*Am. J. Hum. Genet.* 77:346–354, 2005

# An Algorithm to Construct Genetically Similar Subsets of Families with the Use of Self-Reported Ethnicity Information

Andrew D. Skol,[1] Rui Xiao,[1] Michael Boehnke,[1] and Veterans Affairs Cooperative Study 366 Investigators*

[1]Department of Biostatistics, University of Michigan, Ann Arbor

**We present a simple algorithm that uses self-reported ethnicity information, pedigree structure, and affection status to group families into genetically more homogeneous subsets. This algorithm should prove useful to researchers who wish to perform genetic analyses on more-homogeneous subsets when they suspect that ignoring heterogeneity could lead to false-positive results or loss of power. We applied our algorithm to the self-reported ethnicity information of 159 families from the Veterans Affairs Cooperative Study of schizophrenia. We compared these estimates of population membership with those obtained using the program *structure* in an analysis of 378 microsatellite markers. We found excellent concordance between family classifications determined using self-reported ethnicity information and our algorithm and those determined using genetic marker data and *structure*; 158 of the 159 families had concordant classifications. In addition, the degree of admixture estimated using our algorithm and self-reported ethnicity information correlated well with that predicted using the genotype information.**

## Introduction

Many genetic studies of psychiatric illness and other complex disorders are performed with samples from ethnically stratified populations. Linkage and case-control association analyses are sensitive to population heterogeneity of disease etiology and marker allele frequencies (Curtis and Sham 1996; Deng 2001). Using an averaged allele-frequency distribution can result in an overestimate or underestimate of the degree of sharing between affected relatives when parental genotypes are not available (Curtis and Sham 1996). Given possible heterogeneity between populations, it may be advisable to perform analyses separately for the families from each population, in addition to performing a joint analysis in which all families are analyzed together.

In genetic studies of psychiatric disease, a common instrument used to determine a subject's psychiatric diagnosis is the Diagnostic Interview for Genetic Studies (DIGS) (Nurnberger et al. 1994). The interview includes a question regarding the ethnicity of the subject's parents (appendix A). Generally, studies that have used the DIGS or a similar instrument and have performed analyses separately by ethnic or racial group have not disclosed how they defined or determined group membership of families, but they likely used imprecise definitions and ad hoc methods to assign them (Cloninger et al. 1998; Funke et al. 2004). Here, we propose an algorithm that uses the information from the DIGS, family structure, affection status, and a simple decision rule to determine group membership for each family for the purposes of genetic analysis.

The effectiveness of using self-reported ethnicity data to generate genetically homogeneous samples is as dependent on the subjects' interpretation of the interviewer's ethnicity question as it is on how genetically heterogeneous the ethnicities are. Sankar and Cho (2002) offer a useful discussion on the need for researchers, when they are performing studies that incorporate ethnicity as an explanatory variable, to understand the factors that influence which ethnic groups a subject identifies with.

To assess the performance of our algorithm, we applied it to data from a recently completed genetic linkage study of schizophrenia (Tsuang et al. 2000; Faraone et al., in press). We compared the estimates of ancestral background obtained by using our algorithm and the DIGS data with those obtained by using the program *structure* and the analysis of data on 378 microsatellite markers (Pritchard et al. 2000; Falush et al. 2003), which estimates population membership with the use of genotype data. When we used a simple rule that classified a family as European American (EA) or African American (AA) when the probability of the family belonging to that population was estimated to be >50%, 158 of the 159 families were classified concordantly by

Received March 31, 2005; accepted for publication June 15, 2005; electronically published July 14, 2005.

Address for correspondence and reprints: Dr. Andrew Skol, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: askol@umich.edu

* VA Cooperative Study 366 Investigators are listed in the acknowledgments.


0002-9297/2005/7703-0003$15.00

our algorithm and by *structure* ($\kappa = 0.99$). In addition, families that were estimated to be admixed by our algorithm were also shown to have substantial admixture by *structure* analysis.

## Methods

The purpose of our algorithm is to divide samples into genetically more homogenous groups of families. This goal was motivated by a desire to perform linkage and family-based association analyses of our Veterans Affairs (VA) schizophrenia data. When parental genotypes are unavailable, evidence for linkage is sensitive to misspecification of parental allele frequencies (Curtis and Sham 1996; Deng 2001), hence the importance of separating families into homogeneous subsets. Populations that historically have been separated by large distances or geographical boundaries are often genetically heterogeneous, and the longer these populations have been isolated and the more varied their environments, the more heterogeneous they are likely to be (Zhivotovsky et al. 2003). The ethnic or racial groups that individuals are operationally assigned to in epidemiological contexts are usually correlated with the continental region of a person's ancestors in the immediate pre-Columbian era, and so the percentages of ancestry tracing to different continental regions are reasonable variables to use to try to construct more-homogenous subsets of families.

Our algorithm summarizes self-reported ethnicity information from family members into a family-specific ancestral-group classification in four steps. First, we convert each pedigree member's response to the DIGS ethnicity question into maternal and paternal ancestry classification vectors. Second, we estimate the ancestry classification vectors of the pedigree's founders from the information of their descendants. Third, we calculate the family ancestry classification vector as a weighted average of the founders' vectors, where a founder's weight is the sum of the kinship coefficients (Wright 1984) between the founder and affected individuals. Fourth, we assign a summary family ancestry based on values in the family's classification vector.

We present our algorithm with the assumption that the ethnicity information is coming from the DIGS, since this is a likely source of diagnostic and ethnicity information for psychiatric genetic studies. It should be straightforward to adjust our algorithm to use information from other sources of self-reported ethnicity.

### *Step 1: Convert DIGS Ethnicity Responses into Vectors of Ancestry Proportions*

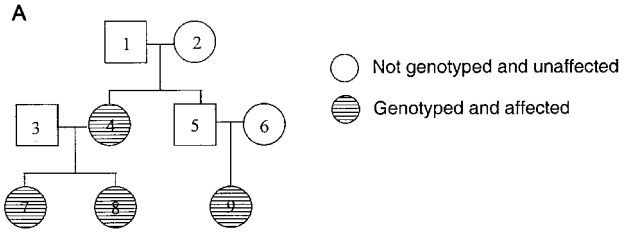
The DIGS question regarding ethnic makeup is given in appendix A. The question asks about the ethnicity of the subject's parents. We store the subject's responses to the question in two vectors, one for each parent. The elements of these vectors contain, in principle, the probability that a randomly chosen gene from the parent traces to a particular continental population. Since our VA sample is almost exclusively of European or African origin, we categorized the ethnicities listed in appendix A into AA (15), EA (1–7), and other (OT) (8–14 and 16). Each respondent is assigned a maternal and paternal ancestry vector, each composed of the proportion of responses corresponding to AA, EA, and OT. Each pedigree member that did not complete the DIGS is assigned a single ancestry vector initialized to [0, 0, 0]. Unlike those who complete the DIGS, these nonrespondents do not require maternal and paternal ancestry vectors; rather, they require only a single vector, since information inferred about their ancestral makeup is an amalgam of their parents' information and cannot be decomposed into parent-specific information.

Figure 1*A* contains a sample pedigree. Figure 1*B* shows the response of the pedigree members to the DIGS ethnicity question and the resulting parental ancestry vectors. For example, individual 8 reported maternal and paternal ethnicities of 15, 15, 1 and 15, 15, 14, respectively, resulting in ancestry vectors [.67, .33, 0] and [.67, 0, .33].

### *Step 2: Infer the Founders' Ancestry Vectors*

In this step, we estimate the ancestry vectors of the founders from the information provided by their descendants. We repeat the following five actions until the founders are reached. First, we identify all individuals with parents but no offspring in the current version of the pedigree. The second, third, fourth, and fifth actions are performed only on these individuals. Second, the ancestry vectors are normalized, if necessary, by dividing each vector element by the sum of the elements; this will not be required for individuals with no offspring in the original pedigree or those who completed the DIGS. Third, each ancestry vector is multiplied by a weight. For our sample, affected individuals, because of possible impaired cognition, communication, and social interaction, are considered to have less reliable responses and are given smaller weights than unaffected individuals. We assign weight 1 to the ancestry vectors of unaffected offspring and weight 0.001 to those of affected offspring. This scheme allows use of the affected offspring responses when no unaffected offspring information is available and virtually ignores this information when responses from unaffected offspring are available. Fourth, for each individual who completed the DIGS, the maternal (or paternal) ancestry vector is added to the mother's (or father's) vector, provided that the mother (or father) did not complete the DIGS. For example, if a mother and her son both completed the DIGS,

A




B

| Study | Response to DIGS | | Inferred Race Vectors | |
|---|---|---|---|---|
| | Maternal | Paternal | Maternal | Paternal |
| 4 | 15, | 15, 14 | [1 0 0] | [.5 0 .5] |
| 7 | 15, 15, 1, 14 | 15, 15, 1, 14 | [.50 .25 .25] | [.50 .25 .25] |
| 8 | 15, 15, 1 | 15, 15, 14 | [.67 .33 0] | [.67 0 .33] |
| 9 | 15 | 15 | [1 0 0] | [1 0 0] |
| 1 | | | [.75 0 .25] | |
| 2 | | | [1 0 0] | |
| 3 | | | [.58 .13 .29] | |
| 5 | | | [1 0 0] | |
| 6 | | | [1 0 0] | |

C

Kinship coefficients

| Pedigree member | 1 | 2 | 3 | 6 |
|---|---|---|---|---|
| 4 | .250 | .250 | 0 | 0 |
| 5 | .250 | .250 | 0 | 0 |
| 7 | .125 | .125 | .250 | 0 |
| 8 | .125 | .125 | .250 | 0 |
| 9 | .125 | .125 | 0 | .250 |

**Figure 1** Example of the ethnicity algorithm applied to a family. *A*, Pedigree diagram. *B*, Responses to DIGS and the corresponding ancestry vectors for pedigree members. Individuals 1, 2, 3, and 6 are founders. The three elements of the ancestry vectors are the AA, EA, and OT proportions, in that order. *C*, Kinship coefficients between founders and affected family members.

then the information the son reports about his mother's ethnicity is ignored, and only the information the mother provided is used. For individuals who do not complete the DIGS, their ancestry vector is added to their mother's (or father's) vector, provided that she (or he) did not complete the DIGS. Note that a respondent, who has both maternal and paternal ancestry vectors, provides parent-specific ethnicity information, whereas a nonrespondent, who has only a single race vector, provides, at best, an average of his or her parents' ethnicity information, for both parents. Fifth, the individuals processed from the current pedigree are removed, and we return to the first action. When the founders are reached, their ancestry vectors are normalized. If a founder completed the DIGS, the maternal and paternal ancestry vectors are averaged.

We demonstrate this procedure for the pedigree in figure 1*A*. We begin with individuals 7, 8, and 9, whose self-reported maternal and paternal ancestry vectors are given in figure 1*B*. None of these individuals require the normalization step. We multiply the maternal and paternal ancestry vectors of individuals 7, 8, and 9 by 0.001, since they are affected. We add the weighted maternal and paternal ancestry vectors of individual 9 to the ancestry vectors of her mother and father, individuals 5 and 6. Both parents' ancestry vectors become [.001, 0, 0]. The information for individuals 7 and 8 is added to the ancestry vector of their father (individual 3) but not to that of their mother (individual 4), because she completed the DIGS. Summing the paternal ancestry vectors of individuals 7 and 8 gives us individual 3's updated vector, $0.001\,[.5, .25, .25] + 0.001[.67, 0, .33] = 0.001\,[1.17, .25, .58]$. We remove individuals 7, 8, and 9 from the pedigree structure and find the individuals with no offspring in the reduced pedigree—individuals 4 and 5. Individual 4's vector does not need to be normalized; individual 5's normalized vector is [1, 0, 0]. We multiply individual 4's ancestry vectors by 0.001 and add the vectors of individuals 4 and 5 to those of their parents' (individuals 1 and 2). Individual 1's ancestry vector becomes $0.001[.5, 0, .5] + [1, 0, 0] = [1.0005, 0, .0005]$; individual 2's becomes $0.001[1, 0, 0] + [1, 0, 0] = [1.001, 0, 0]$. Removing individuals 4 and 5 from the pedigree structure leaves only founders. Founders 1, 2, 3, and 6's normalized vectors are [1, 0, 0], [1, 0, 0], [.58, .13, .29], and [1, 0, 0], respectively.

In the example, affected individuals' information was downweighted. Had we chosen to weight each individual's information equally, the information provided by individual 4 would have been weighted the same as that from her brother, individual 5. This would have given classification vectors [.75, 0, .25] and [1, 0, 0] for founders 1 and 2, respectively, rather than [1, 0, 0] for each.

### *Step 3: Calculate the Family's Ancestry Classification Vector*

We calculate the family's classification vector as a weighted average of the founders' ancestry vectors. For weights, we sum the kinship coefficients between the founder and all affected family members. The kinship coefficient between relatives $x$ and $y$ ($\phi_{xy}$) is the probability that an allele selected at random from an arbitrary locus in $x$ and an allele selected at random from the same locus in $y$ are identical by descent. For a family with $F$ founders and $A$ affected pedigree members, we write the family's race classification vector as

$$\mathbf{R} = \frac{\sum_{i \in F} \phi_i \mathbf{r}_i}{\sum_{i \in F} \phi_i},$$

where $\phi_i = \sum_{j \in A} \phi_{ij}$ and $\mathbf{r}_i$ is the classification vector for

**Table 1**

**Responses to DIGS Ethnicity Question (with Counts >10)**

| Reported Ethnicities | Count |
|---|---|
| 15, –, –, – | 584 |
| 3, –, –, – | 128 |
| 1, –, –, – | 102 |
| 14, 15, –, – | 62 |
| 1, 3, –, – | 62 |
| 15, 15, 15, 15 | 51 |
| 3, 3, –, – | 35 |
| 1, 1, –, – | 30 |
| 4, 4, 4, 4 | 19 |
| 14, –, –, – | 14 |
| 3, 14, –, – | 13 |
| 1, 15, –, – | 12 |
| 1, 4, –, – | 11 |

founder *i*. The values for $\phi_{ij}$ and $\mathbf{r}_i$ can be found in figure 1*B* and 1*C*. For our example, the family ancestry classification vector is [.82, .03, .15]. Generally, analyses will be most sensitive to misspecification of the allele frequencies of the affected members' parents. For this reason, we chose the above weights so that emphasis would be placed on the founders that contribute the most genetically to the affected individuals.

### *Step 4: Assign Family Ancestry*

We assign a family to an ancestry if the family's vector element for that ancestry exceeds a cutoff value. For our VA study, we assigned an ancestry of AA (or EA) if the AA (or EA) element of the family ancestry vector was >0.50. The ancestry vector of the family in figure 1 has an AA element of 0.82, leading us to assign the family an AA classification.

### *Application*

We applied our algorithm to the data of 159 families from our VA Cooperative Studies schizophrenia-study sample. The families had an average of 2.7 affected individuals (range 2–6) and 4.6 genotyped individuals (range 2–14) per family. The markers used were primarily from the ABI Prism Linkage Mapping set, version 2.5, HD5 (Applied Biosystems; see Web Resources).

### *Comparison with* structure

The program *structure* (Pritchard et al. 2000; Falush et al. 2003) implements a Bayesian approach to estimate the proportion of each individual's genome derived from each of *K* populations, where *K* is user specified. Also, *structure* estimates the joint distribution of the marker allele frequencies of the *K* populations, the admixture of each individual, and the population origin of each allele by use of Markov chain–Monte Carlo simulation. Details of the algorithm can be found in Falush et al. (2003). We ran *structure* on our autosomal genome scan data of 378 microsatellite markers, using $K = 2$ and $K = 3$. We ran the analyses once using all 732 genotyped individuals from the 159 families and again using only the 187 genotyped founders. The use of only founders is consistent with *structure*'s assumption that sample individuals are unrelated. The use of all individuals takes advantage of all available information and so is better able to overcome *structure*'s prior distribution that individuals are an equal admixture of the *K* populations, but it overestimates the actual information present.

When we used all genotyped individuals in the *structure* analysis, we estimated family race vectors by averaging the estimated proportion of population membership for the affected members. When we used only founders, we estimated family ancestry vectors using the estimates from *structure* and our algorithm as described above.

## Results

Table 1 displays the frequency with which the most common ethnicity responses were given. Figure 2 shows the ancestry vectors for the 159 schizophrenia-affected VA families after the DIGS responses were submitted to our algorithm. The majority of families cluster at the EA and AA vertices—51 on each. When the families classified as EA were admixed, it was always with OT (specifically, Native American/Native Alaskan), whereas families clas-

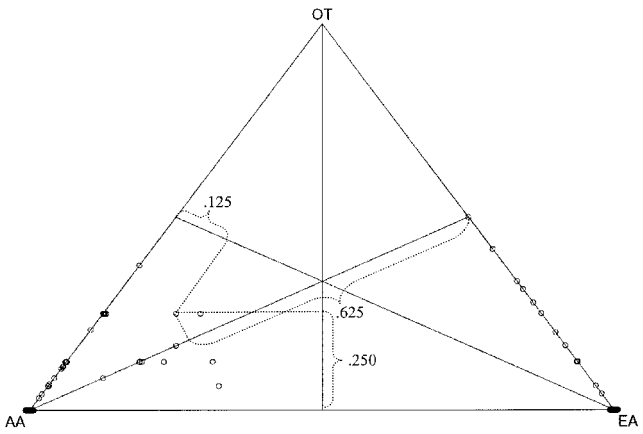



**Figure 2** Distribution of family ancestry vectors estimated from self-reported data. Each of the 159 points represents a family. The proportion AA (or EA or OT) of a family ancestry classification vector can be found by drawing a perpendicular line from the point to the AA (or EA or OT) projection line (i.e., the line emanating from the AA [or EA or OT] vertex). The proportion of the line intersected equals the proportion AA (or EA or OT) estimated. This is illustrated for one of the points.

sified as AA reported admixture of EA, OT, or a combination of EA and OT. There was more variation in the ancestry vectors of families that were predominantly AA than in those that were predominantly EA (fig. 2 and table 2), where we define a family as predominantly EA (or AA) if the EA (or AA) element of the family vector is >0.50 and <1. Predominantly EA families averaged 20% OT admixture. Predominantly AA families averaged nearly equal proportions of EA and OT of ~18% each; for these families, the admixture tends to be either entirely OT or a mixture of OT and EA.

Figure 3 contains bar plots showing *structure*'s estimates of the proportion of the two populations from which an individual's genome is derived, when $K = 2$ populations were assumed. The bars for individuals from AA families, as classified via self-reported information, are on the left. The impressive concordance between a family's ancestry assignment and the proportion admixture from each continental group strongly suggests that the groups identified by *structure* represent European and African populations. When *structure* was run with $K = 3$, no evidence for an additional population was seen. This suggests either that the ethnicities labeled OT are genetically homogeneous to the African or European population or that too little OT genetic information is available for *structure* to distinguish the OT group from the African and European populations. It is possible that, if *structure* could allow the allele frequencies of two populations to be correlated and the third uncorrelated, we would be able to detect heterogeneity between the OT samples and the others, or it may be that too little information is available to distinguish genetic differences between the OT group and the EA and AA groups. Figure 3*A* shows the results for founders only; figure 3*B,* for all individuals. Table 3 reports the average admixture proportions estimated by *structure* when all individuals or only founders are used. It is clear from both figures 3 and 4 and table 3 that the AA sample is much more admixed than the EA sample is—EA and AA families have, on average, 1% and 18% admixture, respectively. The proportion of EA genes in AA individuals is highly variable. In addition, we found that the proportion of admixture estimated is greater when only founders are used than when all available genotyped subjects are used, which is consistent with more data being required to overcome *structure*'s prior that individuals are an equal admixture of each population.

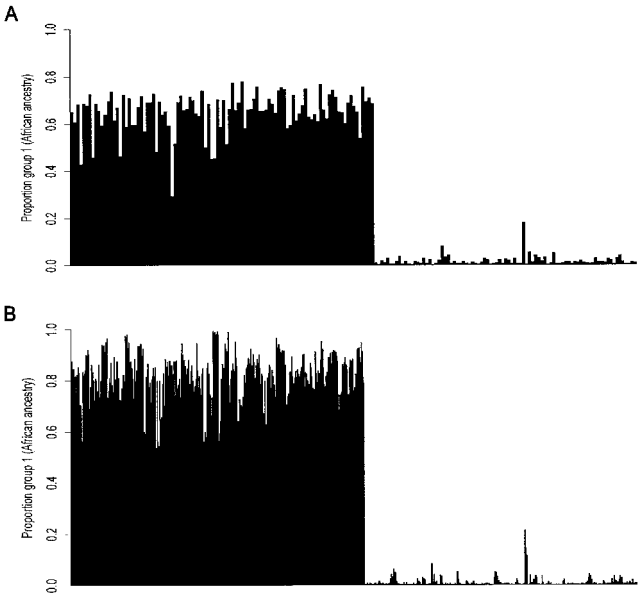



**Figure 3** Population membership estimates for individuals derived by *structure* with the assumption of two populations. Each vertical bar represents the proportion of the individual's genome estimated to be from group 1 (African ancestry). The white area represents the proportion from group 2 (European ancestry). *A,* Estimates with the use of founders only. *B,* Estimates with the use of all family members (and with familial relationships ignored).

We next compared the family race vector estimates from our algorithm with the use of the self-reported data to those estimated by *structure* with the use of genotype information (table 4 and fig. 4). Families reporting more admixture tend to be more genetically admixed according to *structure*. This concordance between self-reported and genetic evidence is shown in more detail in figure 4, which plots the EA proportion estimated using our algorithm and self-reported information versus that estimated using the genotype information and *structure*. The Spearman rank correlation between the EA proportion estimated using *structure* and that estimated using our algorithm is 0.85. When we measured the correlation within only those families that are predominantly EA according to *structure* (EA > 0.5), we found a surprisingly small correlation of just 0.06, whereas correlation within families that are predominantly AA, according to *structure,* is 0.85.

**Table 2**

**Family Race Vector Estimates from Self-Reported Information for Families with EA or AA Proportion >50%**

| | Mean (± SD) Race Vector Estimate (%) | | |
|---|---|---|---|
| Sample | AA | EA | OT |
| EA > 50% | 0 ± 0 | 96 ± 9 | 4 ± 9 |
| 50% < EA < 100% | 0 ± 0 | 80 ± 11 | 20 ± 11 |
| AA > 50% | 92 ± 15 | 2 ± 8 | 6 ± 9 |
| 50% < AA < 100% | 63 ± 22 | 19 ± 12 | 18 ± 12 |

**Table 3**

**Family Race Vector Estimates from *structure* for Families with EA or AA Proportion >50%**

| *structure* Sample | Mean (± SD) Race Vector Estimate (%) | |
|---|---|---|
| | AA | EA |
| EA > 50%: | | |
| All members | 1 ± 2 | 99 ± 2 |
| Founders only | 4 ± 11 | 96 ± 11 |
| AA > 50%: | | |
| All members | 81 ± 8 | 19 ± 8 |
| Founders only | 66 ± 6 | 34 ± 6 |

## Discussion

We presented a simple algorithm that transforms self-reported population affiliation, pedigree structure, affection status, and perceived validity of the self-reported information into a family ancestry classification. Our algorithm should prove useful to researchers who wish to perform linkage analyses or other family-based analyses when population heterogeneity is a potential factor that could lead to excess false positives or loss of power. Although our algorithm was designed for ethnicity data reported from the DIGS, in which subjects report their parents' ethnicities, it can easily be adapted to other information formats, such as those in which individuals report their own ethnicities. If such data are collected, our algorithm could be implemented by assuming that the individual responded that both parents' ethnicities are the same as his or her own.

If a large amount of genotype information is already available from the families, our algorithm can be used in concert with *structure* as we demonstrated above. If no or little genotype information is available and a subset of families with ethnicity information is available to choose from (e.g., National Institute of Mental Health [NIMH] Schizophrenia Genetics Initiative; see Web Resources), our algorithm could be applied to select a collection of families that have a desired distribution of ethnic backgrounds.

We found almost perfect concordance between population assignments determined with the use of our algorithm and those determined with *structure*; 158 of 159 families were classified concordantly ($\kappa = 0.99$). We also discovered that, for predominantly AA families, the amount of admixture predicted from the self-reported information was consistent with that found by *structure* (Spearman correlation coefficient = 0.85). In addition, *structure*'s estimate that AA individuals have, on average, 18% EA admixture is relatively consistent with the findings of Parra et al. (1998) and Shriver et al. (2003), who found that an AA sample from the Washington, D.C., area had genetic contribution estimates of 79%, 18%, and 3% from AA, EA, and Native American populations, respectively. In contrast, for predominantly EA families, we found little correlation ($r = 0.06$). Figure 4 shows that there is little variation in the proportion of EA predicted using *structure* and substantially more variation in that predicted from the self-reported information. To understand why the variability is so much greater in the self-reported EA proportion, recall that the admixture predicted from the self-reported data is entirely due to ethnicities categorized as OT. In our predominantly EA families, the ethnicity leading to an OT classification was Native American/Native Alaskan. At least two explanations exist for why the amounts of Native American/Native Alaskan ancestry estimated via self-reported versus genetic data differ. One possibility is that, even if all individuals accurately reported the amount of Native American ancestry, little genotype data exists in our sample to allow *structure* to distinguish a Native American population from the others; recall that running *structure* with $K = 3$ populations did not show evidence of an additional ancestral group. A second possibility is that individuals overreported the amount of their Native American ancestry. Regardless of the reason, because there is less heterogeneity between Native Americans and European Americans than there is between Native Americans and African Americans (Risch et al. 2002), Native American ancestry is more likely to be identified by *structure* as European, resulting in an underestimation of the heterogeneity in individuals with Native American/Native Alaskan ancestry.

Our algorithm likely performed as well as it did in part because the major groups into which we classified ethnicities—AA and EA—are quite genetically disparate. We also expect our algorithm might work well with Hispanic American data. Distinguishing more subtle ethnic differences may prove challenging and will

**Table 4**

**Comparison of EA and AA Proportions Estimated by *structure* versus Self-Reported Ethnicity Data**

| Self-Reported Estimate | *structure* Estimate (%) | |
|---|---|---|
| | AA | EA |
| EA = 100% | 1 | 99 |
| 50% < EA < 100% | 2 | 98 |
| 50% < AA < 100% | 67 | 33 |
| AA = 100% | 82 | 18 |

likely depend on how recently the groups of interest became isolated or admixed. For example, it is possible that individuals from some Hispanic population can accurately report the amount of Native American and European admixture of their ancestors (Bonilla et al. 2004). In this instance, our algorithm should perform well. However, it is less clear how our algorithm would perform if we were trying to discern the subtle regional ethnic subdivisions of the Icelandic population (Helgason et al. 2005). Subjects from this population are less likely to be aware of these more subtle divisions and so would likely provide less precise responses to ethnicity questions that try to capture the ethnic differences that exist between regions. Additionally, there is likely to be much less genetic variation between subjects from the different regions, leading to groups that, even if respondents report accurately, would be difficult to distinguish, even with methods such as those employed by *structure* (Rosenberg et al. 2002).

There are a number of ways in which our algorithm could be extended or modified. We could weight offspring information on the basis of the number of generations since the information was initially reported, where information from more recent generations would be weighted more heavily than information originating from more distant generations. This would penalize information coming from distant relatives—a reasonable action, given that half the information about an individual is lost in each generation.

Recall that, in our algorithm, information about a father (or mother) provided by his (or her) offspring would be ignored if the father (or mother) provided information about his (or her) own parents. One could argue that the offspring's information should be used. This could be accommodated by adding weighted race vectors of the offspring to the parent's race vector. However, we feel that ignoring the offspring's information is logical, since, in general, self-reported information has better reliability than "secondhand" information. Further, as noted above, parents report information about both of their parents, whereas offspring report about the aggregate of their parents. One scenario in which this approach would be prudent is when an affected parent and at least one of his or her unaffected children complete the DIGS. This would then allow the information from the unaffected child to supersede, or at least contribute, to the possibly less reliable information from the parent.

We assigned families to AA and EA populations on the basis of the simple rule that the AA or EA element of the family ancestry vector must exceed 0.5. If researchers wish to be more discriminating, they could increase the threshold for being assigned to an ancestral group to a value >0.5. The decision rule could also employ a measure of the variance of the family ancestry vector using a cross-validation–type method.

We chose to calculate a weighted average of the founders' classification vectors to determine a family-specific ancestral group. An alternative to identifying an ancestry for each family is to determine the most likely pop-

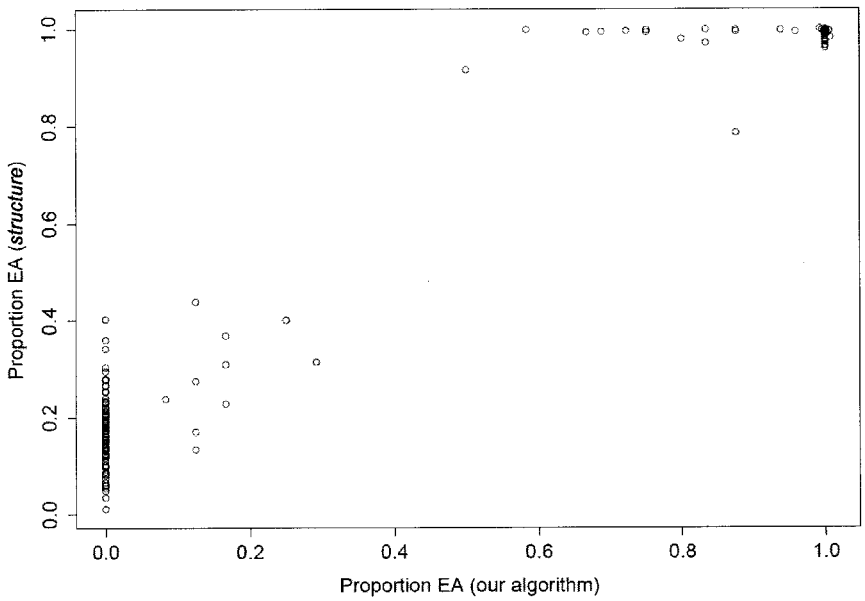



**Figure 4** Comparison of family ancestry estimates calculated from genotype versus self-reported information. The vertical axis represents the proportion of a family's EA ancestry estimated by *structure*. The horizontal axis represents the proportion of a family's EA ancestry estimated by our algorithm.

ulation to which each founder belongs. This would allow us to use population allele-frequency estimates that are appropriate for each founder and, in principle, would allow an even more accurate calculation of the pedigree likelihood. However, no widely used linkage analysis package allows the use of different allele-frequency estimates for a pedigree's founders.

Our algorithm could also be implemented to estimate family-averaged phenotypes. This could be useful if an investigator wishes to separate families on the basis of some phenotype before performing another type of analysis. Ordered subset analysis (OSA) (Hauser et al. 2004) is one such analysis. OSA sorts families by a suspected confounding variable, and a dichotomy in the evidence for linkage between those with high values of the variable and those with low values is tested.

In summary, we have developed a simple, practical, and flexible algorithm that allows researchers to assign family-specific ancestries when only self-reported ethnicity information is available, such as when the DIGS is used. In addition to including the relationships among family members, the algorithm can incorporate the confidence the investigator has in the accuracy of the collected information. Further, this algorithm is not restricted to ethnicity data—it can also be applied to any quantitative or categorical traits.

## Acknowledgments

Veterans Affairs Cooperative Study 366 investigators are C. Baldwin, S. Bingham, J. Collins, S. V. Faraone, S. L. Haverstock, T. Keith, F. Mena, A. S. Menon, S. Prabhudesai, F. Sautter, G. D. Schellenberg, D. W. Tsuang, M. T. Tsuang, D. Weiss, and K. A. Young.

We thank Jodi vanden Eng, for her valuable discussions and input at the early stages of this project, and Jeff Long, for his careful reading of our manuscript and valuable comments, which resulted in a more thoughtful handling of ethnicity and race issues. We gratefully acknowledge support from National Institutes of Health grant HG00376 (to M.B.), the Department of Veterans Affairs Cooperative Studies Program, and a Veterans Affairs Merit Review.

## Appendix A

### DIGS Ethnicity Question

Question: "What is the ethnic background of your biological parents?"

Interviewer instruction: "Code up to four ethnicities on maternal and paternal sides, if possible."

Note that interviewers were directed to ask the question as written and to provide explanations and prompting if necessary. For example, if the respondent replies "white" or "black," then the interviewer might say, "Let's start with your mother; what is her ethnic background? What is the ethnic background of her parents? Do you know the ethnic background of their parents?" These questions would then be repeated for the paternal side. Multiple entries of the same ethnicity value within the mother's (or father's) response column implies that both of the mother's (or father's) parents identified themselves, at least partially, with this ethnicity.

Response was recorded with the following ethnicities:

01 = Anglo-Saxon
02 = Northern European (e.g., Norwegian)
03 = West European (e.g., French, German)
04 = East European, Slavic
05 = Russian
06 = Mediterranean
07 = Ashkenazi Jew
08 = Sephardic Jew
09 = Hispanic (not Puerto Rican)
10 = Puerto Rican Hispanic
11 = Mexican Hispanic
12 = Asian
13 = Arab
14 = Native American/Alaskan Native
15 = African American
16 = Other, Specify:_____UU = Unknown

| MOTHER | FATHER |
|---|---|
| a)_____ | e)_____ |
| b)_____ | f)_____ |
| c)_____ | g)_____ |
| d)_____ | h)_____ |

## Web Resources

The URLs for data presented herein are as follows:

Applied Biosystems, http://home.appliedbiosystems.com/ (for ABI Prism Linkage Mapping set, version 2.5, HD5)
NIMH Schizophrenia Genetics Initiative, http://zork.wustl.edu/nimh/sz.html

## References

Bonilla C, Parra EJ, Pfaff CL, Dios S, Marshall JA, Hamman RF, Ferrell RE, Hoggart CL, McKeigue PM, Shriver MD (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

Cloninger RC, Kaufmann CA, Faraone SV, Malaspina D, Svrakic DM, Harkavy-Friedman J, Suarez BK, Matise TC, Shore D, Lee H, Hampe CL, Wynne D, Drain C, Markel PD, Zambuto CT, Schmitt K, Tsuang MT (1998) Genome-wide search for schizophrenia susceptibility loci: the NIMH genetics initiative and millennium consortium. Am J Med Genet 81:275–281

Curtis D, Sham PC (1996) Population stratifications can cause

false positive linkage results if founders are untyped. Ann Hum Genet 60:261–263

Deng HW (2001) Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 159:1319–1323

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Faraone SV, Skol AD, Tsuang DW, Young KA, Haverstock SL, Prabhudesai S, Mena F, Menon AS, Leong L, Sautter F, Baldwin C, Bingham S, Weiss D, Collins J, Keith T, vanden Eng JL, Boehnke M, Tsuang MT, Schellenberg GD. Genome scan of schizophrenia families in a large veterans affairs cooperative study sample: evidence for linkage to 18p11.32 and for racial heterogeneity on chromosomes 6 and 14. Am J Hum Genet (in press)

Funke B, Finn CT, Plocik AM, Lake S, DeRosse P, Kane JM, Kucherlapati R, Malhotra AK (2004) Association of the *DTNBP1* locus with schizophrenia in a U.S. population. Am J Hum Genet 75:891–898

Hauser ER, Watanabe RM, Duren WL, Bass MP, Langefeld CD, Boehnke M (2004) Ordered subset analysis in genetic linkage mapping of complex traits. Genet Epidemiol 27:53–63

Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37: 90–95

Nurnberger JI, Blehar MC, Kaufmann CA, Yorkcooler C, Simpson SG, Harkavy-Friedman J, Severe JB, Malaspina D, Reich T, Miller M, Bowman ES, Depaulo JR, Cloninger CR, Robinson G, Modlin S, Gershon ES, Maxwell E, Guroff JJ, Kirch D, Wynne D, Berg K, Tsuang MT, Faraone SV, Pepple JR, Ritz AL (1994) Diagnostic interview for genetic studies—rationale, unique features, and training. Arch Gen Psychiatry 51:849–859

Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD (1998) Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet 63: 1839–1851

Pritchard JK, Stephens M, Donnelly PJ (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Risch N, Burchard E, Ziv E, Hua T (2002) Categorization of humans in biological research: genes, race and disease. Genome Biol 3:2007.1–2007.12

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Sankar P, Cho MK (2002) Toward a new vocabulary of human genetic variation. Science 298:1337–1338

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C, Pfaff C, Jones C, Massac A, Cameron N, Baron A, Jackson T, Argyropoulos G, Jin L, Hoggart CJ, McKeigue PM, Kittles RA (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399

Tsuang MT, Faraone SV, Bingham S, Young K, Prabhudesai S, Haverstock SL, Mena F, Menon AS, Pepple J, Johnson J, Baldwin C, Weiss D, Collins J (2000) Department of Veterans Affairs Cooperative Studies Program genetic linkage study of schizophrenia: ascertainment methods and sample description. Am J Med Genet 96:342–347

Wright S (1984) Evolution and the genetics of populations, volume 2: theory of gene frequencies. University of Chicago Press, Chicago

Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186