

# TWO-LOCUS POPULATION GENETICS

## Introduction

So far in this course we've dealt only with variation at a single locus. There are obviously many traits that are governed by more than a single locus in whose evolution we might be interested. And for those who are concerned with the use of genetic data for forensic purposes, you'll know that forensic use of genetic data involves genotype information from multiple loci. I won't be discussing quantitative genetic variation for a few weeks, and I'm not going to say anything about how population genetics gets applied to forensic analyses, but I do want to introduce some basic principles of multilocus population genetics that are relevant to our discussions of the genetic structure of populations before moving on to the next topic. To keep things relatively simple *multilocus* population genetics will, for purposes of this lecture, mean *two-locus* population genetics.

## Gametic disequilibrium

One of the most important properties of a two-locus system is that it is no longer sufficient to talk about allele frequencies alone, even in a population that satisfies all of the assumptions necessary for genotypes to be in Hardy-Weinberg proportions at each locus. To see why consider this. With two loci and two alleles there are four possible gametes:<sup>1</sup>

Gamete	$A_1B_1$	$A_1B_2$	$A_2B_1$	$A_2B_2$
Frequency	$x_{11}$	$x_{12}$	$x_{21}$	$x_{22}$

If alleles are arranged randomly into gametes then,

$$\begin{aligned}x_{11} &= p_1p_2 \\x_{12} &= p_1q_2 \\x_{21} &= q_1p_2 \\x_{22} &= q_1q_2 \quad ,\end{aligned}$$

---

<sup>1</sup>Think of drawing the Punnett square for a dihybrid cross, if you want.

where  $p_1 = \text{freq}(A_1)$  and  $p_2 = \text{freq}(A_2)$ . But alleles need not be arranged randomly into gametes. They may covary so that when a gamete contains  $A_1$  it is more likely to contain  $B_1$  than a randomly chosen gamete, or they may covary so that a gamete containing  $A_1$  is less likely to contain  $B_1$  than a randomly chosen gamete. In that case

$$\begin{aligned}x_{11} &= p_1p_2 + D \\x_{12} &= p_1q_2 - D \\x_{21} &= q_1p_2 - D \\x_{22} &= q_1q_2 + D \quad ,\end{aligned}$$

where  $D = x_{11}x_{22} - x_{12}x_{21}$  is known as the *genetic disequilibrium*.<sup>2</sup>  $D \neq 0$  does not imply that the loci are linked in the physical, genetic sense, only that the alleles at the two loci are not arranged randomly into gametes.  $D$  measures *statistical* association, not *physical* association. Similarly,  $D = 0$  does not imply that the loci are unlinked, only that the alleles at the two loci are arranged into gametes independently of one another.

## A little diversion

It probably isn't obvious why we can get away with only one  $D$  for all of the gamete frequencies. The short answer is:

There are four gametes. That means we need three parameters to describe the four frequencies.  $p_1$  and  $p_2$  are two.  $D$  is the third.

Another way is to do a little algebra to verify that the definition is self-consistent.

$$\begin{aligned}D &= x_{11}x_{22} - x_{12}x_{21} \\&= (p_1p_2 + D)(q_1q_2 + D) - (p_1q_2 - D)(q_1p_2 - D) \\&= (p_1q_1p_2q_2 + D(p_1p_2 + q_1q_2) + D^2) \\&\quad - (p_1q_1p_2q_2 - D(p_1q_2 + q_1p_2) + D^2) \\&= D(p_1p_2 + q_1q_2 + p_1q_2 + q_1p_2) \\&= D(p_1(p_2 + q_2) + q_1(q_2 + p_2)) \\&= D(p_1 + q_1) \\&= D \quad .\end{aligned}$$

---

<sup>2</sup>You will sometimes see  $D$  referred to as the linkage disequilibrium, but that's misleading. Alleles at different loci may be non-randomly associated even when they are not linked.

## Transmission genetics with two loci

I'm going to construct a reduced version of a mating table to see how gamete frequencies change from one generation to the next. There are ten different two-locus genotypes (if we distinguish coupling,  $A_1B_1/A_2B_2$ , from repulsion,  $A_1B_2/A_2B_1$ , heterozygotes as we must for these purposes). So a full mating table would have 100 rows. If we assume all the conditions necessary for genotypes to be in Hardy-Weinberg proportions apply, however, we can get away with just calculating the frequency with which any one genotype will produce a particular gamete.<sup>3</sup>

Genotype	Frequency	Gametes			
		$A_1B_1$	$A_1B_2$	$A_2B_1$	$A_2B_2$
$A_1B_1/A_1B_1$	$x_{11}^2$	1	0	0	0
$A_1B_1/A_1B_2$	$2x_{11}x_{12}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
$A_1B_1/A_2B_1$	$2x_{11}x_{21}$	$\frac{1}{2}$	$0\frac{1}{2}$	0	0
$A_1B_1/A_2B_2$	$2x_{11}x_{22}$	$\frac{1-r}{2}$	$\frac{r}{2}$	$\frac{r}{2}$	$\frac{1-r}{2}$
$A_1B_2/A_1B_2$	$x_{12}^2$	0	1	0	0
$A_1B_2/A_2B_1$	$2x_{12}x_{21}$	$\frac{r}{2}$	$\frac{1-r}{2}$	$\frac{1-r}{2}$	$\frac{r}{2}$
$A_1B_2/A_2B_2$	$2x_{12}x_{22}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$A_2B_1/A_2B_1$	$x_{21}^2$	0	0	1	0
$A_2B_1/A_2B_2$	$2x_{21}x_{22}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$A_2B_2/A_2B_2$	$x_{22}^2$	0	0	0	1

### Where do $\frac{1-r}{2}$ and $\frac{r}{2}$ come from?

Consider the coupling double heterozygote,  $A_1B_1/A_2B_2$ . When recombination doesn't happen,  $A_1B_1$  and  $A_2B_2$  occur in equal frequency ( $1/2$ ), and  $A_1B_2$  and  $A_2B_1$  don't occur at all. When recombination happens, the four possible gametes occur in equal frequency ( $1/4$ ). So the recombination frequency,<sup>4</sup>  $r$ , is half the crossover frequency,<sup>5</sup>  $c$ , i.e.,  $r = c/2$ . Now the results of crossing over can be expressed in this table:

Frequency	$A_1B_1$	$A_1B_2$	$A_2B_1$	$A_2B_2$
$1 - c$	$\frac{1}{2}$	0	0	$\frac{1}{2}$
$c$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
Total	$\frac{2-c}{4}$	$\frac{c}{4}$	$\frac{c}{4}$	$\frac{2-c}{4}$
	$\frac{1-r}{2}$	$\frac{r}{2}$	$\frac{r}{2}$	$\frac{1-r}{2}$

<sup>3</sup>We're assuming random union of gametes rather than random mating of genotypes.

<sup>4</sup>The frequency of recombinant gametes in double heterozygotes.

<sup>5</sup>The frequency of cytological crossover during meiosis.

## Changes in gamete frequency

We can use this table as we did earlier to calculate the frequency of each gamete in the next generation. Specifically,

$$\begin{aligned}x'_{11} &= x_{11}^2 + x_{11}x_{12} + x_{11}x_{21} + (1-r)x_{11}x_{22} + rx_{12}x_{21} \\ &= x_{11}(x_{11} + x_{12} + x_{21} + x_{22}) - r(x_{11}x_{22} - x_{12}x_{21}) \\ &= x_{11} - rD \\ x'_{12} &= x_{12} + rD \\ x'_{21} &= x_{21} + rD \\ x'_{22} &= x_{22} - rD \quad .\end{aligned}$$

## No changes in allele frequency

We can also calculate the frequencies of  $A_1$  and  $B_1$  after this whole process:

$$\begin{aligned}p'_1 &= x'_{11} + x'_{12} \\ &= x_{11} - rD + x_{12} + rD \\ &= x_{11} + x_{12} \\ &= p_1 \\ p'_2 &= p_2 \quad .\end{aligned}$$

Since each locus is subject to all of the conditions necessary for Hardy-Weinberg to apply at a single locus, allele frequencies don't change at either locus. Furthermore, genotype frequencies at each locus will be in Hardy-Weinberg proportions. But the two-locus gamete frequencies change from one generation to the next.

## Changes in $D$

You can probably figure out that  $D$  will eventually become zero, and you can probably even guess that how quickly it becomes zero depends on how frequent recombination is. But I'd be astonished if you could guess exactly how rapidly  $D$  decays as a function of  $r$ . It takes a little more algebra, but we can say precisely how rapid the decay will be.

$$\begin{aligned}D' &= x'_{11}x'_{22} - x'_{12}x'_{21} \\ &= (x_{11} - rD)(x_{22} - rD) - (x_{12} + rD)(x_{21} + rD) \\ &= x_{11}x_{22} - rD(x_{11} + x_{12}) + r^2D^2 - (x_{12}x_{21} + rD(x_{12} + x_{21}) + r^2D^2)\end{aligned}$$

$$\begin{aligned}
&= x_{11}x_{22} - x_{12}x_{21} - rD(x_{11} + x_{12} + x_{21} + x_{22}) \\
&= D - rD \\
&= D(1 - r)
\end{aligned}$$

Notice that even if loci are unlinked, meaning that  $r = 1/2$ ,  $D$  does not reach 0 immediately. That state is reached only asymptotically. The two-locus analogue of Hardy-Weinberg is that gamete frequencies will *eventually* be equal to the product of their constituent allele frequencies.

## Population structure with two loci

You can probably guess where this is going. With one locus I showed you that there's a deficiency of heterozygotes in a combined sample even if there's random mating within all populations of which the sample is composed. The two-locus analog is that you can have gametic disequilibrium in your combined sample even if the gametic disequilibrium is zero in all of your constituent populations. To see why requires some more algebra. Let

$$\begin{aligned}
D_i &= x_{11,i}x_{22,i} - x_{12,i}x_{21,i} \\
D_t &= \bar{x}_{11}\bar{x}_{22} - \bar{x}_{12}\bar{x}_{21} \quad ,
\end{aligned}$$

where  $\bar{x}_{kl} = \frac{1}{I} \sum_{i=1}^I x_{kl,i}$ . Given this, we're now going to calculate  $\bar{D}$  the gametic disequilibrium in the combined sample.

$$\begin{aligned}
\bar{D} &= \frac{1}{k} \sum D_i \\
&= \frac{1}{k} \sum (x_{11,i}x_{22,i} - x_{12,i}x_{21,i}) \\
&= \frac{1}{k} \sum ((x_{11,i} - \bar{x}_{11} + \bar{x}_{11})(x_{22,i} - \bar{x}_{22} + \bar{x}_{22}) - (x_{12,i} - \bar{x}_{12} + \bar{x}_{12})(x_{21,i} - \bar{x}_{21} + \bar{x}_{21})) \\
&= \text{Then a miracle occurs.} \\
&= \frac{1}{k} \sum ((x_{11,i} - \bar{x}_{11})(x_{22,i} - \bar{x}_{22}) - (x_{12,i} - \bar{x}_{12})(x_{21,i} - \bar{x}_{21})) + D_t \\
&= \text{Then another miracle occurs.} \\
&= D_t - \text{Cov}(p_1, p_2) \quad ,
\end{aligned}$$

where  $\text{Cov}(p_1, p_2)$  is the covariance in allele frequencies across populations.<sup>6</sup> Suppose  $D_i = 0$  for all subpopulations. Then  $\bar{D} = 0$ , too (obviously). But that means that

$$\begin{aligned}\bar{D} = 0 &= D_t - \text{Cov}(p_1, p_2) \\ D_t &= \text{Cov}(p_1, p_2) \quad .\end{aligned}$$

So if allele frequencies covary across populations, i.e.,  $\text{Cov}(p_1, p_2) \neq 0$ , then there will be non-random association of alleles into gametes in the sample, i.e.,  $D_t \neq 0$ , even if there is random association alleles into gametes within each population.<sup>7</sup>

## Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

---

<sup>6</sup>Neither of the miracles above is particularly difficult to justify. There's just a lot of tedious algebra. If you're really interested in seeing it, ask me nicely, and I might show it to you.

<sup>7</sup>Well, duh! Covariation of allele frequencies across populations means that alleles are non-randomly associated across populations. What other result could you possibly expect?