

# Exact Tests for Association Between Alleles at Arbitrary Numbers of Loci <sup>1</sup>

D. Zaykin<sup>1</sup>, L. Zhivotovsky<sup>2</sup> and B.S. Weir<sup>1</sup>

<sup>1</sup>Program in Statistical Genetics, Department of Statistics  
North Carolina State University, Raleigh NC 27695-8203

<sup>2</sup>Institute of Gene Biology, Russian Academy of Sciences  
34/5 Vavilov Street, Moscow 117334, RUSSIA

Received 1 July, 1994 Accepted 1 December, 1994

February 23, 2006

<sup>1</sup>Published in GENETICA 96:169–178, 1995

## **Abstract**

Associations between allelic frequencies, within and between loci, can be tested for with an exact test. The probability of the set of multi-locus genotypes in a sample, conditional on the allelic counts, is calculated from multinomial theory under the hypothesis of no association. Alleles are then permuted and the conditional probability calculated for the permuted genotypic array. The proportion of arrays no more probable than the original sample provides the significance level for the test. An algorithm is provided for counting genotypes efficiently in the arrays, and powers of the test presented for various kinds of association. The powers for the case when associations are generated by admixture of several populations suggest that exact tests are capable of detecting levels of association that would affect forensic calculations to a significant extent.

**Keywords:** Exact tests, allelic association, Hardy-Weinberg, Linkage disequilibrium.

## INTRODUCTION

In the absence of evolutionary forces such as drift, selection, migration or mutation, genotypic frequencies are expected to be given by the products of corresponding allelic frequencies. Even if these forces are known to be present, however, it may be that genotypic frequencies are very close to the allelic frequency products. There are situations when it is convenient to be able to invoke this “product rule” for multilocus genotypes, as in the use of genetic profiles for human identification. A specific multilocus genotype is unlikely to have been seen in samples collected for the purpose of estimating frequencies, even though all the constituent alleles are present, and then the product rule offers a means of providing an estimate. Of course, it is necessary first to test for consistency of genotype frequencies to products of allele frequencies and such tests are covered in this paper.

With many loci and many alleles per locus, there are very many possible associations among the frequencies of subsets of the alleles. Even for two alleles at two loci, for example, there are six pairs of genes, four triples and one set of four genes to be considered (Weir and Cockerham, 1989). If the only issue is whether allele frequencies can be used to construct genotype frequencies, all these associations are tested for simultaneously in a single test. This has the advantage of avoiding problems with multiple tests, and the power of the test sometimes increases with the number of loci and alleles per locus. If there is interest in some of the individual associations, then specific tests can be constructed, and will be discussed here.

The primary purpose of this paper is to examine the use of “exact” tests, meaning tests based on the probabilities of sets of alleles conditional on observed counts of subsets of the alleles. In general these tests are expected to perform well and to avoid the problems faced by chi-square goodness-of-fit tests when expected numbers are small. Because of the large number of possible multi-allelic arrays, it is not possible to examine them all to compute significance levels for the tests. Samples of arrays are generated by permutation, following the suggestion of Guo and Thompson (1992).

## GENERAL METHOD

The general hypothesis is that there is no association among the frequencies of constituent genes of a genotype. For locus  $l$ , each of the two alleles received by an individual at that locus has probability  $p_{l_i}$  of being allelic type  $A_{l_i}$ . If  $P_{1_i 1_j, 2_i 2_j, \dots, L_i L_j}$  is the population frequency of the genotype  $A_{1_i} A_{1_j} A_{2_i} A_{2_j} \cdots A_{L_i} A_{L_j}$ , then the hypothesis can be expressed as

$$P_{1_i 1_j, 2_i 2_j, \dots, L_i L_j} = 2^H \prod_l p_{l_i} p_{l_j} \quad (1)$$

where  $H$  is the number of loci that are heterozygous. When only one locus is being considered and  $L = 1$ , Equation 1 is just the Hardy-Weinberg law. In other cases, rejection of the hypothesis does not indicate whether it is allelic frequencies within or between loci that are associated.

The development of a testing strategy is based on the multinomial distribution, meaning that each member of a population is assumed to be equally likely to be sampled, and that there is the same probability for each sample member having a particular genotype. Sample sizes are therefore considered to be very much smaller than the population sizes. If  $\mathbf{A}$  indicates a multilocus genotype, and  $n_{A_g}$  is the number of individuals of type  $\mathbf{A}_g$  in a sample of size  $n$  from a population in which those genotypes have frequency  $P_{A_g}$ , then

$$\Pr(n_{A_1}, n_{A_2}, \dots, n_{A_G}) = \frac{n!}{\prod_{g=1}^G n_{A_g}!} \prod_{g=1}^G (P_{A_g})^{n_{A_g}}$$

The quantity  $G$  is the number of different genotypes possible.

Under the null hypothesis of Equation 1, allelic counts  $n_{l_i}$  at locus  $l$  are multinomially distributed with sample size  $2n$  and probabilities  $p_{l_i}$ . Furthermore, under the hypothesis these distributions are independent over loci. The joint probability of the sets of allelic counts  $\{n_{l_i}\}$  is therefore

$$\Pr(\{n_{l_1}\}, \{n_{l_2}\}, \dots, \{n_{l_L}\}) = \prod_{l=1}^L \left( \frac{(2n)!}{\prod_i n_{l_i}!} \prod_i (p_{l_i})^{n_{l_i}} \right)$$

and the probability of the genotypic counts conditional on the allelic counts is

$$\Pr(\{n_{A_g}\} | \{n_{l_i}\}) = \frac{n!}{\prod_{g=1}^G n_{A_g}!} \prod_{g=1}^G (P_{A_g})^{n_{A_g}} \prod_{l=1}^L \frac{\prod_i n_{l_i}!}{(2n)! \prod_i (p_{l_i})^{n_{l_i}}}$$

Under the null hypothesis of complete independence of allele frequencies

$$\Pr(\{n_{A_g}\} | \{n_{l_i}\}) = \frac{n! \prod_{g=1}^G 2^{n_{A_g} H_g}}{\prod_{g=1}^G n_{A_g}!} \prod_{l=1}^L \frac{\prod_i n_{l_i}!}{(2n)!} \quad (2)$$

where count  $H_g$  is the number of heterozygous loci in genotype  $\mathbf{A}_g$ , of which there are  $n_{A_g}$  copies. Note that the unknown allelic frequencies  $p_{l_i}$  have cancelled out of this expression.

Genotypic arrays  $\{n_{A_g}\}$ , generated by permutation, with conditional probabilities equal to or less than that of the observed sample array contribute to the probability with which the null hypothesis would be rejected if it was true. This is the significance level, or  $p$ -value. In general, it is not feasible to calculate this quantity exactly because of the prohibitively large number of genotypic count arrays for a given array of allelic counts.

Gail and Mantel (1977) discussed methods for determining the numbers of two- and three-dimensional contingency tables with fixed marginals. Another approximate method for setting a lower bound on the number will suffice to show that the number is indeed prohibitive. If  $P_1$  is the largest value of all the array probabilities, then the number of arrays must be at least  $1 + (1 - P_1)/P_1 = 1/P_1$ . This would be the actual number if all arrays had probability  $P_1$ . Similarly, if  $P_1$  is the smallest of the  $T$  largest probabilities, the number of arrays must be at least  $T + (1 - \sum_{i=1}^T P_i)/P_1$ . For cases when all these  $P$ 's are small, the lower bound is given essentially by  $(1 - \sum_{i=2}^T P_i)/P_1$ . Since the ordering of all possible  $P$ 's is unknown, this bound is calculated by finding the set of  $T$  largest  $P$ 's for a large number of permuted arrays. Having  $T > 1$  protects against under-estimation. The procedure was applied to STR data from a sample of 182 people typed at loci with 6, 6, 9 and 14 alleles. With  $T = 1$  and 65,000 permutations, the lower bound was estimated to be of the order of  $10^{748}$  in each of three separate determinations. Obviously, it is not possible to examine all possible arrays.

Instead of identifying all arrays with lower conditional probabilities than the sample, a set of arrays is generated randomly by permuting those alleles hypothesized to be independent. For the hypothesis in Equation 1, this means that alleles are permuted among individuals within loci, and independent permutations performed for each locus. The proportion of permuted arrays as probable or less probable than the sample forms an estimate of the significance level. If the true significance level is  $\alpha$ , then with probability 0.95 the estimate will be within  $\delta$  of that value after  $m$  permutations if  $m \approx 4\alpha(1 - \alpha)/\delta^2 \leq 1/\delta^2$ . Hence, with 95% probability, 10,000 permutations give an estimate accurate to two decimal places. Further discussion of this approach was given by Guo and Thompson (1992).

Applying the permutation method to estimate significance levels can be performed very efficiently. As all the arrays have the same allelic counts, for purposes of comparison it is necessary to compute only

$$P_s = \prod_g \frac{2^{n_{A_g} H_g}}{n_{A_g}!} \quad (3)$$

although it is the logarithm of  $P_s$  that is computed in practice, using an algorithm of Press et al. (1988). Furthermore, it is not necessary to step through all possible multilocus genotypes, since only those with non-zero counts contribute to Equation 3. From now on, this will be indicated by  $gez$ , meaning that only those  $g$  values for which  $n_{A_g} > 0$  are considered. The computer storage requirements therefore depend on the sample size rather than the number of possible genotypes.

## Algorithm

The greatest saving in computing time is made in the way of counting the number of times each genotype appears in one of the genotypic arrays generated by permuting alleles. A naïve way would be to assign each genotype a numerical identifier, sort the identifiers and then count how many times each one occurs. For locus **B** with alleles numbered 1 to  $n_B$ , a possible identifier  $s_B$  for genotype  $B_i B_j$ ,  $j \leq i$  is

$$s_B = \frac{i(i-1)}{2} + j$$

Values of this quantity range from 1 for  $B_1 B_1$  to  $S_B = n_B(n_B + 1)/2$  for  $B_{n_B} B_{n_B}$ . If there is a second locus **C** with  $n_C$  alleles, then the identifier  $s_C$  ranging from 1 to  $S_C$  for genotypes at that locus can be defined similarly, and the identifier for two-locus genotypes is defined by

$$s_{BC} = S_C(s_B - 1) + s_C$$

which ranges from 1 for  $B_1 B_1 C_1 C_1$  to  $S_B S_C$  for  $B_{n_B} B_{n_B} C_{n_C} C_{n_C}$ . The extension to multiple loci is straightforward. If necessary, the genotype can be recovered from the identifier. In the two-locus case, for example,

$$s_C = \frac{i_C(i_C - 1)}{2} + j_C = s_{BC} \pmod{S_C}$$

$$s_B = \frac{i_B(i_B - 1)}{2} + j_B = \frac{(s_{BC} - s_C)}{S_C} + 1$$

and the one-locus genotypes can be recovered from the one-locus identifiers by

$$\begin{aligned} i &= 1 + \text{integer part of } [(\sqrt{8s_B + 1} - 1)/2] \\ j &= s_B - i(i - 1)/2 \end{aligned}$$

The problem with this approach is the need to sort as many identifiers as there are individuals in the sample. The binary search tree method now described is very much faster.

The data set  $\mathbf{D}$  has an element for every individual in the sample. A tree is constructed with nodes for every distinct genotype in  $\mathbf{D}$ . Each element of  $\mathbf{D}$  is placed on the tree, either at an existing node or at a new node, by comparing its identifier to the identifiers for the previously placed elements. This placement procedure begins at the root node of the tree, and at each node, the tree is followed in one of two directions depending on whether the identifier is greater than or less than the identifier for that node.

The algorithm is as follows:

1. Build the binary search tree
  - a. Set counter  $i = 1$ . Take the element of  $\mathbf{D}$  and insert it at the root node as the node identifier ( $ID$ ). Set the root internal counter  $C$  (the number of times that genotype occurs in the sample) to 1 and the number of nodes  $NN$  to 1.
  - b. Increment  $i$ . Take the  $i$ th element  $r_i$  from  $\mathbf{D}$  and recursively traverse the tree. If  $r_i$  is equal to the  $ID$  from some existing node, taking into account that for each locus genotype  $AA'$  is equal to  $A'A$ , increment the internal counter  $C$  at that node by 1. Otherwise, if any outer node (the top) is reached without finding an equal  $ID$ , insert  $r_i$  into the tree as a new node, setting the node for that  $ID$  to  $r_i$ , setting its count  $C$  to 1, and incrementing  $NN$  by 1.
  - c. If  $i < n$ , repeat the previous step.
2. Calculate  $\ln P_s$  in Equation 3.
  - a. Set  $\ln P_s = 0$ .
  - b. For each node  $j$  in the tree, calculate  $C_j H_j \ln 2 - \ln(C_j!)$  and add this to  $\ln P_s$ .



- c. If  $\mathbf{D}$  is the original sample, set  $P_O = P_s$  and set  $K = 0$ . If  $\mathbf{D}$  is a permuted array, increment  $K$  by 1 if  $\ln P_s \leq \ln P_O$ .

**3.** Permutation stage.

- a. For each locus, randomly permute the  $2n$  alleles in  $\mathbf{D}$  for that locus. Return to the binary search tree step. Do this step while the number of permutations is less than the required number  $NR$ .

**4.** The estimated significance level is  $K/NR$ .

The binary search tree method for storing and retrieving the numbers of each multilocus genotype in a sample performs best when the genotypes are not sorted. This will certainly be the case for the permuted arrays. In the worst case, when the genotypes have been sorted, the method degenerates to a sequential search. Since the genotypes in the tree are stored in sorted order, it is guaranteed that no parts of the tree other than the current sub-tree can contain the particular genotype being sought. This means that only half the remainder of the tree needs be considered after each comparison. The maximum number of nodes in the tree cannot exceed the sample size, so that the average branch length remains the same as the numbers of loci and alleles increase and the time to complete the algorithm is therefore affected very little by these two numbers. The most time-consuming part of the algorithm is the permutation stage. Its speed depends on the sample size and number of loci, but not the total number of alleles per locus.

The binary search tree method has been programmed in C++, taking advantage of features of that language. It may be helpful, however, to illustrate the nature of the algorithm using the more primitive genotype-identifiers described above. Suppose a sample of 11 multilocus genotypes has been reduced to integer identifiers by the method described above, and these identifiers are 24, 8, 37, 95, 24, 6, 28, 15, 23, 94, 27. The root node of the tree is 24, and at that stage  $C = 1, NN = 1$ . The tree can be drawn under the rule that the direction of travel is up to the left for identifiers smaller than that at the current node and up to the right for larger identifiers. The resulting tree is shown in Figure 1. A 12th genotype with identifier 25 would be located after only three steps: 25 is greater than 24, less than 37, and less than 28.

## OTHER TESTS

More specific tests can be performed in order to characterize associations among subsets of the alleles within genotypes, or to increase the power of detecting specific associations. Details will now be given for tests on two-locus data.

### Products of one-locus frequencies

There may be interest in whether two-locus genotypic frequencies can be represented as products of one-locus frequencies without assuming Hardy-Weinberg equilibrium. The null hypothesis can be written as

$$P_{A_i A_j B_k B_l} = P_{A_i A_j} P_{B_k B_l} \quad (4)$$

In the human identification case, such a situation might be appropriate when there was evidence of Hardy-Weinberg disequilibrium at each locus, and two-locus genotypic frequencies were to be constructed as products of observed one-locus genotypic frequencies.

The sample counts can be written in a two-way table with row totals being **A**-locus genotype counts  $n_{ij}$  and column totals being **B**-locus counts  $n_{kl}$ . The table cell entries are the two-locus counts  $n_{ijkl}$ . The probability of the two-locus array conditional on the two one-locus arrays, under the null hypothesis in Equation 4, is

$$\Pr(\{n_{ijkl}\}|\{n_{A_{ij}}\}, \{n_{B_{kl}}\}) = \frac{\prod_{i,j} n_{A_{ij}}! \prod_{k,l} n_{B_{kl}}!}{n! \prod_{i,j} \prod_{k,l} n_{ijkl}!}$$

In this notation, the subscripts  $i, j$  range over all **A**-genotypes in the sample and so include each of the rows in the table, and the subscripts  $k, l$  range over all **B**-genotypes and so include all the columns in the table. In estimating the significance level for the exact test of Equation 4, the only quantity to be calculated is  $P_s = 1 / \prod_{g \in z} n_g!$ , where  $g$  indexes each of the two-locus genotypes with a non-zero count. Permutation proceeds by keeping the one-locus genotypes intact and permuting these genotypes among individuals at one of the two loci.

## Products of genotypic frequencies at one locus and two allele frequencies at other locus

To see if the two-locus genotypic frequencies  $P_{A_i A_j B_k B_l}$  can be represented as the product of the genotype frequency  $P_{A_i A_j}$  at one locus and the product of two allele frequencies  $p_{B_k} p_{B_l}$  at the other locus, the hypothesis is

$$P_{A_i A_j B_k B_l} = 2^{H_{kl}} P_{A_i A_j} p_{B_k} p_{B_l} \quad (5)$$

where  $H_{kl}$  is 1 if  $k \neq l$  and is 0 if  $k = l$ . A human identification setting for this test may be when one of two loci show departures from Hardy-Weinberg frequencies.

The conditional probability of a sample under the hypothesis in Equation 5 is

$$\Pr(\{n_{ijkl}\} | \{n_{A_{ij}}\}, \{n_{B_k}\}) = \frac{2^{H_B} \prod_{i,j} n_{A_{ij}}! \prod_k n_{B_k}!}{(2n)! \prod_{i,j} \prod_{k,l} n_{ijkl}!}$$

where  $H_B$  is the number of individuals in the sample that are heterozygous at locus **B**. Estimating the significance level in this case requires comparisons among values of  $2^{H_B} / \prod_{g \in z} n_g!$ , where  $n_g$  is still the number of occurrences of the  $g$ th two-locus genotype. Permutation proceeds by holding the **A**-locus genotypes intact and shuffling all  $2n$  **B**-locus alleles among individuals.

## DISCARDING SOME GENOTYPES

There are systems of genetic markers, such as VNTRs, where there is difficulty in assigning genotypes unambiguously (Weir, 1992). A single band on an electrophoretic gel may represent a homozygote, a heterozygote for two alleles with similar copy numbers, or a heterozygote with an allele giving a band outside the scoring region of the gel. It is only individuals scored as heterozygotes for which there is no ambiguity and for which tests of association are required. Only the heterozygotes in the original sample are used, and only permuted arrays where all resulting individuals are heterozygous are used. Since every individual in every array is heterozygous,  $H_g = L$  for every genotype  $A_g$ , and the conditional probability in Equation 3 for  $L$  loci becomes

$$P_s = \prod_{g \in z} \frac{2^{L n_{A_g}}}{n_{A_g}!}$$

## USING GOODNESS OF FIT MEASURES

Hypotheses have been tested here by determining the proportion of a large number of permuted arrays that have a conditional probability no larger than that of the sample. Instead of using conditional probabilities, arrays could be compared to the hypothesized or expected array on the basis of a goodness of fit statistic such as chi-square. The significance level would be the proportion of arrays with a larger statistic than that of the sample.

For multilocus genotype  $\mathbf{A}_g$ , the observed count is  $n_g$  and the expected count under the hypothesis of interest is written as  $e_g$ . The chi-square statistic is the usual

$$X^2 = \sum_g \frac{(n_g - e_g)^2}{e_g}$$

It is necessary to sum only over the observed genotypes, and use can be made of the fact that, since  $\sum_g n_g = \sum_g e_g = n$ , the statistic is  $[\sum_{g \in z} (n_g^2/e_g)] - n$ . It has been suggested (Anscombe, 1981) that Karl Pearson may have invented this statistic as an approximation for the exact test only because of the lack of computing resources at that time.

The difference between goodness of fit tests and tests based on conditional probabilities was explored in some detail by Maiste (1993). The difference in the present case is greatest when there are so many different genotypes possible that each one is likely to be unique in a sample. Certainly this is the situation for forensic databases once five or more loci are scored. If each  $n_g$  equals 1, the relevant part of the chi-square test statistic is the sum of reciprocals of expected counts, each of which will be less than 1. The test statistic tends to increase with the number of possible genotypes. The conditional probability, by contrast, depends on the reciprocal of the product of factorial counts, each of which is 1, and this product does not change with the number of possible genotypes. As an extreme example, consider the  $20 \times 20$  array in Table 1 which represents a sample of 38 two-locus genotype counts. The row and column totals are the one-locus counts. The sample size is much less than the 400 possible two-locus genotypes, and each two-locus genotype seen is unique in the sample. Other arrays, generated by permuting the genotypes at one locus holding both sets of one-locus counts constant, must have the same (when all two-locus counts remain at 0 or 1) or smaller (when some counts are greater than 1) conditional probability so that the significance level for the conditional probability test of no association is 1. However, the sampled

individuals fall into cells with low expected frequencies and this is reflected by the goodness of fit test. Based on 17,000 shuffled arrays with the same marginals, the significance level for the  $X^2$  test statistic was found to be 0.01. The same type of difference can happen in sparse arrays when each sampled individual is not unique.

## NUMERICAL RESULTS

The performance of the various test statistics and strategies has been investigated by applying them to simulated data sets.

### Two Loci

Cockerham and Weir (1973) expressed two-locus genotypic frequencies in terms of allelic frequencies and a set of disequilibrium coefficients. For example

$$P_{AABB} = p_A^2 p_B^2 + 2p_A D_{ABB} + 2p_B D_{AAB} + 2p_A p_B \Delta_{AB} \\ + p_A^2 D_B + p_B^2 D_A + D_{AB}^2 + D_{A/B}^2 + D_A D_B$$

For either two or four equally frequent alleles, genotypic frequencies were constructed with each disequilibrium coefficient in turn set to one-quarter of its maximum value, the other disequilibria being zero. Samples were taken from populations with these genotypic frequencies. For locus **A**, for example, with two equally frequent alleles, the coefficient  $D_A$  is bounded by  $\pm 0.25$  so that a population would be constructed with  $D_A = 0.0625$ . One thousand replicate samples of size 100 were drawn from each population. For each sample and each test, significance levels were computed from 10,000 permuted arrays. Although this number of permutations corresponds to the procedure used to analyze real data sets, it is not really necessary for simulation studies such as this. Oden (1991) and D.D. Boos (personal communication) have shown that the number of permutations can be set to a number very much smaller than the number of simulated data sets.

The entries in Table 2 are the proportions of the 1,000 simulated populations in which the significance level was estimated to be less than or equal to 0.05, and so are the powers of the tests for a 5% significance level. These powers are labelled  $P_1, P_2, P_3, P_4$  for the four hypotheses considered:

- $P_1$ : two-locus genotype frequency is product of four allele frequencies
- $P_2$ : two-locus genotype frequency is product of two **A** allele frequencies and **B** genotype frequency
- $P_3$ : two-locus genotype frequency is product of **A** genotype frequency and two **B** allele frequencies
- $P_4$ : two-locus genotype frequency is product of two one-locus genotype frequencies

Table 2 shows that the test for overall association, as indicated by  $P_1$ , is working well in all situations, and is most sensitive to departures from Hardy-Weinberg disequilibrium. The power for detecting Hardy-Weinberg disequilibrium is greater for four than for two alleles. As the number of alleles increases beyond four, however, it has been found that power decreases when only one of the two loci has departures from Hardy-Weinberg. For the tests where genotype frequencies at one locus are held constant, the values of  $P_2$  and  $P_3$  show that Hardy-Weinberg disequilibrium at the other locus can be detected. The test characterized by  $P_4$  is the most powerful of the conditional probability tests but, of course, does not detect departures from Hardy-Weinberg disequilibrium. Apart from the Hardy-Weinberg tests, the powers of tests for four-allele data are less than those for two-allele data. Larger sample sizes are needed to detect associations, although the loss of power with number of alleles varies among the tests.

Powers for the two-allele case can be verified by the method described by Fu and Arnold (1992). They discussed tests for  $2 \times 2$  tables, and showed how powers could be calculated by restricting attention to those tables that make a significant contribution to power. For hypothesis  $P_1$  when only  $D_{AB}$  is non-zero, their method can be used for the table of four gametes  $A_1B_1, A_1B_2, A_2B_1, A_2B_2$  with frequencies  $p_{A_1}p_{B_1} + D_{AB}, p_{A_1}p_{B_2} - D_{AB}, p_{A_2}p_{B_1} - D_{AB}, p_{A_2}p_{B_2} + D_{AB}$ . Values very similar to those in Table 2 are obtained. The method should be able to be extended to other situations covered in the table.

It should be noted that the power values shown in Table 2 are, to some extent, dependent on the method of simulating data sets. In real populations, where evolutionary forces may create complex patterns of association, the powers may well be higher. For example,  $P_4$  could be high because of

the non-zero values of several of the disequilibrium coefficients, whereas only one coefficient at a time was allowed to be non-zero in the simulations.

## Many Loci

For more than two loci, power was studied only for the test of the hypothesis in Equation 1 of no association between any of the alleles. Populations under the alternative hypothesis were simulated in a different way from previously, and in such a way to address the concerns of several authors (e.g. Lewontin and Hartl, 1991) that tests of association have low power as tests for population substructure.

Populations were constructed as amalgamations of several subpopulations that were subjected to drift, and were simulated by the coalescent process (Hudson, 1990) using a program written by P.O. Lewis. The degree of drift was specified by the coancestry coefficient  $\theta \equiv F_{ST}$  (Weir, 1990) that measures the probability of any two allelic genes within a randomly mating population being identical by descent (relative to identity between populations), and serves as a measure of divergence of populations from an ancestral population. Apparent associations between alleles were created by constructing samples as composites of samples of size 20 from 10 such populations. These admixed samples serve to address the issue of subpopulation structure that has caused concern over forensic calculations (e.g. Nichols and Balding, 1991) and correspond to the situation where a population consists of a collection of somewhat distinct subpopulations, but a sample is taken from the population as a whole.

Only unlinked loci were simulated, and either 5 or 10 equally frequent alleles at 1 to 100 loci were used. The number of replicate simulated populations used to determine power values ranged from 500 to 1,000 and the number of permuted arrays used to determine the significance level for each replicate was 3,200. Power values in the case where  $\theta = 0$  (no association) should be 0.05.

When  $\theta \neq 0$ , power always increased with the number of loci, and with the number of alleles per locus. When  $\theta = 0.05$ , for example, the power shown in Table 3 is 0.969 for four loci and 10 alleles and 0.765 for 5 alleles, but it drops to 0.552 for 2 alleles (data not shown). The explanation is as follows. As the number of loci and alleles per locus increases, there is an increasing chance that each multilocus genotype in a sample becomes unique and the value of  $P_s$  in Equation 3

reduces to  $\prod_{g \in z} 2^{H_g}$ . Because of the Wahlund effect of reducing heterozygosity by amalgamating 10 populations to provide the original simulated samples, the  $P_s$  value is expected to be smaller for this sample than for a permuted array.

The same does not apply for tests on heterozygotes only since the number of heterozygotes is the same in the sample and the permuted arrays. Indeed, the entries in Table 4 show a decrease in power, and poor performance of the test when  $\theta = 0$  for more than two loci. As Equation 3 tends to a value of  $2^L$ , zero power when  $\theta = 0$  suggests a very low chance of obtaining a sample with more than one copy of any four-locus genotype. When  $\theta \neq 0$ , power values are greater than zero, since then the samples contain genotypes homozygous at some loci and these individuals are discarded. This increases the chance of having some duplicate genotypes in the remaining data.

## DISCUSSION

Testing for associations between alleles, within and between loci, can be performed satisfactorily with exact tests conditional on allelic counts. Significance levels can be found by permutation procedures. These tests are an alternative to those based on normal statistics constructed as estimated disequilibrium coefficients divided by their estimated standard deviations (Weir and Cockerham, 1989). In general, Maiste (1993) found that conditional tests performed better than unconditional tests, of which the variance-based tests are an example.

In the current forensic setting, any associations between neutral genetic markers are most likely to be due to population substructure. Comparing multi-locus genotypic frequencies with the products of corresponding allelic frequencies is likely to detect these associations with a power that increases with the number of alleles per locus and/or the numbers of loci. Specifically, samples of size as small as 100 individuals have high powers for detecting the associations accompanying  $\theta$  values of 0.05 when several loci are used. This  $\theta$  value is the one suggested by Nichols and Balding (1991) as being an upper bound on actual values in human populations. The fact that associations are generally not found when exact tests are applied (e.g. Evett et al. 1995) suggests that  $\theta$  is less than this upper bound.

Conversely, Table 3 shows that the power is low when  $\theta$  is 0.01. What is the effect of not detecting this level of population substructure? The strength of the evidence of a matching genotype



between a evidentiary sample and a person suspected of having contributed that sample can be measured as a likelihood ratio. This is the ratio of the probability of the matching genotypes conditional on the suspected person being the contributor to the probability when another person is the contributor. The ratio, termed the forensic index by Weir (1994) by analogy to the usual paternity index, changes very little for small  $\theta$  values. For a heterozygous matching genotype, when both alleles have frequencies of 0.05, the one-locus index changes from 200 when  $\theta = 0$  to 145 when  $\theta = 0.01$  (Weir, 1994). The change is greater for smaller allelic frequencies or larger  $\theta$  values, but these are less likely for the loci currently being used in human identification.

If the product rule is to be employed to estimate multi-locus genotypic frequencies, it is necessary to check for associations between the constituent allelic frequencies. If exact tests do not detect associations, it appears appropriate to base forensic calculations on the product rule. Associations due to population substructure can be accommodated by modifications of the type proposed by Evett et al. (1995) or Balding and Nichols (1993).

## **ACKNOWLEDGEMENTS**

This work was supported in part by NIH grant GM43544. Helpful comments were provided by Dr M.A. Asmussen. The possibility of calculating exact powers was suggested by Dr J. Arnold, who also provided the Anscombe reference.

## REFERENCES

- Anscombe, F.J. 1981. Computing in Statistical Science through APL. Springer-Verlag, New York.
- Balding, D.J. and R.A. Nichols. 1993. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Sci. Int.* 64:125–140.
- Cockerham, C.C. and B.S. Weir. 1973. Descent measures for two loci with some applications. *Theor. Pop. Biol.* 4:300-330.
- Evett, I.W., P.D. Gill, J.K. Scrannage and B.S. Weir. 1995. Establishing the robustness of STR statistics for forensic applications. (submitted)
- Fu, Y.X. and J. Arnold. 1992. A table of exact sample sizes for use with Fisher's exact test for  $2 \times 2$  tables. *Biometrics* 48:1103–1112.
- Guo, S-W. and E.A. Thompson. 1992. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48:361–372.
- Hudson, R.R. 1990. Gene genealogies and the coalescent process. In D. Futuyma and J. Antonovics (Eds.) *Oxford Surveys in Evolutionary Biology*, pp 1–44.
- Lewontin, R.C. and D.L. Hartl. 1991. Population genetics in forensic DNA typing. *Science* 254:1745–1750.
- Maiste, P.J. 1993. Comparison of Statistical Tests for Independence at Genetic Loci with Many Alleles. Ph.D. Thesis, North Carolina State University, Raleigh, NC.
- Nichols, R.A. and D.J. Balding. 1991. Effects of population structure on DNA fingerprint analysis in forensic science. *Heredity* 66:297–302.
- Oden, N.L. 1991. Allocation of effort in Monte Carlo simulation for power of permutation tests. *J. Am. Stat. Assoc.* 86:1074–1076.
- Press, W.H., S.A. Teukolsky, W.T. Vetterling and B.P. Flannery. 1988. *Numerical Recipes in C*.

- The Art of Scientific Computing. 2nd Edition. Cambridge Univ. Press, New York.
- Weir, B.S. 1990. Genetic Data Analysis. Sinauer Associates, Sunderland, MA.
- Weir, B.S. 1992. Independence of VNTR alleles defined as fixed bins. *Genetics* 130:873–887.
- Weir, B.S. 1994. The effects of inbreeding on forensic calculations. *Ann. Rev. Genet.* 28:597–621.
- Weir, B.S. and C.C. Cockerham. 1989. Complete characterization of disequilibrium at two loci. In M.W. Feldman (Ed.) *Mathematical Evolutionary Theory*, Princeton Univ. Press, Princeton, pp 86–110.

**Table 1** Sample two-locus array. Marginal totals are one-locus counts.

0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	4	
0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2	
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	2	
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
1	1	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	7	
1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	1	4	
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	
0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	
5	2	1	4	2	1	1	1	1	1	1	1	5	1	2	1	2	2	2	2	1	38

**Table 2** Empirical powers of four tests for two loci with  $m = 2$  or  $m = 4$  equally frequent alleles when significance level is 0.05.

Nonzero					
disequilibria	m	$P_1$	$P_2$	$P_3$	$P_4$
None	2	0.049	0.052	0.051	0.056
	4	0.060	0.048	0.061	0.060
$D_A$	2	0.462	0.484	0.058	0.051
	4	0.951	0.976	0.041	0.056
$D_A + D_B$	2	0.778	0.486	0.498	0.065
	4	1.000	1.000	0.995	0.051
$D_{AB}$	2	0.708	0.744	0.742	0.783
	4	0.150	0.216	0.218	0.239
$\Delta_{AB}$	2	1.000	1.000	1.000	1.000
	4	0.382	0.455	0.491	0.545
$D_{AAB}$	2	0.980	0.990	0.970	0.980
	4	0.275	0.351	0.294	0.324
$D_{AAB} + D_{ABB}$	2	1.000	1.000	1.000	1.000
	4	0.555	0.631	0.625	0.724
$\Delta_{AABB}$	2	0.676	0.714	0.714	0.749
	4	0.204	0.226	0.212	0.259

**Table 3** Empirical powers (with standard deviations) for exact multilocus test with  $m = 5$  or  $m = 10$  equally frequent alleles per locus, when significance level is 0.05.

$\theta$	$m$	Number of 5-allele loci									
		1	2	3	4	10	15	25	50	75	100
0	5	0.050	0.048	0.052	0.044	0.052	0.056	0.047	0.051	0.047	0.049
		(.01)	(.01)	(.01)	(.01)	(.02)	(.01)	(.01)	(.01)	(.01)	(.01)
	10	0.050	0.049	0.051	0.052	0.051	0.048	0.050	0.050	0.050	0.049
		(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.01)	(.02)
0.005	5	0.053	0.062	0.074	0.089	0.098	0.147	0.145	0.244	0.275	0.354
		(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.03)	(.02)	(.03)
	10	0.065	0.067	0.097	0.126	0.129	0.192	0.240	0.327	0.493	0.600
		(.01)	(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.02)	(.02)
0.007	5	0.055	0.082	0.098	0.096	0.127	0.197	0.226	0.370	0.459	0.561
		(.01)	(.01)	(.01)	(.01)	(.01)	(.02)	(.01)	(.03)	(.03)	(.02)
	10	0.067	0.097	0.106	0.125	0.225	0.274	0.360	0.584	0.740	0.823
		(.01)	(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.03)	(.04)	(.03)
0.010	5	0.065	0.078	0.104	0.119	0.217	0.270	0.342	0.519	0.682	0.809
		(.01)	(.01)	(.01)	(.01)	(.03)	(.02)	(.03)	(.02)	(.03)	(.02)
	10	0.068	0.104	0.141	0.193	0.298	0.411	0.583	0.845	0.942	0.986
		(.01)	(.01)	(.01)	(.02)	(.02)	(.02)	(.02)	(.02)	(.01)	(.01)
0.050	5	0.148	0.323	0.661	0.770	0.989	1.000	1.000	1.000	1.000	1.000
		(.02)	(.02)	(.02)	(.02)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)
	10	0.237	0.774	0.894	0.969	0.999	1.000	1.000	1.000	1.000	1.000
		(.01)	(.02)	(.01)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
0.100	5	0.434	0.856	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(.02)	(.02)	(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)
	10	0.757	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(.01)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)	(.00)

**Table 4** Empirical power of exact tests with homozygotes excluded when significance level is 0.05.

	Number of 10-allele loci			
$\theta$	1	2	3	4
0	0.052	0.047	0.005	0.000
0.01	0.062	0.050	0.006	0.002
0.05	0.104	0.065	0.019	0.007
0.10	0.334	0.292	0.083	0.014

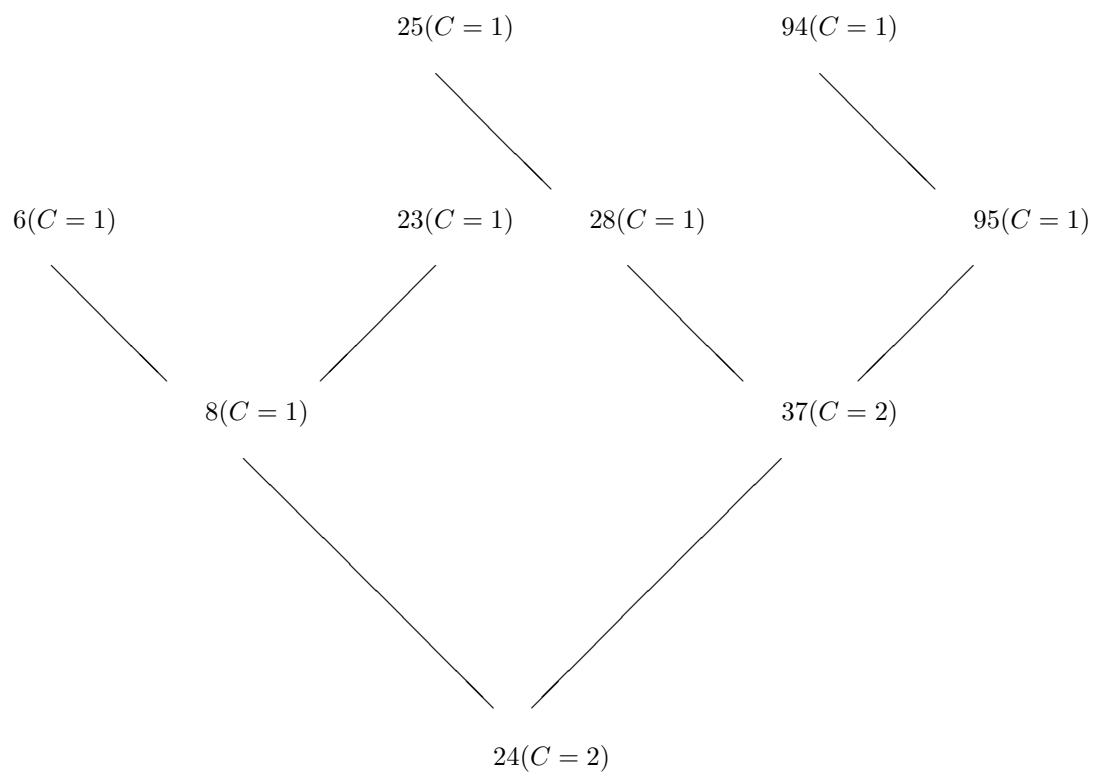


Figure 1: Binary Search Tree.