Editorial

Genetic Ancestry and Molecular Epidemiology

Thomas A. Sellers

It is hard to believe that as recently as a decade ago, some of us were finding it virtually impossible to get an epidemiologic study funded from the National Cancer Institute if it required collection and analysis of DNA. Now, it is virtually impossible to get anything funded if you do not! The rapid transformation of epidemiology into the molecular world has been so swift and dramatic that by some metrics, we are still reeling from the impact and struggling to adjust. Having served as an Associate Editor for the journal for roughly 4 years and as Senior Editor for 1 year, I have seen this played out repeatedly in manuscripts assigned to me for review. Given this opportunity to write an editorial of my choosing, I decided to share some perspectives on where we often tend to fall short.

Let me preface my comments with an acknowledgement of my biases. Although my doctorate is in epidemiology, I have post-doctoral training in statistical genetics under the tutelage of Robert C. Elston. I have also been involved in genetic epidemiology long before it became fashionable to do so—back before a single major cancer susceptibility gene had been cloned. In fact, I remember well the signal important discovery of inherited mutations in the retinoblastoma gene (1) and therefore the proof-of-principle of Knudson's two-hit hypothesis (2). Thus, my bias is about the importance of an understanding of genetics and the fundamental laws of inheritance that is an essential prerequisite to molecular-genetic epidemiology. Consequently, I would like to share some simple suggestions designed to help improve the quality and relevance of our work.

1. *Genetic models*—Aside from rare genetic disorders and translocations, we inherit pairs of chromosomes: one from mother, one from father. Therefore, at a given locus, there are two copies of a gene. When there is more than one allele of a gene, it is said to be polymorphic. The effect of genetic variation at that locus on the phenotype may depend on the number of variant alleles (0, 1, or 2). For rare alleles, the frequency of homozygous carriers is likely to be low, but for alleles with a frequency of, say 0.20, then 4% of the control population may be homozygous for the variant—and possibly a much higher proportion of the cases. Thus, all genetic models (dominant, recessive, additive, co-dominant) should be explored rather than just assuming that carrying a copy of the putative high-risk allele (a ''dominant'' genetic model) is all that matters. Indeed, there are thousands of genetic disorders where inheritance of two copies (a ''recessive'' genetic model) is necessary to affect the phenotype.
2. *Hardy-Weinberg equilibrium (HWE)*—Although it is really just a special case of the binomial theorem for $n = 2$, this became one of the foundations of population genetics in 1908 when G.H. Hardy, a British mathematician, and W. Weinberg, a German physician, independently pointed out its utility. If there are only two alleles in the population, p and q, then the sum of the frequencies of the alleles is 1 ($p + q = 1$). Because humans are diploid, the binomial expansion $(p + q)^2 = p^2 + 2pq + q^2$ can be used to calculate the frequency of the three genotypes in the population. It is straightforward to estimate allele frequencies and then generate the expected distribution of genotypes in the population. A comparison of observed *versus* expected genotypes follows a χ^2 distribution, thus allowing inspection of whether or not the population is in HWE. There are two assumptions underlying HWE: random mating and absence of selection. If the observed distribution of genotypes differs from expectation, further investigation is warranted. Absence of HWE may signify population stratification (the population contains a mixture of subpopulations that differ in their allele frequencies), non-random mating, *in utero* selection, sampling error, or genotyping error. Stratification could be detected by estimation of allele frequencies by race and ethnicity and handled through a stratified analysis. Genotyping error could imply either the existence of a pseudogene or lack of specificity in the primers created to amplify the DNA region of interest. This may require that genotyping be repeated.
3. *Genetic effects*—It is generally held that if there are two forms of a disease, one genetic, the other not, the genetic form of the disease will tend to have an earlier age at onset. Some of the true success stories in gene mapping have capitalized on this axiom. Indeed, many of the statistical genetic analysis programs are predicated on the existence of a different age at onset distribution for each genotype. The custom among epidemiologists is to *control* age with every tool at our disposal. We restrict on age range when defining the sampling frame, match on age in the design phase, or adjust for it in the analysis phase. This mindset could very well make it virtually impossible to detect the effect of a genetic polymorphism should it exist. Old tools are being applied in new ways to remedy this situation, including Accelerated Failure Time models (3). Very simple approaches include comparison of mean age at onset

Received 2/16/04; accepted 2/16/04.

Requests for reprints: Thomas A. Sellers, Cancer Prevention and Control, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612-9497. Phone: (813) 632-1315; Fax: (813) 632-1334. E-mail: SellerTA@moffitt.usf.edu

for cases carrying high-risk alleles to those who do not or estimation of allele frequencies by age categories. We should be more receptive to how potential genetic influences may manifest on age at onset.

4. *Association or linkage disequilibrium?*—Epidemiologists are usually trained to be very careful to not confuse cause and association because of the virtually infinite number of ways that observational studies can get the wrong answer. We do not seem to be quite so careful when genes are the exposures of interest, forgetting in our interpretation of the data that the polymorphism that has been genotyped may be the culprit, or may merely be inherited along with the causative variation either within the gene targeted for study or in another gene close by on the same chromosome (linkage disequilibrium). This is especially challenging in the absence of information on whether a particular allele has lower functional activity or decreased level of expression (when the genetic variant resides in the regulatory non-coding region). We should be cautious in our interpretation of study results in light of this.
5. *Gene x environment or false positive association?*—Geneticists are interested in finding genes and/or determining whether variation within specific genes causes disease. It is probably not an overstatement that molecular epidemiologists are primarily interested in how genetic factors modulate the response to environmental factors. This focus makes it somewhat difficult to reconcile the common practice of looking only for ''main effects'' of genes and designing them with adequate power for such (especially if the proper genetic model is recessive rather than dominant). An additional problem comes from well-intended efforts to estimate measures of risk within strata defined by risk factor levels. The power is greatly reduced and observed positive findings are almost always false. Wacholder presented a False Positive Report Probability (FPRP) at the Key Biscayne conference in September of 2003 to help evaluate this problem (4). It should become a routine part of data interpretation in molecular epidemiology and we should refrain from stratified analyses unless guided by strong prior hypotheses. The intent is not to completely discourage ''hypothesis generating'' analyses, but rather to insist that we be honest when this occurs and present the findings accordingly.

In closing, the mapping of the human genome in many respects is just the beginning. Our discipline must play a major role in determining the relevance of genetic variation on disease risk. The early efforts to explore this arena have been characterized as disappointing, as very few findings have been replicated with any degree of consistency (5, 6). Hopefully, the suggestions raised herein will contribute to an improvement in our batting average.

References

1. Cavenee WK, Hansen MF, Nordenskjold M, et al. Genetic origin of mutations predisposing to retinoblastoma. Science, 1985;228:501–3.
2. Knudson AG Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci, 1971;68:820–3.
3. Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. J Comput Graph Stat, 2003;12:156–75.
4. Rebbeck TR, Ambrosone CB, Bell DA, et al. SNP's, haplotypes and cancer: applications in molecular epidemiology report of a conference held at Key Biscayne, FL, September 13–17, 2003. Cancer Epidemiol Biomark Prev. In press.
5. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DC. Replication validity of genetic association studies. Nat Genet, 2001;29:306–9.
6. Vieland VJ. The replication requirement. Nat Genet, 2001;29:244–5.

Genetic Ancestry and Molecular Epidemiology

Thomas A. Sellers

Cancer Epidemiol Biomarkers Prev 2004;13:499-500.