A Primer of Microsatellite DNA Diversity in Rare Horse Breeds

E. Gus Cothran, Charlene R. Couch, and D. Philip Sponenberg

DNA testing in horses is commonly used to verify the parentage of individual animals. In addition, the same DNA technology can be used to measure important genetic characteristics of breeds, or populations. DNA analysis of populations can reveal levels of **inbreeding** (mating between relatives) and **genetic diversity** (a measure of genetic adaptability). Small populations, such as rare breeds, often suffer from high levels of inbreeding and low genetic diversity due to isolation and to limited numbers of breeding animals in the population. These circumstances result in increased levels of mating between related individuals, thereby increasing levels of inbreeding. Severe inbreeding and diminished genetic diversity can have negative consequences for a population, causing loss of reproductive ability and decreased hardiness and adaptability. Understanding the measures of inbreeding and genetic diversity within a rare breed can help to guide conservation decisions and population management.

One specific technique for measuring parentage or genetic diversity is the use of **microsatellite DNA** testing. DNA is composed of two strands of **nucleotides**, called A, T, C and G. Microsatellites are small regions of DNA that have repeated nucleotide sequences (like CTCTCTCTCT…CT). These microsatellite regions (or **loci**) are scattered throughout the DNA of all animals, including horses. Different microsatellites occur at many loci throughout the animal's genetic material. They tend to be quite variable in their number of repeats. These length variations, also called **alleles**, make them useful **genetic markers** for identifying parentage or measuring genetic diversity within and among breeds, much like DNA fingerprints.

Before pursuing any study of horse genetics, it is useful to define two concepts, **genotype** and **phenotype**. The genotype is the actual genetic material the animal carries, like a microsatellite DNA sequence. The **phenotype** describes the outward physical characteristics of the animal, like color or height, or even performance. Phenotype is determined by the combination of both genotype and the environment. Conditions like nutrition and training can influence phenotype. A slightly more challenging concept to understand is that many genes, not just one or a few, usually influence the traits we observe. Two animals can have similar phenotypes but very different *overall* genotypes. For instance, a Paint horse and a paint-colored Choctaw horse do not share the same genetic background despite sharing similar color genes.

Table 1, at the end of this document, outlines the microsatellite DNA diversity for one common breed (Thoroughbred) and 33 rare breeds. This data set was excerpted from a larger study by Dr. Cochran. Most of the genetic variability measures described in the table may seem to be rather similar, and to some degree they are. These measures are simply several different ways of looking at diversity within the breeds. However, each measure tells us something slightly different that is important. Some definitions may help in interpreting the table information:

The first column is **BREED**. The relatively common Thoroughbred breed is used as a comparison for all the rare breeds listed in the table. Because rare breeds are small in population size, when they are

compared to the more numerous Thoroughbred they may show higher levels of inbreeding **(F_{IS})** and lower levels of genetic diversity **(H_e)**. These measures are explained in greater detail below.

The second column, labeled ***N***, is the number of samples that were tested. The greater the number of horses tested, the more likely the technique is to detect variation. As a consequence, if a relatively large number of horses is tested (large *N*) and shows relatively low measurements of diversity, this is likely to be a real indication that there is low genetic variation in that breed. In contrast, a relatively small number of horses tested (small *N*) means that the results need to be interpreted very cautiously. Additional sampling may well reveal more variation in the breed than was found within the relatively few samples.

Generally speaking, a larger sample size gives a more accurate picture of what is actually going on within the breed. Smaller sample sizes should be interpreted with more caution. A sample size of 20 is generally the smallest number we can use for realistic interpretation of the genetic measures, but numbers nearer to 50 will give truer estimates for each measure. Larger sample sizes are particularly important for landrace breeds (like the Newfoundland Pony) that can be highly variable between family bloodlines. A similar example is strains of horses that have been isolated from each other for many years, like the Colonial Spanish horses, which share a common origin but have diverged from each other through many years of separation. Therefore sampling methods need to target all these important reservoirs of genetic variation in order to reveal the truest measures of genetic variation in a breed.

The other columns in Table 1 relate to the findings based on microsatellite length variants (or alleles) at the specific sites (loci) that were tested. Because each horse has two alleles at every microsatellite site, one from its mother and one from its father, it can either have two alleles that are the same (**homozygous**) or two alleles that are different (**heterozygous**). Therefore, **heterozygosity** is the condition where the two alleles at a locus are different from one another. Heterozygosity of a population is an important indicator of genetic diversity.

In Table 1, the measure ***H_o*** is **observed heterozygosity**. The H_o number represents the average fraction of all the microsatellite loci that are heterozygous per individual. This measure is based upon the actual number of heterozygous genotypes observed in the breed. The H_o can vary a bit based on the number of individuals sampled and the number of different microsatellite loci used for the analysis.

H_e is **expected heterozygosity**. It reflects the number of loci that *should* be heterozygous given a population of this size. This is based on the number of loci used and on the frequency of each allele found in the population, rather than the number of loci that actually are heterozygous (H_o). *High H_e represents high expected heterozygosity and is an indication of high genetic diversity*. For example, H_e measures above about 0.750 indicate relatively high genetic diversity. Lower H_e indicates lower genetic diversity. Most of the breeds in Table 1 have moderate to high H_e.

The measure of H_e, when compared to H_o, can help us to spot disruptions in the breeding system of a population. For example, high H_e with a low H_o may indicate inbreeding within the population. It also could be a result of population subdivision or sampling error in the individuals tested so that a truly representative sample from the breed was not examined. The opposite condition, high H_o with low H_e might indicate mixing between breeds or populations. Most of the breeds in Table 1 have relatively

similar H_o and H_e, but statistical analysis of several populations might reveal a few differences. Both H_o and H_e can range from 0 up to 1.0, with 0.5 being average heterozygosity.

F_{IS} or **inbreeding coefficient** is just what it sounds like—an estimate of the inbreeding level in the population. If the observed heterozygosity matches the expected heterozygosity, the result is no inbreeding. Positive values indicate that some degree of inbreeding is occurring, and negative values indicate little to no inbreeding. The American Cream Draft breed in this study, for instance, has a low inbreeding coefficient of -0.019. This reflects the fact that, even in a rare breed, protocols for reducing inbreeding can be quite effective. Inbreeding measures that are significantly less than zero may indicate sampling error, or it may suggest some level of mixing between populations.

A_e is the **effective number of alleles**. This number measures the diversity (number of alleles at each locus) of the marker system that is being used (in this case, the different microsatellite loci). Lower A_e indicates lower heterozygosity of the marker loci. The A_e is generally smaller than the total number of alleles.

N_A is the **number of alleles**. This is a total of the number of microsatellite alleles found across all the loci tested for the breed. The smallest number of alleles found in this study was in the Santa Cruz horse population. Low numbers of alleles can represent lower diversity, as is also reflected in the Santa Cruz measure of H_e.

MNA is **mean number of alleles.** This is the average number of microsatellite alleles across all of the loci tested. This reflects the overall genetic diversity relative to the number of loci that were examined.

RARE refers to **rare alleles**, which are those that occur at a frequency of 0.05 or less. These are the alleles that are most likely to be lost unless conservation strategies target those animals with those specific variants. A high value for *RARE* indicates a high proportion of rare alleles. The highest proportion of rare alleles in this study was found in the Lippitt Morgan breed, 0.404, compared to the Thoroughbred population at 0.271. Rare alleles can be an indication of valuable breed-specific diversity that should not be lost. Rare alleles point to the need for targeted conservation breeding efforts.

One final note: when comparing data across different genetic studies for a breed, it is important to understand how microsatellite loci behave genetically compared to other types of gene marker systems that may be used. Microsatellites have a higher mutation rate as compared to many other marker DNA types. This is because of the molecular structure of microsatellite DNA repeat sequences. When factors (like inbreeding or population mixing) cause change in the genetic variation of a population, microsatellites can lose alleles at a greater rate than they lose heterozygosity, and in many cases this makes variables related to allele numbers more informative than heterozygosities. A key point is that all analyses must compare apples to apples, or similar marker data to each other. The absolute values of specific measures are not as important as comparing the same measures among different populations.

DNA testing is a powerful tool, but as we have seen, it requires much careful interpretation. Sample size, marker type, and population subdivision are all variables that must be taken into account for any genetic analysis of breeds. That said, DNA genetic analysis is a valuable starting point for any effective breed conservation.

Table 1. Microsatellite DNA diversity of rare and common breeds of horses.

BREED	N	Ho	H_e	F_{IS}	A_e	N_A	MNA	RARE
THOROUGHBRED	139	0.651	0.708	0.0803	3.694	85	5.667	0.271
AKHAL TEKE	29	0.726	0.643	-0.129	3.341	81	5.400	0.222
AMERICAN CREAM DRAFT	25	0.725	0.711	-0.019	3.735	92	6.133	0.261
ARCHIVAL MORGAN HORSE	50	0.638	0.743	0.141	4.257	110	7.333	0.336
LIPPITT MORGAN	63	0.624	0.623	-0.001	2.915	89	5.933	0.404
BACA HERD NM	34	0.687	0.611	-0.125	2.893	80	5.333	0.313
BELGIAN DRAFT	20	0.653	0.656	0.004	3.419	83	5.533	0.253
CANADIAN HORSE	76	0.701	0.702	0.001	3.886	99	6.600	0.343
CASPIAN HORSE IRAN	57	0.775	0.775	0.001	4.701	124	8.267	0.347
CASPIAN HORSE USA	89	0.645	0.737	0.124	4.463	111	7.400	0.324
CHOCTAW	25	0.807	0.762	-0.058	4.447	81	6.750	0.259
CLEVELAND BAY	90	0.629	0.618	-0.016	2.870	73	4.867	0.274
CLYDESDALE	32	0.602	0.582	-0.035	2.563	66	4.400	0.258
COROLLA	38	0.540	0.522	-0.034	2.422	60	4.000	0.133
DALES PONY	86	0.712	0.688	-0.035	3.653	92	6.133	0.283
DARTMOOR PONY	75	0.712	0.671	-0.061	3.319	95	6.333	0.326
EXMOOR PONY	46	0.598	0.625	0.043	2.841	79	5.267	0.329
FELL PONY	55	0.722	0.716	-0.008	3.763	88	5.867	0.250
FLORIDA CRACKER	58	0.654	0.696	0.061	3.557	97	6.467	0.340
FRIESIAN	143	0.475	0.519	0.084	2.524	76	5.067	0.342
GALICENO	32	0.765	0.756	-0.011	4.505	104	6.933	0.260
GOTLAND RUSS TOTAL	73	0.643	0.646	0.005	3.058	85	5.667	0.365
GOTLAND RUSS SWEDEN	43	0.640	0.643	0.005	3.020	82	5.467	0.341
GOTLAND RUSS US	30	0.647	0.578	-0.118	2.649	54	3.600	0.130
HACKNEY HORSE	26	0.672	0.652	-0.029	3.046	82	5.467	0.305
LIPIZZANNER	77	0.605	0.717	0.155	3.634	101	6.733	0.347
MARSH TACKY	23	0.751	0.718	-0.045	3.852	96	6.400	0.292
MOUNTAIN PLEASURE HORSE	75	0.752	0.758	0.008	4.535	115	7.667	0.296
NEWFOUNDLAND PONY	58	0.724	0.736	0.0162	4.234	110	7.333	0.309
SANTA CRUZ ISLAND	50	0.541	0.480	-0.128	2.195	46	3.067	0.043
SHACKLEFORD BANKS	354	0.655	0.657	0.003	3.465	105	7.000	0.381
SHIRE	32	0.685	0.654	-0.047	3.134	81	5.400	0.321
SUFFOLK PUNCH	50	0.734	0.704	-0.043	3.796	82	5.467	0.171
WILBER CRUCE	47	0.689	0.626	-0.101	2.958	68	4.533	0.176