Analytical Aspects of Population-Specific DNA Fingerprinting for Individuals

P. E. Smouse and C. Chevillon

An emerging problem of some interest is whether we can determine the population membership of a single individual, using a population-specific "genetic fingerprint." The levels of accuracy and precision required are beyond the reach of allozyme analysis, and attention has shifted to DNA polymorphisms. There are different types of DNA markers available for population surveys: RFLPs, mini- and microsatellites, and RAPDs, and each type has its own strengths and weaknesses. We present a generic analysis that relates gene pool separation to our ability to assign individuals, an analysis that does not depend on the type of marker. We then review strengths and weaknesses of different DNA markers, in the context of DNA fingerprinting. Codominant loci are best. It is possible to gain more information per marker with multiallelic loci, but diminishing returns set in rapidly, and it is better to add loci. A modest number of independent loci is best, each locus with a modest number of alleles and with each allele in modest frequency.

The use of genetic data for population structure analysis is based on the idea that polymorphic genetic markers can be used to distinguish among individuals and/or populations. Until the development of protein electrophoresis in the 1960s (Hubby and Lewontin 1966; Lewontin and Hubby 1966), such studies were limited to organisms that exhibited polytene chromosome polymorphisms (Dobzhansky et al. 1954; Lewontin and White 1960) or morphologically detectable variation (Cain and Sheppard 1954; Kettlewell 1955). Protein assay allowed serious "population structure" analysis in a number of species for the first time, and the literature on the subject has since become voluminous.

Although assay protocols for a great many allozymic loci are available (Davidson et al. 1989), we seldom have more than a dozen loci in any particular organism that are sufficiently polymorphic to be useful for population structure analysis. Although that is sufficient for population structure analysis, one of the problems of growing interest is whether we can genetically determine the population membership of a single individual, thus allowing us to define a population-specific "fingerprint." The levels of genetic accuracy and precision required are far beyond the reach of routine allozyme analysis, so inasmuch as DNA-based markers are commonly thought to be available in almost limitless supply, recent attention has shifted to them.

While DNA-based markers have increased our genetic resolution, they are not without their own limitations, some of which can be traced to deeper analytic issues. Our purpose here is to explore some of those issues, by way of providing a reality check on the use of DNA markers for population work. In particular we will (1) elucidate an organizing principle that relates the separation of gene pools to our ability to assign individuals, (2) discuss various implications of increased genetic and/or population sampling relative to population structure analysis, and (3) review characteristics and limitations of different DNA methodologies in that context.

Gene Pools as Probability Clouds

Molecular Confetti

The natural inclination is to generate a vast array of polymorphic markers, based on the idea that with enough markers, every individual becomes genetically unique. But consider a set of 25 codominant, unlinked, polymorphic loci, each with two alleles ($p = 0.3$, $q = 0.7$). The most likely genotype is the 25-locus homozygote aabbcc . . . , with a frequency of $(q^2)^{25} < 2 \times 10^{-8}$; the rarest genotype, aabbcc . . . has a frequency of $(p^2)^{25} < 10^{-26}$. Thus every genotype has a frequen-

From the Center for Theoretical and Applied Genetics, Cook College, Rutgers University, New Brunswick, NJ 08903–0231 (Smouse) and the Institut des Sciences de l'Evolution, Universite de Montpellier II, Montpellier, France (Chevillon). The authors wish to thank S. Kaufman, C. J. Kobak, B. S. Weir, and an anonymous reviewer, all of whom contributed helpful critique on the manuscript. The work was supported by the New Jersey Agricultural Experiment Station, USDA-NJAES-32106. Address correspondence to P. Smouse at the address above or e-mail: smouse@aesop.rutgers.edu.

Journal of Heredity 1998;89:143–150

cy in the range $[10^{-27} < \text{fr(genotype)} < 2 \times 10^{-8}]$. Now draw a random genetic sample of 100 individuals. The expected outcome is a collection of unique genotypes, none of which is aabbcc . . . ; draw a second random sample of 100—the most likely outcome is the same. In fact, the two samples (from the same gene pool) will probably not share a single genotype in common. The gene pool is a thinly dispersed probability cloud; the probability of drawing any particular genotype, even the most likely genotype, is virtually nil. Any particular sample will contain an unpredictable collection of unique genotypes.

Now imagine a second gene pool having the same 25 polymorphic loci and the same alleles, but with the allele frequencies reversed ($p = 0.7$, $q = 0.3$). The most likely genotype in this second population is the 25-locus homozygote AABBCC . . . , with a frequency of $(q^2)^{25} < 2 \times 10^{-8}$; the rarest genotype—aabbcc . . . —has an expected frequency of $(p^2)^{25} < 10^{-26}$. The sampling implications for this second gene pool are the same as those for the first, an unpredictable mix of unique genotypes. Given two samples, we should anticipate no genetic overlap, whether they are drawn from the same or from different gene pools. How are we to distinguish between one gene pool, sampled twice, and two gene pools, sampled once each? The uniqueness of the sampled genotypes does not help, and viewing the gene pool as a collection of molecular confetti is not the answer.

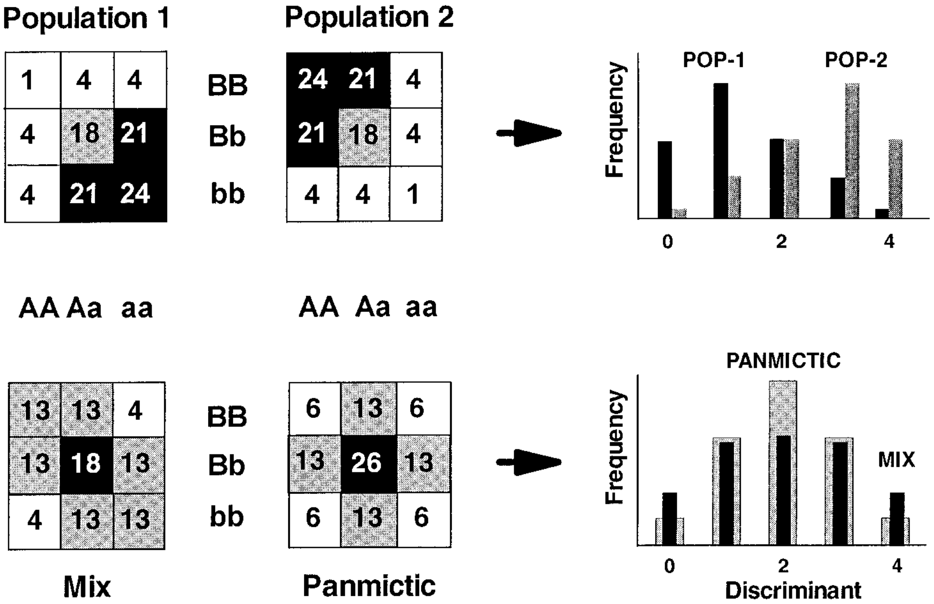

Figure 1. Two-locus genotype frequency arrays for a pair of gene pools: upper panel—separate panmictic equilibria, where the first gene pool has $p_A = p_B = 0.3$ and the second gene pool has $p_A = p_B = 0.7$, also plotted as a histogram, where the ordinate is a count of the number of cap-alleles (A or B); lower panel—a 50:50 mixture of the separate gene pools (on the left), a panmictic gene pool with the same allele frequencies, $p_A = p_B = 0.5$, in the middle, and a histogram of the frequencies along the same discriminant axis (on the right).

An Organizing Principle

We need an organizing principle that tells us when we are sampling from one probability cloud and when we are sampling from two different clouds. We cannot predict the precise genotype of a random individual, but we can determine which part of the probability space it occupies, and from that we can determine its most likely gene pool of origin, post hoc. To do that convincingly, we require a large number of well-behaved and statistically independent genetic markers that exhibit sufficient allele-frequency divergence between gene pools.

Return to the example, but consider just the first two loci, with frequencies $p = 0.3$ and $q = 0.7$ in the first gene pool and $p = 0.7$ and $q = 0.3$ in the second. Since the loci are segregating independently, each population is in two-locus panmictic equilibrium; we have Hardy–Weinberg equilibrium for both loci, as well as gametic equilibrium between them. All nine genetic phenotypes (10 genotypes, but AB/ab and Ab/aB are not usually distinguishable) are present in each population, but most of the probability mass is in the lower right corner of the first population and in the upper left corner of the second (Figure 1, upper panel). That probability separation provides a key to the solution of our identification problem.

If we know the frequency composition of the two gene pools, we can use discriminant analysis to assign individuals to one or the other gene pool (Smouse et al. 1982; Spielman and Smouse 1976). A picture of the analytic situation is provided by projection of the probabilities onto the principal axis of the two-dimensional genotype array (right-hand side of Figure 1). That diagonal axis is a discriminant function, a count of the number of cap-alleles (A and B) within a genotype: AABB (4), AABb and AaBB (3), AAbb, AaBb, and aaBB (2), Aabb and aaBb (1), and aabb (0). Scanning along the diagonal (discriminant) axis, we see a partial separation of the two gene pools.

Of course, we may be in ignorance of the allele frequencies within the separate gene pools; we may not even realize that we have more than one gene pool. In such cases, the procedure is the same, except that we have to extract the principal axis from the mixed-population data themselves. For instance, consider a 50:50 mixture of the two gene pools (lower panel, Figure 1). By contrivance, the allele frequencies for both loci are $p = 0.5 = q$ for the mix. For this simple example, the principal axis is unchanged, and the projection of the frequencies onto that axis yields the result on the right-hand side of the lower panel of Figure 1. For a single panmictic gene pool with the same ($p = 0.5 = q$) allele frequencies, the genotype array is frequency symmetric around the central (AaBb) genotype; that translates into a unimodal, roughly normal distribution along the principal axis. By contrast, the genotype array of the two-population mixture is overdispersed along the principal diagonal axis relative to the bell-shaped panmictic reference distribution. The mix shows too many homozygotes (Wahlund effect), as well as gametic phase disequilibrium (Smouse and Neel 1977; Smouse et al. 1983). Both features are intrinsic signatures of a heterogeneous mixture; the magnitudes of both are increasing functions of the allele-frequency divergence between the two gene pools.

Now extend the argument to four loci, each with the same 30:70 versus 70:30 allele-frequency split. We create a four-dimensional cube, with the first gene pool having its probability mass concentrated near the aabbccdd corner, and the second gene pool with its mass near the AABBCCDD corner. Again using a principal axis rotation, we portray the frequency profiles in the upper left corner of Figure 2. Divergence of the two gene pools is

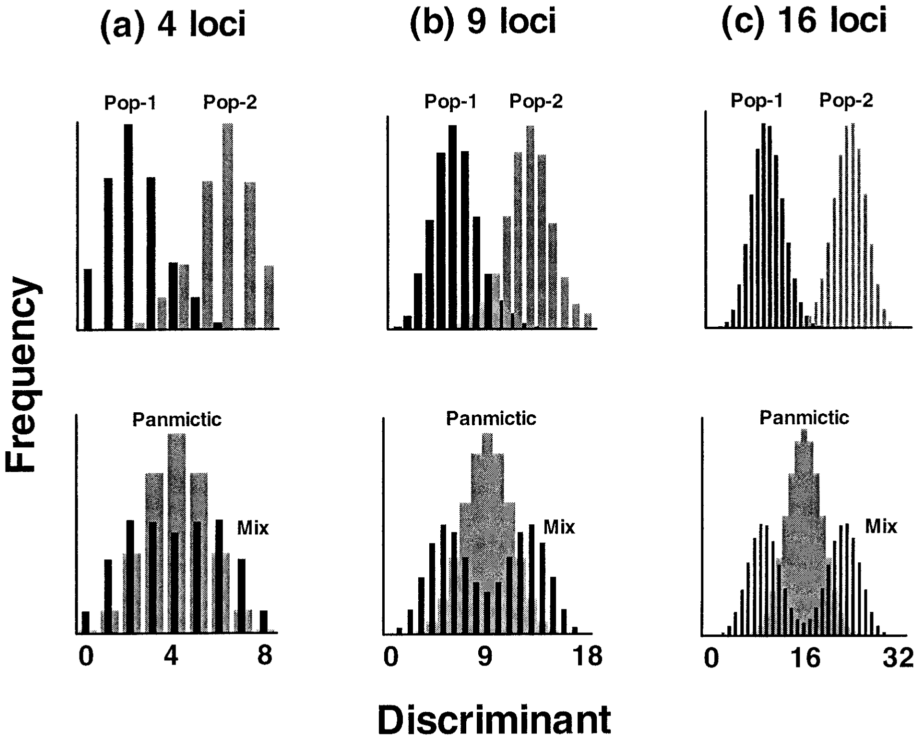

Figure 2. A projection of frequency profiles from two separate gene pools (top panel), their 50:50 mix, and a panmictic reference (bottom panel) onto the discriminant axis. Each locus has the same 30:70 versus 70:30 frequency split of Figure 1: left-hand pair, 4 loci; middle pair, 9 loci; right-hand pair, 16 loci. Resolution improves with the number of loci.

even more evident with four loci than with two, and the mixture shows overt evidence of genetic bimodality (lower left corner). Four loci are better than 2, 9 loci are better than 4 (middle entries of Figure 2), and 16 are better than 9 (right-hand side of Figure 2). Each locus, viewed in isolation, shows substantial overlap of gene pools, but with as many as 25 loci, there would be little doubt that the collection of genotypes represents two multidimensional probability clouds, whether we knew their probability distributions in advance or not, and in spite of the fact that every genotype in the sample was unique. Separation would emerge clearly from the data themselves.

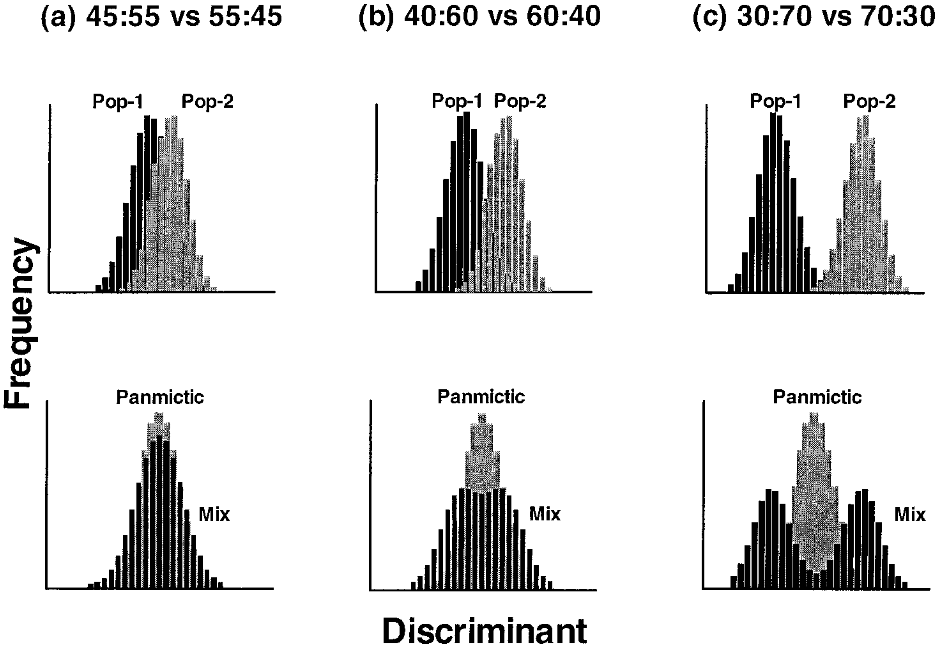

Figure 3. A projection of frequency profiles from two separate gene pools (top panel), their 50:50 mix, and a panmictic reference (bottom panel) onto a 16-locus discriminant axis: **(a)** loci with a 45:55 versus 55:45 split; **(b)** loci with a 40:60 versus 60:40 split; **(c)** loci with a 30:70 versus 70:30 split. Resolution improves with allele frequency divergence.

The example is an illustration, but the real world involves loci with smaller divergence among gene pools. To see the frequency implications, compare the 16-locus results of a 45:55 versus 55:45 split (Figure 3, left-hand side), or those from a 40:60 versus 60:40 split (Figure 3, middle entries), with those from our 30:70 versus 70:30 split (Figure 3, right-hand side). With the 30:70 versus 70:30 split, the separation is self-evident and almost obvious; with the 40:60 versus 60:40 split, it is evident but less than categorical; with the 45:55 versus 55:45 split, we will need a priori gene pool identification; there is little hope of discerning the divergence from the data themselves. In addition, real organisms present loci with allele frequencies that are not conveniently balanced around $p = 0.5 = q$, multiple alleles, multiple populations, mixture fractions that are not equal, and myriad other departures from the simple assumptions of our example. Statistical aggravations aside, however, the basic principle remains. Each gene pool is a probability cloud, and the greater the allele-frequency divergence among gene pools, the less the overlap of those probability clouds, and the more likely we are to be able to assign individuals to the correct population, based on genetic criteria.

A Formal Translation

We need a formal translation that converts allele frequency divergence into a measure of genetic overlap and/or the efficacy of assigning individuals to their proper gene pools. This treatment should convert different sorts of genetic data into a common statistical currency, allowing comparison of different molecular methodologies. We note that if populations do not overlap at all, we can use genetic data to assign an individual to its correct population with certainty. To the extent that the probability clouds overlap, we will make some mistakes; the more overlap we have, the more mistakes we will make. Allele frequency divergence, probability cloud overlap, and the probability of correct allocation are three closely related measures of the same thing, and we can exploit that relationship to our advantage.

The probability of correctly allocating

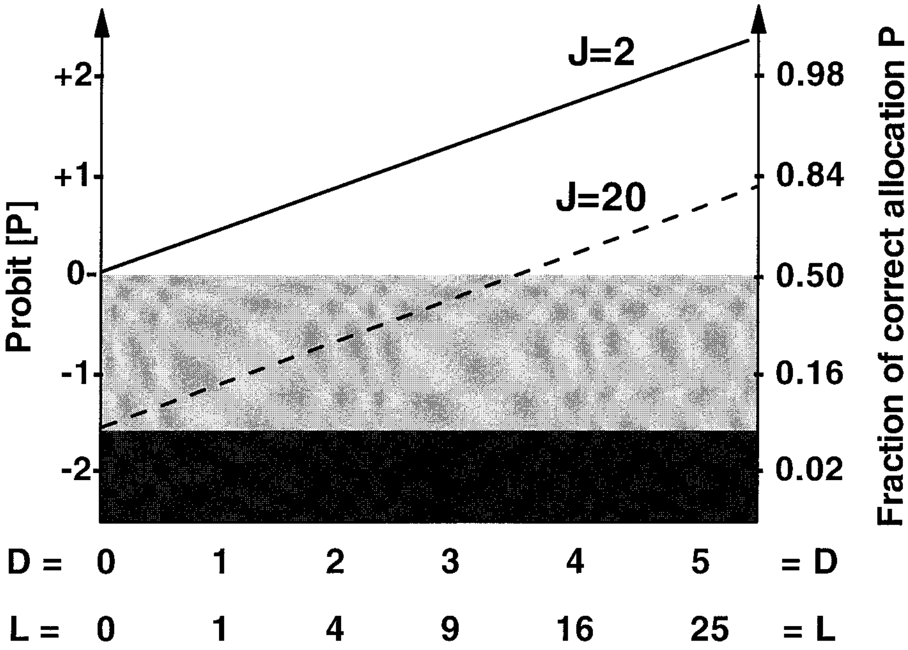

Figure 4. Linear relationship between the probit of correct allocation rate and the average genetic distance between populations. For loci of equal discriminatory power, average genetic distance increases with the square root of the number of loci. Shaded areas are correct allocation fractions lacking genetic data on population divergence.

an individual is directly related to the average pairwise genetic distance between populations, defined as

$$D = [\{2J(J - 1)\}^{-1} \Sigma_j\Sigma_l\Sigma_k\{(\bar{y}_{jk} - \bar{y}_{lk})^2/\bar{y}_{\bullet k}\}]^{1/2} \quad (1)$$

where J is the number of candidate populations, and where $\bar{y}_{jk}$, $\bar{y}_{lk}$, and $\bar{y}_{\bullet k}$ are estimated frequencies of the kth allele in the jth and lth populations and the global collection, respectively. The relationship between allocation success and average genetic distance is shown in Figure 4, where the probit of the correct allocation fraction is shown on the ordinate and the average genetic distance (D) on the abscissa. [Recall that probit [p] is the number of standard deviations, left or right of the mean (zero) of an N(0,1) distribution, that corresponds to an upper tail probability of $(1 - p)$. Thus, for example, probit [0.50] $= 0$, and probit $[0.05] = -1.645$.] The slope of the line is 1/2, irrespective of the number of candidate populations. The intercept is probit [$1/J$]; if there are J candidate sources, there are $(J - 1)$ ways to misallocate. In the absence of genetic divergence, the probability of correct allocation is $1/J$. The more candidate populations there are, the harder it is to allocate accurately for a given value of D (Smouse et al. 1982). For example, while a D value of 2 would indicate minimal overlap for a pair ($J = 2$) of populations and would provide impressive allocation success; that same D value for $J = 20$ populations would not provide much resolution (Figure 4). The more candidate populations we have, the greater the value of D we will need to achieve the same degree of discriminability.

Multiple Loci and Multiple Alleles

While the probit of the correct allocation fraction is linear in D, it is linear in the square root of the number of two-allele loci, provided each locus contributes equally and independently. The linear progression is thus from 1 locus to 4 loci, to 9 loci, to 16 loci, to 25 loci, etc. This is also shown in Figure 4, for the special case where each locus contributes $D = 1$. We will need a daunting number of polymorphic loci, especially with multiple candidate populations. The addition of nonindependent loci (linkage disequilibria within populations) reduces genetic resolution; the marginal information provided by a second locus derives from discrimination within genotypes of the first locus. Addition of statistically correlated loci is informationally redundant.

It is also possible to extract more information per unit locus by using a technology that uncovers multiple alleles, a strategy that also reduces the cost per unit assay. It can be shown that increasing the number of alleles increases the value of D to some degree, but beyond some point it is more efficacious to deploy another locus than to add more alleles. The precise trade-off point will depend on the (unpredictable) allele-frequency profiles of the loci being used. In addition, with increasing numbers of rare alleles, we encounter some sampling difficulties that are difficult to circumvent.

The Sample Size Problem

To elucidate the sample size problem, we begin with the distance between the ith individual in the jth population and the average of the lth population,

$$D_{ij,l} = [\Sigma_k(y_{ijk} - \bar{y}_{lk})^2/\bar{y}_{\bullet k}]^{1/2}. \quad (2)$$

Now, y_{ijk} is the frequency of the kth allele in the ith individual of the jth population (1 if homozygous, ½ if heterozygous, 0 if absent); $\bar{y}_{lk}$ and $\bar{y}_{\bullet k}$ are the estimated frequencies of the kth allele for the lth population and total collection, respectively. The idea is to assign the ith individual to that population for which D is smallest, that is, to the genetically most similar population.

The difficulty is that the procedure is biased in favor of correct allocation. This is especially a problem for rare alleles, which not uncommonly appear as single heterozygotes for the entire study. An allele whose frequency is $r = 0.001$ is unlikely to occur even once in a sample of 50 diploid individuals. An allele whose frequency is $q = 0.01$ will probably occur once or not at all. Conversely an allele whose frequency is $p = 0.10$ will almost surely be seen more than once in a sample of 50 individuals. The absence of a rare allele from a sample may indicate its absence from the gene pool, but it may equally well be interpreted as the expected sampling outcome for a rare allele that really is present in the gene pool.

Because we use the y_{ijk} to produce the $\bar{y}_{jk}$, singletons are invariably closest to their observed sample of occurrence, and they are thus invariably allocated "correctly." Unfortunately the procedure is circular; we have "stacked the deck" in favor of correct allocation. With a large number of rare alleles, the upward bias in the estimated success rate will be substantial. We have (Smouse et al. 1982) shown that to avoid this allocation bias, one removes the individual from its own population before computing $D_{ij,j}$, asking instead how close the individual is to others in its population. For polymorphic alleles, removal of an index individual simply removes the bias, but removing a singleton amounts to removal of the rare allele entirely from the study, and we have nowhere to allocate it. Singletons are utterly useless.

We need to ensure that $2N_j \times p_{jk} \geq 2$ for all alleles, so that removal of the index individual creates no confusion. We can do that if we increase sample sizes. For example, with a sample size of $N_j = 1000$, relatively low-frequency alleles will be seen more than once. Of course, such sample sizes represent severe overkill for the common alleles that make up the bulk of the sample, and on which the overall success of the enterprise depends. The alternative is to reduce the numbers of alleles. With many loci to choose from, we can concentrate on those with allele frequency spectra compatible with limited sample sizes. Alternatively we can pool alleles into a smaller number of higher frequency "allelic classes." We need a modest number of alleles per locus, each of modest frequency. Large numbers of rare alleles per locus should not be viewed as a substitute for large numbers of loci.

Beyond this concern with rare alleles, there are some minimal sample size requirements for the description of populations in multidimensional genetic space. We need large numbers of genetic characters for large numbers of candidate populations, but it would be silly to have many more characters sampled than individuals per population. The number of characters, K, is the total number of alleles at all loci minus the number of loci. We need to put a tight confidence ellipse around each population mean vector, of length K, if we are to establish population divergence credibly. The degrees of freedom for an estimate of the within-population covariance matrix are $(N_j - 1)$ for the jth population (Anderson 1958). If we use $K = 100$ characters, for example, we need a bare minimum sample size of $N_j = 101$ to have enough degrees of freedom to estimate a non-singular within-population covariance matrix for the jth population. If we can make the assumption that the within-population covariance matrices from the J populations are homogeneous, we can pool the separate within-population matrices, none of them estimated very well, to obtain an estimate of the average within-population matrix, with $(N - J)$ degrees of freedom. At the very least, we need $(N - J) > K$ degrees of freedom and a sample size such that $(N - J) > 2K$ would not be excessive, since the variances of variance/covariance terms are fourth order. There is no justification for viewing large numbers of characters as a substitute for inadequate sample sizes. Indeed, they make larger sample sizes imperative.

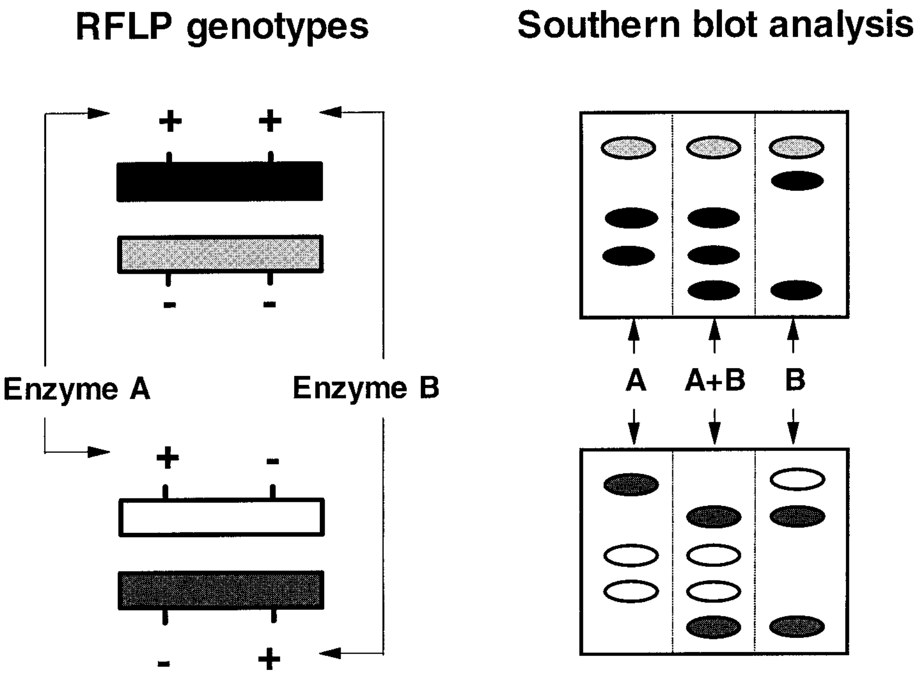

Figure 5. The gametic phase ambiguity of a two-site RFLP locus, and how that can be resolved with a double-digest procedure.

Competing Molecular Methodologies

RFLP Markers

There are several alternative molecular methodologies we might employ, each with its own strengths and weaknesses, vis-à-vis population identification. Consider first a single restriction fragment length (RFLP) marker, assayed with one molecular probe and one restriction enzyme, yielding a pair of codominant alleles, the informational equivalent of a two-allele, codominant allozyme locus. The cost of developing RFLP assays is high, so the usual strategy is to assay for multiple restriction sites detectable with the same probe, often requiring two or more restriction enzymes. With additional recognition sites, we can detect multiple haplotypes, increasing the level of polymorphism per unit locus.

The difficulty, of course, is that double-marker heterozygotes are linkage-phase ambiguous. It is possible, using double digest methods (Figure 5) or combination probes, to resolve some of that phase ambiguity. In other cases, data on relatives can be used to accomplish the same end. With multiple markers, however, all the multiple-site heterozygotes are phase ambiguous. Moreover, since all the markers are tightly linked, linkage disequilibrium guarantees high correlations within the set. The marginal information added by each successive marker of the set decreases rapidly; we quickly reach the point where it is better to deploy another probe with a small number (1–3) of informative sites (see Smouse and Chakraborty 1986). The consequence is a meaningful increase in the cost per unit information.

Mini- and Microsatellite Markers

Minisatellite methods also reveal multiple length alleles, but without phase ambiguity. Such markers have proven useful in human forensic analysis (Budowle et al. 1994; Chakraborty et al. 1992; Devlin 1993), as well as in population surveys of other organisms (Bentzen and Wright 1993; Dias et al. 1996; Kempenaers et al. 1992; Rave et al. 1995), and there are some meaningful cost advantages. Problems arise when the sheer number of length alleles is large and their separation is difficult to establish in routine assay. The usual strategy is to enforce some sort of binning (Weir and Gaut 1993), which reduces the ambiguity to manageable levels. Given the informational redundancy inherent in multiple alleles, we are better served by a small set of "allele classes," all of moderate frequency, than we are by a very large number of rare alleles, whose identities and frequencies are difficult to establish precisely. Modest numbers of alleles are preferable.

An attractive variation on this theme is

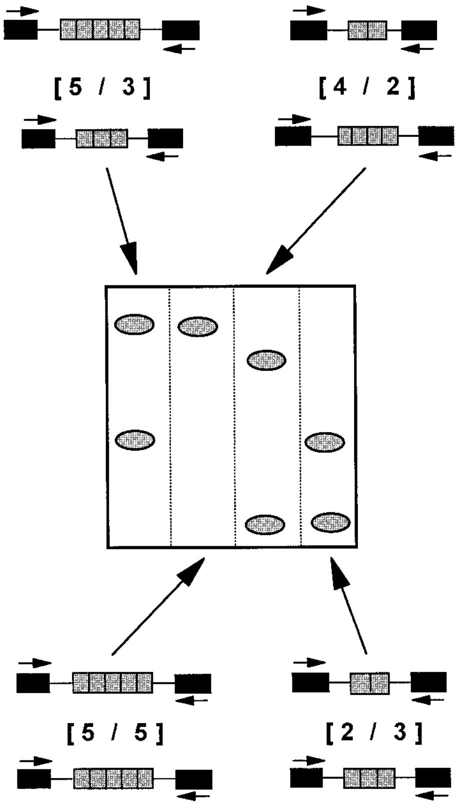

Figure 6. A schematic of diploid, codominant phenotypic resolution for a typical microsatellite locus with multiple alleles, showing an absence of phase ambiguity.

provided by microsatellite loci, based on polymerase chain reaction (PCR) assay. The idea is that small repeating motifs are flanked by a particular pair of primer recognition sequences (Hammond et al. 1994). With PCR amplification we can assay a series of unambiguous length alleles (Figure 6). There are a large number of assayable polymorphic loci available with this technique, and with appropriate preliminary work we can obtain single-locus assay systems with simple banding patterns. Mini- and microsatellite systems have been useful in humans (Chakraborty et al. 1992; Gomolka et al. 1994), other mammals (Dallas et al. 1995; Ellegren et al. 1992), birds (Dias et al. 1996; Dixon et al. 1994), fish (Brooker et al. 1994), insects (Estoup et al. 1995; Peters et al. 1995), and fungi (Stendid et al. 1994).

RAPD Markers

The DNA methods described thus far require major R&D investments to support high- volume survey work. There are interesting organisms, still in the early stages of exploration and/or domestication, for which a large-scale R&D effort is neither timely nor cost effective. If we can accept answers that are approximate, polymerase chain reaction (PCR) methods can be employed to produce large numbers of random amplified polymorphic DNA (RAPD) markers. Given that a primer recognition sequence is present on one strand of the double helix, a PCR reaction will begin in one direction. If the same sequence is found downstream, but on the other strand of the double helix and in the opposite orientation, the reaction will go in both directions. Amplification occurs everywhere in the genome that the flanking sequences are just right. If some individuals have an unrecognizable flanking sequence (due to mutation), amplification fails; the locus becomes polymorphic and useful (Welsh and McClelland 1990; Williams et al. 1990).

We begin with an oligonucleotide primer of arbitrary sequence. Hundreds of such primers are commercially available, each with its own profile of amplification products, to a first approximation—one product per locus. We use the term "locus" here advisedly; we observe amplification products whose genetic inheritance remains unexplored. PCR amplification tends to be an all-or-nothing phenomenon. It is difficult to distinguish between homozygous (+ +) and heterozygous (+ −) individuals; the (− −) individuals are unambiguous, but dominance transforms the example in Figure 1 (re-created as Figure 7b) to the form shown in Figure 7a. It takes a lot more dominant than codominant markers to achieve the same population resolution, but we can assay a great many loci for minimal effort (Huff et al. 1993).

Haploid Genetic Systems

All of the preceding is based on the use of diploid loci. There are haploid organisms and organelles that can be examined in much the same fashion. Most work to date has been devoted to animal mtDNA (Avise et al. 1987; Cann et al. 1987; Moritz et al. 1987), which shows large amounts of variation within many species. The major advantage of assaying a haploid genome is that dominant markers and linkage phase become transparent. In the absence of recombination, however, mtDNA markers are not independent and multimarker mtDNA haplotypes are best treated as multiple alleles, with all of the inherent statistical limitations. The major limitation of haploid analysis, however, is that we lose the considerable advantages of the diploid, Hardy–Weinberg reference frame (Xu et al. 1994). To put that in context, consider two populations with four mtDNA haplotypes and our (by now familiar) 30:70 versus 70:30 polymorphic frequency split. Known gene pools will show obvious differences (Figure 7c), but mixtures leave no obvious signature, since single gene pools are themselves out of linkage equilibrium, and there is no Wahlund effect to detect. The separate probability clouds exist, but we need a great deal of mitochondrial divergence to have any realistic hope of clear-cut separation.

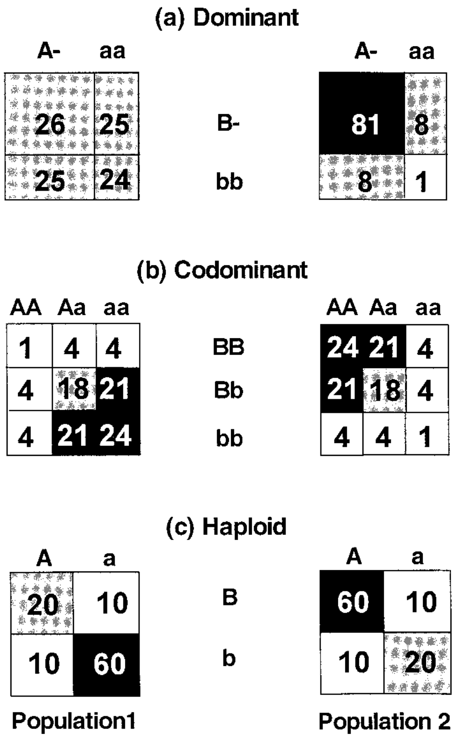

Figure 7. Expected frequency distributions for different types of biallelic genetic markers at each of two loci. Population 1 is characterized by $p_A = p_B = 0.3$, while population 2 is characterized by $p_A = p_B = 0.7$ in each case: **(a)** the frequency distribution of dominant (RAPD) phenotypes for a diploid organism; **(b)** the distribution of codominant (RFLP or microsatellite) phenotypes for a diploid organism; **(c)** the distribution of alleles for a haploid (mtDNA or chloroplast) genome.

On the other hand, we can usually relate the haplotypes to each other by single-step (mutation) links. That fact can sometimes provide us with useful information on population membership, the idea being that populations may evolve and diverge along the network (Excoffier et al. 1992). The ability to measure along the net provides an alternative organizing principle, and can make up for the loss of the Hardy–Weinberg reference frame (Excoffier and Smouse 1994; Takahata and Palumbi 1985; but see Epifanio et al. 1995). Plant chloroplast and mitochondrial genomes are now coming under scrutiny for population work (Birky 1988; Harris and Ingram 1991; Mason-Gamer et al. 1995; Soltis et al.

1992), and we can anticipate the same sorts of problems and opportunities as we have encountered with animal mitochondria. Haploid analysis, either alone or in parallel with diploid analysis, is also possible for conifers, fungi, ferns, and other organisms with a pronounced alternation of generations. The combination of haploid and diploid analysis can be useful in resolving otherwise intractable linkage phase problems (Adams 1992; Gibson and Hamrick 1991).

Conclusions

We need a large number of polymorphic markers to assign individuals to their correct populations; allozyme methods have proven inadequate, but newer DNA methods are capable of providing the resolution required. There is a natural tendency to assay myriad markers, so that each individual is genetically unique, but chopping the genome into molecular confetti is counterproductive. We need an organizing principle to extract information from those genotypes, and it is provided by the multiple-locus Hardy–Weinberg equilibrium.

Probit $[p]$ is essentially linear in the average genetic distance between pairs of populations, and theory leads to several conclusions: (1) The greater the allele frequency divergence among populations, the better is our discriminatory power. (2) The more candidate populations we have, the harder it is to assign correctly for a particular level of genetic divergence, and the greater is the number of loci required. (3) Probit $[p]$ is linear in the square root of the number of loci, everything else being equal. (4) The informational requirement for strong discrimination of a large number of candidate populations is severe.

It is possible to extract more information per unit locus by using multiallelic methods, but diminishing returns set in quickly with more than two alleles. Moreover, if one uses sample sizes that are too small or loci with too many alleles, rare alleles occur as singletons, and singletons are useless. Beyond some point, it is better to deploy another locus than to distinguish among more alleles. We need sample sizes that are sufficient ($N_j > K$) to describe the candidate gene pools credibly.

There are several alternative molecular methodologies we might employ. RFLP methods can be used to describe single loci with multiple, codominant alleles, but linkage disequilibria are large and gametic phase ambiguous. Mini- and microsatellite methods can also generate multiallelic loci, but without the linkage phase ambiguity; the numbers of alleles are often so large that one must pool alleles into length classes. RAPDs can be generated with minimal cost and R&D investment, but their limitation is that they are dominant, and we require many more dominant than codominant markers to accomplish the same resolution. We can also examine haploid organisms and organelles, avoiding dominance problems altogether. Two major disadvantages are the loss of the diploid Hardy–Weinberg reference frame and the existence of tight linkage among markers. We prefer a modest number of independently segregating, multiallelic, codominant loci, each locus with a small number of alleles, with each allele in moderate frequency.

References

Adams WT, 1992. Gene dispersal within forest tree populations. New Forest 6:217–240.

Anderson TW, 1958. An introduction to multivariate statistical analysis. New York: John Wiley.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, and Saunders NC, 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu Rev Ecol Syst 18:489–522.

Bentzen P and Wright JM, 1993. Nucleotide sequences and evolutionary conservation of a minisatellite variable number tandem repeat cloned from Atlantic salmon, *Salmo salar.* Genome 36:271–277.

Birky CWJ, 1988. Evolution and variation in plant chloroplast and mitochondrial genomes. In: Plant evolutionary biology (Gottlieb L and Jain S, eds). New York: Chapman and Hall; 25–53.

Brooker AL, Cook D, Bentzen P, Wright JM, and Doyle RW, 1994. Organization of microsatellites differs between mammals and cold-water teleost fishes. Can J Fish Aquat Sci 51:1959–1966.

Budowle B, Monson KL, Giusti AM, and Brown B, 1994. The assessment of frequency estimates of *Hae-III* generated VNTR profiles in various databases. J Foren Sci 39:319–352.

Cain AJ and Sheppard PM, 1954. Natural selection in *Cepea.* Genetics 39:89–116.

Cann RL, Stoneking M, and Wilson AC, 1987. Mitochondrial DNA and human evolution. Nature 325:31–36.

Chakraborty R, DeAndrade M, Daiger SP, and Budowle B, 1992. Apparent heterozygote deficiencies observed in DNA typing data and their implications in forensic applications. Ann Hum Genet 56:45–57.

Dallas JF, Dod B, Boursot P, Prager EM, and Bonhomme F, 1995. Population subdivision and gene flow in Danish house mice. Mol Ecol 4:311–320.

Davidson WS, Birt TP, and Green JM, 1989. A review of genetic variation in Atlantic salmon, *Salmo salar* L., and its importance for stock identification, enhancement programmes and aquaculture. J Fish Biol 34:547–560.

Devlin B, 1993. Forensic inference from genetic markers. Stat Meth Med Res 2:241–262.

Dias PC, Verheyen GR, and Raymond M, 1996. Source-sink populations in Mediterranean blue tits: evidence using single-locus minisatellite probes. J Evol Biol 9: 965–978.

Dixon A, Ross D, O'Malley SLC, and Burke T, 1994. Paternal investment inversely related to degree of extra-pair paternity in the reed bunting. Nature 371:698–700.

Dobzhansky TH, Pavolvsky O, and Green JM, 1954. Interaction of the adaptive values in polymorphic experimental populations of *Drosophila pseudoobscura.* Evolution 8:335–349.

Ellegren H, Johansson M, Sandberg K, and Andersson L, 1992. Cloning of highly polymorphic microsatellites in the horse. Anim Genet 23:133–142.

Epifanio JM, Smouse PE, Kobak CJ, and Brown BL, 1995. Mitochondrial DNA divergence among populations of American shad (*Alosa sapidissima*): how much variation is enough for mixed stock analysis? Can J Fish Aquat Sci 52:1688–1702.

Estoup A, Garnery Z, Solignac M, and Cornuet J-M, 1995. Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics 140:679–695.

Excoffier L and Smouse PE, 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. Genetics 136:343–359.

Excoffier L, Smouse PE, and Quattro JM, 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction sites. Genetics 131:479–491.

Gibson JP and Hamrick JL, 1991. Heterogeneity in pollen allele frequencies among cones, whorls, and trees of table mountain pine (*Pinus pungens*). Am J Bot 78: 1244–1251.

Gomolka M, Hundrieser J, Nurnberg P, Roewer L, Epplen JT, and Epplen C, 1994. Selected di- and tetranucleotide microsatellites from chromosomes 7, 12, 14 and Y in various Eurasian populations. Hum Genet 93:592–596.

Hammond HA, Jin L, Zhong Y, Caskey CT, and Chakraborty R, 1994. Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet 55:175–189.

Harris SA and Ingram R, 1991. Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plasmid transmission. Taxon 40:393–412.

Hubby JL and Lewontin RC, 1966. A molecular approach to the study of genetic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura.* Genetics 54:577–594.

Huff DR, Peakall R, and Smouse PE, 1993. RAPD marker variation within and among natural populations of outcrossing Buffalograss (*Buchloë dactyloides* (Nutt.) Engelm.). Theor Appl Genet 86:927–934.

Kempenaers B, Verheyen GR, Van den Broeck M, Burke T, van Broeckhoven C, and Dhondt AA, 1992. Extra-pair paternity results from female preference for high-quality males in the blue tit. Nature 357:494–496.

Kettlewell HDB, 1955. Selection experiments on industrial melanism in the Lepidoptera. Heredity 9:323–342.

Lewontin RC and Hubby JL, 1966. A molecular approach to the study of genetic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura.* Genetics 54:595–609.

Lewontin RC and White MJD, 1960. Interaction between inversion polymorphisms of the two chromosome pairs in the grasshopper, *Moraba scurra.* Evolution 14: 116–129.

Mason-Gamer RJ, Holsinger KE, and Jansen RK, 1995. Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae). Mol Biol Evol 12:371–381.

Moritz C, Dowling TE, and Brown WM, 1987. Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu Rev Ecol Syst 18:269–292.

Peters JM, Queller DC, Strassmann JE, and Solis CR,

1995. Maternity assignment and queen replacement in a social wasp. Proc R Soc Lond B 260:7–12.

Rave EH, Fleischer RC, Duvall F, and Black JM, 1995. Genetic analyses through DNA fingerprinting of captive populations of Hawaiian geese. Conserv Biol 8:744–751.

Smouse PE and Chakraborty R, 1986. The use of restriction fragment length polymorphisms in paternity analysis. Am J Hum Genet 38:918–939.

Smouse PE and Neel JV, 1977. Multivariate analysis of gametic disequilibrium in the Yanomama. Genetics 85: 733–752.

Smouse PE, Neel JV, and Liu W, 1983. Multiple-locus departures from panmictic equilibrium within and between village gene pools of Amerindian tribes at different stages of agglomeration. Genetics 104:133–153.

Smouse PE, Spielman RS, and Park M-H, 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am Nat 119:445–463.

Soltis DE, Soltis PS, and Milligan BG, 1992. Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. In: Molecular systematics of plants (Soltis P, Soltis D, and Doyle J, eds). New York: Chapman and Hall; 117–150.

Spielman RS and Smouse PE, 1976. Multivariate classification of human populations. I. Allocation of Yanomama Indians to villages. Am J Hum Genet 28:317–331.

Stendid J, Karlsson J-O, and Högberg N, 1994. Intraspecific genetic variation in *Heterobasidion annosum* revealed by amplification of minisatellite DNA. Myco Res 98:57–63.

Takahata N and Palumbi SR, 1985. Extranuclear differentiation and gene flow in the finite island model. Genetics 109:441–457.

Weir BS and Gaut BS, 1993. Matching and binning DNA fragments in forensic science. Jurimetr J 34:9–19.

Welsh J and McClelland M, 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, and Tingey SV, 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535.

Xu S, Kobak C, and Smouse P, 1994. Constrained least squares estimation of mixed population stock composition from mtDNA haplotype frequency data. Can J Fish Aquat Sci 51:417–425.

Received January 17, 1997
Accepted July 8, 1997

Corresponding Editor: Stephen J. O'Brien