

# Cornell Institute for Biology Teachers

Copyright Cornell Institute for Biology Teachers 1998, 2000.  
This work may be copied by the original recipient from CIBT to provide copies for users working under the direction of the original recipient. All other redistribution of this work without the written permission of the copyright holder is prohibited.

Lab issue/rev. date: 10/98

**Title:**

**Statistics and Probability in Evaluation of DNA Evidence**

**Authors:**

Martha Hamblin, Cornell University, Ithaca, NY

edited by: Nancy Wright, Honeoye Central School, Honeoye, NY

**Appropriate  
Level:**

Life Science, High School, Honors, or Advanced Placement Biology

**Abstract:**

In this lab students will apply statistical analysis to the DNA profiling results of an actual rape case. This lab is an extension of DNA profiling labs which students have previously completed. In part I of this exercise, students will use sets of candy to represent alleles in a particular population at a particular locus. They will sample "alleles" from "populations" at three "loci", estimate frequencies of alleles in those samples and calculate probabilities of particular combinations of alleles.

**Time Required:** **Teacher prep:** minimal.

**In class:** Two 40-minute periods. This time will also include a discussion of principles with your students.

**National Science  
Standards:**

The entire contents can be found at : <http://books.nap.edu/html/nses/html/>

CONTENT STANDARD C: As a result of their activities in grades 9-12, all students should develop understanding of molecular basis of heredity

CONTENT STANDARD E: As a result of activities in grades 9-12, all students should develop understandings about science and technology.

# Additional Teacher Information

## Information with which students should be familiar:

- Process of DNA profiling and interpretation of profiles
- Mendelian genetics
- Probability and statistics: students should be able to calculate a frequency.
- Vocabulary: Loci, allele, gene frequencies, genotype, phenotype
- knowledge of scientific notation is helpful

## Materials:

1 bag of M& M's per lab group (or 40 -60 M& M's in a cup). Skittles may also be used, they have 5 different colors that each occur with equal frequency.

Xerox Copies of the autorads from the real rape case.

A photograph or genuine autorad would be a nice addition.

## Helpful hints:

There are two student versions of this lab, one that examines 3 loci, another that examines 6. For three loci, the numbers stay small enough that students will be able to do the calculations without using scientific notation on a standard calculator. For 6 loci, scientific calculators, and a working knowledge of scientific notation are needed.

Remind students not to eat the candies until they are directed to do so.

According to the M&M/MARS corporation, on average the mix of colors for plain M&M's chocolate candies is 30% brown, 20% yellow, 20% red, 10% orange, 10% green and 10% blue.

## Answers to questions:

### Part I

- 1) Are your allele frequencies the same as the frequencies of the other groups?

*Different samples of the same size will vary by chance.*

- 2) Are your allele frequencies the same as the class average? Why or why not?

*Larger samples are more accurate.*

## Part II

- 1) At any one locus, the probability of a particular combination of two different alleles is  $2ab$ . Why do we multiply by 2?

*There are two different ways to get that combination of  $a$  and  $b$ . ( $A$  from mom and  $B$  from dad or  $A$  from dad and  $B$  from mom)*

- 2) What if the two alleles are the same (do you still need to multiply by 2)?

*You do not need to multiply by two.*

### Practice example:

Assuming that the three loci are unlinked (i.e., independent), calculate the probability of observing this particular combination of alleles (i.e., multilocus genotype) in this population.

You have drawn the following:

at locus A: alleles 1 & 2, at locus B: alleles 2 & 4 and at locus C, alleles 1 & 3

The frequencies of those alleles are as shown below:

Locus A	Freq.	Locus B	Freq.	Locus C	Freq.
Allele 1	0.55	Allele 2	0.10	Allele 1	0.25
Allele 2	0.20	Allele 4	0.40	Allele 3	0.40

$$A = 0.22 (2.2 \times 10^{-1})$$

$$B = 0.08 (8.0 \times 10^{-2})$$

$$C = .20 (2.0 \times 10^{-1})$$

So, the probability for this multilocus genotype is  $= 0.242 \times 0.08 \times 0.20 = .00352 = 3.52 \times 10^{-3}$

**Interpretation:** in a population of 10,000 with these allele frequencies, about 3 or 4 individuals will have this same multilocus genotype.

## Part III

- 1) Are the shorter bands at the top or the bottom?

*bottom*

- 2) What is the reason for including positive controls (human DNA unrelated to the case) on the gel?

*The positive control is used to make sure that the correct probe was used and that the digest ran correctly. The positive control should always produce a band of known size with a particular probe. If it doesn't something is wrong.*

- 3) What was the purpose of the test from the underpants where there was no sperm?

*This was done to prove that some chemical or structure in the underpants did not cause the banding pattern.*

- 4) Which two lanes on the gel are from the evidence?

*6 & 7*

- 5) How were these DNA samples treated before being run on the gel?

*They must be digested with a restriction enzyme to create fragments. In this case *HindIII* was used.*

- 6) What is the smallest number of bins at any one locus? The largest?

*13 bins at locus *D17S79**

*26 bins at locus *D14S13**

- 7) Find the bin with the highest frequency for each locus. Write the frequency next to the locus name below.

**D10S28:** 0.087      **D14S13:** 0.228      **D2S44:** 0.124

**D17S79:** 0.263      **D1S7:** 0.079      **D4S139:** 0.191

- 8) Which loci are more valuable for creating a DNA profile to identify an individual? Why?

*Loci with more bins usually have lower frequencies for any particular bin, so there is less likely to be a match just by chance, Locus *D17S79*, for example, has the smallest number of bins, and the bin with the highest frequency among all six loci. At this locus, the suspect and the victim share both alleles just by chance, because those alleles are actually common in the population. In contrast, *D10S28* has no bin frequency over 0.087, so a match by chance is much less likely, and this locus is more valuable.*

- 9) How do the numbers and frequencies of "alleles" in your M & M exercise compare to actual numbers of alleles at VNTR loci in humans?

*The number of alleles is much larger and the frequencies are often much lower.*

**Data Table 2:**

Locus	Allele	Matches suspect Y or N	Between marker _____ and marker _____	Frequency of allele	Frequency of genotype
<b>DS244</b>	<b>1</b>	Y	12/13	0.083	0.014
	<b>2</b>	Y	16/17	0.086	$(1.4 \times 10^{-2})$
<b>D17S79</b>	<b>1</b>	Y	6/7	0.224	0.089
	<b>2</b>	Y	9/10	0.199	$(8.9 \times 10^{-2})$
<b>D10S28</b>	<b>1</b>	Y	11/12	0.047	0.0018
	<b>2</b>	Y	13/14	0.019	$(1.8 \times 10^{-3})$
<b>D14S13</b>	<b>1</b>	Y	7/8	0.053	0.0038
	<b>2</b>	Y	20/21	0.036	$(3.8 \times 10^{-3})$
<b>D4S139</b>	<b>1</b>	Y	15/16	0.006	0.0016
	<b>2</b>	Y	25/26	0.131	$(1.6 \times 10^{-3})$
<b>DIS7</b>	<b>1</b>	Y	20/21	0.067	0.0039
	<b>2</b>	Y	14/15	0.029	$(3.9 \times 10^{-3})$

According to the rule of multiplication, the probability that two or more independent events will occur in combination is the product of the individual probabilities. Calculate the probability for the multilocus genotype for two loci. Do this by multiplying together the frequencies of each of the two loci.

**Data table 3:**

Locus	Frequency 2ab	Multi locus frequency (use scientific notation)
<b>D2S44</b>	0.014 $(1.4 \times 10^{-2})$	
<b>D17S79</b>	0.089 $(8.9 \times 10^{-2})$	D2S44 x D17S79 = $1.24 \times 10^{-3}$
<b>D10S28</b>	0.0018 $(1.8 \times 10^{-3})$	D2S44 x D17S79 x D10S28 = $2.48 \times 10^{-6}$
<b>D14S13</b>	0.0038 $(3.8 \times 10^{-3})$	D2S44 x D17S79 x D10S28 x D14S13 = $9.92 \times 10^{-9}$
<b>D4S139</b>	0.0016 $(1.6 \times 10^{-3})$	D2S44 x D17S79 x D10S28 x D14S13 x D4S139 = $1.98 \times 10^{-11}$
<b>DIS7</b>	0.0039 $(3.9 \times 10^{-3})$	D2S44 x D17S79 x D10S28 x D14S13 x D4S139 x DIS7 = $7.9 \times 10^{-14}$

- 10) State in words what the multi locus probability (for all six loci) means.

*This is the probability that two individuals will share, by random chance, all six loci.*

*Regents version: 2.48 people out of one million would have this combination of alleles by chance.*

*AP version: 7.9 people out of 100,000,000,000,000 people would have this same combination of alleles by chance.*

- 11) How does this probability compare with the size of the human population (about 6 billion)?

*Regents: It is larger. In this example, 4 loci would equal the world population. But, the more loci that match, the less likely it becomes that someone else will have the same profile.*

*AP: It is MUCH smaller.*

- 12) Do you think that the suspect is the person who left the sperm evidence at the crime scene?

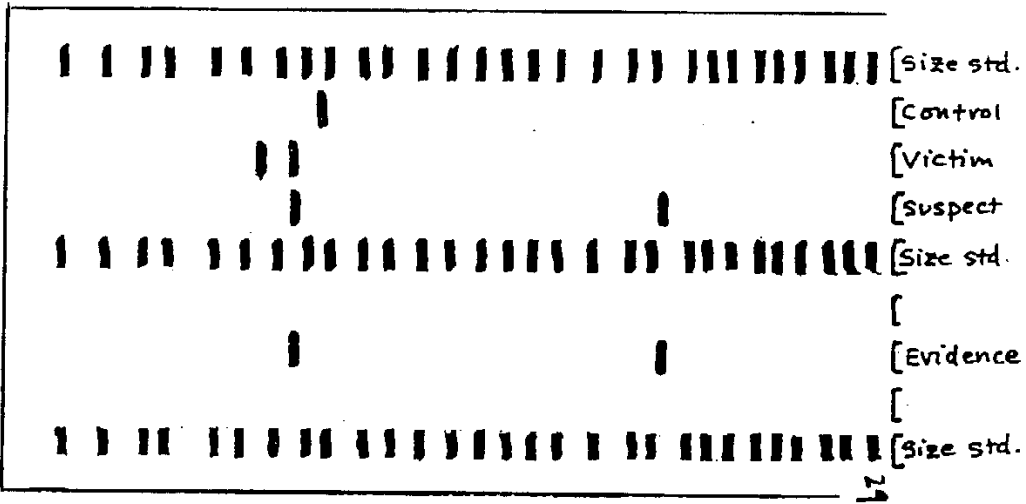
*Yes*

- 13) How many loci does it take to prove, “beyond the shadow of a doubt”, that the suspect is guilty.

*In this case, 4.*

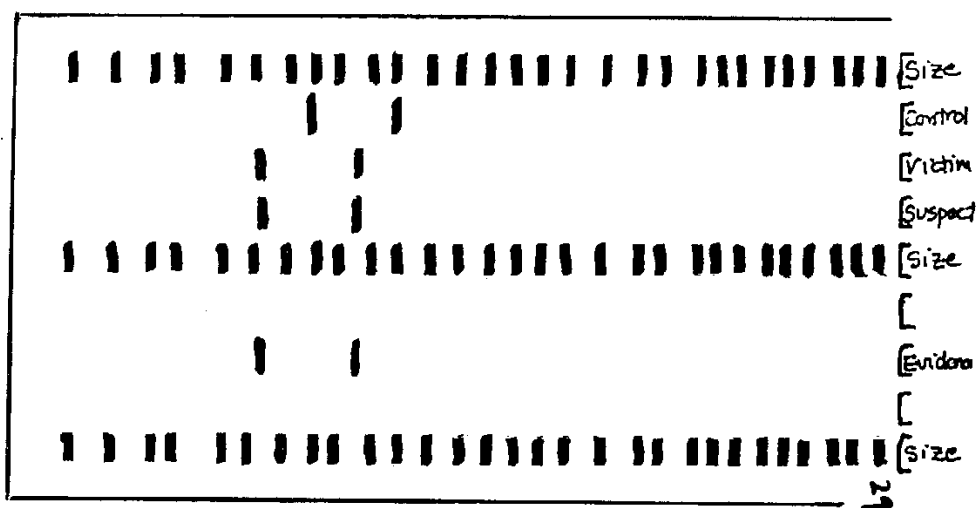
# Locus D14S13

between markers	Count	Fraction
< 4	6	0.004
4-6	6	0.004
6-7	85	0.057
7-8	80	0.053
8-9	342	0.228
9-10	120	0.080
10-11	216	0.144
11-12	46	0.031
12-13	122	0.081
13-14	41	0.027
14-15	55	0.037
15-16	38	0.025
16-17	37	0.025
17-18	46	0.031
18-19	45	0.030
19-20	39	0.026
20-21	54	0.036
21-22	25	0.017
22-23	45	0.030
23-24	20	0.013
24-25	8	0.005
25-27	10	0.007
27-29	7	0.005
> 29	9	0.006
Totals	1584	1.000



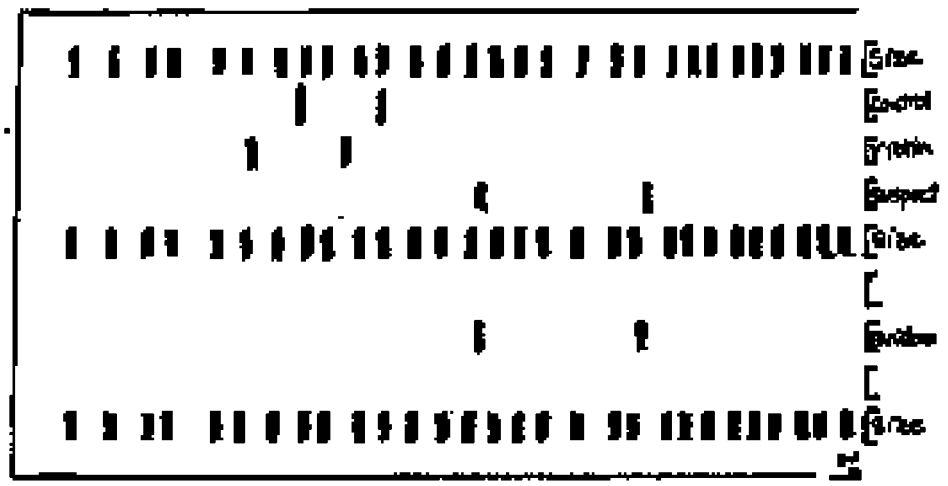
# Locus D17S79

between markers	Count	Fraction
< 1	16	0.010
1 - 2	5	0.003
2 - 3	11	0.007
3 - 5	6	0.004
5 - 6	23	0.015
6 - 7	348	0.224
7 - 8	307	0.198
8 - 9	408	0.263
9 - 10	309	0.199
10 - 11	44	0.028
11 - 12	50	0.032
12 - 13	16	0.010
> 13	9	0.006
Totals	1552	0.999



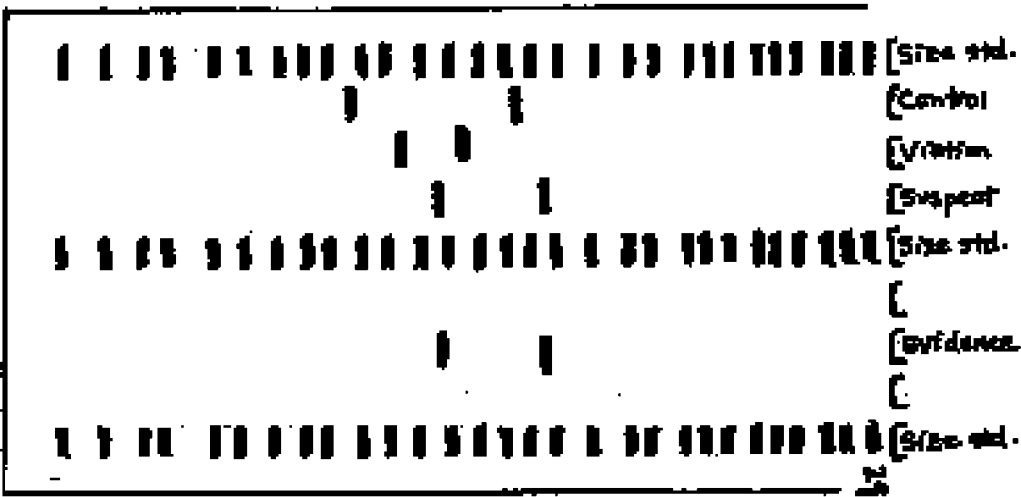
# Locus D1S7

between markers	Count	Proportion
<5	5	0.004
5-6	7	0.006
6-7	11	0.009
7-8	14	0.012
8-9	13	0.011
9-10	16	0.013
10-11	12	0.010
11-12	34	0.029
12-13	24	0.020
13-14	16	0.013
14-15	34	0.029
15-16	37	0.031
16-17	55	0.046
17-18	81	0.068
18-19	66	0.055
19-20	74	0.062
20-21	80	0.067
21-22	65	0.055
22-23	71	0.060
23-24	75	0.063
24-25	94	0.079
25-26	92	0.077
26-27	91	0.076
27-28	39	0.033
28-29	23	0.019
> 29	61	0.051



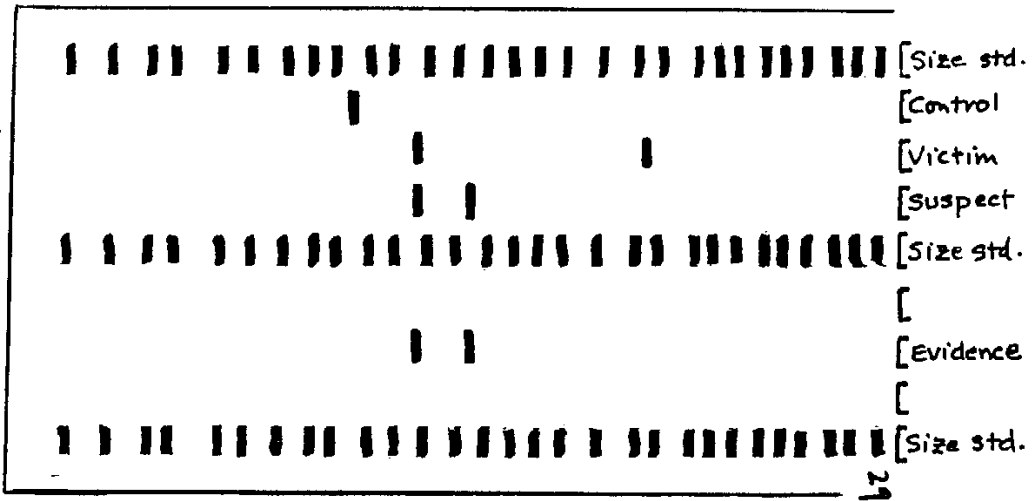
# Locus D2S44

between markers	Count	Fraction
< 3	8	0.005
3-4	5	0.003
4-5	24	0.015
5-6	30	0.024
6-7	73	0.046
7-8	85	0.055
8-9	197	0.124
9-10	170	0.107
10-11	131	0.083
11-12	79	0.050
12-13	131	0.083
13-14	60	0.038
14-15	65	0.041
15-16	63	0.040
16-17	136	0.086
17-18	141	0.089
18-19	119	0.075
19-20	36	0.023
20-21	27	0.017
21-24	13	0.008
> 24	12	0.008
<b>Totals</b>	<b>1584</b>	<b>1.000</b>



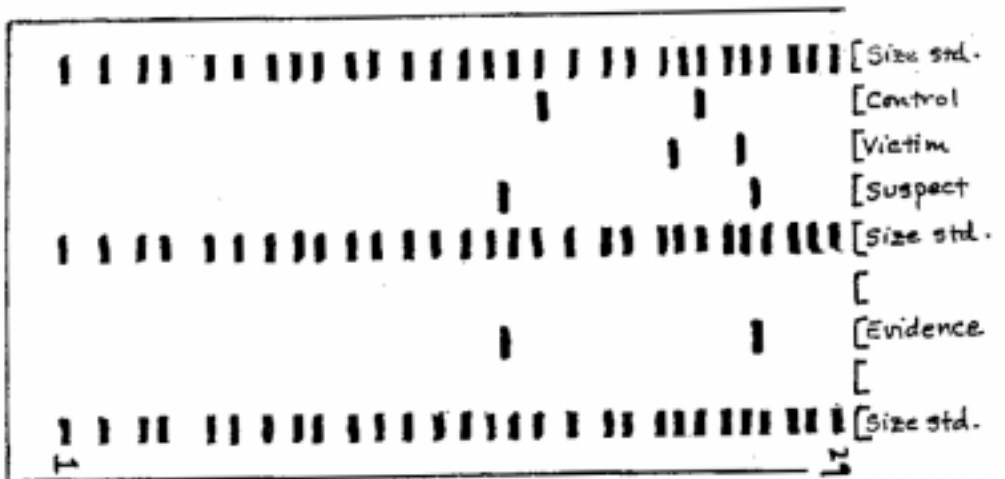
# Locus D10S28

between markers	Count	Fraction
< 4	13	0.015
4-5	44	0.051
5-6	38	0.044
6-7	15	0.017
7-8	34	0.040
8-9	67	0.078
9-10	75	0.087
10-11	71	0.083
11-12	40	0.047
12-13	51	0.059
13-14	16	0.019
14-15	14	0.016
15-16	36	0.042
16-17	42	0.049
17-18	41	0.048
18-19	56	0.065
19-20	39	0.045
20-21	62	0.072
21-22	58	0.068
22-23	12	0.014
23-24	6	0.007
24-25	23	0.027
> 25	5	0.006
Totals	858	0.999

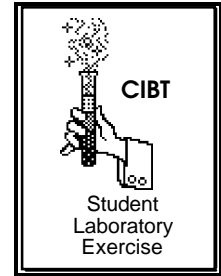


# Locus D4S139

between markers	Count	Fraction
< 13	5	0.004
13-14	12	0.010
14-16	7	0.006
16-17	17	0.014
17-18	37	0.031
18-19	27	0.023
19-20	47	0.040
20-21	56	0.047
21-22	64	0.054
22-23	86	0.072
23-24	128	0.108
24-25	227	0.191
25-26	156	0.131
26-27	113	0.095
27-28	43	0.036
28-29	42	0.035
>29	121	0.102
Totals	1188	0.999



# Statistics and Probability in Evaluation of DNA Evidence



## Introduction:

Many questions in biology have simple yes-or-no answers: for example, "Is light necessary for photosynthesis?" Does this gene code for a certain protein?

Other questions are less straightforward, because there is a lot of variation. We need to look for patterns in the variation, and decide whether there are real patterns, or whether the variation is simply random. The science of statistics provides us with a systematic way to answer such questions.

"Statistics deals with techniques for collecting, analyzing, and drawing conclusions from data.... The basic ideas in statistics assist us in thinking clearly about the problem, provide some guidance to the conditions that must be satisfied if sound inferences are to be made, and enable us to detect many inferences that have no good logical foundation." (Statistical Methods, Snedecor and Cochran 1989).

Many important questions in everyday biology require the use of statistical analysis. Some examples are:

- Mapping of genetic traits such as human genetic disorders, disease and insect resistance in crop plants
- Determining the effect of diet on human health
- Determining the effectiveness of medical treatments and procedures.

DNA evidence is increasingly being used as evidence in murder and rape trials, as well as for other forms of human identification. If you read about these cases in the news, you will come across numbers like "There is a 1 in 1 million chance that the DNA was not that of the suspect", or "1 in 35,000 Caucasians will have the genotype of the victim." These are examples of numbers that are the results of statistical analysis of DNA profiling data.

In previous classes, you have learned about the molecular biology of VNTR loci, and how they can be detected by Southern blotting, with the end product of an autoradiogram. Today, we are going to focus on the interpretation of those autoradiograms. You will be given some copies of actual autoradiograms from a rape case that took place near here last

year, and will learn how to calculate a probability that the suspect is in fact the perpetrator of the crime. In order to do this accurately, you need to apply some principles from statistics and population genetics. So, before we look at the data from the rape case, we are going to do a simple exercise that should demonstrate some of the basic concepts that are important in many kinds of statistical analysis.

Some technical terms that you will encounter in this exercise:

**Allele:** an alternate form of a gene at a given locus.

**Estimate:** a measurement of a property in the sample.

**Frequency:** the number of items occurring in a given category. It can be expressed as a Decimal % between 0 and 1. The frequencies of all the categories added together must be equal to 1.

**Genotype:** the genetic make-up of an organism.

**Locus (Loci):** a particular place along a given chromosome where a given gene is located.

**Phenotype:** the physical and physiological traits of an organism.

**Population:** a large group of individuals about which we wish quantitative information.

**Probability:** the relative frequency with which an event occurs by chance

**Sample:** a set of items or individuals selected from a larger group, the population.

## **Part I: What are the odds of getting a red M & M?**

In this exercise, you will use different colored M & Ms to represent alleles in a particular population at a particular locus. You will:

- 1) Estimate frequencies of alleles.
- 2) Calculate probabilities of particular combinations of alleles.

Each pair of students will have a container of M&Ms. Each container represents the alleles in the sample. There are six different colored alleles in each container, present in different **frequencies**.

Our first challenge is to estimate the frequency of the alleles. In other words, we need to calculate the frequency of each of the six colors of M & Ms. Count the number of M&Ms of each color, then calculate what part of the total number they represent. Report your data to the class and calculate the class average frequency for each color.

$$\text{Frequency} = \frac{\text{\# of a color}}{\text{total \# M\&Ms}}$$

Table 1:

Color	Number	Frequency	Class average frequency
Red			
Green			
Blue			
Yellow			
Brown			
Orange			
<b>TOTAL:</b>			

- 1) Are your allele frequencies the same as the frequencies of the other groups?
  
- 2) Are your allele frequencies the same as the class average? Why or why not?

Different samples of the same size will vary by chance. Larger samples are more accurate. Two estimates are significantly different if they are more different than we would expect by chance alone.

### Part II: What are the odds of getting a red and a green M & M?

To interpret DNA forensic data correctly, we need to know the probability that two randomly chosen individuals will share, purely by chance, the same combination of alleles at VNTR loci. This, in turn, requires knowing the frequencies of those alleles in the population. You will do a simple example of this type of calculation; frequency data will be used to calculate the **probability** of a particular genotype at a single locus in your population (container) of candy.

Without looking, randomly draw two M&Ms from your container. This will represent the alleles you received from mom (egg) and dad (sperm).

Record the colors of the two candies. This is your randomly chosen genotype:

(Color 1) a = \_\_\_\_\_ (Color 2) b= \_\_\_\_\_

Below, enter the frequencies of these alleles from the class data chart. Use the actual frequencies for the population (container) from which you picked your alleles. We will call one allele a and the other b.

Frequency (color 1) a = \_\_\_\_\_ frequency (Color 2) b= \_\_\_\_\_

- 1) At any one locus, the probability of a particular combination of two different alleles is  $2ab$ . Why do we multiply by 2?
- 2) What if the two alleles are the same (do you still need to multiply by two)?

**Multilocus Genotypes practice example:**

In DNA data, we don't just look at one locus. Several different loci are combined to calculate a multi locus genotype. Assuming that the three loci are unlinked (i.e., inherited independently), calculate the probability of observing this particular combination of alleles (i.e., multilocus genotype) in this population.

You have drawn the following alleles:

at locus A: alleles 1 & 2, at locus B: alleles 2 & 4 and at locus C, alleles 1 & 3

**The frequencies of those alleles are as shown below:**

Locus A	Freq.	Locus B	Freq.	Locus C	Freq.
Allele 1	0.55	Allele2	0.10	Allele 1	0.25
Allele 2	0.20	Allele 4	0.40	Allele 3	0.40

A = 2 ( \_\_\_ X \_\_\_ )

B = 2 ( \_\_\_ X \_\_\_ )

C = 2 ( \_\_\_ X \_\_\_ )

A = \_\_\_\_\_

B = \_\_\_\_\_

C = \_\_\_\_\_

You have calculated the frequencies of individuals having that combination of alleles at each of the three loci listed above.

Now we want to know the frequency of individuals which would have all 3 of these alleles. This is called the **multilocus genotype**. The multilocus genotype is the product of the frequencies for each locus. In other words, the frequency at A x frequency at B x frequency at C = \_\_\_\_\_.

This means that \_\_\_\_\_ out of \_\_\_\_\_ individuals would have exactly the same alleles at these three loci.

### **Part III: Analysis of actual data from a rape case in NY**

In 1996, a man lured a 13 year old girl into a cabin and allegedly raped her. She went home, bathed and changed her clothes. Her mother came home and the girl told her what had happened. The mother took the girl to the hospital, where a vaginal swab did not show any sperm present. However, the girl's dirty underpants and shorts had semen stains on them. DNA was extracted from the semen and used to make a DNA profile of the accused rapist by RFLP analysis. You will be given copies of autoradiograms showing the VNTR alleles from DNA extracted from semen from the girl's clothing, and DNA of the suspect.

Find the following lanes on your autoradiogram:

**Lanes 1: Size standards,**

These are 29 marker bands of known size. The smallest band is 639 base pairs; the longest is 12, 830 base pairs.

**Lane 2:** Positive controls. A human DNA sample from an individual who is unrelated to the crime. They are used because their DNA creates bands of known size.

**Lane 3:** DNA from the **victim** (13 year old girl).

**Lane 4:** DNA from the **suspect**.

**Lane 5:** Size standard.

**Lane 6:** sample the girl's underpants where there was no sperm.

**Lane 7:** DNA extracted from sperm on the underpants.

**Lane 9:** size standard.

**Answer the following questions before you go on:**

- 1) On a DNA gel, are the shorter bands at the top or the bottom?
  
- 2) What is the reason for including positive controls (the human DNA unrelated to the case) on the gel?
  
- 3) What was the purpose of the test from the underpants where there was no sperm?
  
- 4) Which two lanes on the gel are from the evidence?
  
- 5) How were these DNA samples treated before being run on the gel?

Six different probes were used to make these autoradiograms. The lab did not need to run 6 different gels. The same blot was probed six times with 6 different radioactive markers. This is fairly time-consuming, compared to doing PCR. The six loci are identified by number and letter combinations. The number identifies the chromosome on which the locus is found.

Your first task is to decide whether the alleles of the suspect match the alleles of the DNA evidence for each locus. You were given copies of 3 of the six autorads. Examine the 3 autorads you were given and determine if the alleles appear to be the same. Use the same technique you used in the DNA profiling lab. Do the alleles in the evidence match the suspect's alleles for each locus? Record your answer, for each allele, in data table 2.

At this point we do not know whether this is very strong evidence that the suspect committed the crime. We need to know whether such a match is likely to have occurred simply by chance, or whether it is reasonable to conclude that the multilocus genotype

matches because the evidence was in fact left by the suspect. So we will calculate the probability of such a three-locus genotype occurring by chance in the population.

We will use a process called “**binning**”. In human DNA, because there are so many different sized alleles, some of which are very similar in size, it would be impossible to try to distinguish every different allele size on a gel of this kind. To get around this problem, bands that fall between adjacent markers are classified as all belonging to the same “bin”, even if they differ somewhat in size.

You have been given tables of actual allele frequencies from a Caucasian reference population for the three different VNTR loci used in this case, presented on the same page with the autoradiogram showing the pattern observed at that locus. These tables show the alleles sizes that were observed when these loci were scored in large samples of individuals. You can see that the number of possible combinations of alleles is very large.

The column labeled “between markers” tells you the size class (“bin”) of the alleles, the column labeled “count” tells you the number of people in the sample population who had that allele. And, the column labeled “fraction” tells you the frequency of that bin in the sample.

Examine the frequency tables for the 3 loci you were given.

- 6) What is the smallest number of bins at any one locus? \_\_\_bins at locus \_\_\_\_\_  
The largest? \_\_\_\_\_ bins at locus \_\_\_\_\_
  
- 7) Find the bin with the highest frequency for each locus. Write the frequency next to the locus name below.  
  
D2S44: \_\_\_\_\_ D17S79: \_\_\_\_\_ D10S28: \_\_\_\_\_
  
- 8) Which loci are more valuable for creating a DNA profile to identify an individual?  
Why?
  
- 9) How do the numbers and frequencies of the "alleles" in your M & M exercise compare to actual numbers and frequencies of alleles at VNTR loci in humans?

Now you will calculate the actual probability of the multilocus genotype observed in the suspect and the evidence in this case:

The first thing you need to do is evaluate if the suspect's DNA matches the evidence for each allele at every locus. Determine if there is a visual match at each locus and record your results in table 2.

For each locus, place a straight-edge across the markers which run down both sides of the autorad. Find the two marker bands that are above and below the alleles in the evidence. **Count up from the bottom** and write the number of the marker bands on your worksheet (table 2). Look up the bins on the tables of allele counts from the reference population; this will be the column marked "between markers". Write the frequencies (fraction) for those bins on your worksheet. If an allele seems to fall right at the position of a marker band, use whichever bin has the higher frequency. Once you have the frequency of the individual alleles, calculate the probability that a randomly chosen individual from that population would have the same combination of alleles at that locus. Use the same formula we used to calculate the probability of drawing a red and green Skittle. Record this in the far right column of table 2. **Remember that you have to multiply by two!**

**Data Table 2:**

<b>Locus</b>	<b>Allele</b>	<b>Matches suspect Y or N</b>	<b>Between marker ___ and marker ___</b>	<b>Frequency of allele</b>	<b>Frequency of genotype</b>
<b>DS244</b>	<b>1</b>		<b>/</b>		
	<b>2</b>		<b>/</b>		
<b>D17S79</b>	<b>1</b>		<b>/</b>		
	<b>2</b>		<b>/</b>		
<b>D10S28</b>	<b>1</b>		<b>/</b>		
	<b>2</b>		<b>/</b>		

According to the **rule of multiplication**, the probability that two or more independent events will occur in combination is the product of the individual probabilities. Calculate the probability for the multilocus genotype for all three loci. Do this by multiplying

together the frequencies of each of the three loci. Record your values as you go in data table 2.

**Data table 3:**

<b>Locus</b>	<b>Frequency</b>	<b>Multi locus frequency</b>
<b>D2S44</b>		
<b>D17S79</b>		D2S44 x D17S79=
<b>D10S28</b>		D2S44 x D17S79 xD10S28=

10) State in words what the multi locus probability (for all three loci) means.

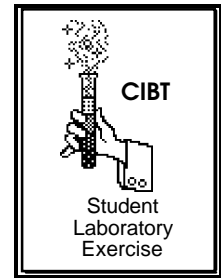
\_\_\_\_\_ out of \_\_\_\_\_ people would have this same combination of alleles by chance.

11) How does this probability compare with the size of the human population (about 6 billion)? How many more alleles would it take to equal or exceed a one in a billion chance?

12) Do you think that the suspect is the person who left the sperm evidence at this crime scene?

13) How many loci does it take to prove “beyond a shadow of a doubt” that a person is guilty.

# Statistics and Probability in Evaluation of DNA Evidence



## Introduction:

Many questions in biology have simple yes-or-no answers: for example: Is light necessary for photosynthesis? Does this gene code for a certain protein?

Other questions are less straightforward, because there is a lot of variation. We need to look for patterns in the variation, and decide whether there are real patterns, or whether the variation is simply random. The science of statistics provides us with a systematic way to answer such questions.

“Statistics deals with techniques for collecting, analyzing, and drawing conclusions from data.... The basic ideas in statistics assist us in thinking clearly about the problem, provide some guidance to the conditions that must be satisfied if sound inferences are to be made, and enable us to detect many inferences that have no good logical foundation.” (Statistical Methods, Snedecor and Cochran 1989).

Many important questions in everyday biology require the use of statistical analysis. Some examples are:

- Mapping of genetic traits such as human genetic disorders, disease and insect resistance in crop plants
- Determining the effect of diet on human health
- Determining the effectiveness of medical treatments and procedures.

DNA evidence is increasingly being used as evidence in murder and rape trials, as well as for other forms of human identification. If you read about these cases in the news, you will come across numbers like “There is a 1 in 1 million chance that the DNA was not that of the suspect”, or “1 in 35,000 Caucasians will have the genotype of the victim”. These are examples of numbers that are the results of statistical analysis of DNA profiling data.

In previous classes, you have learned about the molecular biology of VNTR loci, and how they can be detected by Southern blotting, with the end product of an autoradiogram. Today, we are going to focus on the interpretation of those autoradiograms. You will be given some copies of actual autoradiograms from a rape case that took place near here last

year, and will learn how to calculate a probability that the suspect is in fact the perpetrator of the crime. In order to do this accurately, you need to apply some principles from statistics and population genetics. So, before we look at the data from the rape case, we are going to do a simple exercise that should demonstrate some of the basic concepts that are important in many kinds of statistical analysis.

Some technical terms that you will encounter in this exercise:

**Allele:** an alternate form of a gene at a given locus.

**Estimate:** a measurement of a property in the sample.

**Frequency:** the number of items occurring in a given category. It can be expressed as a Decimal % between 0 and 1. The frequencies of all the categories added together must be equal to 1.

**Genotype:** the genetic make-up of an organism.

**Locus (Loci):** a particular place along a given chromosome where a given gene is located.

**Phenotype:** the physical and physiological traits of an organism.

**Population:** a large group of individuals about which we wish quantitative information.

**Probability:** the relative frequency with which an event occurs by chance

**Sample:** a set of items or individuals selected from a larger group, the population.

## **Part I:** What are the odds of getting a red M & M?

In this exercise, you will use different colored M & Ms to represent alleles in a particular population at a particular locus. You will

- 1) Estimate frequencies of alleles
- 2) Calculate probabilities of particular combinations of alleles

Each pair of students will have a container of M&Ms. Each container represents the alleles in the sample. There are six different colored alleles in each container, present in different **frequencies**.

Our first challenge is to estimate the frequency of the alleles. In other words, we need to calculate the frequency of each of the six colors of M & Ms. Count the number of M&Ms of each color, then calculate what part of the total number they represent. Report your data to the class and calculate the class average frequency for each color.

$$\text{Frequency} = \frac{\text{\# of a color}}{\text{total \# M\&Ms}}$$

**Table 1:**

Color	Number	Frequency	Class average frequency
Red			
Green			
Blue			
Yellow			
Brown			
Orange			
<b>TOTAL:</b>			

- 1) Are your allele frequencies the same as the frequencies of the other groups?
  
- 2) Are your allele frequencies the same as the class average? Why or why not?

Different samples of the same size will vary by chance. Larger samples are more accurate. Two estimates are significantly different if they are more different than we would expect by chance alone.

**Part II: What are the odds of getting a red and a green M & M?**

To interpret DNA forensic data correctly, we need to know the probability that two randomly chosen individuals will share, purely by chance, the same combination of alleles at VNTR loci. This, in turn, requires knowing the frequencies of those alleles in the population. You will do a simple example of this type of calculation; frequency data will be used to calculate the **probability** of a particular genotype at a single locus in your population (container) of candy.

Without looking, randomly draw two M & Ms from your container. This will represent the alleles you received from mom (egg) and dad (sperm).

Record the colors of the two candies. This is your randomly chosen genotype:

(Color 1) a = \_\_\_\_\_ (Color 2) b= \_\_\_\_\_

Below, enter the frequencies of these alleles from the class data chart. Use the actual frequencies for the population (container) from which you picked your alleles. We will call one allele a and the other b.

Frequency (Color 1) a = \_\_\_\_\_ Frequency (Color 2) b= \_\_\_\_\_

- 1) At any one locus, the probability of a particular combination of two different alleles is  $2ab$ . Why do we multiply by 2?
  
- 2) What if the two alleles are the same (do you still need to multiply by two)?

**Multilocus Genotypes practice example:**

In DNA data, we don't just look at one locus. Several different loci are combined to calculate a multi locus genotype. Assuming that the three loci are unlinked (i.e., inherited independently), calculate the probability of observing this particular combination of alleles (i.e., multilocus genotype) in this population.

You have drawn the following alleles:

at locus A: alleles 1 & 2, at locus B: alleles 2 & 4 and at locus C, alleles 1 & 3

**The frequencies of those alleles are as shown below:**

Locus A Freq.		Locus B Freq.		Locus C Freq.	
Allele 1	0.55	Allele 2	0.10	Allele 1	0.25
Allele 2	0.20	Allele 4	0.40	Allele 3	0.40

A= 2 ( \_\_\_ X \_\_\_ )                      B= 2 ( \_\_\_ X \_\_\_ )                      C= 2 ( \_\_\_ X \_\_\_ )  
A = \_\_\_\_\_                              B= \_\_\_\_\_                              C= \_\_\_\_\_

You have calculated the frequencies of individuals having that combination of alleles at each of the three loci listed above.

Now we want to know the frequency of individuals which would have all 6 of these alleles. This is called the **multilocus genotype**. The multilocus genotype is the product of the frequencies for each locus. In other words, the frequency at A x frequency at B x frequency at C = \_\_\_\_\_.

This means that \_\_\_\_\_ out of \_\_\_\_\_ individuals would have exactly the same alleles at these three loci.

### **Part III: Analysis of actual data from a rape case in NY**

In 1996, a man lured a 13 year old girl into a cabin and allegedly raped her. She went home, bathed and changed her clothes. Her mother came home and the girl told her what had happened. The mother took the girl to the hospital, where a vaginal swab did not show any sperm present. However, the girl's dirty underpants and shorts had semen stains on them. DNA was extracted from the semen and used to make a DNA profile of the accused rapist by RFLP analysis. You will be given copies of autoradiograms showing the VNTR alleles from DNA extracted from semen from the girl's clothing, and DNA of the suspect.

Find the following lanes on your autoradiogram:

**Lanes 1: Size standards,**

These are 29 marker bands of known size. The smallest band is 639 base pairs; the longest is 12, 830 base pairs.

**Lane 2:** Positive controls. A human DNA sample from an individual who is unrelated to the crime. They are used because their DNA creates bands of known size.

**Lane 3:** DNA from the **victim** (13 year old girl).

**Lane 4:** DNA from the **suspect**.

**Lane 5:** Size standard.

**Lane 6:** sample the girl's underpants where there was no sperm.

**Lane 7:** DNA extracted from sperm on the underpants.

**Lane 9:** size standard.

**Answer the following questions before you go on:**

- 1) On a DNA gel, are the shorter bands at the top or the bottom?
  
- 2) What is the reason for including positive controls (the human DNA unrelated to the case) on the gel?
  
- 3) What was the purpose of the test from the underpants where there was no sperm?
  
- 4) Which two lanes on the gel are from the evidence?
  
- 5) How were these DNA samples treated before being run on the gel?

Six different probes were used to make these autoradiograms. The lab did not need to run 6 different gels. The same blot was probed six times with 6 different radioactive markers. This is fairly time-consuming, compared to doing PCR. The six loci are identified by number and letter combinations. The number identifies the chromosome on which the locus is found.

Your first task is to decide whether the alleles of the suspect match the alleles of the DNA evidence for each locus. You were given copies of the six autorads. Examine the 6 autorads you were given and determine if the alleles appear to be the same. Use the same technique you used in the DNA profiling lab. Do the alleles in the evidence match the suspect's alleles for each locus? Record your answer, for each allele, in data table 2.

At this point we do not know whether this is very strong evidence that the suspect committed the crime. We need to know whether such a match is likely to have occurred simply by chance, or whether it is reasonable to conclude that the multilocus genotype matches because the evidence was in fact left by the suspect. So we will calculate the probability of such a three-locus genotype occurring by chance in the population.

We will use a process called “**binning**”. In human DNA, because there are so many different sized alleles, some of which are very similar in size, it would be impossible to try to distinguish every different allele size on a gel of this kind. To get around this problem, bands that fall between adjacent markers are classified as all belonging to the same “bin”, even if they differ somewhat in size.

You have been given tables of actual allele frequencies from a Caucasian reference population for the six different VNTR loci used in this case, presented on the same page with the autoradiogram showing the pattern observed at that locus. These tables show the alleles sizes that were observed when these loci were scored in large samples of individuals. You can see that the number of possible combinations of alleles is very large. **The column labeled “between markers” tells you the size class (“bin”) of the alleles, the column labeled “count” tells you the number of people in the sample population who had that allele. And, the column labeled “fraction” tells you the frequency of that bin in the sample.**

Examine the frequency tables for the 6 loci you were given.

- 6) What is the smallest number of bins at any one locus? \_\_\_ bins at locus \_\_\_\_\_ The largest? \_\_\_\_\_ bins at locus \_\_\_\_\_
- 7) Find the bin with the highest frequency for each locus. Write the frequency next to the locus name below.  
  
D2S44: \_\_\_\_\_ D17S79: \_\_\_\_\_ D10S28: \_\_\_\_\_  
D14S13: \_\_\_\_\_ D4S139: \_\_\_\_\_ D1S7: \_\_\_\_\_
- 8) Which loci are more valuable for creating a DNA profile to identify an individual? Why?
- 9) How do the numbers and frequencies of the "alleles" in your M & M exercise compare to actual numbers and frequencies of alleles at VNTR loci in humans?

Now you will calculate the actual probability of the multilocus genotype observed in the suspect and the evidence in this case:

The first thing you need to do is evaluate if the suspect’s DNA matches the evidence for each allele at every locus. Determine if there is a visual match at each locus and record your results in table 2.

For each locus, place a straight-edge across the markers which run down both sides of the autorad. Find the two marker bands that are above and below the alleles in the evidence. ***Count up from the bottom*** and write the number of the marker bands on your worksheet (table 2). Look up the bins on the tables of allele counts from the reference population; this will be the column marked “between markers”. Write the frequencies (fraction) for those bins on your worksheet. If an allele seems to fall right at the position of a marker band, use whichever bin has the higher frequency. Once you have the frequency of the individual alleles, calculate the probability that a randomly chosen individual from that population would have the same combination of alleles at that locus. Use the same formula we used to calculate the probability of drawing a red and green Skittle. Record this in the far right column of table 2. **Remember that you have to multiply by two!**

**Data Table 2:**

<b>Locus</b>	<b>Allele</b>	<b>Matches suspect Y or N</b>	<b>Between marker ___ and marker ___</b>	<b>Frequency of allele</b>	<b>Frequency of genotype</b>
<b>DS244</b>	<b>1</b>		/		
	<b>2</b>		/		
<b>D17S79</b>	<b>1</b>		/		
	<b>2</b>		/		
<b>D10S28</b>	<b>1</b>		/		
	<b>2</b>		/		
<b>D14S13</b>	<b>1</b>		/		
	<b>2</b>		/		
<b>D4S139</b>	<b>1</b>		/		
	<b>2</b>		/		
<b>D1S7</b>	<b>1</b>		/		
	<b>2</b>		/		

According to the **rule of multiplication**, the probability that two or more independent events will occur in combination is the product of the individual probabilities. Calculate the probability for the multilocus genotype for all six loci. Do this by multiplying together the frequencies of each of the six loci. As you do this multiplication, fill in your answers in Data Table 3

**Data Table 3:**

<b>Locus</b>	<b>Frequency</b>	<b>Multi locus frequency (use scientific notation)</b>
<b>D2S44</b>		
<b>D17S79</b>		D2S44 x D17S79 =
<b>D10S28</b>		D2S44 x D17S79 x D10S28 =
<b>D14S13</b>		D2S44 x D17S79 x D10S28 x D14S13 =
<b>D4S139</b>		D2S44 x D17S79 x D10S28 x D14S13 x D4S139 =
<b>D1S7</b>		D2S44 x D17S79 x D10S28 x D14S13 x D4S139 x D1S7 =

10) State in words what the multi locus probability (for all six loci) means.

\_\_\_\_\_ out of \_\_\_\_\_ people would have this same combination of alleles by chance.

11) How does this probability compare with the size of the human population (about 6 billion)?

12) Do you think that the suspect is the person who left the sperm evidence at this crime scene?

13) In your opinion, how many loci does it take to prove “beyond a shadow of a doubt” that a person is guilty?