Am. J. Hum. Genet. 68:1189–1196, 2001

Regression Models for Linkage Heterogeneity Applied to Familial Prostate Cancer

Daniel J. Schaid,[1,2] Shannon K. McDonnell,[2] and Stephen N. Thibodeau[3]

Departments of [1]Health Sciences Research, [2]Medical Genetics, and [3]Experimental Pathology, Mayo Clinic/Foundation, Rochester, MN

Linkage heterogeneity frequently occurs for complex genetic diseases, and statistical methods must account for it to avoid severe loss in power to discover susceptibility genes. A common method to allow for only a fraction of linked pedigrees is to fit a mixture likelihood and then to test for linkage homogeneity, given linkage (admixture test), or to test for linkage while allowing for heterogeneity, using the heterogeneity LOD (HLOD) score. Furthermore, features of the families, such as mean age at diagnosis, may help to discriminate families that demonstrate linkage from those that do not. Pedigree features are often used to create homogeneous subsets, and LOD or HLOD scores are then computed within the subsets. However, this practice introduces several problems, including reduced power (which results from multiple testing and small sample sizes within subsets) and difficulty in interpretation of results. To address some of these limitations, we present a regression-based extension of the mixture likelihood for which pedigree features are used as covariates that determine the probability that a family is the linked type. Some advantages of this approach are that multiple covariates can be used (including quantitative covariates), covariates can be adjusted for each other, and interactions among covariates can be assessed. This new regression method is applied to linkage data for familial prostate cancer and provides new insights into the understanding of prostate cancer linkage heterogeneity.

Introduction

Genetic heterogeneity creates significant challenges to efforts to discover the genetic basis of complex genetic diseases. Although the causes of heterogeneity may be varied, locus heterogeneity, which occurs when only a subset of families demonstrate linkage to a chromosomal region of interest, can be particularly damaging. If linkage heterogeneity is ignored when the analysis is performed, the power to detect linkage is dramatically reduced.

A widely used method to account for locus heterogeneity is based on a likelihood composed of a mixture of family types—linked and nonlinked (Smith 1961). Likelihood-ratio tests can then be constructed either to test for linkage homogeneity given the existence of linkage (admixture test), or to test for linkage while allowing for heterogeneity (heterogeneity $\log_{10}$ odds [HLOD] score) (Ott 1999). This likelihood method and its various extensions, such as allowance for a trait locus to be linked to markers on any number of chromosomes (Bhat et al. 1998), are available in the widely used suite of HOMOG programs (Ott 1999). The additional heterogeneity parameters in the mixture likelihood can make the method robust to model misspecification, which is appealing for linkage analysis of complex human diseases. For example, a mixture likelihood gives a good approximation to full two-locus models (Schork et al. 1993); however, this approach has limitations. If the probability that a pedigree is the linked type depends on characteristics of the phenotype (e.g., age at disease onset), then the results from the admixture likelihood may be biased (Janssen et al. 1997). An obvious limitation is that characteristic features, such as mean age at disease onset, that may help to distinguish linked from nonlinked pedigree types are not directly incorporated into the likelihood methods. To address this limitation, pedigrees can be stratified according to their features, and then each subset can be analyzed separately. Subset analyses, however, introduce several problems: multiple testing can inflate the number of false-positive conclusions, the comparison of linkage information across subsets can be cumbersome, the combination of multiple pedigree features can lead to a small number of families in some subsets, and quantitative pedigree features must be split into categories. Although stratification can increase the power to detect linkage when the stratification factor adequately represents the locus heterogeneity, stratification on insignificant features can reduce power (Leal and Ott 2000).

To overcome the limitations of subset analyses, we

Received December 14, 2000; accepted for publication March 15, 2001; electronically published April 13, 2001.

Address for correspondence and reprints: Dr. Daniel J. Schaid, Harwick 775, Section of Biostatistics, Mayo Clinic/Foundation, Rochester, MN 55905. E-mail: schaid@mayo.edu

0002-9297/2001/6805-0013$02.00

developed a regression-based extension of the mixture likelihood. For this method, pedigree features are used as covariates that determine the probability that a pedigree is the linked type. We apply this regression method to two published studies of linkage analyses of familial prostate cancer (MIM 176807) to illustrate some of the strengths of this approach.

Methods

Regression Model for Linkage Heterogeneity

To present the mixture likelihood for each pedigree, let α denote the probability that a pedigree is the linked type, and let $L_i(\theta)$ denote the likelihood for the ith pedigree at recombination fraction θ. Then, the mixture likelihood for a pedigree is $L_i(\alpha,\theta) = \alpha L_i(\theta) + (1 - \alpha)L_i(0.5)$, and the likelihood for the collection of all N pedigrees is $L(\alpha,\theta) = \prod_{i=1}^{N} L_i(\alpha,\theta)$. But, because it is common practice to compute LOD scores for the pedigrees, the likelihood can also be written in terms of LOD scores, by dividing $L_i(\alpha,\theta)$ by $L_i(0.5)$ and substituting $L_i^*(\alpha,\theta) = [\alpha 10^{LOD_i(\theta)} + (1 - \alpha)]$ for $L_i(\alpha,\theta)$.

The mixture likelihood can be extended by allowing the α to depend on covariates that characterize the features of the pedigree. Let x_i denote a vector of pedigree features for the ith pedigree, with the first element of x_i equal to 1, for the intercept. We model the pedigree features by the logistic regression

$$\alpha(\beta|x_i) = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} ,$$

where the vector of regression parameters, β, captures the linkage heterogeneity information provided by the pedigree features. The resulting mixture likelihood for the ith pedigree depends on both β and θ according to $L_i^*(\beta,\theta|x_i) = [\alpha(\beta|x_i)10^{\mathrm{LOD}_i(\theta)} + 1 - \alpha(\beta|x_i)]$, and the total log likelihood for all families is $\ln L(\beta,\theta) = \sum_{i=1}^{N} \ln L_i^*(\beta,\theta|x_i)$.

To estimate the parameters β and θ, we maximize the mixture likelihood by using the expectation-maximization (EM) algorithm. To illustrate this procedure, first consider the simpler situation of modeling the effects of pedigree covariates on the probability of linkage when it is known which pedigrees are linked. In this case, we would let the dependent value y_i have a value of 1 if linked and of 0 if not linked, and then we would simply use logistic regression. But, because pedigree type (linked or nonlinked) is unknown, the expectation step of the EM algorithm is used to estimate the expectation that a pedigree is a linked type. This is achieved by using current values of β and θ to compute the posterior probability that a family is a linked type,

$$post_i = \frac{\alpha(\beta|x_i)10^{\mathrm{LOD}_i(\theta)}}{\alpha(\beta|x_i)10^{\mathrm{LOD}_i(\theta)} + 1 - \alpha(\beta|x_i)} .$$

Then, the maximization step entails maximizing a weighted logistic-regression function, with the values of the posterior probabilities serving as weights. That is, each family is used twice, once as linked and once as nonlinked, so each family contributes the following to the logistic regression: the vector (1,0) of "dependent" y_i values, the vector of repeated covariates (x_i, x_i), and the vector of corresponding weights ($post_i$,[$1 - post_i$]). Standard software can be used to maximize this logistic-regression function, to determine updated values for β. These new β values are then used to determine $\alpha(\beta|x_i)$ for each family, which, in turn, are used to determine the value of θ that maximizes $\ln L(\beta,\theta)$. Then, the updated values of β and θ are used for the expectation step, and the EM cycle is repeated until the change in $\ln L(\beta,\theta)$ is small.

Likelihood-ratio test (LRT) statistics can be computed to test hypotheses about the β parameters. The LRT is twice the difference between a full-model log likelihood and a reduced-model log likelihood. Several types of hypotheses can be formulated: (1) a test for linkage, allowing for heterogeneity; (2) a test for homogeneity (i.e., all pedigrees linked), given linkage; and (3) a test of whether the pedigree features discriminate, at a statistically significant level, between linked and nonlinked pedigrees, given linkage. If linkage heterogeneity exists but is not explained by the covariates, then only the intercept, β_0, differs from 0. In this case, our regression model provides the same information as the usual HLOD computed by HOMOG, because $\beta_0 = \log[\alpha/(1 - \alpha)]$. In this report, we focus on the third type of hypothesis testing. It should be noted, however, that the LRT for linkage, allowing for heterogeneity, has a complex distribution. When there is no linkage, the fraction of linked pedigrees is not defined; and even in the simplest case, with only an intercept in the regression model, the asymptotic distribution of the HLOD is a complex mixture of χ^2 distributions with mixing probabilities that depend on the linkage information provided by the pedigrees (Chiano and Yates 1995) and the assumed genetic model (Huang et al. 2000). However, if linkage exists, then the LRT is expected to follow a χ^2 distribution. Therefore, to test whether the pedigree features discriminate, at a statistically significant level, between linked and nonlinked pedigrees, given linkage, we compute probability values for the LRTs according to the usual χ^2 distribution. The heterogeneity regression analyses were conducted by S-PLUS software (Insightful, Corp.), and locally written functions, which can be loaded into S-PLUS and are available at the authors' Web site.

Chromosome 1 Data Set

An international collaborative effort to characterize the evidence for linkage of hereditary prostate cancer to chromosome region 1q24-25 reported an HLOD of 1.40 and an estimated 6% of the families linked to this region (Xu and International Consortium for Prostate Cancer Genetics 2000). This collaborative study was based on 772 families that were contributed by nine groups of investigators. Because of informed-consent issues, it was not possible to consider data for individual members of each pedigree. Therefore, only the information summarizing the pedigree, including pedigree characteristics and LOD scores, was pooled. Multipoint LOD scores were computed by GENEHUNTER (Kruglyak et al. 1996), with six markers, spanning a region of 24.1 cM, and an assumed-dominant model (see details in the report by Xu and International Consortium for Prostate Cancer Genetics 2000). The software HOMOG (Ott 1999) was used to compute the HLOD for the entire group of pedigrees, as well as for multiple subsets. Subsets were defined by the presence or absence of male-to-male disease transmission of prostate cancer, by the mean age at diagnosis of prostate cancer among pedigree members, by the number of affected family members, and by combinations of male-to-male disease transmission with the other two features. The conclusion from this report was that *HPC1,* the putative gene in the region 1q24-25, accounts for a significant fraction of hereditary prostate cancer in the subset of families characterized by at least five affected family members, younger mean age at diagnosis (<65 years), and male-to-male disease transmission. This conclusion was based on 48 families, with an HLOD of 2.25 and $\alpha = .29$.

Chromosome 20 Data Set

Another study of hereditary prostate cancer performed similar subset analyses for chromosome 20 linkage in 162 families with prostate cancer (Berry et al. 2000). The strongest evidence for linkage was detected in the subset of families that was mutually exclusive of the subset used in the chromosome 1q24-25 combined analysis, that is, the families characterized by fewer than five affected family members, older mean age at diagnosis ($\geq$66 years), and no male-to-male disease transmission. The HLOD from these 19 families was 2.34, with $\alpha = .75$.

Our regression model for linkage heterogeneity was applied to the combined data for 772 families studied for chromosome 1 linkage and the data for 162 families studied for chromosome 20 linkage, to explore the contribution and interaction of each covariate. The information available for each pedigree included multipoint LOD scores and information about factors that were used as covariates. For each data set, the contribution of mean age at diagnosis was evaluated as a binary covariate, as presented in the original reports (using a cutoff of age 65 years for the chromosome 1 data and a cutoff of age 66 years for the chromosome 20 data), as well as by other models (linear age in years, quadratic age effect [incorporating both age and age^2 in the model], log of age, and four age categories). The other covariates were included as binary terms (five or more vs. fewer than five affected family members and presence vs. absence of male-to-male disease transmission).

Simulations

To gain some insights into the statistical properties of the likelihood-ratio statistic, we performed a limited set of simulations. A total of 100 families, each with four affected siblings, were simulated by SLINK (Weeks et al. 1990), according to an autosomal dominant model (Xu and International Consortium for Prostate Cancer Genetics 2000), with a rare mutant–allele frequency of .003 and with penetrances of .001 and 1.0 for noncarriers and carriers, respectively. Conditional on disease status, marker genotypes were simulated such that the disease locus was midway between two markers spaced 10 cM apart. To consider linkage heterogeneity, two types of pedigrees were simulated; 50 pedigrees linked with probability α_1 and 50 pedigrees linked with probability α_2. Multipoint linkage analyses were performed by GENEHUNTER (Kruglyak et al. 1996), using the simulation genetic model for analyses. A covariate, having a value of 0 if a type 1 pedigree and a value of 1 if a type 2 pedigree, was created for each pedigree. This covariate, along with the multipoint LOD scores per pedigree, was used to create an LRT to test the null hypothesis that $\beta = 0$, which implies that the linkage heterogeneity is not explained by the covariate. The criterion for statistical significance was LRT > 3.84 and was based on a χ^2 distribution with 1 df and $P < .05$. The simulation process was repeated 100 times to compute the type I error rate and the power.

Results

Chromosome 1 Results

Before multiple covariates were assessed, we evaluated different ways to model the mean age at diagnosis, and the results are presented to demonstrate the flexibility of the regression method (table 1). The LRT was used to test each model versus the model with only an intercept. As demonstrated in table 1, none of the methods for including mean age at diagnosis resulted in a statistically significant effect of age. Although the linear model has a positive β_1, suggesting that the probability of a

Table 1

Chromosome 1 Linkage Heterogeneity Regression Models for Mean Age at Diagnosis

Model	Parameter	ln L	LRT[a] (df)	P
Intercept only	$\beta_0 = -2.67$	3.216	reference	
Linear	$\beta_0 = -10.97$	3.565	.70 (1)	.40
	$\beta_1 = .12$			
Quadratic	$\beta_0 = 56.83$	4.812	3.19 (2)	.20
	$\beta_1 = -1.81$			
	$\beta_2 = .01$			
Categories:				
<60 years	$\beta_0 = -1.75$	5.059	3.69 (3)	.30
60–65 years	$\beta_1 = -.42$			
65–70 years	$\beta_2 = -2.63$			
⩾70 years	$\beta_3 = -.27$			
Binary:				
<65 years	$\beta_0 = -2.08$	3.918	1.40 (1)	.24
⩾65 years	$\beta_1 = -1.08$			

NOTE.—For all models, the estimated $\theta = .06$.

[a] LRT against model with intercept only.

linked type of pedigree increases with age, there is strong evidence that the effect of age is actually not linear. The more flexible model with four age categories suggests that the youngest age group (<60 years) has the highest fraction of linked pedigrees, because the negative coefficients for the older age groups are added to the intercept and hence decrease the probability that they are linked. In fact, the third age group (65–70 years) had the lowest estimated fraction of linked pedigrees, demonstrating a nonlinear effect of age. The models with the smallest P values provided evidence that they may be the best discriminators of linkage heterogeneity. Although the quadratic effect of age may be the best, binary coding, which was used in the original published report, seems to capture the effect of age just as well and provides a simpler interpretation. For this reason, the binary coding was used for all subsequent models. On the basis of results of the binary age effect, the predicted probability that a pedigree is a linked type is 11% when mean age at diagnosis is <65 years and is 4% when mean age is ⩾65. These results agree with the published subset analyses, because whenever a single categorical covariate is included in the regression model, the fraction of linked pedigrees estimated by the regression model agrees with the fractions calculated by HOMOG for each of the category subsets. This is not necessarily true when multiple covariates are included in the regression model.

To evaluate the simultaneous effects of the three covariates (number affected, mean age at diagnosis, and male-to-male disease transmission), a backwards stepwise regression was performed. A full model with all three covariates was fitted, and then the least-significant covariates were eliminated in a stepwise fashion. The advantage of evaluating covariates simultaneously is that the influence of each covariate is adjusted for the effects of the others. For the chromosome 1 data, male-to-male transmission was positively correlated with a greater number of affected family members (odds ratio 3.02 [$P < .001$] for the binary covariates). At the first step, the tests for male-to-male disease transmission, mean age at diagnosis, and number affected (each covariate adjusted for the other two) resulted in $P = .01$, $P = .10$, and $P = .29$, respectively. After excluding number affected, the tests for male-to-male disease transmission and mean age at diagnosis resulted in $P = .01$ and $P = .13$, respectively. These stepwise models (summarized in table 2) suggest that only male-to-male disease transmission is a statistically significant ($P = .02$) predictor of whether a pedigree is the linked type. On the basis of the final model, the predicted probability that a pedigree is a linked type is near 0% for pedigrees without male-to-male disease transmission and is 11% for those with male-to-male disease transmission. Although our results suggest that a younger mean age at diagnosis is more likely to be of the linked type—because the regression coefficient for older age is negative (−1.14)—this covariate does not achieve statistical significance.

The report by Xu and International Consortium for Prostate Cancer Genetics (2000) concluded that the evidence for linkage was strongest in the subset of pedigrees meeting all three criteria—male-to-male disease transmission, younger mean age at diagnosis, and ⩾5 affected family members—which implies interaction of the covariates. To examine whether a combination of pedigree features provides significant discrimination between linked and nonlinked types of pedigrees, we fitted a series of interaction models. All models included the three main effects but used different interaction terms; three models included each of the two-way interaction terms, and another model included the three-way interaction. For these interaction models, the highest LRT

Table 2

Chromosome 1 Linkage Heterogeneity Stepwise Regression of Pedigree Covariates

Step and No.[a]	Intercept	Regression Coefficients for			
		Transmission[b]	No. Affected[c]	Age[d]	ln L
1, 3	−13.00	11.17	.83	−1.28	7.63
2, 2	−11.28	9.85		−1.14	7.06
3, 1	−12.34	10.29			5.92

[a] No. of covariates.

[b] Male-to-male disease transmission.

[c] No. of pedigree members affected (<5 vs. ⩾5).

[d] Mean age at diagnosis (<65 years vs. ⩾65 years).

Table 3

Chromosome 20 Linkage Heterogeneity Regression Models for Mean Age at Diagnosis

Model	Parameter	ln L	LRT[a] (df)	P
Intercept only	$\beta_0 = -1.95$	2.480	Reference	
Linear	$\beta_0 = -22.26$	4.165	3.37 (1)	.07
	$\beta_1 = .30$			
Quadratic	$\beta_0 = 23.26$	4.269	3.58 (2)	.17
	$\beta_1 = -1.04$			
	$\beta_2 = .01$			
Categories:				
<60 years	$\beta_0 = -18.16$	4.142	3.32 (3)	.34
60–65 years	$\beta_1 = 14.90$			
65–70 years	$\beta_2 = 16.51$			
⩾70 years	$\beta_3 = 17.42$			
Ranked	$\beta_0 = -4.14$	4.038	3.12 (1)	.08
	$\beta_1 = 1.17$			
Binary:				
<66 years	$\beta_0 = -4.51$	3.213	1.47 (1)	.23
⩾66 years	$\beta_1 = -1.08$			

NOTE.—For all models, the estimated $\theta = .06$.

[a] LRT against model with intercept only.

score for interaction was .42 ($P = .52$). Given that only male-to-male disease transmission was significant in the stepwise selection and given that no tests demonstrated significant interactions, it is unlikely that there is any three-way interaction. This, in turn, suggests that the evidence of strongest linkage in the subset meeting all three criteria is driven mainly by the effect of male-to-male disease transmission.

Chromosome 20 Results

To explore the effects of mean age at diagnosis on chromosome 20 linkage results, we first evaluated age by fitting several models (table 3), similar to the approach illustrated above for chromosome 1 analyses. Although none of the models demonstrated a statistically significant effect of age, there is a strong suggestion that an increasing mean age at diagnosis increases the probability that a family is the linked type. This is demonstrated by positive regression coefficients for increasing age for almost all models. The model with four age categories is fairly robust, and the values of the estimated regression coefficients suggest that the log odds of the probability that a family is the linked type increases approximately linearly across the three older age groups. Hence, a model was fitted with an age rank of 0, 1, 2, or 3, according to whether the mean age at diagnosis was <60, 60–65, 65–70, or ⩾70 years. The resulting P value for this ranked age–category model ($P = .08$) was close to that found for the linear age model ($P = .07$), yet the ranked age–category model may be more robust.

To illustrate the importance of appropriately accounting for the influence of a covariate, we performed stepwise selection of the regression coefficients for the three covariates (number affected, mean age at diagnosis, and presence or absence of male-to-male transmission); however, in one analysis we used the binary coding for age, and in a second analysis we used the ranked-age category (table 4). For the binary age covariate, the tests for male-to-male transmission, mean age at diagnosis, and number affected, with each covariate adjusted for the other two, resulted in P values of .0005, .13, and .25, respectively. After excluding number affected, the tests for male-to-male transmission and mean age at diagnosis resulted in P values of .0008 and .28, respectively. This stepwise procedure resulted in a final model that included only male-to-male disease transmission as a statistically significant predictor ($P < .001$) of whether a pedigree is the linked type. In contrast, when mean age at diagnosis was included as the ranked-age category, it was retained in the stepwise regression ($P = .004$), along with male-to-male disease transmission ($P < .001$); only the number affected was nonsignificant ($P = .24$). Note that the influence of age was more striking when it was adjusted for male-to-male disease transmission ($P = .08$ when age-rank was tested by itself vs. $P =$

Table 4

Chromosome 20 Linkage Heterogeneity Stepwise Regression of Pedigree Covariates: Comparison of Binary Versus Ranked Age Category

			Regression Coefficients for			
Age Covariate and Step	No. of Covariates	Intercept	Male-to-Male Transmisson (Yes vs. No)	No. Affected (<5 vs. ⩾5)	Mean Age at Diagnosis (Years)	ln L
Binary:						
1	3	−1.06	−9.68	−.91	2.29	9.55
2	2	−.41	−3.75		1.20	8.89
3	1	.23	−5.36			8.31
Ranked:						
1	3	−24.58	−15.38	−1.79	12.70	13.07
2	2	−28.59	−17.60		14.35	12.37

Table 5

Correlation of Male-to-Male Disease Transmission with Mean Age at Diagnosis

	Pedigrees with Mean Age at Diagnosis of			
M to M[a]	<60 Years *n* (%)	60–65 Years *n* (%)	65–70 Years *n* (%)	⩾70 Years *n* (%)
No	5 (38)	13 (27)	16 (28)	12 (28)
Yes	8 (62)	36 (73)	41 (72)	31 (72)
Total	13 (100)	49 (100)	57 (100)	43 (100)

[a] Male-to-male disease transmission.

.004 when age-rank was adjusted for male-to-male disease transmission). This masking of the age effect occurred because families with a earlier mean age at diagnosis (<60 years) had a lower frequency of male-to-male disease transmission compared with families with a later age at diagnosis (table 5). This highlights the advantage of simultaneously evaluating various pedigree features. Pairwise interactions were explored in a manner similar to that used for chromosome 1 analyses, both for the binary and the ranked-category methods of including age. The largest LRT for interaction over all models was .58 ($P = .45$), suggesting that there is no interaction between any of the three covariates.

Simulation Results

Our limited simulations, for a single covariate, are presented in table 6. These results indicate that the LRT is conservative (with simulated type I error rates of 1% and 4%) and that, unless the covariate has a large influence on linkage heterogeneity, the power of the LRT is weak. Further simulations are required to evaluate the influence of multiple covariates, as well as interactions.

Discussion

We have presented a regression-based method that allows evaluation of whether the amount of linkage heterogeneity differs with values of pedigree covariates. The regression procedure provides several benefits. It allows use of a flexible manner to model the effects of covariates (including continuous covariates); the contribution of each covariate is adjusted for the other covariates (useful for correlated covariates), the relative contribution of the covariates can be easily compared, and interaction terms can be evaluated. In contrast, subset analyses are often based on ad hoc groupings chosen to maximize the HLOD, which can lead to an increased chance of false-positive conclusions and, perhaps, misleading interpretations. Furthermore, stratification does not provide a mechanism to test whether linkage heterogeneity varies over the strata. Finally, if the fraction of linked families does not differ dramatically over the various subsets, then splitting the data into more subsets than necessary can diminish the power to detect linkage (Leal and Ott 2000).

The strength of the regression-based approach, in contrast to the creation of subsets, is illustrated by the results for both chromosomes 1 and 20. That is, if a single covariate is a strong predictor whether families are of the linked type (male-to-male disease transmission for chromosome 1), then combining that covariate with an insignificant covariate, in an attempt to refine the subsets, can cause misleading conclusions that the combined covariate effects are important. Random variation is likely to cause the amount of linkage heterogeneity due to male-to-male disease transmission to vary over levels of an insignificant covariate. Therefore, we are very likely, by chance alone, to find stronger linkage evidence, as measured by HLOD, in one of the subsets created by combining male-to-male transmission with the insignificant covariate. In contrast, the regression-based approach allows direct testing of whether the combined effect of covariates (i.e., covariate interaction) is statistically significant.

A limitation of our regression method is that a moderate number of pedigrees are likely to be required to provide numerical stability in the estimation procedure. Although the β parameters provide a flexible method to account for linkage heterogeneity explained by the pedigree covariates, one should be cautious about their interpretation. It is unlikely that the fraction of linked pedigrees in a population can be estimated without bias, because the parameter estimates depend on the assumed

Table 6

Simulation Results for LRT to Test whether a Covariate Explains Linkage Heterogeneity

α_1 (%)	α_2 (%)	Significant LRT (%)
0	5	4
	10	0
	20	12
	50	68
5	5	1[a]
	10	8
	20	11
	50	60
10	5	3
	10	4[a]
	20	8
	50	45

Note.—Data simulate 50 pedigrees linked with probabilty α_1 and 50 pedigrees linked with probabilty α_2.

[a] Type I error rate.

genetic model and on how the pedigrees were ascertained. If the assumed genetic model is wrong (for example, if the pedigrees of the linked type do not have the same genetic model—i.e., allele frequency and penetrance—as the pedigrees that are not linked), then the estimated fraction of linked pedigrees will be biased (Logue and Vieland 2000)

Our results suggest that linkage heterogeneity for prostate cancer may be partially explained by families with male-to-male disease transmission, which are more likely to be linked to chromosome 1, and families without male-to-male disease transmission, which are more likely to be linked to chromosome 20. On the basis of the chromosome 1 final model, the predicted probability that a pedigree is a linked type is near 0% for pedigrees without male-to-male transmission and is 11% for those with male-to-male transmission. For the final model developed for the chromosome 20 data, which included both male-to-male disease transmission and ranked age categories, the model predictions are as follows: for pedigrees without male-to-male disease transmission, the fraction of linked pedigrees is near 0% if mean age at diagnosis is <65 years, 53% if age is 65–70 years, and 100% if age is ⩾70 years; for pedigrees with male-to-male disease transmission, the predicted fraction of linked pedigrees is near 0% when mean age at diagnosis is <70 years and is 4% when mean age at diagnosis is ⩾70 years. Note that the effect of mean age at diagnosis was in opposite directions for chromosomes 1 and 20; later mean age is less likely linked to chromosome 1 and more likely linked to chromosome 20.

A direct method to evaluate the simultaneous influence of linkage to two different chromosomes is to fit a mixture model with three types of pedigrees; a fraction linked to chromosome 1, a fraction linked to chromosome 20, and the remainder not linked to either chromosome (as implemented in HOMOG3R [Ott 1999]). Although the combined analysis of 772 families for chromosome 1 did not have chromosome 20 markers available, we did apply HOMOG3R to our subset of 142 families from the Mayo Clinic study that had both chromosome 1 and 20 data. The HLOD score was 1.43, with ~10% of the families linked to chromosome 1 and ~18% linked to chromosome 20. However, this approach does not consider the features of the pedigrees. It is possible to extend our regression-based method to allow for multiple loci by introducing a regression equation for each locus; however, it is likely that this extension would require a large number of families to obtain reliable estimates of the many regression parameters. Nonetheless, our proposed regression method should allow better exploration of linkage heterogeneity, as described by features of the pedigrees. As with any regression analysis, covariates should be coded with care, and the number of multiple regression models should be restricted to avoid the potential for an inflated rate of false-positive findings.

Acknowledgments

The authors gratefully acknowledge J. Xu and the International Consortium for Prostate Cancer Genetics (see listing of members in Xu and International Consortium for Prostate Cancer Genetics, 2000), for contributing the chromosome 1 data set, and Mayo Clinic investigators (R. Berry, J. Cunningham, J. J. Schroeder, A. J. French, and M. Brumm), for contributing the chromosome 20 data set. This research was supported by the Mayo Clinic Comprehensive Cancer Center and by National Institutes of Health contract grants DE13276, CA72818, and CA15083.

Electronic-Database Information

Accession number and URLs for data in this article are as follows:

Authors' Web site, http://www.mayo.edu/statgen/ (for functions used in these analyses)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for familial prostate cancer [MIM 176807])

References

Berry R, Schroeder JJ, French AJ, McDonnell SK, Peterson BJ, Cunningham JM, Thibodeau S, Schaid D (2000) Evidence for a prostate cancer–susceptibility locus on chromosome 20. Am J Hum Genet 67:82–91

Bhat A, Heath SC, Ott J (1999) Heterogeneity for multiple disease loci in linkage analysis. Hum Hered 49:229–231

Chiano MN, Yates JRW (1995) Linkage detection under heterogeneity and the mixture problem. Ann Hum Genet 59: 83–95

Huang J, Vieland V, Wang K (2000) The null distribution of the heterogeneity LOD score (HLOD) does depend on the assumed genetic model for the trait. Genet Epidemiol 19: 253

Janssen B, Halley D, Sandkuijl L (1997) Linkage analysis under locus heterogeneity: behavior of the A-test in complex analyses. Hum Hered 47:223–233

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

Leal SM, Ott J (2000) Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. Am J Hum Genet 66:567–575

Logue M, Vieland V (2000) The heterogeneity LOD cannot be used to estimate the population proportion of linked families. Genet Epidemiol 19:259

Ott J (1999) Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore

Schork NJ, Boehnke M, Terwilliger JD, Ott J (1993) Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53:1127–1136

Smith C (1961) Homogeneity test for linkage data. Proc Sec Int Congr Hum Genet 1:212–213

Weeks D, Ott J, Lathrop G (1990) SLINK: a general simulation program for linkage analysis. Am J Hum Genet Suppl 47: A204

Xu J, International Consortium for Prostate Cancer Genetics (2000) Combined analysis of hereditary prostate cancer linkage to 1q24-25: results from 772 hereditary prostate cancer families from the International Consortium for Prostate Cancer Genetics. Am J Hum Genet 66:945–957