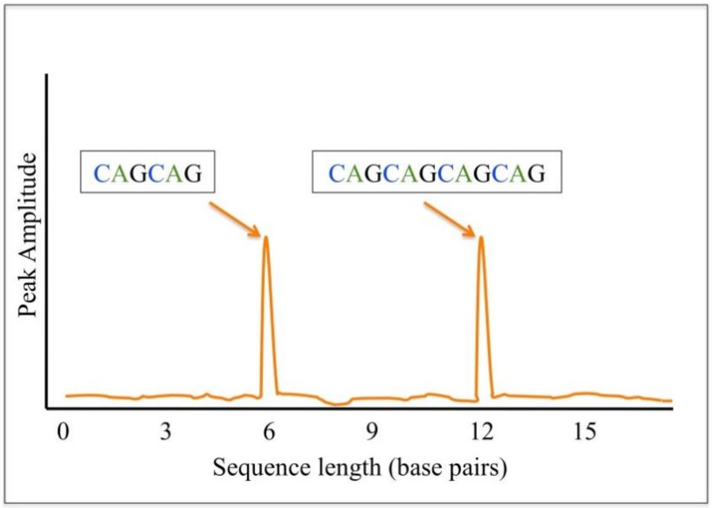

Figure 1: Example of peaks from analysis of microsatellites – indicative of length, but not content

QUESTIONS:

1. What are some of the advantages to using microsatellite markers?
2. Can you think of any potential disadvantages?
3. For which sorts of studies are microsatellites most appropriate? Why?

ANSWERS:

1. Microsatellites are very polymorphic because of their rapid mutation rate thus they can provide a lot of genetic information. Microsatellites are also useful because they are neutral markers therefore selection should not interfere with our ability to infer population history. Finally, microsatellites are relatively inexpensive to quantify, even in non-model organisms, so they can be applied to a wide variety of systems.
2. Microsatellites are limiting in that they sometimes violate the stepwise mutation process, a basic assumption for most population genetic analyses. Furthermore, since mutations can either add or remove tandem repeat units, homoplasy between alleles is possible, particularly when comparing distantly related groups.
3. Due to the potential for homoplasy between alleles, microsatellites are most appropriate for studies on recent evolutionary timescales such as recent population subdivision or paternity assignment studies.

What is a Heterozygote?

Diploid organisms have two copies of each chromosome, one from each parent. If the each parent contributes a chromosome with the same allele for a particular microsatellite the offspring will have two identical copies of this allele and is called a *homozygote*. A *heterozygote* is an individual with two different alleles for the same microsatellite and arises when the two parents contribute different versions of an allele (figure 2).

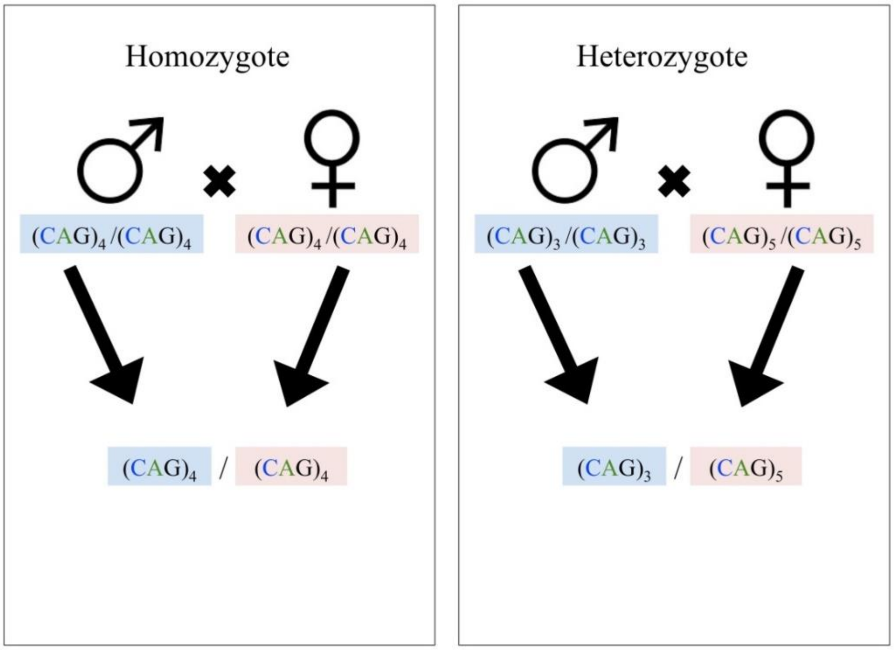

Figure 2: Formation of homozygous and heterozygous diploid offspring.

If we know the frequency of two alleles in a population (where p is the frequency of allele one and q is the frequency of allele two), we can estimate the expected frequency of heterozygotes and homozygotes in the population using the Hardy-Weinberg relationship:

$$p^2 + 2pq + q^2 = 1$$

p^2 is the proportion of the population homozygous for allele one
$2pq$ is the proportion of heterozygotes in the population
q^2 is the proportion of the population homozygous for allele two

For example if allele $(CAG)_3$ (or "p") has a frequency of **0.7** and $(CAG)_4$ (or "q") has a frequency of **0.3** we simply plug in the numbers:

$$(0.7)^2 + 2(0.7)(0.3) + (0.3)^2 = 1$$

As long as the assumptions for Hardy-Weinberg equilibrium are met (random mating, infinite population size, no selection, no new mutations, and no migration) we expect 49% of the population to be homozygous $(CAG)_3/(CAG)_3$, 42% of the population to be heterozygous $(CAG)_3/(CAG)_4$ and 9% to be homozygous $(CAG)_4/(CAG)_4$.

QUESTION:

1. Which allele frequencies for p and q will maximize the Hardy-Weinberg expected proportion of heterozygotes? Which allele frequencies will minimize the expected proportion of heterozygotes?

ANSWER:

1. Heterozygosity is maximized when the allele frequencies p and q are 0.5 (2pq= 2(0.5)(0.5)= 0.50) and minimized when either p or q are at low frequencies. For example if the frequency of p is 0.99 and the frequency of q is 0.01 the expected proportion of heterozygotes is approximately 2%.

What is a Population Bottleneck?

When populations are under strong natural selection or artificial selection, only a subset of individuals in the population will reproduce therefore relatively few individuals contribute alleles to subsequent generations. Alleles for gene-regions that are not under selection are present in the post-selection population as a random subset of the original allelic diversity. The probability of an allele being present in subsequent generations is equivalent to its frequency in the original population therefore high frequency alleles have a greater probability of being present in the post-selection population than low frequency alleles.

If selection pressure lasts for many generations, rare alleles will be lost simply by chance resulting in a post-selection population with fewer alleles and lower heterozygosity than the original population. This process is referred to as a *population bottleneck* (Figure 3). The overall loss of genetic diversity is proportional to the strength and duration of the selection event.

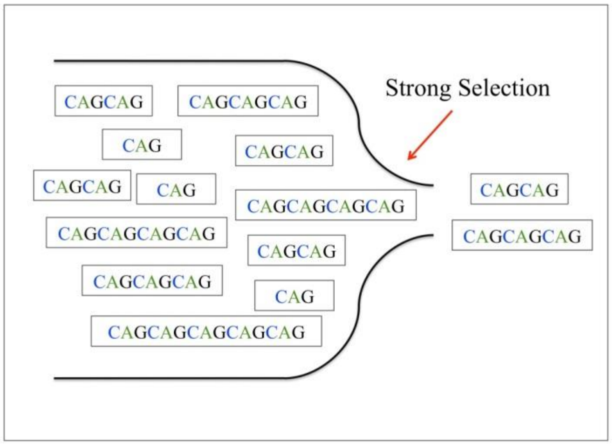

Figure 3: Strong selection in a population leads to dramatic shifts in allele frequencies in the post-selection population.

We can detect population bottlenecks in populations using molecular markers such as microsatellites to observe changes in allelic diversity and heterozygosity within populations. In particular we expect fewer alleles, monomorphic loci, smaller allelic size ranges, lower genetic diversity and less heterozygosity in populations that have undergone severe, prolonged bottlenecks.

QUESTION:

1. What processes other than artificial or natural selection could lead to a population bottleneck?

ANSWER:

1. Demographic processes, such as a population that is founded by a small number of individuals (i.e. an island colonization event) or a population that experiences strong reduction in population size due to a natural phenomenon (i.e. any random, mass mortality event) may result in a loss of low frequency alleles and a decrease in heterozygosity as a result of genetic drift.

Part II: Tracing the Domestication of Dogs using Molecular Markers

Population bottlenecks in Dogs: Dogs were probably domesticated from Eurasian wolves approximately 15,000-40,000 years ago. From the time of domestication, strong artificial selection for specific traits has resulted in the hundreds of dog breeds we recognize today.

Unfortunately it is difficult to understand the process of domestication and the genetics of breed formation without knowing anything about the genetic composition of the original domesticated, or *ancestral,* dog population. Populations of semi-feral, village or *indigenous* dogs from across the world may provide the best representation of the ancestral gene pool. In contrast with a population that is simply a mix of multiple dog breeds, we expect a true indigenous dog population would show genetic variation that is not present in any of the current dog breed populations.

Since we expect indigenous dog populations have not undergone artificial selection and we know dog breeds have undergone extreme artificial selection, we can compare genetic diversity across populations to check for the presence of severe population bottlenecks in breed dogs. We can also compare different sized populations of indigenous dogs to check whether there is evidence of minor population bottlenecks in smaller populations due to demographic processes.

QUESTION:

1. Based on the evolutionary histories of the following three dog populations, which do you expect will have the lowest genetic diversity? Which will have the highest genetic diversity? Why?
    1. Namibia-North Population: a large population of village dogs in Namibia
    2. Egypt-Kharga Population: a small, isolated population of village dogs in the Kharga Oasis of Sarahan Egypt
    3. Basenji Breed: one of the most ancient dog breeds, originates from Central Africa

ANSWER:

1. We expect Namibia-North have the greatest genetic diversity because it is a large population of indigenous dogs that has probably been minimally affected artificial selection and genetic drift. Egypt-Kharga should have an intermediate level of genetic diversity because though it has not undergone artificial selection, it has probably experienced genetic drift because of its degree of isolation and small population size. Finally, we expect the Basenji breed to exhibit the least genetic diversity because it has sustained strong selection and small population size.

Putting it all together (Using Arlequin) :

Note: Download Arlequin at http://cmpg.unibe.ch/software/arlequin35/. It is available for both Mac and PC. This program is integrated software for population genetics data analysis.

****This program has been cited above 10,000 times since their publication in both 1995 and 2000. This means they are widely used in actual research being published!*

Now let's test the predictions with real data. The attached input file consists of 20 microsatellite markers, which have been typed for five individuals from each of the three populations. Using the population genetic program **Arlequin**, obtain estimates of the following parameters to recreate the evolutionary history of each of these three dog populations.

1. Mean number of alleles: it is the number of copies of a particular allele divided by the number of copies of all alleles at the genetic place (locus) in a population.
2. Number of polymorphic loci: it is the number of loci which show polymorphism (ie. multiple forms) versus monomorphism (ie. one form)
3. Allelic size range: it is the size range of allele size (for microsatellite). The range comprises of the lowest size to the highest size.
4. Expected heterozygosity: it is the expected level of heterozygotic individuals in a population based on the Hardy-Weinberg equation.
5. Observed heterozygosity: it is the actual (observed) level of heterozygotic individuals in a population (individuals that have two different alleles at one loci).
6. Fst values: The fixation index (F_{ST}) is a measure of population differentiation due to genetic structure. It is frequently estimated from genetic polymorphism data, such as single-nucleotide polymorphisms (SNP) or microsatellites. It is one of the most commonly used statistics in population genetics.

Note: Download the program at http://cmpg.unibe.ch/software/arlequin35/. It is available for both Mac and PC. This program is integrated software for population genetics data analysis. The program has been cited above 10,000 times since it was published in 2000. This means that it is widely used by the scientific community!

(FYI) Relevant Results:

Population	Mean # Alleles	# Polymorphic Loci	Allelic Size Range	Expected Heterozygocity	Observed Heterozygosity
Namibia-North	3.65	20	10.5	0.62111	0.61000
Egypt-Kharga	2.85	20	7.75	0.53778	0.56000
Basenji	2.05	15	8.867	0.35667	0.45333

QUESTION:

1. Do the results satisfy your predictions? Why or why not?
2. Is there evidence of a strong population bottleneck in any of the populations?
3. (a) What did the Boyko et al (2009) paper using a larger data set find? Did they draw similar conclusions about the origins of modern domesticated dogs?
   (b) What did prior research suggest, and how was Boyko et al.'s results different?
   (c) What have we learned about the domestication of dogs recently (Consult "Old Dogs Teach a New Lesson About Canine Origins")? What are the issues with the research that has been performed thus far?

ANSWER:

1. Yes, our predictions matched our results. This is because the prediction was that the Namibia-North group would have the highest genetic diversity (highest mean allele, polymorphic loci, allele size range, observed heterozygocity), followed by Egypt-Kharga (intermediate in all values) then Basenji (lowest).
2. Although we did not run analyses specifically testing for the presence of a bottleneck, it is obvious that the Basinji breed has undergone a reduction in genetic diversity compared to the other populations. Ostensibly due to their small population size and continued breeding via breeders (ie. inbreeding, plus artificial selection).
3. (a) Their numbers were slightly different due to more markers (our data set had 20, while theirs had 89), but overall our results mimic their overall results [See Table 3 in Boyko et al., 2009). They were not able to rule conclusively that the origins of domestication occurred in Africa, but their results were able to cast doubt on the hypothesis that they were domesticated in Eastern Asia.
   (b) Prior research suggested that the domestication of dogs occurred in East Asia, while Boyko et al. (2009) results suggest non-East Asian ancestry, but possibly African origins.
   (c) New research suggests that the domestication of dogs may have occurred in Europe, pre-dating the agricultural revolution, suggesting hunter-gatherers were the initiators. This research is exciting because they utilized ancient DNA. However, main issues include incomplete sampling regimes.

Part III: Artificial Selection in Dogs (Village Dogs to Purebred dogs)

As humans migrated around the planet, a variety of dog forms migrated with them.
The agricultural revolution and subsequent urban revolution led to an increase in the dog population and a demand for specialization. These circumstances would provide the opportunity for selective breeding to create specialized types of working dogs and pets.

Through genetic mapping, scientists have been able to figure out which genes were artificially selected for during the domestication of the various breeds of dogs. Not all results of domestication were positive, and therefore a number of purebred dogs have multiple congenital defects. For this reason and others, dogs are fast becoming a model organism for human genetic diseases.

Please consult the following papers:
(1) Shearin AL and EA Ostrander. 2010. Leading the way: canine models of genomics and disease. *Disease Models and Mechanisms*. 3: 27-34.

**This article discusses how dogs are good model systems for the study of human diseases

(2) Parker HG, Shearin AL and EA Ostrander. 2010. Man's Best Friend Becomes Biology's Best in Show: Genome Analyses in the Domestic Dog. *Annual Reviews in Genetics*. 44:309-336.

**This article lists a number of genes putatively responsible for various traits in domesticated dogs.

Partial list of genes: IGF1 (body size), MSTN (muscle mass), FOX13 and FGF5 (coat characteristics), BRCA1 (canine mammary cancer), CDH2 (obsessive-compulsive disorder)

Use NCBI (http://www.ncbi.nlm.nih.gov/gene) to discover additional information about these genes in *Canis lupus familiaris*. Fill out information about each of these genes with the template below, and pick an additional gene from the "Man's Best Friend Becomes Biology's Best in Show" article.

Answers: The results of this section are subject to change since NCBI is constantly updating information.

Gene name: ___________________

(1) Gene description:

(2) Gene type:

(3) Location (on chromosome):

(4) Sequence:

(5) Homology – in what other organisms is this gene found? :

(6) mRNA and Protein(s) – how many different variants does this gene produce? :

(7) Describe the gene (use Related articles in PubMed):

(8) How might this gene be involved in a character trait which humans 'selected' during domestication?

Download Arlequin from here -
http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html

Arlequin Instructions:

(1) Search for Arlequin 3.5.1.3 (Go to Google or another search engine), the first link which shows up is "Arlequin 3.5 – Downloads". Click it! If you have Windows, then download "winarl35.zip" onto your desktop. If you have a Mac OS X, then download "arlecore_macosx.zip" onto your desktop. Unzip the files.
(2) Move the MicroLesson.arp file into the WinArl35 folder
(3) Open "WinArl35" within the folder
(4) Go to "Open Project" on the top left hand corner, and click
(5) Find the MicroLessor.arp file and open it
(6) It will bring up a page which looks like this, where you can see the three subpopulations of dogs we're analyzing data from
(7) Go to the "Project wizard" tab, and under "Data Type", choose "MICROSAT" from the drop-down selection.

Now we are going to tell the program what analyses to run…

(8) Go to settings tab, click on "Molecular Diversity Indices" under the Linkage Disequilibrium heading; check the "Standard diversity indices box". Don't change any of the other settings under this heading.
(9) Go to settings tab again, click on "Population Comparisons" under the Genetic Structure heading; check the "Compute Pairwise FST". Don't change any of the other settings under this heading.
(10) Now you should see there are red dots highlighting the following items in the Settings tab in this order: Calculation Settings, Genetic Structure, Population Comparisons, and Molecular Diversity Indices. If there are any others highlighted, click on them and then un-click the boxes.
(11) Go to the top right hand corner of Arlequin, and then click Start (which looks like a play button). The output from this analysis will show up in your web browser, and it should only take a few seconds to finish running.
(12) There will be a lot of information which is not relevant to this lab, only look at the "Standard diversity indices" for each of the three sample sets. Also "Summary of computations done within populations" subheading "Expected Heterozygosity", "Observed Heterozygosity", "Number of alleles", "Allelic size range"