Am. J. Hum. Genet. 73:652–655, 2003

Report

A Full-Likelihood Method for the Evaluation of Causality of Sequence Variants from Family Data

Deborah Thompson,[1] Douglas F. Easton,[2] and David E. Goldgar[1]

[1]International Agency for Research on Cancer, Lyon, France; and [2]Strangeways Research Laboratories, University of Cambridge, Cambridge, United Kingdom

In many disease genes, a substantial fraction of all rare variants detected cannot yet be used for genetic counselling because of uncertainty about their association with disease. One approach to the characterization of these unclassified variants is the analysis of patterns of cosegregation with disease in affected carrier families. Petersen et al. previously provided a simplistic Bayesian method for evaluation of causality of such sequence variants. In the present report, we propose a more general method based on the full pedigree likelihood, and we show that the use of this method can provide more accurate and informative assessment of causality than could the previous method. We further show that it is important that the pedigree information be as complete as possible and that the distinction be made between unaffected individuals and those of unknown phenotype.

The identification of specific genes involved in a number of common diseases has brought genetic testing into clinical practice. For many of these genes, the sequence variants that have been identified include known deleterious mutations (often protein truncating), recognized polymorphisms, and rare variants (usually missense changes). The last category poses problems for genetic counselling, since tested individuals and their families are given an uninformative result unless sufficient evidence is available that a given missense change is deleterious. In the case of *BRCA1* and *BRCA2,* these so-called "unclassified variants" account for approximately one-half of all variants (other than common polymorphisms) detected (see, e.g., Frank et al. 2002) and were present in 13% of the women tested in the study by Frank et al. (2002).

Various types of evidence may help to classify such variants as deleterious or neutral with respect to the disease of interest; these include the nature and position of the amino acid substitution, the degree of conservation among species, and the results of functional assays. However, the evidence most relevant to the genetic-counselling problem comes from two main sources. First, there may be epidemiological data on the frequency of the variants in families with different personal and family history of disease as compared with their frequency in the general population. The difficulty is that often the variants of interest are quite rare, precluding assessment in even large case-control series. Second—and more readily available in practice—is evidence obtained through examination of the extent of cosegregation of the variant with disease in families in which a given sequence variant has been identified.

The general design is that in which a single individual from a family (usually affected) has been tested and found to harbor a sequence variant of unknown significance in a particular gene. Subsequently, additional family members are tested for the specific variant. The goal is to evaluate the evidence for the causality associated with the specific variant with reference to the disease of interest. Typically, this involves testing the hypothesis that the variant confers some specified risk (e.g., that conferred by known deleterious mutations) against the hypothesis of complete neutrality (i.e., wherein the variant is not associated with any increased risk of the disease). In principle, one could consider a more general model in which the variant confers some intermediate risk, but, except for relatively common founder variants such as *BRCA1* 185delAG, it is rarely possible to ac-

Received April 18, 2003; accepted for publication June 23, 2003; electronically published July 29, 2003.

Address for correspondence and reprints: Dr. David E. Goldgar, Unit of Genetic Epidemiology, International Agency for Research on Cancer, 150 Cours Albert Thomas, 69008 Lyon, France. E-mail: goldgar@iarc.fr

0002-9297/2003/7303-0019$15.00

cumulate enough data to estimate the average risks associated with a specific variant.

To address this problem, Petersen et al. (1998) developed a method for calculating what they termed "Bayes factors of causality," which are based on the degree of relationship to the proband and genotype/carrier status at the presumed causal variant of each tested relative. In situations in which the penetrance/relative risk of the causal allele is unknown, Petersen et al. (1998) proposed the use of a Bayesian approach, with an assumed prior distribution for the penetrances conferred by the variant in question. Although this approach can be applied to a number of practical problems, their specific formulation suffers from some limitations. Most restrictive is that it does not easily allow for different risks (absolute or relative) for individuals as a function of, for example, their age and sex. Furthermore, tested unaffected individuals are not included in their method. For high-risk alleles, the respective genotypes of older unaffected individuals may contribute information on causality. Last, there is some information to be gained from the phenotypes of individuals who have not been genotyped but whose genotype may be partially inferred from the genotypes of their close relatives. Petersen et al. (1998) provided a simple S-plus program to implement their method in the case of a rare disease allele. As noted by Petersen et al. (1998), this simplification also ignores the dependence between multiple tested individuals in the same pedigree. To overcome these limitations, we propose here a more general formulation that is easily implemented in existing genetic analysis programs and that reduces to the method of Petersen et al. (1998) in those cases in which the latter can be used.

We note that the relevant likelihood ratio (or Bayes factor) is of the form

$$B = \frac{L(\mathbf{V}|\mathbf{P},V_p,C = 1)}{L(\mathbf{V}|\mathbf{P},V_p,C = 0)} .$$

Here, **P** is the vector of disease phenotypes within the family, **V** is the vector of the variant genotypes, and V_p is the variant genotype of the proband. Following Petersen et al. (1998), $C = 1$ when the variant is disease causing, and $C = 0$ when the variant is neutral. By "causality," we mean that a given variant has a penetrance or relative risk of disease (perhaps age and/or sex specific) equal to that of proven deleterious mutations (or some other prespecified value). Since all variants at the disease locus are either deleterious or neutral, these likelihoods can be computed using a standard two-allele model, with a hypothetical susceptibility allele *A* (corresponding to all deleterious alleles), with some frequency *p,* and a normal allele *a* (corresponding to all neutral alleles). The hypothesis $C = 1$ can be modeled by treating the variant as a genetic marker allele that is in complete linkage disequilibrium, *D,* with the proposed susceptibility allele *A* (i.e., $\theta = 0$; $D = 1$). Under the hypothesis $C = 0$, the variant segregates independently of the disease, equivalent to the same model but with $\theta = 1/2$ and $D = 0$. Thus,

$$B = \frac{L(\mathbf{V}|\mathbf{P},V_p,\theta = 0,D = 1)}{L(\mathbf{V}|\mathbf{P},V_p,\theta = 1/2,D = 0)}$$

$$= \frac{L(\mathbf{P},\mathbf{V}|\theta = 0,D = 1)}{L(\mathbf{P},\mathbf{V}|\theta = 1/2,D = 0)} \times \frac{L(\mathbf{P},V_p|\theta = 1/2,D = 0)}{L(\mathbf{P},V_p|\theta = 0,D = 1)} .$$

If information is available from multiple families carrying the same variant, then the overall evidence of causality is the product of the individual Bayes factors over these (independent) families. When assessing the causality of a given variant over multiple families, we implicitly assume that the variant under study is not in linkage disequilibrium with a hidden deleterious mutation in some but not all families being examined.

The first of these likelihood ratios is similar to the antilogarithm of the LOD score for linkage between the variant and disease but with the additional information arising from the fact that the variant is assumed to be in phase with the susceptibility allele. The second ratio is a correction for the fact that the proband is known to carry the variant. These likelihoods can be calculated using standard pedigree-analysis software. If there are no other genes of interest (e.g., modifier loci affecting disease risk in carriers of the variant), then the likelihood ratio can be computed straightforwardly in, for example, Linkage (Lathrop et al. 1984). If necessary, likelihoods based on more-complex models that also incorporate other genes or known environmental risk factors can be computed using, for example, the Pedigree Analysis Package (Hasstedt 2002) or Mendel (Lange et al. 1988). The formulation is essentially exact, providing that the model (in particular, the penetrances) is correctly specified. Thus, the complexity of the genetic model is limited only by the software used for the analysis, with the only restriction being that the software be capable of incorporating linkage disequilibrium. We have chosen to use the Linkage package (Lathrop et al. 1984) for the analyses described below. Sample Linkage parameter files required in order to implement this approach are available from the authors on request.

The Bayes factor can also be used as a test of the hypothesis of causality, by comparing *B* with its distribution under the hypothesis of neutrality. This distribution does not depend on the true penetrance parameters, and the test is therefore valid whether or not the penetrances are correctly specified. The power of this

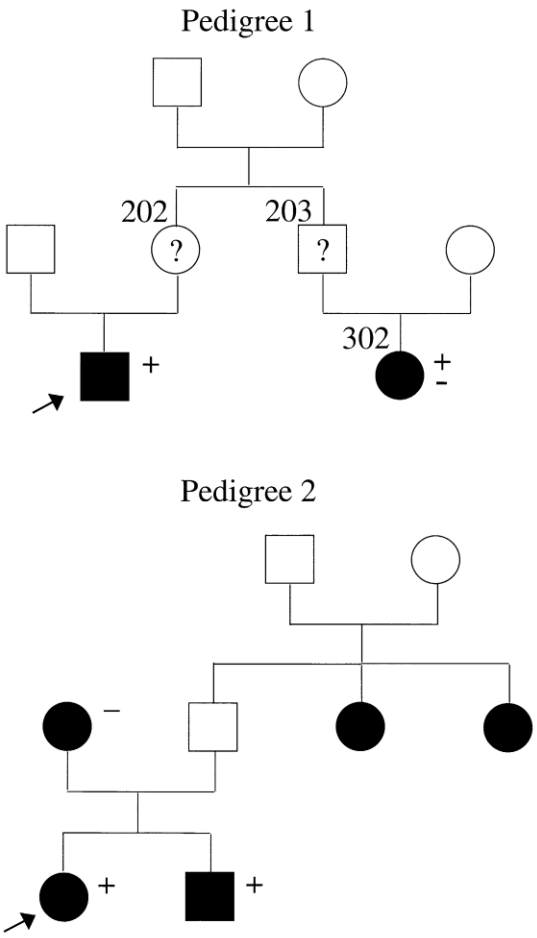

Figure 1 Two sample pedigrees used to demonstrate calculation of odds of causality (shown in table 1). Blackened symbols indicate affected individuals; question marks (?) indicate phenotype changed in table 1; plus (+) and minus (−) symbols indicate carrier status at the sequence variant under assessment. Numbers are given for those individuals specifically referred to in table 1.

test will clearly depend on the accuracy of the penetrance estimates. To illustrate the advantages of our implementation, consider the two pedigrees shown in figure 1. The comparisons between the rare-allele method outlined by Petersen et al. (1998) and our proposed implementation are shown in table 1, assuming a genetic model with allele frequency 0.001 and penetrances of 0.7 in carriers of the variant and 0.05 in noncarriers. Several interesting features of the method can be seen. First, it is reassuring that, when the two parents (202 and 203) are of unknown phenotype, the two methods agree quite closely. Second, we note the rather broad range of Bayes factors obtained using our proposed method as a function of the phenotype of the untested (but likely obligate carriers) parents. Note that if, in the Petersen et al. (1998) rare-allele method, these are assumed to be affected carriers, the evidence of causality is heavily overstated (Bayes factor corresponding to odds of 32.8:1 in favor of causality) because it does not properly take into account the dependence of the tested individuals. In our method, if these individuals were tested positive, then the odds in favor of causality are only slightly higher, 7.4:1. Last, the results for pedigree 2 show the effect of including untested individuals in the analysis, with odds *against* causality of 5:1 if they are not considered, compared with roughly even odds if the two affected untested individuals are included. Somewhat surprisingly, untested unaffected individuals can have an effect as well, if the assumed penetrance is high. Thus, it is important that the pedigree information be as complete as possible and that the distinction be made between unaffected individuals and those of unknown phenotype. In contrast to the method of Petersen et al. (1998), the method described here uses all available genotype information from the family, including any unaffected individuals who have been tested. However, since neither of the above pedigrees contains any tested unaffected members, the difference in results is not due to the inclusion of extra genotypes. In their article, Petersen et al. (1998) analyzed two published data sets; we have reanalyzed one of them by using all the information provided in the pedigrees (Barker et al. 1996) but assuming the same genetic model as Petersen et al. (1998). For the two pedigrees carrying the R841W variant in *BRCA1,* Bayes factors of 1.73 and 2.96 were obtained by Petersen et al. (1998), using the exact method outlined in their article. Our reanalysis of these pedigrees, including the tested unaffected individuals and untested breast cancer cases, results in values of 0.66 and 0.89, respectively. Combining the data from the two pedigrees, Petersen et al. (1998) conclude that the hypothesis of causality (under the specified penetrance model) is five times more likely than noncausality; however, using all

Table 1

Bayes-Factor Comparison of the Two Approaches for Hypothetical Data

Pedigree and Genotype (Individual 302)/Phenotype (Individuals 202 and 203)	Petersen et al. (1998)	Full Likelihood
Pedigree 1:		
+/Affected	5.33 (32.8[a])	7.30 (7.37[a])
+/Unknown	5.33	5.28
+/Unaffected	5.33	1.83
−/Affected	.38	.09
−/Unknown	.38	.38
−/Unaffected	.38	.88
Pedigree 2	.21	1.03

NOTE.—For details of the two approaches, see figure 1.

[a] Value obtained if one assumes that 202 and 203 are carriers of the variant.

the data provided in the pedigrees, our analysis under the same genetic model finds odds of 1.7:1 *against* causality from these data. If the unaffected individuals used in our method but not the Petersen et al. (1998) approach are not considered in the analysis, application of our method results in essentially even odds of causality (Bayes factor of 1.06), compared with the 5:1 found by Petersen et al. (1998). We also repeated the analysis using a model for inherited breast cancer (Easton et al. 1993) that takes into account both the ages at diagnosis of the affected individuals and the ages at last observation of the unaffected individuals (assumed to be similar to their affected sisters). Under this perhaps-more-realistic model, the odds are very slightly in favor of causality (1.2:1). Note that this amino acid residue, R841, shows considerable variation among the various orthologues for which sequence data are available, with two alternative amino acids present (neither of which are the observed change, *W*) and no conservation other than in chimpanzee (Orelli et al. 2001). This analysis, as well as the illustrative examples shown in table 1, point out the risks in using the simplistic approach of Petersen et al. (1998), which in some circumstances could result in spurious classification of unknown sequence variants. In the case of unknown penetrance, Petersen et al. (1998) used a Bayesian approach, integrating over the prior distribution for the penetrance (usually expressed as the risk ratio or hazard ratio between disease risk in carriers as compared with noncarriers). This approach could also easily be implemented in the full-likelihood approach through the use of a shell program, for example.

In summary, we have proposed a simple likelihood-based framework and implementation, to provide more informative and accurate assessment of causality for DNA sequence variants of unknown functional significance. This information can then be used, in conjunction with other lines of evidence, to assess the causality of variants of uncertain clinical significance.

Acknowledgments

We wish to acknowledge the support of grant 1R01 CA-81203 (to D.E.G.) from the National Institutes of Health and Cancer Research U.K. D.F.E. is a Principal Research Fellow of Cancer Research U.K. The work reported was undertaken during the tenure of a Postdoctoral Fellowship (to D.T.) from the International Agency for Research on Cancer.

References

Barker DF, Almeida ER, Casey G, Fain PR, Liao SY, Masunaka I, Noble B, Kurosaki T, Anton-Culver H (1996) *BRCA1* R841W: a strong candidate for a common mutation with moderate phenotype. Genet Epidemiol 13:595–604

Easton DF, Bishop DT, Ford D, Crockford GP, Breast Cancer Linkage Consortium (1993) Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. Am J Hum Genet 52:678–701

Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC (2002) Clinical characteristics of individuals with germline mutations in *BRCA1* and *BRCA2:* analysis of 10,000 individuals. J Clin Oncol 20:1480–1490

Hasstedt SJ (2002) PAP: Pedigree Analysis Package, version 5.0. Department of Human Genetics, University of Utah, Salt Lake City

Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, dGENE. Genet Epidemiol 5: 471–472

Lathrop GM, Lalouel J-M, Julier C, Ott J (1984) Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446

Orelli BJ, Logsdon JMJ Jr, Bishop DK (2001) Nine novel conserved motifs in BRCA1 identified by the chicken orthologue. Oncogene 20:4433–4438

Petersen GM, Parmigiani G, Thomas D (1998) Missense mutations in disease genes: a Bayesian approach to evaluate causality. Am J Hum Genet 62:1516–1524