Commentary

The Predominance of the Environment over Genes in Cancer Causation: Implications for Genetic Epidemiology

Loïc Le Marchand
Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii

Introduction

It has generally been accepted that most cancers result from the combined effects of environmental factors and inherited susceptibilities and that only few cancers are “solely” of genetic origin (1, 2). Yet, the current strategies for identifying genetic risk factors for common cancers are still almost entirely based on the premise that there are two forms of the disease, one that is genetic and the other that is not. The importance of the environment in cancer causation has implications for both gene discovery and gene characterization that may not have sufficiently been appreciated or emphasized in genetic epidemiology.

The Importance of Host Susceptibility

The fact that not all similarly exposed individuals (e.g., smokers) get the same disease (e.g., lung cancer, coronary heart disease), or any disease at all, is central to improving our ability to assess risk at the individual level. To date, public health and regulatory policies have been based on the assumption that all individuals in a population have the same biological response to a specified level of a cancer-causing exposure (3). However, molecular epidemiology approaches now enable us to assess the role played by host factors, particularly genetic factors, in this interindividual variation in response. Because, typically, the marked reduction or complete elimination of a deleterious exposure is socially and economically costly, it may be important to prioritize susceptible groups for targeted intervention. Moreover, further progress in exposure avoidance (e.g., smoking) and screening adherence may be made if one’s genetic susceptibility to developing cancer can clearly be linked to a specific exposure or risk behavior although the evidence for this is still limited. Most importantly, the improved biological understanding that will come from better defining host susceptibility should result in more effective prevention and therapeutic strategies.

Genetic Susceptibility to Cancer

The explosion of knowledge about the human genome has already made possible the molecular elucidation of inherited cancer syndromes. Linkage studies among large multicase families have been instrumental in mapping the mutations responsible for a proportion of a number of malignant phenotypes, such as retinoblastoma, Wilms’ tumor, Li-Fraumeni syndrome, Von Hippel-Lindau disease, familial adenomatous polyposis, hereditary nonpolyposis colorectal cancer, and breast and ovarian cancers. These mutations are rare in populations but carry a high disease risk (high penetrance) and, thus, can be used for predictive genetic testing in situations where preventive interventions are possible.

However, the tumors responsible for most of the cancer morbidity and mortality are likely to result not from mutations in “major cancer genes,” such as the ones mentioned above, but from the interplay of multiple environmental and genetic factors acting over several decades. A significant interindividual variability also exists in tumor progression and response to treatment. As a result of this etiologic complexity, malignant diseases that do not follow a simple Mendelian pattern of inheritance in a substantial number of families have remained refractory to mapping by linkage.

Consequently, the research emphasis in genetic epidemiology has shifted toward exploring common, low-penetrance variants that may affect the risks of common malignant diseases (4). Association studies (typically, case-control studies among unrelated individuals) are conducted to compare genotype or allele frequencies between affected and unaffected participants. Low-penetrance variants have limited application in screening and risk prediction (except perhaps in the future, as part of multigenic models) because the changes in risk associated with each are expected to be modest at best. The importance of these associations should not be underestimated, however, as they may contribute critical evidence in support of a risk factor or biological mechanism.

Current Trends in Genetic Epidemiology

The initial candidate gene association studies were often limited to a single polymorphism and typically focused on variants with known functional significance, as shown by assessing gene expression, or measuring the gene product or an intermediate marker (e.g., enzyme activity or levels of a metabolite or hormone). These studies were building on laboratory or clinical data and were typically based on a strong biological rationale. Unfortunately, these initial investigations were often mired with problems (e.g., inadequate control group, insufficient statistical power, and/or selective reporting), led to a plethora of replication studies, often sharing similar limitations, and resulted in poor reproducibility of these early findings (5). Despite a recent increase in study size and quality, it is probably fair to say that a lack of reproducibility, many null findings, and only a handful of reproducible associations characterized these early studies. Also, it is notable that the magnitude of the (main) effects observed in these studies has been low, with odds ratios typically in the 1.3 to 1.5 range. Another general observation...
from these studies is the marked differences that exist across ethnic/racial groups in allele frequency for most common genetic variants studied.

The recent availability of extensive sequence data on the human genome (and, particularly, on the 0.1% of bases that differ across individuals) and high-throughput genotyping methods has moved the field away from targeting functional polymorphisms and toward conducting wide searches for susceptibility loci. Studies are now focusing on several candidate variants in a single gene and/or on several genes in a single pathway, perceived as most critical. The most recent wave of studies is attempting to capitalize on the linkage disequilibrium that exists along stretches (“blocks”) of DNA in the human genome to systematically search for associations with common haplotypes and the observed and unobserved variants that they carry (6). A few attempts are also been made to conduct genome scans searching for associations with large numbers (100,000+) of markers systematically distributed throughout the genome. The assumption is that the critical work of zooming in on a causal polymorphism, if it exists, and documenting its biological function would follow the establishment of a reproducible association. Given the cost and large scale of these endeavors, the extent to which replication will be practical remains unclear. Moreover, the large number of comparisons made in these studies introduces sizable methodologic challenges and the weak effects typically observed have been difficult to interpret.

The Predominant Role of the Environment in Cancer Causation

As early as the 1970s and 1980s, migrant and temporal trend studies have clearly shown that major changes in cancer incidence can occur over a few decades and as early as in the first or second generation of migrants (1, 2). Also, substantial variation exists in cancer rates among countries with populations of similar ethnic/racial background. These observations have been made for all common cancers (e.g., breast, prostate, colorectal, and lung). Two major conclusions are drawn from these data: (a) fast changing environmental factors, and not the genetic makeup of a population, play a predominant role in determining cancer risk and (b) substantial prevention opportunities should result from the identification of key modifiable risk factors in the environment (usually taken as including lifestyle, medications, environmental pollutants, radiation, and infectious agents).

This tenet of cancer epidemiology has led to several decades of intensive investigation aimed at identifying environmental risk factors. Some major successes were achieved, such as the establishment of the roles of smoking, alcohol, obesity, hormone replacement therapy, and physical activity, in the etiology of a number of important cancers. Indeed, for most major cancers, the majority of the variation in risk across individuals can now be explained by known environmental risk factors (80-90% for lung cancer, 70-85% for colorectal, over 50% for breast cancer; refs. 2, 7, 8). However, and in particular in the area of diet, suggestive laboratory and ecological evidence for the role of a number of other factors has not been confirmed in analytic epidemiologic studies. This relative “failure” of epidemiology reflects a number of limitations, including the difficulties in measuring long-term exposures and the lack of exposure heterogeneity in most studies. It also probably reflects the insufficient attention given to host factors.

Also, some well-documented cancer risk differentials (e.g., differences in breast cancer among similarly exposed populations argue in favor of a major role for modifying factors, most likely genetic in nature. For example, the >2-fold difference in the strength of the association between smoking and lung cancer that exists among ethnic/racial groups (9) has spurred the investigation of genetic factors in lung cancer susceptibility. Also, the marked excess risk of some ethnic/racial groups for certain cancers have remained unexplained when considering known and measured lifestyle risk factors, such as the high prostate cancer risk of African Americans and the high colorectal cancer risk of Japanese Americans.

Implications for Genetic Epidemiology

It can be argued that the predominance of the environment over genes in the causation of cancer has several major implications for genetic epidemiology that have not sufficiently been taken into consideration.

Important genetic factors are primarily effect modifiers. If environmental exposures, and principally lifestyle factors, play a predominant role in the etiology of cancer, then genetic factors of potential importance would be expected to display interactions with environmental risk factors (i.e., the effect of the environmental exposure(s) should be limited to, or much stronger in, the subset of individuals with the high-risk genotype). Consequently, these gene-environment (G × E) interactions should not be viewed as exploratory analyses as they currently tend to be. They should be embraced as core analyses, driving the design, analysis, and interpretation of genetic epidemiology studies. New approaches for incorporating prior biological knowledge into our analysis plans are needed. For example, kinetic and pharmacologic data from animal models and clinical studies could be used, when available, to build mechanistic models for these analyses.

Although studies of G × E interactions require larger sample sizes and are more expensive, they are likely to identify factors that are amenable to intervention and, thus, of greater public health significance. The recent establishment of very large collaborative studies (e.g., Cooperative Family Registries1 and Cohort Consortium2) offers new opportunities for innovative approaches to investigating G × E interactions.

The main effects of genes incorporate the effects of environmental factors and other genetic factors. If common genetic factors act primarily as modifiers, their main effects, “averaged” over all levels of exposure to the interacting factor(s), are most certainly weak. Even if the relevant environmental exposures are assessed and correctly modeled with the genetic factor of interest, these effects are likely to be “diluted” by the existence of other unmeasured environmental and genetic factors. This is because most genes are expected to act through a complex biological pathway where cofactors and competing or compensatory mechanisms exist. Our recent findings on metabolic genes, well-done meat intake, and colorectal cancer illustrate this point (10). Only weak main effects (odds ratio, 1.2-1.3) were observed in this study; however, an odds ratio of 8.8 was found for subjects who were both exposed and genetically susceptible. Several environmental factors (consumption of well-done meat and smoking) and genes (NAT2 and CYP1A2), originally hypothesized to be part of the same pathologic pathway, had to be considered to identify this effect. No lower-order combinations of these factors showed any substantial association with risk. This pattern is consistent with past studies that almost uniformly considered only two of these factors (NAT2 and well-done meat) and observed at best weak associations (11).

Heterogeneity in the main effects of genes should be expected across populations. The strength of a main effect between a gene and a specific cancer should be expected

\[1 \text{http://www.ccr.epicr.uc.edu/} \]

\[2 \text{http://epi.grants.cancer.gov/BPC3/} \]
to vary across populations because the size of the genetically susceptible group is likely to vary based on ethnic/racial background and/or because the interacting environmental exposures are not present to the same extent. A gene effect may not be detectable in populations with few susceptible individuals and/or little exposure to the relevant environmental factors. Heterogeneity in the main effects of genes should be particularly high for cancers with few known environmental risk factors (e.g., prostate cancer) because those cannot be taken into consideration and are likely to vary as well across populations. This variability is likely to weaken the power of pooled analyses of genetic association studies conducted in dissimilar populations. Stratifying these pooled analyses by ethnic/racial group and population risk levels may be helpful in interpreting the data.

**Genome scans should be conducted in samples from high-risk populations.** Because the main effects of (modifying) genes are expected to be weak, then they should be easiest to detect in populations with an exceptionally high risk for the cancer of interest. A genetic susceptibility that increases risk by amplifying the effect of single or multiple environmental exposures should more clearly manifest itself in a population where these exposures are prevalent. In a high-risk population, it is more likely that such exposures are occurring. Thus, wide gene discovery efforts, such as whole genome association studies, may be best conducted in high-risk populations. This is analogous to the accepted (but still rarely used) practice of oversampling cases with a positive family history or early age at onset to enrich the sample for genetic factors in gene discovery efforts.

When studying a cancer for which the environmental etiology has been well characterized (e.g., colorectal cancer), one could increase the efficiency of a genome scan by selecting both case and control groups on their having a high-risk profile based on a comprehensive modeling of lifestyle risk factors (7). Contrasting cases and controls with similarly high-risk lifestyles should increase the power to detect the effects of genes. A low-risk population would only be interesting if it had low incidence rates despite substantial exposure to known risk factors.

By selecting high-risk populations to conduct genome scans, one would not be able to identify a negative $G \times E$ interaction through which the subgroup with the minor allele would be protected (e.g., the nullifying effect of alcohol intake on the inverse association between the MTHFR 677TT genotype and colorectal cancer; ref. 12). However, one can argue that identifying these relationships is not a public health priority as the recommendation made to the general population to reduce exposure to the environmental factor would not be altered by the identification of such a negative interaction.

**Taking full advantage of genetic epidemiology findings for public health interventions will require improving the measurement of environmental exposures.** Although genotype can be assessed with little measurement error with modern techniques, most environmental risk factors are still measured with considerable inaccuracy. The resulting misclassification decreases the power of studies to detect $G \times E$ interactions (13) and, thus, is an impediment to correctly characterizing the risk associated with genes. Although, by selecting a high-risk population, susceptibility genes should be identifiable through their main effects, correctly assessing $G \times E$ interactions will be needed before this knowledge can be translated into valid individualized interventions. In the meantime, public health recommendations are likely to continue to be made for the population as a whole.

**Cohort studies are preferable over case-control studies.** As pointed out many times, case-control studies are open to differential recall when collecting questionnaire information and to reverse causation when assessing a biomarker. In contrast, these biases are minimized in cohort studies because biological specimens and information on environmental exposures are obtained before diagnosis. Thus, cohort studies should be preferred for studying $G \times E$ interactions.

**Conclusion**

The predominant role of the environment in the etiology of cancer has been recognized for decades. Now that we have the ability to study the role of common genetic variants in disease etiology, priority should be given to unraveling the joint effects of lifestyle factors and genetic susceptibilities. Not only this new research area should be recognized as a critical component of genetic epidemiology, but it should also be taken into account in the design and interpretation of gene discovery efforts (e.g., by oversampling high-risk populations for genome scans or stratifying pooled analyses by population-risk level). Most importantly, focusing on $G \times E$ interactions should greatly facilitate the translation of genetic epidemiology findings into major public health advances. It may also help to change the public’s perception of the role of susceptibility genes in cancer from one of determinism to one that mainly reflects our maladaptation to changing life habits.

**References**

The Predominance of the Environment over Genes in Cancer Causation: Implications for Genetic Epidemiology

Loïc Le Marchand