Editorial

Complementing the Genome with an ''Exposome'': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology

Christopher Paul Wild

Molecular Epidemiology Unit, Centre for Epidemiology and Biostatistics, Leeds Institute of Genetics, Health and Therapeutics, Faculty of Medicine and Health, University of Leeds, Leeds, United Kingdom

The sequencing and mapping of the human genome provides a foundation for the elucidation of gene expression and protein function, and the identification of the biochemical pathways implicated in the natural history of chronic diseases, including cancer, diabetes, and vascular and neurodegenerative diseases. This knowledge may consequently offer opportunities for a more effective treatment and improved patient management. Genetic research of this kind captures the public imagination in a positive manner and attracts political attention. For example, the 2003 UK Government White Paper on genetics (1), entitled ''Our inheritance, our future: realising the potential of genetics in the National Health Service,'' highlighted the opportunities for tailored drug treatments and gene therapy flowing from the sequencing and mapping of the human genome. However, it was notable that this influential document also drew attention to the use of genetic testing to identify individuals at greatest susceptibility to chronic diseases and the possibilities this raises for disease prevention. Application of genetics in this context moves away from the clinic towards the population and the observational methods of epidemiology. The fruit of this research in terms of reduced morbidity and mortality is therefore to be predominantly realized through public health measures.

It is well known that the majority of genetic variants or single nucleotide polymorphisms in the human genome are of low penetrance, including genes implicated in metabolism of environmental chemicals, immunity, lipid metabolism, and hemostasis among others. The high prevalence of these single nucleotide polymorphisms means that despite their low penetrance, they may substantially contribute to population disease burden (2). Nevertheless, the majority will do so only in the presence of specific environmental exposures that in themselves are of low penetrance. Environmental exposures are acknowledged to play an overwhelmingly important role in those common chronic diseases mentioned above, which constitute the major health burden in economically developed countries (3, 4). Despite this, many exposure-disease associations remain ill defined and the complex interplay with genetic susceptibility is only beginning to be addressed. This raises the question as to whether fundamental knowledge about genetics will improve understanding of disease etiology at the population level.

The new generation of mega-cohort studies, including the UK Biobank or similar proposed US and Asian cohorts (5-8), provides the framework for such investigations of genetic variation, environment, lifestyle, and chronic disease. At the same time, they represent substantial investment. For example, UK Biobank will recruit half a million people at a cost of around £60 million ($110 million) in the initial phase. The proposal to establish a ''Last Cohort'' of 1 million people in the United States (7) or a similar-sized Asian cohort (8) would presumably exceed this sum. In each case, the high cost is heavily influenced by the collection and banking of biological material. This expense is predicated on the assumption that biochemical and molecular measures on this material will resolve the etiologic questions alluded to above. It is self-evident that unraveling of complex environmental and genetic aetiologies demands that both environmental exposures and genetic variation are reliably measured. Advances in statistical methods and in bioinformatics in relation to large data sets are also of critical importance.

The development of biomarkers is one avenue to improved measurement of exposure, susceptibility (including genetic), and disease in population studies, and the field of molecular cancer epidemiology is founded on this trinity of biomarker categories (9). The characterization of disease can be improved by modern histopathology, particularly the application of sophisticated molecular markers to subcategorize tumors to homogenous groups within which etiology or therapy can be better defined (10, 11). In parallel to advances in disease classification, however, the balance of efforts to improve the measurement of genetic and environmental risk factors has dramatically changed over the last 15 years in a way that has been less than ideal.

The ease of genotyping following the introduction of the PCR in the late 1980s and 1990s saw a major shift in effort and resources, away from development of accurate exposure biomarkers towards the conduct of gene-disease association studies. Allied to the relatively trivial technical demands of genotyping, these measures, by consequence of the fact that they were unchanging over time, were well suited to classic case-control study design. For laboratories involved in molecular cancer epidemiology, gene-disease association studies offered rapid gains in research output. The literature is now replete with meta-analyses of these data. The studies that have been conducted have, by some accounts, yielded only a modicum of success with relatively few reproducible findings (see for example ref. 12). More recently, improvements in study design have been suggested, notably by increasing subject numbers and by analyzing multiple polymorphisms, of functional relevance (13). A more comprehensive coverage of the genome and the possibility to examine the interplay between single nucleotide polymorphisms are now feasible through the application of microarray technology (14). It is predictable that as costs decrease, there will emerge analyses of existing studies on a grander scale. The consequence may not be greater clarity but a greater number of chance findings and an increasing difficulty of dealing with the sheer volume of data in the absence of parallel advances in data analysis. Things may get worse before they get better.

Cancer Epidemiol Biomarkers Prev 2005;14(8):1847–50

Grant support: National Institute of Environmental Health Sciences (USA) grant no. ES06052.

doi:10.1158/1055-9965.EPI-05-0456

Partially as a consequence of the emphasis on genotyping, the accurate assessment of many environmental exposures remains an outstanding and largely unmet challenge in cancer epidemiology. As measurement of one half of the gene:environment equation continues to be refined, the other remains subject to a large degree of misclassification (15). In this imbalance, we risk adopting the image of the male fiddler crab (*Uca pugnax*), a species characterized by one huge well-developed claw (growing to 65% of its total body weight) waved proudly to attract a mate, but contrasting almost comically with its second, apparently underdeveloped one; the analogy breaks down of course in that the claws of the fiddler crab are adapted to purpose.

The imbalance in measurement precision of genes and environment has consequences, most fundamentally in compromising the ability to fully derive public health benefits from expenditure on the human genome and the aforementioned cohort studies. There is a desperate need to develop methods with the same precision for an individual's environmental exposure as we have for the individual's genome. I would like to suggest that there is need for an ''exposome'' to match the ''genome.'' This concept of an exposome may be useful in drawing attention to the need for methodologic developments in exposure assessment.

At its most complete, the exposome encompasses life-course environmental exposures (including lifestyle factors), from the prenatal period onwards. Developing reliable measurement tools for such a complete exposure history is extremely challenging. Unlike the genome, the exposome is a highly variable and dynamic entity that evolves throughout the lifetime of the individual. It is not without good cause that progress has been limited in meeting this goal. However, the methodologic challenge may not be more daunting than the one faced two decades ago of investigating an estimated 10 million single nucleotide polymorphisms in the human genome. In addition, as with the genome, even a partial, targeted understanding of exposure can provide substantial advances; a prime example is the contribution of aflatoxin exposure biomarkers to the assignment of hepatocellular carcinoma risk in developing countries (16). Of course, biomarkers will only be part of the solution and will need to be coupled with increasingly refined questionnaire-based approaches. Nevertheless, it is the prospective cohort study design that best suits the exposure biomarker approach, providing as it does opportunities for repeat sampling to enable a broader timeframe of exposure assessment, coupled to the avoidance of reverse causation by the collection of samples in advance of disease onset.

Infectious disease epidemiology has extensively employed exposure biomarkers. In the case of infections and cancer, developmental work to establish validated laboratory assays for antibodies to viral or bacterial antigens (e.g., hepatitis viruses, human papilloma virus, and *Helicobacter pylori*) has been central to understanding the etiologic role of these agents in epidemiologic studies. For example, in the case of *H. pylori* and gastric cancer, only 11 years elapsed between identification of the organism in 1983 and its classification as a human carcinogen by the IARC in 1994 (17). Similarly, in the case of hepatitis B virus and liver cancer, it was the availability of a sensitive and specific marker of infection (hepatitis B surface antigen) that transformed the epidemiology that had earlier relied on less precise markers of exposure, such as clinical history. These examples illustrate the value of precise biomarkers of exposure in epidemiology. The first generation of exposure biomarkers in molecular cancer epidemiology was predominantly derived from the paradigm of chemical carcinogenesis. This led to emphasis on genotoxic chemicals and methods to measure carcinogen-DNA adducts or carcinogen-protein adducts (as surrogates for DNA adducts) and carcinogen metabolites in body fluids (18, 19). There has been less emphasis on biomarkers of exposure in light of other nongenotoxic, mechanistic paradigms, although many environmental risk factors act through such mechanisms (e.g., receptor-mediated effects or alterations in cell proliferation).

The emphasis on DNA and protein adducts has undoubtedly contributed substantially to establishing the biological plausibility of exposure-disease associations. Examples include investigations of the association between genotypes for carcinogen-metabolizing enzymes and adduct levels, or the use of biomarkers as modifiable end points in short-term intervention studies (20). In contrast, the application of exposure biomarkers of this nature to etiologic studies is far more limited. Aflatoxin is perhaps the prime example in which the exposure biomarker permitted categorization of this environmental agent as a human carcinogen (21, 22), others include polycyclic aromatic hydrocarbon-DNA adducts in lung cancer (23) and arylamine-hemoglobin adducts in bladder cancer (24). Valuable progress is still being made in this area and some chemical-specific markers have the advantage of combining information on exposure with insights into individual susceptibility and mechanistic pathways (25). However, even accepting that researchers in the exposure arena might have been enticed away to genotype, it is still striking that the number of successful applications of environmental exposure biomarkers into etiologic studies is relatively modest. In addition to the call for more commitment, there is also a need for novel approaches.

In this context, it is pertinent to ask whether the new ''omics'' technologies of transcriptomics, proteomics, and metabonomics can help unlock the problem of environmental exposure assessment. Currently, these methods are mainly applied to the understanding of disease mechanisms and diagnosis. Their validity for exposure assessment primarily revolves around whether specific exposures will be reflected by altered levels of mRNA, proteins, or metabolites. Will the ''omics'' with their thousands of composite parts offer specific signatures or fingerprints of environmental exposures across a broad spectrum of mechanisms of action, both genotoxic and nongenotoxic? Is this new technology in a position to permit a step-change in exposure assessment?

There are indications that this is a fruitful area of research, albeit one that to date is relatively unexplored. For example, with respect to mRNA expression, studies of naturally occurring or industrial compounds with estrogen activity do alter the expression of similar genes *in vitro*. A panel of 172 genes were selected to create a customized DNA microarray that responded to exposure to this class of compounds in a cell model system (26). *In vitro* ionizing irradiation of human lymphocytes induced alterations in expression of specific genes (27). Subsequent studies in lymphocytes of patients undergoing whole body irradiation revealed specific patterns of gene induction and some of the affected genes were the same as those identified from the *in vitro* studies (28). Microarray data were confirmed by reverse transcription-PCR for specific genes, illustrating the way in which initial screening can be translated to low-density arrays of gene subsets for more routine analysis. It was noteworthy that some genes responded to the first dose of radiation but not the second, and there was considerable heterogeneity in response between patients; both these observations are important when thinking of application to environmental exposures. Application of DNA microarray in a population setting was shown when a panel of 36 candidate reporter genes were identified that were able to discriminate between smokers and nonsmokers (29). Importantly, expression of the reporter genes was not correlated with potential confounding factors such as alcohol, aspirin use, and vegetable intake. As with the study of ionizing radiation, analyses of mRNA were done in peripheral blood cells, suggesting that such technology might be applicable at the population level.

Both these studies provide some encouragement to the strategy of identifying panels of genes responsive to specific exposures.

Proteomics, like transcriptomics, is being primarily explored with a view to disease diagnosis and prognosis, identifying panels of proteins that permit those with and without disease to be discriminated. Currently, applications to human exposure assessment do not seem to be common although investigations *in vitro* with mobile-phone radiation have been reported (30).

Examination of the small molecule metabolites that constitute the metabolome also offers opportunities to address exposure. Metabolomics currently employs ^{1}H nuclear magnetic resonance or liquid chromatography coupled to mass spectrometry, and it is estimated that the metabolome comprises of the order of 3,000 major metabolites (31). One approach is to simply scale up a *metabolic profile* by increasing the number of specific known metabolites analyzed for a given chemical or class of chemical exposure. The alternative is to use the metabolome to identify altered patterns of metabolites, a *metabolic fingerprint*, in analogous fashion to the gene and protein expression arrays discussed above. An example comes from a study of mice exposed to *Schistosoma mansoni* in which differences in urinary metabolite fingerprints were obtained (32). The data pointed to specific pathways (e.g. glycolysis and amino acid metabolism), being affected by infection, illustrating the potential to also obtain information on disease mechanisms using these methods. Of particular interest is a recent application to individuals who changed from a non-soy to a soy-containing diet (33). As with the patients subject to irradiation mentioned earlier, there was considerable interindividual and temporal heterogeneity in the plasma metabolome and this initially obscured any effect of the dietary change. Principal component analysis was used (as in the mouse study) to examine the intra-individual variation with consumption of the two diets, and by this method differences in plasma profiles were seen. Following further spectral filtering of the data to control for interindividual variation, effects of the soy diet on plasma metabolome began to emerge, including changes in metabolites relating to the lipoprotein profile. The potential application of ''omics'' technologies to characterize dietary exposures (34) therefore receives some support from this early investigation.

This new generation of technologies requires extensive research before judging whether they can provide the step-change that is required in human exposure assessment. It remains to be seen whether, in principle, mRNA, protein, or metabolite expression can be specific and sensitive enough to define exposures at low levels in human populations. It will be important to understand whether complex mixtures or families of chemicals act through the same pathways and can be represented by common targets at the mRNA, protein, or metabolite level on common mechanistic pathways. However, the dynamic nature of each of these systems may militate against long-term exposure assessment, unless some of the changes prove stable over time. In addition to the necessity for proof of principle, there is also the need for considerable investment in technology with an eye from the outset to their eventual application to large numbers of samples, of often limited quantity. Purification procedures in the case of metabolomics and proteomics will be essential to measure rarer, possibly more informative, proteins or metabolites among the background of quantitatively more dominant species. Although not explicitly discussed here, the need for sophisticated statistical analysis emerges as crucial to any eventual application. As with the earlier generation of exposure biomarkers, a carefully planned strategy, starting with model systems and small-scale human studies, is likely to be most successful (20).

In conclusion, by their nature, prospective cohort studies take time as well as money; given the challenges outlined here, some of this time would seem to be well spent in developing reliable exposure assessment tools. The concept of an exposome may serve to highlight this requirement and to balance the effort going towards characterization of the genome. An extension of the current generation of biomarkers, together with an evaluation of the new generation of ''omics'' technologies, has a crucial role to play in this regard (35). However, advances will require increasing collaboration between epidemiologists, biostatisticians, experts in bioinformatics, and laboratory and environmental scientists. In addition, funding agencies must take a medium- to long-term view and encourage research that focuses on improved measures of environmental risk factors, an area that currently seems to be less of a priority for support than many others in the broad domain of medical research.

Acknowledgments

The author would like to thank his colleagues David Forman, Janet Cade, Mark Gilthorpe, and Tricia McKinney for their comments on the text.

References

1. Our inheritance, our future: realising the potential of genetics in the National Health Service http://www.dh.gov.uk/assetRoot/04/01/92/39/04019239.pdf.
2. Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. Lancet 2001;357:709–12.
3. Peto J. Cancer epidemiology in the last century and the next decade. Nature 2001;411:390–5.
4. Cappuccio FP. Commentary: epidemiological transition, migration, and cardiovascular disease. Int J Epidemiol 2004;33:387–8.
5. Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475–7.
6. Barbour V. UK Biobank: a project in search of a protocol? Lancet 2003;361:1734–8.
7. Potter JD. Toward the last cohort. Cancer Epidemiol Biomarkers Prev 2004;13:895–7.
8. Cyranoski D, Williams R. Health study sets sights on a million people. Nature 2005;434:812.
9. Schulte PA, Perera FP. Molecular epidemiology: principles and practice. San Diego, CA, USA: Academic Press, Inc.; 1993.
10. David RE, Staudt LM. Molecular diagnosis of lymphoid malignancies by gene expression profiling. Curr Opin Hematol 2002;9:333–8.
11. 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.
12. Caporaso NE. Why have we failed to find the low penetrance genetic constituents of common cancers? Cancer Epidemiol Biomarkers Prev 2002;11:1544–9.
13. Rebbeck TR, Martinez ME, Sellers TA, Shields PG, Wild CP, Potter JD. Genetic variation and cancer: improving the environment for publication of association studies. Cancer Epidemiol Biomarkers Prev 2004;13:1985–6.
14. Kennedy GC, Matsuzaki H, Dong SL, et al. Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233–7.
15. Vineis P. A self-fulfilling prophecy: are we underestimating the role of the environment in gene-environment interaction research? Int J Epidemiol 2004;33:945–6.
16. Wild CP, Turner PC. The toxicology of aflatoxins as a basis for public health decisions. Mutagenesis 2002;17:471–81.
17. IARC. Monographs on the evaluation of carcinogenic risks to humans: schistosomes, liver flukes and *Helicobacter pylori*. Vol. 61. Lyon, France: IARC; 1994.
18. Poirier MC. Chemical-induced DNA damage and human cancer risk. Nat Rev Cancer 2004;4:630–7.
19. Hecht SS. Tobacco carcinogens, their biomarkers and tobacco-induced cancer. Nat Rev Cancer 2003;3:733–44.
20. Groopman JD, Kensler TW. The light at the end of the tunnel for chemical-specific biomarker: daylight or headlight? Carcinogenesis 1999;20:1–11.
21. Qian GS, Ross RK, Yu MC, et al. A follow-up study of urinary markers of aflatoxin exposure and liver cancer risk in Shanghai, Peoples Republic of China. Cancer Epidemiol Biomarkers Prev 1994;3:3–10.
22. Wang LY, Hatch M, Chen CJ, et al. Aflatoxin exposure and risk of hepatocellular carcinoma in Taiwan. Int J Cancer 1996;67:620–5.
23. Tang DL, Phillips DH, Stampfer M, et al. Association between carcinogen-DNA adducts in white blood cells and lung cancer risk in the physicians health study. Cancer Res 2001;61:6708–12.
24. Gan JP, Skipper PL, Gago-Dominguez M, et al. Alkylaniline-hemoglobin adducts and risk of non-smoking-related bladder cancer. J Natl Cancer Inst 2004;96:1425–31.
25. Carmella SG, Chen M, Yagi H, Jerina DM, Hecht SS. Analysis of

phenanthrols in human urine by gas chromatography-mass spectrometry: potential use in carcinogen metabolite phenotyping. Cancer Epidemiol Biomarkers Prev 2004;13:2167–74.
26. Terasaka S, Aita Y, Inoue A, et al. Using a customized DNA microarray for expression profiling of the estrogen-responsive genes to evaluate estrogen activity among natural estrogens and industrial chemicals. Environ Health Perspect 2004;112:773–81.
27. Amundson SA, Do KT, Shahab S, et al. Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. Radiat Res 2000;154:342–6.
28. Amundson SA, Grace MB, McLeland CB, et al. Human *in vivo* radiation-induced biomarkers: gene expression changes in radiotherapy patients. Cancer Res 2004;64:6368–71.
29. Lampe JW, Stepaniants SB, Mao M, et al. Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 2004;13:445–53.
30. Nylund R, Leszczynski D. Proteomics analysis of human endothelial cell line EA.hy926 after exposure to GSM 900 radiation. Proteomics 2004;4:1359–65.
31. Dettmer K, Hammock BD. Metabolomics — A new exciting field within the ''omics'' sciences. Environ Health Perspect 2004;112:A396–7.
32. Wang YL, Holmes E, Nicholson JK, et al. Metabonomic investigations in mice infected with *Schistosoma mansoni*: an approach for biomarker identification. Proc Natl Acad Sci U S A 2004;101:12676–81.
33. Solanky KS, Bailey NJC, Beckwith-Hall BM, et al. Application of biofluid H-1 nuclear magnetic resonance-based metabonomic techniques for the analysis of the biochemical effects of dietary isoflavones on human plasma profile. Anal Biochem 2003;323:197–204.
34. Davis CD, Milner J. Frontiers in nutrigenomics, proteomics, metabolomics and cancer prevention. Mutat Res 2004;551:51–64.
35. Weis BK, Balshaw D, Barr JR, et al. Personalized exposure assessment: promising approaches for human environmental health research [cited 2005 March 3]. doi:10.1289/ehp.7651. Available from: http://dx.doi.org/.