Letters to the Editor

Am. J. Hum. Genet. 65:1195–1197, 1999

SPERMSEG: Analysis of Segregation Distortion in Single-Sperm Data

To the Editor:

Single-sperm typing has proved to be a valuable tool for detection of segregation distortion at a variety of loci in males (Williams et al. 1993; Leeflang et al. 1996; Takiyama et al. 1997; Girardet et al. 1998; Grewal et al. 1999; Takiyama et al. 1999). Sperm typing can provide the large sample sizes needed to detect even small deviations from 50:50 segregation that might arise during meiosis or that are due to differential sperm viability during or immediately after spermiogenesis (Leeflang et al. 1996). Furthermore, the problem of ascertainment bias, a serious concern in family-based studies of segregation distortion, is circumvented by sperm typing. However, in order to analyze such sperm data, one must model experimental errors—such as failure of alleles to amplify to a detectable level, deposit of 0 or >1 sperm in a sample, and contamination by exogenous DNA (Cui et al. 1989). Here I describe SPERMSEG, software programmed in C to analyze segregation in single-sperm data. This likelihood-based software is very flexible, allowing for any number of one- and two-marker data sets from one or more donors, with the capabilities to fit virtually any identifiable submodel of interest, to provide confidence intervals for all parameters, and to perform a wide range of hypothesis tests, including simulation-based goodness-of-fit tests.

For a likelihood analysis of segregation distortion using single-sperm data, the basic study design involves one or more two-marker data sets from each of several donors. By a two-marker data set, I mean that, for a given donor, two markers, for which the donor is heterozygous and which are linked to the locus of interest, are typed on each of a number of sperm. The reason two markers are typed is that if only one marker were typed on each sperm, it would not be possible to estimate the error parameters in the sperm-typing model. However, with additional assumptions, one-marker data sets can be included in the analysis, in addition to two-marker data sets. Such additional assumptions could include equality, between one-marker and two-marker data sets, of some of the error parameters. Only markers for which the donor is heterozygous can be included in the SPERMSEG analysis. Data from markers for which the donor is homozygous contain no information on segregation distortion (although they may contain a very small amount of information on the error parameters). Thus, it is assumed that each donor's sperm are typed only for markers for which the donor is heterozygous, with those markers allowed to differ among donors, and with possibly different pairs of markers typed for different subsets of sperm from the same donor.

Let G be the locus of interest, with alleles *G* and *g* in a given donor. Each two-marker data set involves sperm typed at a pair of markers A and B, at which a given donor has alleles *A*/*a* and *B*/*b*, respectively, linked to G. Assume that the donor haplotypes are known, say *GAB*/*gab*, and assume that the three recombination probabilities θ_{GA}, θ_{GB}, and θ_{AB}, between G and A, G and B, and A and B, respectively, are known. This parametrization of the recombination probabilities is completely general, to allow for interference. Special cases in which one or both of A and B are completely linked to G are allowed and lead to simplified calculations. The observed data for a given donor and pair of markers are assumed to be multinomial, with 16 possible outcomes: ––––, –––*b*, ––*B*–, ––*Bb*, ––*a*–, –*a*–*b*, –*aB*–, –*aBb*, *A*–––, *A*––*b*, *A*–*B*–, *A*–*Bb*, *Aa*––, *Aa*–*b*, *AaB*–, and *AaBb*, where, for example, –––– means that no allele was amplified to a detectable level, and, for example, –aBb means that alleles *a*, *B*, and *b* were detected, but *A* was not. For each donor, there may be several such two-marker data sets, with one or both of the linked markers differing among data sets. Different donors may be typed at different markers and, in general, will have alleles different from each other. In SPERMSEG, there is no limit on the overall number of markers or number of alleles, except that each sperm is assumed to be typed at no more than two markers.

Consider a single two-marker data set. The segregation-distortion model for the two-marker data set was described by Leeflang et al. (1996) for the special case when G, A, and B are completely linked, and it is also a good approximation when there is very tight linkage. This model includes segregation parameter $s = P$(sperm has allele *G*), with $1 - s = P$(sperm has allele *g*). It includes sperm-deposit parameters γ_i, which allow for the

possibility that, instead of one sperm being present in a given sample, either zero or two sperm are present, where $\gamma_i = P(i$ sperm present in a given sample), i = 0,1,2, with the assumption $\gamma_0 + \gamma_1 + \gamma_2 = 1$. The model includes amplification parameters α_A, α_a, α_B, and α_b, where, for example, α_A is the probability that allele *A* is amplified to a detectable level by PCR, given that it is present on a single sperm. If two sperm are deposited, both with allele *A*, then the two *A* alleles are assumed to amplify independently of one another, each with probability α_A. The contamination parameters are β_A, β_a, β_B, and β_b, where, for example, β_A is the probability that allele *A* is falsely detected because of contamination by exogenous DNA. This model is very close to the original model developed by Cui et al. (1989) for estimation of a recombination fraction, which was extended to three loci by Goradia et al. (1991). Both of these models include deposit parameters γ_i, i = 0,1,2,3,4, assumed to sum to 1, and amplification and contamination parameters as given above. Cui et al. (1989) have an unknown recombination fraction in their model, whereas Goradia et al. (1991) have two unknown recombination fractions with an unknown interference coefficient, a parametrization that is equivalent to our trio of recombination probabilities. Both models assume $s = .5$. The segregation-distortion model described here is a three-locus model, as in the article by Goradia et al. (1991), but s is allowed to vary, the amplification and contamination parameters for locus G are effectively set to 0, and the recombination probabilities are assumed to be known instead of estimated. Furthermore, since the parameters γ_3 and γ_4 are always estimated as 0 in the articles by Cui et al. (1989) and Goradia et al. (1991), I follow the lead of Lazzeroni et al. (1994), in setting them to 0. Cui et al. (1989) and Goradia et al. (1991) have not shown that their model including γ_3 and γ_4 is actually identifiable. If it is assumed that it is identifiable, there is certainly very little information, in a reasonably sized data set, with which to estimate γ_3 and γ_4, and the fact that γ_2 is typically estimated at just a few percent suggests that these values are close to 0 in any case.

Now, suppose that several two-marker data sets are to be analyzed simultaneously. For instance, one might have two two-marker data sets from donor 1 (perhaps with different markers; i.e., some sperm are typed at markers A and B, and other sperm are typed at markers C and D), a two-marker data set from donor 2, and a two-marker data set from donor 3. One might wish to analyze the data by using a model with, say, donor-specific segregation parameters, experiment-specific deposit parameters, allele-specific amplification parameters, and locus-specific contamination parameters. SPERMSEG is designed to be very flexible in allowing the user to specify such models and will maximize the likelihood subject to these constraints. Any segregation parameters may be set equal to each other or to fixed values—similarly for deposit, amplification, and contamination parameters. This is especially useful for testing hypotheses of interest, such as whether there is segregation distortion at all in a collection of data sets; whether, among donors within a phenotypic class, there is heterogeneity in the segregation ratio; and whether, among phenotypic classes, there is heterogeneity in the segregation ratio. Parameter estimates and the maximized log-likelihood are output for each model that the user selects. SPERMSEG allows the user to calculate confidence intervals for all estimated parameters under any of these models as well.

In addition to two-marker data sets, there may be one or more single-marker data sets. Each single-marker data set involves sperm typed at a single marker C, at which a given donor has alleles *C* and *c*, linked to G with known recombination fraction (possibly 0). The four possible multinomial observations are then –, -*c*, C-, C*c*. There are only three freely varying observed counts, so the segregation, deposit, amplification, and contamination parameters (seven parameters in all, in this case) cannot be estimated from such a data set alone. However, either in combination with two-marker data sets from which these parameters can be estimated, or with some of the error parameters assumed to be known, the single-marker data sets provide additional information. Thus, of the seven parameters in a one-locus data set, most of them either must be set equal to comparable parameters in some two-locus data set that is also included in the analysis or must be set equal to fixed values, if appropriate values are known. SPERMSEG allows for any number of one-marker data sets to be included in the analysis, in addition to the two-marker data sets, with the user specifying which parameters are to be set equal to each other or to fixed values, so that the model is identifiable.

SPERMSEG uses the expectation-maximization (EM) algorithm of Dempster et al. (1977) to maximize the likelihood. For a single one- or two-marker data set, the complete-data likelihood is simply a product of binomials and multinomials. For the published and simulated sperm-typing data sets so far analyzed for segregation distortion, the EM algorithm has quickly converged to the global maximum of the likelihood, even from starting points relatively far from the maximum. SPERMSEG allows the user to specify different starting points, if desired, to help determine that a global maximum has been reached.

In maximum-likelihood analysis of sperm-typing data, it is common to have some parameters estimated on the boundary of the parameter space. These are either contamination parameters (β) or probabilities of two sperm deposited (γ_2) that are estimated to be 0. When this occurs, the gradient of the log-likelihood of the data at

the maximum-likelihood estimate is not necessarily 0, and it is not appropriate to estimate the standard errors of the parameter estimates by calculating the Fisher information. Instead, the SPERMSEG software inverts the likelihood-ratio test to obtain confidence intervals for the parameter estimates. Confidence intervals obtained by inverting the likelihood-ratio test are generally more accurate than those obtained from the Fisher information, even when the maximum-likelihood estimate is in the interior of the parameter space.

One can perform a χ^2 goodness-of-fit test to make sure that the model used to analyze the sperm-typing data actually fits the data. However, when some parameters are estimated on the boundary of the parameter space, the appropriate number of df for the χ^2 test is no longer clear. SPERMSEG has a built-in simulation routine to calculate a *P* value, for the goodness-of-fit test, that will be valid even when some parameters are estimated on the boundary.

In order to make full use of single-sperm typing as a valuable tool for the study of segregation distortion, flexible software must be available to analyze the resulting data. SPERMSEG allows for any number of one- and two-marker data sets from one or more donors. It performs full likelihood analysis of the data, using models of the user's choice. Log-likelihoods are output for use in hypothesis testing, and confidence intervals based on inverting the likelihood-ratio test and simulation-based goodness-of-fit tests are calculated, both of which are reliable even when parameters are estimated on the boundary.

Acknowledgments

This work is supported by National Institutes of Health/National Health Genome Research Institute grant 1 R29 HG01645.

MARY SARA MCPEEK

Department of Statistics
University of Chicago
Chicago

Electronic-Database Information

The URL for data in this article is as follows:

SPERMSEG, http://galton.uchicago.edu/~mcpeek/software/spermseg (for SPERMSEG software and documentation)

References

Cui X, Li H, Goradia TM, Lange K, Kazazian HH Jr, Galas D, Arnheim N (1989) Single-sperm typing: determination of genetic distance between the $^{G}\gamma$-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. Proc Natl Acad Sci USA 86:9389–9393

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc [B] 39:1–38

Girardet A, Leeflang EP, McPeek MS, Munier F, Arnheim N, Claustres M, Pellestor F (1998) Analysis of meiotic segregation at the retinoblastoma locus using sperm typing technique. Am J Hum Genet Suppl 63:A362

Goradia TM, Stanton VP Jr, Cui X, Aburatani H, Li H, Lange K, Housman DE, et al (1991) Ordering three DNA polymorphisms on human chromosome 3 by sperm typing. Genomics 10:748–755

Grewal RP, Cancel G, Leeflang EP, Dürr A, McPeek MS, Draghinas D, Xiang Y, et al (1999) French Machado-Joseph disease patients do no exhibit gametic segregation distortion: a sperm typing analysis. Hum Mol Genet 8:1779–1784

Lazzeroni LC, Arnheim N, Schmitt K, Lange K (1994) Multipoint mapping calculations for sperm-typing data. Am J Hum Genet 55:431–436

Leeflang EP, McPeek MS, Arnheim N (1996) Analysis of meiotic segregation, using single-sperm typing: meiotic drive at the myotonic dystrophy locus. Am J Hum Genet 59: 896–904

Takiyama Y, Sakoe K, Amaike M, Soutome M, Ogawa T, Nakano I, Nishizawa M (1999) Single sperm analysis of the CAG repeats in the gene for dentatorubral-pallidoluysian atrophy (DRPLA): the instability of the CAG repeats in the DRPLA gene is prominent among the CAG repeat diseases. Hum Mol Genet 8:453–457

Takiyama Y, Sakoe K, Soutome M, Namekawa M, Ogawa T, Nakano I, Igarashi S, et al (1997) Single sperm analysis of the CAG repeats in the gene for Machado-Joseph disease (MJD1): evidence for non-Mendelian transmission of the MJD1 gene and for the effect of the intragenic CGG/GGG polymorphism on the intergenerational instability. Hum Mol Genet 6:1063–1068

Williams C, Davies D, Williamson R (1993) Segregation of ΔF508 and normal CFTR alleles in human sperm. Hum Mol Genet 2:445–448

Address for correspondence and reprints: Dr. Mary Sara McPeek, Department of Statistics, University of Chicago, 5734 South University Avenue, Chicago, IL 60637. E-mail: mcpeek@galton.uchicago.edu

0002-9297/1999/6504-0029$02.00

Am. J. Hum. Genet. 65:1197–1199, 1999

Cultural Difference and the Eugenics Law

To the Editor:

Mao recently reported results of a survey of Chinese geneticists' views on ethical issues in genetic testing and screening, which are quite different from those of their Western counterparts (Mao 1998). Although this report provides a welcome opportunity to further illuminate the East-West controversy that surrounds the Chinese