ALLELIC INDEPENDENCE

Testing for Allelic Independence

What is the probability a person has a particular DNA profile? What is the probability a person has a particular profile if it has already been seen once?

The first question is a little easier to think about, but difficult to answer in practice: it is very unlikely that a profile will be seen in any sample of profiles. Even for one STR locus with 10 alleles, there are 55 different genotypes and most of those will not occur in a sample of a few hundred profiles.

For locus D3S1358 in the African American population, the FBI frequency database shows that 31 of the 55 genotype counts are zero. Estimating the population frequencies for these 31 types as zero doesn't seem sensible.

D3S1358 Genotype Counts

Observed	<12	12	13	14	15	16	17	18	19	>19
<12	0									
12	0	0								
13	0	0	0							
14	0	0	0	2						
15	0	0	1	19	15					
16	1	1	1	15	39	19				
17	0	0	2	10	26	24	9			
18	1	0	1	2	6	10	3	0		
19	0	0	0	1	0	0	1	0	0	
>19	0	0	0	0	1	0	0	0	0	0

The number in row i and column j is the observed count of indivuals with alleles i and j.

Hardy-Weinberg Law

A solution to the problem is to assume that the Hardy-Weinberg Law holds. For a random mating population, expect that genotype frequencies are products of allele frequencies.

For a locus with two alleles, A, a:

$$\begin{aligned} P_{AA} &= (p_A)^2 \\ P_{Aa} &= 2p_A p_a \\ P_{aa} &= (p_a)^2 \end{aligned}$$

For a locus with several alleles A_i:

$$\begin{aligned} P_{A_iA_i} &= (p_{A_i})^2 \\ P_{A_iA_j} &= 2p_{A_i}p_{A_j} \end{aligned}$$

D3S1358 Hardy-Weinberg Calculations

The allele counts for D3S1358 in the African-American sample are:

										Total
Allele	<12	12	13	14	15	16	17	18	19	>19	
Count	2	1	5	51	122	129	84	23	2	1	420

If the Hardy-Weinberg Law holds, then we would expect to see $n\tilde{p}_{13}^2 = 210 \times (5/420)^2 = 0.03$ individuals of type 13,13 in a sample of 210 individuals.

Also, we would expect to see $2n\tilde{p}_{13}\tilde{p}_{14} = 420 \times (5/420) \times (51/420) = 0.61$ individuals of type 13,14 in a sample of 210 individuals.

Other values are shown on the next slide.

D3S1358 Observed and Expected Counts

		<12	12	13	14	15	16	17	18	19	>19
<12	Obs.	0									
	Exp.	0.0									
12	Obs.	0	0								
	Exp.	0.0	0.0								
13	Obs.	0	0	0							
	Exp.	0.0	0.0	0.0							
14	Obs.	0	0	0	2						
	Exp.	0.2	0.1	0.6	3.1						
15	Obs.	0	0	1	19	15					
	Exp.	0.6	0.3	1.5	14.8	17.7					
16	Obs.	1	1	1	15	39	19				
	Exp.	0.6	0.3	1.5	15.7	37.5	19.8				
17	Obs.	0	0	2	10	26	24	9			
	Exp.	0.4	0.2	1.0	10.2	24.4	25.8	8.4			
18	Obs.	1	0	1	2	6	10	3	0		
	Exp.	0.1	0.1	0.3	2.8	6.7	7.1	4.6	0.6		
19	Obs.	0	0	0	1	0	0	1	0	0	
	Exp.	0.0	0.0	0.0	0.2	0.6	0.6	0.4	0.1	0.0	
>19	Obs.	0	0	0	0	1	0	0	0	0	0
	Exp.	0.0	0.0	0.0	0.1	0.3	0.3	0.2	0.1	0.0	0.0

Testing for Hardy-Weinberg Equilibrium

A test of the Hardy-Weinberg Law will somehow decide if the observed and expected numbers are sufficiently similar that we can proceed as though the law can be used.

In one of the first applications of Hardy-Weinberg testing in a US forensic setting:

> "To justify applying the classical formulas of population genetics in the Castro case the Hispanic population must be in Hardy-Weinberg equilibrium. Applying this test to the Hispanic sample, one finds spectacular deviations from Hardy-Weinberg equilibrium."

Lander ES. 1989. DNA fingerprinting on trial. Nature 339: 501-505.

VNTR "Coalescence"

Forensic DNA profiling initially used minisatellites, or VNTR loci, with large numbers of alleles. Heterozygotes would be scored as homozygotes if the two alleles were so similar in length that they coalesced into one band on an autoradiogram. Small alleles often not detected at all, and this is a likely cause of Lander's finding (Devlin et al, Science 249:1416-1420.) .

Considerable debate in early 1990s on alternative "binning" strategies for reducing the number of alleles (Science 253:1037-1041, 1991).

Typing has moved to microsatellites with fewer and more easily distinguished alleles, but testing for Hardy-Weinberg equilibrium continues. There are still reasons why the law may not hold.

Population Structure can Cause Departure from HWE

If a population consists of a number of subpopulations, each in HWE but with different allele frequencies, there will be a departure from HWE at the population level. This is the Wahlund effect.

Suppose there are two equal-sized subpopulations, each in HWE but with different allele frequencies, then

	Subpopn 1	Subpopn 2	Total Popn
p_A	0.6	0.4	0.5
p_a	0.4	0.6	0.5
P_{AA}	0.36	0.16	$0.26 > (0.5)^2$
P_{Aa}	0.48	0.48	$0.48 < 2(0.5)(0.5)$
P_{aa}	0.16	0.36	$0.26 > (0.5)^2$

Population Structure

Effect of population structure taken into account with the "theta-correction." Matching probabilities allow for a variance in allele frequencies among subpopulations.

$$\Pr(AA|AA) = \frac{[3\theta + (1-\theta)p_A][2\theta + (1-\theta)p_A]}{(1+\theta)(1+2\theta)}$$

where p_A is the average allele frequency over all subpopulations. We will come back to this expression.

Population Admixture

A population might represent the recent admixture of two parental populations. With the same two populations as before but now with 1/4 of marriages within population 1, 1/2 of marriages between populations 1 and 2, and 1/4 of marriages within population 2. If children with one or two parents in population 1 are considered as belonging to population 1, there is an excess of heterozygosity in the offspring population.

If the proportions of marriages within populations 1 and 2 are both 25% and the proportion between populations 1 and 2 is 50%, the next generation has

	Population 1	Population 2
P_{AA}	$0.09 + 0.12 = 0.21$	0.04
P_{Aa}	$0.12 + 0.26 = 0.38$	0.12
P_{aa}	$0.04 + 0.12 = 0.16$	0.09
	0.75	0.25

Exact HWE Test

The preferred test for HWE is an "exact" one. The test uses the conditional probability of the genotypic counts (n_{AA}, n_{Aa}, n_{aa}) given the allelic counts (n_A, n_a) and given HWE:

$$\Pr(n_{AA}, n_{Aa}, n_{aa} | n_A, n_a, \text{HWE}) = \frac{n!}{n_{AA}! n_{Aa}! n_{aa}!} \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

Reject the Hardy-Weinberg hypothesis if this probability is unusually small.

Exact HWE Test Example

Reject the HWE hypothesis if the probability of the genotypic array, conditional on the allelic array, is among the smallest probabilities for all the possible sets of genotypic counts for those allele counts.

As an example, consider $(n_{AA} = 1, n_{Aa} = 0, n_{aa} = 49)$. The allele counts are $(n_A = 2, n_a = 98)$ and there are only two possible genotype arrays:

AA	Aa	aa	$\Pr(n_{AA}, n_{Aa}, n_{aa} \mid n_A, n_a, \text{HWE})$
1	0	49	$\frac{50!}{1!0!49!}\frac{2^0 2!98!}{100!} = \frac{1}{99}$
0	2	48	$\frac{50!}{0!2!48!}\frac{2^2 2!98!}{100!} = \frac{98}{99}$

Exact HWE Test

The probability of the data on the previous slide, conditional on the allele frequencies and on HWE, is $1/99 = 0.01$. This is less than the conventional 5% significance level.

In general, the p-value is the (conditional) probability of the data plus the probabilities of all the less-probable datasets. The probabilities are all calculated assuming HWE is true.

Permutation Test

For large sample sizes and many alleles per locus, there are too many genotypic arrays for a complete enumeration and a determination of which are the least probable 5% arrays.

A large number of the possible arrays is generated by permuting the alleles among genotypes, and calculating the proportion of these permuted genotypic arrays that have a smaller conditional probability than the original data. If this proportion is small, the Hardy-Weinberg hypothesis is rejected.

Permutation Test

Mark a set of five index cards to represent five genotypes:

Card 1:	A	A
Card 2:	A	A
Card 3:	A	A
Card 4:	a	a
Card 5:	a	a

Tear the cards in half to give a deck of 10 cards, each with one allele. Shuffle the deck and deal into 5 pairs, to give five genotypes.

Permutation Test

The permuted set of genotypes fall into one of four types:

AA	Aa	aa	Number of times
3	0	2	
2	2	1	
1	4	0	

Permutation Test

Check the following theoretical values for the proportions of each of the three types, from the expression:

$$\frac{n!}{n_{AA}!n_{Aa}!n_{aa}!} \times \frac{2^{n_{Aa}} n_A! n_a!}{(2n)!}$$

AA	Aa	aa	Conditional Probability
3	0	2	$\frac{1}{21} = 0.048$
2	2	1	$\frac{12}{21} = 0.571$
1	4	0	$\frac{8}{21} = 0.381$

These should match the proportions found by repeating shufflings of the deck of 10 allele cards.

Permutation Test for D3S1358

For an STR locus, where $\{n_g\}$ are the genotype counts and $n = \sum_g n_g$ is the sample size, and $\{n_a\}$ are the alleles counts with $2n = \sum_a n_a$, the exact test statistic is

$$\Pr(\{n_g\}|\{n_a\}, \text{HWE}) = \frac{n! 2^H \prod_a n_a!}{\prod_g n_g! (2n)!}$$

where H is the count of heterozygotes.

This probability for the African American genotypic counts at D3S1358 is 0.6163×10^{-13}, which is a very small number. But it is not unusually small if HWE holds: a proportion 0.81 of 1000 permutations have an even smaller probability. We do not reject the HWE hypothesis in this case.

HWE in NIST Database

Recent work by Graffelman looked more closely at NIST database. Graffelman J, Weir BS. FSI:Genetics 58:102680 (2022)

Work prompted in part by NGS sequencing of STR alleles that revealed more alleles than length-based alleles. For TH01 there are 10 SB alleles but 8 LB alleles:

SB allele	LB allele
$(AATG)_5$	5
$(AATG)_6$	6
$(AATG)_7$	7
$(AATG)_7$-rs1051822965	
$(AATG)_8$	8
$(AATG)_9$	9
$ATTC(AATG)_8$	
$(AATG)_6ATG(AATG)_3$	9.3
$(AATG)_{10}$	10
$(AATG)_{11}$	11

Deleting Alleles

If a locus is in Hardy-Weinberg equilibrium, then deleting an allele and all the genotypes with that allele preserves HWE. For example, if the alleles are (A,B,C) and these are reduced to A,B:

Genotype	Count	Frequency	Genotype	Count	Frequency
AA	36	0.36	AA	36	4/9
AB	36	0.36	AB	36	4/9
BB	9	0.09	BB	9	1/9
AC	12	0.12			
BC	6	0.06			
CC	1	0.01			
Total	100	1.00	Total	81	1.00
A	120	0.60	A	108	2/3
B	60	0.30	B	54	1/3
C	20	0.10			
Total	200	1.00	Total	162	1.00

Combining Alleles

If a locus is in Hardy-Weinberg equilibrium, then combining two or more alleles and all the genotypes with those allele preserves HWE. For example, if the alleles are (A,B,C) and these are reduced to combine to D=A+C, B:

Genotype	Count	Frequency	Genotype	Count	Frequency
AA	36	0.36	DD	36	49/100=0.49
AB	36	0.36	DB	36	42/100=0.42
BB	9	0.09	BB	9	9/100=0.09
AC	12	0.12	DD	12	
BC	6	0.06	DB	6	
CC	1	0.01	DD	1	
Total	100	1.00	Total	100	1.00
A	120	0.60	D	140	0.7
B	60	0.30	B	60	0.3
C	20	0.10			
Total	200	1.00	Total	200	1.00

NIST Data

NIST TH01

Each allele is removed in turn and HWE tested for the remaining data. Data across all four goups was combined. Allele 9.3 appears to be the reason for HWE departures.

SB allele	p-value	LB allele	p-value
None	0.0001	None	0.0003
$(AATG)_5$	0.0001	5	0.0001
$(AATG)_6$	0.0001	6	0.0000
$(AATG)_7$	0.0036	7	0.0035
$(AATG)_7$-rs1051822965	0.0004		
$(AATG)_8$	0.0001	8	0.0001
$(AATG)_9$	0.0047	9	00030
$ATTC(AATG)_8$	0.0001		
$(AATG)_6ATG(AATG)_3$	0.2338	9.3	0.1958
$(AATG)_{10}$	0.0001	10	0.0004
$(AATG)_{11}$	0.0002	11	0.0002

TH01 Allele Frequencies

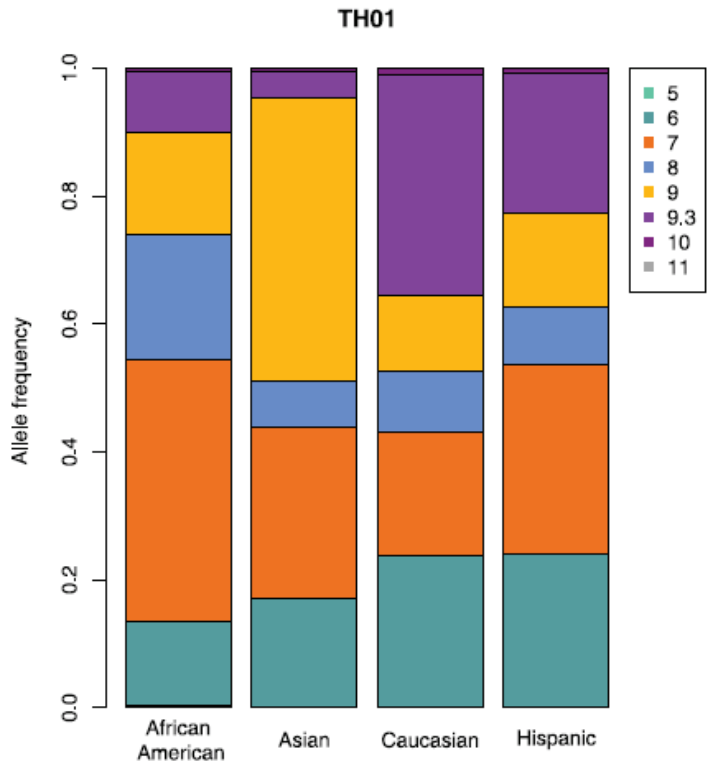


Fig. 1. Allele frequencies of TH01 for four ethnicities.

Both alleles 9 and 9.3 differ over groups. Their combination differs less, and combining those alleles makes TH01 conform to HWE.

Multiple Testing

When multiple tests are performed, each at significance level α, a proportion α of the tests are expected to cause rejection even if all the hypotheses are true.

Bonferroni correction makes the overall (experimentwise) significance level equal to α by adjusting the level for each individual test to α'. If α is the probability that at least one of the L tests causes rejection, it is also 1 minus the probability that none of the tests causes rejection:

$$\begin{aligned} \alpha &= 1 - (1 - \alpha')^L \\ &\approx L\alpha' \end{aligned}$$

provided the L tests are independent.

If $L = 10^6$, the "genome-wide significance level" is 5×10^{-8} in order for $\alpha = 0.05$.

All NIST Loci

Testing for HWE at all 29 STR loci in all four groups gives a multiple testing problem: the usual Bonferroni procedure of requiring p-values to be less than 0.05/(29x4)=0.0004 is very conservative and may obscure real HWE departures. Better to test with combined data but conducting permutations separately within each group still shows all 29 loci to conform to HWE overall.

Locus SE33 for Hispanics and African Americans, and locus D22S1045 for Asians, do show apparent departure from HWE because of unexpected homozygotes for rare alleles. This suggests some of these homozygotes may actually be heterozygotes. There have been reports of heterozygote imbalance for D22S1045.

Linkage Disequilibrium

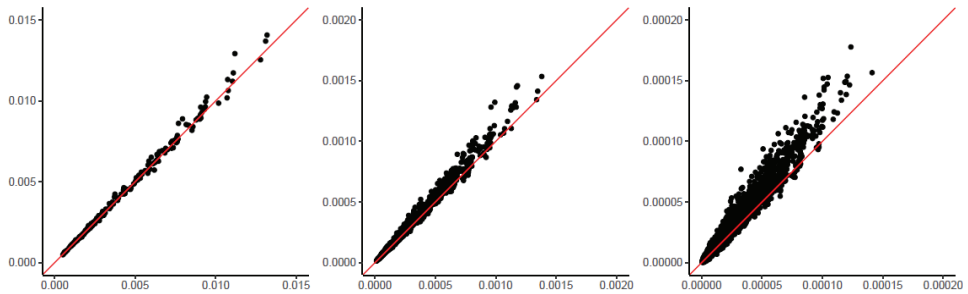
This term is generally reserved for association between pairs of alleles – one at each of two loci. In the present context, it may simply mean some lack of independence of profile or match probabilities at different loci.

Unlinked loci are expected to be almost independent.

However, if two profiles match at several loci this may be because they are from the same, or related, people and so are likely to match at additional loci.

Linkage Disequilibrium

We have examined a set of 2849 20-locus profiles constructed by merging the NIST 1036 set with 1813 FBI profiles, after checking for duplicates. For each set of 2-,3- or 4 loci we compared the proportion of matching pairs of the four million or so pairs of multilocus profiles with the products of the corresponding one-locus matching proportions. This figure shows that the product over loci clearly does less well with more loci.



Two loci Three Loci Four Loci

Multi-locus match proportions (Y-axis) vs products of single-locus proportions (X-axis).