New Lung Cancer Susceptibility Locus Identified: Significance and Implications for Other Genome-Wide Association Studies

Thomas A. Sellers1 and Y. Ann Chen2

Summary: Genome-wide association studies have identified more than 150 loci that influence the risk of cancer. In this issue of Cancer Discovery, Shi and colleagues report that a variant in RAD52 is a risk factor for squamous cell lung carcinoma. This work is important not only for its potential implications on control of this dreaded malignancy but also for its methodologic contributions that can advance the field of molecular-genetic epidemiology. Cancer Discovery. 2(2):110-1. ©2012 AACR.

Commentary on Shi et al., p. 131 (4).

Lung cancer has historically been cited as the prototype of a malignancy that is determined solely by the environment, principally tobacco exposure. Although familial clusters of lung cancer had been reported in the literature, the fact that smoking habits are familial in nature prevented any conclusions about whether there was a genetic influence on the disease. That changed in 1990, with a statistical genetic analysis of families that provided strong evidence for the existence of a major genetic influence, even after adjustment for individual-level smoking (1). A major gene locus was identified in family studies some 14 years later at 6q23-25 (2), with fine mapping of the locus revealing that RGS17 was the most likely gene influencing susceptibility (3).

Because this single locus could not account for all of the interindividual variation in risk and the frequency of the allele is low, a strategy using genome-wide association studies (GWAS) of unrelated individuals to identify common variants was adopted. Conducted primarily among populations of European ancestry, GWAS results have been quite consistent in their implication of 3 polymorphic variants at 15q25.1 (CHRNA5-CHRNA3-CHRNB4), 5p15.33 (TERT-CLPTM1), and 6p21.33 (BAT3-MSH5) as risk factors for lung cancer. The report by Shi and colleagues (4) in this issue of Cancer Discovery now adds a new locus at 12p13.33, most likely encompassing the homologous recombination gene RAD52 as a susceptibility gene for squamous cell lung carcinoma.

Although any progress on cancer is important, especially lung cancer, which is the leading cause of cancer death among men and women with more than a million deaths per year worldwide, the report by Shi and colleagues (4) is also notable for its important methodologic advances. The first is that the analysis was restricted to a particular histologic subset of lung cancer. Like many other cancers, lung cancer histologic subtypes can have dramatically different clinical behaviors. There is increasing recognition that this may reflect differences in environmental influences, but the current study reinforces the notion that there may be differences in genetic susceptibility as well. At present, such approaches consider histology or stratification on the basis of cellular receptors. It is reasonable to speculate that future studies in the field might incorporate more global categorizations such as gene expression profiles or genomewide methylation profiles to define subsets of the disease that are more etiologically homogeneous.

The second methodologic contribution of the report by Shi and colleagues (4) is how they addressed one of the fundamental limitations of GWAS. In particular, on a gene chip with 500,000 genetic markers, one would expect approximately 25,000 P values less than 0.05. Most of these associations are attributable to chance rather than underlying biology, which is why a very stringent threshold for statistical significance (P value of $10^{-8}$) is mandated before a GWAS variant is declared significant.

However, there are almost certainly other bona fide loci lurking beneath this threshold. To discover them, we cannot always merely rely on larger sample sizes. For continuous trait phenotypes, this is less of an issue, but for dichotomous traits, this can become a rate-limiting step. Thus, the application of novel statistical approaches to decrease the degrees of freedom that one expends can help illuminate novel risk loci. The strategy that Shi and colleagues (4) used was to focus on a specific pathway of genes involved in the inflammatory response, thereby reducing the 591,928 single-nucleotide polymorphisms (SNP) in the original GWAS to 19,082 SNPs that mapped to 917 genes. Moreover, this approach leverages information from other investigations and publicly available resources to further inform the value of the exercise. Thus, by taking into consideration published findings on the inflammation genes and the strength of the association, they significantly enhanced the plausibility of the observed findings.

The pathway approach is more than just a means to reduce the degrees of freedom penalty associated with an agnostic search for susceptibility loci. Importantly, such a strategy enhances the ability to leverage the large amounts of biologic knowledge accumulated in the past few decades. In particular, pathway-based methods have recently been developed to evaluate the joint effects of multiple alleles within predefined functional gene sets and pathways on cancer risk. On the basis of the formulation of hypotheses,
these methods are generally grouped into 2 major categories: competitive or self-contained null hypotheses. A competitive method compares the magnitude of the association between the genes within a gene set and the disease phenotype with the genes from the rest of the genome (e.g., ref. 5). The study by Shi and colleagues (4), however, uses a self-contained method that does not include such a comparison. On the basis of their expert knowledge of lung cancer, they focused the analyses solely on genes in the inflammation pathway. A potential next step would be to explore additional biologic pathways.

Although the current study highlights the ability to identify novel associations by applying pathway analyses, there are caveats that should be considered. Different data-processing steps and different statistical methods have their own pros and cons. Some of these are well-recognized, such as different ways to map SNPs to genes, effects of gene set size, formulation of the null hypothesis (competitive vs. self-contained), linkage disequilibrium patterns, overlapping genes, construction of the test statistics, and assessment of statistical significance (6).

However, there are at least 2 additional major challenges for agnostic pathway analyses that are discussed less frequently and deserve more attention. First, some of the predefined gene sets contain hundreds (or thousands) of genes. It is likely that a small portion of these hundreds of genes within a gene set are associated with the disease phenotype, and sometimes because of the “noise” of additional genes contained within the gene set, no significant association is detected. Second, most of the methods evaluate only one pathway at a time (like the report by Shi and colleagues; ref. 4) rather than evaluate the joint effects of pathways and functional sets. Statistical methods that address both challenges by performing marker selection across multiple levels (e.g., pathways, genes) and assessing the joint effects of selected markers across multiple levels can enable association testing in a holistic manner (7).

There have been highly publicized criticisms of the GWAS approach, primarily because of the perception that they have failed to account for a significant fraction of the heritability of the disease. One of the reasons for this is that the causal variant is almost never identified in a GWAS, just a variant that is near the actual risk locus. In most cases, fine mapping of the region is needed, and the magnitude of risk can be expected to be larger. Although Shi and colleagues (4) did not perform fine mapping of the region, they did impute unobserved genotypes by using data from HapMap Phase III and 1000 Genomes Project, but no SNPs were identified that were more strongly associated than what had been originally identified (rs6489769). It is also noteworthy that in the study by Shi and colleagues (4), the minor allele was inversely associated with risk. Thus, the major (common) allele is associated with greater risk, independent of smoking history. With an estimated minor allele frequency of 0.38, this implies that 85% of the population carries at least one risk allele and 38% carries 2 copies of the risk allele. Therefore, although the associated risk increase of 1.20 may be modest, the high frequency of carriers means the population-attributable risk is significant.

It is also appreciated that identification of a locus through a GWAS is merely the beginning. This signifies the start of rigorous laboratory testing of the candidate genes and variants for functional effects (8). Future work is needed to help understand the mechanisms underlying the association of the 12p13.33 locus with risk of squamous cell lung carcinoma and identify transnational potential for risk prediction, risk reduction, and improved therapies.

Given the size of the discovery population and the replication efforts, it is unclear whether additional DNA variants can be identified for lung cancer. Certainly, there are other biologic pathways, functional gene sets, or regulatory regions that could be examined using the analytic approach showcased by Shi and colleagues (4). In the future, most GWAS findings are not actually within genes, but in intergenic regions that may signal the location of DNA polymorphisms associated with control of translation (e.g., ref. 9). Of course, there remains optimism that there is an undiscovered cache of uncommon and rare variants in exomes for which greater effect sizes will be evident (10). Making sense of all of these data through an integrative epidemiologic framework can help realize the potential to make measurable progress against cancer.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Received December 27, 2011; accepted December 29, 2011; published online February 13, 2012.

REFERENCES
New Lung Cancer Susceptibility Locus Identified: Significance and Implications for Other Genome-Wide Association Studies

Thomas A. Sellers and Y. Ann Chen