Simulation of population sampling and allele frequency, linkage equilibrium, and random match probability calculations

Cynthia B. Zeller[1*] and Kelly M. Elkins[1]*

*[1]Chemistry Department, Forensic Science Program, Towson University, 8000 York Road, Towson, MD 21252, *corresponding authors: czeller@towson.edu, kmelkins@towson.edu*

Abstract: Population sampling and analysis are necessary for performing DNA typing statistics. In this activity, we present a population sampling simulation using candy as a manipulative for hands-on learning to empirically develop a population database and promote calculations including allele frequency, polymorphism, linkage equilibrium, Hardy-Weinberg equilibrium, and random match probability. Candy of different types are assigned to genetic loci and colors are assigned to represent allelic variations. The activity has been used for several years in a Forensic Molecular Biochemistry course to aid in teaching this topic.

Keywords: Forensic Science, Population, Statistics, Allele Frequency, Linkage Equilibrium, Random Match Probability, Candy, Undergraduate/Graduate, Analogies/Transfer, Hands-On Learning/Manipulatives

.

Introduction

The topic of a population database (1) used in forensic statistics calculations is abstract and therefore difficult to conceptualize without data. Creating population databases is essential to estimating the rarity of a genetic profile for an individual – such as a perpetrator of a crime or missing person – in the relevant population. Allele frequencies are used in random match probability (RMP) and likelihood ratio (LR) statistics. Students can be introduced to calculations using published population databases to analyze their laboratory data (2) but still may not conceptualize how to prepare and test a population database.

Manipulatives have been demonstrated to help students learn difficult genetics concepts. For example, socks have been used to teach mitosis and meiosis (3). Candy has been used to teach Punnett Squares (4) and genetic drift (5). Relatedly, candy has been used to build a DNA model (6), chocolate has been employed in a crime scene investigation activity (7), and one black candy sprinkle in a million candy sprinkles has been used to demonstrate the concept of one part per million (ppm) (8).

In this activity, students are provided candy of different types and colors to represent different genetic loci and alleles at those loci to teach difficult concepts including creating a population database and computing statistics for allele frequency, minimum allele frequency (MAF), linkage equilibrium, Hardy-Weinberg equilibrium (HWE), and RMP for an individual profile.

Materials and Methods

Materials

The instructor should obtain 5-6 packages of candies, snack bags and cups before the classroom session. The use of fun size packs of each type is a mechanism to reduce contamination. Care should be taken prior to the exercise to survey students to avoid potential allergens that may affect the particular population of students in the class.

In the example in this paper, Good & Plenty® Licorice Candy, Smarties®, Skittles®, Starburst, Plain M&M's®, and chocolate chips were obtained. Other candy products such as Dum Dums Lollipops, Gumballs, Dots, Mentos, Mike and Ike, and gummi bears have also been used and other regional candy varieties could be substituted.

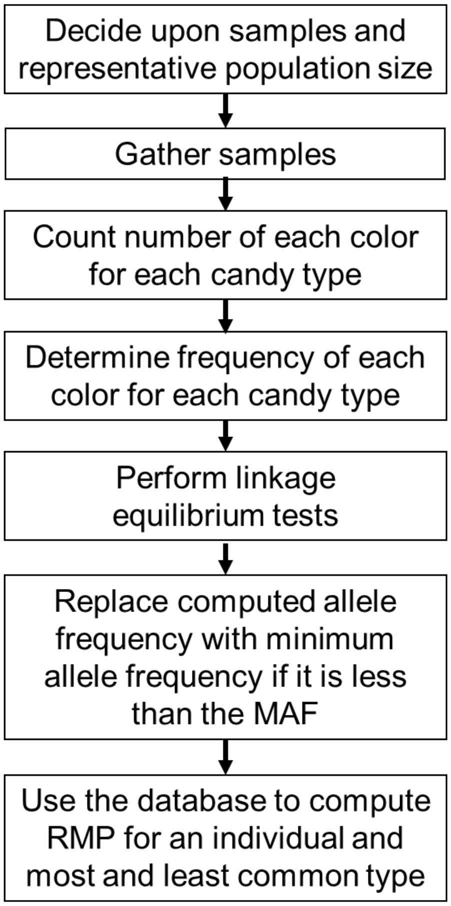
Population Simulation

To simulate the population sampling process, the students are instructed to take two of each candy and place them in a snack bag. The instructor asks each student to describe the candy type and color contents in their bag and tabulates them in a chart on a chalkboard or whiteboard or in a shared spreadsheet using an overhead projector. For smaller class sizes, it is recommended that each student assemble two or three bags in order to obtain more samples.

Results

An overview of the process for developing the population database and statistical tests is shown in Figure

1. The instructor provided Good & Plenty®, Skittles®, M&M's®, Starburst, and Smarties® candy. The instructor filled containers with the different candy types and placed them at each student's seat. Alternatively, a bag or box of each candy type could be placed on a table at the front of the room or the students could be provided snack size bags. The candy types are used to represent genetic loci. The colors are used to represent alleles.


Decide upon samples and representative population size

Gather samples

Count number of each color for each candy type

Determine frequency of each color for each candy type

Perform linkage equilibrium tests

Replace computed allele frequency with minimum allele frequency if it is less than the MAF

Use the database to compute RMP for an individual and most and least common type

FIGURE 1 Process for developing a population database and using it to perform statistical tests.

The composition of each bag sampled by the students represented the genetic makeup, or genotype, of an individual in the population. In a recent class, the instructor instructed the 22 students to each assemble two bags containing two of each candy type. This resulted in a total of forty-four bags representing a sampled population consisting of 44 individuals (N=44). Consistent with the diploid nature of human autosomes, each locus had two alleles.

The students reported the results for each individual in the population represented by the candy in the bag (Figure 2). The resulting tabulation of the candy types and colors represents a population database assembled with the class. It visually shows how often each allele occurs at each locus in a population and was used for genotype frequency calculations. From the tabulated data, the total number of each color for each type of candy was totaled. The allele frequency is computed by dividing the number of sampled alleles by the total number of alleles in the population.

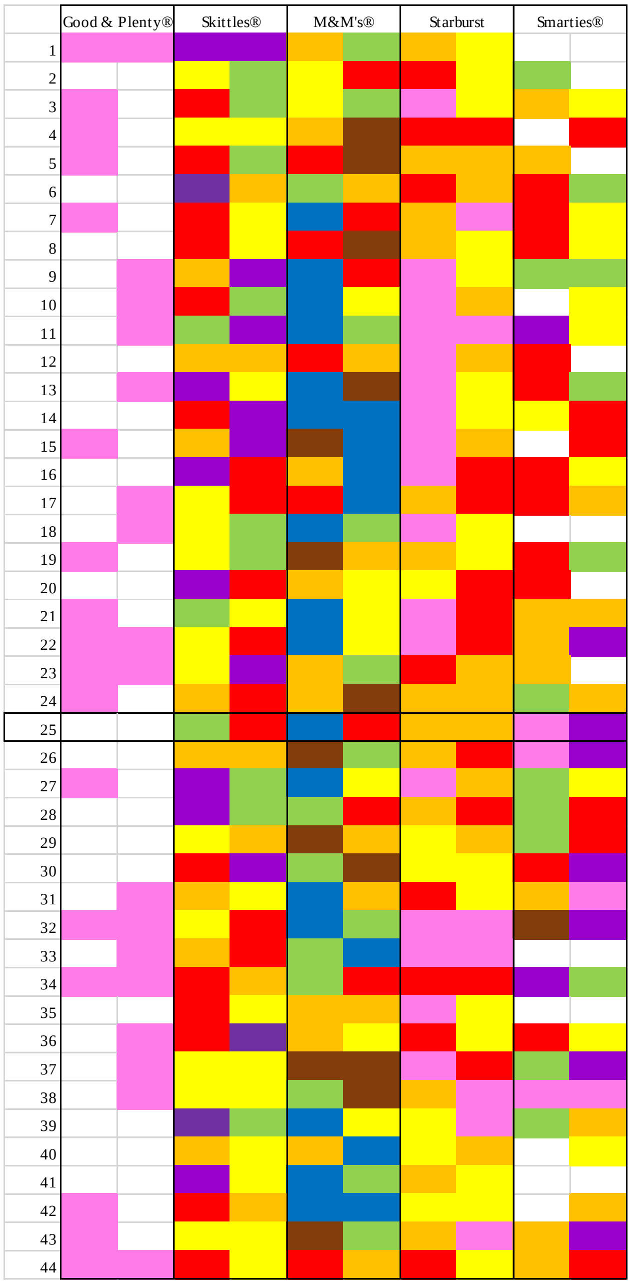


FIGURE 2 Color data for each candy by individual.

For example, the class counted 34 pink and 54 white Good & Plenty® alleles for computed frequencies of 0.38636 and 0.61364, respectively. The allele frequencies were computed for each color of each candy (Table 1A and Supplementary Information).

Next linkage equilibrium tests were performed. The linkage disequilibrium test is a likelihood ratio test to determine if two loci are inherited independently or not. The students were asked to choose gametes from Good & Plenty® candies and chocolate chips. The test compares the gamete frequency to the allele frequency from the population database and uses the constant D to determine the linkage. Data for a simple calculation based on the two allele system is shown in Table 1.

TABLE 1 Allele frequencies (A) and gamete data (B) counts and fractions for Good and Plenty® and chocolate chips (N=44) for linkage equilibrium tests

(A)

Good & Plenty®		Chocolate chips	
Pink	0.38636	Brown	0.75
White	0.61364	Ecru	0.25

(B)

Pink Brown	Pink Ecru	White Brown	White Ecru
18 (0.4090)	6 (0.1363)	15 (0.3409)	5 (0.1136)

In order to calculate the linkage equilibrium coefficient, D, the gamete frequency data is compared with the allele frequency data. If the genes, in this case the candies, are not linked then the product of the gametes formed from the most common allele of each locus and the gametes formed from the least common of the two loci minus the product of the gamete frequencies of the other two gametes should equal zero. In this example, D= (0.4090)(0.1363)-(0.3409)(0.1136) = 0.0017. Due to the limited number of gametes used in this example, the number is close to, but not zero as would be expected.

The data was also used to determine if the loci were in Hardy-Weinberg Equilibrium. A $\chi2$ test was used to test the hypothesis that the observed genotypes are the product of a random union of gametes. In order to do this, the students used the data obtained to compare the actual numbers of each genotype to the predicted numbers and determine how much difference exists (Table 2). The $\chi2$ statistic can be computed mathematically using the $\chi2$ equation or using software such as Excel or SPSS and interpreted by calculating the probability (p) value. In Excel, the function CHITEST(actual_range, expected_range) will result in the test statistic. The degrees of freedom is one less than the number of categories, or 2 in this example. A low $\chi2$ value indicates a high correlation between the actual and expected values. If the test statistic is greater than the p value using the significance level of 0.05, there is no significant departure from the HWE. Alternatively, a significance level of 0.05 indicates no significant linkage equilibrium. In Excel, CHISQ.DIST.RT(x, degree_freedom) returns the right tail P value. The computed p = 0.635092 meaning the result is not significant at $p < 0.05$. The population database should only be used for reporting if there is no significant linkage equilibrium and if the database conforms to the Hardy-Weinberg equilibrium.

TABLE 2 Example of Hardy-Weinberg Equilibrium chi square test using the Good & Plenty® data

Genotype	Actual	Expected
pp	6	6.5682
pw	22	20.8636
ww	16	17.1875
Chi square		0.9080
Probability (right-tail)		0.6351

The database is reviewed and the MAF is assigned for any allele for which the sampled allele frequency is lower than the MAF. The MAF compensates for the sampling of rare alleles in the population database due to small sampling size. A standard MAF method requires that a minimum of 5 copies of an allele are used for the allele frequency calculation. The MAF is computed by dividing 5 by 2N where N is the number of individuals in the database. For example, the observed frequency for pink smarties is at the MAF threshold. The counted number of pink smarties was 5 and the number of alleles for N=44 is 88. Dividing 5 by 88 results in 0.05682 as shown in Table 3; the value of the MAF equals 5/2N or 5/(2*44).

Finally, using the computed or MAF allele frequencies at all of the candy loci in the population database and the product rule, the instructor demonstrated calculations of the RMP for each allele and for the overall genotype for one of the individuals used to build the population database. Table 3 shows the computation of the random match probability using HWE for a selected set of candy loci corresponding to individual 25 in Figure 2.

TABLE 3 RMP calculation of individual 25 from the population database

Locus	allele	frequency	allele	frequency	probability	1 in	Combined
Good & Plenty®	white	0.61364	white	0.61364	0.37655	2.66	2.66E+00
Skittles®	green	0.13636	red	0.22727	0.06198	16.13	4.28E+01
M&M's®	blue	0.26136	red	0.12500	0.06534	15.30	6.56E+02
Starburst	orange	0.27273	orange	0.27273	0.07438	13.44	8.82E+03
Smarties®	pink	0.05682	purple	0.10227	0.01162	86.04	7.59E+05

The allele frequencies in the population database sum to 1 as shown in Table 1A. The alleles in an individual are represented by p and q. If there are two events in the probability space, p+q=1 is represented by $(p+q)^2=1$ and binomial expansion results in $p^2+2pq+q^2=1$. Thus, the genotype probability of the allele combination at each locus is computed using 2pq for heterozygotes and p^2 or q^2 for homozygotes. For each locus, the rarity of the probability is shown by taking the reciprocal of the probability to yield the 1 in value. The combined RMP is shown in the last column and is computed using the product rule. The RMP for this profile using the population database is 1 in 0.759 million. For comparison, the most common genotype is computed using the highest frequency alleles at each locus and the least common genotype could be computed using the MAF for each locus. These values bookend the range of RMP values for the theoretical least and most common genotypes using the loci and population database.

Discussion

Population databases enable criminalists to determine how often a particular genotype may be encountered and how the allele frequencies determined using population database data are used to compute statistics including RMP and LR. The database can be used to determine the RMP for a hypothetical individual, the most common genotype and the least common genotype.

The time needed for the activity will vary depending upon class size, number of individuals tabulated for the population, the number of candies used, and how many RMP calculations are performed. Typically, a population database will have at least 100 individuals or more. The activity may require a 50-minute class period or more if the instructor demonstrates many examples or if the class size is large. The population database described in this paper is small and a similar database may result in loci that are not in linkage equilibrium or HWE.

The activity engages the students in a hands-on activity to demonstrate an otherwise abstract concept of a population database. Furthermore, the students practice the following terms: individual, population, locus, allele, genotype, minimum allele frequency, linkage equilibrium, and random match probability. The Forensic Molecular Biochemistry (FRSC 601) course at Towson University is enrolled by 12-24 students each year. The activity can be performed with small and large classes. Preparing more candy bags to represent individuals should more closely represent the allele frequency in the population. Care must be taken to correctly identify the color of the candy; color-blindness may alter the perceived and reported color. The candy materials used in this activity are inexpensive, accessible, and non-toxic. The students are more engaged when working with a manipulative, especially an edible one such as candy. After the data is collected, the students are invited to indulge in eating the candy. The activity poses no chemical hazards although students may report a stomachache if they eat too much of the candy.

This activity can also be used to teach inclusion/exclusion of individuals to K-12 students. For example, sixth grade students were asked to take two Good & Plenty® candy. Approximately half of the class selected each color. Then, they selected two Skittles® candy. The class was asked who had both a pink Good & Plenty® and a red Skittles® candy and the students were engaged in a discussion of how many individuals were included and excluded. Finally, each student selected a Dum Dums Lollipop. The students were asked if anyone had a pink Good & Plenty®, red Skittles® and a butterscotch Dum Dums. This activity demonstrates that it is important to use several “traits” in the analysis.

Conclusion

The activity engages students with candy manipulatives to teach abstract genetics concepts and terminology by creating a population database with a class and computing genotype statistics.

Acknowledgements

The authors thank the Towson University students that enrolled in FRSC 601 from 2007-2019 and North Harford middle school sixth and eighth grade classes from 2007-2009 for their participation.

References

1. Butler JM. Forensic DNA Typing. 2nd ed. Burlington, MA: Elsevier Academic Press, 2005.
2. Elkins KM. Forensic DNA Biology: A Laboratory Manual. Waltham, MA: Elsevier Academic Press, 2013.
3. Chinnici JP, Neth SZ, Sherman LR. Using “Chromosomal Socks” To Demonstrate Ploidy in Mitosis & Meiosis. The American Biology Teacher 2006;68(2): 106-109.
4. Baker WP, Thomas CL. Gummy Bear Genetics. The Science Teacher 1998:25-27.
5. Staub NL. Teaching Evolutionary Mechanisms: Genetic Drift and M&M's®. BioScience 2002;52(4):373–77.
6. Have your DNA and Eat it Too. https://teach.genetics.utah.edu/content/dna/HaveYourDNAandEatItToo.pdf (accessed April 20, 2020).
7. Marle PD, Decker L, Taylor V, Fitzpatrick K, Khaliqi D, Owens JE, Henry RM. CSI–Chocolate Science Investigation and the Case of the Recipe Rip-Off: Using an Extended Problem-Based Scenario To Enhance High School Students’ Science Engagement. J Chem Educ 2014;91(3):345-50.
8. Meloan ML, Meloan JM, Meloan CE. Candy Sprinkles To Illustrate One Part Per Million. J Chem Educ 1994;71(8)658.

Supplementary Information

Population Database Allele Frequency Tables

Good & Plenty®	Simulated population (N=44)
Pink	0.38636
White	0.61364

Skittles®	Simulated population (N=44)
Purple	0.19318
Yellow	0.28409
Red	0.22727
Green	0.13636
Orange	0.15909

M&M's®	Simulated population (N=44)
Brown	0.15909
Yellow	0.10227
Red	0.12500
Green	0.17045
Orange	0.18182
Blue	0.26136

Starburst	Simulated population (N=44)
Pink	0.27273
Yellow	0.25000
Red	0.20455
Orange	0.27273

Smarties®	Simulated population (N=44)
White	0.22727
Yellow	0.11364
Red	0.19318
Green	0.14773
Orange	0.14773
Pink	0.05682
Purple	0.10227
Brown	0.01136