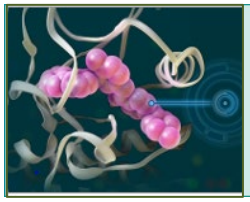


OVERVIEW

In this activity, students explore single nucleotide polymorphisms (SNPs) that are associated with different traits in dogs to help identify genes associated with those traits. The activity is based on actual genome-wide association studies (GWAS) with dogs. First, students learn about GWAS and SNPs, then answer questions about a news release that describes a real GWAS. Students then engage with a hands-on card activity to identify associations between certain phenotypes and SNPs in dogs. The activity includes an optional extension in which students use chi-square analysis to determine whether the associations are statistically significant.

This activity complements the 2013 Holiday Lecture "Dog Genomics and Dogs as Model Organisms," in which biologist Elinor Karlsson discusses how dogs can be used as model organisms for genomic studies, such as GWAS. The activity includes actual sequence data from DNA isolated from dog saliva, which was obtained and analyzed by Karlsson and her colleagues.

Additional information related to pedagogy and implementation can be found on this resource's webpage, including suggested audience, estimated time, and curriculum connections.

KEY CONCEPTS

- Comparing DNA sequences of many individuals reveals common variations across the genome.
- Some DNA variations occur more frequently in individuals with one form of a trait than another.
- Variations associated with a trait can point to the location of the gene (or genes) responsible for that trait.
- DNA sequences closer together on a chromosome tend to get inherited together and will often stay together over evolutionary time.

STUDENT LEARNING TARGETS

- Describe how a genome-wide association study (GWAS) works and what questions it can be used to investigate.
- Explain how GWAS uses single nucleotide polymorphisms (SNPs) to identify genes that affect a trait of interest.
- Identify and interpret patterns in real genomic data.

PRIOR KNOWLEDGE

Students should be familiar with:

- the concept of genes and mutations
- the relationship between genotype and phenotype
- the relationship between a single gene, a chromosome, and the genome
- the difference between coding and regulatory regions in the genome

MATERIALS

- copies of the "Student Handout"
- both sets of "SNP Cards" (curly-straight, short-long)
- (optional) access to the video lecture "Dog Genomics and Dogs as Model Organisms"

TEACHING TIPS

Running the Activity

- Students can work individually or in groups.
- Time estimates for each part of the "Student Handout" are as follows:
  - Part 1 may take about 30 to 40 minutes to complete, depending on the student's reading ability. Parts 2–5 may take about 20 minutes of in-class time, depending on the amount of collaboration. You may need additional time for answering questions and class discussion.
  - The "Extension" section may take another 20 minutes, depending on students' backgrounds.
- After **Part 1**, you may want to discuss the readings in class and address student questions or points of confusion. Suggested discussion questions include:
  - What is the relationship between a complete genome, a chromosome, and a gene?
  - How are SNPs used in a GWAS?
  - How can GWAS be used to find genes that cause disease?
  - What is the difference between a SNP and a mutation?
- For **Parts 3 and 4**, you will need to print complete sets of "SNP Cards" for each student or group. You could laminate the cards for repeat use.
  - Use the coat length cards ("short-long") for Part 3. These cards show SNP alleles at seven loci on chromosome 32 in 12 dogs with either short or long coats.
  - Use the coat texture cards ("straight-curly") for Part 4. These cards show SNP alleles at six loci on chromosome 27 in 10 dogs with either straight or curly coats.
- The "**Extension"** section is optional and is most useful for students with some background in chi-square analysis. It can be used to practice statistical methods.

Clarifications and Caveats

- The "Student Handout" uses the term **"allele"** to refer to variations of SNPs, which may also be called "genetic variants" or "SNP variants." Some scientists use the term "allele" to refer only to variations that cause changes in phenotypes, which is not always the case for variations in SNPs.
- The conceptual framework of GWAS relies on the concept of **genetic linkage**, which is typically taught during Mendelian genetics. This activity does not explore genetic linkage in depth.
  - If students are not familiar with genetic linkage, you may want to briefly summarize the following points so that they will understand the principles behind GWAS better:
    - Two DNA sequences on the same chromosome are said to be "linked" if they tend to be inherited together during meiosis.
    - If meiotic crossing-over occurs between the two sequences, they will not be inherited together.
    - The chance of crossing-over occurring between the two sequences depends on the physical distance between them. If they are closer together, the chance is smaller.
    - Therefore, two sequences on the same chromosome that are close together tend to be inherited together more often than those that are far apart.
    - Over evolutionary time, two sequences that are closer together tend to stay together.
  - You could take the opportunity to discuss genetic linkage when describing the difference between a SNP that causes a trait (causative SNP) and an associated SNP that does not cause a trait. Even if a SNP shows strong association with a trait in GWAS, it takes further investigation to see if it is a causative SNP (i.e., in the noncoding or coding region of the gene) or if it's simply closely linked to the causative part of the genome.
- Although this activity shows only a few SNPs on two chromosomes, a real GWAS uses millions of SNPs across the entire genome. A computer then sifts through the data to find associations. This activity uses simple

examples that can be done without a computer to demonstrate the concepts involved.

- Although the human genome was obtained by pooling sequence data from several individuals, the reference dog genome consists of data from just one dog (Tasha). Scientists have also sequenced dogs from other breeds and put all the differences they found compared to Tasha's genome in the public database dbSNP.

Supplements and Extensions

- The optional video lecture "Dog Genomics and Dogs as Model Organisms," which provides an introduction to GWAS in dogs, is useful for enhancing the activity.
  - Chapters 2–6 (time **1:56 to 12:15**) are the segments of the video most relevant to the activity. In particular, starting around 10:15, the video shows how to analyze SNP data similar to those in Parts 2–4 of the activity. Consider having students watch these segments as homework or in class.
  - In this video, Dr. Karlsson uses the term **"correlated"** to discuss the relationship found in the GWAS. This is a colloquial use of the term "correlated." In a strict statistical sense, correlations describe relationships between two quantitative variables. Here, because the variables are categories and not numbers, the term that should be used is **"associated"** (which is used throughout the "Student Handout"). You may need to clarify this statistical language with your students.
- The study featured in the news release in Part 1 of the activity is Cadieu et al. (2009). This is the same study described in Part 5 of the activity. If your students are comfortable reading and analyzing scientific papers, consider having them read the original paper to learn more about the study and its findings.
- Consider showing students how GWAS data is plotted; an example can be found in the "Integrative Genomics Viewer" from the Broad Institute. GWAS data is often shown as a "Manhattan plot" with SNPs represented as dots and color-coded by chromosome.
  - In a Manhattan plot, the *x*-axis shows the locations of the SNPs in the genome, and the *y*-axis is typically the negative logarithm of the *P* value for the association between the SNP and the trait. Stronger associations have smaller *P* values and thus larger negative logarithms of the *P* values. So, the higher the dot is on the *y*-axis, the more significant the SNP's association with the trait.
- Consider showing students scientific papers describing human genomics studies that use GWAS to identify genes that cause disease. You can also point out additional areas of research in which GWAS are used, such as in agriculture.
- This activity could serve as a lead-in to evolution since dogs are an excellent example of artificial selection.
- Students could write an analysis of the activity as a blog post (on a class blog or individual student blogs). Students could comment on each other's blogs to facilitate peer review and discussion.
- Students could create a "Dog Genotype and Phenotype" infographic, intended for a brochure or a handout in a veterinarian's office. This could be done using free online infographic creation tools, such as Easelly or Piktochart.

ANSWER KEY

PART 1: Introduction to GWAS

1. In general, why do you think GWAS is useful? What kinds of problems could GWAS be used to solve?
   Student answers will vary. Based on the reading, they will likely recognize that GWAS is useful for finding genes that affect traits of interest. This could be used to, for example, identify genetic risk factors for disease, design treatments for gene therapy, inform agricultural breeding programs, etc.
2. List the three combinations of alleles (C and A) that a dog could have for the SNP shown in Figure 2.
   CC, AA, CA (or AC)
3. Why do you think SNPs are referred to as "markers" or "signposts"?
   A SNP that is associated with a particular trait (meaning that it occurs more frequently in individuals with that trait) is often near a region of the genome that affects the trait. So, the location of the SNP acts as a

"marker" or "signpost" for a part of the genome (such as a gene or regulatory region) that affects the trait. Scientists can examine the region of the genome around the SNP in order to find such genes or regulatory regions.

4. Consider the different types of SNPs shown in Figure 3: associated, unassociated, and causative (including both noncoding and coding).

   a. Which types of SNPs affect protein production or function for the gene of interest?
   causative SNPs (noncoding and coding SNPs)

   b. Which types of SNPs might be identified in a GWAS?
   Any of the SNPs shown in Figure 3 could be identified in a GWAS, but only the associated SNPs and causative SNPs are likely to appear associated with the trait of interest.

5. How many genes account for the wide variety of coat types in dogs?
   three genes

6. In two or three sentences, describe how scientists identified these genes.
   The scientists did a GWAS that looked at SNPs in thousands of dogs from a variety of breeds. They compared their results to the coat types of the different dogs and found three genetic variants that accounted for the majority of coat types. The locations of the variants were used to find the three genes.

7. Why do you think it is important to analyze the DNA of many dogs when doing this research?
   It's important to analyze the DNA of many dogs to find variations that are generally common in dogs with a specific trait compared to dogs without that trait. If you look at the variations in only a few dogs, it would be hard to know whether those variations are generally associated with your trait of interest rather than other traits in those few dogs.

8. Humans have SNPs too. In general, how might GWAS studies with dogs benefit humans?
   Student answers will vary. They may say that the human genome shares similarities with the dog genome, so some of the genes we learn about in dogs could have similar effects in humans. They may also say that refining the GWAS approach in dogs can help us conduct better GWAS for humans. For example, if we can find the genes responsible for coat types in dogs, we can use a similar approach to identify genes involved in complex diseases in humans.

PART 2: Applying GWAS to Dog Fur Color

9. Give **two** possible reasons for why a SNP would be associated with a trait like fur color.
   The SNP could be in a gene (or regulatory region for the gene) that affects fur color, or it could be located near that gene on the chromosome.

10. Which SNP in Table 1 do you think is completely associated with fur color? Explain the reasoning for your choice.
    The correct answer is 7. It is the only SNP where all the dogs with black fur share the same alleles (TT) and all the dogs with white fur share different alleles (AA).

11. Which SNPs in Table 1 do you think are completely unassociated with fur color? Explain the reasoning for your choices. (*Hint*: There are five in total.)
    The correct answers are 1, 4, 9, 11, and 15. For all of these except 4, all the dogs in both groups have the same alleles. For 4, each group of dogs has the same total numbers of each allele (7 G and 1 C). (If students struggle with recognizing 4 as a completely unassociated SNP, consider revisiting this question after students learn how to calculate the strengths of associations in Part 3.)

12. Which SNP in Table 1 do you think has the *next strongest* association with fur color, after the completely associated SNP you identified in Question 10? Explain the reasoning for your choice.

The correct answer is 5. (Student answers may vary since they do not learn how to calculate the strengths of associations until Part 3. Encourage them to come up with an answer on their own, but don't penalize them for a wrong initial answer. This practice encourages students to reflect on their prior ideas and check their understanding, a useful metacognitive habit.)

PART 3: Identify Associations Using Real Data

13. Which SNP on the cards do you think has the *strongest* association with dog coat length? Explain the reasoning for your choice.
    Student answers may vary. Encourage them to come up with an answer on their own, but don't penalize them for a wrong initial answer. They will revisit their answer in Question 18, after they have learned more about how to calculate the strengths of associations.
14. Complete the SNP tables below using the method described above. The first two tables have been completed for you as examples.

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
T	***3***	***12***	***9***
G	***9***	***0***	***9***
		Total number of differences	***18***

chr32 7479580

Allele	Short Coat	Long Coat	Difference
T	***9***	***12***	***3***
C	***3***	***0***	***3***
		Total number of differences	***6***

chr32 7482867

Allele	Short Coat	Long Coat	Difference
A	***5***	***0***	***5***
G	***7***	***12***	***5***
		Total number of differences	***10***

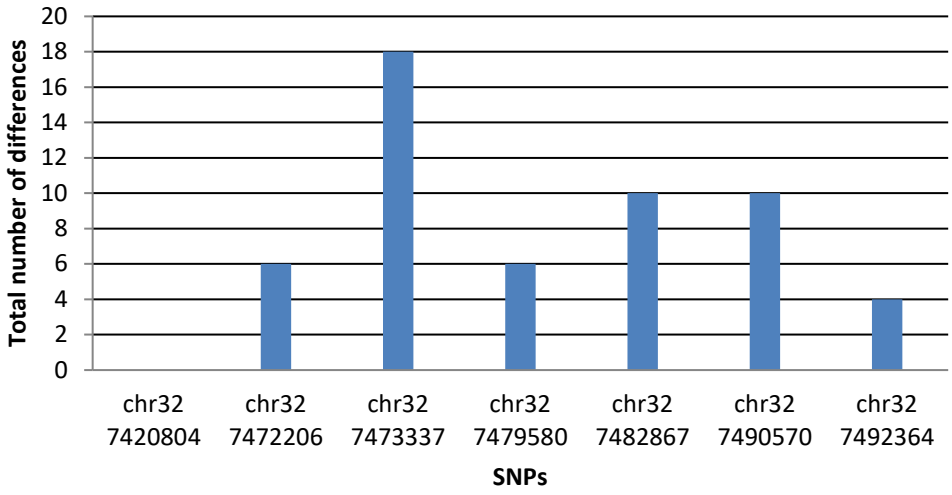
chr32 7490570

Allele	Short Coat	Long Coat	Difference
T	***6***	***11***	***5***
C	***6***	***1***	***5***
		Total number of differences	***10***

chr32 7492364

Allele	Short Coat	Long Coat	Difference
C	***3***	***1***	***2***
G	***9***	***11***	***2***
		Total number of differences	***4***

15. Using your results from the tables above, graph the **total number of differences** in alleles for each of the seven SNPs.



16. Which of these SNPs has the *strongest* association with coat length?
chr32 7473337

17. Which of these SNPs has the *weakest* association with coat length?
chr32 7420804

18. Based on these data, would you revise your answer to Question 13? Why or why not?
Student answers will vary depending on their answer to Question 13.

19. Based on these seven SNPs, where would you search if you wanted to find a gene involved in dog coat length?
The most promising place to search would be near the location of the SNP that has the strongest association with coat length. This would be on chromosome 32 near nucleotide 7473337.

20. Based on what you've learned in Part 3, would you revise your answer to Question 12 from Part 2? Why or why not?
Student answers will vary depending on their answer to Question 12.

PART 4: Identifying Associations for Dog Coat Texture

21. Fill out the following tables using the same method as before.

chr27 5525002

Allele	Curly Coat	Straight Coat	Difference
C	***4***	***5***	***1***
T	***6***	***5***	***1***
		Total number of differences	***2***

chr27 5541113

Allele	Curly Coat	Straight Coat	Difference
C	***7***	***4***	***3***
T	***3***	***6***	***3***
		Total number of differences	***6***

chr27 5542806

Allele	Curly Coat	Straight Coat	Difference
C	***4***	***10***	***6***
T	***6***	***0***	***6***
		Total number of differences	***12***

chr27 5545082

Allele	Curly Coat	Straight Coat	Difference
A	***4***	***8***	***4***
G	***6***	***2***	***4***
		Total number of differences	***8***

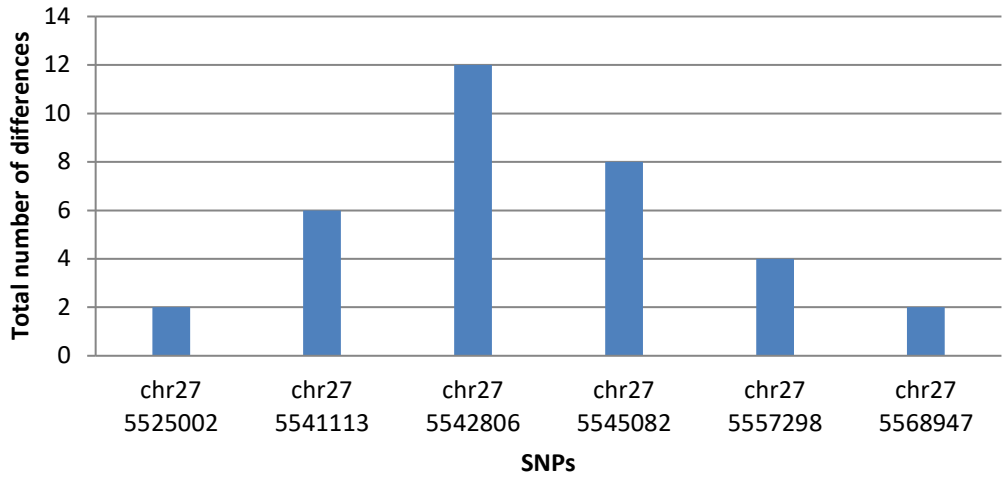
chr27 5557298

Allele	Curly Coat	Straight Coat	Difference
C	***4***	***2***	***2***
T	***6***	***8***	***2***
		Total number of differences	***4***

chr27 5568947

Allele	Curly Coat	Straight Coat	Difference
C	***5***	***6***	***1***
T	***5***	***4***	***1***
		Total number of differences	***2***

22. Using your results from the tables above, graph the **total number of differences** in alleles for each of the six loci.



23. Which of these SNPs has the *strongest* association with coat texture?
chr27 5542806

24. Based on these six SNPs, where would you search if you wanted to find a gene involved in dog coat texture?
The most promising place to search would be near the location of the SNP that has the strongest association with coat texture. This would be on chromosome 27 near nucleotide 5542806.

PART 5: Which Genes Determine Dog Coat Traits?

25. Which dog breed in Figure 5 has the ancestral allele for all three genes, similar to gray wolves?
Basset Hound

26. Which dog breeds in Figure 5 have a more recent allele for *FGF5* but an ancestral allele for *KRT71*?
Golden Retriever and Bearded Collie

27. Which coat type is the ancestral allele of the *KRT71* gene associated with?
The ancestral allele of* KRT71 *is associated with straight coats because the dogs with this ancestral (–) allele all have straight coats and the dogs with more recent (+) alleles all have curly coats.

28. How might understanding the functions of genes in dogs help us better understand human health?
Student answers will vary. They may say that certain genes in dogs, such as* FGF5*, play similar roles in humans. So, understanding the functions of these genes in dogs can help us better understand these genes, and the mechanisms of related diseases, in humans. In addition, students may recognize that the many traits in humans are controlled by more than one gene. By figuring out how multiple genes in the dog genome interact to affect certain traits (such as coat type), we may be able to use similar methods to understand the genes underlying complex traits that affect human health (e.g., predisposition to certain diseases).

29. The methods described in this activity can be used to study the genes of many different organisms. Pick an organism other than dogs or humans that you are interested in. Describe a specific problem or question that you could investigate by doing a GWAS with the organism you picked.
Student answers will vary; be open to a range of reasonable responses. Students should recognize that GWAS is used to identify genes associated with traits of interest, so their problems or questions will likely involve determining which genes affect a specific trait.

EXTENSION: Chi-Square Test of Independence Analysis

1. Calculate the expected numbers of each allele for the SNP that had the strongest association with coat length (your answer to Question 16 in Part 3). Show your work.
In Part 3, the SNP that had the strongest association with coat length was the one at chr32 7473337. This SNP had two alleles: T and G. The T allele occurred 3 times in dogs with a short coat and 12 times in dogs with a long coat, so:
total number of T alleles = 3 + 12 = 15
expected number of T alleles = 15/2 = 7.5

The G allele occurred 9 times in dogs with a short coat and 0 times in dogs with a long coat, so:
total number of G alleles = 9 + 0 = 9
expected number of G alleles = 9/2 = 4.5

2. Complete the table below for the SNP that had the strongest association with coat length. (You calculated the expected numbers of alleles for this SNP in Step 1 above.)

Allele	# Observed in Short Coats	# Observed in Long Coats	# Expected in Each Group
T	***3***	***12***	***7.5***
G	***9***	***0***	***4.5***

3. Calculate the chi-square value for this SNP. Show your work.

$$X^2 = \underset{\text{for T in short coats}}{\frac{(O-E)^2}{E}} + \underset{\text{for G in short coats}}{\frac{(O-E)^2}{E}} + \underset{\text{for T in long coats}}{\frac{(O-E)^2}{E}} + \underset{\text{for G in long coats}}{\frac{(O-E)^2}{E}}$$
$$= \frac{(3-7.5)^2}{7.5} + \frac{(9-4.5)^2}{4.5} + \frac{(12-7.5)^2}{7.5} + \frac{(0-4.5)^2}{4.5}$$
$$= \frac{(-4.5)^2}{7.5} + \frac{(4.5)^2}{4.5} + \frac{(4.5)^2}{7.5} + \frac{(-4.5)^2}{4.5}$$
$$= 2.7 + 4.5 + 2.7 + 4.5$$
$$= 14.4$$

4. What is the *P* value for the SNP that had the strongest association with coat length? (You calculated the chi-square value for this SNP in Step 2 above.)

The chi-square value for this SNP is 14.4 and the df is 1. Table 1 shows that this corresponds to a* P *value* less than *0.001.

5. What does this *P* value tell you?
   The* P *value is less than 0.05, so it is statistically significant. Thus, we can reject the null hypothesis that this SNP is not associated with coat length, meaning that it's likely this SNP* is *associated with coat length.
6. Which SNP had the strongest association with coat texture in Part 4?
   As determined in Part 4 (Question 23), the SNP that had the strongest association with coat texture was the one at chr27 5542806.
7. Is this association statistically significant? What is the evidence for your choice?
   Following the same methods as above, we can show that the chi-square value for this SNP is about 8.57. Again, the df is 1. This corresponds to a* P *value of between 0.005 and 0.002, which is less than 0.05 and thus statistically significant.

REFERENCES

Cadieu, E., M. W. Neff, P. Quignon, K. Walsh, K. Chase, H. G. Parker, B. M. VonHoldt, et al. "Coat variation in the domestic dog is governed by variants in three genes." *Science* 326, 5949 (2009): 150–153. https://doi.org/10.1126/science.1177808.

Karlsson, E. K., I. Baranowska, C. M. Wade, N. H. C. Salmon Hillbertz, M. C. Zody, N. Anderson, T. M. Biagi, et al. "Efficient mapping of mendelian traits in dogs through genome-wide association." *Nature Genetics* 39, 11 (2007): 1321–1328. https://doi.org/10.1038/ng.2007.10.

CREDITS

Written by Melissa Csikari, Laura Bonetta, Eriko Clements, HHMI; Elinor Karlsson, Broad Institute of MIT, MA

Edited by Satoshi Amagai, Esther Shyu, Laura Bonetta, HHMI; Sue Dodge, Robin Heyden, consultants

Scientific review by David McDonald, North Carolina Central University, NC; Vince Buonaccorsi, Juniata College, PA

Illustrations by Heather McDonald

- Figure 3 was adapted from *Learn Genetics* "Making SNPs Make Sense."
- Figure 5 was adapted from Figure 3 in Cadieu et al. (2009).