

## FP-TDI SNP Scoring by Manual and Statistical Procedures: A Study of Error Rates and Types

*BioTechniques* 34:610-624 (March 2003)

**E.J.C.G. van den Oord,  
Y. Jiang, B.P. Riley,  
K.S. Kendler, and X. Chen**  
Virginia Commonwealth  
University, Richmond, VA, USA

### ABSTRACT

*For technologies that are commonly used in ordinary laboratories such as fluorescence-polarization detection with template-directed, dye-terminator incorporation (FP-TDI), SNP genotype scoring is usually done manually. Here we study rates of errors and missing genotypes obtained with this procedure. We also introduce three statistical genotype scoring methods to examine whether they form a viable alternative. Data consisted of eight SNPs typed in about 1400 individuals from 268 pedigrees. The statistical procedures performed better on several internal criteria, such as the number of Mendelian errors, and showed much higher agreement with discrepant genotypes re-scored by two raters. The best results were obtained with the statistical procedure that incorporated information about regularities in the error structure of the FP-TDI data. We estimated that there were about 1.6% more errors if genotypes were scored manually. About 0.6% of these errors could be explained by data manipulation errors, leaving 1% as the result of possible incorrect scoring. There were 3.3% more missing genotypes in the manual scoring due to errors in data manipulation (1.7%) and conservative scoring (1.6%).*

### INTRODUCTION

SNPs are increasingly used in genetic studies. However, the large-scale genotyping of SNPs brings along the problem of genotyping errors. Several authors have discussed the detection of such errors (2–7), studied the effects of errors on linkage analyses (8–19), or proposed statistical tests that can take genotyping errors into account (20,21). Less attention has been paid to studying the causes and rates of genotyping errors and devising procedures to reduce them.

A few studies have examined genotyping error rates (22–25). Estimates varied from 0.08% to 3%. However, these studies pertained to multi-allelic short tandem repeat (STR) or microsatellite markers. Differences between these markers and SNPs exist not only with respect to the genotyping procedures but also in the way the data are processed and the genotypes are scored. For example, in this study we focus on the SNP genotyping technology introduced by Chen et al. (1) labeled fluorescence-polarization detection with template-directed, dye-terminator incorporation (FP-TDI). FP-TDI is commonly used and has proven to be efficient and accurate. When performing FP-TDI genotyping, to our knowledge there are no widely used computer packages, so these are often performed by laboratory technicians.

Scoring genotypes manually introduces possible human errors. The first aim of this article is to study error rates and types as may occur when SNP genotypes are scored manually. The second aim is to examine whether statistical genotype scoring methods form

a viable alternative. Statistical procedures have several potential advantages. First, several steps of data manipulation are required to transform the raw SNP data to a genotype in a database that can be used for subsequent analyses. Automating these steps avoids errors that can result from handling the data manually. Second, ratings by technicians can be subject to variations in scoring rules, confusion, or fatigue. Statistical genotype scoring is not affected by these sources of unreliability. Third, additional information such as genotypes of relatives can be included when scoring the genotypes by computer algorithms. This allows the use of more information in the decision process that could potentially improve the scoring. However, scoring genotypes automatically requires a statistical model. This is the potential disadvantage of statistical procedures. If the model is not an accurate representation of the data, then erroneous decisions about genotypes will be obtained.

In this study, we focus on eight SNPs typed in about 1400 individuals. The genotyping was done as part of a previous larger fine mapping study. Therefore, the technicians were unaware of the aim of the present study, making our results more representative of datasets obtained from the daily routine in a medium-sized human molecular genetics laboratory. We compared the genotypes scored manually with three statistical procedures. The core of all three statistical approaches was a mixture model. Based on the results obtained from fitting mixture models to the FP-TDI data, we classified data points as ambiguous, outliers, failures,

minor allele homozygotes, heterozygotes, and common homozygotes. We also tried to improve the classification by including genotype information of family members and making use of regularities we discovered in the error structure of the model. Results were evaluated in terms of the data quality as measured by several internal criteria such as the occurrence of Mendelian errors. Discrepancies between genotypes scored manually and the statistical procedures were examined in detail. If the source of the discrepancy could not be identified, then the genotypes were re-scored by two raters and agreement measures were computed to validate the scoring by statistical procedures and technicians.

## MATERIALS AND METHODS

### Sample and Genotyping

The sample consisted of 268 multiplex families selected for high density of schizophrenia (26). The number of individuals in the pedigrees was 2368. Most pedigrees consisted of two or three generations. The average number of children per nuclear families was 3.2. DNA was available for 1405 indi-

viduals. The SNPs were genotyped using the technology discussed by Chen et al. (1). The genotyping was performed as part of a larger fine mapping study so that the technicians were unaware of the aim of this study. We focused on eight SNPs that were approximately evenly spaced in a 30-kb region. This implied a total of 117 plates of 96 wells and, after excluding control samples, a total of 11 177 samples. Seventeen of the 117 plates were redone because their quality was considered too poor for genotype scoring.

### Manual Genotype Scoring

The TDI reactions were read using an LJI fluorescence plate reader (Analyst HT; Molecular Devices, Sunnyvale, CA, USA). For each dye used, the reader generates an FP value for each sample. The two FP values can be plotted. The plot normally forms four distinct groups representing failures, minor allele homozygotes, heterozygotes, and common homozygotes. Figure 1 shows an example. Technicians visually inspected the plots to assign the individual points to one of the four groups. Outliers or points that are hard to score are included with the failures. If the segregation is very poor, then the plate

will typically be redone. Although most plots will show a better segregation of groups than the one in Figure 1, these kinds of plots are scored and demonstrate some of the challenges that technicians may face. The failures are always in the lower left corner, the homozygotes in the upper left and lower right corner, and the heterozygotes in the upper right corner. In this example, the groups of heterozygotes and minor allele homozygotes (lower right corner) seem to be fairly well separated. More problematic are points such as 26 and 74 that seem to be somewhere in between the groups of failures and common homozygotes. Technicians will typically not score points when they are segregated from the bulk of the group. Although this guideline would be helpful for point 80, it is less clear how point 96 should be scored. Furthermore, for small groups such as the minor allele homozygotes that may consist of only a few points, it may not always be easy to determine the bulk of the group.

### Statistical Genotype Scoring

The Appendix gives details of the statistical genotype scoring. The first step is to prepare the data so that they are in the proper format. Next, the number of groups need to be determined by visual inspection of the data, starting values have to be provided for the analysis, and a choice has to be made concerning the type of analysis to be used for the scoring. The analysis consists of fitting mixture models to the two FP values for each plate of 96 wells. An advantage of this approach is its flexibility, allowing the inclusion of covariates, all sorts of variables that may affect the classification, and constraints across groups to improve the statistical behavior of the model. The maximum number of groups in this mixture model was four. This is because, for bi-allelic SNPs, the maximum number of genotype groups is three: minor allele homozygotes, heterozygotes, and common homozygotes. In addition, a fourth group can be present, including those cases where the FP reaction failed. Note that not all genotype groups may be present on each plate, and in some instances there

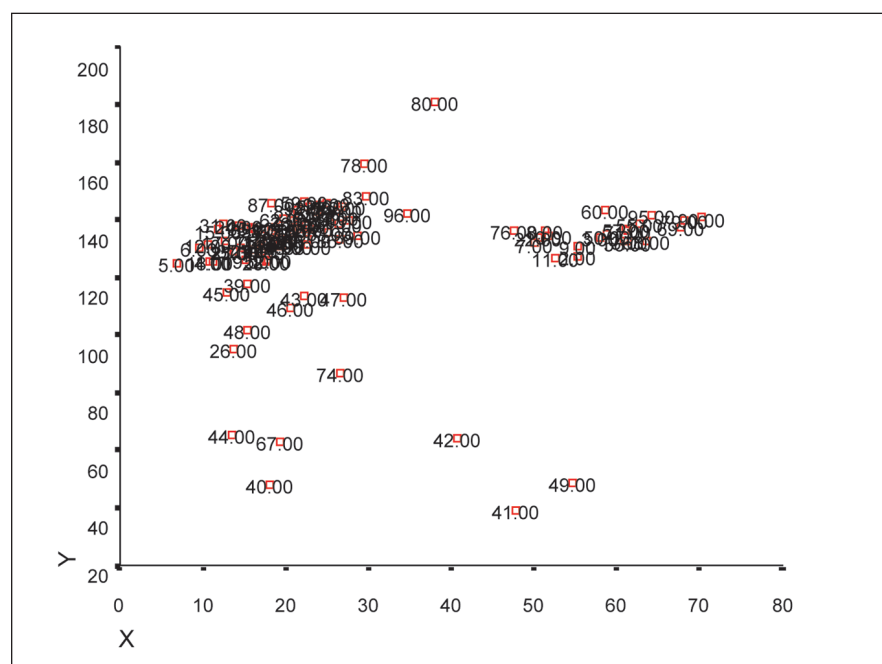


Figure 1. 2-D plot with FP-TDI results.

# Research Report

may be no failures so that four is a maximum. Starting values consist of rough estimates of the mean of each group in the 2-D FP plot. We studied a basic model, a model that included genotypes of relatives, and a model with error deviations as covariates.

**Basic model.** The basic model (Figure 2A) uses the two FP values denoted  $X$  and  $Y$  as input. The groups are indicated with the variable  $c_k$ . Because the groups are not directly observed, they may be better viewed as latent classes. The arrows labeled  $x_k$  and  $y_k$  represent the means of  $X$  and  $Y$  in each group. These parameters are subscripted  $k$  because they can be different for each group. We also need to estimate a variance and a covariance for  $X$  and  $Y$ . We tried different models by fixing the covariance between the FP values to be zero versus estimating this covariance and estimating different covariance matrices in each group versus forcing these matrices to be equal. Effect sizes and fit indices suggested that it was necessary to estimate the covariance between the two FP values. The often very small group sizes of failures and minor allele homozygotes resulted in unstable estimates of the variance-covariance matrices, and we therefore constrained the variance-covariance

matrices to be equal across groups. Therefore, the parameters are not subscripted  $k$  in Figure 2A. The estimates were obtained by maximum likelihood. This assumes that within each group of the mixture model the FP values are multivariate normal distributed.

Based on the results of the analysis (posterior) probabilities can be computed that a case  $i$  belongs to each of the four groups. A case was assigned to a group only if the probability of belonging to that group was higher than 99%, or else it was considered ambiguous. This threshold of 99% was chosen because it seemed to maximize the agreement with the manual scoring and produced results that had face validity after we inspected a large number of plots. When a case is much closer to one group in the 2-D FP plot than the three other groups, it will have a high probability. However, even if it is relatively close to that group, it may have different FP values compared to other members of that group. To determine whether a case was an outlier, we used the Mahalanobis distance that measures the distance of a data point to the “center” of the group. An outlier was defined as a data point with probability lower than 0.001 of belonging to the group to which it most likely belonged.

This threshold of 0.001 was again determined on the basis of agreement with the manual scoring and face validity.

In the first cycle of the analysis, group membership was assumed to be unknown for all cases. In our mixture model approach, variables can be used to indicate the group(s) to which a case can belong. There are as many indicator variables as there are groups. The  $k^{\text{th}}$  indicator variable is defined as  $I_{ik} = 1$  if case  $i$  can belong to group  $k$  and is zero otherwise. In Figure 2A, this indicator variable is allowed to have an effect on  $c_k$ , thus the way subjects are assigned to one of the classes of the mixture. Indicator variables were fixed to one for cases that were assigned to a group. The model was then rerun to see whether this helped to classify the cases that were ambiguous. This cycle of rerunning analyses was repeated until there were no more changes in group assignments.

**Including genotypes of family members.** Indicator variables were also used to include information about genotypes of family members. This may help to score outliers or ambiguous cases. For biallelic SNPs, there are three possible genotypes. If three individuals would be unrelated, then there are  $3 \times 3 \times 3 = 27$  possible combinations of their genotypes. However, in triads of two

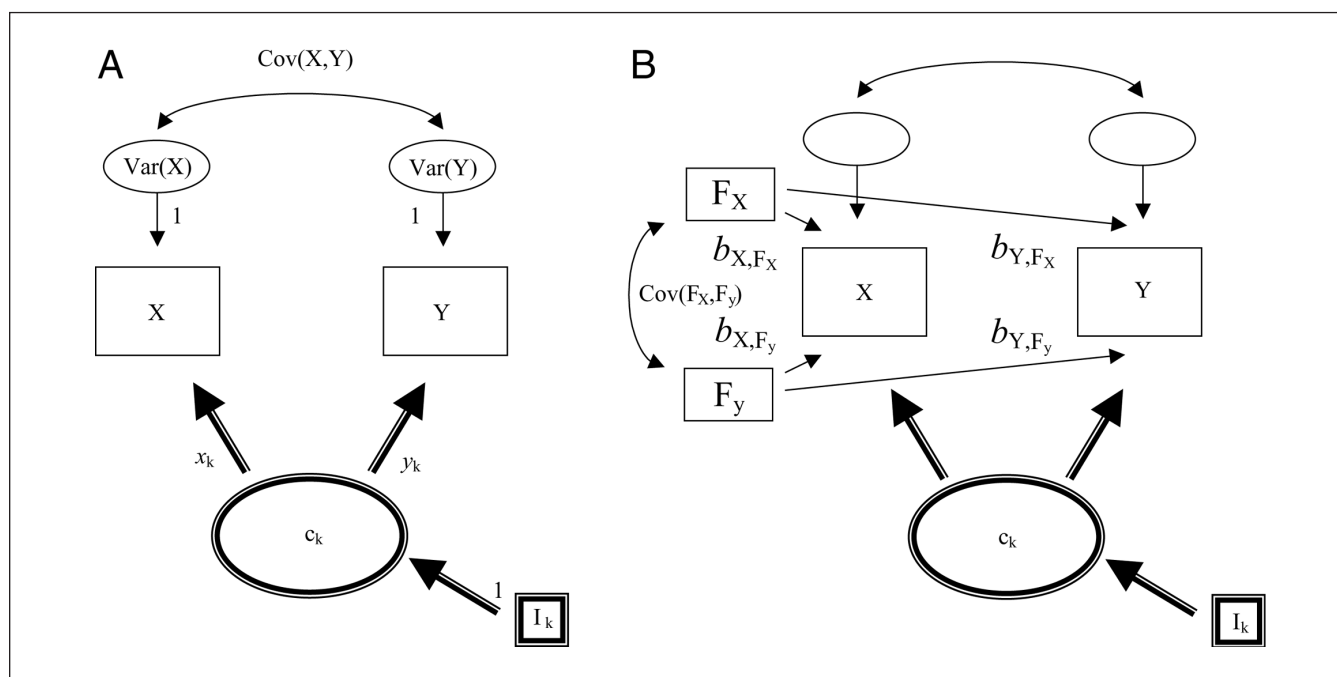


Figure 2. Path diagrams for the mixture models. (A) Basic model/family members. (B) Model with covariates.

**Table 1. Factor and Multiple Regression Analyses on Residual Scores**

	Factor Analysis		Multiple Regression	
	X	Y	r	r
SNP 1	0.484	0.450	0.575	0.558
SNP 2	0.731	0.531	0.755	0.615
SNP 3	0.656	0.603	0.716	0.688
SNP 4	0.480	0.510	0.573	0.600
SNP 5	0.577	0.486	0.661	0.580
SNP 6	0.338	0.181	0.429	0.252
SNP 7	0.599	0.489	0.677	0.586
SNP 8	0.301	0.329	0.383	0.433

parents and a child, the laws of inheritance limit the number of combinations to 15 (27). For example, if a parent and a child are both homozygous for the minor allele, then the genotype of the other parent has to have at least one copy of the minor allele. These constraints may help to classify ambiguous cases that would now be based on both the FP data and the genotypes of family members. This genotype information can be incorporated in the model by giving the case a value of zero for the indicator variable that pertains to the group of common homozygotes and values of one for the indicator variables that correspond to the other groups.

To reduce the possible genotypes for ambiguous cases/outliers, we started with the cases who were parents by using possible genotypes of the spouse in combination with each of the children. Next, possible genotypes of the parents of the index cases were used. The search was stopped if only one genotype was left or all possible triads had been considered. Genotype information could either be an assigned genotype in one or both other members of the triad or multiple genotypes when one triad member was an ambiguous case or outlier itself where one of its genotypes was eliminated in a previous step. Thus, information from the whole pedigree was used. This genotype reduction was only tried in families that passed the Mendelian check. If a Mendelian error was found, then the indicator variables for all members from that family were set to one so that the genotype scoring was completely redone. Information from family members cannot be used to exclude the possibility that a case is a

failure. Therefore, the indicator variable corresponding to the group of failures was always set to one. Note that genotypes are not only assigned on the basis of the genotypes of family members but also on the FP data. Thus, if the genotypes of family members would suggest certain genotypes that are incompatible with the FP data because of genotyping errors in the relatives or the misspecification of family relations, then that case would remain ambiguous.

**Error deviations as a covariate.** To reduce the “error” term and obtain a more accurate classification, covariates can be added to the model. We were not able to identify variables that explained between group variation. However, relative to the group mean, individual data points appeared to have similar positions or deviations in the 2-D FP-plot across SNPs. For each plate  $p$ , we fitted the basic mixture model to estimate the group means  $x_{kp}$  and  $y_{kp}$  and obtain a pooled estimate of the standard deviation:  $SD(X)_{kp} = SD(X)_p$  and  $SD(Y)_{kp} = SD(Y)_p$ . Next, for each case  $i$  that was assigned to group  $k$  on plate  $p$ , deviation scores were computed and divided by the pooled standard deviation to standardize across plates:  $X' = (x_{ikp} - x_{kp})/SD(X)_p$  and  $Y' = (y_{ikp} - y_{kp})/SD(Y)_p$ . Correlations between standardized deviation scores across different SNPs were on average 0.266 for  $X'$  and 0.192 for  $Y'$ . The deviations from the eight SNPs were also submitted to an unweighted least squares factor analysis. Table 1 displays the results. Factors were extracted using the traditional “eigenvalue larger than 1” criterion. For both  $X'$  and  $Y'$ , only one factor was extracted. This suggested that a single common factor

# Research Report

can explain the correlations between the deviations across SNPs. A possible candidate is DNA concentration that might affect the polarization for different SNPs in the same way. This is suggested by the literature where fluorescence polarization is used to measure the concentration of substrate—in our case, the concentration of PCR products and the fact that many different systems have been developed using FP principles to quantify the amount of substrates or products (28–31).

To obtain overall measures to be included in the mixture model analyses to score genotypes, we computed a factor score for the eight deviation scores  $X'$  as well as a factor score for the eight deviations  $Y'$  from the y-axis. The factor scores were regressed on the deviations of the individual SNPs to give an impression of their explanatory power. Results are shown in the second part of Table 1. The multiple correlations varied between 0.252 and 0.755 and were on average 0.568, suggesting that over 30% of the within-group variance could be explained by the factor scores. Although the correlation between the two factor scores was 0.757, in most cases, both factor scores made significant contributions so that both were allowed to affect both FP values in the mixture model. Figure 2B displays the model where these factor scores were included. In addition to the parameters shown in Figure 2A (that were left out for sake

of simplicity), we now also estimated the effects  $b$  of the factor scores on X and Y. These effects were constrained to be equal across the groups and are therefore not subscripted  $k$  in Figure 2B.

## Programs

For the statistical genotype scoring, we wrote a Pascal program that controlled the cycles of fitting mixture models, read the output, scored the genotypes, and wrote the genotypes plus group and outlier probabilities to an output file. The mixture models were fit by calling the program Mplus (32). Our program plus documentation can be downloaded from the software library at [www.BioTechniques.com](http://www.BioTechniques.com) and from <http://www.vipbg.vcu.edu/~edwin>.

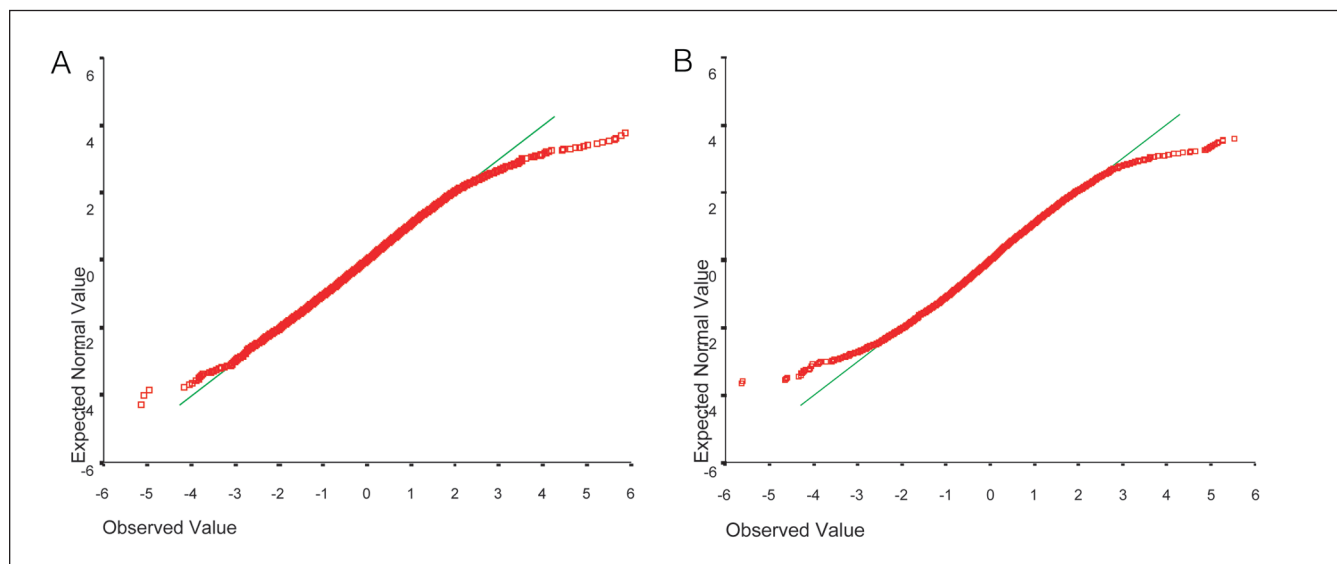
## RESULTS

To check the assumption of normality made by our mixture model, in Figure 3 we plotted the quantiles of the error terms computed as described in the section “Error deviations as a covariate” for the basic model against the quantiles of the normal distribution. If the normality assumption is correct, then the points should cluster around a straight line. Figure 3 shows that this match is pretty good and supports the assumption of normal distributions. Some deviation is observed but only

for z scores greater than three or less than -3 that represent merely 0.2% of the most extreme scores.

## Scoring by Manual and Statistical Procedures

Table 2 shows the frequency distributions of the genotypes obtained with the three statistical procedures and the technicians. In addition, several indices are reported that reflect the quality of the scored genotypes. The first index was the percentage of scored genotypes. Higher percentages are better because they imply fewer missing values for subsequent analyses. We also computed the number of Mendelian errors and double recombinants. Only certain genotyping errors will result in Mendelian inconsistencies (e.g., two common homozygous parents cannot have heterozygous or minor allele homozygous offspring) (4,15). Mendelian errors are also not exclusive indicators for genotyping errors and could, for instance, be the result of incorrectly specified biological relations between family members. Although the absolute number may be difficult to interpret, differences between the number of Mendelian errors between genotypes scored with different approaches can only reflect genotyping errors. The number of double recombinants was estimated via the program Simwalk (33). We first estimated haplotypes in



**Figure 3. Quantile-quantile plots to check the normality assumption.** (A) Deviations from x-axis in FP plot. (B) Deviations from y-axis in FP plot.

# Research Report

**Table 2. Results for Different Genotype Scoring Methods**

	Mixture Model Approach			Manual
	Basic	Family Members	Covariates	
Outlier	87	87	113	
Ambiguous	127	119	76	
Failure	212	212	213	783*
Minor allele homozygote	258	258	258	235
Heterozygote	2449	2456	2461	2356
Common homozygote	8036	8037	8048	7803
Total % scored	96.19%	96.26%	96.40%	93.0%
No. Mendelian errors	5	5	5	9
No. double recombinants	13	12	9	13
No. common haplotypes	7	7	7	9
% with common haplotype	93.64%	93.65%	93.74%	91.40%

Note that \* means these failures include outliers and ambiguous points.

our pedigrees. A haplotype is a piece of DNA that contains genetic variations linked so closely that it is nearly always inherited as a unit. Two nearby microsatellites were included in these analyses to avoid the possibility that a substantial proportion of the haplotypes would be uncertain. Haplotype ambiguity is a big problem with SNPs because they have two alleles only (34). Microsatellites have many alleles so that it is much easier to establish from which parent chromosome the haplotype was derived. Although it tries to minimize them, the haplotype estimation allowed for recombination. Because the markers were at a very short distance from each other, the presence of two or more recombinations is very likely to reflect genotyping errors. The absolute number of double recombinations may again be difficult to interpret because of factors such as genotyping errors in the microsatellites or the incorrect estimation of haplotypes by Simwalk. However, differences in the number of double recombinants between methods are likely to be the result of errors in genotype scoring. Recombinations are more likely at specific locations. Therefore, individuals from different families may still share part of the haplotype from very distant ancestors. Indeed, recent studies confirm a structure consisting of haplotype blocks in which a large percentage of a sample can typically be character-

ized by few common haplotypes that cover relatively long regions of the genome (35,36). Our final indices assessed the consistency of the haplotypes with this model. We reported the number of common haplotypes, considering the SNPs only, defined as haplotypes with a prevalence larger than 1% in the founders. In addition, the percentage of founder haplotypes that belonged to one of the common haplotypes was calculated. The idea is that genotyping errors and missing genotypes result in a more diffuse haplotype structure so that genotype scoring is likely to be better when there are a few common haplotypes that account for a relatively large percentage of the founder haplotypes in the sample.

The percentage of genotypes that could be scored with the basic mixture model was 96.19%. These rates range from 91.7 to 98.5, indicating that that some SNPs are easier to score than others. Five Mendelian errors and 13 double recombinants were found in the 11 177 genotypes. Seven common haplotypes accounted for 93.64% of the founder haplotypes in the sample. Only minor differences were found with the two other automated procedures. These differences involved the distribution of the various categories, the “Total % scored”, the number of common haplotypes, and “% with common haplotype”. The differences are unlikely to be merely chance findings. For in-

stance, the distribution of outliers and ambiguous points was different for the “family member” and “covariate” model (Chi-square = 12.2,  $df=1$ ,  $P < 0.001$ ). Furthermore, the pattern of differences seems consistent and sensible. For instance, eight previously ambiguous cases could be assigned genotypes when the model included the genotypes of relatives. This seems to make sense because you add further information that is likely to be correct and therefore helps the classification. The fact that the use of genotypes from family members makes little difference may be because, in most cases, the use of genotype data from other family members excludes merely one homozygous genotype. Therefore, in principle, it would be most helpful to classify ambiguous points that could be one of the two homozygotes. However, in the FP plot, the two homozygous genotypes are always in opposite corners so that it is unlikely to have data points that are in the middle of these two groups. The use of the overall within-group residual terms as covariates yielded the most favorable results in terms of the total number of genotypes scored. Three observations suggested that these changes represented a small improvement. First, the number of double recombinants was reduced. Second, the same seven haplotypes accounted for a slightly larger percentage of the founder haplotypes in the sample. Third, the eight ambiguous cases that could be assigned genotypes by including information about family members were assigned the same genotypes when the covariates were included in the model.

The manual genotype scoring consistently yielded less favorable results compared to the statistical scoring. It resulted in greater than 3%–3.4% more missing genotypes. The number of Mendelian errors and double recombinants equaled those found by the basic mixture model but were poorer than the mixture model approach with covariates. Finally, two more common haplotypes were found, and this larger number of haplotypes accounted for a smaller percentage of founder haplotypes in the sample. This suggested that genotyping errors produced a less clear and consistent haplotype structure.

A case was assigned to a group if the

# Research Report

probability of belonging to that group was higher than 99%; otherwise, it was considered ambiguous. An outlier was defined as a data point with probability less than 0.001 of belonging to the group to which it most likely belonged. We studied how altering these criteria affected the scoring. The vast majority of cases had probabilities greater than 0.99 of belonging to the groups to which they were assigned. Depending on which method was used, lowering this threshold to 0.95 would increase the number of scored genotypes with 41–59 cases. Lowering the threshold even more would increase the number of scored genotypes with 35–68 observations, which is a very modest increase for what seems to be a substantial increase in uncertainty. Decreasing the outlier criterion from 0.001 to 0.0005 would only enable us to score 11–15 more genotypes. Being more conservative by increasing the outlier criterion from 0.001 to 0.005 would have a relatively larger impact and result in 59–64 more outliers. In general, however, changing these parameters had only fairly modest effects, and the initial chosen values of 0.99 and 0.001 seem to offer a compromise between the risk of misclassification on the one hand and loss of data on the other.

## Statistical Procedures versus Manual Scoring

In 94%–95% of the cases, the technicians and the three statistical procedures scored classified the samples in the same group (failures, two homozygotes, and heterozygotes). This corresponded to a Kappa (37), a measure between zero and one for agreement between raters that corrects for agreement due to coincidence, of about 0.88. When we excluded the failures from both methods, the average agreement increased to 99.1% and the Kappa was 0.978. Thus, about 95% of the genotypes were identical and therefore likely to be correct. The disagreements in the remaining 5% most often involved data points that were not scored by one procedure but were assigned a genotype by the other.

We inspected all 748 data points that were classified differently by statistical procedures versus manual scoring. In 200 cases, the differences could be iden-

**Table 3. Agreement between Different Methods for Scoring Samples that Were Scored Differently by Statistical and Manual Procedures**

	Rater 1	Rater 2	Basic	Family Members	Covariates	Manual
<b>Raters</b>						
1	—	0.991	0.986	0.986	0.993	0.536
2	0.616	—	1.00	1.00	1.00	0.247
<b>Mixture model</b>						
Basic	0.545	0.619	—	1.00	1.00	n/a
Family members	0.539	0.613	0.977	—	1.00	n/a
Covariates	0.506	0.505	0.747	0.763	—	n/a
<b>Manual</b>	0.003	-0.086	n/a	n/a	n/a	—

Note: above diagonal is Kappa when failures are excluded from both methods, and below diagonal is Kappa with failures included. n/a is not applicable because the genotypes were selected to be discordant. Cases for which the cause of the discrepancy could be identified were not included.

tified and were the result of data handling errors. That is, for two plates of 96 samples, the wrong FP data were used for the genotype scoring, and, in one situation, the wrong FP data were used for part of the plate. In addition, two plates of scored genotypes were mistakenly excluded from the final data file, which explains about 190 missing genotypes.

After excluding these identified differences, the remaining 548 discrepant genotypes were independently scored by two raters (Y.J. and X.C.). The raters disagreed about 132 points. Table 3 shows Kappa statistics. The Kappas above the diagonal show that there was perfect agreement among the statistical procedures and between the statistical procedures and rater 2. That is, there were no instances where different genotypes were assigned to the same individual. Agreement between the statistical procedures and rater 2 versus rater 1 was very close to one. The Kappas below the diagonal show a much lower agreement. This is because disagreement nearly always involved unclear data points—that is, data points that were scored as outliers or ambiguous points by one procedure and assigned a genotype by another. The statistical procedures tended to show higher agreement than the two raters. This was particularly true for the basic and family member method because the only difference in scoring between these meth-

ods was that eight previously ambiguous cases could be assigned genotypes when the model included the genotypes of relatives. The agreement between raters and statistical procedures decreased when additional information about family members and covariates was included. This may not necessarily point to poorer results. These models use information that is not available to the raters so that the lower disagreement could reflect the different inputs used to score the genotypes.

The agreement between the original manual genotype scoring and the two raters was clearly lower than between raters and statistical procedures. This is true for the Kappas above and below the diagonal. This finding is more remarkable when one considers the fact that technicians and raters are more likely to use similar rules to score genotypes. Therefore, such method effects may be expected to favor the manual scoring. The conclusion of these findings is that the discrepancies between statistical procedures and manual scoring were more likely to reflect errors in the manual scoring.

## A Combined Procedure

We also studied whether genotype scoring could be improved by having technicians/raters score difficult data points. For this purpose, we created a

dataset where all discrepant points (i.e., those scored differently by statistical procedures and technicians) were assigned the genotype that was most frequent in the scoring by the two raters and statistical procedure that included covariates. In comparison to the statistical scoring alone, this combined procedure, although it increased the number of missing genotypes by 49, reduced the number of Mendelian errors and double recombinants by one. Thus, a slight reduction of errors seems possible by having technicians score difficult data points.

It is unclear whether this improvement was the result of systematic differences between raters and statistical procedure or the general phenomenon that raters tended to take fewer risks at the expense of having more missing genotypes. To study this, we selected cases where statistical procedures and raters differed. The statistical procedure scored more aggressively: in 92% (57/62) of the cases, the statistical procedure scored a genotype, whereas the raters classified these cases as failures. However, the differences concentrated on certain plates. These plates generally showed relatively poor separation of the groups. In these situations, the raters were apparently more cautious and less likely to assign genotypes. The raters were also more inclined to score data points as outliers when they were disconnected from the rest of group. A variation on this theme involved the group of minor allele homozygotes that typically consisted of very few data points. As a result, there was more often a point that looked disconnected. Our model that assumes equal variances in all groups would score such points as homozygous, whereas the raters were more inclined to score them as outliers. This phenomenon may also account for the finding in Table 2 that the frequency of the minor allele tended to be somewhat lower in the manual scoring (13.6%) than with the statistical procedure (13.8%).

## DISCUSSION

We studied error rates and types when SNP genotypes are scored manually and explored the possibility of automating SNP genotype scoring. For

0.6% of the total number of scored genotypes, the wrong FP data were used in the manual procedure as a result of data-handling errors. These errors are not inherent in the data. However, assuming that our laboratory is representative of other laboratories, the errors are inherent to manual genotype scoring. For another 1%, differences were found between manual genotype scoring and statistical procedures, but the source of the errors could not be identified. However, there were two indications that the genotypes scored manually were more likely to be wrong. First, the statistical procedures performed better with respect to several internal criteria such as the number of Mendelian errors. Second, when discrepant genotypes were re-scored by two raters, the genotypes scored by the statistical procedure showed much higher agreement with the raters, even though possible method effects should have favored the manual genotype scoring. Assuming that the statistical procedure was correct in the 1% where technicians scored the genotypes differently and errors could not be identified, an overall estimate of the error rate equals  $1 + 0.6 = 1.6\%$ . Another aspect of data quality involved the finding that there were 3.3% more missing genotypes when scoring was done manually. This percentage can be broken up into 1.7% caused by errors in data manipulation where data got lost so that these genotypes were not present in the final dataset. This percentage is again not inherent in the data but, to the extent our laboratory is representative, is inherent to genotype scoring where data are handled manually. The other 1.6% was due to conservative scoring by technicians who were more likely to classify less clear data points as failures.

Our model made several assumptions such as multivariate normality, equal group variances, and that there are no external processes that may cause more than four groups. Even if, strictly speaking, some of these assumptions may not hold, results suggested that the scoring method is robust to violations of these assumptions. The manual genotype scoring can be viewed as a “non-parametric” method. After eliminating the data manipulation

# Research Report

---

errors where the wrong FP was used by technicians, greater than 95% of the samples were scored identically. If the automated procedures were sensitive to violations of the assumptions, then we would not have observed such a large agreement. The disagreements in the remaining percentage involved “difficult” data points that were not scored by one procedure but were assigned a genotype by the other. As judged by a wide variety of criteria, our analyses suggested that the automated procedures scored these difficult points better than the technicians. Thus, although, strictly speaking, some model assumptions may not hold, even for those difficult points, the statistical procedures yielded better results than those obtained via manual scoring.

Another kind of question is whether the automated procedure could be further improved by making other assumptions. A first remark is that the room for

improvement seems very limited. One indication that there were only 62 cases where statistical procedures and the two raters (who re-scored the “difficult points”) differed consistently. Furthermore, it seems likely that part of this systematic difference has to do with a reluctance of raters to score difficult points rather than a systematic error caused by the normality assumption of the automated procedure. Another indication is that only five Mendelian errors were left after the genotypes were scored by the statistical procedures. It should be noted that this number is likely to be an underestimate of the true number of errors (3,4,15). Douglas et al. (15) found that, particularly with SNPs, genotype errors are difficult to detect via Mendelian inconsistencies. Assuming a random-allele-error model, they suggested that Mendelian checks identify 13%–75% of the genotype errors. However, even these percentages

would still imply merely 0.06%–0.34% remaining errors that could be eliminated. In addition, 3.6% of the genotypes were failures or could not be scored. This percentage gives the upper bound for a further possible reduction of the number of missing genotypes.

Even if one assumes that a systematic error is introduced by some of the model assumptions, it is an open question whether the best way to solve this problem would be to change the assumptions. Estimating more complex models and more parameters may have a negative impact on the statistical behavior of the models and the precision of the parameters estimates. Therefore, the challenge would be to improve the classification of these possible systematic errors and maintain the same accuracy in classifying the other points at the same time. We also found that the possible systematic differences concentrated on plates that showed relatively poor separation of the groups and situations where points looked disconnected from the rest of group. In addition to changing model assumptions, another option could be to find and incorporate the factors that account for these possible anomalies into the model. A more rigorous option involves duplication or replication with a different SNP genotyping technology of the problematic points. However, because of the increase in cost/effort and limited room for improvement, it may be difficult to justify this approach. Furthermore, because of the regularities in the error structure, difficult data points will remain difficult to score using the same genotyping method. Assuming that it eliminates this phenomenon, replication with a different technology may therefore be better. There was some evidence that genotyping errors may be slightly reduced by having technicians score plates that showed relatively poor separation of the groups. The explanation may be that in these situations the more conservative manual scoring is justified by the lower data quality. Therefore, a final option is to let difficult data points score independently by raters.

It was encouraging to find that SNP genotype scoring could be statistical. In addition to fewer errors and missing

# Research Report

genotypes, statistical procedures have the obvious advantage of being much less time consuming and expensive because they save manpower. The inclusion of family members to aid the genotype scoring had little effect. Merely eight additional cases could be assigned genotypes when the genotypes of family members were included. The model that included covariates reflecting regularities in the error structure of FP scores across SNPs scored these eight points in the same way. Because the latter model is much easier to implement, it may not be worth the trouble of fitting the model with family members, and the choice may either be the basic model or the model with covariates. The covariate model produced the best results with respect to the internal criteria. That is, it reduced the number of double recombinants and increased the number of scored genotypes. However, the differences with the basic model were small. It seems to be a matter of individual preference whether the slightly better results are judged to be worth performing the more elaborate procedure. In this study, we used a fairly large sample and included eight different SNPs. Furthermore, results were fairly robust against changes in the probabilities that determined ambiguous cases and outliers. Although the program we used has that flexibility, under ordinary circumstances we would therefore not think that dataset-specific tuning would be required in other datasets.

## APPENDIX

### Mixture Models

Let  $n_y$  be the number of variables (two for the basic model and four for the model that includes the covariates),  $y_i$  be the  $n_y$ -dimensional random vector with the FP data plus possible covariates for case  $i$  ( $i = 1..N$ ), and  $t_i$  the 4-D vector of indicator variables where  $t_{ik} = 1$  if  $i$  can belong to group  $k$  and zero otherwise. The density function can then be written as a mixture of four multivariate normals:

$$f(y_i, t_i; p, \Sigma, \mu) = \sum_{k=1}^4 (t_{ik} p_k) g_k(y_i; \Sigma_k, \mu_k)$$

where  $\Sigma = [\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4]$ ,  $\mu = [\mu_1, \mu_2, \mu_3, \mu_4]$  and  $p = [p_1, p_2, p_3, p_4]$ . The  $n_y \times n_y$  matrices  $\Sigma_k$  comprise the variances and covariances between the two data vectors, and the vector  $\mu_k$  the means in the groups. The mixing proportions in  $p_k$  determine the relative contribution in terms of the observations to each group within the mixture. They are subject to the constraints:  $p_k \geq 0$  and  $\sum p_k = 1$ . Within each group,  $g_k$  is defined by:

$$g_k(y_i; \Sigma_k, \mu_k) = 2\pi^{-1/2 n_y} |\Sigma_k|^{-1/2} \exp[-1/2 (y_i - \mu_k)^t \Sigma_k^{-1} (y_i - \mu_k)],$$

where superscript  $t$  indicates transposition, and  $|\Sigma_k|$  and  $\Sigma_k^{-1}$  denotes the determinant and inverse of  $\Sigma_k$ . Let  $\theta$  be a vector comprising the parameters used to model the covariances as  $\Sigma_k(\theta)$  and means as  $\mu_k(\theta)$ . Maximum likelihood estimates of  $p$  and  $\theta$  given the observed continuous data and indicator variables are then obtained by first summing the individual likelihoods over the groups:

$$L_i = \sum_{k=1}^4 (t_{ik} p_k) g_k(y_i; \Sigma_k(\theta), \mu_k(\theta))$$

Next, the logarithm of the individual likelihoods is taken, and the sum of all individual log-likelihoods is maximized:

$$\ln L(\theta; p; y_i, t_i) = \sum_{i=1}^N \ln(L_i)$$

Given the ML parameter estimates, it is possible to assign the cases by calculating the posterior probability that a given case belongs to a given group. By Bayes' theorem, the posterior probability that case  $i$  belongs to component  $k$  is:

$$\text{prob}(k | y_i) = p_k g_k(y_i; \Sigma_k\{\theta_k\}, \mu_k\{\theta_k\}) / f(y_i; p, \Sigma\{\theta\}, \mu\{\theta\}).$$

For outlier detection, we used the Mahalanobis Distance using the ML estimates:

$$d_i^2 = (y_i - \mu_k)^t \Sigma_k^{-1} (y_i - \mu_k).$$

The distance  $d_i^2$  can be expressed as:

$$d_i^2 = z_i^t z_i = \sum_{j=1}^2 z_{ij}^2$$

and  $z_i = A^t (y_i - \mu_k)$  and  $A$  is the whitening transformation (38). Since the mean vector of  $z$  are  $[0, 0]^t$  and covariance matrix of  $z$  the identity matrix, the  $z_{ij}$ s

are independent random variables with zero mean and unity variance. Thus, if  $y$  is normal, then the Mahalanobis distance is a chi-square random variable with two degrees of freedom. Therefore, the probability that observation  $y_i$  belongs to group  $k$  is  $\chi^2(k > d_i^2 | k, \theta)$ ,  $df. = 2$ .

## REFERENCES

- Chen, X., L. Levine, and P.Y. Kwok. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res.* 9:492-498.
- Miller, C.R., P. Joyce, and L.P. Waits. 2002. Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics* 160:357-366.
- Gordon, D., S.M. Leal, S.C. Heath, and J. Ott. 2000. An analytic solution to single nucleotide polymorphism error-detection rates in nuclear families: implications for study design. *Pac. Symp. Biocomput.* 17:663-674.
- Gordon, D., S.C. Heath, and J. Ott. 1999. True pedigree errors more frequent than apparent errors for single nucleotide polymorphisms. *Hum. Hered.* 49:65-70.
- Ehm, M.G., M. Kimmel, and R.W. Cottingham. 1996. Error detection for genetic data, using likelihood methods. *Am. J. Hum. Genet.* 58:225-234.
- Douglas, J.A., M. Boehnke, and K. Lange. 2000. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *66:1287-1297.*
- Sobel, E., J.C. Papp, and K. Lange. 2002. Detection and integration of genotyping errors in statistical genetics. *Am. J. Hum. Genet.* 70:496-508.
- Goring, H.H. and J.D. Terwilliger. 2000. Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am. J. Hum. Genet.* 66:1107-1118.
- Terwilliger, J.D., D.E. Weeks, and J. Ott. 2002. Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimates of the recombination fraction. *Am. J. Hum. Genet.* 71:A201.
- Shields, D.C., A. Collins, K.H. Buetow, and N.E. Morton. 1991. Error filtration, interference, and the human linkage map. *Proc. Natl. Acad. Sci. USA* 88:6501-6505.
- Ott, J. 1977. Linkage analysis with misclassification at one locus. *Clin. Genet.* 12:119-124.
- Goldstein, D.R., H. Zhao, and T.P. Speed. 1997. The effects of genotyping errors and interference on estimation of genetic distance. *Hum. Hered.* 47:86-100.
- Cherny, S.S., G.R. Abecasis, W.O. Cookson, P.C. Sham, and L.R. Cardon. 2001. The effect of genotype and pedigree error on linkage analysis: analysis of three asthma genome scans. *Genet. Epidemiol.* 21(Suppl):S117-S122.
- Buetow, K.H. 1991. Influence of aberrant observations on high-resolution linkage analysis

# Research Report

- outcomes. *Am. J. Hum. Genet.* 49:985-994.
15. **Douglas, J.A., A.D. Skol, and M. Boehnke.** 2002. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am. J. Hum. Genet.* 70:487-495.
16. **Goring, H.H. and J.D. Terwilliger.** 2000. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am. J. Hum. Genet.* 66:1310-1327.
17. **Abecasis, G.R., S.S. Cherny, and L.R. Cardon.** 2001. The impact of genotyping error on family-based analysis of quantitative traits. *Eur. J. Hum. Genet.* 9:130-134.
18. **Gordon, D., T.C. Matisse, S.C. Heath, and J. Ott.** 1999. Power loss for multiallelic transmission/disequilibrium test when errors introduced: GAW11 simulated data. *Genet. Epidemiol.* 17Suppl1:S587-S592.
19. **Akey, J.M., K. Zhang, M. Xiong, P. Doris, and L. Jin.** 2001. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am. J. Hum. Genet.* 68:1447-1456.
20. **Gordon, D., S.C. Heath, X. Liu, and J. Ott.** 2001. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.* 69:371-380.
21. **Gordon, D. and J. Ott.** 2001. Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. *Pac. Symp. Biocomput.* 18:18-29.
22. **Brzustowicz, L.M., C. Merette, X. Xie, L. Townsend, T.C. Gilliam, and J. Ott.** 1993. Molecular and statistical approaches to the detection and correction of errors in genotype databases. *Am. J. Hum. Genet.* 53:1137-1145.
23. **Ewen, K.R., M. Bahlo, S.A. Treloar, D.F. Levinson, B. Mowry, J.W. Barlow, and S.J. Foote.** 2000. Identification and analysis of error types in high-throughput genotyping. *Am. J. Hum. Genet.* 67:727-736.
24. **Ghosh, S., Z.E. Karanjawala, E.R. Hauser, D. Ally, J.I. Knapp, J.B. Rayman, A. Musick, J. Tannenbaum, et al.** 1997. Methods for precise sizing, automated binning of alleles, and reduction of error rates in large-scale genotyping using fluorescently labeled dinucleotide markers. FUSION (Finland-U.S. Investigation of NIDDM Genetics) Study Group. *Genome Res.* 7:165-178.
25. **Weeks, D.E., Y.P. Conley, R.E. Ferrell, T.S. Mah, and M.B. Gorin.** 2002. A tale of two genotypes: consistency between two high-throughput genotyping centers. *Genome Res.* 12:430-435.
26. **Kendler, K.S., M.C. Neale, and D. Walsh.** 1995. Evaluating the spectrum concept of schizophrenia in the Roscommon Family Study. *Am. J. Psychiatry* 152:749-754.
27. **Weinberg, C.R.** 1999. Allowing for missing parents in genetic studies of case-parent triads. *Am. J. Hum. Genet.* 64:1186-1193.
28. **Bicamumpaka, C. and M. Page.** 1998. Development of a fluorescence polarization immunoassay (FPIA) for the quantitative determination of paclitaxel. *J. Immunol. Methods* 212:1-7.
29. **Heyduk, T., Y. Ma, H. Tang, and R.H. Ebright.** 1996. Fluorescence anisotropy: rapid, quantitative assay for protein-DNA and protein-protein interaction. *Methods Enzymol.* 274:492-503.
30. **Sun, S., L.H. Nguyen, R.O. Harold, G.F. Hollis, and R. Wynn.** 2002. Quantitative analysis of c-myc-tagged protein in crude cell extracts using fluorescence polarization. *Anal. Biochem.* 307:287-296.
31. **Ye, B.C., K. Ikebukuro, and I. Karube.** 1998. Quantitative analysis of polymerase chain reaction using anisotropy ratio and relative hydrodynamic volume of fluorescence polarization method. *Nucleic Acids Res.* 26:3614-3615.
32. **Muthén, L.K. and B.O. Muthén.** 1998. Mplus. Muthén and Muthén. Los Angeles, CA.
33. **Sobel, E. and K. Lange.** 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* 58:1323-1337.
34. **Hodge, S.E., M. Boehnke, and M.A. Spence.** 1999. Loss of information due to ambiguous haplotyping of SNPs. *Nat. Genet.* 21:360-361.
35. **Daly, M.J., J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander.** 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29:229-232.
36. **Patil, N., A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, et al.** 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719-1723.
37. **Cohen, J.** 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20:37-46.
38. **Fukunaga, K.** 1990. Introduction to Statistical Pattern Recognition. Academic Press, Boston.

Received 15 July 2002; accepted 16 January 2003.

#### Address correspondence to:

Dr. Edwin van den Oord  
Department of Psychiatry  
Virginia Commonwealth University  
P.O. Box 980126  
Richmond VA 23298-0126, USA  
e-mail: ejvandenoord@vcu.edu