Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs

John D. Potter

Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N. MP-900, Seattle, WA 98109-1024, USA

Different disciplines approach cancer with different study designs, techniques and established bodies of knowledge. This article identifies some established epidemiological data and methods, which are useful for cross-disciplinary molecular and genetic studies of cancer but which are ignored by some researchers. First, the international variation in cancer risk is accounted for extensively by variation in environmental exposures; it is unlikely that even minute characterization of individual genomes will provide the best assessment of cancer risk in the absence of comparably detailed information on the environment. Second, epidemiological study-design methods are sometimes the most appropriate to answer molecular questions, particularly when using techniques such as expression microarrays or proteomics to establish differences among cancer subtypes or biomarkers in the setting of a non-experimental study. In such studies, it is essential to avoid bias, control confounding and undertake accurate replication. Established epidemiological data and methods will contribute to the best use of the new molecular technology.

Epidemiological studies

Epidemiology involves the study of disease patterns and their determinants in populations. It is largely an observational science, like astronomy and evolutionary biology, and thus is distinguishable from experimental sciences. The objects of study are people both with and without disease, unlike most clinical research. Comparisons between these classes require calculation of risks (i.e. the probability of disease given exposure) and rates (i.e. the frequency of disease per unit of population per unit time). The most common of these measurements is the rate ratio, also called the relative risk (i.e. the rate of disease in the exposed divided by the rate of disease in the unexposed).

Epidemiology flourished initially in infectious disease research, using field methods (data collection in specified populations) and laboratory methods, and studying affected and unaffected individuals to determine the time and place of exposure to the disease vector, and characterizing the biology of the disease and causative microorganism. As epidemiological methods were applied to chronic diseases, measurable intermediate biology (e.g. serum lipids in cardiovascular disease) also enabled field and laboratory scientific approaches to be integrated.

Initially, cancer epidemiology did not take this road for several reasons: cancer is a collection of diseases with different causes and biology and there were no easily measured and relevant intermediate markers. Furthermore, studies of lung cancer identified an exposure – smoking – that was associated with a large relative risk [1], encouraging the belief that identifying the causes of cancer would not be difficult. Demonstrating that radiation markedly increased cancer risk further reinforced this belief. It has so far proved illusory for most cancers, although other potent causes [e.g. specific viruses, such as hepatitis B virus (HBV) and human papillomavirus (HPV)] have been identified.

Unlike clinical trials, epidemiology is an observational science: the researcher does not allocate those being studied to exposed and unexposed conditions. Thus, other characteristics might correlate with a specific exposure. Also, because people make choices about participation, the study groups (those with and without disease or exposed individuals versus unexposed) might or might not be comparable with the wider population about which we wish to make inferences, which can bias estimates of risk. Solutions to these study-design problems are central to good epidemiological practice – and ultimately to the use of specific techniques that need observational methods, including the use of microarrays.

Cancer arises from multiple genetic and epigenetic changes in the genome. High-throughput techniques that rapidly characterize biological material are increasingly
available. Accordingly, there is a growing capacity to define at a molecular level, in substantial numbers, the susceptibility, intermediate processes and subsets of cancers themselves. As a result, sciences that developed independently, epidemiology and genetics in particular, now have overlapping interests. Hence, the need for improved cross-talk between the communities.

**Genes and environment**

Cancer is a disease of the genome, however, in most cases, it is somatic DNA damage that is the hallmark of cancer — genes are usually the target, not the origin, of the cancer process. Individuals, in whom the cause of the disease lies in damaged germline DNA, constitute only a small proportion of all cancers (considerably <10%) and there is a large body of evidence suggesting that environmental variation determines the patterns and pattern changes in cancers that are observed worldwide.

Doll first discussed the implications of the marked variation in the international incidence of cancer [2]; it varies by tenfold to >200-fold. For example, around 1990 breast cancer incidence varied ∼14-fold between Kangwha, Korea (7.1 per 10^5 individuals) and Los Angeles, California (103.7 per 10^5); the incidence of melanoma varied ∼165-fold between Osaka, Japan (0.2 per 10^5) and New South Wales, Australia (33.1 per 10^5) [3]. Figure 1 shows the variation in colon cancer incidence around the world by age 70. One consequence of increasing specialization in the study of cancer is that this massive variation is ignored, or worse not recognized, by many.

Furthermore, the incidence of several cancers has changed markedly, both up and down, over the course of the 20th century. For example, in the UK, mortality from lung cancer per 10^5 men rose from approximately two per 10^5 individuals in 1911–1915 to a peak of 102 per 10^5 in 1966–1967 [4]. In marked contrast, stomach cancer mortality in men and women in the USA has fallen from 28.8 per 10^5 in 1930 to around 3.5 per 10^5 in 2000 [5,6]. Whereas the cause of the lung cancer epidemic is known, the causes of the decline of stomach cancer are less clear, although the declining prevalence of infection with *Helicobacter pylori* is central to this reduction. Finally, there are instances in stable populations, of rates of a specific cancer changing rapidly. For example, the incidence of colorectal cancer in men in Miyagi, Japan, has risen almost tenfold in <50 years; Japan now has the highest rate in the world (Figure 2).

The first observation — the large international variation — is compatible with either genetic or environmental explanations. The second observation — the rapid change — is difficult to explain by genetics alone. What makes the choice between the gene and environment explanations easier are the data on migrating populations. Rates of cancer in migrants always approximate those of the host country, often within the generation of those migrating, and always within two subsequent generations [7–17].

Taken together, these data establish a dominant role for environmental factors in the causation of cancer. Any explanation for the aetiology of cancer must take this into account, even as we seek the interesting biology that is inherent in genetically determined risk or, more importantly, as we explore the genetic variation that distinguishes those susceptible from those resistant to the environmental exposures. It is unlikely that even the minute characterization of individual genomes will provide a reliable prediction of cancer risk in the absence of comparably detailed information on exposures.

![Figure 1](http://tigs.trends.com)
appropriate design for true experiments [23] and to Chipping Forecast II [22], careful thought is given to described. In the recent guide to the state of the art The experiments when both observations and experiments are being primary header again refers only to microarray exper-
mected.org/Workgroups/MIAME/miame.html), where the (MIAME) now required by many journals ( http://www.
microarray gene expression data’ (MGED) ontology[21], and experiment but still has as the superclass which acknowledges the distinction between observation and experiment. To make the difference absolutely clear, no experiment is being conducted, exposures are not difference between cancer and non-cancer tissues. In this expression profiles to identify subsets of cancers or the between two stars, no one would make the error of calling

Study design – experiments and observations
Molecular genetics converges with epidemiology, particularly in studies that use microarrays. Expression micro-
arrays enable comparisons between two conditions across tens of thousands of specific mRNAs. There are two different ways in which expression microarrays are used. First, they are central to experiments designed to understand cell behaviour under different conditions, including establishing patterns of gene expression during the cell cycle and identifying related sets of upregulated and downregulated genes following treatment with specific agents [18–20]. Second, they are used to compare expression profiles to identify subsets of cancers or the difference between cancer and non-cancer tissues. In this case, no experiment is being conducted, exposures are not being assigned randomly to two identical model systems: rather, observations are being compared.

Although these are not ‘experiments’, they are consistently described that way, even in the now well-accepted ‘microarray gene expression data’ (MGED) ontology [21], which acknowledges the distinction between observation and experiment but still has as the superclass descriptor: ‘ExperimentDesign’ (http://mged.sourceforge.
et/ontologies/MGEDontology.php) and in the related ‘minimal information about a microarray experiment’ (MIAME) now required by many journals (http://www.
mged.org/Workgroups/MIAME/miame.html), where the primary header again refers only to microarray exper-
iments when both observations and experiments are being described. In the recent guide to the state of the art The Chipping Forecast II [22], careful thought is given to appropriate design for true experiments [23] and to validation [24], but none to differentiating observation from experiment. To make the difference absolutely clear, if an astronomer compares the Fraunhoffer spectra between two stars, no one would make the error of calling this an experiment – this is a comparison of observations. The importance of this, in the biological setting, is that the study-design methods that should be employed for the comparison of observations are those of epidemiology, not experimental biology – therefore, confounding, bias and reproducibility are relevant.

Confounding
Confounding addresses a problem that is inherent in seeking causal interpretations: that of mistaking potential causes with factors that are associated both with a real causal factor and the disease itself. It is possible to formally identify confounding in a study by noting a change in the risk ratio, following the addition of another variable to the analysis. For example, a risk ratio for vegetable consumption in relation to lung cancer might be 0.3 (this indicates a threefold reduction in risk). However, after the inclusion of smoking data in the analysis, this becomes 0.5 – a halving of the risk. Thus, although the consumption of vegetables is truly associated with a reduced risk, this reduction is not as large as the analysis that ignores smoking suggests. The difference in estimates occurs because smokers eat fewer vegetables than non-smokers; that is, the relationship between vegetable consumption and lung cancer is confounded by smoking (Figure 3a,b).

In many microarray studies, investigators appear to forget that they have not used an experimental design [25] and ignore the possibility that the differences they see in

Figure 3. Confounding and bias in observational studies using cDNA microarray and proteomics. (a) The inference, in an observational study, that a particular association with an outcome is causal or biologically meaningful must be tem-
pered by the possibility that the association is the consequence of confounding; that is, the variable of interest might be statistically associated with both the real causal variable and the disease of interest. (b) The possible protective effect of veg-
etables against a cancer is explained in part by the fact that smokers consume fewer vegetables than non-smokers – see main text. (c) The inference that the difference in expression-array patterns might be biologically meaningful could be suspect if the cancer cases and normal controls have different age, sex, distri-
butions (confounding) or if the subjects and samples have been handled differ-
ently (bias) – see main text.
expression profiles between one set of cancers and another might be due to unmeasured confounding by age, sex (the largest genetic differences between humans are those between the sexes [26,27]), or other exposures (Figure 3c).

Indeed, observational studies are common in which tumour samples, often of unknown provenance, are assembled haphazardly, analyzed for some biological characteristics, and reported on as if they were representative of all tumours [28–30].

Uncontrolled confounding is just one of the consequences of failing to distinguish between observational and experimental studies. In a standard in vitro experiment, cells are treated or remain untreated and the expression profiles are compared. This design enables strong inferences to be made about treatment being the cause of the differences in outcome. In a clinical trial (also an experimental design), groups of individuals are assigned randomly to treatment A versus treatment B. Inferences regarding causality are strong only if a crossover design is used, with each subject acting as their own control; even if subjects are matched for sex, age, and other factors, the likelihood that other inherited or acquired differences might influence expression profiles is high and can be unpredictable from one study to another.

In a cross-sectional observational study, tissue from, for example, alcohol drinkers is compared with that from non-drinkers. A simple attribution of differences in expression patterns between groups to differences in drinking is inappropriate if, for example, the differences are a result of more of the drinkers being smokers; that is, if there is confounding. In the in vitro counterpart of this study design, cancer samples, for example, are compared with normal tissue (and sometimes, as noted, ill-described as an experiment). Attributing differences in gene expression to differences in disease state is inappropriate in the absence of data regarding the age, sex, genetic profile, histology and treatment of the person from whom the sample came. This involves not just the failure to control confounding, but often the failure even to measure any of the relevant confounders. Even if unaffected tissue from the same patient is used as a comparison, there are still problems because of field defects (i.e. changes across a whole field of cells, either inherited, such as germline mutations in BRCA1, or acquired, such as Barrett esophagus), as a result of which, non-cancer tissue is more like cancer tissue than normal tissue [31].

**Bias**

Confounding is the consequence of the clustering of specific behaviours and characteristics in nature. Confounding is not an error of study design. Bias, by contrast, arises in observational studies because study-design is not optimized. An obvious example is treating samples from cases and controls differently: for example, all case samples are collected in the operating room and sit at room temperature for several hours; all control samples are collected in the laboratory and frozen immediately. In these circumstances, the proteome differences seen between cases and controls are the consequence of handling differences as well as biological ones (Figure 3c).

In observational microarray studies, as in other observational studies, another kind of bias – selection bias – is also a problem, irrespective of the strength of the signal, lack of confounding or study size. Selection bias results from the failure to collect all, or a well-defined random subset, of the cases of interest. Such a procedure means that the samples collected do not include all of the possible subsets, or they include them in the proportions that do not represent the ‘universe’ of all such tumours. Many published microarray studies fail the test of proper study design. The use of haphazardly (as opposed to randomly) collected samples to reveal distinguishable gene-expression profiles works reasonably well if there are only a small number of qualitative on-or-off patterns to be detected. The distribution frequency of these in the population will not be revealed but each individual sample can be described and the complete spectrum of expression patterns found. If there are large numbers of expression phenotypes that are being explored in a small number of samples, neither the full set of expression patterns nor their distribution frequency will be identified.

To reduce the likelihood of selection bias, tissue samples from every case of a particular tumour type for a specified time period might be collected at one centre. However, if this institution is a tertiary referral hospital, these samples might still not be representative of the wider universe of all cancers of this classification, reducing the generalizability of the findings.

Although some of the most common patterns for specific diseases might be emerging, we clearly do not have a good handle on the distribution frequency.

It might be argued that, when signals are clear and consistent, confounding and bias are unlikely to explain the observed difference. This is a strong point but it is undermined when different studies of the same phenomena are not consistent [32,33] and when the patterns are not reproducible.

**Reproducibility**

In observational studies with few biological samples and large numbers of sample observations, what would be surprising would not be finding ‘statistically significant’ differences between the sets being compared, but rather failing to find such differences. Although it seems likely that there are biologically meaningful data to be derived from such comparisons, one of the central tenets in epidemiological practice for decades has been the reproducibility of findings from one study to another. An obvious point to make here is that expression patterns might strongly distinguish between two sets of cancers, whether identified by prior classification [hypothesis-driven; e.g., what are the mRNA-expression differences between estrogen receptor positive (ER +) and ER – breast cancers] or allowing the patterns to emerge from the data (i.e. discovery-driven) and still be the result of chance and multiple comparisons. There are a variety of established and new statistical techniques to deal with this issue but these only help us to estimate and delimit the element of chance; they enable probabilistic not deterministic conclusions. Reproducibility is central to drawing causal inferences from both experimental and observational science.
Accordingly, when a pattern of differences is defined, using microarray technology or a proteomics approach, such as mass spectrometry, the resulting pattern should become the hypothesis for a second comparison. This could occur within the framework of the same study, using split but not overlapping samples [34,35]. Alternatively, the results of one study can be tested in a separate study and population [36]. In the absence of such an approach, the interpretation is more problematic [37]. We have proposed elsewhere a structured approach for the development of biomarkers more generally, using both observational and experimental (clinical trial) methods [38].

Applying what we know
The reason for the failure to readily distinguish between observational and experimental designs might be that, although observational scientists are trained in experimental methods, the reverse is seldom true [39]. Nonetheless, scientists know how to draw a random sample from populations; it is known that there are variables that enable a better definition and classification of specimens – factors that influence risk, pathology, and prognosis, and that allow greater insight into the patterns that these observational microarray expression studies reveal. The observations are subject to all the difficulties that induce bias, confounding and imprecision in epidemiological studies: what is the origin of the samples? How Comparable are those individuals with and without disease? What statistical power exists to make inferences?

Conclusions
There are data that provide us with clear evidence that the greatest variation in cancer risk – most obvious in the comparisons across, rather than within, countries – is accounted for by a variation in environmental exposures. Because the major role of genetic variation in relation to cancer is to modify susceptibility to environmental exposure at the individual level, it seems unlikely that even minute characterization of the individual genome will provide a good prediction of risk in the absence of comparably detailed information on environmental exposure.

We need to understand the difference between observational and experimental sciences. When observational methods are the most appropriate to answer a biological question, the relevant study design and methods of data analysis are those of epidemiology, not of experimental biology: we need to avoid bias, measure confounders and ensure reproducibility.

For other observational sciences, advances have occurred as a result of improvements in technology. As astronomy evolved from naked-eye observations, through the early telescopes of Galileo, to tools that facilitate observations across a wide spectrum of wavelengths, understanding of the universe has increased exponentially. The tools of genomics and proteomics are likely to provide similar growth in understanding the epidemiology of cancer. However, these tools will provide only part of the required data. Genotypes will be more finely described but the history of exposure and better characterization of phenotypes are also essential; for these, we will continue to rely on clinical and epidemiological methods, again augmented with the new biology.

References

http://tigs.trends.com
40 Doll, R. et al. (1966) Cancer Incidence in Five Continents (Vol. 1), International Union Against Cancer

Articles of interest in Trends and Current Opinion journals

Turning germ cells into stem cells
Peter J. Donovan and Maria P. de Miguel
Current Opinion in Genetics and Development 13, 463–471

The microRNA world: small is mighty
Peter Nelson, Marianti Kiriakidou, Anup Sharma, Elsa Maniataki and Zissimos Mourelatos
Trends in Biochemical Sciences 28, 534–540

Functional genomics in virology and antiviral drug discovery
Victor DeFilippis, Camilo Raggio, Ashlee Moses and Klaus Früh
Trends in Biotechnology 21, 452–457

DNA-based methods for pedigree reconstruction and kinship analysis in natural populations
Michael S. Blouin
Trends in Ecology and Evolution 18, 503–511

Transcriptionally targeted gene therapy to detect and treat cancer
Lily Wu, Mai Johnson and Makoto Sato
Trends in Molecular Medicine 9, 421–429

Cholesterol-receptor-mediated genomics in health and disease
Deepak Kaul
Trends in Molecular Medicine 9, 44–449

Transcription factors controlling the beginning and end of B-cell differentiation
Kristen Johnson and Kathryn Calame
Current Opinion in Genetics and Development 13, 522–528

Alternative splicing and proteome diversity in plants: the tip of the iceberg has just emerged
Kemal Kazan
Trends in Plant Science 8, 46–471

Genes that shape the course of ageing
Thomas B.L. Kirkwood
Trends in Endocrinology and Metabolism 14, 34–347

Genes within genes within bacteria
Christopher Ellis and James W. Brown
Trends in Biochemical Sciences 28, 521–523