Let us turn back the clock just over 3 years to January 2007 and imagine that we are reading an editorial confidently predicting that more than 100 new reproducible risk factors for various cancers would be discovered by 2010. Given the incremental progress over the previous decade in robustly establishing new risk factors with the candidate gene approach at the expense of a sea of unreplicated reports, we doubt that editorial would have been considered credible.

In this issue of the Journal, Ioannidis et al. (1) review the success of genome-wide association studies (GWAS) to discover common alleles associated with cancer risk.

In the 1960s, Sir Peter Medawar published The Art of the Soluble (2) and in later comments pointed out that scientific progress often results from “devising some means of quantifying phenomena or states” (3). It is hard to think of a better example in cancer epidemiology than GWAS—the newfound ability to assess case–control associations for more than 300,000 loci across the genome. The availability of a catalog of common genetic variants in the genome [courtesy of the HapMap (4)] was leveraged by the production of dense microchips that permitted simultaneous genotyping of thousands of single-nucleotide polymorphisms (SNPs). Think back a decade when a week in the laboratory could genotype a single favorite SNP in a few hundred samples, and the idea of accurately testing 300,000 SNPs in a single sample seemed like a fantasy.

The prospect of at least 300,000 $P$ values derived from a single case–control study forced GWAS practitioners to confront the sobering reality of the challenge of multiple comparisons. It was not possible to pretend we were “hypothesis testing” or to sweep the facts of post hoc cut points, elastic data definitions, or subgroup analyses under the carpet. Nor was it possible to publish “findings” without substantial evidence for replication (5). Publication standards for GWAS have been set high to ensure that the new signals are sufficiently conclusive to warrant the extensive follow-up work required. Consequently, required sample sizes for case patients and control subjects increased dramatically, leading to the design of multistage GWAS and consortium formation on an unprecedented scale in epidemiology.

What has been learned from the first 3 years of conducting “agnostic” GWAS? (An agnostic GWAS is one in which no SNP is thought, a priori, to have a higher probability of being associated with the phenotype of interest than any other SNP.) Common susceptibility alleles have been identified for both common and uncommon cancers alike. As the initial smoke clears, some trends have emerged. Most estimates of relative risk per allele are in the 1.10–1.40 range, underestimating to many geneticists accustomed to highly penetrant mutations found in classic single-gene disorders. However, these estimates fall in a comfort zone familiar to epidemiologists; many epidemiological risk factors for cancer are in this range (notably excepting smoking and infectious causes). In nearly every chromosomal region, the identified marker is a surrogate for the variants that biologically explain the association signal. Mapping the regions and then nominating the optimal common variants for functional studies represent the next frontier in pursuing the discoveries of GWAS.

Finding the functional variants in genes or intergenic regions promises to be a major new source of insight into the biology of specific cancers (6).

Known variants account for a small fraction of heritability for each cancer site. Similarly, when analyzed in combination, they do not yet enable adequate discrimination between case patients and control subjects, particularly for clinical decision making. Although this lack of predictive power is the glass half empty to some, it may be too early to render judgment. In fact, one could argue that the glass may be half full because SNP-based scores do at least as well as previously used risk prediction models, such as the Gail model for breast cancer (7). It could be argued that we have discovered more risk factors predictive of breast cancer in the past 3 years than in the previous three decades combined.

What more can be gained through the current approaches to GWAS? As Ioannidis et al. (1) demonstrate in simulations, most currently published GWAS have limited statistical power to detect any specific variant, particularly for per-allele risk estimates below 1.2. Thus, supersizing GWAS should discover additional variants in common and uncommon cancers alike.

If epidemiology is “The Art of the Soluble” (8) (as coined by Sir Michael Marmot using Medawar’s phrase), what may be possible in the future? A new generation of SNP chips, leveraging information from the 1000 Genomes Project, will genotype uncommon SNPs (ie, minor allele frequency [MAF] of 0.5%–10%), extending our reach beyond the common variants (>10% MAF) of current GWAS. Even larger sample sizes will be needed for agnostic genome-wide searches at these low MAFs. Therefore, the argument should be made for enrollment of new case patients and control subjects into appropriately consented studies that can be combined in consortia. The harder question is when will GWAS transition to whole-genome sequencing. The timing will depend on the cost of the promised “$1000 genome” (ie, the ability to fully sequence a human genome) and the computational capacity to handle the data. More daunting is the challenge of how to achieve the sample sizes needed for multiple-comparison-corrected statistical significance in analyses of rare variants. Such analyses will likely require innovative research models that use genetic information acquired for personal or clinical use, as well as stand-alone research studies.
The pace of discovery appears to be accelerating as more scans are launched and pooled. But in the first generation of studies, the tendency has been to lump cancer subtypes into a site-specific entity to find common elements, raising the questions of when and how to interrogate cancer subtypes, such as estrogen receptor–negative breast cancer and how to incorporate newer expression-array definitions of subtype. The current crop of GWAS studies has mostly not collected these types of data. Thus, to integrate the advances in somatic characterization, it will be necessary to fuse efforts largely directed at prognostic studies (eg, expression and/or mutation patterns predictive of prognosis), with etiologic studies focused on inherited germline variation. Similarly, GWAS have also highlighted challenges in gene–environment interactions, most notably the difficulty in determining whether the 15q25 locus and its association with lung cancer is attributable to primary carcinogenesis or to the genetics of smoking, a strong risk factor for lung cancer (9). Clearly, statistically well-powered studies are needed to address gene–gene and gene–environment interactions. The coming years promise to be an exciting time for the genetic epidemiology of cancer, even if they have more of a taste of consolidation, rather than the revolutionary flavor of the past 3 years.

References

Affiliations of authors: Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, Boston, MA (DJH); Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD (SJJC).