Hypothesis/Commentary

Design Considerations for Genomic Association Studies: Importance of Gene-Environment Interactions

Loïc Le Marchand and Lynne R. Wilkens

Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii

Introduction

Several recent reviews and editorials have highlighted the limited success obtained to date in unraveling the joint effects of genes (G) and the environment (E) in the causation of cancer and other complex diseases (1-3). Some have questioned the wisdom of focusing on G × E interactions (1); others have even warned that funding agencies might already be losing interest in this critical area of research (3). With the advent of genotyping platforms that can be used to interrogate large parts, if not the entire genome, for association with disease, investigators have been quick to exploit these new technologies in search of common susceptibility genes. The statistical challenges raised by these new types of association studies are being vigorously debated; however, epidemiologic considerations, particularly the importance that should be given to environmental/lifestyle factors, are rarely discussed. This is regrettable considering that less-than-optimal approaches for genomic association studies may result in reinforcing this perception of unfulfilled promises.

E>G in the Etiology of Cancer

Several well-documented epidemiologic observations point to the predominant role of the environment in the causation of cancer: (*a*) the considerable variation in cancer incidence rates that exists among populations of similar ancestry, (*b*) the shift in disease risk experienced by migrants toward that of their host population, (*c*) the sharp secular trends observed for many cancers, and (*d*) the greater cancer risk explained by nonshared environmental exposures than shared heritable factors in twins (4). This evidence strongly suggests that the environment (mostly lifestyle) exerts a greater overall influence than genes on the causation of cancer. This tenet of cancer epidemiology has been reaffirmed many times in the past 30 years; yet, it is rarely taken into consideration in association studies of common predisposing genes because these studies are almost unvaryingly designed to test main effects of genes (see below).

Importance of G × E Interactions

On the other hand, there is also abundant evidence that environmental factors fail to completely explain cancer risk and that genetic susceptibility may play a crucial role in determining who among similarly exposed individuals will develop the disease. All common cancers show significant heritability (4) even after excluding known familial forms of the disease. For certain malignancies for which much is known of their environmental etiology (e.g., lung and colorectal cancers), there remain major ethnic/racial differences in risk after taking lifestyle exposures into account. Genetic variants interacting with environmental factors are likely to play a role in these ethnic differences. This is of course the case of skin cancer, the occurrence of which varies with skin pigmentation, a genetically determined trait. Other examples include the lung cancer risk associated with smoking, which was shown to vary over 2-fold among ethnic/racial groups in the United States, after adjusting for change in smoking status, smoking dose and duration, and other risk factors (5, 6).

Similarly, cancer incidence rates for migrants from historically low-risk countries have in some cases surpassed those of their similarly exposed host populations, suggesting the existence of G × E interactions as a cause for their high susceptibility to the disease (e.g., colorectal cancer in Japanese Americans; ref. 7). Finally, there are several examples in which genetic variants were reproducibly found to modify the effect of specific lifestyle factors on cancer risk (e.g., *NAT2*, smoking, and bladder cancer; *MTHFR* C677T, folate status, and colorectal cancer).

Other observations indicate that environmental factors play an important role in modifying the penetrance of cancer susceptibility genes, even those displaying a Mendelian mode of inheritance. For example, it has been shown that the breast cancer risk of *BRCA1* mutation carriers is greater for recent birth cohorts compared with older birth cohort, age for age (8). Similarly, the main cancer phenotype observed in Lynch syndrome families has changed from stomach cancer to colorectal cancer during the last century, following the trends seen in the general population (9).

Cancer Epidemiol Biomarkers Prev 2008;17(2):263–7

Received 5/2/07; revised 10/6/07; accepted 11/21/07.

Grant support: National Cancer Institute/U.S. Department of Health and Human Services grants CA72520, CA85997, and CA74806.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Loïc Le Marchand, Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Suite 407, 1236 Lauhala Street, Honolulu, HI 96813. Phone 808-586-2988. E-mail: loic@crch.hawaii.edu

doi:10.1158/1055-9965.EPI-07-0402

Genetic Variants as Biomarkers

A practical consideration worth noting about genetic variants is that, unlike lifestyle exposures, they do not change over the lifetime and thus necessarily pertain to the period that is relevant to carcinogenesis. They also are characterized with very little error and are less subject to certain biases (e.g., reverse causation) than other biomarkers. Population stratification is a concern when using unrelated controls, but it can be detected and controlled for by several genomic methods that have recently been described. Thus, genetic markers contribute valid, time-independent, biologically based information to epidemiologic studies and are thus important research tools. One of their main limitations, however, is that it is often difficult to determine whether they are causal or merely markers for another tightly linked variant.

Implications of E>G for Genetic Epidemiology Research

The low-penetrance susceptibility polymorphisms and the exposures with which they may interact are often common in the population. In addition, the magnitude of these interactions can be substantial. Thus, G × E interactions, when they exist, may be responsible for a large fraction of the cases in the population. Their identification provides important causal evidence for the environmental factors involved and improves our understanding of the underlying biological mechanisms. When the environmental exposure can be modified, this knowledge can be directly translated into prevention. The effect size of a preventive intervention can then be expected to be large in the subgroup with the at-risk genotype. This is in sharp contrast to the identification of genes that contribute to risk independently from the environment. Although these genes may be relatively common in the population, their effects on risk are typically small and their potential for prevention is low, possibly limited to contributing to a multigene risk prediction model. Thus, one can take the view that the genetic factors of main importance to chronic diseases are those that modify environmental risk factors.

Current Approach to Identifying G × E Interactions

The study of G × E interactions in cancer emerged from two approaches: that of traditional cancer epidemiology, which historically focused on environmental risk factors, and that of genetic epidemiology, a discipline that was built on the elucidation of rare Mendelian syndromes, where the role of the environment could reasonably be ignored. Because both fields realized the need to study common low-penetrance variants in the much more common nonfamilial forms of cancers, and because linkage studies have proven less powerful than association studies for studying these variants, the traditional case-control study have become the standard study design. Lifestyle exposure information is typically obtained through a questionnaire and genomic DNA is characterized for genetic variants, either selected for their being functional or serving as markers for tightly linked untyped variants. The great majority of case-control studies have used unrelated controls, but family-based designs (e.g., using discordant sibpairs) offer a notable alternative. Although family-based studies are somewhat less powerful than case-control studies using unrelated controls because of ''overmatching'' on genotype and on several environmental exposures, they can be more powerful for testing G × E (and G × G) interactions (10). They also provide absolute protection from population stratification bias (i.e., confounding by ethnicity/race), to which conventional case-control studies are subject, although the magnitude of this bias is still hotly debated (11, 12).

In a typical data analysis of a case-control study, main effects are investigated separately for each environmental risk factor and for each gene variant. The practice has been that it is only when a genetic variant is statistically significantly associated with disease that interactions with environmental factors are then tested and vice versa (Fig. 1). (This approach also applies to G × G interactions.) The reason is of course to reduce the number of independent tests and the likelihood of chance findings (inflated type I error). This view seems to have been enforced by grant review panels and journal reviewers.

However, this approach discounts the possibility that an important association may be limited to a small subset of subjects, resulting only in weak overall main effects that would require an unusually large sample size to be detected as statistically significant. To determine the magnitude of relative risks for main effects in the presence of a large gene-environment interaction and little effect of the genotype or environmental exposure alone, we did a simulation with 1,000 iterations from a case-control study, where the odds ratios (OR) were 1.2 for each factor individually and 10 jointly, where each factor was observed in 10% of the population, and where the dominant model was assumed for G with an allele frequency of 5% (see Figure 2). Data sets were created based on a multinomial distribution across the cells defined by case-control status, gene and environmental factors, and used to estimate ORs for the main effects of the factors. The median OR was 1.36 for both the main (marginal) effect of the gene and that of the environmental exposure. With a sample size of 2,500 cases and 2,500 controls, the power to detect these main effect ORs as significant at a critical level of 0.05 was 87%

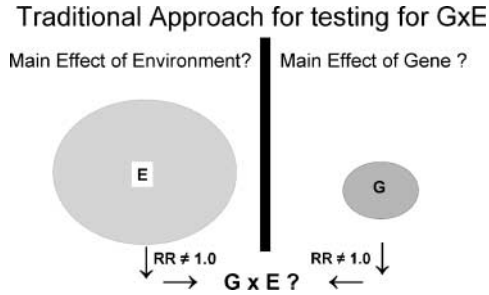

Figure 1. In typical association studies, main effects are investigated separately for each environmental risk factor and for each gene variant. The practice has been that it is only when a factor is statistically significantly associated with disease that interactions between genes and environmental factors are tested.

Exposure

Gene	No (90%)	Yes (10%)	**Main Effect of G** (over all exposure levels)
w/o (90%)	1.0 (ref.)	1.2	1.0 (ref.)
w. (10%)	1.2	10.0	1.36 (median OR)
Main Effect of E (over all genotypes)	1.0 (ref.)	1.36 (median OR)	

Figure 2. Simulation of a nested case-control study. The median ORs were based on 1,000 data sets of 2,500 cases and 2,500 controls randomly created according to the criteria in the table and the multinomial distribution. The power to detect these main effects as statistically significant is 87% for a study with 2,500 cases and 2,500 controls but only about 50% for a study with 1,000 cases and 1,000 controls. However, a sample size of only 1,250 cases and 1,250 controls is needed to detect the interaction at the 0.05 level and a power of 80% in the 2 × 2 table. Thus, an interaction of this type would be missed in most current association studies if tested only after finding a significant main effect. These power estimates are not corrected for multiple comparisons. Note that we are assuming a dominant genetic model with an allele frequency of 5%.

(with no multiple comparison adjustment). The power was around 80% with 2,000 cases and 2,000 controls and 50% with 1,000 cases and 1,000 controls. This is in contrast to the sample size of 1,250 cases and 1,250 controls needed to detect the interaction effect in the 2 × 2 table of ORs as significant with an 80% power. Even if error in measuring E attenuates the interaction OR by 10% or 20%, the sample size needed remains at 1,367 and 1,504 case-control pairs, respectively.

Also of interest is whether using information on G × E interactions in the first stage of a two-stage study would improve power to detect those variants that act through interaction with environmental factors in a small subset of subjects. In multistage studies, the current practice is also to test for G × E interactions only in the latter stages for variants found to have an overall marginal main effect. We simulated a two-stage study of G × E interactions using optimized parameters based on Wang et al. (13) and based on the joint distributions of ORs in Fig. 2, with 1,000 iterations, an overall P value of 5×10^{-8} (Bonferroni correction of 0.05 based on 1 million single nucleotide polymorphisms), and 8,000 participants: 1,000 cases and 1,000 controls in stage I and 3,000 cases and 3,000 controls in stage II. This sample size and design can detect as significant, with an overall power of 80%, a marginal main effect OR of 1.57 for G assuming a dominant model and of 1.46 assuming a log-additive model (14). However, the power to detect the G × E interaction in this two-stage study, with promotion rules based on the main effect of G in stage I, is only 49%. Using as promotion criteria the significance of the gene main effect or the G × E effect, and adjusting the critical value further for multiple testing, increases the power to 62%. Moreover, additional consideration of the significance of the main effect of E for promotion, with the relevant adjustment for multiple testing, increases the power to 88%. Using a case-only comparison to test for G × E interactions changes these respective powers to 64%, 93%, and 96%. Thus, in this example, only relying on a significant marginal main effect of G to detect G × E interactions is clearly inefficient in a one-stage or two-stage design and consideration of the G × E effect in the promotion rules of a two-stage design may improve the power substantially.

Large interactions leading to modest main effects would be important to identify as they would be informative as to the etiology of the disease. Further, this interaction scenario is not only realistic but might actually be common, if not the norm. Lifestyle risk factors are very prevalent, but cancer remains a rare disease. This is no doubt because cancer results from a complex interplay of risk and protective factors acting through competing pathways. As more becomes known of the etiology of a specific cancer, it becomes increasingly clear that multiple exposures and candidate genes act in the same pathway to increase or decrease risk, resulting in small ''susceptible'' subgroups.

Our recent findings on metabolic genes, well-done meat, and colorectal cancer illustrate this point (15). Heterocylic amines are carcinogenic compounds that are formed when meat is cooked at high temperature. After absorption, they require biochemical transformation, first in the liver, primarily by CYP1A2, then in the colon, by NAT2 before they can bind to DNA and, in the absence of DNA repair, cause a mutation that may lead to colorectal cancer. Both NAT2 and CYP1A2 are polymorphic and CYP1A2 is inducible by smoking. In our study, only weak main effects (ORs of 1.2, 1.3) were found; however, an OR of 8.8 was found for subjects who were both exposed and genetically susceptible. This effect was observed in the small subgroup hypothesized *a priori* to be at increased risk, those who were exposed to both the carcinogen (preference for well-done meat) and the inducer (smoking), and who carry the high-activity genotype for both activating genes (*CYP1A2* and *NAT2*). No lower-order combinations of these factors showed any substantial association with risk. This finding was consistent with past studies and what is known of the pathway.

Another example is the well-reproducible interactions between folate intake, alcohol, and the *MTHFR* C677T variant on the risk of colorectal cancer (reviewed in ref. 16). The TT genotype has been shown to result in a marked reduction in the activity of the enzyme and has been associated with a reduced risk of colorectal cancer, with this effect being limited to individuals with a high folate status. This association with the TT genotype was only present at low intake level of ethanol (a folate antagonist), further reducing the size of the protected subgroup. The plausibility of this association is supported by the central role played by MTHFR in regulating the flow of folate between two important pathways affecting cancer risk: the production of thymidylate and purines for DNA synthesis and the supply of methyl groups for the methylation of DNA and other important proteins.

Use of ''Enriched'' Populations to Increase Statistical Power to Detect G × E Interactions

If indeed G × E interactions typically occur in small subsets of the population, it would be advantageous to

design studies to maximize statistical power by enriching the study sample for the gene(s) and/or environmental exposure(s) of interest. For example, one could select a population with a demonstrated susceptibility to the environmental causes of the disease (e.g., a population where a sharp increase in incidence has occurred independently of screening, or a population with an unexpectedly high risk, given an exposure level). Another approach is to select a population with a high allele frequency (when following a candidate gene approach) or, as more often done in family-based studies, to oversample for family history or early age at onset as an attempt to blindly enrich the study with gene carriers.

One could also focus on cases and controls highly exposed to the known environmental risk factor(s) of interest (e.g., study only breast cancer cases and controls that used hormone replacement therapy when testing for genes involved in estrogen synthesis and metabolism). In the earlier simulation study with 5,000 subjects, among 500 individuals exposed to the environmental factor, the power was ≥95% for detection of the genetic factor. (The power would be similar for studying exposure among 500 individuals carrying the susceptibility gene.) For a disease for which not much is known about environmental risk factors (e.g., prostate cancer), it may be assumed that a population with extraordinarily high rates (e.g., African Americans) would be exposed to those unknown factors and thus would constitute an enriched choice of study population.

Correctly Measuring E, Avoiding Biases, Refining the Phenotype, and Analyzing Pathways

Except for a few exceptions (e.g., smoking, alcohol, and body mass index), environmental factors are notoriously difficult to measure in population-based research. Because measurement error decreases the power of tests that include environmental factors (test of E or G × E interactions), it is important in genomic association studies to optimize the quality of these measurements. It is also helpful to use multiple related measures (in particular biomarkers) with uncorrelated errors to decrease, or correct for, measurement error. For example, the finding of an interaction on the risk of colorectal cancer for the *MTHFR* 677TT genotype and both plasma and dietary folate strengthens the likelihood that this interaction is real and specific to folate status. NIH, in its Genes, Environment and Health Initiative, has recognized this need by funding research on new methods for monitoring environmental exposures that interact with genetic variation (http://www.gei.nih.gov/exposurebiology/).

Moreover, as in any association studies and as noted many times, it is important to avoid biases, such as selection, survival, screening, and recall bias. In that regard, prospective studies are superior to retrospective design, because these biases are less common or avoided in cohort studies. It is also important to reduce misclassification on disease status by reducing heterogeneity in the phenotype (e.g., only including ER-positive cases in the breast cancer example above because estrogens, including hormone replacement therapy, are more strongly associated with this subtype of breast tumor).

As pointed out above, as we learn more about specific biological mechanisms, new approaches and methodologic tools to model environmental and genetic factors, and their interactions, are needed to better characterize these complex pathways and better assess individual risk. The predictability of the risk model in the well-done meat and colorectal cancer example above could theoretically be improved by taking into account genetic variants in other parts of the pathway, such as the detoxification of heterocyclic amines and DNA repair. Assessing multiple types of biomarkers (e.g., intermediate metabolites, DNA adducts, and methylation levels) may also be useful in pathway-driven analyses to better characterize biological relationships (17). In the folate, *MTHFR*, and colorectal cancer example, measuring serum homocysteine and assessing gene methylation and purine/pyrimidine synthesis might provide information on the respective contribution of each part of the pathway to risk.

Testing for G × E Interactions in Genomic Association Studies

Current studies are now testing several hundreds or thousands of single nucleotide polymorphisms (close to 1 million with newly released gene arrays) for association with disease. The genome-wide association studies reported to date have followed the practice noted above of focusing on overall main effects. Presumably, these studies will, at a later date, test for G × E interactions using gene variants found consistently associated with disease in validation and replication steps. Little consideration seems to have been given to the large environmental risk component that has been shown for all major cancers. Given the high cost and potentially low power of these studies, a judicious choice of study population as discussed above may significantly increase power. Moreover, environmental factors should be considered in the genome-wide data analysis. Instead of limiting a genome-wide association study to the testing of the main effects of genes, separate analyses could be added to identify genetic markers interacting with relevant lifestyle risk factors on a genome-wide basis followed by validation and replication of the most significant genetic main effects and G × E interactions in appropriate studies. This proposal is similar to that of Marchini et al. (18) and Evans et al. (19) in the context of testing for G × G interactions on a genome-wide basis. These authors showed that in analyzing genome-wide association studies, fitting models that explicitly allow for interactions between loci can add substantially to single locus-by-locus searches despite the penalty introduced by multiple testing. They showed that these interaction-based searches can be more powerful than single-locus approaches (18) and that an exhaustive search involving all pairwise combinations of markers across the genome decreases the risk of missing interaction loci that contribute little to the main effect (19).

For example, a multistage genome-wide association study of colorectal cancer could test in its first stage main effect associations with genes as well as G × E interactions with several environmental factors, such as well-done meat, smoking, alcohol, folate, and calcium, for which interactions with single nucleotide polymorphisms have already been suggested by past

studies (15, 16, 20, 21). A case-only analysis could be considered as it offers improved power for detecting G × E interactions (14). This approach assumes independence between genetic and environmental risk factors, an assumption that can be checked among the controls. As noted above, a prospective study would provide the strongest design to test for these G × E interactions. The most strongly associated single nucleotide polymorphisms, for both main effects of G, and possibly E, and interaction effects with E, would be carried forward to the subsequent stage(s), and their overall evaluation for association or effect modification can be conducted using the combined data from all stages. Then, replications of main effects and interactions would be conducted in other cohort studies to validate the findings.

Conclusion

The promises of genetic susceptibility studies, especially in terms of prevention and public health, will not be fulfilled unless we successfully identify the genetic variants that modify the effect of lifestyle and the environment on cancer risk. Genomic association studies provide new tools for making great strides toward these goals, even with our current limited understanding of cancer biology. The power of these tools has recently been shown by the identification of the first truly reproducible genetic associations for prostate cancer with single nucleotide polymorphisms located at 8q24, a chromosomal region that was not suspected previously to play a role in cancer (22). This finding has tremendous potential for improving our understanding of the biology of prostate cancer as well as, possibly, that of its environmental causes that have eluded us for so many years. It is imperative that environmental exposures be considered in the design, particularly in the selection of study populations, and in the analysis of genomic association studies to make full use of the data and obtain the best return on the investment. The practice of relying on significant main effects of G to investigate G × E interactions is not always adequate, and genome-wide testing for these interactions is indicated despite the penalty introduced by multiple testing.

References

1. Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001; 358:1356–60.
2. Brennan P. Gene-environment interaction and aetiology of caner: what does it mean and how can we measure it? Carcinogenesis 2002; 23:381–7.
3. Sellers TA. The beginning of the end of the epidemiologic focus on gene-environment interactions? Cancer Epidemiol Biomarkers Prev 2006;15:1059–60.
4. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343: 78–85.
5. Le Marchand, Wilkens LR, Kolonel LN. Ethnic differences in the lung cancer risk associated with smoking. Cancer Epidemiol Biomarkers Prev 1992;1:103–7.
6. Haiman CA, Stram DO, Wilkens LR, et al. Ethnic and racial differences in the smoking-related risk of lung cancer. N Engl J Med 2006;354:333–42.
7. Le Marchand L. Combined influence of genetic and dietary factors on colorectal cancer incidence in Japanese Americans. Monogr Natl Cancer Inst 1999;26:101–5.
8. King MC, Marks JH, Mandell JB; New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 2003;302:643–6.
9. Potter JD. Colorectal cancer: molecules and populations. J Natl Cancer Inst 1999;91:916–32.
10. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 1999; 149:693–705.
11. Thomas DC, Witte JS. Point: Population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002;11:505–12.
12. Wacholder S, Rothman N, Caporaso N. Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 2002; 11:513–20.
13. Wang H, Thomas DC, Pe'er I, Stram DO. Optimal two-stage genotyping designs for genome-wide association scans. Genet Epidemiol 2006;30:356–68.
14. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 2002;155:478–84.
15. Le Marchand L, Hankin JH, Wilkens LR, et al. Combined effects of well-done red meat, smoking and rapid NAT2 and CYP1A2 phenotypes in increasing colorectal cancer risk. Cancer Epidemiol Biomarkers Prev 2001;10:1259–66.
16. Le Marchand L, Wilkens LR, Kolonel LN, Henderson BE. The *MTHFR* C677T polymorphism and colorectal cancer: the Multiethnic Cohort Study. Cancer Epidemiol Biomarkers Prev 2005; 14:1198–203.
17. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005;14:557–9.
18. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex disease. Nat Genet 2005;37:413–7.
19. Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. PLOS Genet 2006;2:1424–32.
20. Cortessis V, Siegmund K, Chen Q, et al. A case-control study of microsomal epoxide hydrolase, smoking, meat consumption, glutathione *S*-transferase M3, and risk of colorectal adenomas. Cancer Res 2001;61:2381–5.
21. Kim HS, Newcomb PA, Ulrich CM, et al. Vitamin D receptor polymorphism and the risk of colorectal adenomas: evidence of interaction with dietary vitamin D and calcium. Cancer Epidemiol Biomarkers Prev 2001;10:869–74.
22. Witte JS. Multiple prostate cancer risk variants on 8q24. Nat Genet 2007;39:579–80.