MAPPING GENES TO TRAITS IN DOGS USING SNPs

INTRODUCTION

Figure 1. Tasha. Scientists sequenced the first canine genome using DNA from a boxer named Tasha.

Meet Tasha, a boxer dog (**Figure 1**). In 2005, scientists obtained the first complete dog genome sequence using Tasha's DNA. Like all dogs, Tasha's genome consists of a sequence of 2,400,000,000 pairs of nucleotides (A, C, T, and G) located on 39 pairs of chromosomes.

What do scientists do with this information? This activity will introduce you to an approach for using genome sequence data to identify genes associated with an organism's characteristics, or phenotypes, called genome-wide association studies (GWAS). GWAS involve scanning DNA sequences across the genomes of a large number of individuals—in this case, many different dogs—to find differences, or variations, associated with particular phenotypes, to then guide the identification of the responsible genes. You will learn about the science behind GWAS by working through different exercises to find variations associated with a dog's coat color, length, and texture.

SNPs Are Common Variations in DNA Sequences

GWAS use DNA "markers" across the genome called single-nucleotide polymorphisms, or SNPs (pronounced "snips"). A SNP is a variation in a single nucleotide at a particular position in the genome (**Figure 2**). Not all single-nucleotide changes are SNPs. To be classified as a SNP, the change must occur in more than 1% of the population.

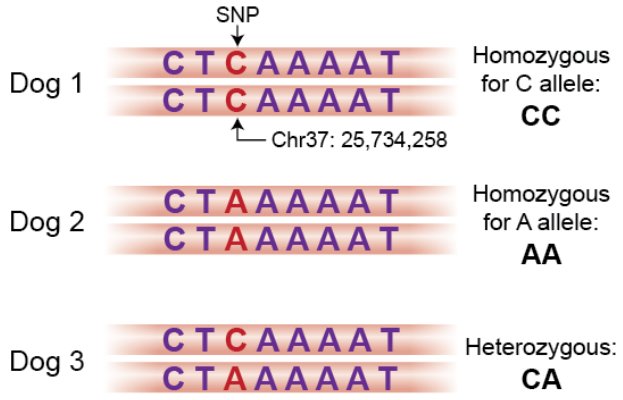


Figure 2. Example of a common SNP in dogs. This illustration shows three short segments of DNA sequence from chromosome 37 of the dog genome. There are two sequences per dog, one from the maternal chromosome and one from the paternal. Most dogs have either a C or an A at nucleotide position 25,734,258. Each nucleotide position on a given chromosome is assigned a reference number, counting from one end of the chromosome to the other. (Note that for each chromosome, the DNA sequence of only one strand is shown; the sequence of the complementary DNA strand is not shown.)

SNPs Are Identified by Comparing Many Genomes

After sequencing Tasha's genome, scientists sequenced the genomes of many dogs from a variety of breeds, comparing them to one another. They identified millions of common variations among these genomes and their locations on chromosomes. Specific locations are denoted by the chromosome number followed by the nucleotide number along the chromosome. For example, at a particular location some dogs have an A, while other dogs have a C (**Figure 2**). Dogs can have two copies of the C allele in this location, two copies of the A allele, or one of each allele.

SNPs Are Used to Find the Locations of Genes Associated with Particular Traits

Once we know where the SNPs are located in an organism's genome, they can be used to home in on the genes of interest. In a GWAS, scientists typically compare SNPs in two groups of individuals: one with one version of a trait (for example, dogs with long fur) and one with another version of the trait (for example, dogs with short fur). They then identify SNPs that are found to occur significantly more frequently in dogs with one version of the trait than the other (for example, short versus long fur). Those SNPs serve as "markers" for the region of the dog genome where a gene responsible for determining coat length resides.

Why are certain SNPs correlated, or associated, with certain traits? SNPs that occur within a gene or in a regulatory area near a gene may directly affect that gene's function and the resulting trait. For example, the change from an A to C in the **Figure 2** example may itself be responsible for a dog having long rather than short hair. However, a SNP does not have to be responsible for causing a trait in order to be correlated with that trait. If a SNP is close enough to a trait-causing allele, they may be inherited together (**Figure 3**).

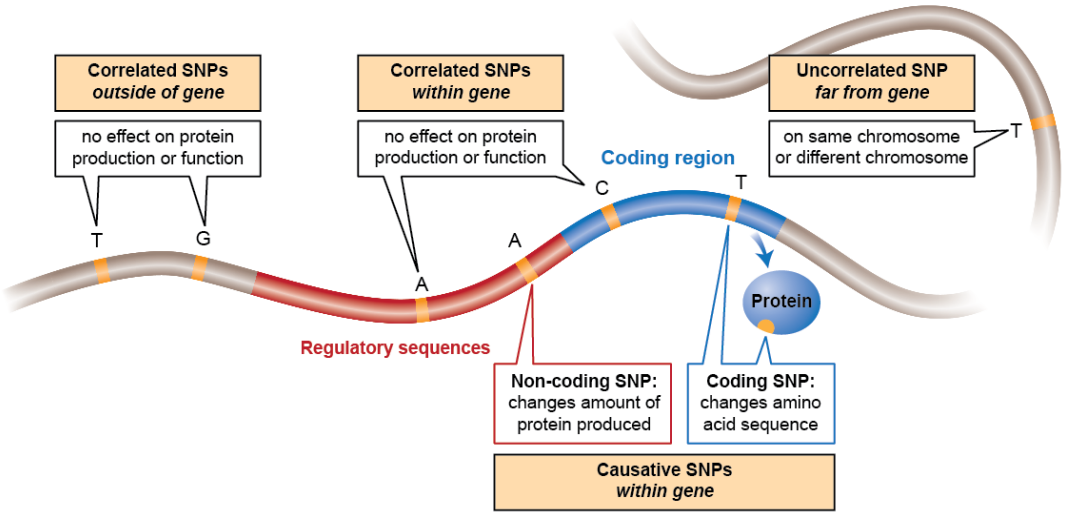


Figure 3. SNPs serve as markers for the locations of genes. SNPs can be associated with a trait because they occur within or close to the gene that causes that trait. The locations of the SNPs in the genome serve as markers or signposts for the locations of a gene correlated with a particular trait. (Graphic adapted from *Learn Genetics* "Making SNPs Make Sense.")

PROCEDURE

Part 1: Read and Answer Questions

To begin, read the article below and answer the questions that follow to check your understanding.

Variants in Three Genes Account for Most Dog Coat Differences

Variants in just three genes acting in different combinations account for the wide range of coat textures seen in dogs — from the poodle's tight curls to the beagle's stick-straight fur. A team led by researchers from the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health, reports these findings today in the advance online issue of the journal *Science*.

"This study is an elegant example of using genomic techniques to unravel the genetic basis of biological diversity," said NHGRI Scientific Director Eric Green, M.D., Ph.D. "Genomics continues to gain new insights from the amazing morphological differences seen across the canine species, including many that give clues about human biology and disease."

Until now, relatively little was known about the genes influencing the length, growth pattern and texture of the coats of dogs. The researchers performed a genome-wide scan of specific signposts of DNA variation, called single-nucleotide polymorphisms, in 1,000 individual dogs representing 80 breeds. These data were compared with descriptions of various coat types. Three distinct genetic variants emerged to explain, in combination, virtually all dog hair types.

"What's important for human health is the way we found the genes involved in dog coats and figured out how they work together, rather than the genes themselves," said Elaine A. Ostrander, Ph.D., chief of the Cancer Genetics Branch in NHGRI's Division of Intramural Research. "We think this approach will help pinpoint multiple genes involved in complex human conditions, such as cancer, heart disease, diabetes and obesity."

Artificial selection, at the heart of breeding for desirable traits in domesticated animals, has yielded rapid change in a short span of canine history. While researchers estimate that modern dog breeds diverged from wolves some 15,000 years ago, the genetic changes in the dog genome that create multiple coat types are more likely to have been pursued by breeders in just the past 200 years. In fact, short-haired breeds, such as the beagle, display the original, more wolf-like versions of the three genes identified in the study.

Modern dog breeds are part of a unique population structure, having been selectively bred for many years. Based on this structure, the researchers were able to break down a complex phenotype — coat — into possible genetic variations. "When we put these genetic variants back together in different combinations, we found that we could create most of the coat varieties seen in what is among the most diverse species in the world — the dog," Dr. Ostrander said. "If we can decipher the genetic basis for a complex trait such as the dog's coat, we believe that we can do it as well with complex diseases."

(Excerpt from NIH publication, 27 August 2009)

1. How many genes account for the wide variety of coat textures in dogs?

2. In two or three sentences, describe how scientists went about identifying these genes.

3. In this reading, why are SNPs referred to as "genetic markers" or "signposts"?

4. Why do you think it is important to analyze the DNA of many dogs when doing this research?

5. Do humans have SNPs?

6. How might the dog genome project benefit humans?

Part 2: Introduction to a GWAS

Let's explore how a GWAS works using a simple example. Scientists compare the SNPs in two groups of dogs: dogs with white fur and dogs with black fur. If one type of SNP is found much more frequently in dogs with white fur than in dogs with black fur, the SNP is said to be "correlated," or "associated," with the white coat color. The correlated SNPs mark a region of the genome that could contain a gene involved in determining the white coat color.

Table 1 shows alleles at 17 loci in the genome sequences of eight different dogs, four with white fur and four with black fur. Each locus is represented by a nucleotide at that particular location in the genome. (For simplicity, only one letter is shown, but in reality, each SNP locus has two nucleotides, one from each parental chromosome.)

To determine whether any of these loci have SNPs correlated with white coat color, you will compare the SNPs of the dogs with black fur and the dogs with white fur displayed in **Table 1**.

Table 1. Nucleotides at 17 Different Loci in Two Groups of Dogs.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	C	A	C	G	G	T	T	T	G	A	T	G	A	A	C	G	T
	C	T	C	C	A	T	T	C	G	A	T	T	G	A	C	T	A
	C	A	A	G	A	C	T	C	G	A	T	G	G	G	C	G	T
	C	A	C	G	A	T	T	C	G	G	T	G	G	A	C	G	T
	C	A	C	G	A	T	A	C	G	A	T	G	G	A	C	G	T
	C	T	A	G	A	T	A	T	G	A	T	G	A	A	C	G	A
	C	T	C	C	A	T	A	C	G	A	T	T	A	A	C	T	T
	C	A	C	G	G	T	A	C	G	G	T	G	A	A	C	G	T

1. A particular SNP is 100% correlated with coat color when *all* dogs with white fur have the same nucleotide at a particular position, while *all* dogs with black fur have another nucleotide. If a SNP is 100% correlated with a phenotype, the SNP may be located within the gene responsible for the phenotype and cause the difference in nucleotides that is responsible for the difference in traits. Draw a circle around the SNP(s) in Table 1 that are 100% correlated with coat color. What is the locus number(s) for the 100% correlated SNP(s)? _____

2. In general, the closer two DNA sequences are to one another, the more likely they are to be inherited together. So, if a SNP is close to a gene, it is likely to be inherited with that gene. The degree of correlation predicts how close a particular SNP is to a gene responsible for that phenotype. Draw a rectangle around the SNP(s) in Table 1 that are "somewhat correlated" to coat color. What is the locus number (or numbers) for the somewhat correlated SNP(s)? _____

3. Uncorrelated SNPs occur with about equal frequency in dogs with black fur and dogs with white fur. What is the locus number (or numbers) for the uncorrelated SNP(s)? _____

4. What are two possible explanations for why a SNP is correlated with a phenotype like coat color?

Part 3: Identify Correlations Using Real Data

When researchers scan the genome using thousands of markers to identify variations that are correlated with particular phenotypes, they need techniques to evaluate the strength of correlations so that they know which SNPs are closest to the locations of the genes of interest. How is this done?

Dog Coat Length

Your instructor will give you 12 *SNP Cards* for dog coat length (**Figure 4**). Each card represents sequence data obtained from one dog. The sequences show SNPs at seven loci on chromosome 32. The dog DNA was obtained from saliva samples collected by student dog owners. The DNA sequence was then determined by scientists at the Broad Institute in Cambridge.

Figure 4. An example of the SNP Cards. These cards show SNP alleles at seven loci in order along chromosome 32. Six of the dogs have a short coat and six have a long coat. Each locus provides two nucleotides—one on each chromosome 32. The topmost card shows the locations of the SNPs; it provides the chromosome number followed by a seven-digit nucleotide number (e.g., Chr32 7479580).

Take a few moments to look at the cards. Group all the sequences from the dogs with long coats and those from the dogs with short coats. Now, compare the two groups.

1. Looking at these cards, predict which SNPs are strongly correlated with the long- and short-coat phenotypes. Write your answer below and explain your reasoning.

In this next part of the activity, you will examine the data to measure which SNPs have the strongest correlation.

For each SNP locus on the cards, count the number of times an allele appears on the cards. Record the number in the appropriate box in the tables below. Note that if the SNP at a particular locus is homozygous (e.g., TT), you will count the T allele twice. If it is heterozygous (e.g., TC) you will count the T allele once and the C allele once.

Once you've recorded the number of times each allele appears, calculate the difference in alleles between the two groups of dogs and record that number. Add the differences for each allele and record the total number of differences in the last box of each table. **The largest total difference indicates the strongest correlation between that particular SNP and the phenotype.**

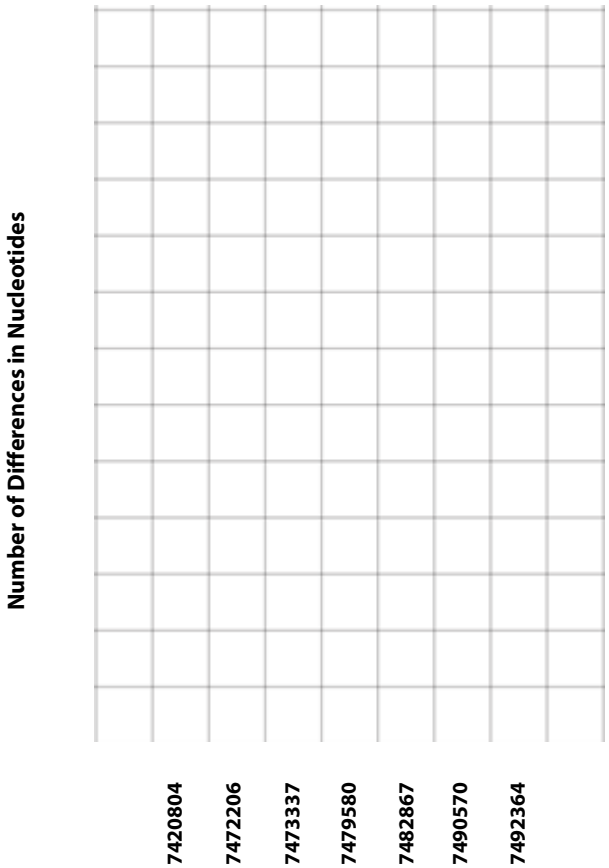
The first two loci have been completed for you. Complete the next five on your own.

Chr32 7420804	Allele	Short Coat	Long Coat	Difference
	T	4	4	0
	C	8	8	0
			Total number of differences	0

Chr32 7472206	Allele	Short Coat	Long Coat	Difference
	A	9	12	3
	G	3	0	3
			Total number of differences	6

Chr32 7473337	Allele	Short Coat	Long Coat	Difference
			Total number of differences	
Chr32 7479580	**Allele**	**Short Coat**	**Long Coat**	**Difference**
			Total number of differences	
Chr32 7482867	**Allele**	**Short Coat**	**Long Coat**	**Difference**
			Total number of differences	
Chr32 7490570	**Allele**	**Short Coat**	**Long Coat**	**Difference**
			Total number of differences	
Chr32 7492364	**Allele**	**Short Coat**	**Long Coat**	**Difference**
			Total number of differences	

Create a graph that plots the number of differences in nucleotides counted from your SNP cards for each of the seven loci.



1. Examine the graph. Which SNP (or SNPs) is most strongly correlated with coat length? (Write the SNP locus or loci in the space.) _____

2. Which SNP (or SNPs) is least strongly correlated with coat length? _____

Dog Coat Texture

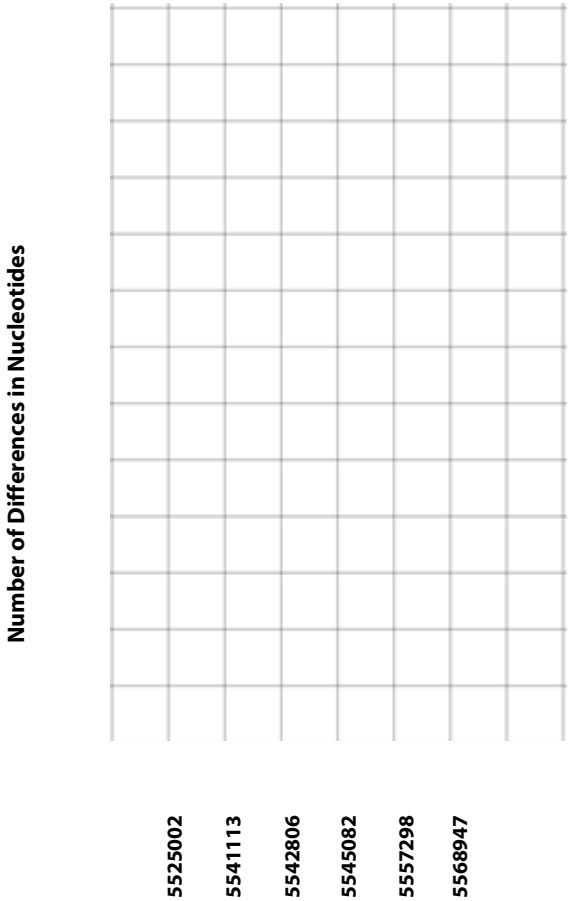
Next, your instructor will give you 10 *SNP Cards* for coat type. The data displayed on the coat type card shows SNPs at six loci on chromosome 27 in five dogs with a curly coat and five with a straight coat.

Using the same method, count the number of times an allele appears for each of the six loci and record the numbers in the appropriate boxes. Calculate the differences at each SNP locus.

Chr32 5525002	Allele	Curly Coat	Straight Coat	Difference
			Total number of differences	
Chr32 5541113	**Allele**	**Curly Coat**	**Straight Coat**	**Difference**
			Total number of differences	
Chr32 5542806	**Allele**	**Curly Coat**	**Straight Coat**	**Difference**
			Total number of differences	
Chr32 5545082	**Allele**	**Curly Coat**	**Straight Coat**	**Difference**
			Total number of differences	
Chr32 5557298	**Allele**	**Curly Coat**	**Straight Coat**	**Difference**
			Total number of differences	

Chr27 5568947	Allele	Curly Coat	Straight Coat	Difference
			Total number of differences	

Create a graph that plots the number of differences counted from your SNP cards against the six loci.



3. Examine the graph. Which SNP (or SNPs) shows the strongest correlation to coat texture?

_____

4. If you were to look for a gene involved in coat length, on which region of the genome would you focus your analysis? What about a gene for coat texture? Write your answer below.

_____

_____

_____

_____

Part 4: Chi-Square Analysis (Optional)

This part of the activity is optional. Complete it if it was assigned to you by your instructor.

You just identified SNPs that appear to be correlated with coat length and coat texture. How do you determine whether these correlations are real? It is possible that the frequency of alleles would differ between the two groups of dogs just by chance. On the other hand, if these differences did not occur by chance, the correlations might indicate that the SNPs mark the positions of genes involved in coat length and texture.

To determine whether these correlations are statistically significant (that is, unlikely to be due to chance alone), you will conduct a chi-square analysis.

The formula for calculating a chi square is as follows:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

In the equation, O and E refer to observed and expected counts. The greater the difference between the observed and expected values, the larger the chi-square value. How large does a differnce need to be before it is considered "significant"? For an experiment with 1 degree of freedom, differences with a chi-square value larger than 3.84 are considered significant.

Let's look at coat length. The null hypothesis (H_0) is that there is no difference in allele distribution between dogs with a short coat and dogs with a long coat and that any difference between the two groups occurred by chance.

Determine Expected Values

To determine whether you can reject the null hypothesis for the **SNP with the strongest correlation with coat length**, you will need to calculate the expected allele distribution. The expected value is equal to the total number of times an allele occurs at that locus divided by 2.

Let's look at an example using the SNP at chromosome 32: 7492364. That SNP has two alleles, C and G. The C allele occurs three times in dogs with a short coat and one time in dogs with a long coat, so four times in total. So, if there is no difference in distribution between dogs with a long coat and dogs with a short coat, you would not expect any difference in the two groups. The expected C allele distribution would be 4/2 = 2. If you do the same calculations for the G allele, the answer is 10.

SNP Locus: Chr32 7492364
Total number of C alleles = Short coat + Long coat = 3 + 1 = 4 Expected number of C alleles = 4/2 = **2**
Total number of G alleles = Short coat + Long coat = 9 + 11 = 20 Expected number of G alleles = **10**

1. Use the space below to calculate the expected allele distribution for the SNP that was ranked "most correlated" for coat length.

SNP Locus:

Calculate Chi-Square Values

You will now **calculate the chi-square value** using the observed number of alleles for the "most correlated" SNP above and the expected value you just calculated.

Allele	Short Coat Observed	Long Coat Observed	Expected
C	**3**	**1**	**2**
G	**9**	**11**	**10**

The calculation for the chi-square value for SNP locus Chr32 7492364 is shown below:

Short $(3 - 2)^2/2$ + Short $(9 - 10)^2/10$ + Long $(1 - 2)^2/2$ + Long $(11 - 10)^2/10$ =

0.5 + 0.1 + 0.5 + 0.1 = 1.2

2. Complete the table below for the SNP with the strongest correlation for coat length.

3. Now calculate the chi-square value (show your work):

4. Record the chi-square value: _____.

Determine the *p*-Value

Now, examine the chi-square distribution table below to find the *p*-value for the chi-square value you calculated.

In this study, there are two alleles, so degrees of freedom (df = #Allele categories – 1)(#Coat catgories – 1) is equal to (2 – 1)(2 – 1) = 1.

Allele	Short	Long	Expected

For the example, using the Chr32 7492364 locus:
$X^2 = 0.52$ with 1 degree of freedom. For an experiment with 1 degree of freedom, a chi-square value smaller than 3.841 indicates that the difference between two groups is not significant. For this particular locus, we cannot reject the null hypothesis that there is no difference between the two groups of dogs. In the table below, the position of the chi-square value X^2 of 0.52 has a *p* value between 0.975 and 0.2. This indicates that the probability of getting a difference as large as that observed by chance alone is much higher than our 5% cutoff.

	P										
DF	**0.995**	**0.975**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**	**0.002**	**0.001**
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515

5. Record the *p*-value for the SNP with the strongest correlation: _____

6. What does this *p*-value tell you?

7. Which SNP has the strongest correlation for curly versus noncurly coat?

8. Is this correlation statistically significant?

9. What is the evidence for your choice?

Part 5: What Are the Genes?

Now you know the basic concept of how SNPs are used to find correlations with traits. Once correlated SNPs are identified, researchers look in the region of the SNPs to find the genes of interest.

Using this technique, researchers analyzed the DNA from several gray wolves (species name: *Canis lupus*) and over 1,000 dogs from 80 recognized breeds. That analysis allowed researchers to identify three SNPs correlated with coat types. These SNPs occur within the genes *FGF5*, *RSPO2*, and *KRT71* (Figure 5). These are the genes that you read about in the article at the beginning of this document.

Gray wolves, the ancestors to modern domesticated dogs, have short, straight hair without furnishings. (Furnishings are tufts of hair over the eyes and around the mouth.) The gray wolves' versions of the *FGF5*, *RSP02*, and *KRT71* genes represent the "ancestral alleles," shown with a minus

sign in the chart in Figure 5. (The Basset Hound is most like gray wolves and has ancestral alleles of those three genes.)

More recent alleles found in some dog breeds have single-nucleotide changes in the DNA when compared to ancestral alleles. They are denoted by plus signs. Various combinations of these ancestral and more recent alleles account for the seven major coat types of purebred dogs (Figure 5). If you look at the versions of the *FGF5* gene in different dog breeds, you can see that the ancestral allele (minus sign) is associated with short fur and the allele with the more recent variation (plus sign) is associated with long fur.

Dog	Phenotype	Gene		
		FGF5	*RSPO2*	*KRT71*
Basset Hound 	Short	-	-	-
Border Terrier 	Wire	-	+	-
Airedale Terrier 	Wire and Curly	-	+	+
Golden Retriever 	Long	+	-	-
Bearded Collie 	Long with Furnishings	+	+	-
Irish Water Spaniel 	Curly	+	-	+
Bichon Frisé 	Curly with Furnishings	+	+	+

Figure 5. In the chart, phenotypes are listed in the left column and the genes associated with each phenotype are listed along the top. A minus sign (–) indicates the ancestral allele and a plus sign (+) indicates the more recent variant. (Graphic adapted from Cadieu et al., *Science*, 326(5949): 150–153,2009.)

1. What phenotype is the ancestral allele of the *KRT71* gene associated with in dogs?

_____

2. Which breed of dog shown above exhibits the phenotype for a variant allele for *FGF5* but an ancestral allele for *KRT71*? _____

Identifying which genes affect the type of fur a dog has is interesting and important for dog breeders, but are there benefits beyond breeding? Absolutely! These genes are not just associated with coat differences. It turns out that they produce proteins that regulate a variety of processes in all mammals, not just coat variations in dogs. For example, fibroblast growth factor 5 (*FGF5*) plays a role in embryonic development, cell growth, morphogenesis, tissue repair, and tumor growth. Understanding the functions of various genes in dogs is helping scientists better understand the mechanisms of many diseases that affect both dogs and humans. In addition, many traits in humans (e.g., height or predisposition to diabetes or cardiovascular disease) are controlled by more than one gene. By figuring out how to sort through the dog genome for the genes responsible for particular traits, we may be able to apply the same methods to human diseases.

REFERENCES

Cadieu et al., *Science* **326**: 150–153 (2009).
Karlsson et al., *Nature Genetics* 39:1321-1328 (2007).

AUTHORS

This activity was written by Melissa Csikari, Colonial Forge High School, Stafford, VA; Elinor Karlsson, Ph.D., Broad Institute of MIT, Cambridge, MA; Laura Bonetta, Ph.D., and Eriko Clements, Ph.D., HHMI.

They were edited by Laura Bonetta, Ph.D., HHMI and Robin Heyden, consultant; copyedited by Linda Felaco.

Reviewers: David McDonald, Ph.D., North Carolina Central University; Vince Buonaccorsi, Ph.D., Juniata College; Elinor Karlsson, Ph.D., Broad Institute.