Where did my M&Ms come from?

Bryant McAllister
University of Iowa

This classroom exercise demonstrates the logic behind the modeling applied to estimate ancestry composition from an individuals' SNP data.

Learning Objectives

- Identify the discriminating information provided by allele frequency differences among reference populations for ancestry predictions.
- Evaluate evidence to make predictions of potential source populations of an individual.
- Develop the concept that an estimate of ancestry composition including admixture is produced from fine-grained fitting of many SNP markers across multiple populations.

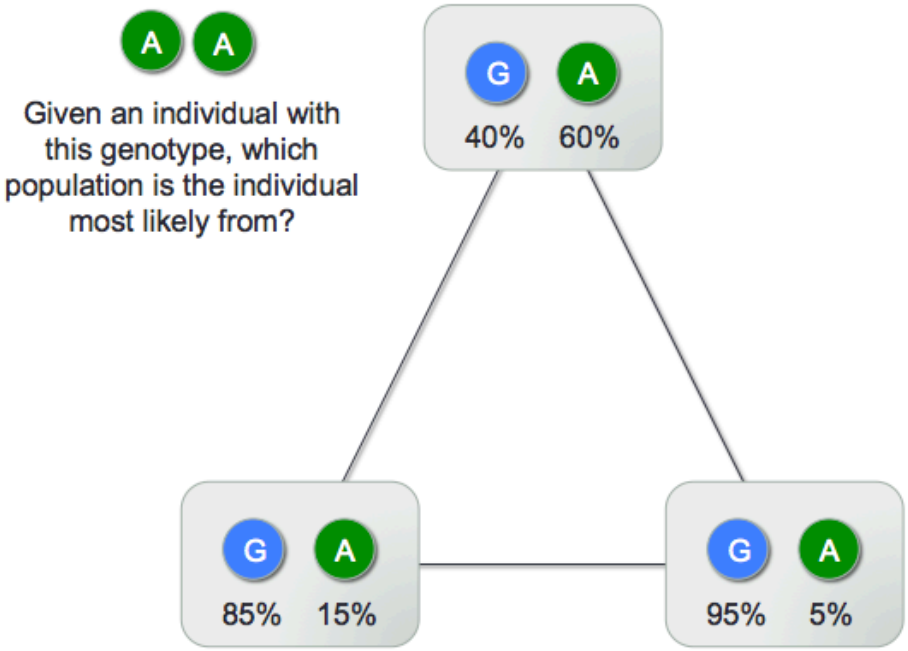
Materials

A bag of M&Ms (or another button-sized candy with four or more color varieties)
Small bags for populations of M&Ms
Marker for labeling plastic bags

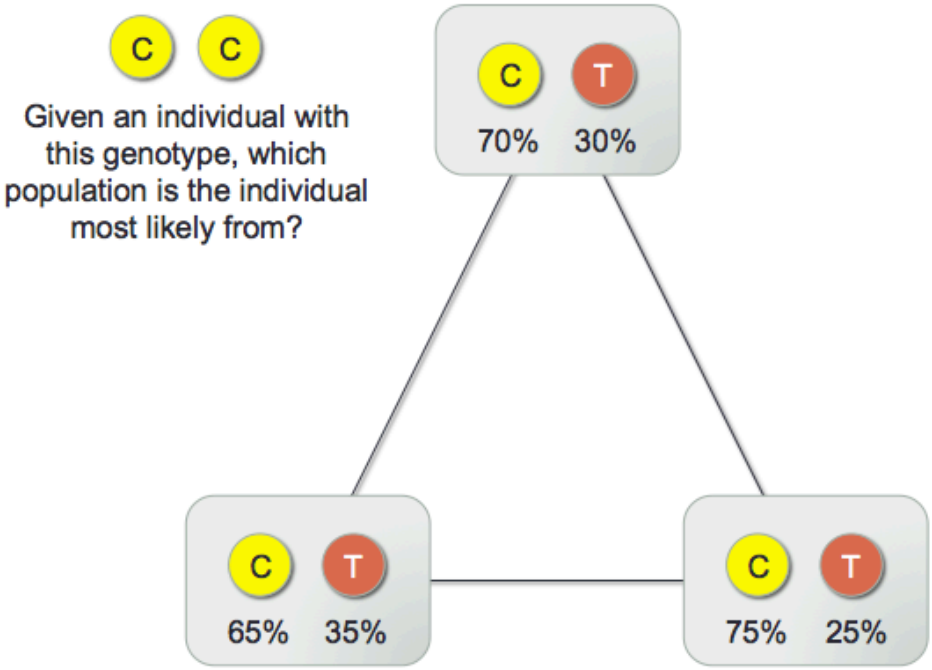
Methods (this example is used in a class of 18 students sitting at 3 tables)
Construct bags each with 20 M&Ms representing allele frequencies at six different SNPs in three different populations. Label each bag with the SNP ID (e.g., SNP 1) and population ID (e.g., Pop A). The following table presents a possible construction of three populations.

	Population A		**Population B**		**Population C**	
SNP 1	**blue**	**green**	**blue**	**green**	**blue**	**green**
	85%	15%	40%	60%	95%	5%
	17	3	8	12	19	1
SNP 2	**yellow**	**red**	**yellow**	**red**	**yellow**	**red**
	65%	35%	70%	30%	75%	25%
	13	7	14	6	15	5
SNP 3	**green**	**yellow**	**green**	**yellow**	**green**	**yellow**
	100%	0%	30%	70%	100%	0%
	20	0	6	14	20	0
SNP 4	**blue**	**red**	**blue**	**red**	**blue**	**red**
	5%	95%	40%	60%	95%	5%
	1	19	8	12	19	1
SNP 5	**green**	**blue**	**green**	**blue**	**green**	**blue**
	15%	85%	90%	10%	20%	80%
	3	17	18	2	4	16
SNP 6	**yellow**	**red**	**yellow**	**red**	**yellow**	**red**
	45%	55%	0%	100%	100%	0%
	9	11	0	20	20	0

Students should be familiar with the concept that 23andMe data represent bi-allelic SNPs distributed throughout the genome. Introduce students to the concept of predicting the origins of an individual based on their genotype compared to a set of reference populations by focusing on a single SNP. For example, the image below poses the question of which population an individual with an AA genotype (i.e., two green M&Ms) at SNP 1 most likely originated considering the allele frequencies observed in the reference populations.



In this case, because of the high frequency of the A allele, the population with 60% A alleles is the most likely source of the individual. The image below illustrates SNP 2 – one with similar allele frequencies across populations – so that the origin of the individual is unclear.



Each group of students will need to examine the frequencies of different color M&Ms in the six bags representing each SNP in a population. These frequencies can be used to predict which of three individuals below most likely originated from their population.

Which individual most likely originated from your candies?

Individual	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6
John	A A	C C	A C	G T	A A	T T
Kelly	G G	C T	A A	G G	G G	C C
Mark	A A	C T	A A	T T	G G	C T

Depending on the level of students, different approaches may be applied to make the inference. Entry-level students may simply examine the frequencies finding SNPs with dramatically different frequencies, or absence of an allele, to make the inference. For example, John doesn't appear to originate from population A due to his AC genotype at SNP 3, and yellow M&Ms (C alleles) are absent at SNP 3 in population A. Both Kelly and Mark could originate from population A; however, Mark's genotype fits better than Kelly's with the allele frequencies of population A. Advanced students could model this difference applying a likelihood approach and using allele frequencies to develop a probability of the observed composite genotypes in the context of the Hardy Weinberg model.

The most likely sources of the individuals follow:
John – Population B
Kelly – Population C (less likely from Population A)
Mark – Population A

Points of Discussion
Discuss admixture as a way to infer origins from multiple populations due to SNPs in different chromosomal regions exhibiting likely origins from different reference population.

Use the Global Similarity Map in the 23andMe Ancestry Tools to display allele composition of the reference populations in a 2-dimensional image. Placement of an individual's sample is inferred within the context of the spatial display of reference individuals.