INVESTIGATION 2
MATHEMATICAL MODELING: HARDY-WEINBERG*

How can mathematical models be used to investigate the relationship between allele frequencies in populations of organisms and evolutionary change?

■ BACKGROUND

Evolution occurs in populations of organisms and involves variation in the population, heredity, and differential survival. One way to study evolution is to study how the frequency of alleles in a population changes from generation to generation. In other words, you can ask *What are the inheritance patterns of alleles, not just from two parental organisms, but also in a population?* You can then explore how allele frequencies change in populations and how these changes might predict what will happen to a population in the future.

Mathematical models and computer simulations are tools used to explore the complexity of biological systems that might otherwise be difficult or impossible to study. Several models can be applied to questions about evolution. In this investigation, you will build a spreadsheet that models how a hypothetical gene pool changes from one generation to the next. This model will let you explore parameters that affect allele frequencies, such as selection, mutation, and migration.

The second part of the investigation asks you to generate your own questions regarding the evolution of allele frequencies in a population. Then you are asked to explore possible answers to those questions by applying more sophisticated computer models. These models are available for free.

This investigation also provides an opportunity for you to review concepts you might have studied previously, including natural selection as the major mechanism of evolution; the relationship among genotype, phenotype, and natural selection; and fundamentals of classic Mendelian genetics.

---

* Transitioned from the *AP Biology Lab Manual* (2001)

Learning Objectives

- To use a data set that reflects a change in the genetic makeup of a population over time and to apply mathematical methods and conceptual understandings to investigate the cause(s) and effect(s) of this change

- To apply mathematical methods to data from a real or simulated population to predict what will happen to the population in the future

- To evaluate data-based evidence that describes evolutionary changes in the genetic makeup of a population over time

- To use data from mathematical models based on the Hardy-Weinberg equilibrium to analyze genetic drift and the effect of selection in the evolution of specific populations

- To justify data from mathematical models based on the Hardy-Weinberg equilibrium to analyze genetic drift and the effects of selection in the evolution of specific populations

- To describe a model that represents evolution within a population

- To evaluate data sets that illustrate evolution as an ongoing process

General Safety Precautions

There are some important things to remember when computer modeling in the classroom. To avoid frustration, periodically save your work. When developing and working out models, save each new version of the model with a different file name. That way, if a particular strategy doesn't work, you will not necessarily have to start over completely but can bring up a file that had the beginnings of a working model.

If you have difficulty refining your spreadsheet, consider using the spreadsheet to generate the random samples and using pencil and paper to archive and graph the results.

As you work through building this spreadsheet you may encounter spreadsheet tools and functions that are not familiar to you. Today, there are many Web-based tutorials, some text based and some video, to help you learn these skills. For instance, typing "How to use the SUM tool in Excel video" will bring up several videos that will walk you through using the SUM tool.

THE INVESTIGATIONS

Getting Started

This particular investigation provides a lab environment, guidance, and a problem designed to help you understand and develop the skill of modeling biological phenomena with computers. There are dozens of computer models already built and available for free. The idea for this laboratory is for you to build your own from scratch. To obtain the maximum benefit from this exercise, you should not do too much background preparation. As you build your model and explore it, you should develop a more thorough understanding of how genes behave in population.

To help you begin, you might want to work with physical models of population genetics, such as simulations that your teacher can share with you. With these pencil-and-paper simulations, you can obtain some insights that may help you develop your computer model.

Procedure

It is easy to understand how microscopes opened up an entire new world of biological understanding. For some, it is not as easy to see the value of mathematics to the study of biology, but, like the microscope, math and computers provide tools to explore the complexity of biology and biological systems — providing deeper insights and understanding of what makes living systems work.

To explore how allele frequencies change in populations of organisms, you will first build a computer spreadsheet that models the changes in a hypothetical gene pool from one generation to the next. You need a basic familiarity with spreadsheet operations to complete this lab successfully. You may have taken a course that introduced you to spreadsheets before. If so, that will be helpful, and you may want to try to design and build your model on your own after establishing some guidelines and assumptions. Otherwise, you may need more specific guidance from your teacher. You can use almost any spreadsheet program available, including free online spreadsheet software such as Google Docs or Zoho (*http://www.zoho.com*), to complete the first section of your investigation.

In the second part of the investigation, you will use more sophisticated spreadsheet models or computer models to explore various aspects of evolution and alleles in populations. To understand how these complex tools work and their limitations, you first need to build a model of your own.

Building a Simple Mathematical Model

The real world is infinitely complicated. To penetrate that complexity using model building, you must learn to make reasonable, simplifying assumptions about complex processes. For example, climate change models or weather forecasting models are simplifications of very complex processes — more than can be accounted for with even the most powerful computer. These models allow us to make predictions and test hypotheses about climate change and weather.

By definition, any model is a simplification of the real world. For that reason, you need to constantly evaluate the assumptions you make as you build a model, as well as evaluate the results of the model with a critical eye. This is actually one of the powerful benefits of a model — it forces you to think deeply about an idea.

There are many approaches to model building; in their book on mathematical modeling in biology, Otto and Day (2007) suggest the following steps:

1. Formulate the question.

2. Determine the basic ingredients.

3. Qualitatively describe the biological system.

4. Quantitatively describe the biological system.

5. Analyze the equations.

6. Perform checks and balances.

7. Relate the results back to the question.

As you work through the next section, record your thoughts, assumptions, and strategies on modeling in your laboratory notebook.

Step 1 Formulate the question.

Think about a recessive Mendelian trait such as cystic fibrosis. Why do recessive alleles like cystic fibrosis stay in the human population? Why don't they gradually disappear?

Now think about a dominant Mendelian trait such as polydactyly (more than five fingers on a single hand or toes on a foot) in humans. Polydactyly is a dominant trait, but it is not a *common* trait in most human populations. Why not?

How do inheritance patterns or allele frequencies change in a population? Our investigation begins with an exploration of answers to these simple questions.

Step 2 **Determine the basic ingredients.**

Let's try to simplify the question *How do inheritance patterns or allele frequencies change in a population?* with some basic assumptions. For this model, assume that all the organisms in our hypothetical population are diploid. This organism has a gene locus with two alleles — *A* and *B*. (We could use *A* and *a* to represent the alleles, but *A* and *B* are easier to work with in the spreadsheet you'll be developing.) So far, this imaginary population is much like any sexually reproducing population.

How else can you simplify the question? Consider that the population has an infinite gene pool (all the alleles in the population at this particular locus). Gametes for the next generation are selected totally at random. What does that mean? Focus on answering that question in your lab notebook for a moment — it is key to our model. For now let's consider that our model is going to look only at how allele frequencies might change from generation to generation. To do that we need to describe the system.

Step 3 **Qualitatively describe the biological system.**

Imagine for a minute the life cycle of our hypothetical organism. See if you can draw a diagram of the cycle; be sure to include the life stages of the organism. Your life cycle might look like Figure 1.
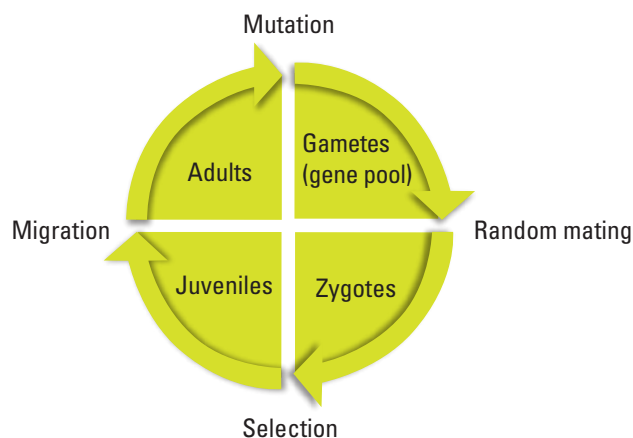

Figure 1. Life Stages of a Population of Organisms

To make this initial exploration into a model of inheritance patterns in a population, you need to make some important assumptions — all the gametes go into one infinite pool, and all have an equal chance of taking part in fertilization or formation of a zygote. For now, all zygotes live to be juveniles, all juveniles live to be adults, and no individuals enter or leave the population; there is also no mutation. Make sure to record these assumptions in your notebook; later, you will need to explore how your model responds as you change or modify these assumptions.

Step 4 Quantitatively describe the biological system.

Spreadsheets are valuable tools that allow us to ask *What if?* questions. They can repeatedly make a calculation based on the results of another calculation. They can also model the randomness of everyday events. Our goal is to model how allele frequencies change through one life cycle of this imaginary population in the spreadsheet. Use the diagram in Figure 1 as a guide to help you design the sequence and nature of your spreadsheet calculation.

Each part of the life cycle can be represented by a spreadsheet operation.

1. Set allele frequencies (assign a value to a cell).

2. Use the random function (RAND) to generate a random number which will be compared to the allele frequency from Step 1.

3. Compare the random number to allele frequency and assign the appropriate allele.

4. Repeat Steps 1–3 for the second allele.

5. Use the CONCATENATE function to combine the two alleles to form a zygote.

6. Copy this procedure (Steps 2–5) for multiple offspring.

Let's get started. The first step is to randomly draw gametes from the gene pool to form a number of zygotes that will make up the next generation.

To begin this model, let's define a couple of variables.

Let

$$p = \text{the frequency of the } A \text{ allele}$$
$$\text{and let } q = \text{the frequency of the } B \text{ allele}$$

Bring up the spreadsheet on your computer. The examples here are based on Microsoft® Excel, but almost any modern spreadsheet can work, including Google's online Google Docs (*https://docs.google.com*) and Zoho's online spreadsheet (*http://www.zoho.com*).

Hint: If you are familiar with spreadsheets, the RAND function, and using IF statements to create formulas in spreadsheets, you may want to skip ahead and try to build a model on your own. If these are not familiar to you, proceed with the following tutorial.

Somewhere in the upper left corner (in this case, cell D2), enter a value for the frequency of the *A* allele. This value should be between 0 and 1. Go ahead and type in labels in your other cells and, if you wish, shade the cells as well. This blue area will represent the gene pool for your model. (Highlight the area you wish to format with color, and right-click with your mouse in Excel to format.) This is a spreadsheet, so you can enter the value for the frequency of the *B* allele; however, when making a model it is best to have the spreadsheet do as many of the calculations as possible. All of the alleles in the gene pool are either *A* or *B*; therefore $p + q = 1$ and $1 - p = q$. In cell D3, enter the formula to calculate the value of q.

In spreadsheet lingo it is

=1-D2

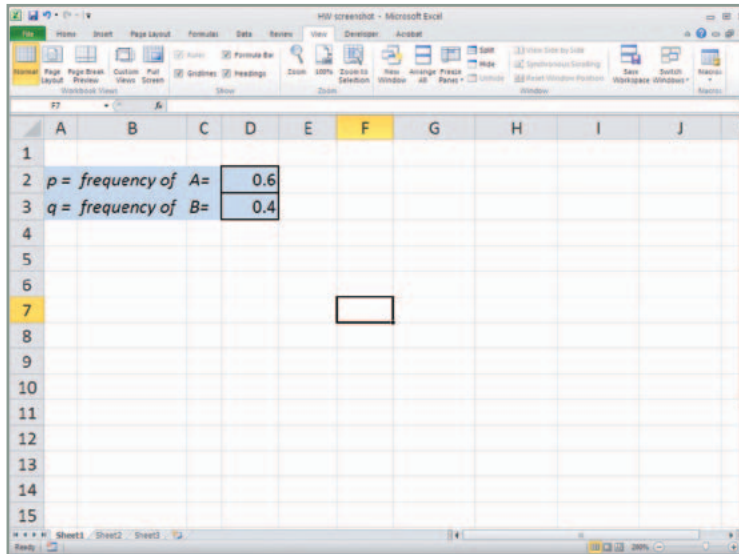Your spreadsheet now should look something like Figure 2.

Figure 2

Let's explore how one important spreadsheet function works before we incorporate it into our model. In a nearby empty cell, enter the function (we will remove it later).

=Rand()

Note that the parentheses have nothing between them. After hitting *return*, what do you find in the cell? If you are on a PC, try hitting the F9 key several times to force recalculation. On a Mac, enter *cmd +* or *cmd =*. What happens to the value in the cell? Describe your results in your lab notebook.

The RAND function returns random numbers between 0 and 1 in decimal format. This is a powerful feature of spreadsheets. It allows us to enter a sense of randomness to our calculations if it is appropriate — and here it is when we are "randomly" choosing gametes from a gene pool. Go ahead and delete the RAND function in the cell.

Let's select two gametes from the gene pool. In cell E5, let's generate a random number, compare it to the value of *p*, and then place either an *A* gamete or a *B* gamete in the cell. We'll need two functions to do this, the RAND function and the IF function. Check the help menu if necessary.

Note that the function entered in cell E5 is

$$=IF(RAND()<=D\$2,\text{“A”, “B”})$$

Be sure to include the $ in front of the *2* in the cell address D2. It will save time later when you build onto this spreadsheet.

The formula in this cell basically says that if a random number between 0 and 1 is less than or equal to the value of *p*, then put an *A* gamete in this cell, or if it is not less than or equal to the value of *p*, put a *B* gamete in this cell. IF functions and RAND functions are very powerful tools when you try to build models for biology.

Now create the same formula in cell F5, making sure that it is formatted exactly like E5. When you have this completed, press the recalculate key to force a recalculation of your spreadsheet. If you have entered the functions correctly in the two cells, you should see changing values in the two cells. (This is part of the testing and retesting that you have to do while model building.) Your spreadsheet should look like Figure 3.

Try recalculating 10–20 times. Are your results consistent with what you expect? Do both cells (E5 and F5) change to *A* or *B* in the ratios you'd expect from your *p* value? Try changing your *p* value to 0.8 or 0.9. Does the spreadsheet still work as expected? Try lower *p* values. If you don't get approximately the expected numbers, check and recheck your formulas now, while it is early in the process.
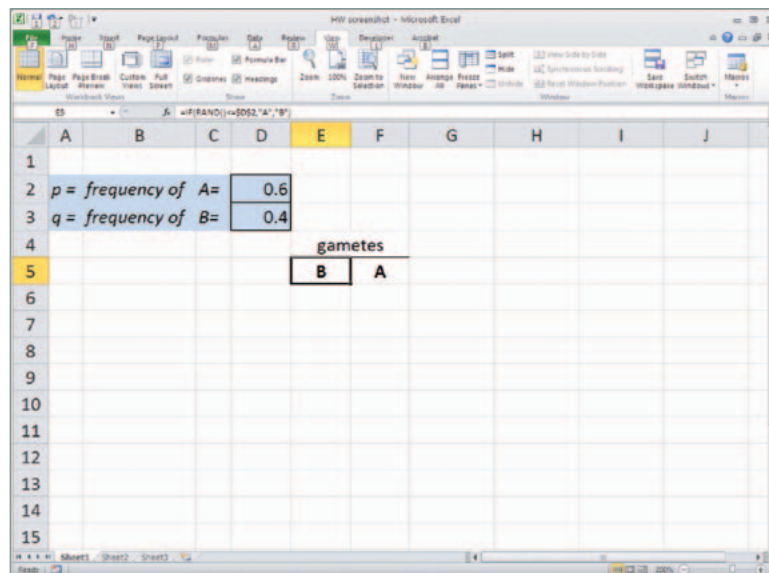

Figure 3

You could stop here and just have the computer recalculate over and over — similar to tossing a coin. However, with just a few more steps, you can have a model that will create a small number or large number of gametes for the next generation, count the different genotypes of the zygotes, and graph the results.

Copy these two formulas in E5 and F5 down for about 16 rows to represent gametes that will form 16 offspring for the next generation, as in Figure 4. (To copy the formulas, click on the bottom right-hand corner of the cell and, with your finger pressed down on the mouse, drag the cell downward.)
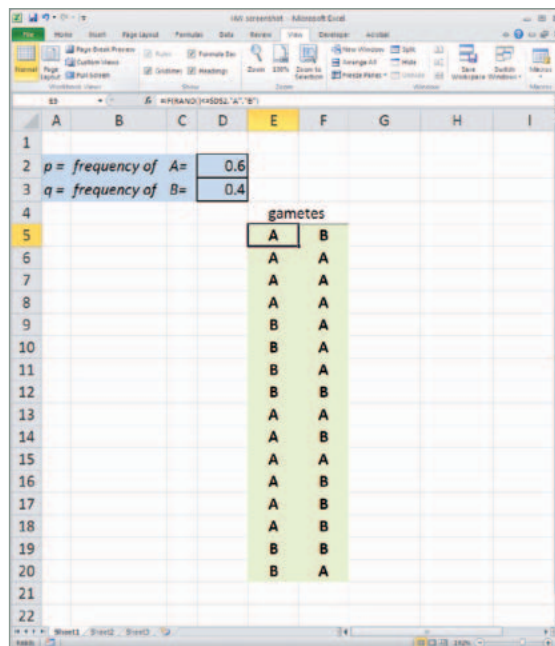
Figure 4

We'll put the zygotes in cell G5. The zygote is a combination of the two randomly selected gametes. In spreadsheet vernacular, you want to concatenate the values in the two cells. In cell G5 enter the function

$$=CONCATENATE(E5,F5)$$

Copy this formula down as far down as you have gametes, as in Figure 5 on the next page.
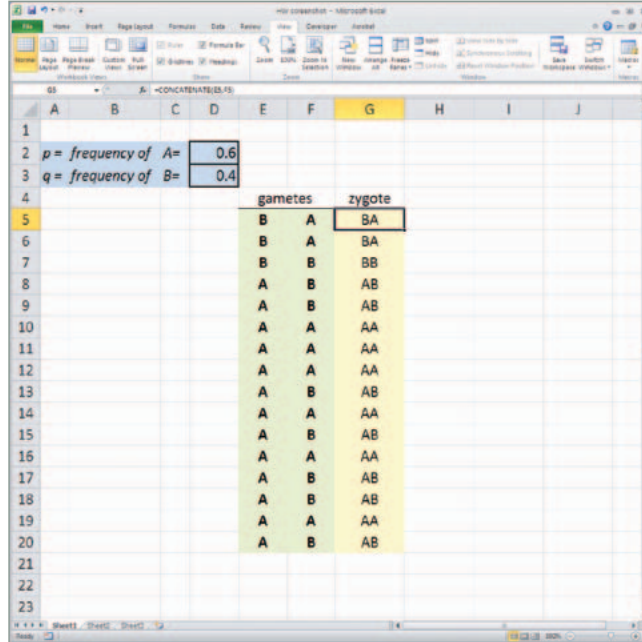
Figure 5

The next columns on the sheet, H, I, and J, are used for bookkeeping — that is, keeping track of the numbers of each zygote's genotype. They are rather complex functions that use IF functions to help us count the different genotypes of the zygotes.
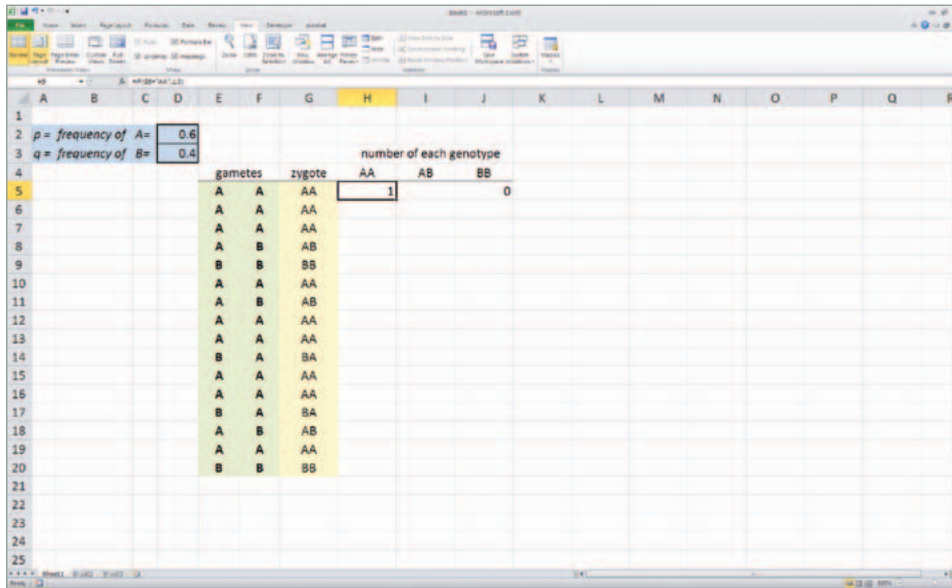
The function in cell H5 is

$$=IF(G5="AA",1,0)$$

This basically means that if the value in cell G5 is AA, then put a 1 in this cell; if not, then put a 0.

Enter the following very similar function in cell J5: $=IF(G5="BB",1,0)$

- Can you interpret this formula?
- What does it say in English?

Your spreadsheet now should resemble Figure 6.

Figure 6

Now let's tackle the nested IF function. This is needed to test for either *AB* or *BA*.

In cell I5, enter the nested function:

=IF(G5="AB",1,(IF(G5="BA",1,0)))

This example requires an extra set of parentheses, which is necessary to nest functions. This function basically says that if the value in cell G5 is exactly equal to *AB*, then put a 1; if not, then if the value in cell G5 is exactly *BA*, put a 1; if it is neither, then put a 0 in this cell. Copy these three formulas down for all the rows in which you have produced gametes.

Enter the labels for the columns you've been working on — *gametes* in cell E4, *zygote* in cell G5, *AA* in cell H4, *AB* in cell I4, and *BB* in cell J4, as shown in Figure 7 on the next page.
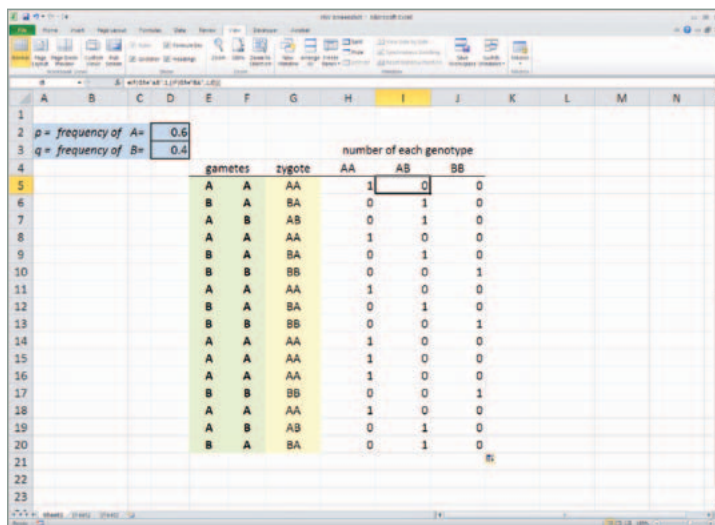
Figure 7

As before, try recalculating a number of times to make sure everything is working as expected. What is expected? If you aren't sure yet, keep this question in mind as you complete the sheet. You could use a p value of 0.5, and then you'd see numbers similar to the ratios you would get from flipping two coins at once. Don't go on until you are sure the spreadsheet is making correct calculations. Try out different values for p. Make sure that the number of zygotes adds up. Describe your thinking and procedure for checking the spreadsheet in your lab notebook.

Now, copy the cells E5 through J5 down for as many zygotes as you'd like in the first generation. Use the SUM function to calculate the numbers of each genotype in the H, I, and J columns. Use the genotype frequencies to calculate new allele frequencies and to recalculate new p and q values. Make a bar graph of the genotypes using the chart tool. Your spreadsheet should resemble Figure 8.
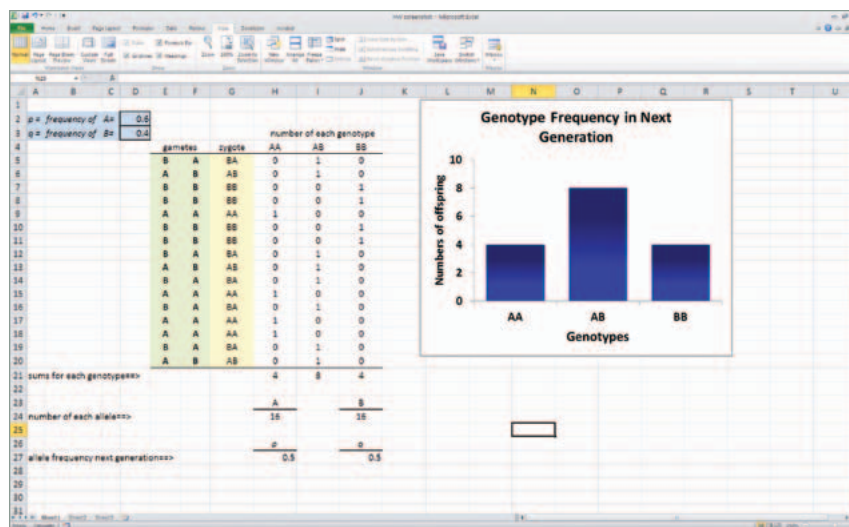

Figure 8

Testing Your Mathematical Model

You now have a model with which you can explore how allele frequencies behave and change from generation to generation. Working with a partner, develop a plan to answer this general question: *How do inheritance patterns or allele frequencies change in a population over one generation?* As you work, think about the following more specific questions:

- What can you change in your model? If you change something, what does the change tell you about how alleles behave?

- Do alleles behave the same way if you make a particular variable more extreme? Less extreme?

- Do alleles behave the same way no matter what the population size is? To answer this question, you can insert rows of data somewhere between the first row of data and the last row and then copy the formulas down to fill in the space.

Try out different starting allele frequencies in the model. Look for and describe the patterns that you find as you try out different allele frequencies. Develop and use a pattern to select your values to test and organize your exploration. In particular, test your model with extreme values and intermediate values. In your lab notebook, describe your observations and conclusions about the population inheritance patterns you discover.

Try adding additional generations to your model to look at how allele frequencies change in multiple generations. To do this, use your newly recalculated p and q values to seed the next generation. Once you've included the second generation, you should be able to copy additional generations so that your model looks something like Figure 9, with each new generation determining the new p and q values for the next.
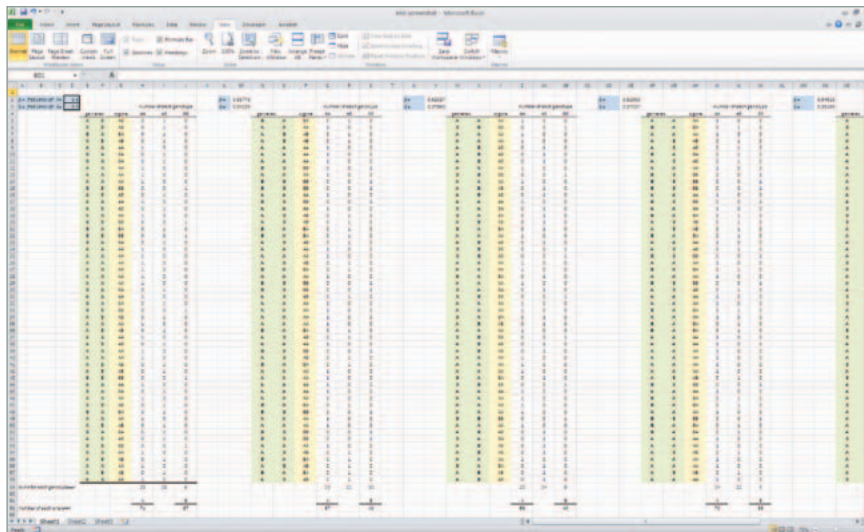

Figure 9

Try to create a graph of p values over several generations, for different-sized populations. See if you can detect a pattern of how population size affects the inheritance pattern. Be sure to try out both large and small populations of offspring.

This model relies on the RAND function to randomly select gametes from an infinite gene pool.

- What would happen if there were no randomness to this selection?
- What kind of pattern of genotypes would you expect in the next generation?

Creating a Formula that Predicts the Genotypes of the Next Generation

Here are two approaches to develop the formula. You might first try a graphical approach. Create a Punnet square, like Figure 10 and similar to what you might use to solve a Mendelian genetics problem. In this case, however, plot the values of p and q. Scale each side of the square based on the magnitude of the p or q values. Place this diagram in your lab notebook, and fill in the squares with variables and values, as in Figure 10.

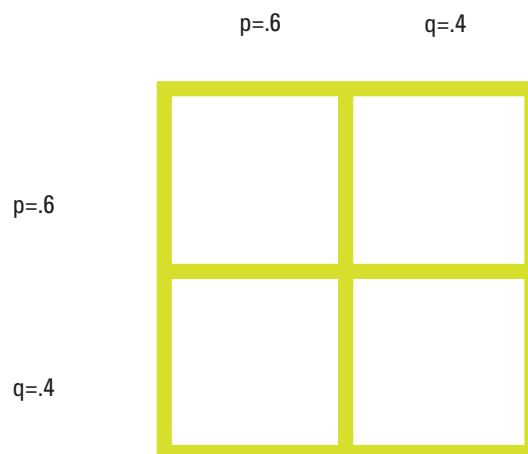p=.6   q=.4

p=.6

q=.4

Figure 10

Of course, you could also calculate the expected results for the next generation.

Remember that $p + q = 1$

The probability of two A alleles combining in one organism in the next generation is p^2. The probability of two B alleles combining is q^2. The probability of a combination of AB is $p * q$, as is the probability of combination of BA alleles, for a total of $2pq$.

For the next generation, the formula that predicts genotypes is

$$(p + q)^2 = 1, \text{ which works out to: } p^2 + 2pq + q^2$$

Based on the calculations you made while testing your model, how would you answer the following questions?

- In the absence of random events (an infinitely large population), are the allele frequencies of the original population expected to change from generation to generation?

- How does this compare to a population that has random gamete selection but is small?

- What happens to allele frequencies in such a population? Is it predictable?

This mathematical model can predict allele frequencies from generation to generation. In fact, it is a *null* model. That is, in the absence of random events or other real-life factors that affect populations, the allele frequencies do not change from generation to generation. This is known as the Hardy-Weinberg equilibrium (H-W equilibrium). The H-W equilibrium is a valuable tool for population biologists because it serves as a baseline to measure changes in allele frequencies in a population. If a population is not in H-W equilibrium, then something else is happening that is making the allele frequencies change.

What factors can cause allele frequencies to change in a population? (Hint: There are many.) How could you model these factors using your spreadsheet?

◼ Designing and Conducting Your Investigation

By this point you've been able to use your model to explore how random chance affects the inheritance patterns of alleles in large and small populations. Perhaps you've also been able to find some interesting patterns in how alleles behave across generations.

At the end of the last section you were asked what factors can cause allele frequencies to change in a population and how you would model them. Choose one of your answers, and try it out using your spreadsheet. This may involve adding multiple columns or rows along with a few extra operations. Keep the life cycle of your hypothetical population in mind as you develop additional strategies.

With your new spreadsheet model, generate your own questions regarding the evolution of allele frequencies in a population. From these questions (noted in your lab notebook), you need to develop hypotheses that you can test — those that allow you to easily manipulate the parameters of population size, number of generations, selection (fitness), mutation, migration, and genetic drift. Collect sufficient data by running your model repeatedly. Analyze your data. Formulate your conclusions and present a miniposter that supports your claim with sound reasoning and evidence to the class. Your teacher may have some ideas for questions to investigate.

■ Where Can You Go from Here?

An excellent extension to this laboratory is the following investigation:

McMahon, K. A. 2008. Supertasters—Updating the Taste Test for the A & P Laboratory. Pages 398–405, in Tested Studies for Laboratory Teaching, Volume 29 (K.L. Clase, Editor). Proceedings of the 29th Workshop/Conference of the Association for Biology Laboratory Education (ABLE).

Your teacher will provide the lab, or you can google "ABLE proceedings + supertaster" to access the lab.

There are few human traits that express the intermediate dominance necessary for testing for the null hypothesis. The supertaster trait described in this laboratory does express an intermediate phenotype; therefore, it creates an exemplary investigative population genetics laboratory.

■ REFERENCE

Otto, S. P. and T. Day (2007). A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press.
http://www.zoology.ubc.ca/biomath/