Editorial

Toward the Last Cohort

John D. Potter

Fred Hutchinson Cancer Research Center, Seattle, Washington

Two essential elements contribute to the risk of all diseases, especially cancer: genetic variation and environmental exposures. The completion of a solid draft of the human genome will allow researchers to characterize genotypes to ever-finer degrees of detail, and to continue to identify genetic traits associated with disease in individuals and families. However, much of the genetic variation that is associated with cancer seems to modify risk only in the presence of environmental variability, both for exposures that increase, and those that decrease, risk. There are 10- to 200-fold differences in rates of disease across different geographic locations around the world and, over 50 years, up to 10-fold differences when comparing the same geographic location over time. These cannot be explained by differences in genes, but only by differences in exposures, modified by the interaction between genetic variation and such exposures.

Exposures and Genotypes

A broad array of exposures influences the presence or absence of disease in humans. The accurate characterization and measurement of many of the environmental exposures is difficult but there is extensive experience in the epidemiologic community. The human species has adapted to cultures, diets (both marginal and excessive), microorganisms and parasites, toxic exposures, and bad habits, in environments from the equator to the poles. We are an outbred species that, nonetheless, passed through a genetic bottleneck perhaps 150,000 years ago; the degree of fit between phenotype and environment is still undergoing rapid change. Thus, there is a wide variety of genetic susceptibilities to, and protection against, these exposures. It follows that an evaluation, both of variation in environmental exposures and of genetic variation, is essential to establishing the full picture of the causes of disease.

Disease Heterogeneity: A Rose Is Not a Rose Is Not a Rose...

Unlike the increasing ease with which genotypes (considered at small scale or at genomic level) can be generated, characterizing disease phenotypes remains problematic. There are myriad classification schemes that range from precise molecular characterization, such as that of *bcr-abl* leukemia, to vague, very heterogeneous, syndromes, such as schizophrenia. This imprecision and inconsistency is important for three reasons. First, specific molecularly defined subsets of diseases have been shown to be associated with particular exposures. For example, smoking is associated with hyperplastic polyps, in both the presence and absence of adenomas but not with a ''pure adenoma'' phenotype (1) and with MSI+ but not MSI− colon cancer (2). Second, variation in specific molecularly defined subsets of cancers explains, in part at least, variation by time and place. For example, the international variation in breast cancer is explained extensively by variation in ER+/PR+ breast cancer, with risk of other receptor-defined subsets being less varied (3). Third, there exists extensive evidence that molecularly defined subsets of disease also carry very different responses to therapy and prognoses (4, 5).

For each of these reasons, it is essential to work toward a better—molecular—classification of disease, particularly of cancer, rather than continue to rely extensively on histopathology. The key to this problem is the collection of fresh tissue at the time of diagnosis or treatment, ensuring the ability to characterize tumors, especially, using protein profiles or mRNA expression—and in a setting that allows these findings to be related to exposure and genetics on the one hand, and response to therapy and outcome on the other.

The greater the degree of precision of molecular phenotypic classification—and, thus, the greater the homogeneity of disease subsets—the higher the likelihood of being able to reduce susceptibility or increase resistance (prevention), to detect early disease, and, thus, treat at the earliest opportunity. As more outcomes accumulate with time, in a well-characterized and well-followed cohort and as molecular classification schemes improve, there will be an increasing capacity to define and refine homogeneous disease subsets.

Early Detection

Early detection of disease, again especially cancer, is important because, by and large, a disease caught early is treated more readily, more simply, with fewer complications, and with better survival. Markers of an early disease need a variety of characteristics (6). Among the most important is specificity, which in this context, has two aspects: first, distinguishing disease from nondisease, and second, distinguishing disease that will present clinically from disease that remains indolent

Cancer Epidemiol Biomarkers Prev 2004;13(6):895–7

Requests for reprints: John D. Potter, Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104. Phone: (206) 667-4683. E-mail: jpotter@fhcrc.org

beyond the life span of the individual. It is clear that some current early detection methods work well in achieving the first kind of specificity but may be doing more poorly than we thought at the second; a recent article of Zahl et al. (7) provides the best evidence yet that methods for early detection need the capacity not only to detect disease but also to distinguish between disease that will progress and disease that will remain indolent.

Among the best opportunities to develop early detection profiles that achieve this aim is to follow a large number of individuals over a prolonged period, sampling blood and other body fluids at multiple time points to establish, say, protein profiles with predictive power derived from both a cross-sectional picture and change over time (8).

The Last Cohort

To attempt to establish the complete pattern of human disease susceptibility and resistance, to establish longitudinal profiles as early-detection markers, and to identify more precise phenotypes, what is needed is a study of a very large number of ethnically diverse individuals who are well characterized genetically, whose exposures are diverse and well mapped, and whose illness pattern and mortality can be monitored. This cohort, if we do it correctly, could be ''The Last Cohort''; it would examine the impact of exposures on causes and rates of disease and study the interaction of these exposures with genetic variation. Blood samples would be collected from healthy individuals, along with detailed information about each individual—including behaviors (e.g., diet, smoking, exercise), medical history, reproductive history, family history, demographics, and geographic location. Over time, a predictable proportion of these healthy individuals will develop diseases. The blood samples collected from the healthy donors—plus additional exposure data on these donors—will facilitate the search for early markers of disease. The Last Cohort, therefore, represents the opportunity to learn more about the genetic and environmental etiology of tightly defined disease entities; about early detection; about response to therapy and outcome; and, ultimately, about prevention.

The size of the cohort is determined by the degree of human genetic variation, the degree of variation in exposures, and the size of the specific homogeneous sets of outcomes to identify; perhaps 1,000,000 Last Cohort participants would be needed to achieve the desired results. After 6 years of follow-up, for instance, a healthy cohort of this size in the United States, ages 50 to 75, would experience about 80,000 cancers and 110,000 deaths.

This large, long-term study of individuals, characterized using collected data, genomics, and proteomics on pre-diagnostic serum, with the capacity to follow up for outcomes of interest, would answer several questions:

- What is the association between specific exposures and molecularly defined disease risk? (Paradigm: smoking and lung cancer)
- What is the association between specific allele variants and disease risk? (Paradigm: BRCA1 truncation mutation and breast cancer)
- What aspects of the interaction between exposure and genetic variants influence disease risk? (Paradigm: folate/MTHFR variants and colon cancer)
- What aspects of the proteome profile subsequently distinguish those with and without disease? (Paradigm: PSA and prostate cancer)
- What molecular characteristics of host and disease distinguish good response to therapy and good outcome from their opposites? (Paradigm: specific mRNA expression patterns in breast cancer)

The examples given, and the structure of this editorial more generally, are focused on cancer but the arguments can be generalized, *mutatis mutandis*, to all disease, which indeed, is also the proper realm of inquiry for The Last Cohort.

Adults or Children?

It is very attractive to think of recruiting a cohort of children (even *in utero*, or pre-pregnancy; ref. 9) and following them for many decades, taking into account not only genetic variation but also the earliest life exposures, to establish their impact on disease outcome. Unfortunately, there are a variety of problems with such a design. To name two, it will be decades before outcome events accumulate in sufficient numbers unless the cohort is numbered in the tens of millions; and the capacity to follow-up infants, children, adolescents, and young adults in isolation is markedly more difficult than tracking adults.

Nonetheless, there are designs that could incorporate the tempting opportunity presented here, namely by recruiting adults, say 40 to 70, and subsequently their children or grandchildren. What this would allow is first, a cohort with early payoff as well as very long viability; second, a better opportunity to track the more mobile younger group because of multiple contacts in the family; and third, the possibility of incorporating family and genetic linkage studies.

Existing Studies or New Cohort?

The most difficult and expensive part of establishing such a cohort from scratch is collecting fresh specimens for proteomics, mRNA expression, etc., to define cancer and other outcomes as precisely as possible. Most existing cohort studies have no infrastructure to collect fresh tumor specimens. It is likely that doing it piecemeal for each cohort will be particularly complex and expensive. It may be more rational to devise a network of cohort centers that are designed, from the outset, to collect fresh tumors and to incorporate a uniform set of procedures for doing so into the overall design. The proposed NCI National Biospecimen Network and establishing the cohort itself within HMOs may provide the most cost-effective strategy and the best setting for a very large cohort that collects fresh tissue (10).

A second difficulty with agglomerating existing cohorts is the lack of uniformity of data collection procedures. One solution to this, of course, would be to institute a new wave of data collection consistent across

all cohorts. This would be cheaper than starting recruitment from scratch, but would still not readily solve the problem of reducing outcome heterogeneity, using characteristics of fresh tissue.

International Collaboration

The United States may be the best society in the world in which to undertake this study because of the extensive genetic and ethnic variation and mixing, and the wide variety of lifestyles, from the most abstemious to the most sybaritic. On the other hand, other societies also provide significant advantages because they can expand contrasts in genetic variability, in diet and lifestyle, in disease rates, as well as offering greater ease in identifying individuals and following them over time, Accordingly, The Last Cohort is almost certainly best undertaken as an international collaboration.

Access for All

Perhaps the most attractive feature of a new cohort, in addition to establishing it with fresh-tissue collection built-in, is the ability to structure it so that access to the complete data set is widespread across the research community. As with all epidemiologic studies, data collected via questionnaires, etc. would be electronic. More importantly, as the generation of genomic data becomes increasingly cheaper, it is not fanciful to imagine that individual genomes (characterized by, say, 500,000 haplotype-tagging SNPs) will not only be possible across the whole cohort but sufficiently inexpensive that each of these, too, becomes an electronic datum. Similar considerations, albeit on a longer time scale, apply to the data on serum-protein profiles that can provide the basis for widespread early detection, as well as to the molecular profiles of tumors and outcomes following therapy.

Accordingly, access to the Last Cohort data would be widespread and not rely solely on access to study participants, blood, or specimens themselves. This has major advantages for international collaboration, decentralization of specific lab work, and the best use of the data for understanding genes and environment in etiology, developing early detection markers, and tailoring therapy toward both host and tumor phenotype.

We are in some danger of having much of the right technology in place over the next 5 to 10 years to do high-throughput genome sequencing (for susceptibility and resistance) and proteomics (for screening, early detection, and to define homogeneous disease subsets) on very large populations but having no study infrastructure in place to best exploit those gains to understand human disease, its prevention, and its early detection. The last cohort is here proposed to fill that gap.

References

1. Morimoto LM, Newcomb PA, Ulrich CM, et al. Risk factors for hyperplastic and adenomatous polyps: evidence for malignant potential? Cancer Epidemiol Biomarkers & Prev 2002;11:1012-8.
2. Slattery ML, Curtin K, Anderson K, et al. Associations between cigarette smoking, lifestyle factors, and microsatellite instability in colon tumors. J Natl Cancer Inst 2000;92:1831-6.
3. Yasui Y, Potter JD. The shape of age-incidence curves of female breast cancer by hormone-receptor status. Cancer Causes & Control 1999;10:431-7.
4. Armstrong SA, Staunton JE, Silverman LB, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41-7.
5. van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415: 530-6.
6. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93:1054-61.
7. Zahl P-H, Strand BH, Mæhlen J. Incidence of breast cancer in Norway and Sweden during introduction of nationwide screening: prospective cohort study. BMJ 2004;328:921-4.
8. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. Nat Rev Cancer 2003;3:243-52.
9. Brown JE, Potter JD, Jacobs DR, et al. Maternal waist-to-hip ratio as a predictor of newborn size: results of the DIANA project. Epidemiology 1996;7:62-6.
10. Potter JD. Chapter 7: National Biospecimen Network and Public Health. In: Friede A, Grossman R, Hunt R, et al., editors. National Biospecimen Network Blueprint. Durham, NC: Constella Group, Inc.; 2003.