Will Investments in Large-Scale Prospective Cohorts and Biobanks Limit Our Ability to Discover Weaker, Less Common Genetic and Environmental Contributors to Complex Diseases?

Morris W. Foster[1] and Richard R. Sharp[2]

[1]Department of Anthropology, University of Oklahoma, Norman, Oklahoma, USA; [2]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas, USA

Increasing the size of prospective cohorts and biobanks is one approach to discovering previously unknown contributors to complex diseases, but it may come at the price of concealing contributors that are less common across all the participants in those larger studies and of limiting hypothesis generation. Prospective cohorts and biobanks constitute significant, long-term investments in research infrastructure that will have ongoing consequences for opportunities in biomedical research for the foreseeable future. Thus, it is important to think about how these major additions to research infrastructure can be designed to be more productive in generating hypotheses for novel environmental contributors to complex diseases and to help identify genetic and environmental contributors that may not be common across the larger samples but are more frequent within local or ancestral subsets. Incorporating open-ended inquiries and qualitative information about local communal and ecologic contexts and the political, economic, and other social structures that affect health status and outcome will enable qualitative hypothesis generation in those localized contexts, as well as the collection of more detailed genealogic and family health history information that may be useful in designing future studies. Using communities as building blocks for larger cohorts and biobanks presents some practical and ethical challenges but also enhances opportunities for interdisciplinary, multilevel investigations of the multifactorial contributors to complex diseases. *Key words:* **biobanks, communities, complex disease, gene–environment interaction, prospective cohorts, qualitative methods, research design.** *Environ Health Perspect* **113:119–122 (2005). doi:10.1289/ehp.7343 available via *http://dx.doi.org/* [Online 4 November 2004]**

Of the approximately 30,000 genes in the entire human genome, > 1,500 genetic variants have been discovered in which a single allele (either as a homozygote or heterozygote) is sufficient for a single gene or Mendelian disorder such as Huntington's disease to develop (National Center for Biotechnology Information 2004). However, relatively few variants have been confirmed for complex diseases such as cancer, heart disease, and diabetes in which both susceptibility genes and environmental contributors are required for the disease to develop (Botstein and Risch 2003; Hirschhorn et al. 2002). The slow pace in identifying and confirming genetic contributors for complex diseases is due primarily to the difficulties of detecting relatively weak, incremental genetic effects as well as to the possibility that even moderate or strong effects involving a genetic contributor may require the co-occurrence of one or more environmental contributors (Hodgson and Popat 2003).

Similarly, although the identity and function of some environmental contributors to complex diseases such as cancer are well known (toxicants such as asbestos, behaviors such as smoking, viruses such as human papilloma virus), almost all of these known contributors have been identified as such because they have relatively strong effects on disease susceptibility. At the same time, however, a significant proportion of environmental contributors remain unknown for many complex diseases. For example, only one-third of the breast cancer cases in the United States can be accounted for by known risk factors (Stevens 2002). The overwhelming remainder involves either candidate risk factors that are known but have not yet been confirmed as such (which raises the cases accountable to ~50%) or risk factors that are not recognized as such at all. Moreover, even already-identified risk factors for disease such as diet, tobacco, and hormones each are composed of complicated combinations of behaviors and toxicants whose roles in carcinogenesis are not well understood (Brennan 2002). Smoking, for instance, is a contextually shaped behavior that can take a variety of often culturally specific forms as it exposes those who perform it (and others around them) to > 300 different toxicants (Chassin et al. 2000; Frohlich et al. 2002).

In response to these current limitations, a number of researchers have suggested scaling up research sample sizes to provide greater statistical power for identifying and confirming genetic and environmental contributors to complex diseases (Caporaso 2002; Collins 2004; Little et al. 2003; Millikan 2002). Efforts in scaling up sample sizes involve significant national and private investments in research infrastructure. Governmental and nonprofit funding agencies as well as for-profit ventures in various countries are in the process of planning or assembling larger scientific resources to meet that perceived need.

Some of these larger sample collections are in the form of prospective cohorts that recruit healthy participants with the intention of following their health status over a number of years. For example, the National Institute of Child Health and Human Development along with the National Institute of Environmental Health Sciences, the Centers for Disease Control and Prevention, and the U.S. Environmental Protection Agency has been planning a National Children's Study designed to follow 100,000 children and their parents over multiple decades (National Children's Study 2004), and the National Cancer Institute (NCI) has recently issued a new call for proposals for funding large prospective cohorts (NCI 2003). The NCI already funds the Black Women's Cohort (64,500 participants) and the California Teachers Study (133,479 participants) among other large prospective studies (NCI 2004). The NCI announcements of funding for prospective cohorts explicitly contrast them with previous investments in cross-sectional or case–control studies, characterizing cohorts as more flexible, longer-lasting investments in research infrastructure. Most recently, the National Human Genome Research Institute, in collaboration with the National Heart, Lung, and Blood Institute, has requested information from researchers in planning a national cohort of 500,000 participants (National Institutes of Health 2004). In Europe, there is a long tradition of birth cohort studies that extend decades into adulthood, with recent investments in new birth cohorts by the United Kingdom and planning for a "mega" cohort by the European Union (Kogevinas 2002).

Address correspondence to M.W. Foster, Department of Anthropology, 455 W. Lindsey, Room 505C, University of Oklahoma, Norman, OK 73019 USA. Telephone: (405) 325-2491. Fax: (405) 325-7386. E-mail: morris.w.foster-1@ou.edu

This publication was made possible by grant ES11174 from the National Institute of Environmental Health Sciences (NIEHS) and grant HG02691 from the National Human Genome Research Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, National Human Genome Research Institute, or National Institutes of Health.

The authors declare they have no competing financial interests.

Received 22 June 2004; accepted 3 November 2004.

A related kind of resource, often called biobanks, incorporates members of national or regional populations for which extensive retrospective medical records, DNA samples, and other health-related information are available to researchers (Austin et al. 2003). Some biobanks also function as prospective cohorts. The deCODE project, for instance, already has isolated genes that appear to contribute to osteoporosis, stroke, diabetes, and several other complex diseases using historical and contemporary health information and DNA samples from more than 100,000 residents of Iceland (deCODE Genetics 2004), although some of those findings may turn out to be limited to rarer familial factors. Similar biobanks are being assembled in Estonia (an open-ended number of participants), the United Kingdom (500,000 participants), Quebec (60,000 participants), and Japan (300,000 participants). In the United States, Howard University has announced the formation of a biobank with samples from participants who identify themselves as African Americans (Kaiser 2003).

Investments made today in prospective cohorts and biobanks that are projected to be used (and funded) for decades to come will have significant consequences for determining both the opportunities and the limits of future research into genetic and environmental contributors to complex diseases. Although it will be possible to establish new cohorts and biobanks in the future, it will be several years before prospectively recruited participants develop diseases of interest in sufficient numbers for analysis. Moreover, funding for additional cohorts in future years will compete with the costs of maintaining ongoing cohorts, which likely will limit future growth in this research infrastructure. Consequently, as cohorts and biobanks are being planned, it is important to consider the methodologic implications that their increased scales may have for identifying genetic and environmental contributors that may be more locally variable in effect. Locally variable or less common contributors nonetheless can have significant effects on health disparities, raising questions about the equitable distribution of research benefits in the case of large, expensive cohorts that may not be designed to attend to smaller-scale contexts.

Is Bigger Always Better?

Although larger cohorts or biobanks likely will help identify many genetic and environmental contributors that are more common among their members, they will be less likely to help identify those less common contributors that are rare among most participants. Indeed, the additional power that a larger cohort provides to detect weaker common effects simultaneously can mask those contributors that are localized primarily within subsets of the larger sample, depending on how cohort information is collected and analyzed. For instance, a genetic variant that is more frequent among individuals of a particular ancestry but rare among others may not be detected in a sample of 100,000 participants recruited using such inclusion criteria as regional or national residence or occupation. Similarly, an environmental contributor that is specific to exposures resulting from a local ecologic feature or a locally specific behavior also could be lost in a large, multisite cohort, even though it may be a significant determinant of disease. This means that the ways in which participants are categorized and recruited for a particular cohort or biobank and in which their information is collected and analyzed will affect what studies using that resource may find as well as what they may miss.

A criticism of the UK Biobank, for instance, has been that it has no specific plans to incorporate a familial component into its recruitment strategy (Wright et al. 2002). Family members (particularly sibling pairs and parents) provide greater power for separating genetic effects from the background noise of nongenetic effects. In addition, there also tend to be correlations in common environmental and gene–environment interactions among close relatives compared with random, unrelated individuals. Thus, the larger size of a cohort may not necessarily increase its power to detect genetic or environmental contributors to complex diseases.

That situation is complicated further by the possibility that the same complex disease may have multiple genetic and environmental contributors that are neither necessary nor sufficient for a similar phenotype to be expressed (Smith and Lusis 2002). In the cases of type 2 diabetes and systemic lupus erythematosus, for example, different candidate genes have been proposed from studies of geographically and ancestrally differing patient populations, although some of those will not be confirmed (Kelly et al. 2002; Stern 2002). In the case of breast cancer (as for the vast majority of other cancers), not all confirmed environmental contributors need be present for the disease to develop. With the additional variable of gene–environment interactions, it may well be that some significant (although still minority) proportions of the incidence of most complex diseases are attributable to intersections of locally varying combinations of genetic and environmental contributors some or even many of which may not be detectable in large multisite samples. To the extent that those polygenic and polyenvironmental contributors are nonrandomly distributed among and across populations, a large cohort or biobank may fail to detect some or even most of these unless it is structured to support more intensive study of subsets of participants.

The greater cost of larger cohorts, however, tends to mean that fewer and often less precise measures are obtained for each participant, a situation that actually can reduce the power of a larger sample (Wong et al. 2003). Sampling costs also can reduce the ability to collect information that is most productive for hypothesis generation. Because it is expensive to investigate family histories and environmental exposure histories for large numbers of participants (Barbour 2003), large cohort studies tend to collect participant information through closed-ended questions—that is, by giving participants a range of predetermined answers to predetermined questions and forcing them to choose among them (UK Biobank 2002). For environmental exposures, closed-ended questions are useful in testing hypotheses about established or suspected contributors but are of limited value in identifying previously unsuspected contributors whether those are localized or more common (Foster and Aston 2003). With respect to ancestry, some studies allow participants to indicate more than one ethnic or racial background but without eliciting additional information that may be more informative about how genetic variants are distributed in the extensive middle ground between immediate family members and large population categories such as European American or African American.

These limited, closed-ended responses frequently are used as proxies for a shared population history (in the case of ancestry) or for shared environmental exposures (or both) for purposes of sample stratification. The difficulty, however, is that such broad, decontextualed proxies often are treated as units of analysis rather than as heuristic means to disambiguate or discover specific ancestral and environmental contributors to disease or to provide a degree of diversity within the sample frame.

Identity alone, however, is not causal and may not even necessarily be predictive. First, not all factors linked to a given identity necessarily contribute to disease expression or to the expression of the same diseases. Second, only some environmental and ancestral factors are shared among those with a common identity. Third, only some of those with a common identity necessarily share those linked factors. Social identity does become a more powerful predictor, however, when it intersects with ecology in a specific locality. Sharing both a social identity and a locality increases the likelihood that and the extent to which a social community will regulate the actions of its members according to some standard of appropriateness (and, hence, manifest many of the same behavioral environmental factors), the likelihood that community members are exposed to many of the same ambient factors in the physical environment, and the degree of access to prevention, surveillance, and treatment available to community members. Locality also may limit significantly the number of ancestries shared

among co-residents while increasing the likelihood that some are related by more immediate genealogic connections.

Communities as Building Blocks

These critiques suggest that a significant challenge in constructing large-scale cohorts and biobanks is to design a study with a large number of participants that nonetheless gathers rich data on individuals and the contexts that affect their health, providing flexibility for discovering unanticipated data fields and new categories within existing fields. One solution may be to use the local communities in which individuals are everyday members—a naturally occurring social middle ground between single participants and very large ethnic and other categories—as building blocks for constructing large prospective samples. Local communities also would be appropriate contexts for recruiting parents and siblings to enrich the familial component of cohorts and biobanks.

Local communities, of course, may be quite variable in form, ranging from relatively well-defined residential clusters or towns in rural areas to neighborhoods or social networks within large metropolitan areas. What defines a localized community, however, is that its members share similar interactional conventions, a consequence of their everyday encounters with one another, as well as similar ambient or background exposures due to the local physical environment.

The idea that locality or place may affect health is not new (Durkheim 1951). However, the last decade has seen a revival of interest in theorizing and conceptualizing that relationship (Curtis and Rees-Jones 1998; Kearns and Joseph 1993; Macintyre et al. 1993; Tunstall et al. 2004). In contrast with the prevailing epidemiologic focus on individual risk factors, this revival has emphasized collective or contextual effects that may mediate the effects of individual-level variables such that the health status of individuals depends to some extent on the social and physical environments in which individuals grow up and live (Schwarz 1994; Susser 1994). The proponents of this approach argue that collective or contextual "area effects" are complex, multilevel interactions involving phenomena or forces ranging from global, national, or regional social structures that determine opportunities and limitations for well-being (including economic systems and conditions, health care systems and access, political structures and equity, and widespread cultural beliefs and social practices) to more localized communal beliefs, practices, and conditions to diverse intracommunity patterns of individual agency (Macintyre et al. 2002; Popay et al. 1998). Thus, rather than adopt the traditional epidemiologic practice of isolating and testing one environmental factor at a time while attempting to control for the effects of others, a more appropriate method of analysis may be to embrace the complexity of multilevel collective or contextual contributors.

Fine-grained information about contextual effects in local communities offers two primary advantages in studies of environmental contributors to disease susceptibility. First, those data provide additional background information that can be used to better interpret responses to standardized questions, but in ways that still allow comparison across the larger sample. For instance, the same ethnic identity or household income level can indicate differing health risks and outcomes depending on such locally variable contributors as beliefs about health and illness, familial and communal social dynamics and networks, and political and economic structures (Krieger 2001; Williams 2003). Each of these parameters (along with others) helps shape everyday life in ways that can have differing consequences for behaviors that may expose individuals to environmental toxins and may be further differentiated by local variations in physical environments and the ambient exposures that those offer. Detailed investigations of these local differences can augment an understanding of the pathways by which social and ecologic factors contribute to disease susceptibility or can explain why a risk factor does not appear to be as predictive for a specific subpopulation (Frohlich et al. 2001).

Second, detailed local investigations allow many more opportunities for hypothesis generation, which then can be tested across the larger sample. Epidemiologic tests for the statistical significance of associations between established proxies such as ethnicity or socioeconomic status and disease incidence or mortality offer few opportunities for generating novel hypotheses about environmental contributors, mainly because those proxies summarize rather than disaggregate specific environmental factors. In contrast to proxies that summarize information, a community-specific approach that produces large amounts of in-depth information about a broad range of aspects of everyday life provides many specific possibilities for generating hypotheses (Brown 2003; Thompson and Gifford 2000). Indeed, generating hypotheses in small-scale contexts is preferable to doing so across large multisite samples because the former is more amenable to qualitative studies of the different ways in which a large number of factors interact with one another, whereas the latter is more suited to testing hypotheses about a limited number of well-defined, measurable data points.

One of the primary difficulties in using large samples to detect gene–environment interactions is that most nongenetic influences are difficult to measure such that they often are dismissed as being beyond investigation in large samples (Wright et al. 2002). Rather than simply ignore those influences in a larger sample because they cannot be measured accurately or efficiently using existing metrics, qualitative, community-specific approaches offer the possibility of developing a functional understanding of how their effects are achieved, which may help develop accurate, efficient measures that then can be applied in quantitative analyses of larger samples. For example, qualitative data gathered using a "life course" approach can be analyzed to identify biologic and social factors that affect health throughout life in a cumulative manner (both independently and interactively), develop measures of their effects, and describe chains or pathways of risk by which linked exposures raise the likelihood of disease expression (Ben-Shlomo and Kuh 2002; Hallqvist et al. 2004; Kuh and Ben-Shlomo 1997; Kuh et al. 2003). Thus, qualitative methods such as ethnography may become an interdisciplinary companion to epidemiology (Kaufman and Cooper 2001; O'Campo 2003).

Practical and Ethical Challenges

Taking a community-specific approach does raise several logistical and ethical issues that may be problematic. An immediate reaction to our proposal is likely to be concern about the additional cost of recruiting participants to comprise both local community units and the overall cohort, as well as the additional cost of in-person elicitation of open-ended ethnographic and genealogic information. However, given the significant investments already required by very large prospective cohorts or retrospective biobanks, incurring additional costs to enrich the information collected, particularly with respect to hypothesis generation, should be seen as enhancing the value of what will become long-term investments in biomedical infrastructure. A less expensive alternative could be to recruit some but not all participants as members of community units, with the idea that hypothesis generation need not involve all cohort or biobank participants. Indeed, community units may be selected within the larger scale of the study as a whole in two ways: as models that are representative of most study participants (and so have a likelihood of generating hypotheses that may be tested quantitatively across most participants to identify more common contributors) or as efforts to make the cohort more diverse by including participants whose identities contain elements (e.g., ancestry, residence, occupation, household income) that may evidence some contributors to disease differing from those that are more common within the larger cohort. Both strategies add value to the cohort as a whole, albeit in different ways.

With respect to the latter strategy, a frequent problem in making a participant pool more diverse is that including subjects who

may represent minority experiences of disease does not necessarily ensure sufficient power to stratify the sample to quantitatively analyze the less common contributors that may affect those more diverse participants. However, by recruiting some of those minority participants as members of community units, they can be oversampled by the greater detail of information collected rather than by attempting to recruit larger numbers of participants who fit those less common inclusion criteria.

A community-specific strategy also presents several ethical challenges. For example, investigators will need to consider when the additional subject interactions become overly burdensome on particular populations—for example, minority communities that may have been studied extensively in the past. Collecting large amounts of in-depth data about participants, family members, and local communities also presents somewhat greater ethical challenges than do responses to closed-ended questions. Although maintaining confidentiality is a requirement in both cases, it is more difficult to anticipate the risks that might accrue from open-ended inquiries. Moreover, gathering additional information about communities as wholes and about third-party relatives may entail the potential for risks to others than just study participants. For example, published indications of greater genetic susceptibility to a disease among individuals of a specific ancestry or of greater environmental risks to those who reside in a particular place or pursue a particular lifestyle may put those with that ancestry, residence, or lifestyle at a greater risk for discrimination or stigmatization.

At the same time, community-specific investigation often creates a stronger relationship between researchers and participants that should tend to produce greater trust and, hence, more extensive and accurate responses as well as reduced attrition in multiyear and multidecade studies. Emphasizing communities makes it possible to engage pre-existing social organizations and networks in evaluating (and possibly modifying) ethical protections and recruitment strategies, in assisting in participant recruitment and liaison, in actually collecting some study information, and in helping construct local interpretations of the information collected (Sharp and Foster 2000). This greater attention to local contexts should result in greater participant influence in shaping how research is done and greater investigator awareness of local community needs.

Conclusion

The future of biomedical research should reside both in "small science" and in "big science." The two approaches are not necessarily mutually exclusive, although the larger scale of the latter may limit the scale of information that is collected from participants. We believe that larger cohorts and biobanks need not preclude smaller, finer-grained investigations of community-specific influences on disease. In fact, qualitative, community-specific investigations are not only possible within the context of those increasingly large-scale investigations but can provide opportunities for additional hypothesis generation as well as facilitate the multilevel analysis of individual, contextual, and structural factors that contribute to complex diseases.

References

Austin MA, Harding S, McElroy C. 2003. Genebanks: a comparison of eight proposed international genetic databases. Community Genet 6(1):37–45.

Barbour V. 2003. UK Biobank: a project in search of a protocol? Lancet 361:1734–1738.

Ben-Shlomo Y, Kuh D. 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges, and interdisciplinary perspectives. Int J Epidemiol 31:285–293.

Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33(suppl):228–237.

Brennan P. 2002. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? Carcinogenesis 23(3):381–387.

Brown P. 2003. Qualitative methods in environmental health research. Environ Health Perspect 11:1789–1798.

Caporaso NE. 2002. Why have we failed to find the low penetrance genetic constituents of common cancers? Cancer Epidemiol Biomarkers Prev 11(12):1544–1549.

Chassin L, Presson CC, Pitts SC, Sherman SJ. 2000. The natural history of cigarette smoking from adolescence to adulthood in a midwestern community sample: multiple trajectories and their psychosocial correlates. Health Psychology 19(3):223–231.

Collins FS. 2004. The case for a US prospective cohort study of genes and environment. Nature 429(6990):475–477.

Curtis S, Rees-Jones I. 1998. Is there a place for geography in the analysis of health inequality? Soc Health Illness 20:645–672.

deCODE Genetics. 2004. From genes to drugs. Reykjavik, Iceland:deCODE Genetics. Available: http://www.decode.com/ [accessed 14 October 2004].

Durkheim E. 1951 [1897]. Suicide: A Study in Sociology (Spaulding JA, Simpson G, trans). New York:Free Press.

Foster MW, Aston CE. 2003. A practice approach for identifying previously unsuspected environmental contributors to systemic lupus erythematosus and other complex diseases. Environ Health Perspect 111:593–597.

Frohlich KL, Corin E, Potvin L. 2001. A theoretical proposal for the relationship between context and disease. Sociol Health Illness 23(6):776–797.

Frohlich KL, Potvin L, Gauvin L, Chabot P. 2002. Youth smoking initiation: disentangling context from composition. Health Place 8(3):155–166.

Hallqvist J, Lynch J, Bartley M, Lang T, Blane D. 2004. Can we disentangle life course processes of accumulation, critical period, and social mobility? An analysis of disadvantaged socio-economic positions and myocardial infarction in the Stockholm Heart Epidemiology Program. Soc Sci Med 58:1555–1562.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genet Med 4(2):45–61.

Hodgson SV, Popat S. 2003. Polymorphic sequence variants in medicine: a challenge and an opportunity. Clin Med 3(3):260–264.

Kaiser J. 2003. African-American population biobank proposed. Science 300:1485.

Kaufman JS, Cooper RS. 2001. Commentary: considerations for use of racial/ethnic classification in etiologic research. Am J Epidemiol 154:291–298.

Kearns RA, Joseph AE. 1993. Space in its place: developing the link in medical geography. Soc Sci Med 37:711–717.

Kelly JA, Moser KL, Harley JB. 2002. The genetics of systemic lupus erythematosus: putting the pieces together. Genes Immunity 3(suppl 1):S71–S85.

Kogevinas M. 2002. Expression of interest for an integrated project: European Union birth-cohort on child health and human development. Luxembourg:Community Research and Development Information Service. Available: http://eoi.cordis.lu/docs/int_28461.doc [accessed 14 October 2004].

Krieger N. 2001. Theories of social epidemiology in the 21st century: an ecosocial perspective. Int J Epidemiol 30:668–677.

Kuh D, Ben-Shlomo Y, eds. 1997. A Life Course Approach to Chronic Disease Epidemiology. Oxford, UK:Oxford University Press.

Kuh D, Ben-Shlomo Y, Lynch J, Hallqvist J, Power C. 2003. Life course epidemiology. J Epidemiol Community Health 57:778–783.

Little J, Khoury MJ, Bradley L, Clyne M, Gwinn M, Lin B, et al. 2003. The Human Genome Project is complete. How do we develop a handle for the pump? Am J Epidemiol 157(8):667–673.

Macintyre S, Ellaway A, Cummins S. 2002. Place effects on health: how can we conceptualise, operationalise, and measure them? Soc Sci Med 55:125–139.

Macintyre S, Mciver S, Soomans A. 1993. Area, class, and health: should we be focusing on places or people. J Soc Policy 22:213–234.

Millikan R. 2002. The changing face of epidemiology in the genomics era. Epidemiology 13(4):472–480.

National Center for Biotechnology Information. 2004. Online Mendelian Inheritance of Man. National Center for Biotechnology Information. Available: http://www.ncbi.nlm.nih.gov/Omim/mimstats.html [accessed 14 October 2004].

National Children's Study. 2004. What Is the National Children's Study? Available: http://nationalchildrensstudy.gov/about/mission/overview.cfm [accessed 14 October 2004].

National Institutes of Health. 2004. Request for Information: Design and Implementation of a Large-Scale Prospective Cohort Study of Genetic and Environmental Influences on Common Diseases. NOT-OD-04-041. Bethesda, MD:National Institutes of Health. Available: http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-04–041.html [accessed 14 October 2004].

NCI. 2003. Cohort Studies in Cancer Epidemiology (PAR-04-011). National Cancer Institute. Available: http://grants1.nih.gov/grants/guide/pa-files/PAR-04-011.html [accessed 14 October 2004].

NCI. 2004. Consortium of Cohorts: Comparison of Selected Characteristics among Large Cohorts. Available: http://ospahome.nci.nih.gov/cohort/table.htm [accessed 14 October 2004].

O'Campo P. 2003. Invited commentary: advancing theory and methods for multilevel models of residential neighborhoods and health. Am J Epidemiol 157:9–13.

Popay J, Williams G, Thomas C, Gatrell A. 1998. Theorising inequalities in health: the place of lay knowledge. Sociol Health Illness 20:619–644.

Schwarz S. 1994. The fallacy of the ecological fallacy—the potential misuse of a concept and the consequences. Am J Pub Health 85:819–824.

Sharp RR, Foster MW. 2000. Involving study populations in the review of genetic research. J Law Med Ethics 28:41–51.

Smith DJ, Lusis AJ. 2002. The allelic structure of common disease. Hum Mol Genet 11:2455–2461.

Stern MP. 2002. The search for type 2 diabetes susceptibitity genes using whole-genome scans: an epidemiologist's perspective. Diabetes Metab Res Rev 18(2):106–113.

Stevens RG. 2002. Lighting during the day and night: possible impact on risk of breast cancer. Neuroendocrinol Lett 23(suppl 2):57–60.

Susser M. 1994. The logic in ecological. 1. The logic of analysis. Am J Public Health 84:825–829.

Thompson SJ, Gifford SM. 2000. Trying to keep a balance: the meaning of health and diabetes in an urban Aboriginal community. Soc Sci Med 51(10):1457–1472.

Tunstall HVZ, Shaw M, Dorling D. 2004. Places and health. J Epidemiol Community Health 58:6–10.

UK Biobank. 2002. Protocol for the UK Biobank: a study of genes, environment and health. Available: http://www.ukbiobank.ac.uk/documents/draft_protocol.pdf [accessed 14 October 2004].

Williams GH. 2003. The determinants of health: structure, context and agency. Sociol Health Illness 25:131–154.

Wong MY, Day NE, Luan JA, Chan KP, Wareham NJ. 2003. The detection of gene-environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? Int J Epidemiol 32(1):51–57.

Wright AF, Carothers AD, Campbell H. 2002. Gene-environment interactions—the BioBanks UK study. Pharmacogenomics J 2:75–82.