Am. J. Hum. Genet. 68:1457–1462, 2001

Evaluation of Candidate Genes in Case-Control Studies: A Statistical Method to Account for Related Subjects

S. L. Slager and D. J. Schaid

Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Traditional case-control studies provide a powerful and efficient method for evaluation of association between candidate genes and disease. The sampling of cases from multiplex pedigrees, rather than from a catchment area, can increase the likelihood that genetic cases are selected. However, use of all the related cases without accounting for their biological relationship can increase the type I error rate of the statistical test. To overcome this problem, we present an analysis method that is used to compare genotype frequencies between cases and controls, according to a trend in proportions as the dosage of the risk allele increases. This method uses the appropriate variance to account for the correlated family data, thus maintaining the correct type I error rate. The magnitude of the association is estimated by the odds ratio, with the variance of the odds ratio also accounting for the correlated data. Our method makes efficient use of data collected from multiplex families and should prove useful for the analysis of candidate genes among families sampled for linkage studies. An application of our method, to family data from a prostate cancer study, is presented to illustrate the method's utility.

Introduction

There is a strong need for analytic methods that can be used to compare candidate-gene frequencies in biologically related cases versus those in unrelated controls. The main reason for this need is the efficiency gain that is obtained when case-control methods are chosen over other association designs, such as family-based association studies. At times, this gain can be substantial, providing a two- to sixfold increase in efficiency (Risch and Teng 1998; Teng and Risch 1999; Risch 2000). Most of the explanation for this gain comes from an increased allele-frequency difference between related cases and unrelated controls. When the number of affected relatives sampled increases, the expected frequency of the high-risk allele increases but the expected frequency remains constant for the unrelated controls (Risch and Teng 1998), thus resulting in a larger difference in allele frequencies. In contrast, when family-based related controls are used, the frequency of the high-risk allele also increases among these controls, thereby decreasing the expected allele-frequency difference between cases and related controls. The factors that influence this gain in efficiency include the use of unrelated controls versus related controls (i.e., family-based controls) (Khoury and Yang 1998; Morton and Collins 1998; Risch and Teng 1998; Teng and Risch 1999; Risch 2000), the sampling of cases from multiplex families versus cases from simplex families, and the sampling, in multiplex families, of all affected relatives versus only one affected relative (Risch and Teng 1998; Teng and Risch 1999; Risch 2000).

Other reasons for the need to develop methods for case-control studies with sampled relatives have to do with the capability to analyze data that are readily available, either from linkage studies or from other resources. For example, candidate-gene studies often follow up on promising linkage findings. The ability to use the same multiplex-family data for association analyses would be economical; all that will then be required of the researcher is the sampling of unrelated controls, who are less expensive to ascertain than are cases. Furthermore, nationwide registries of families at high risk for particular traits currently exist, providing researchers with access to lists of individuals and/or families that otherwise are not currently available; for example, the National Cancer Institute has developed the Cancer Genetics Network, the Cancer Family Registries for Breast and Colorectal Cancer Studies, and the Chronic Lymphocytic Leukemia Family Registry.

Although there are many advantages to using all the cases from multiplex families, the correlations among relatives must be accounted for in the statistical analysis, to avoid an increase in the type I error rate of the statistical test. Previous work by Teng and Risch (1999) introduced, for family-based study designs, a transmission/disequilibrium-test–like statistic, T_{DS}, that com-

Received February 28, 2001; accepted for publication April 2, 2001; electronically published May 15, 2001.

Address for correspondence and reprints: Dr. Susan Slager, Harwick 6, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905. E-mail: slager@mayo.edu

0002-9297/2001/6806-0017$02.00

pares the allele frequency in the affected child (or children) with that in the parents. In situations in which parental data are absent, unaffected sibs or unrelated controls can provide an estimate of the allele frequency in the parents. The T_{DS} statistic, however, can only be applied to nuclear-family data, which is too restrictive for studies of late-onset diseases such as prostate cancer, for which many of the sampled pedigrees contain cousin data in addition to sibling data.

In the present article we propose statistical methods that account for the sampling of biologically related subjects (e.g., siblings, cousins, aunts, etc.) when candidate-gene association analyses are performed. Our statistical test is based on the Armitage (1955) test for trend in proportions and includes a variance that appropriately accounts for family relationships. The Armitage test for trend has two advantages over other statistical tests of association. First, the test does not require that the genotype frequencies comply with Hardy-Weinberg proportions (HWP). Earlier work by Sasieni (1997) demonstrated that statistical tests based on the comparison of allele frequencies—rather than genotype frequencies—between unrelated cases and controls can have an increased rate of false-positive conclusions when genotype frequencies do not fit HWP. Second, the Armitage test for trend provides a flexible analysis method, because different scores can be used to test the dosage of the high-risk allele, depending on whether there is prior knowledge of the disease. The parameter-estimation part of our method uses the traditional odds ratio as an estimate of the relative risk, but the variance of the odds ratio accounts for the correlated data. At the end of this article, we illustrate the utility of our statistic by applying it to a prostate cancer study.

Methods

Setup

Suppose that S controls and R cases are sampled, for a total of N subjects, where the subjects may or may not be sampled from the same family. For simplicity, each person is genotyped for a candidate gene that comprises two alleles: all high-risk alleles, A, and all other alleles, a. Below, we discuss how to extend the analysis to more than two alleles. The data can be summarized in a 2×3 contingency table, as shown in table 1, according to each person's genotype and disease status. Note that the subscripts in table 1 denote the number of high-risk alleles, A, in a genotype.

Table 1

Notation for Genotype Counts and Conditional Probabilities, for Cases and Controls

	Genotype Count (Conditional Probability) for Genotype			
Sample	*a/a*	*a/A*	*A/A*	Total
Cases	r_0 ($\pi_{0\|1}$)	r_1 ($\pi_{1\|1}$)	r_2 ($\pi_{2\|1}$)	R
Controls	s_0 ($\pi_{0\|2}$)	s_1 ($\pi_{1\|2}$)	s_2 ($\pi_{2\|2}$)	S
Total	n_0	n_1	n_2	N

Statistical Test

When there is no association between genotype and disease, the genotype frequencies will be similar in cases and controls. Although a general test to detect any difference in genotype frequencies can be derived, a trend test that is sensitive to monotonic differences may be the most powerful for genotypic data. We use Armitage's (1955) test for trend, which measures a trend in proportions according to a general measure of genetic dosage, x_i. The values of x_0, x_1, and x_2 are used to weight the counts of genotypes *a/a, A/a,* and *A/A,* respectively. We assume that $x_i = i$—that is, we assume that x_i is the number of high-risk alleles that a person has—although other values of $\mathbf{x}$ can be used, such as (0,1,1), which are powerful for a dominant effect, or (0,0,1), which are powerful for a recessive effect. The χ^2 test statistic has the form $U^2/\mathrm{Var}\,(U)$, where U is a sum of weighted differences, of genotype counts, between cases and controls, $U = \sum_{i=0}^{2} x_i[(S/N)r_i - (R/N)s_i]$, where r_i and s_i are, respectively, the number of cases and controls with i high-risk alleles, as defined in table 1.

Since subjects may be biologically related, we need to account for their correlations when we compute the variance of U, or $\mathrm{Var}\,(U)$. To do so, we represent the genotype counts, r_i and s_i, as sums of indicator values, because it is straightforward to determine the variances and covariances of the elements of summations. Let $\mathbf{y_i} = (y_{i0}, y_{i1}, y_{i2})$ denote a vector of genotype indicators for the ith case, with elements $y_{ij} = 1$ if the ith case has the jth genotype; otherwise, $y_{ij} = 0$. A similar indicator vector for the ith control is denoted as "$\mathbf{z_i}$." The sums of these vectors for cases and controls, respectively, are $\mathbf{r} = \sum_i^R \mathbf{y_i}$ and $\mathbf{s} = \sum_i^S \mathbf{z_i}$, where $\mathbf{r} = (r_0, r_1, r_2)$ and $\mathbf{s} = (s_0, s_1, s_2)$. The numerator of our test statistic, U, can then be written in vector notation, as

$$U = \mathbf{x}'[(1 - \phi)\mathbf{r} - \phi\mathbf{s}] \ , \qquad (1)$$

where $\phi = R/N$, the proportion of subjects that are cases. The general formula for $\mathrm{Var}\,(U)$ is derived by the following sequence of steps:

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}\left\{\mathbf{x}'[(1-\phi)\mathbf{r}-\phi\mathbf{s}]\right\} \\
&= \mathrm{Var}\left[(1-\phi)\mathbf{x}'\sum_i \mathbf{y}_i - \phi\mathbf{x}'\sum_i \mathbf{z}_i\right] \\
&= \mathrm{Var}\left[(1-\phi)\mathbf{x}'\sum_i \mathbf{y}_i\right] + \mathrm{Var}\left(\phi\mathbf{x}'\sum_i \mathbf{z}_i\right) \\
&\quad -2\,\mathrm{Cov}\left[(1-\phi)\mathbf{x}'\sum_i \mathbf{y}_i, \phi\mathbf{x}'\sum_i \mathbf{z}_i\right] \\
&= (1-\phi)^2\mathbf{x}'\mathrm{Var}\left(\sum_i \mathbf{y}_i\right)\mathbf{x} + \phi^2\mathbf{x}'\mathrm{Var}\left(\sum_i \mathbf{z}_i\right)\mathbf{x} \\
&\quad -2\phi(1-\phi)\mathbf{x}'\mathrm{Cov}\left(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i\right)\mathbf{x} \\
&= (1-\phi)^2\mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{y}_i) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{y}_i,\mathbf{y}_j)\right]\mathbf{x} + \\
&\quad \phi^2\mathbf{x}'\left[\sum_i \mathrm{Var}(\mathbf{z}_i) + 2\sum_{i<j}\sum \mathrm{Cov}(\mathbf{z}_i,\mathbf{z}_j)\right]\mathbf{x} \\
&\quad -2\phi(1-\phi)\mathbf{x}'\left[\sum_i\sum_j \mathrm{Cov}(\mathbf{y}_i,\mathbf{z}_j)\right]\mathbf{x} \ . \qquad (2)
\end{aligned}$$

Note that the three covariance terms in the last step account for the biological relationships among the subjects: $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$, for correlation between the ith and jth cases; $\mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j)$, for correlation between the ith and jth controls; and $\mathrm{Cov}(\mathbf{y}_i, \mathbf{z}_j)$, for correlation between the ith case and the jth control.

Under the null hypothesis—that is, when there is no association of genotype with disease—$\mathrm{Var}(\mathbf{y}_i) = \mathrm{Var}(\mathbf{z}_i)$, which is equal to the multinomial covariance matrix with elements σ_{ij}, where $\sigma_{ii} = p_i(1-p_i)$ and $\sigma_{ij} = -p_i p_j$, where p_i denotes the probability for the ith genotype g_i. In addition, the covariance matrix for the indicator vectors of any two subjects—say, $\mathbf{w}_i$ and $\mathbf{w}_j$, where i and j can be cases, controls, or one of each—is

$$\begin{aligned}
\mathrm{Cov}(\mathbf{w}_i,\mathbf{w}_j) &= E(\mathbf{w}_i\mathbf{w}_j') - E(\mathbf{w}_i)E(\mathbf{w}_j)' \\
&= P(g_i,g_j) - \mathbf{p}\mathbf{p}' \ , \qquad (3)
\end{aligned}$$

where $P(g_i, g_j)$ is the matrix of the joint genotype probabilities for subjects i and j and where $\mathbf{p}\mathbf{p}'$ is the cross-product of the marginal genotype probabilities $\mathbf{p}' = (p_0, p_1, p_2)$, for genotypes *a/a*, *A/a*, and *A/A*. The joint genotype-probability matrix is a function of the identical by descent (IBD) probabilities (for details, see the Appendix), which can be estimated by GENEHUNTER (Kruglyak et al. 1996). These IBD-probability calculations are conditional on the marker genotypes of the subjects and typically assume that the distribution of genotypes in the population follows HWP.

We have so far assumed a biallelic marker, although this need not be the case. There are several general ways to extend our statistic to markers with more than two alleles. One approach is to create a vector of trend statistics, where an arbitrary allele is chosen as a baseline and where the "dosage" of each of the other alleles is contrasted between the cases and the controls. To construct such a statistic for K alleles, with the Kth allele as the baseline, we create a matrix $\mathbf{X}$ such that (*a*) the jth column in $\mathbf{X}$ corresponds to the jth allele, ($j = 1, 2, \ldots, K-1$) and (*b*) an element in the jth column is the number of alleles of type j. This matrix $\mathbf{X}$ has $K(K+1)/2$ rows, the total number of possible genotypes. The vector of trend statistics is $U = \mathbf{X}'[(1-\phi)\mathbf{r} - \phi\mathbf{s}]$, where the vectors $\mathbf{r}$ and $\mathbf{s}$ are defined as before but with length equal to the number of distinct genotypes (i.e., $K(K+1)/2$). The $\mathrm{Var}(U)$ is easily derived by replacing the vector $\mathbf{x}$ by the matrix $\mathbf{X}$ in equation (2). It follows, then, that the statistic $U'[\mathrm{Var}(U)]^{-1}U$ has an approximate χ^2 distribution with $K-1$ df. Another way to account for multiple alleles is to emphasize homozygotes, by scoring them as 1 and by scoring heterozygotes as 0, and then, with these scores placed in a vector $\mathbf{x}$, to compute U and $\mathrm{Var}(U)$ according to equations (1) and (2), respectively.

Parameter Estimation

The odds ratio ψ is used as a measure of association between the candidate gene and the disease. It approximates the relative risk of disease for a genotype with i high-risk alleles, relative to a genotype with 0 high-risk alleles. This measure, however, is not a measure of the odds ratio in the general population, since, as shown below, it does not account for the oversampling of cases from multiplex families. However, it is an estimate of the association in families similar to the sampled families. Note, however, that, to measure the effect in the general population, the ascertainment criteria would need to be modeled or corrected. The odds ratio, determined by rows 1 and 2 and columns i and 0, is defined to be $\psi_i = (\pi_{i|1}\pi_{0|2})/(\pi_{0|1}\pi_{i|2})$, where $\pi_{i|j}$ is the conditional probability of the ith genotype, given the jth disease category, as defined in table 1. The estimate of ψ_i, when the cell counts found in table 1 are used, is $\hat{\psi}_i = r_i s_0 / r_0 s_i$.

For convenience, we will work with logarithms to calculate the variance of $\log(\hat{\psi}_i)$. Replacing the r's and s's in $\hat{\psi}_i$ by the sums of indicator vectors $\mathbf{y}_i$ and $\mathbf{z}_i$, respectively, we have

$$\mathrm{Var}[\log(\hat{\psi}_i)] = \mathrm{Var}\left[\log\left(\frac{\mathbf{c}_i'\Sigma_i\mathbf{y}_i}{\mathbf{c}_0'\Sigma_i\mathbf{y}_i}\right) - \log\left(\frac{\mathbf{c}_i'\Sigma_i\mathbf{z}_i}{\mathbf{c}_0'\Sigma_i\mathbf{z}_i}\right)\right],$$

where $\mathbf{c}_i$ and $\mathbf{c}_0$ are indicator vectors indicating genotypes i and 0, respectively. Through use of Taylor se-

ries approximation (Rao 1973), an approximation of $\mathrm{Var}[\log(\hat{\psi}_i)]$ can be shown to be

$$\widehat{\mathrm{Var}}[\log(\hat{\psi}_i)] = \frac{\mathbf{c}_i'\mathrm{Var}(\sum_i \mathbf{y}_i)\mathbf{c}_i}{r_i^2} + \frac{\mathbf{c}_0'\mathrm{Var}(\sum_i \mathbf{y}_i)\mathbf{c}_0}{r_0^2} - 2\frac{\mathbf{c}_i'\mathrm{Var}(\sum_i \mathbf{y}_i)\mathbf{c}_0}{r_i r_0} + \frac{\mathbf{c}_i'\mathrm{Var}(\sum_i \mathbf{z}_i)\mathbf{c}_i}{s_i^2} + \frac{\mathbf{c}_0'\mathrm{Var}(\sum_i \mathbf{z}_i)\mathbf{c}_0}{s_0^2} - 2\frac{\mathbf{c}_i'\mathrm{Var}(\sum_i \mathbf{z}_i)\mathbf{c}_0}{s_i s_0} - 2\frac{\mathbf{c}_i'\mathrm{Cov}(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i)\mathbf{c}_i}{r_i s_i} + 2\frac{\mathbf{c}_i'\mathrm{Cov}(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i)\mathbf{c}_0}{r_i s_0} + 2\frac{\mathbf{c}_0'\mathrm{Cov}(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i)\mathbf{c}_i}{r_0 s_i} - 2\frac{\mathbf{c}_0'\mathrm{Cov}(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i)\mathbf{c}_0}{r_0 s_0} . \quad (4)$$

Similar to what has been seen in the calculations for $\mathrm{Var}(U)$, $\mathrm{Var}(\sum_i \mathbf{y}_i)$ in equation (4) is equivalent to $\sum_i \mathrm{Var}(\mathbf{y}_i) + 2\sum\sum_{i<j}\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$ where $\mathrm{Var}(\mathbf{y}_i)$ is the multinomial covariance matrix and where $\mathrm{Cov}(\mathbf{y}_i, \mathbf{y}_j)$ is the correlation matrix between the ith and jth cases and is calculated as in equation (3). The calculation of $\mathrm{Var}(\sum_i \mathbf{z}_i)$ among the controls and of $\mathrm{Cov}(\sum_i \mathbf{y}_i, \sum_i \mathbf{z}_i)$ among the cases and controls follows analogously. Note, however, that the variance and covariance terms in equation (4) are calculated under the alternative hypothesis, implying that the genotype frequencies differ between the cases and controls. We estimate these frequencies by r_i/R and s_i/S, for the cases and controls, respectively, for $i = 0$, 1, or 2. Furthermore, for the situation in which the subjects are independent (i.e., unrelated), all of the covariance terms equal 0, and the variance approximation in equation (4) reduces to the standard variance estimate of the log-odds ratio—that is, the variance approximation for the log-odds ratio becomes $1/r_i + 1/r_0 + 1/s_i + 1/s_0$.

A confidence interval for the log-odds ratio is found as follows: Let $z_{\alpha/2}$ denote the quantile, from the standard normal distribution, with a right-tail probability of $\alpha/2$. An approximate $100(1-\alpha)\%$ confidence interval (CI) for $\log(\hat{\psi}_i)$ is then

$$\log(\hat{\psi}_i) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\log(\hat{\psi}_i)]} .$$

Taking the antilogs of the endpoints gives an approximate $100(1-\alpha)\%$ CI for $\hat{\psi}_i$.

Data Application

To illustrate the use of our trend test, we applied it to a prostate cancer study of 948 white men, of whom 445 had prostate cancer and 503 were population controls. The 445 cases were enrolled in a genomewide linkage study of prostate cancer. To be eligible for the linkage study, a pedigree was required to have at least three closely related men with prostate cancer, at least two of whom were required to give a blood sample. A total of 163 pedigrees were ascertained. In addition, 18 "singleton" pedigrees were provided, in which there were at least three affected relatives but in which only one affected man provided a blood sample. Controls were randomly sampled from Olmsted County in southeastern Minnesota. Each man was genotyped for the variant of the luteinizing hormone β (LHβ) gene that exists on chromosome 19. The genotype distribution for the data is summarized in table 2, with the variant allele denoted as "*A*."

In our sample, only the cases were biologically related, resulting in 409 affected relative pairs. We used GENEHUNTER to estimate the IBD-sharing probabilities, and we programmed an S plus function (Statistical Sciences) to calculate our trend test. If weights are assumed to be equal to the number of high-risk alleles a person possesses, the value of our trend statistic is $\chi_1^2 = 2.61$ ($P = .11$). If family relationships are ignored, the naive Armitage trend test is $\chi_1^2 = 3.10$ ($P = .08$), which is biased and therefore illustrates the importance of accounting for related subjects. Segregation analyses of prostate cancer have suggested the presence of an autosomal dominant susceptibility gene. If we use a weighting scheme in which a person is scored as 1 if they have one or two variant alleles and as 0 otherwise, corresponding to a dominant scoring scheme, our test that accounts for family relationships is $\chi_1^2 = 3.61$ ($P = .06$), in contrast to the naive Armitage trend test $\chi_1^2 = 4.27$ ($P = .04$). If the *a/a* genotype is used as the baseline genotype, the odds ratio for a person with one high-risk allele is $\hat{\psi}_1 = 1.51$, with a 95% CI of 1.03–2.20, whereas the 95% CI that ignores the familial relationships is 1.06–2.14. The odds ratio for one or more high-risk alleles versus no high-risk alleles is found to be 1.44, with a correct 95% CI of 0.99–2.10 and an incorrect naive 95% CI of 1.02–2.04. The odds ratio for the homozygous *A/A* genotype, $\hat{\psi}_2$, is 0, because none of the cases were ho-

Table 2

Distribution for the LHβ Genotype

	No. of Subjects with Genotype[a]			
Sample	*a/a*	*a/A*	*A/A*	Total
Cases	362	83	0	445
Controls	434	66	3	503
Total	796	149	3	948

[a] *A* is the high-risk allele.

mozygous for the variant allele. Although both methods of analysis result in nonsignificant associations, our test correctly accounts for the correlation among the cases, resulting in larger variances—and, hence, smaller χ^2 statistics and wider confidence intervals—than those resulting from the incorrect, naive method.

Discussion

Similar to traditional association methods, the analysis method that we have presented is used to compare the genotype frequencies in cases versus those in controls. However, unlike the traditional methods, our test allows for the sampling of biologically related individuals, such as related cases from multiplex pedigrees, because it accounts for the correlation among family members. Ignoring the biological relationships will inflate the type I error rate. Randomly sampling one case from a multiplex family circumvents the correlation among cases but will lead to a loss of information. Our method overcomes this loss by using all subjects. Furthermore, the more distant the sampled related cases are, the less they are correlated, thereby potentially increasing the power of our trend test. If no related subjects are ascertained, then our method reduces to the usual Armitage test for trend.

Although unaffected family members (i.e., familial controls) can be used in our method, the optimal study design is one that uses unrelated controls. Advantages to the use of unrelated controls versus the use of family-based controls include the ease with which it is possible to match subjects according to important confounders, such as age and gender, and the larger pool of available controls. A disadvantage of using unrelated controls, however, is the possible increase in false positives that is due to population stratification or admixture. This potential problem can be circumvented through use of family-based controls, although the robustness against the effect of population stratification depends on the choice of familial controls; parental and sibling controls provide unbiased estimates of the genetic association, whereas cousin controls may bias the estimates if the cousins are not from the same gene pool as the cases (Witte et al. 1999). Nevertheless, research has shown that traditional case-control studies are more powerful and efficient than are their family-based counterparts (Khoury and Yang 1998; Morton and Collins 1998; Teng and Risch 1999; Risch 2000) and that the greater power of case-control studies may outweigh the potential increase in false positives that is caused by population stratification. Furthermore, the extent of the bias caused by population stratification is arguable. Work presented by Bacanu et al. (2000) and Wacholder et al. (2000) and in figure 1 of Witte et al. (1999) demonstrates that only with extreme differences in both allele frequency and disease prevalence does the effect of population stratification reach unacceptable levels. Rather than population stratification, a more likely explanation for the spurious results from case-control studies is the lack of a stringent significance criterion (Risch 2000), since many markers are tested, each with a small prior probability of association. In any case, a well designed case-control study that either samples cases and controls from homogeneous populations or matches subjects according to major confounding factors, including ethnic background, is an absolute necessity, as are appropriate statistical analyses. Both should minimize the effects of population stratification, if it is present.

If, however, one has reason to suspect that a given sample is subject to population stratification, then one can use the genomic control method as a possible correction (Devlin and Roeder 1999). This method uses null genetic markers (i.e., markers that are believed to be independent of the candidate gene and disease, as well as of each other) to estimate the background association due to population structure. A tendency for a positive association of disease with the null markers is indicative of population stratification, since each null marker is presumed to be unrelated to the disease. This tendency can be summarized over all null markers by a parameter denoted as "λ." The candidate-gene–association χ^2 statistic can then be adjusted for population stratification, by dividing it by λ. In the absence of population stratification, λ is expected to be 1; otherwise, λ will tend to be >1 and therefore will reduce the numerical value of the χ^2 test for association of the candidate gene.

Note that the conditional (or posterior) IBD-sharing probabilities, which are used in the calculation of the covariance terms for biologically related individuals, were calculated by GENEHUNTER, which assumes HWP. Although deviations from HWP may bias IBD-sharing estimates, we expect this to have little impact on the results of our trend statistic; however, further work is necessary to verify this. An alternative approach would be to use the prior IBD-sharing probabilities—that is, the probabilities that ignore the genotype information; however, this approach could result in a loss of power.

Moreover, we note that the odds-ratio estimates will be consistent estimators, even though these estimates do not take into account the oversampling of subjects from the same family. This is because the estimates, which are of the form *ad/bc,* are from a model (i.e., the unconditional logistic-regression model) that is a special case of the generalized linear models considered by Liang and Zeger (1986), who showed the consistency of the estimates for correlated longitudinal data. Our work differs from the problem considered by Liang and Zeger (1986), in that we present variances that are based

on IBD-sharing probabilities rather than on the sandwich estimator, thereby producing more-efficient estimators, although we plan to investigate this further in future work.

In summary, the use of multiplex families in a case-control setting provides a powerful approach to testing for genetic association (Teng and Risch 1999; Risch 2000). Our trend statistic, which is an extension of the Armitage (1955) test for trend, provides an analysis method for these designs. It uses all the available biologically related cases from multiplex families, as well as unrelated population controls and even unaffected family members. Furthermore, it offers flexibility in the analysis, allowing the user to choose different ways to score the genotype. Finally, the magnitude of the association can be estimated by the usual odds ratio, but with the variance of the odds ratio accounting for the correlated data.

Acknowledgments

This research was supported by the United States Public Health Service, National Institutes of Health contract grant number DE13276.

Appendix

The joint genotype-probability matrix $P(g_i,g_j)$ for the ith and jth subjects can be calculated by the ITO method (Li 1955). The ITO method uses three stochastic matrices: I, T, and O. Each row of each matrix corresponds to the conditional probability of genotype m, given both genotype l and IBD status for subjects i and j. Thus, the rows of each matrix sum to unity. The matrix I, which is the identity matrix, gives the conditional probabilities when the relatives share two alleles IBD; matrix O gives the conditional probabilities when the relatives share no alleles IBD, in which the elements of each row of O are the genotype probabilities p_0, p_1, and p_2; and, finally, the matrix T gives the conditional probabilities when the relatives share one allele IBD, and has the form

$$T = \begin{array}{cc|ccc|} & & & \mathbf{y}_i & \\ & & a/a & A/a & A/A \\ \hline \mathbf{y}_i & a/a & q & p & 0 \\ & A/a & q/2 & 1/2 & p/2 \\ & A/A & 0 & q & p \\ \hline \end{array} .$$

Thus, when Bayes's theorem and the ITO method are used, the matrix $P(g_i,g_j)$ is given by

$$\begin{aligned} P(g_i,g_j) &= P(g_i)P(g_j|g_i) \\ &= P(g_i)\sum_k P(g_j|g_i,\text{IBD} = k)P(\text{IBD} = k) \\ &= P(g_i)\left(\pi_2 I + \pi_1 T + \pi_0 O\right) , \end{aligned}$$

where π_k is the probability that the ith and jth subjects share k alleles IBD and where $P(g_i)$ is a matrix with genotype probabilities, p_i, along the diagonal, and 0 elsewhere.

References

Armitage P (1955) Tests for linear trends in proportions and frequencies. Biometrics 11:375–386

Bacanu S-A, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Khoury MJ, Yang Q (1998) The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 9:350–354

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Li CC (1955) Population genetics. University of Chicago Press, Chicago

Liang K, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Morton NE, Collins A (1998) Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci USA 95:11389–11393

Rao CR (1973) Linear statistical inference and its applications. 2d ed. John Wiley & Sons, New York

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405:847–856

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. Genome Res 8:1273–1288

Sasieni PD (1997) From genotypes to genes: doubling the sample size. Biometrics 53:1253–1261

Teng J, Risch N (1999) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Res 9:234–241

Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 92:1151–1158

Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149:693–705