Updated 03/26/2024

TEACHING THE GENOME GENERATION

Sequence Comparison in Ancestry Testing (spreadsheet-based)

The Jackson Laboratory
Leading the search for tomorrow's cures

NIH National Institutes of Health
Turning Discovery Into Health
SEPA SCIENCE EDUCATION PARTNERSHIP AWARD
SUPPORTED BY THE NATIONAL INSTITUTES OF HEALTH

Sam's Story: How does ancestry testing work?

Sam is a high school student who wants a DNA ancestry test for their 18th birthday. At school, Sam approaches their biology teacher Dr. V to share their excitement about taking this at-home genetic test. But Sam is surprised to learn the test doesn't work the way they thought it would.

Read the scene below from Sam's life. If you're in a class or group, assign one person as Sam, and one person as Dr. V, and read the scene out loud.

Sam: *Dr. V! I asked for a DNA ancestry test for my 18th birthday! I am so excited to get all of my DNA sequenced!*

Dr. V: *That's not exactly how these tests work, Sam. Only a portion of our genomes are examined in ancestry tests.*

Sam: *What?! What are just a few DNA sites going to tell you?!*

Dr. V: *Let's do an exercise together to learn how these tests really work.*

Prediction

From their discussion in class, Sam might not fully understand how ancestry tests work. How do you think an ancestry test works? Create a flowchart outlining the steps involved in an ancestry test. Think about sample collection, what information is analyzed by the company, and what results are shared.

Background

Dr. V wants to teach the class how ancestry tests work. Before getting into the details of the test, first she prepares the class with some important background information. Dr. V will review concepts from Biology class on genetics and inheritance, and also explore new topics on alleles and sequencing.

Bolded words are defined in the Glossary at the end of this lesson.

Genetics & Inheritance

DNA is a biological molecule comprised of a four-letter code of the **nucleotides** adenine (A), thymine (T), cytosine (C) and guanine (G). The human **genome** is the complete set of DNA in an individual that encodes the instructions for making each of us who we are.

Human genomes are 99.9% identical between individuals. Only 0.1% of our DNA is different between individuals! Variation in this 0.1% of DNA are the **genotypic** differences that leads to the **phenotypic** differences we see around us. When only one DNA nucleotide varies, this is called a **single nucleotide polymorphism (SNP)** or **single nucleotide variant (SNV)**.

Humans have **diploid** genomes, which means they have two copies of every piece of DNA. One copy is passed from each **gamete:** one copy from the **haploid** sperm cell, and one copy is passed from the haploid egg cell. Thus, an individual inherits two copies, one from each biological parent, for any given genomic location.

Let's look at an example in Figure 1. Here, the blue DNA on the left represents the sperm haploid genome, and the yellow DNA on the right represents the egg haploid genome. These cells come together to make an offspring, shown here as an individual with a diploid genome inherited from both the sperm and the egg.

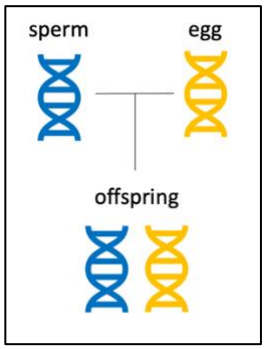


Figure 1. Haploid sperm and eggs cells come together to make a diploid offspring.

At any location in the genome, the genotype of the offspring will depend on the DNA they inherited from both biological parents. An individual inherits two copies of DNA, one from the sperm and one from the egg. Each copy, or version of DNA, is called an **allele**. At any DNA **locus**, if the two inherited alleles are the same, the individual is **homozygous** at that locus. If the alleles are different, the individual is **heterozygous**.

Let's examine one specific DNA locus in the genome that is known to contain a SNP in Figure 2. Here at the circled locus, we observe the sperm cell carries a G allele, and the egg cell carries an A allele. Since the offspring inherits one allele from each parent, the offspring's genotype is heterozygous at this locus.

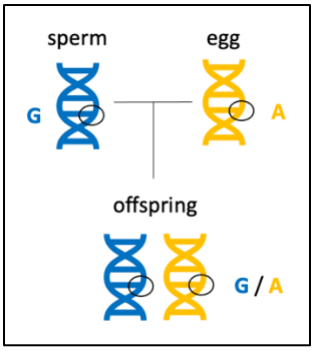


Figure 2. An offspring's genotype at one locus is determined by what alleles they inherit.

Each site of our genomes can be examined in the same way to observe which individuals carry which DNA variants. Looking at another SNP at a different DNA locus in Figure 3, we observe that here the sperm cell carries an A allele, and the egg cell carries a T allele, resulting in an offspring that is heterozygous at this locus.

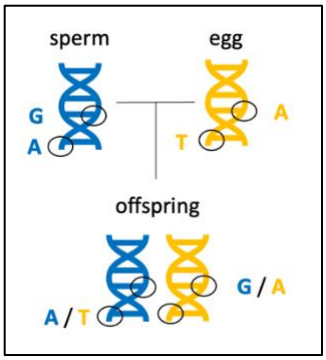


Figure 3. Different regions of the genome can carry different alleles, or variants, across individuals.

Allele Frequencies

Counting alleles in individuals reveals **allele frequencies**. This is a measure of the distribution of all DNA variations (alleles) in a **population**. It can be calculated very simply once you know the number of each allele observed in your population, as well as the total number of individuals in your population. The formula is:

$$Allele\ Frequency = \frac{v}{2N}$$

Where the **variable** v is the number of each variant allele detected in the dataset, and the variable N is the total number of diploid individuals in the dataset. Written another way without variables, the formula for allele frequency is:

$$Allele\ Frequency = \frac{Allele\ count\ of\ a\ specific\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population}$$

This allele frequency number can either be reported as a fraction or as a percent. To report as a fraction, the calculation ends with the formulas as shown above. To report as a percent, the number is multiplied by 100:

$$Allele\ Frequency\ (percent) = \frac{v}{2N} * 100$$

$$Allele\ Frequency\ (percent) = \frac{Allele\ count\ of\ a\ specific\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} * 100$$

Let's work through an example of this by returning to our example. In Figure 4, the genomes of the two parents of the offspring are now revealed, and the genotypes at the first DNA site are shown for all three individuals.

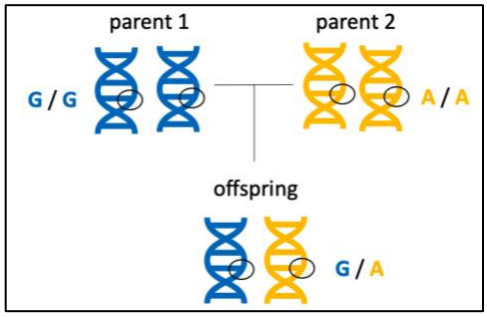


Figure 4. The genomes of the parents are revealed and the genotypes of all 3 individuals are known at this DNA site.

We can count how many of each variant allele we detect in this dataset and calculate allele frequencies for each allele at DNA site 1. Recall that at every DNA locus, there are four possible variant alleles for the four DNA nucleotides: either an A, T, C, or G will be present. In this example, there are 3 individuals total which is the population of this dataset. There are no C or T alleles. A total of 3 G alleles and 3 A alleles are present. This can be recorded in a table, like in Figure 5.

	Allele count at DNA SNP site 1
C	0
T	0
A	3
G	3

Dataset: N = 3 individuals

Figure 5. Each variant allele from Figure 5 at DNA SNP site 1 is counted and entered into the table.

Now the allele frequency fraction for each variant allele can be calculated as follows:

$$C\ Allele\ Frequency\ at\ SNP\ 1 = \frac{Allele\ count\ of\ C\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} = \frac{0}{6} = 0$$

$$T\ Allele\ Frequency\ at\ SNP\ 1 = \frac{Allele\ count\ of\ T\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} = \frac{0}{6} = 0$$

$$A\ Allele\ Frequency\ at\ SNP\ 1 = \frac{Allele\ count\ of\ A\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} = \frac{3}{6} = 0.5$$

$$G\ Allele\ Frequency\ at\ SNP\ 1 = \frac{Allele\ count\ of\ G\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} = \frac{3}{6} = 0.5$$

Remember that these allele frequencies can also be reported as percentages. The allele frequency percent for each variant allele would be:

$$C\ Allele\ Frequency\ at\ SNP\ 1 = \ 0 * 100 = 0\ \%$$

$$T\ Allele\ Frequency\ at\ SNP\ 1 = \ 0 * 100 = 0\ \%$$

$$A\ Allele\ Frequency\ at\ SNP\ 1 = \ 0.5 * 100 = 50\ \%$$

$$G\ Allele\ Frequency\ at\ SNP\ 1 = \ 0.5 * 100 = 50\ \%$$

Recall that allele frequencies represent the proportion of each DNA variant allele observed in the population. But, what is that population? In our example, we know our dataset is only 3 individuals, and they are biologically related which means they share DNA. Allele frequency data can change based on the population dataset size and the diversity of individuals.

For example, let's examine the same DNA SNP site 1 in two datasets with a different number of individuals in Figure 6. Our example with 3 individuals gives very different allele frequency numbers than looking at 50 individuals in a larger, more diverse dataset. Graphing this data as a pie chart gives us another representation of the same data, which shows clearly the drastic different in frequencies of each allele with a bigger and more diverse dataset.

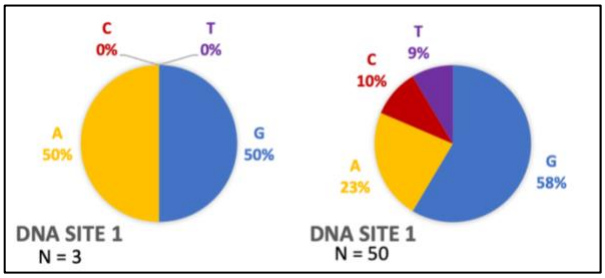


Figure 6. The allele frequencies at any DNA site can change based on the population examined.

DNA Sequencing

Ancestry testing companies typically perform this type of SNP genotyping at DNA sites with known and high variation. Interestingly, some medical tests also use similar technology to detect known variants. Whole genome sequencing is not typically performed in ancestry tests.

Spreadsheets and Tables

Spreadsheets and tables are used to display information from different analyses. Each unit of a spreadsheet or table is a **cell**, and information such as numbers, dates, or words can be entered. When information from multiple cells is arranged **vertically**, it forms a **column**. When information from multiple cells is arranged **horizontally**, it forms a **row**. **Headers** are cells with descriptions that provide additional information about what is in a row or in a column.

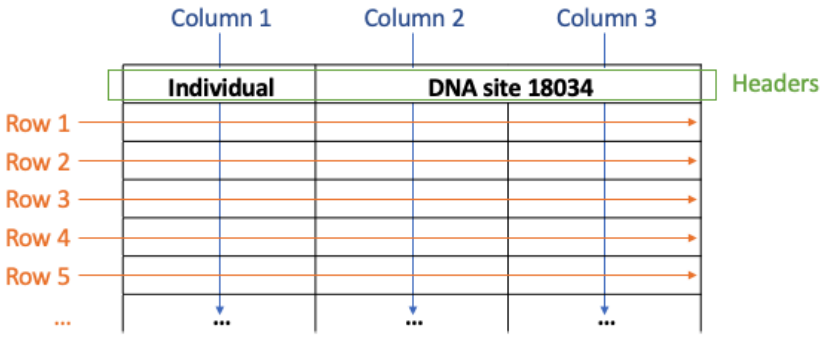


Figure 7. A table or spreadsheet has many cells of information. Columns, rows, and headers are highlighted.

We already saw an example of a table in Table 1. Each DNA nucleotide is a row with one piece of information that corresponds to it: an allele count number. The allele count number is a column, and has a descriptive header. Go back to review that table to notice the features: it has 2 columns of information, one with a header, and there are 4 rows of data.

Quick Knowledge Check

Check your understanding of the background material by answering these questions.

1. What statement best describes alleles?
    a. Haploid gamete cells with a single set of chromosomes.
    b. Two alternative forms of a gene that arose from DNA variations.
    c. Diploid genomes with two sets of chromosomes.
    d. Individuals with various genotypes in a population.

2. True or false? Allele frequency calculations must be calculated for each allele separately.

3. Draw a table with 2 columns and 3 rows. How many cells total are in this table?

Activity

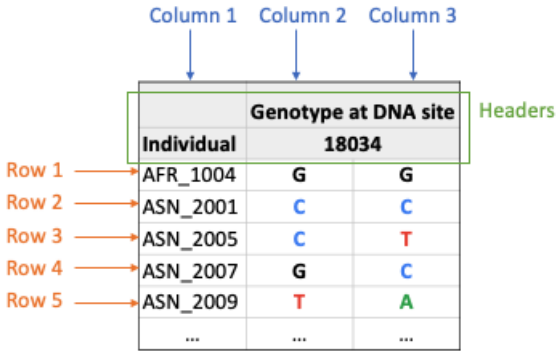
Dr. V shows Sam and the rest of the class an example of an ancestry testing company called *23Chromosomes* that is building a reference DNA database. *23Chromosomes* selected 50 individuals who volunteered to share their genetic information with the company. These carefully selected individuals represent 5 populations from around the world: African, Asian, European, Native American, and Pacific Islander. *23Chromosomes* will use the genetic data from these individuals to build their own DNA **reference database**.

Follow along with Dr. V's example to analyze the data from *23Chromosomes* and learn how DNA ancestry tests work. Calculate the allele frequencies for different alleles across populations and identify which alleles are associated with which populations.

Part 1. Explore the data

The genomic technologies division of *23Chromosomes* performed SNP genotyping on 50 individuals from 5 regions around the world. They delivered the genotype data in spreadsheets and sent it to the computational biology team for analysis.

First, let's observe part of the data in the spreadsheets you chose to learn how it is organized. This image shows part of Table 1 for DNA Site 18034, with the rows, columns, and headers identified. Use this table shown to answer the questions that follow.



a. Record the name of the Column 1 header.

b. How many individuals are displayed in the portion of the table shown?

c. The individuals are distinguished from one another by a 3-letter code, an underscore, and a 4-number code. Based on what you know about the origin of this data, what do the 3 letters represent?

d. Which geographic regions are represented in the individuals shown in this part of the table?

e. Record the name of the header for Columns 2 and 3.

f. Why are there two columns for this header?

g. Record the genotype for the individual in Column 1, Row 1.

h. Choose another individual and record their information:
Population:
Individual number:
Genotype at DNA site 18034:

Part 2. Organize the data

Now that we're more familiar with the data, it's time to start working with the data from *23Chromosomes*.

Find all data from *23Chromosomes* in the linked Google Spreadsheets. There are five spreadsheets, each with genotype data from one of the five DNA sites (SNPs) in all 50 individuals. There is also one master spreadsheet with all of the data. Choose one of the five DNA sites to start and focus on **Table 1 Genotype Data**.

23Chromosomes DNA Site 18034
23Chromosomes DNA Site 46754
23Chromosomes DNA Site 53134
23Chromosomes DNA Site 95005
23Chromosomes DNA Site 123030
- or -
23Chromosomes All Data

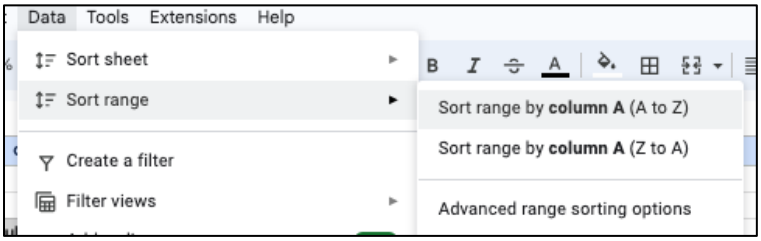
Table 1 contains the genotype data from 50 individuals in all five populations at one DNA site. This data is not organized. Your first task is to sort the table named Table 1 so that the individuals are grouped by population.

One approach is to select the data in the table and use a spreadsheet feature called sort. Using your cursor or mouse, select all of the individuals and all of the genotype data. Do not select any of the headers. Make sure you scroll and capture all the data that may be outside of your current view!

5	(1) First, organize Table 1 based on population.		
6			
7	Table 1. Genotype data		
8			
9		Genotype at DNA site	
10	Individual	18034	
11	AFR_1004	G	G
12	ASN_2001	C	C
13	ASN_2005	C	T
14	ASN_2007	G	C
15	ASN_2009	T	A
16	EUR_3002	G	G
17	EUR_3009	C	A
18	NAM_4001	A	T

In Table 1, select all of the individual and genotype data, not the headers.

Navigate to the top menu Data, then select Sort range, then select Sort range by column A (A to Z).



Sort the selected data range.

Now look back at Table 1. You should see the data changed in the table. To check and ensure you sorted properly, find the individuals whose information you recorded previously. Are their genotypes the same? If yes, then you sorted properly! If not, try again by going back to the original spreadsheet.

Part 3. Count alleles

The genotype data in Table 1 should now be organized with all individuals grouped by population. Now, the alleles for all individuals can be counted more easily.

In your spreadsheet, find **Table 2 Allele Count Data** for the DNA site you are working on. Next, use the sorted Table 1 to count each allele, for the given DNA site, in each population. Record your results in Table 2. Continue counting each variant for all 10 individuals in the population. Repeat this for all five regions. Total your counts and enter them into the appropriate cells in Table 2.

Part 4. Calculate allele frequency

Now that you calculated how many variants are detected in the individuals at each DNA site, you can calculate the allele frequency, or the proportion of each variant, in each population.

In your spreadsheet, find **Table 3 Allele Frequency Percent Data** for the DNA site you are working on. Use the allele count data you previously generated in Table 2 to calculate the allele frequency percent in each population.

Recall that the formula to calculate allele frequency percent is:

$$Allele\ Frequency = \frac{Allele\ count\ of\ a\ specific\ allele\ in\ the\ population}{Total\ allele\ count\ in\ the\ population} * 100$$

Continue your calculations for each variant for all 10 individuals in each population. Repeat this for all five regions. Record your results in the appropriate cells in Table 3.

Part 5. Compare allele frequency data

Look at the allele frequency data from all five DNA sites in all five populations. First, compare the numbers for the allele frequencies at one DNA site across all populations. Then, compare the numbers for allele frequencies in one population across all DNA sites. What do you notice? Choose at least three of the following questions and record your answers below.

1. What biological concepts can explain these observations?
2. Why don't all individuals from the same population have the same genotypes?
3. Why do some populations have the same frequencies at some DNA sites?
4. Would you expect a different ancestry testing company to have the same allele frequency data for these populations?
5. What would the data look like if we had 500 individuals? What about 500,000 individuals?
6. What would make you more confident in the data?

Part 6. Ethics

The company *23Chromosomes* used data from individuals from 5 populations: African (AFR), Asian (ASN), European (EUR), Native American (NAM), and Pacific Islander (PIS).

Using the map provided, circle and label the areas that correspond to each region. If you're in a group or class, work independently, then compare your answers with your peers when everyone has finished. Then choose at least three of the following questions and record your answers below.

1. If you compared with your peers, did you have different answers? Which version is correct?
2. What areas of the world are not covered?
3. What about countries that span more than one continent?
4. How did you define Asian? Native American? European?
5. How large is each region?
6. How many modern-day countries are in each region?
7. How would you have defined these regions 100 years ago? How might these regions change 100 years from now?

Part 7. Careers

Explore jobs and career paths that relate to this activity.

1. Navigate to any job search site. Some recommended ones are
   - Indeed: https://www.indeed.com/salaries
   - Zippa: https://www.zippia.com/careers
   - JobViz: https://www.galacticpolymath.com/jobviz

2. Search for the jobs from this activity. You can also find more jobs using keywords from the activity or explore jobs and categories on the site.

   Bioinformaticians analyze genomic data. Similar jobs include Data Scientist and Bioinformatics Analyst.

   Genomic Technologists perform DNA sequencing assays. Similar jobs include Laboratory Technician and Research Technician.

   Population Geneticists study the genetics of groups of individuals, like humans. Similar jobs include Computational Geneticist and Statistical Geneticist.

3. Fill in the table below with 2 jobs that interest you. Record the job title, degree(s) and training needed, and the salary estimate. You can also write a description of the work and any other notes about why you found the job interesting. Continue filling the table with more jobs if you want to.

Job title	Degree(s)/training needed	Salary estimate	Description of work	Notes

4. Hear from people in some of the jobs you just discovered. Choose 1 resource, then use the table below to record the resource you explored, as well as your thoughts about the people and the jobs they do. Continue filling the table with more jobs if you want to.
    - JAX Career Chats: Bioinformatics Analyst
    - JAX Career Spotlight: Genomic Technologist
    - I Am A Scientist: Population Geneticist

Resource you explored	Person name and job title	In your own words, describe the work they do	Which of their traits match your skills and interests?	What do you want to learn more about?

Glossary

Allele – One of several alternative forms of a gene. Alleles for a given gene have different DNA sequences, which can lead to different phenotypes. In a diploid cell, each cell has two alleles for each gene (one from each parent). These two alleles will be found at the same position (locus) on homologous chromosomes.

Allele frequency – The fraction of all chromosomes in a population that carry a specific allele over the total population.

Cell – In a spreadsheet or table, a cell is one unit where information such as numbers or words can be entered.

Column – In a spreadsheet, when information from multiple cells is arranged vertically.

Deoxyribonucleic acid (DNA) – A biological molecule of which the primary role is the storage of genetic information. DNA is made of deoxyribonucleotides. The nitrogenous bases found in DNA include adenine (A), guanine (G), cytosine (C) and thymine (T).

Gamete – A gamete is a reproductive cell of an animal or plant, such as egg and sperm cells. In animals, these cells are haploid and carry only one copy of each chromosome.

Genome – The complete set of chromosomes, or genetic material, of an organism or cell.

Genotype – The genetic makeup of an organism (whereas the term phenotype describes the physical traits of an organism). The term genotype can be used to describe which alleles an individual has for a single gene. When describing one particular gene, genotype refers to the pair of alleles inherited by the organism for that gene. The term genotype can also be used to describe the complete set of a cell's or an organism's genes. For example, humans have about 20,000 genes, so a genotype would indicate which alleles are present for each of those genes.

Header – In a spreadsheet, cells that have descriptions that provide additional information about what is in a row or column.

Homozygous – Having two identical alleles for a given gene (ie, the same allele was inherited from each parent).

Horizontal – A direction that is parallel to the plan of the horizon, at right angles to the vertical.

Locus – A locus in genomics is a physical location within the genome. It can be a region like a gene or portion of a gene, or a specific nucleotide or set of nucleotides.

Nucleotide – The building blocks of nucleic acids (DNA and RNA) comprised of a nitrogenous base, five-carbon sugar, and phosphate.

Phenotype – The observable characteristics of a cell or organism.

Population – In biology, a population is the total number of individuals in a specific area.

Reference genome – A template genome for organism. This genetic resource is a digital nucleic acid sequence database, assembled by scientists as a representative example of the genomic features in one idealized individual.

Row – In a spreadsheet, when information from multiple cells is arranged horizontally.

Sequencing – Sequencing is a technique that is used in molecular biology to determine the order of nucleotides in a particular DNA or RNA molecule. It involves identifying the specific order of nucleotides that make up the molecule, which can provide important information about the genetic code of an organism. Sequencing can be done using a variety of methods, including Sanger sequencing, next-generation sequencing, and single-molecule sequencing.

Single nucleotide polymorphism – A single nucleotide polymorphism (abbreviated as SNP, pronounced as *snip*) is a type of genomic variant when one single nucleotide in the DNA is changed. Also known as a single nucleotide variant (abbreviated as SNV).

Variable – In mathematics, a variable is a symbol (such as v, N, or x) that represents an object (such as a number).

Variant – A change or difference in the DNA sequence of a cell or organism.

Vertical – A direction where the top is directly above the bottom, at right angles to the horizontal.