REPORT

Estimating Ethnic Admixture from Pedigree Data

Janet S. Sinsheimer,[1,2,3,*] Christopher L. Plaisier,[1] Adriana Huertas-Vazquez,[1] Carlos Aguilar-Salinas,[4] Teresa Tusie-Luna,[5] Päivi Pajukanta,[1] and Kenneth Lange[1,2]

This paper introduces a likelihood method of estimating ethnic admixture that uses individuals, pedigrees, or a combination of individuals and pedigrees. For each founder of a pedigree, admixture proportions are calculated by conditioning on the pedigree-wide genotypes at all ancestry-informative markers. These estimates are then propagated down the pedigree to the nonfounders by a simple averaging process. The large-sample standard errors of the founders' proportions can be similarly transformed into standard errors for the admixture proportions of the descendants. These standard errors are smaller than the corresponding standard errors when each individual is treated independently. Both hard and soft information on a founder's ancestry can be accommodated in this scheme, which has been implemented in the genetic software package Mendel. The utility of the method is demonstrated on simulated data and a real data example involving Mexican families of mixed Amerindian and Spanish ancestry.

Determining the ethnic admixture of individuals with genotype data has become very popular. Estimated admixture proportions are helpful in understanding population histories,[1] satisfying people's curiosity about their family origins,[2,3] and adjusting for ethnic admixture in genetic association studies.[4]

When family data are available, an allelic association with a disease can be detected in several ways.[5] For example, if the phenotype is quantitative, then the measured genotype approach treats allelic contributions as fixed effects and environmental and polygenic background as random effects. Although this approach is powerful, it can lead to false associations when population substructure is present but ignored. Family-based methods such as the transmission disequilibrium test (TDT),[6] the gamete competition model,[7,8] and family-based association test FBAT[9,10] are specifically designed to guard against false inferences in studies with ethnically diverse subjects drawn from ancestral populations differing widely in genetic background and disease risk.[10] The price paid by these safeguarded methods is loss of power. As an alternative approach, covariate adjustment of measured genotypes for ethnic admixture can reduce the chance of false inference while maintaining good power.

In this report, we describe a likelihood method that uses ancestry-informative marker (AIM) genotypes from all available family members to estimate the ethnic admixture proportions of the founders. These estimates are then propagated to the nonfounders by a simple averaging process. The standard errors of the founder estimates can likewise be propagated to the nonfounders. For admixture estimation to have a decent chance of success, markers should be chosen with allele frequencies that clearly separate the ancestral populations.

Our ethnic-admixture-estimation method applies to both pedigrees and unrelated individuals. It estimates an individual's ancestry from K ancestral populations by conditioning on the observed genotypes throughout his or her pedigree. So that excessive computation times with family data can be avoided, it is limited to unlinked markers. For random individuals, this assumption can be relaxed to markers in linkage equilibrium. The method requires accurate specification of the ancestral populations, good estimates of AIM allele frequencies, and AIMs that individually discriminate between at least two of the putative ancestral populations.[11] For the inference of ancestry, typing of one or more unlinked highly polymorphic markers per chromosome is ideal; fewer markers can be used at the expense of precision. Microsatellites, indels, or SNPs are all valid genotyping targets. Microsatellite markers are not necessarily better suited to the method than SNPs because the method can treat closely spaced SNPs as supermarkers when the recombinations are small.[8,12] Because modern likelihood calculations are designed to handle markers with dominant and recessive alleles, there is no need prior to analysis for individuals to be haplotyped at a SNP-combination marker.

Assuming the pedigrees are independent, the likelihood method proceeds pedigree by pedigree. Random individuals count as degenerate pedigrees in this process. In the first stage of estimation for a pedigree, the likelihood of all marker phenotypes scattered across the pedigree is maximized with respect to the ancestral admixture proportions of the founders. Although the likelihood depends on the population allele frequencies at each marker in each ancestral population, these frequencies are not parameters. For a pedigree with n people labeled $1,\dots, n$, let X_i and G_i, respectively, denote the multilocus phenotype and genotype of individual i at all S markers. Because some genotypes may be unknown, the likelihood must sum over all possible values of G_i. Starting from Ott's representation,[13] the likelihood L of the pedigree is

[1]Department of Human Genetics, [2]Department of Biomathematics, [3]Department of Biostatistics, The University of California, Los Angeles, Los Angeles, CA 90095; [4]Department of Endocinology and Metabolism, [5]Molecular Biology and Genomic Medicine Unit, Investigaciones Biomédicas de la UNAM, Instituto Nacional de Ciencias Medicas y Nutricion, Salvador Zubiran, CP14000 Mexico City, Mexico
*Correspondence: janet@mednet.ucla.edu

DOI 10.1016/j.ajhg.2007.12.014. ©2008 by The American Society of Human Genetics. All rights reserved.

$$
\begin{aligned}
L(p) &= \sum_{G_1} \cdots \sum_{G_n} \Pr(X_1,\ldots,X_n \mid G_1,\ldots,G_n)\Pr(G_1,\ldots,G_n) \\
&= \sum_{G_1} \cdots \sum_{G_n} \prod_i \mathrm{Pen}(X_i \mid G_i)\Pr(G_1,\ldots,G_n) \\
&= \sum_{G_1} \cdots \sum_{G_n} \prod_i \mathrm{Pen}(X_i \mid G_i) \prod_j \mathrm{Prior}(G_j) \\
&\quad \prod_{\{c,\ell,m\}} \mathrm{Tran}(G_c \mid G_\ell, G_m) \\
&= \prod_s \sum_{G_{1s}} \cdots \sum_{G_{ns}} \prod_i \mathrm{Pen}(X_{is} \mid G_{is}) \prod_j \mathrm{Prior}(G_{js}) \\
&\quad \prod_{\{c,\ell,m\}} \mathrm{Tran}(G_{cs} \mid G_{\ell s}, G_{ms}).
\end{aligned}
\qquad (1)
$$

Here, X_{is} and G_{is} denote the phenotype and genotype of individual i at marker s. The product rule for likelihoods is in effect because the markers are unlinked. The penetrance function Pen is ordinarily 0 or 1, but it could in principle be more complicated and capture genotyping error. Careful structuring of the Pen function permits the use of noncodominant markers such as SNP-combination markers. The function Tran supplies the usual probability for genetic transmission from parents ℓ and m to their offspring c. In view of our simplifying assumption, Tran incorporates Mendel's laws but no recombination effects.[14] A founder j's probability of belonging to each of the K ancestral populations is determined by the Prior function, which is parameterized by j's admixture proportions. If j is known to have a specific ancestry, then the corresponding admixture proportions are fixed rather than estimated.

The form of the Prior function is unusual and deserves more explanation. Let p_{jk} be the proportion of founder j's ancestry attributable to population k. The p_{jk} are nonnegative and satisfy the constraint $\sum_{k=1}^{K} p_{jk} = 1$. In our simple model, nature chooses a genotype for j at marker s by selecting two random alleles from an infinite pool of possible alleles. Allele a with frequency f_{ka} in ancestral population k is drawn with probability $q_a = \sum_{k=1}^{K} p_{jk} f_{ka}$ from the pool. A genotype a/b for j has the Hardy-Weinberg frequency

$$
\mathrm{Prior}(G_{js} = a/b) = \begin{cases} q_a^2 & a = b \\ 2q_a q_b & a \neq b. \end{cases} \qquad (2)
$$

When all the AIMs are codominant and founders are completely genotyped, offspring genotypes are irrelevant in the determination of ancestry. However, if the founders are not genotyped or incompletely genotyped or if the markers are noncodominant, then the admixture estimates are improved when offspring genotypes are taken into account. Details of the maximum-likelihood estimation and the incorporation of prior ancestry information are given in Appendix A.

Once the founders' admixture proportions have been estimated, the nonfounders' admixture proportions can be calculated. Let w_{cj} be the proportion of individual c's genes that derive from founder j. For consistency, we put $w_{jj} = 1$ and $w_{jh} = 0$ for another founder $h \neq j$. The matrix $W = (w_{cj})$ is computed recursively starting with these boundary values. If a child c has parents ℓ and m, then we compute w_{cj} as the average

$$
w_{cj} = \frac{1}{2}\left(w_{\ell j} + w_{mj}\right),
$$

provided $w_{\ell j}$ and w_{mj} are already known. If we number parents before children, then we can compute all of W in a single sweep starting with the founder values in the upper left-hand corner of W. It is no accident that this looks suspiciously like the classical algorithm for the computation of kinship coefficients. In fact, w_{cj} is twice the kinship coefficient between c and founder j. Given the w_{cj}, it makes sense to compute the proportion p_{ck} of c's ancestry due to population k as the weighted average

$$
p_{ck} = \sum_j w_{cj} p_{jk}. \qquad (3)
$$

Again, the p_{ck} are nonnegative and satisfy the constraint $\sum_{k=1}^{K} p_{ck} = 1$. To estimate p_{ck}, we simply substitute the estimate $\hat{p}_{jk}$ of p_{jk} in Equation 3 for each founder j. This can produce results that are slightly odd on first sight. For instance, although two siblings might have inherited different genes from their parents, their estimated admixture proportions are always exactly the same. This apparent anomaly is not worrisome because their ancestral proportions across the entire genome should be identical.

Standard errors for the founders' admixture proportions are computed from the observed information matrix. In view of Equation 3, we have

$$
\mathrm{Var}(\hat{p}_{ck}) = \sum_j w_{cj}^2 \mathrm{Var}(\hat{p}_{jk}) + 2\sum_j \sum_{h<j} w_{cj} w_{ch} Cov(\hat{p}_{jk}, \hat{p}_{hk}) \qquad (4)
$$

for any nonfounder c, where h and j range over all founders. Because $0 \leq w_{cj} \leq 1/2$ for all c and j when there is no inbreeding and $\mathrm{Cov}(\hat{p}_{jk}, \hat{p}_{hk})$ is often nearly zero, the variances of an offspring's estimates are very often less than the weighted average of variances of the founder's estimates.

To demonstrate the utility of the method, we apply it to a real data example involving Mexican families of mixed Amerindian and Spanish ancestry. We then turn to a carefully designed simulation study to test the properties of the method. Readers interested the nuts and bolts of running Mendel on their own data can refer to the Mendel documentation.[12]

We now consider the admixture problem for six multigeneration Mexican families from Mexico City. These families were recruited by the Lipid Clinic of the Instituto Nacional de Ciencias Medicas y Nutricion Salvador Zubiran (INCMNSZ) as part of a study on the genetics of familial combined hyperlipidemia (FCHL). Each subject provided written informed consent as part of the original study, and approval was obtained by the Institutional Committee of Biomedical Research in Humans of the INCMNSZ.[15] By using the pedigree-trimming option of Mendel,[16] we excluded ungenotyped family members who are unnecessary in determining the relationships among genotyped members. The trimmed pedigrees each contain from 15 to 23

Table 1. An Excerpt of the Summary Output from the Program Mendel

Admixture Coefficients Pedigree by Pedigree

PEDIGREE NAME	PERSON NAME	POPULATION NAME	ESTIMATED PROPORTION	STD ERROR OF ESTIMATE
3	53	AM_IND	0.9255	0.2077
3	53	SPANISH	0.0745	0.2077
3	56	AM_IND	0.0638	0.2744
3	56	SPANISH	0.9362	0.2744
3	55	AM_IND	0.4946	0.1462
3	55	SPANISH	0.5054	0.1462
3	54	AM_IND	0.4946	0.1462
3	54	SPANISH	0.5054	0.1462
3	AVERAGE	AM_IND	0.6058	
3	AVERAGE	SPANISH	0.3942	

members and from three to six sibships, for a total of 27 sibships in the entire dataset. Only three of the 76 offspring have no genotypes. In contrast, many of the founders are completely untyped. In family 5, none of the six founders is available for typing. At the other extreme, all seven founders of families 1 and 2 are at least partially genotyped. We checked for genotyping errors by using the mistyping option of Mendel[17] and removed inconsistent genotypes.

All family members are Mestizos whose ancestry is predominantly a mixture of Spanish and Amerindian. The proportion of African ancestry is negligible in the families used in this study.[15] The low likelihood of African ancestry is consistent with previous studies of Mestizos from Mexico City, where the estimated proportion of European ancestry is between 34.8% and 70.8%, the proportion of Amerindian ancestry is between 27.6% and 56.2%, and the proportion of African ancestry is between 0.9% and 6.2%. These previous studies are summarized by Bonilla and coworkers.[18] Thus, we limited the ancestral populations to Spanish and Amerindian.

Bonilla and coworkers assembled an AIM panel to estimate ethnic admixture in Hispanics.[18–21] The ancestral populations with published allele frequencies at these markers include Spaniards, Mayans, Nahuas, and Southwestern Native Americans (Cheyenne, Pima, and Pueblo). All individuals with evidence of admixture were excluded from the calculation of the allele frequencies for the ancestral populations.[19] Allele frequencies and other information on the AIMs are available in dbSNP, submitter id PSU_ANTH.

For our purposes, we selected nine unlinked AIMs that show greater than 30% absolute difference between Spanish and Amerindian allele frequencies and have no direct or indirect connection with FCHL susceptibility. Because the specific Amerindian origin varies among Mestizos, we used the average of the Mayan, Nahua, and Southwestern Native American frequencies in this study. For almost all of the markers, the allele frequencies differ by less than 10% among the three Amerindian reference groups. As a check on our assumption of regional homogeneity, we estimated ethnic admixture using each of the three Amerindian reference group frequencies (Mayan, Nahua, and Southwestern Native Americans) and found essentially no differences in our conclusions (data not shown).

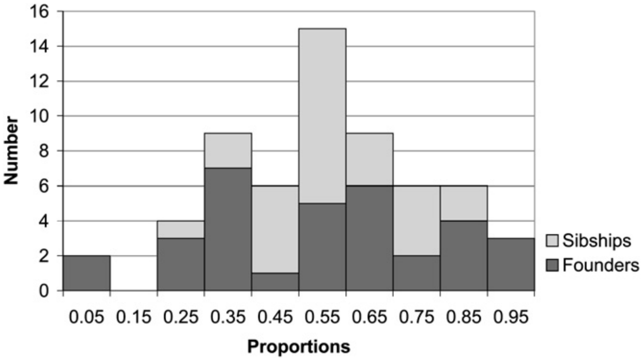

Figure 1. Estimated Amerindian Proportions in Founders and Sibships
The light-colored blocks represent the number of sibships with the indicated Amerindian ancestry proportions, and the darker blocks represent the number of founders with the indicated American ancestry proportions.

If we want to include a Dirichlet prior for each ancestor, then we must convey the prior counts to Mendel. For example, if we suspect that each of these families has slightly more Spanish than Amerindian ancestry, then in a ratio 0:56:0.44, we specify 2.24 counts favoring Spanish ancestry and 1.76 prior counts favoring Amerindian ancestry. The choice of 4 for the sum was selected empirically so that the prior would have a moderately strong effect on the results. In such a situation, the ancestral origins of the families are well known and consistent across families.

The program Mendel produces a summary file that gives the admixture proportions and their standard errors for each person, pedigree-averaged admixture proportions, as well as a new pedigree file that can be used as an input file in further analyses. Table 1 is an excerpt of the summary file. Individuals 53 and 56 are the parents of individuals 54 and 55. Although the genotypes for the siblings 54 and 55 differ at two markers (data not shown), their estimated admixture proportions are identical (Table 1). No genotype data are available for either founder 53 or 56, and all of their offspring are in common. We know that one of these two founders has predominantly Spanish ancestry and the other has predominantly Amerindian ancestry, but we do not know which is which. Therefore, their estimated admixture proportions can be swapped.

Figure 1 displays the distribution of the Amerindian proportions for the 33 founders and the 27 sibships. For the founders, the mean proportion of Amerindian ancestry is 0.553, the median is 0.598, the range is from 0.064 to 0.953, and the average standard error ($\overline{SE}$) is 0.262. The mean and median proportions of 0.576 and 0.538 of Amerindian ancestry for the nonfounders are similar to the corresponding values for the founders. However, nonfounder proportions show much smaller range, from 0.286

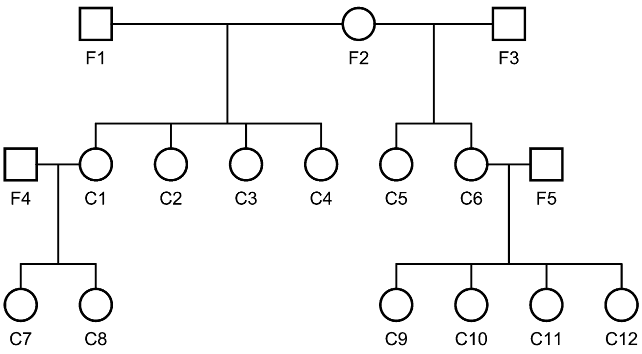

Figure 2. Structure of the Simulated Pedigrees
The presented pedigree structure (with five founders and 12 offspring) was used in all simulations for the creation of genotype data for 500 multigenerational pedigrees.

to 0.898, and tend to have smaller standard errors ($\overline{SE} = 0.146$). The use of a Dirichlet prior with population prior counts of 2.24 for Spanish ancestry and 1.76 for Amerindian ancestry decreases the mean Amerindian proportion for founders to 0.501 and the mean for nonfounders to 0.514. The $\overline{SE}$s[22] of the founder and nonfounder estimates decline to 0.176 and 0.106. The range decreases to span from 0.241 to 0.741 for founders and to span from 0.356 to 0.783 for nonfounders.

To examine the properties of our method, we simulate data under a variety of scenarios. In all scenarios, we use the pedigree structure shown in Figure 2 and create genotype data for 500 multigenerational pedigrees (each with five founders and 12 offspring) by using the gene-dropping option of Mendel.[16] The structure of the pedigrees mimics that of the Mexican pedigrees. We first examine the effects of missing founder genotypes (columns 3–11, Table 2). Then, by using pedigrees where grandparents F1–F3 are untyped, we examine the impact of varying the number of markers (columns 12–14, Table 2), the informativeness of the markers (columns 3–14, Table 3), and allele-frequency misspecification (columns 15–17, Table 3).

Our most informative simulation scenario involves 46 unlinked AIMs with allele frequencies of 0.9 and 0.1 in one population and 0.1 and 0.9 in the other. We choose these allele frequencies according to the criteria of Mao et al.[23] These authors selected AIMs on the basis of the standardized variance of the allele frequencies, $SV = (f_{1a} - f_{2a})^2/4f(1-f)$, where f_{ka} is the a allele frequency in population k and $f = (f_{1a} + f_{2a})/2$. Choosing the top two most informative markers from chromosomes 3–22 and the top three from chromosomes 1 and 2, the average SV for the Mao data is 0.62, consistent with the $\overline{SV} = 0.64$ for our simulated data. Our least informative choice of allele frequencies of 0.75 and 0.25 for one population and 0.25 and 0.75 in the other leads to $\overline{SV} = 0.25$, which is less than the average SV of the nine AIMs used in the Mexican family example ($\overline{SV} = 0.31$).

We find that the precision of the ancestral proportions for founders is highly dependent on whether they are genotyped (columns 3–11, Table 2). Untyped grandparents cause a small reduction in precision for the grandchildren's ancestral proportions even when their parents are fully genotyped (see entries for C1–C12, columns 3–8, Table 2). When analyzed as part of a large pedigree, the number of siblings has little effect on the precision even when all founders are untyped (compare, for example, entries in columns 9–12, Table 2, for C1–C4 to the entries for C5–C6). As predicted by Equation 4, offspring generally have smaller standard errors than founders. Not surprisingly, the precision depends on the number of markers. The average standard error ($\overline{SE}$) and the absolute difference from the actual values ($\overline{d}$) decrease approximately 2-fold as the number of AIMs increases from 9 to 46 (columns 12–14 versus columns 6–8, Table 2). Likewise, as the informativeness of the AIMs increases, these precision measures improve (columns 3–14, Table 3).

We also examined the effects of the misspecification of the AIM allele frequencies by varying the stated frequencies by ± 0.10 from their true values (Table 3). Specifically, we are interested in whether the misspecification of half of

Table 2. Effects of Missing Genotypes and Number of Markers on Estimated Ancestry

		No Missing Genotypes, $S^a = 46$			F1–F3 Untyped, $S = 46$			F1–F5 Untyped, $S = 46$			F1–F3 Untyped, $S = 9$		
Pedigree Member	p_{i1}[b]	$\widehat{p}_{i1}$	$\overline{SE}$[c]	$\overline{d}$[d]	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$
F1	0.750	0.749	0.063	0.050	0.744	0.092	0.074	0.744	0.092	0.074	0.739	0.159	0.145
F2	0.500	0.493	0.065	0.052	0.499	0.090	0.072	0.499	0.090	0.072	0.501	0.168	0.153
F3	0.250	0.256	0.060	0.048	0.249	0.095	0.081	0.249	0.095	0.081	0.256	0.176	0.154
F4	0.500	0.506	0.065	0.051	0.494	0.065	0.051	0.490	0.088	0.070	0.504	0.137	0.104
F5	0.500	0.504	0.065	0.051	0.504	0.066	0.052	0.506	0.074	0.063	0.510	0.104	0.119
C1–C4	0.625	0.621	0.044	0.036	0.621	0.047	0.039	0.621	0.047	0.039	0.619	0.101	0.085
C5–C6	0.375	0.375	0.044	0.034	0.374	0.052	0.043	0.374	0.052	0.043	0.378	0.111	0.095
C7–C8	0.562	0.559	0.031	0.031	0.558	0.040	0.032	0.556	0.041	0.039	0.562	0.088	0.068
C9–C12	0.438	0.439	0.039	0.031	0.439	0.042	0.034	0.439	0.045	0.039	0.444	0.091	0.075

[a] S denotes the number of unlinked markers.
[b] p_{i1} denotes the true proportion of population 1 ancestry, and $\widehat{p}_{i1}$ denotes the estimated proportion of population 1 ancestry.
[c] $\overline{SE}$ denotes the average standard error.
[d] $\overline{d}$ denotes the average absolute difference between calculated and actual ancestral proportions.

Table 3. Estimated Ancestry as a Function of the Marker Informativeness and Allele-Frequency Misspecification

Pedigree Member	p_{i1}[c]	$f_{1a} = 0.9, f_{2a} = 0.1$,[a] No Misspec[b]			$f_{1a} = 0.83, f_{2a} = 0.17$, No Misspec			$f_{1a} = 0.8, f_{2a} = 0.2$, No Misspec			$f_{1a} = 0.75, f_{2a} = 0.25$, No Misspec			$f_{1a} = 0.8, f_{2a} = 0.2$, 0.1 Misspec		
		$\widehat{p}_{i1}$	$\overline{SE}$[d]	$\overline{d}$[e]	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$
F1	0.750	0.744	0.092	0.074	0.750	0.093	0.078	0.734	0.122	0.101	0.746	0.149	0.122	0.712	0.110	0.095
F2	0.500	0.499	0.090	0.072	0.501	0.093	0.076	0.511	0.126	0.101	0.497	0.148	0.119	0.506	0.110	0.092
F3	0.250	0.249	0.095	0.081	0.253	0.104	0.084	0.248	0.132	0.108	0.268	0.164	0.132	0.280	0.122	0.099
F4	0.500	0.494	0.065	0.051	0.503	0.070	0.054	0.501	0.086	0.069	0.502	0.102	0.082	0.499	0.082	0.065
F5	0.500	0.504	0.066	0.052	0.500	0.070	0.052	0.500	0.086	0.072	0.507	0.103	0.083	0.500	0.082	0.069
C1–C4	0.625	0.621	0.047	0.039	0.626	0.050	0.040	0.622	0.064	0.052	0.621	0.078	0.065	0.609	0.060	0.051
C5–C6	0.375	0.374	0.052	0.043	0.378	0.057	0.044	0.378	0.070	0.058	0.383	0.085	0.068	0.393	0.066	0.055
C7–C8	0.562	0.558	0.040	0.032	0.564	0.043	0.034	0.562	0.054	0.043	0.562	0.064	0.054	0.554	0.051	0.042
C9–C12	0.438	0.439	0.042	0.034	0.439	0.045	0.036	0.440	0.054	0.047	0.445	0.066	0.052	0.447	0.053	0.045

[a] f_{1a} denotes the a allele frequency in population 1, and f_{2a} denotes the a allele frequency in population 2.

[b] Misspec signifies the degree of allele-frequency mispecification where 0.1 denotes that the major alleles are mispecified by 0.1 from their true values.

[c] p_{i1} denotes the true proportion of population 1 ancestry, and $\widehat{p}_{i1}$ denotes the estimated proportion of population 1 ancestry.

[d] $\overline{SE}$ denotes the average standard error.

[e] $\overline{d}$ denotes the average absolute difference between calculated and actual ancestral proportions.

the marker frequencies as $f_{1a} = 0.90$, $f_{2a} = 0.10$ and the misspecification of the other half as $f_{1a} = 0.70$, $f_{2a} = 0.30$, when the true marker frequencies in the two populations are $f_{1a} = 0.80$, $f_{2a} = 0.20$ for all the markers, produces a bias as measured as the difference between the mean and the actual ancestral proportion. The ancestral proportions show a meaningful degree of bias with this much misspecification. The bias is slight when the misspecification is ± 0.05 of the true marker frequencies (data not shown).

As further validation of Mendel, we compare our results to those obtained with Structure.[24,25] We simulate data at 46 unlinked AIMs for 20 unrelated individuals from one population (allele frequencies at each marker $f_{1a} = 0.1$ and $f_{1b} = 0.9$), 20 related individuals from another population (allele frequencies at each marker $f_{2a} = 0.9$ and $f_{2b} = 0.1$), and five unrelated, admixed individuals. In the implementation of Structure, the 40 individuals of known ancestry are used for the estimation of allele frequencies, but their ancestries are not estimated. The Structure and Mendel proportions are quite close even when the original allele frequencies are used in Mendel. The average absolute difference between the Structure and Mendel estimates is 0.017, and the absolute differences range from 0.010 to 0.029. The estimates are, in general, even closer when the Structure-derived frequencies are used in Mendel. In this case, the average absolute difference between the Structure and Mendel estimates is 0.012, and the absolute differences range from 0.002 to 0.023.

Because our method can use AIMs in linkage equilibrium (LE) when individuals are unrelated, it is of interest to determine the effect of ignoring family structure. We compare the bias, $\overline{SE}$, and $\overline{d}$ under two methods of estimation and three scenarios. In each scenario, grandparents are untyped. Method R kept the family intact; method U treats the offspring as unrelated. The three scenarios are (1) offspring genotyped at 46 unlinked AIMs with $\overline{SV} = 0.64$ (columns 3–8, Table 4), (2) offspring genotyped at 200 markers in LE with $\overline{SV} = 0.64$ (columns 9–11, Table 4), and (3) offspring genotyped at 200 AIMs in LE with $\overline{SV} = 0.55$ (columns 12–14, Table 4). Scenario 2 is included so that the reader can judge the effects of more AIMs with the same informativeness as the unlinked AIMs. Unfortunately, this comparison is not completely realistic because

Table 4. Treating Data as Pedigrees versus Unrelated Individuals

Pedigree Member	p_{i1}[d]	R[a], S[b] $= 46$, $\overline{SV}$[c] $= 0.64$			U, $S = 46$, $\overline{SV}$[c] $= 0.64$			U, $S = 200$, $\overline{SV}$[c] $= 0.64$			U, $S = 200$, $\overline{SV}$[c] $= 0.55$		
		$\widehat{p}_{i1}$	$\overline{SE}$[e]	$\overline{d}$[f]	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\overline{d}$
C1–C4	0.625	0.621	0.047	0.039	0.620–0.623	0.064	0.049–0.051	0.623–0.628	0.030	0.042–0.044	0.625–0.627	0.033	0.042–0.044
C5–C6	0.375	0.374	0.052	0.043	0.372–0.376	0.063	0.050–0.053	0.373–0.374	0.030	0.041–0.044	0.376–0.378	0.033	0.045–0.046
C7–C8	0.562	0.558	0.040	0.032	0.553–0.556	0.064–0.065	0.052–0.054	0.561–0.566	0.031	0.044–0.045	0.562–0.563	0.035	0.046
C9–C12	0.438	0.439	0.042	0.034	0.438–0.442	0.064	0.051–0.054	0.437–0.438	0.031	0.043–0.044	0.437–0.442	0.034	0.047

[a] R stands for "related" and signifies that the pedigree is analyzed intact; U stands for "unrelated" and signifies that the offspring are treated as though they are unrelated.

[b] S denotes the number of unlinked markers.

[c] $\overline{SV}$ denotes the standardized variance.

[d] p_{i1} denotes the true proportion of population 1 ancestry, and $\widehat{p}_{i1}$ denotes the estimated proportion of population 1 ancestry.

[e] $\overline{SE}$ denotes the average standard error.

[f] $\overline{d}$ denotes the average absolute difference between calculated and actual ancestral proportions.

Table 5. Misspecifying $K = 2$ as $K = 3$

Pedigree Member	p_{i1}[a],p_{i2}	$\widehat{p}_{i1}$	$\overline{SE}$[b]	$\overline{d}$[c]	$\widehat{p}_{i2}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i3}$	$\overline{SE}$	$\overline{d}$
F1	0.750,0.250	0.706	0.120	0.086	0.240	0.099	0.085	0.054	0.076	0.054
F2	0.500,0.500	0.478	0.096	0.071	0.488	0.097	0.072	0.034	0.053	0.034
F3	0.250,0.750	0.194	0.104	0.099	0.719	0.119	0.110	0.087	0.097	0.087
F4	0.500,0.500	0.484	0.074	0.060	0.476	0.074	0.059	0.040	0.045	0.040
F5	0.500,0.500	0.471	0.079	0.067	0.464	0.079	0.065	0.065	0.068	0.065
C1–C4	0.625,0.375	0.592	0.058	0.052	0.364	0.056	0.049	0.044	0.053	0.044
C5–C6	0.375,0.625	0.336	0.062	0.058	0.604	0.065	0.054	0.060	0.063	0.060
C7–C8	0.562,0.438	0.538	0.047	0.042	0.420	0.047	0.039	0.042	0.043	0.042
C9–C12	0.438,0.562	0.404	0.058	0.050	0.534	0.078	0.064	0.062	0.091	0.075

K denotes the number of populations.

[a] p_{ij} denotes the true proportion of population j ancestry, and $\widehat{p}_{ij}$ denotes the estimated proportion of population j ancestry.

[b] $\overline{SE}$ denotes the average standard error.

[c] $\overline{d}$ denotes the average absolute difference between calculated and actual ancestral proportions.

200 AIMs in LE with $\overline{SV} = 0.64$ are currently unavailable for distinguishing between Amerindians and Europeans. By using the supplemental data from Mao et al.,[23] we calculate that the top 200 AIMs have $\overline{SV} = 0.55$, as suggested by scenario 3. When the same unlinked markers are used (scenario 1), method R has greater precision than method U; both show little bias. Detailed inspection of Table 4 suggests that method R under scenario 1 is roughly equivalent to method U under both scenarios 2 and 3.

We next examine the effects of the misspecification of the number of ancestral populations. Misspecification of *K* can occur in two ways. The number of populations can be overestimated, or it can be underestimated. To investigate the effects of assuming too many populations, we reanalyze the scenario 1 data assuming $K = 3$ (Table 5). We assume that for 23 of these AIMs, population 3's allele frequencies are the same as population 1's allele frequencies, and that for the other 23 AIMs, population 3's allele frequencies are the same as population 2's allele frequencies. The ancestral proportions estimated for population 3 average less than 10% for all family members. Hence, the effect of over-assigning the number of ancestral populations, in this case, is that a small fraction of an individual's ancestry is incorrectly attributed to a third population. Because the standard errors are of the same magnitude as the estimated population 3 proportions, most users would be wary of the population 3 assignment.

To determine the effects of under specifying *K*, we generate data for 45 unlinked AIMs where 15 markers separate populations 1 and 2 from population 3, 15 markers separate populations 1 and 3 from population 2, and 15 markers separate populations 2 and 3 from population 1. For each set of 15 markers, two of three populations have the same AIM frequencies of 0.1 for one allele and 0.9 for the other, and these frequencies are reversed in third population. We first analyze the data generated with 3 populations assuming $K = 3$ (columns 3–11, Table 6). Our method R estimates are accurate, but precision has decreased because of the reduction in AIMs that distinguish between each of the populations. There is still sufficient data to differentiate the ancestral proportions for individuals with relativity small degrees of population 3 ancestry (see individuals C1–C12, column 9, Table 6) from the ancestral proportions for individuals with no population 3 ancestry (see individuals F2, F4, and F5,

Table 6. Misspecifying $K = 3$ as $K = 2$

		K[a] $= 3$									$K = 2$			
Pedigree Member	p_{i1}[b],p_{i2},p_{i3}	$\widehat{p}_{i1}$	$\overline{SE}$[c]	$\overline{d}$[d]	$\widehat{p}_{i2}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i3}$	$\overline{SE}$	$\overline{d}$	$\widehat{p}_{i1}$	$\overline{SE}$	$\widehat{p}_{i2}$	$\overline{SE}$
F1	0.500,0.250,0.250	0.499	0.121	0.104	0.277	0.103	0.093	0.223	0.116	0.101	0.610	0.112	0.390	0.112
F2	0.500,0.500,0.000	0.480	0.118	0.093	0.470	0.112	0.087	0.050	0.050	0.050	0.502	0.109	0.498	0.109
F3	0.250,0.500,0.250	0.278	0.129	0.099	0.516	0.134	0.116	0.206	0.111	0.097	0.380	0.125	0.620	0.125
F4	0.500,0.500,0.000	0.487	0.084	0.062	0.487	0.084	0.062	0.026	0.035	0.026	0.500	0.080	0.500	0.080
F5	0.500,0.500,0.000	0.501	0.083	0.059	0.480	0.083	0.059	0.019	0.028	0.019	0.510	0.080	0.490	0.080
C1–C4	0.500,0.375,0.125	0.488	0.066	0.054	0.375	0.065	0.051	0.137	0.055	0.043	0.556	0.060	0.444	0.060
C5–C6	0.375,0.500,0.125	0.378	0.072	0.054	0.495	0.073	0.051	0.127	0.060	0.043	0.441	0.066	0.559	0.066
C7–C8	0.500,0.438,0.062	0.488	0.053	0.038	0.431	0.053	0.041	0.081	0.036	0.030	0.528	0.050	0.472	0.050
C9–C12	0.438,0.500,0.062	0.440	0.055	0.038	0.487	0.050	0.038	0.073	0.037	0.043	0.476	0.052	0.524	0.052

[a] *K* denotes the number of populations assumed in the estimation.

[b] p_{ij} denotes the true proportion of population j ancestry, and $\widehat{p}_{ij}$ denotes the estimated proportion of population j ancestry.

[c] $\overline{SE}$ denotes the average standard error.

[d] $\overline{d}$ denotes the average absolute difference between calculated and actual ancestral proportions.

column 9, Table 6). However, based on our results, it might be preferable in practice to use method U with additional linked AIMs rather than method R with unlinked AIMs in dealing with more than three ancestral populations. If we misspecify the analysis by analyzing data generated from three populations while assuming $K = 2$, we get what appear to be reasonable estimates (columns 12–15, Table 6). This result clearly demonstrates that it is important in using Mendel's ethnic admixture option to be confident about the number and nature of the ancestral populations.

The controversies generated by genetic-association studies stem from the failure of researchers to adjust for ethnic admixture.[26] Making such adjustment easy will encourage better statistical analysis. Most methods that estimate ethnic admixture assume that genotyped individuals are not closely related,[24] with a notable exception[27] that uses self-reported ancestry. Mendel adopts a reasonable likelihood model that takes pedigrees rather than random individuals as the unit of analysis. As our simulation examples illustrate, this can reduce the parameter standard errors by as much as 1/3. Estimating founder admixture proportions first and then propagating these to nonfounders by repeated averaging minimizes the number of primary parameters. The connection with kinship coefficients is both natural and esthetically pleasing. Prior evidence on a society's ethnic background can be exploited by the introduction of Dirichlet priors. This is a good idea when genotyping data are sparse. One caution that should be kept in mind is that the standard errors both for the founders' proportions and for the offspring proportions do not incorporate the uncertainty due to allele-frequency estimation.

The limitations of likelihood-based estimation should be respected. The foremost limitations are that the number and nature of the ancestral populations must be known and that markers that discriminate among them must be employed. Misspecification of the number of ancestral populations can profoundly impact study conclusions. In our simulations, the over-specification of K is less of a problem than is the under-specification of K, but either error can cause confusion. In contrast, minor misspecification of the allele frequencies does not drastically affect the results. In estimating the ethnic admixture of Mestizo families from Mexico City, we used averaged Amerindian allele frequencies that span a number of possible Amerindian ancestries. The AIMs we used in this study differ much more between the Spanish and Amerindians than within Amerindian groups. If suitable reference frequencies had been unavailable, we could have collected an additional sample of unrelated Mestizo individuals and used a program like Structure[24,25] to estimate the number of ancestral groups and the ancestral allele frequencies. Regardless of how ancestral allele frequencies are derived, these can be readily fed into Mendel.

The other important limitation is the assumption of unlinked markers. Relaxing this assumption with pedigree data entails a sharp increase in computational complexity. For isolated individuals, the less demanding assumption of linkage equilibrium can be substituted. In our simulations, the use of 200 linked AIMs in LE and treatment of family members as unrelated gave roughly equivalent results to the use of 46 unlinked AIMs and pedigrees. Even with unlinked AIMs, the computational speed can be too slow for very large pedigrees with marriage loops. Mendel flags pedigrees that are too complex for analysis. These pedigrees can often be broken up into subpedigrees without too much loss of information. A more satisfying solution might be to replace the exact pedigree-likelihood calculations with Markov chain Monte Carlo (MCMC) approximations.

There are several features of using pedigrees in Mendel's implementation that deserve comment. First, the ancestries of individuals' parents can not be inferred if the individuals are treated as unrelated. For example, suppose individual c with parents ℓ and m has ancestral proportions $p_{c1} = 0.5$ and $p_{c2} = 0.5$. When c's is treated as an unrelated individual, the most we know about her parents' ancestries is that $p_{\ell 1} = t$ and $p_{m1} = 1 - t$ where $0 \leq t \leq 1$. That is, an infinite number of combinations of parental ancestries are possible. Second, because Mendel can handle noncodominant markers,[8,12] tightly linked SNPs in linkage disequilibrium can be used with pedigrees. Mendel's SNP combining utility makes this easy. Thus, if several moderately good AIMs are tightly linked, they can be combined to produce an even more informative AIM.

The program Mendel is straightforward to use and produces high-quality estimates of ethnic admixture. Not only are admixture proportions immediately usable in variance components models for association, they are also applicable in penetrance estimation with generalized linear models. It is worth pointing out that a new analysis option of Mendel makes generalized linear models a fruitful avenue of statistical analysis with pedigree data. By itself, inclusion of ethnic admixture will not revolutionize statistical genetics. Seen as another tool in the increasingly sophisticated toolkit of statistical geneticists, it will have an important impact.

Appendix A

Our computer program Mendel maximizes the log likelihood by recursive quadratic programming with quasi-Newton updates to the observed information $-d^2 \ln L(p)$ subject to the constraint $\sum_k p_{jk} = 1$ for each founder j.[12,14] At each iteration, the current approximation to $-d^2 \ln L(p)$ is improved by a rank-one perturbation. Soft prior information on ethnic admixture can incorporated by multiplication of the likelihood (1) by a separate Dirichlet prior for each founder. These independent and identically distributed priors steer maximum a posteriori estimates toward reasonable values when typing is sparse. The prior is multiplied by the likelihood to create the joint likelihood,

$$L_{joint} \propto L(p) \prod_j \prod_k p_{jk}^{v_k+1-1}, \tag{5}$$

where the pseudocount v_k supports ancestry k. In maximum a posteriori estimation, the joint likelihood (5) is maximized in the same manner as the original log likelihood.

Acknowledgments

The authors thank the families for the participation. They also thank Laura Riba and Maribel Rodriguez for recruiting these families, collecting the samples, and maintaining the pedigree database. The research was supported in part by United States Public Health Service grants MH59490, GM53275, T32-HG002536, HL082762, HL-28481, and HL-70150, as well as by American Heart Association grant 0430180N.

Received: May 9, 2007
Revised: October 1, 2007
Accepted: December 11, 2007
Published online: March 6, 2008

Web Resources

The URLs for data presented herein are as follows:

dbSNP, http://www.ncbi.nih.gov/SNP/
Mendel Software, http://www.genetics.ucla.edu/software/

References

1. Martinez Marignac, V.L., Bertoni, B., Parra, E.J., and Bianchi, N.O. (2004). Characterization of admixture in an urban sample from Buenos Aires, Argentina, using uniparentally and biparentally inherited genetic markers. Hum. Biol. *76*, 543–557.
2. Gollust, S.E., Wilfond, B.S., and Hull, S.C. (2003). Direct-to-consumer sales of genetic services on the Internet. Genet. Med. *5*, 332–337.
3. Shriver, M.D., and Kittles, R.A. (2004). Genetic ancestry and the search for personalized genetic histories. Nat. Rev. Genet. *5*, 611–618.
4. Redden, D.T., Divers, J., Vaughan, L.K., Tiwari, H.K., Beasley, T.M., Fernandez, J.R., Kimberly, R.P., Feng, R., Padilla, M.A., Liu, N., et al. (2006). Regional admixture mapping and structured association testing: Conceptual unification and an extensible general linear model. PLoS Genet *25*, e137.
5. Lange, K., Sinsheimer, J.S., and Sobel, E. (2005). Association testing with Mendel. Genet. Epidemiol. *29*, 36–50.
6. Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium - the insulin gene function and insulin-dependent diabetes-mellitus (IDDM). Am. J. Hum. Genet. *52*, 506–516.
7. Sinsheimer, J.S., Blangero, J., and Lange, K. (2000). Gamete-competition models. Am. J. Hum. Genet. *66*, 1168–1172.
8. Sinsheimer, J.S., McKenzie, C.A., Keavney, B., and Lange, K. (2001). SNPs and snails and puppy dogs' tails: Analysis of SNP haplotype data using the gamete competition model. Ann. Hum. Genet. *65*, 483–490.
9. Horvath, S., Xu, X., and Laird, N.M. (2001). The family based association test method: Strategies for studying general genotype-phenotype associations. Eur. J. Hum. Genet. *9*, 301–306.
10. Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nat. Rev. Genet. *7*, 385–394.
11. Parra, E.J., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., and Shriver, M.D. (1998). Estimating African American admixture proportions by use of population specific alleles. Am. J. Hum. Genet. *63*, 1839–1851.
12. Lange, K., Cantor, R., Horvath, S., Perola, M., Sabatti, C., Sinsheimer, J., and Sobel, E. (2001). Mendel version 4.0: A complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am. J. Hum. Genet. *69* (Suppl. 1), 1886.
13. Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am. J. Hum. Genet. *26*, 588–597.
14. Lange, K. (2002). Mathematical and Statistical Analysis for Genetic Analysis (New York: Springer).
15. Huertas-Vazquez, A., Aguilar-Salinas, C., Lusis, A.J., Cantor, R.M., Canizales-Quinteros, S., Lee, J.C., Mariana-Nunez, L., Riba-Ramirez, R.M., Jokiaho, A., Tusie-Luna, T., and Pajukanta, P. (2005). Familial combined hyperlipidemia in Mexicans. Association with upstream transcription factor 1 and linkage on chromosome 16q24.1. Arterioscler. Thromb. Vasc. Biol. *25*, 1985–1991.
16. Lange, K., and Sinsheimer, J.S. (2004). The pedigree triming problem. Hum. Hered. *58*, 108–111.
17. Sobel, E., Papp, J.C., and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. Am. J. Hum. Genet. *70*, 496–508.
18. Bonilla, C., Gutierrez, G., Parra, E.J., Kline, C., and Shriver, M.D. (2005). Admixture analysis of a rural population of the state of Guerrero, Mexico. Am. J. Phys. Anthropol. *128*, 861–869.
19. Bonilla, C., Parra, E.J., Pfaff, C.L., Dios, S., Marshall, J.A., Hamman, R.F., Ferrell, R.E., Hoggart, C.L., McKeigue, P.M., and Shriver, M.D. (2004). Admixture in the Hispanics of the San Luis Valley, Colorado, and its implications for complex trait gene mapping. Ann. Hum. Genet. *68*, 139–153.
20. Bonilla, C., Shriver, M.D., Parra, E.J., Jones, A., and Fernandez, J.R. (2004). Ancestral proportions and their associations with skin pigmentation and bone mineral density in Puerto Rican women from New York City. Hum. Genet. *115*, 57–68.
21. Shriver, M.D., Parra, E.J., Dios, S., Bonilla, C., Norton, H., Jovel, C., Pfaff, C., Jones, C., Massac, A., Cameron, N., et al. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. Hum. Genet. *112*, 397–399.
22. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). Bayesian Data Analysis (Boca Raton, Florida: Chapman and Hall/CRC).
23. Mao, X., Bigham, A.W., Mei, R., Gutierrez, G., Weiss, K.M., Brutsaert, T.D., Leon-Velarde, F., Moore, L.G., Vargas, E., McKeigue, P.M., et al. (2007). A genomewide admixture mapping panel for hispanic/latino populations. Am. J. Hum. Genet. *80*, 1171–1178.
24. Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
25. Falush, D., Stephens, M., and Pritchard, J.K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics *164*, 1567–1587.
26. Redden, D.T., and Allison, D.B. (2003). Nonreplication in genetic association studies of obesity and diabetes research. J. Nutr. *133*, 3323–3326.
27. Skol, A.D., Xiao, R., Boehnke, M., and Veteran Affairs Cooperative Study (2005). An algorithm to construct genetically similar subsets of families with the use of self-reported ethnicity information. Am. J. Hum. Genet. *77*, 346–354.