Missense Mutations in Disease Genes: A Bayesian Approach to Evaluate Causality

Gloria M. Petersen[1]

Giovanni Parmigiani[2]

Duncan Thomas[3]

1. Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, Maryland, 21205.

2. Institute of Statistics and Decision Sciences, Duke University Durham, NC, 27708.

3. Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90033.

Running title: Missense mutations and Bayesian analysis

Correspondence to:

Gloria M. Petersen, Ph.D.

Department of Epidemiology

Johns Hopkins School of Public Health

615 N. Wolfe St.

Baltimore, MD 21205

Office: (410) 955-7497

Fax: (410) 955-0863

email: gpeterse@jhsph.edu

Summary

The problem of interpreting missense mutations of disease-causing genes is an increasingly important one. Because these point mutations result in altering only a single amino acid of the protein product, it is often unclear whether this change alone is sufficient to cause disease. We propose a Bayesian approach that utilizes genetic information on affected relatives in families ascertained through known missense mutation carriers. This method is useful in evaluating known disease genes for common disease phenotypes, such as breast cancer or colorectal cancer. The posterior probability that a missense mutation is disease causing is conditioned on the relationship of the relatives to the proband, the population frequency of the mutation, and the phenocopy rate of the disease. The approach is demonstrated in two cancer datasets: BRCA1 R841W and APC I1307K. In both examples, this method helps establish that these mutations are likely to be disease causing, with Bayes factors in favor of causality of 5.09 and 66.97 respectively, and posterior probabilities of $.836$ and $.985$. We also develop a simple approximation for rare alleles and consider the case of unknown penetrance and allele frequency.

Keywords: Genetic analysis, breast cancer, colon cancer, APC, BRCA1.

1 Introduction

The commitment of the genetics research community to map the human genome and identify disease causing genes has been successful in advancing our knowledge in a number of areas, such as molecular technology leading to more rapid cloning and sequencing of genes (Guyer and Collins, 1995; Savill, 1997; Schuler et al., 1996). These major advances have opened other important areas of investigation, including biochemical mechanisms for disease, and genetic epidemiologic implications of newly discovered genes in patients and populations (Ellsworth et al., 1997).

Of major importance is the interpretation of kinds of mutations that occur in genes and whether these may be involved in causation of disease (Cooper and Krawczak, 1993). It is widely accepted that mutations that result in frameshifts (insertions or deletions) can significantly alter the protein product, as can point mutations that result in splice site alterations or stop codons (nonsense mutations). A problematic type of point mutation is one that results in a substitution of one amino acid for another (missense mutations). These amino acid substitutions can be disease causing if they affect an important functional region of the protein. A well-known example is the substitution of valine for glutamic acid in the β—globin gene at the sixth codon, which results in an abnormal hemoglobin that is less soluble in deoxygenated blood, the basic biochemical defect in sickle cell anemia. Alternatively, an amino acid substitution may occur in a less critical or conserved region of the protein, and be tolerated, such that it results in an isoform of the protein, and genetically is interpreted as an allelic variant or possible polymorphism.

With the identification of disease causing genes, a major challenge to molecular geneticists is detection and characterization of mutations in affected persons (Cotton, 1997). The definitive methodology is DNA sequencing, which will identify point mutations, including missense mutations. The interpretation of a missense mutation is often inconclusive as to whether it is disease-causing. In particular, when the mutation is frequent and carriers exhibit lower penetrance, it poses a greater challenge because one could argue

that it is a polymorphism.

We present a method that utilizes a Bayesian approach in families with multiple affected persons to establish statistically whether a missense mutation in an autosomal dominant gene is disease causing. We apply this method to published data on a missense mutation of the APC gene in familial colon cancer (Laken et al., 1997) and a missense mutation of BRCA1 in hereditary breast cancer (Barker et al., 1996).

2 Methods

Study Design

We propose a study design based on testing affected relatives of probands who are known to be carriers of the missense mutation. This design is efficient and provides meaningful information about the association when there is good *a priori* information about the rate of sporadic disease in the population and about the population frequency of the mutation. If the missense mutation is not disease-causing, we expect to observe a proportion of positive affected relatives that is simply dictated by the degree of relationship between the relatives in the sample and the proband. For example, among first-degree relatives of heterozygous carrier probands of a rare mutation, we would expect about one half to be carriers. If the mutation is disease-causing, we expect to see an increased proportion, beyond one half, depending on the penetrance and the phenocopy rate. Our approach quantifies how many more to expect and weighs the evidence in favor of causality.

In addition to its practical and ethical advantages, a design based on testing only affected relatives can be statistically efficient when reliable information about phenocopy rates is available. Swift *et al* studied a simpler design where a single affected relative per proband is tested, and developed a large sample approximation to the confidence interval on the odds ratio of genotype given disease status. They also compared the statistical efficiency of affected-only designs to that of case-control designs where unaffected relatives are also tested, and concluded that if the disease incidence among non-carriers is known with

reasonable accuracy, testing only affected is substantially more efficient. Although our design is more general, their overall conclusions about efficiency are likely to be applicable here.

Sampling Distributions

We consider an autosomal dominant gene and focus on a single, potentially disease-causing allele (i.e. a missense mutation in a disease-causing gene). We assume that the allele frequency in the population is known, and we denote this by p, with $1 - p = q$. We also assume that the rate of disease among noncarriers (phenocopy rate) and the penetrance among carriers are known. We define the penetrance as $\beta = P\{\text{Disease}|AA\} = P\{\text{Disease}|Aa\}$ and the phenocopy rate as $\varphi = P\{\text{Disease}|aa\}$. We are interested in whether the allele is disease-causing: that is, in whether carriers of the allele are at increased risk. The presence of a causal effect is denoted by the variable C, with $C = 1$ if the disease is caused by the mutation, $C = 2$ if not. So $C = 1$ corresponds to $\beta > \varphi$, while $C = 2$ corresponds to $\beta = \varphi$. The Appendix extends our approach to the case of unknown reduced penetrance and unknown allele frequency, with known rate of disease for the overall population.

To assess the presence of a causal effect, we compute the Bayes factor for $C = 1$ versus $C = 2$ and the *a posteriori* probability that the allele is disease-causing, given the observed mutation test results in one or more pedigrees. If the penetrance, prevalence and phenocopy rate are known, the Bayes factor is simply the ratio of the likelihoods of the observed testing results under the two hypotheses. But using a Bayesian approach, extensions to unknown penetrance, prevalence, and phenocopy rate are straightforward.

For K probands, with genotypes $g_{01}, \ldots, g_{0K}$, where g_{0k} is either AA or Aa, and is fixed by design, we have corresponding n_k affected relatives who are tested for the same mutation. The genotype of relative i of proband k is g_{ik}, and can be AA, aa or Aa. The calculations are also conditioned on the relationship with the proband. To keep the notation concise we refer to the vector of the genotypes for the relatives of proband k as $\mathbf{g}_k = (g_{1k}, \ldots, g_{n_k k})$.

Because of the study design, all probabilities are implicitly conditional on the proband's genotype, on the relative being affected, and on his/her degree of relationship with the proband.

We begin by considering the contribution of a single family in the sample. To evaluate this, we need to be able to evaluate the probabilities $P\{\text{Observed genotypes of relatives}|C\}$ for all possible combinations of observed genotypes and C. If the mutation is not disease-causing ($C = 2$), the conditional probability of observed genotypes of relatives in family k is

$$\gamma_{\textit{non-causal},k} \equiv P\{\mathbf{g}_k|g_{0k}, C = 2\} = P\{\mathbf{g}_k|g_{0k}\}, \tag{1}$$

because affected status does not carry any information about genotype probabilities. $P\{\mathbf{g}_k|g_{0k}\}$ depends on the allele frequency p and also on the degree of relationship with the proband, which is not explicitly incorporated in the notation, but is considered in the calculation, and it can be determined simply based on the degree of relationship with the proband, using standard conditional probability arguments (Elandt-Johnson, 1971). For example, if proband k is Aa and his/her family contains only one other affected relative —a sibling also with genotype Aa, then $n_k = 1$ and

$$P\{g_{1k}|g_{0k}, C = 2\} = P\{Aa|Aa, C = 2\} = \frac{1}{2}(1 + pq).$$

For a heterozygous parent of the proband, we would have

$$P\{g_{1k}|g_{0k}, C = 2\} = P\{Aa|Aa, C = 2\} = \frac{1}{2},$$

for a homozygous sib,

$$P\{g_{1k}|g_{0k}, C = 2\} = P\{AA|Aa, C = 2\} = \frac{1}{4}p(1 + p),$$

and so forth. A general algorithm is presented by Li and Sacks (1954) . The software package LINKAGE (Laboratory of Statistical Genetics, Rockefeller University, 1998) can be used to automate these calculations.

If the mutation is disease-causing ($C = 1$), the expected fraction of carriers among affected relatives is higher than if it is not disease-causing. Conditional on a genotype vector $\mathbf{g}_k$ with n_k^{aa} relatives with aa genotype, and assuming that the disease outcomes of relatives are independent given their genotypes, the probability that all relatives have disease is $\varphi^{n_k^{aa}}\beta^{n_k - n_k^{aa}}$. Using the genotype probabilities $P\{\mathbf{g}_k|g_{0k}\}$ as priors, we can determine genotype probabilities under $C = 1$ via Bayes' rule. For the $C = 1$ case, the probability of the observed genotypes of relatives in family k is

$$\gamma_{\mathit{causal},k} = P\{\mathbf{g}|g_{0k}, C = 1\} = \frac{P\{\mathbf{g}_k|g_{0k}\}\varphi^{n_k^{aa}}\beta^{n_k - n_k^{aa}}}{\sum_{\mathbf{g}} P\{\mathbf{g}|g_{0k}\}\varphi^{n_{\mathbf{g}}^{aa}}\beta^{n_k - n_{\mathbf{g}}^{aa}}}, \tag{2}$$

where $\sum_{\mathbf{g}}$ ranges over all 3^k possible genotype combinations, and $n_{\mathbf{g}}^{aa}$ is the number of aa elements in the vector $\mathbf{g}$. In practice, the number of terms in the summation at the denominator can be reduced by factoring terms that lead to the same value of $n_{\mathbf{g}}^{aa}$. For example, for a proband with one sibling who is Aa, we have

$$\begin{aligned} P\{g_{1k}|g_{0k}, C = 1\} &= P\{Aa|Aa, C = 1\} \\ &= \frac{P\{Aa|Aa\}P\{\text{Disease}|Aa \text{ or } AA\}}{P\{Aa \text{ or } AA|Aa\}P\{\text{Disease}|Aa \text{ or } AA\} + P\{aa|Aa\}P(\text{Disease}|aa\}} \\ &= \frac{\frac{1}{2}(1 + pq)}{1 - \frac{1}{4}q(1 + q) + \frac{1}{4}q(1 + q)\varphi}. \end{aligned}$$

When the disease has a variable age of onset, and age-of-onset data about relatives are available, β and φ should be specified in terms of age-specific survival contributions. This would entail taking each subject's age into account. However, by rewriting expression (2) as

$$\gamma_{\mathit{causal},k} = \frac{P\{\mathbf{g}_k|g_{0k}\}\left(\frac{\varphi}{\beta}\right)^{n_k^{aa}}}{\sum_{\mathbf{g}} P\{\mathbf{g}|g_{0k}\}\left(\frac{\varphi}{\beta}\right)^{n_{\mathbf{g}}^{aa}}}, \tag{3}$$

it can be seen that $\gamma_{\mathit{causal},k}$ depends on β and φ only via their ratio, and therefore is valid as long as the ratio of the penetrance to the phenocopy rate remains constant with respect to age. It is possible to carry out an age-dependent analysis along the lines of the approach described here. However, this requires either a large number of observations or good a priori knowledge of the penetrance function for the mutation under consideration. These were not available in the applications considered in this paper.

Probability of a disease-causing effect

We can now evaluate the probability of a causal effect conditional on the observed mutation tests in the affected relatives (Data). Because the families can be considered independent, in our case

$$\begin{aligned} P\{\text{Data}|C=1\} &= \prod_{k=1}^{K} \gamma_{\textit{causal},k} \\ P\{\text{Data}|C=2\} &= \prod_{k=1}^{K} \gamma_{\textit{non-causal},k} \end{aligned}$$

so that the Bayes factor in favor of the hypothesis of causality is

$$B = \frac{\prod_{k=1}^{K} \gamma_{\textit{causal},k}}{\prod_{k=1}^{K} \gamma_{\textit{non-causal},k}}. \tag{4}$$

Using expressions (1) and (2) and algebraic manipulation we can express the Bayes Factor simply in terms of the genotype coefficients and the penetrance-to-phenocopy-rate ratio:

$$B^{-1} = \prod_{k=1}^{K} \sum_{\mathbf{g}} P\{\mathbf{g}|g_{0k}\} \left(\frac{\varphi}{\beta}\right)^{n_{\mathbf{g}}^{aa} - n_{k}^{aa}}. \tag{5}$$

This incorporates the dependence among relatives of the same proband and the possibility that there is more than one copy of the mutated allele in the same family.

Using Bayes rule, which states

$$P\{C=1|\text{Data}\} = \frac{P\{C=1\}P\{\text{Data}|C=1\}}{P\{C=1\}P\{\text{Data}|C=1\} + P\{C=2\}P\{\text{Data}|C=2\}}, \tag{6}$$

and denoting by $O = P\{C=1\}/P\{C=2\}$ the *a priori* odds in favor of causality, the posterior probability of causality is

$$P\{C=1|\text{Data}\} = \frac{OB}{OB+1}. \tag{7}$$

The choice of *a priori* odds can be important and must reflect the context of the analysis. If the mutation was selected because it resides on a known disease-causing gene, the prior odds may be high, while when screening for disease causing mutation, a low, population based, prior probability of causality may be more appropriate. In any case, analysis of the Bayes factor gives a measure of the weight of evidence of the data in favor of the hypothesis

that does not require specifying prior odds. Kass and Raftery (1995) give an in-depth discussion of interpretation and calibration of Bayes factors. A convenient option is assuming $P\{C = 1\} = 1/2$, which gives $O = 1$, that is even *a priori* odds to the hypothesis of causality.

3 Results

Familial Breast Cancer and BRCA1 R841W

We investigate the causality of the BRCA1 R841W mutation using two pedigrees from Barker et al., 1996. The disease phenotype of interest is either breast or ovarian cancer. There are $K = 2$ probands who have tested affected relatives. Proband 1160 has one heterozygous sibling affected with breast cancer, $\mathbf{g}_1 = g_{11} = Aa$. Proband 728 has two heterozygous siblings affected with breast cancer. Thus $\mathbf{g}_2 = (g_{12}, g_{22}) = (Aa, Aa)$. We assume a phenocopy rate $\varphi = .125$, a penetrance $\beta = .85$,indicated as plausible for other BRCA1 mutations (Ford and Easton, 1995), and an allele frequency of $p = 1/100$. Evidence to estimate age-specific penetrance for this specific mutation is insufficient. Our calculations are based on assuming that the ratio β/φ of penetrance for carriers and noncarriers does not depend on age. We will later assess sensitivity to p and β/φ. Not all affected family members are tested; we assume that the reason why they were not tested is unrelated to their genotype.

Using expressions (1) and (2)

$$
\begin{aligned}
\gamma_{\textit{non-causal},1} &= \frac{1}{2}(1+pq) = 0.505 \\
\gamma_{\textit{non-causal},2} &= \frac{1}{4}(5pq + p^2 + q^2) = 0.257 \\
\gamma_{\textit{causal},1} &= \frac{\gamma_{\textit{non-causal},1}}{1 - \frac{1}{4}q(1+q)(1-\varphi)} = 0.871 \\
\gamma_{\textit{causal},2} &= \frac{\gamma_{\textit{non-causal},2}}{\frac{1}{8}q(1+q)\varphi^2 + p^2 + \frac{1}{4}q^2 + \frac{25}{16}pq + (1 - \frac{1}{8}q(1+q) - p^2 - \frac{1}{4}q^2 - \frac{25}{16}pq)\varphi} \\
&= 0.761
\end{aligned}
$$

The evidence from family 1 is about 1.73 times more likely under the hypothesis of

causality than not. The evidence from family 2 is 2.96 times more likely under the hypothesis of causality than not. This results in a Bayes factor of 5.09, indicating that test results from the two families are about five times more likely under the hypothesis of causality than they are under the hypothesis of no causality. If prior odds for causality are even, the posterior probability of causality is $.836$.

Our results are virtually unchanged as p varies over the range from 0 to $.01$ and decreases to $.82$ if p is set to $.05$. The rare allele approximation discussed in the Appendix gives virtually the same results. The results are only mildly sensitive to assumptions about penetrance. At a penetrance of 1, the probability of causality of $.85$. At a penetrance of $.5$ it is $.8$.

Familial Colon Cancer and APC I1307K

Mutations of the APC gene can be associated with increased risk of colorectal cancer and colorectal adenomas. We investigate the causality of mutation T to A at APC nucleotide 3920, using 8 pedigrees reported by Laken and colleagues (1997). We used a disease rate α of $.2$, based on clinical judgment, and an allele frequency of the mutation p of $.036$, based on Table 1 in Laken (1997); we estimated allele frequencies separately in the disease and control group and then combined using the postulated value of α. Expressions (1) and (2) were computed for each proband. The results are summarized in Table 1.

The resulting Bayes factor is 66.97 —strong evidence in favor of causality. The posterior probability under even prior odds is 0.985. The rare allele approximation discussed in the Appendix produces a Bayes factor of 194.90, leading to a posterior probability 0.995. Because the mutation is relatively common, the rare allele approximation does not work as well as in the breast cancer example. Also, using expression (12) leads to ignoring the dependence between proband's 8 sibling and her offspring and is therefore not appropriate in this case.

Following the development discussed in Appendix 1, we also conducted with unknown penetrance and allele frequency. We assumed that causality ($C = 1$) corresponds to $\beta > \varphi$ and we assume that if the mutation is indeed causal, then all values of β are equally likely *a*

priori. Regarding p, we used information from Table 1 in Laken et al (1997) to specify a prior on the fraction of carriers and then converted that into a distribution for p. Because the fraction of diseased individuals in the sample approximates the overall population fraction, we combined diseased and non-diseased individuals in specifying our prior. The resulting specification is $p = 1 - \sqrt{1-r}$, where r has a Beta distribution with parameters $47 + 22 = 69$ and $766 + 211 - 47 - 22 = 908$.

To analyze the effect of uncertainty about β and p on the Bayes factor and posterior probability of $C = 1$, we first computed the *a posteriori* probability distribution $p(\beta, p|\text{Data}, C = 1)$. This inference is conditional on the assumption of uninformative ascertainment, which may be violated in this data set. While inference on the penetrance and prevalence is not the focus of the methodology proposed here, incorporating uncertainty via $p(\beta, p|\text{Data}, C = 1)$ is an effective strategy to make inferences about causality without relying on strong assumptions on β and p. To illustrate the range of plausible values of β, Figure 1 graphs the *a posteriori* probability distribution of β conditional on $p = .035$. For the purpose of testing for causality, the important aspect of this distribution is that it assigns very low probability to values close to the phenocopy rate and high probability to values far away from it. So even though the actual magnitude of β is not accurately estimated, reliable conclusions about causality can be reached. We computed the Bayes factor and posterior probability of $C = 1$ using expression (11). The Bayes factor is 53.21 and the resulting posterior probability is $.982$. Even in the presence of substantial uncertainty about the exact value of the prevalence parameter, the evidence in favor of causality in this example remains strong. Results are somewhat sensitive to the specification of the phenocopy rate α, but the evidence in favor of causality is not questioned within a broad range of values. At $\alpha = .1$ the posterior probability is $.987$, while at $\alpha = .3$ it is $.965$.

4 Discussion

We propose an approach using Bayesian methods to statistically determine whether a missense mutation in an autosomal dominant disease gene is likely to be disease-causing. Our method is particularly useful in evaluating common disease phenotypes, only a small proportion of which may be due to a known disease causing gene, such as hereditary breast cancer or familial colorectal cancer. This approach employs the testing of other affected relatives for the missense mutation in pedigrees which have been identified through a proband with the missense mutation. The posterior probability that the mutation is disease causing is conditioned on the relationship of the relative to the proband, the frequency of the mutation, and the phenocopy rate of the disease. Because only affected relatives are selected for genetic analysis, this method is efficient, from the perspective of conducting a study. Because unaffected relatives are not studied, this method avoids the pitfalls that relate to recruitment of such persons for research in the current genetic testing environment (Hubbard and Lewontin, 1996; Geller et al., 1997; Holtzman et al., 1997).

However by limiting analysis only to families with at least two affected relatives, there is a potential for biasing the conclusion in favor of causality, as the penetrance may in reality be lower than the sample indicates. This can be partially addressed by varying penetrance estimates in the computations. This design element also may affect our results if there is a sizable proportion of families that have only one affected person (proband). If the unaffected phenotypes in these families are due to reduced penetrance, the magnitude of the effect for specific mutations may only be empirically assessed by genotyping unaffected relatives. This design modification, of course, engenders logistical issues in executing the study.

Another important consideration is the potential bias that may occur in the selection of affected relatives for genotyping. Our approach assumes that inclusion in the study is unrelated to genotype. Ideally, any and all affected biological relatives should be genotyped, but availability of DNA can be affected by numerous factors (relatives who are deceased, unable or unwilling to participate). If availability for study is non-random for whatever

reasons, this may have an effect on the posterior probability, but the effect would depend upon the nature of the selection bias.

We have shown that this method can be usefully applied to empiric data. We analyzed families identified through probands with the APC I1307K mutation (Laken et al., 1997), in which 10/11 affected relatives tested positive for the same mutation, including three second degree relatives. We found a $.985$ posterior probability that the APC I1307K mutation is causally related to the colorectal neoplasia in these families. In the case of BRCA1 R841W (Barker et al., 1996), there were fewer families and persons studied, but we estimated a $.836$ posterior probability that this missense mutation may be causally related to breast cancer or ovarian cancer in these families.

These two examples contrast another challenge in interpreting missense mutations: plausible mechanism of disease causation engendered by the mutation (Ellsworth et al., 1997; Fearon, 1997). In the case of APC I1307K, the T to A transversion in the APC gene results in an $(A)^8$ tract, which appears to engender an inherent instability of the gene, allowing deleterious mutations to occur in this critical gene during subsequent cell division (Laken et al., 1997). In the case of BRCA1 R841W, however, it is difficult to posit a plausible mechanism because the function of the gene remains unclear. Our proposed method may provide additional evidence to support a disease-causing missense mutation in such cases. That a missense mutation may have a plausible mechanism for causing disease can justify a more realistic estimate of the penetrance (whether higher or lower), while a missense mutation in a gene of unknown function may conservatively be justified to be lower.

There is a possibility, though remote, that the missense mutation may be in linkage disequilibrium with a "true" deleterious mutation elsewhere in the allele. When considering the way in which many missense mutations are identified through DNA sequencing, such a possibility would have been uncovered. In the case of our colon cancer example, Laken et al. (1997) reported that the APC I1307K bearing alleles in two carriers were fully sequenced and no other mutations were detected.

This approach is predicated on the assumption that a single missense mutation has been singled out for particular attention on *a priori* grounds. Since disease genes may potentially be highly polymorphic, additional complications arise in interpreting the causality of polymorphisms that may come to attention because of their frequency in multiple case families, as such polymorphisms are more likely than randomly selected polymorphisms to appear to be causal, even if C=2 were true. A hierarchical Bayesian treatment of this problem might entail simultaneous consideration of all known polymorphisms, with additional parameters for variation in β and p between polymorphisms. The model could include "prior covariates," such as the position in the gene or the nature of the particular amino acid substitution that results, whose effects on penetrance are to be estimated empirically. Such an analysis is beyond the scope of this article, but similar methods have been applied in the treatment of multiple exposures problems in epidemiology (Greenland, 1993) and gene-environment interactions (Aragaki et al., 1997).

The design approach of studying affected relatives of a proband was employed by Swift et al. (1990) and applied by Athma et al. (1996) in a study of breast cancer risk in ataxia telangiectasia heterozygotes. This approach differs from the method proposed here in a number of elements. Their design requires ascertainment of one affected relative of an index carrier (who may not be affected with the disease of interest) per family. Inference and testing are based on a large sample approximation to the confidence interval on the odds ratio of genotype given disease status. An important advantage of the Bayesian approach is that it is simple to derive inferences without having to rely on large sample approximations that could be inaccurate, and it is straightforward to incorporate uncertainty about unknown additional parameters, as discussed in the Appendix.

Finally, we have developed a simple program in S-Plus, shown in the Appendix, which provides computation of the Bayes factor and posterior probabilities in the rare allele case. This basic approach can be extended to consider reduced penetrance with variable age at onset and errors in genetic testing. Development of these extensions is currently in progress.

5 Conclusions

The approach proposed here anticipates one of the outgrowths of the Human Genome Project and genetic disease research: a number of families that segregate common disease phenotypes may be identified to carry missense mutations of known disease causing genes. A concomitant problem in interpreting the significance of the missense mutations is determining whether they are disease causing or simple polymorphisms. We have developed a logistically efficient and computationally feasible approach that may more quickly help determine the importance of such mutations.

Acknowledgments

The authors thank Stephen Gruber, Steve Laken, Bert Vogelstein, Kenneth Kinzler, Frank Giardiello, Stan Hamilton, and Susan Booker for their contributions to this study. This research was supported in part by NIH National Cancer Institute grants CA 52862, R01 CA 63721, P50 CA 62924 (SPORE in Gastrointestinal Cancer, Johns Hopkins University), and P50 CA68438 (SPORE in Breast Cancer at Duke University), and the Clayton Fund. Work carried out while Giovanni Parmigiani was visiting the Department of Biostatistics, Johns Hopkins University, whose warm hospitality is gratefully acknowledged.

References

Aragaki C, Greenland S, Probst-Hensch N, Haile RW (1997) Hierarchical modeling of gene-environment interactions: estimating nat2* genotype-specific dietary effects on adenomatous polyps. Cancer Epidemiol Biomarkers & Prevention 6:307–314

Athma P, Rappaport R, Swift M (1996) Molecular genotyping shows that ataxia-telangiectasia heterozygotes are predisposed to breast cancer. Cancer Genetics & Cytogenetics 92:130–4

Barker DF, Almeida ERA, Casey G, Fain PR, Liao SY, Masunaka I, Noble B, et al (1996) BRCA1 R841W: A strong candidate for a common mutation with moderate phenotype. Genetic Epidemiol 13:595–604

Berger JO, Delampady M (1987) Testing precise hypotheses. Statistical Science 2:317–335

Cooper DN, Krawczak M (1993) *Human Gene Mutation*. Oxford: BIOS Scientific Publishers

Cotton RGH (1997) *Mutation Detection*. New York: Oxford University Press

Elandt-Johnson RC (1971) *Probability Models and Statistical Methods in Genetics*. John Wiley & Sons

Ellsworth DL, Hallman DM, Boerwinkle G (1997) Impact of the Human Genome Project on epidemiologic research. Epidemiologic Reviews 19:3–13

Fearon ER (1997) Human cancer syndromes: clues to the origin and nature of cancer. Science 278:1043–1050

Ford D, Easton DF (1995) The genetics of breast and ovarian cancer. Br J Cancer 72:805–812

Geller G, Botkin JR, Green MJ, Press N, Biesecker BB, Wilfond B, Grana G, et al (1997) Genetic testing for susceptibility to adult-onset cancer. the process and content of informed consent. JAMA 277:1467–74

Greenland S (1993) Methods for epidemiologic analysis of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. Stat Med 12:717–736

Guyer MS, Collins FS (1995) How is the Human Genome Project doing, and what have we learned so far? Proc Nat Acad Sci USA 92:10841–8

Holtzman NA, Murphy PD, Watson MS, A BP (1997) Predictive genetic testing: from basic research to clinical practice. Science 278:602–5

Hubbard R, Lewontin RC (1996) Pitfalls of genetic testing. New Engl J Med 334:1192–94

Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90:773–795

Laboratory of Statistical Genetics, Rockefeller University (1998) Genetic linkage analysis. `http://linkage.rockefeller.edu`

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, et al (1997) Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nature Genetics 17:79–83

Li CC, Sacks L (1954) The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10:347–360

Savill J (1997) Molecular genetic approaches to understanding disease. BMJ 314:126–9

Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, et al. (1996) A gene map of the human genome. Science 274:540–6

Swift M, Kupper LL, Chase CL (1990) Effective testing of gene-disease associations. American Journal of Human Genetics 47:266–274

Appendix

Unknown Penetrance and Prevalence

The derivation of the posterior probability of a causal effect can be extended to the case of unknown penetrance and prevalence parameters. Within a Bayesian approach, this is carried out by replacing the fixed values of β and p with prior probability distributions reflecting information from prior studies or other biological evidence. The prior distribution on the unknown β and p is denoted here by $\pi(\beta, p)$ If the penetrance of similar mutation has been previously studied, published data can be used to specify an informative prior about β. Otherwise, an attractive option is to assume *a priori* that all values larger than φ are equally likely, leading to $\pi(\beta) = 1/(1-\varphi)$, $\varphi < \beta \leq 1$. Because the mutation may be responsible for a nonnegligible fraction of the overall cases, the assumption that φ is known also needs to be relaxed. Here we assume that the overall incidence α is known. The overall incidence depends on φ, β and p via the relationship: $\alpha = q^2\varphi + (1-q^2)\beta$, leading to

$$\varphi = \frac{\alpha - (1-q^2)\beta}{q^2}. \tag{8}$$

As in Section 2.2, under $C = 1$ it is natural to assume that $\beta > \varphi$. With this specification, then, we are testing a point null hypothesis against a composite alternative hypothesis. Berger and Delampady (1987) review the Bayesian approach to this problem and compare it to frequentist approaches . For a fixed α, a prior distribution on β and p implies a prior distribution on φ. Also $\varphi < \beta$ whenever $\alpha < \beta$.

Under these assumptions, expression (1) remains the same, while expression (2) can be now interpreted as conditional on β and p, that is:

$$\gamma_{causal,k}(\beta, p) = P\{\mathbf{g}|g_{0k}, C = 1, \beta, p\} = \frac{P\{\mathbf{g}_k|g_{0k}\}\beta^{n_k - n_k^{aa}}\varphi^{n_k^{aa}}}{\sum_{\mathbf{g}} P\{\mathbf{g}|g_{0k}\}\beta^{n_k - n_{\mathbf{g}}^{aa}}\varphi^{n_{\mathbf{g}}^{aa}}}; \tag{9}$$

rewriting φ in terms of α, β and p:

$$\gamma_{causal,k}(\beta, p) = \frac{P\{\mathbf{g}_k|g_{0k}\}\left(\frac{\alpha-(1-q^2)\beta}{q^2\beta}\right)^{n_k^{aa}}}{\sum_{\mathbf{g}} P\{\mathbf{g}|g_{0k}\}\left(\frac{\alpha-(1-q^2)\beta}{q^2\beta}\right)^{n_{\mathbf{g}}^{aa}}}, \tag{10}$$

so that $P\{\text{Data}|C=1\} = \prod_{k=1}^{K} \gamma_{\textit{causal},k}(\beta, p)$.

The Bayes factor for the hypothesis of causality is now

$$B = \frac{\int_0^1 \int_\alpha^1 \prod_{k=1}^{K} \gamma_{\textit{causal},k}(\beta, p)\pi(\beta, p)d\beta dp}{\prod_{k=1}^{K} \gamma_{\textit{non-causal},k}}, \tag{11}$$

and the posterior probability incorporating uncertainty about β and p can be derived in the usual way as $P\{C=1|\text{Data}\} = OB/(1+OB)$, where O is the prior odds in favor of $C=1$. Expression (11) is easily evaluated numerically using a Monte Carlo approach. To evaluate the numerator, it is sufficient to generate a sample for the prior distribution $\pi(\beta, p)$ and compute the average of the resulting values of $\prod_{k=1}^{K} \gamma_{\textit{causal},k}(\beta, p)$.

Expression (9) can be used to derive an *a posteriori* distribution on β and p, conditional on $C=1$:

$$p(\beta, p|\text{Data}, C=1) = \frac{\prod_{k=1}^{K} \gamma_{\textit{causal},k}(\beta, p)\pi(\beta, p)}{\int_0^1 \int_\alpha^1 \prod_{k=1}^{K} \gamma_{\textit{causal},k}(\beta)\pi(\beta, p)d\beta dp}.$$

This provides a valid inference on the penetrance if it can be assumed that the selection mechanism, that is choosing families with at least one affected relative and eliminating index cases, is not inducing a bias in the estimation of penetrance.

Rare Alleles

When the allele is rare, and there are no homozygous carriers in the sample, ignoring the dependence among relatives of the same proband and ignoring the possibility that there is more than one copy of the mutated allele in the same family have a limited impact on the final answer. Expressions (1) and (2) can then be simplified considerably. It is interesting to study the form of these simplified expressions. All the $N = \sum n_k$ relatives can be considered to be approximately independent. For relative j, Let r_j be the degree of relationship with the proband and h_j be an indicator variable of whether relative j is a carrier or not. Then

$$P\{\text{Data}|C=2\} \approx \prod_{j=1}^{N} \left(\frac{1}{2}\right)^{r_j h_j} \left[1 - \left(\frac{1}{2}\right)^{r_j}\right]^{1-h_j} = \prod_{j=1}^{N} \frac{(2^{r_j}-1)^{1-h_j}}{2^{r_j}}.$$

Applying Bayes' rule for each individual, and using independence, we also obtain:

$$P\{\text{Data}|C=1\} \approx \prod_{j=1}^{N} \frac{\left(\frac{1}{2}\right)^{r_j h_j} \beta^{h_j} \left[\left(1-\left(\frac{1}{2}\right)^{r_j}\right)\varphi\right]^{1-h_j}}{\left(\frac{1}{2}\right)^{r_j}\beta + \left(1-\left(\frac{1}{2}\right)^{r_j}\right)\varphi} = \prod_{j=1}^{N} \frac{\beta^{h_j}[(2^{r_j}-1)\varphi]^{1-h_j}}{1+(2^{r_j}-1)\varphi}.$$

After some simple manipulations, these lead to a Bayes factor in favor of a causal effect of

$$B = \prod_{j=1}^{N} \frac{\left(\frac{\beta}{\varphi}\right)^{h_j} 2^{r_j}}{\left(\frac{\beta}{\varphi}\right) + 2^{r_j} - 1}. \tag{12}$$

Each of the N terms is the ratio of the probability of the evidence under $C=2$ to the probability of the evidence under $C=1$, or weight of evidence against $C=2$. For example, for a first-degree relative the ratio is $2\beta/(\beta+\varphi)$, which is greater than one, if a carrier, and is $2\varphi/(\beta+\varphi)$, which is smaller than one, if not a carrier.

In the unknown penetrance case, the expression for the Bayes factor becomes:

$$B = \int_{\alpha}^{1} \prod_{j=1}^{N} \frac{\left(\frac{\beta}{\varphi}\right)^{h_j} 2^{r_j}}{\left(\frac{\beta}{\varphi}\right) + 2^{r_j} - 1} \pi(\beta) d\beta. \tag{13}$$

A function evaluating this expression using a Monte Carlo algorithm is provided below.
In the rare allele case it is simple to develop general-purpose functions for computing the Bayes factors and the posterior probability of causality. Here we present a function, written in the statistical package S-plus, that handles both the known and unknown penetrance case. This function is available on the Internet site `www.isds.duke.edu/~gp`.

The input to the function are: the vector `relations`, with as many elements as there are relatives, and each element a 1,2, etc, for first-, second-degree relatives etc; the vector `genotypes`, again with one element for each relative, either 0 if the relative is aa or 1 if the relative is Aa —use of the rare allele approximation is not recommended when there are homozygous relatives in the sample; the scalar `prevalence`, for the value of the prevalence if known; the Boolean variable `unknown.prevalence`, which can be set to T or F depending on whether the penetrance is known. If `unknown.prevalence` is set to T, the input value of `prevalence` is ignored; the integer `MonteCarlo`, specifying the number of Monte Carlo sample desired in evaluation the Bayes factor when the penetrance is unknown.

Tables

Proband	*Affected Relatives*	*Genotypes of relatives*	$\gamma_{causal,k}$	$\gamma_{non\text{-}causal,k}$	$\frac{\gamma_{causal,k}}{\gamma_{non\text{-}causal,k}}$
1	Two siblings, one niece from a third sibling	Aa, Aa, Aa	0.42	0.08	5.02
2	Sibling	Aa	0.81	0.52	1.56
3	Mother	Aa	0.79	0.50	1.58
4	Sibling	aa	0.18	0.47	0.38
5	Sibling	Aa	0.81	0.52	1.56
6	Grandmother	Aa	0.61	0.28	2.15
7	Mother	Aa	0.79	0.50	1.58
8	One sibling with an offspring	Aa, Aa	0.70	0.26	2.72

Table 1: Data for the colon cancer example. Summary of genetic testing results and contributions of each family to the likelihood ratio. The rightmost column represents the contribution of each family to the calculation of the Bayes factor, which is the product of the values in the column.

Figure Legends

Figure 1. Posterior distribution of the penetrance parameter β.

Figure 2 S-plus function for computing the Bayes Factor and posterior probability of a disease causing effect in the rare allele case. This function is available for downloading at Internet site `www.isds.duke.edu/~gp`

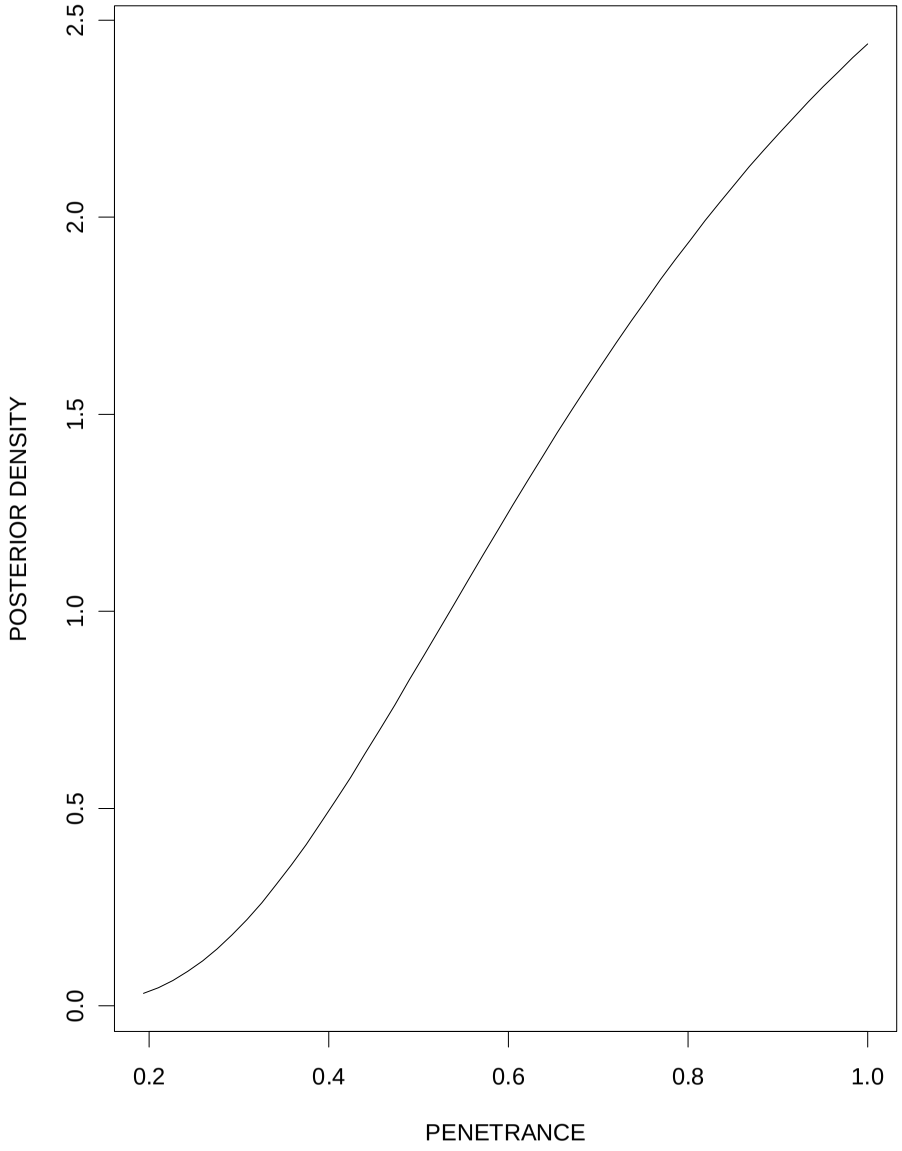

Figure 1:

```
arcs.rare _ function(relations,genotypes,phenocopy.rate,
                     prevalence=1,unknown.prevalence=F,
                     MonteCarlo=1000) {

  rr _ relations
  gg _ genotypes
  phi _ phenocopy.rate
  beta _ prevalence
  M _ MonteCarlo
  NN _ length(relations)
  integrand _ NULL

  if(!unknown.prevalence) {
    bf _ exp ( sum(rr*log(2) +
                   gg*log(beta/phi) -
                   log( beta/phi + 2^rr - 1 ) ) )
  }

  if(unknown.prevalence) {
    beta _ runif(M,phi,1)
    for (m in 1:M) {
      integrand[m] _ exp ( sum (rr*log(2) +
                                gg*log(beta[m]/phi) -
                                log( beta[m]/phi + 2^rr - 1 ) ) )
    }
    bf _ mean(integrand)
  }

  post _ bf / ( bf + 1 )
  return(bf,post)
}
```

Figure 2: