Int. J. Epidemiol. Advance Access published August 19, 2004

International Journal of Epidemiology
DOI: 10.1093/ije/dyh278

REVIEW

The emergence of epidemiology in the genomics age

Muin J Khoury,[1] Robert Millikan,[2] Julian Little[3] and Marta Gwinn[1]

'As genomics and epidemiology begin to intersect, there is the potential for both fields to be altered in ways that are mutually beneficial' (R Millikan[1])

While observational epidemiology is considered to be the scientific foundation of public health[2] it is often viewed as a 'soft science' limited by an inherent inability to fully control for confounding, misclassification, and selection bias.[3] Enter the 'genomics era'.[4,5] With the completion of the Human Genome Project, and the eventual identification of thousands of human genetic variants, scientists are predicting that the use of personalized genomic information will revolutionize the future practice of medicine[6] and public health.[7] In this paper, we discuss the emerging role of epidemiology in the genomics era. In particular, we show how synergistic interaction between genomics and epidemiology is not only mutually beneficial but crucial to the optimal development of each field in the 21st century. Our thesis is twofold:

1. Epidemiology is essential to fulfil the promise of genomics for clinical and public health practice. Epidemiological principles and methods will be applied for gene discovery, gene characterization in populations, as well as evaluation of genetic information in practice; and
2. Genomics can enhance the potential for epidemiology to contribute to multidisciplinary scientific research. Genomic tools will influence epidemiological study design, analysis, and causal inference on 'environmental' causes of disease.

Throughout this paper, we define genomics as 'the study of the functions and interactions of all the genes in the genome'.[6] Although we focus on genomics, epidemiological approaches apply as well to all emerging 'omic' disciplines that are concerned with the study of gene products, expressions, and interactions (e.g. proteomics,[8] transcriptomics,[9] metabonomics,[10] and nutrigenomics[11]). We consider only the human genome although genomes of other organisms are among the 'environmental' factors contributing to human health and disease. Lastly, we discuss only observational epidemiology, keeping in mind that experimental epidemiology is the scientific basis for randomized controlled clinical trials. For additional references about genetics and genomics concepts and terminology, we refer readers to the recently published Encyclopedia of the Human Genome.[12]

The impact of epidemiology on genomics

'Although experimental species are of great value for the initial identification and functional analysis of complex disease genes, final evidence for the involvement of these genes in human diseases must come from extensive epidemiological studies, preferably in different populations' (Peltonen and McKusick[13])

Vision of genomics research in the 21st century

The completion of the Human Genome Project in 2003 has heightened expectations that health benefits will follow quickly. In the US, the National Institutes for Health (NIH) has published a vision for the next generation of genomic research that will span the continuum from biology to health and society.[14] This vision identifies 'six critically important cross-cutting elements': *resources, technology development, computational biology (using mathematical and computational methods to understand complex biological phenomena), training; ethical, legal and social implications*, and *education*. It acknowledges implicitly the need for more epidemiological research (e.g. *resources* to establish a 'healthy cohort' for correlating genetic variants with health and disease), the need to study gene–environment interactions (e.g. *computational biology* to elucidate effects of environmental factors and their interactions with genetic variants), the need to stimulate interdisciplinary collaboration (e.g. *computational biology* to promote 'standardization of data sets'), and *training* to provide scientists with interdisciplinary skills). In addition, NIH recently published a scientific roadmap[15] that addresses the need to uncover the daunting complexity of biological systems and their many interconnected networks of molecules, cells, tissues, their interactions, and regulation.

The need for a 'public health' research strategy in genomics

What is less often recognized but equally crucial in fulfilling the promise of the Human Genome Project is the need for a 'public health' strategy for guiding genetics research and for translating basic research findings into new opportunities for disease prevention, detection, and treatment.[7] A population approach addresses critical gaps in translation: (1) population-level information on the role of genomic variation and its interaction with modifiable risk factors in health and disease; (2) evidence-based processes for assessing the added value of genomic

[1] Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA.

[2] Department of Epidemiology, University of North Carolina School of Public Health, Chapel Hill, North Carolina, USA.

[3] Epidemiology Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, Scotland.

Correspondence: Dr M Khoury. Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, 1600 Clifton Road, Mailstop E82, Atlanta, Georgia, USA 30333. E-mail: mkhoury@cdc.gov

information in diagnosis, treatment, and prevention of most human diseases; and (3) health system capacity to implement genomic applications in practice.[7] The last issue was highlighted by Lenfant in his 2003 Shattuck lecture, 'Lost in Translation,'[16] in which he commented on the poor state of the 'translational highway' for taking findings from basic science and clinical investigations into the practice of medicine and public health. For example, although aspirin is highly effective for secondary prevention among patients with heart disease, it is prescribed for less than a third of patients,[17] prompting his comment: 'If we didn't do it with aspirin, how can we expect to do it with DNA?'[16]

The emergence of human genome epidemiology

As the basic science of population health, epidemiology will have an increasingly important role in addressing these translation gaps, especially in the conduct of population-based research to assess the utility of genomic information in improving health and preventing disease.[18] Over the past few years, we have promoted an epidemiological approach to the human genome–human genome epidemiology,[19] an evolving field of inquiry that uses systematic application of epidemiological methods to assess the impact of human genetic variation on disease occurrence. As shown in Table 1, human genome epidemiology plays an important role in the continuum from gene discovery to the development and applications of genomic information for diagnosing, predicting, treating, and preventing disease. In 1998, the Human Genome Epidemiology Network, or HuGE Net,[20] began as an ongoing global collaboration of individuals and organizations committed to the assessment of the impact of human genome variation on population health. Through collaboration, systematic reviews, training, and information dissemination (Figure 1), HuGE Net applies systematic approaches to build the global knowledge base on population characteristics of genes and their associations with various diseases. An important activity of HuGE Net is the development of guidelines, recommendations and methods for the appraisal and integration of epidemiological data on the human genome along the continuum from genetic research to genetic testing.[21] The synthesis of knowledge is crucial to the evidence-based integration of human genomics into the practice of medicine and public health in the 21st century.

HuGE Net continuously collects information on the published epidemiological literature on human genes. Between 2001 and 2003 (Table 2), we saw an increasing number of published 'epidemiology' papers on the human genome, most of which (86%) were on gene–disease associations.[22] Unfortunately, many observational studies that are lumped under the rubric of 'epidemiology' suffer from serious methodological flaws in study design, subject selection, genotype assessment, measurement of environmental exposures, small sample sizes, and lack of adjustment of potential confounding variables. These flaws have resulted in many unreplicated findings with non-causal explanations.[23–25] Genetic 'association studies' are often inaccurately equated with 'epidemiological studies,' with the implicit if not explicit connotation of sub-optimal study design, as in a recent *Lancet* editorial: 'In the genetics of complex diseases, association is in danger of becoming a rather dirty word.'[26]

Population-based gene discovery and characterization

Epidemiology can make a major contribution to genomics in the next decades by applying a well-disciplined methodological

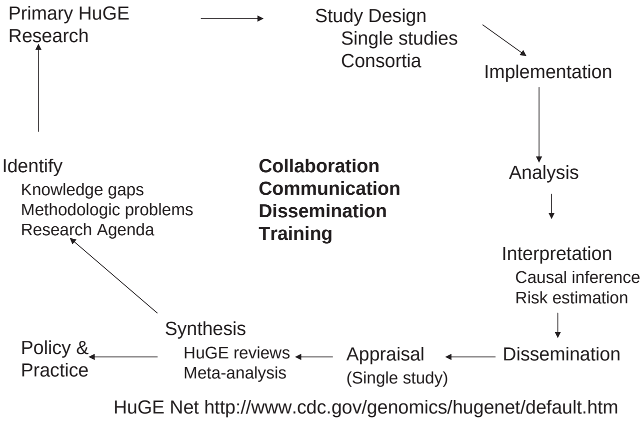

Figure 1 The Human Genome Epidemiology Network: From primary research to synthesis and dissemination for policy and practice

Table 1 Human genome epidemiology:[a] from gene discovery to applications

From discovery to applications	Epidemiologic impact on genomics	Examples
Gene discovery	Epidemiological approaches to sound study design: selection, representativeness, and generalizability	Co-operative Family Registries for Breast and Colorectal Cancer Research[29,30]
Gene characterization	Epidemiological measures of risk (relative, absolute and attributable fraction); Measurement of gene–environment interaction;Methods for validation and adjustment for extraneous factors	Atherosclerosis Research in Communities[33] National Birth Defects Prevention Network[36]
Genomic testing	Epidemiological evaluation of sensitivity, specificity and predictive values of genetic tests. Decision analysis using epidemiological measures	Pharmacogenomics[47] Genomic profiling[41] Mass screening[38]

[a] We use the term human genome epidemiology to refer to the continuum of epidemiological applications to the human genome. The term genetic epidemiology is more established in the literature and often has been used in the context of studies of gene discovery, most notably family studies.

Table 2 Number of published[a] human genome epidemiological papers, by type of paper and year of publication

	Year		
Type of paper	2001	2002	2003
Population prevalence of genetic variants	308	347	307
Genotype–disease associations	2137	2793	2920
Gene–gene and gene–environment interaction	436	573	587
Evaluation of genetic tests and newborn screening	95	92	94
All[b]	2478	3202	3415

[a] Abstracts of published papers are available on the CDC online searchable database [22] (GDP Info). Search was conducted on 20 March 2004.

[b] Total number of papers per year exceeds the sum of individual categories because of overlap in types of studies.

approach to the design, conduct, and analysis of 'association studies'[19] and the process of causal inference from these studies. Because genomics employs laboratory methods, some may have the false impression (or hope) that findings from observational studies in genomics represent 'experimental' data and therefore offer a higher level of evidence of causation than other epidemiological studies, irrespective of study quality. As remarked by Potter,[27] most genomic studies are observational in nature and therefore have the same limitations as other observational studies, thus deserving a rigorous epidemiological design, analysis, and interpretation.

A deliberate epidemiological approach can support simultaneous gene discovery and population-based inference of risks. For example, case-control studies of population-based incident disease cases and their families provide a platform for conducting family-based linkage and association studies to discover new genes, and permit inferences regarding the contribution of these genes to the burden of disease in the underlying population.[28] The National Cancer Institute (NCI) sponsors Co-operative Family Registries for Breast and Colorectal Cancer Research that reflect this philosophy.[29,30] Population-based case registries can support a number of study designs, including extended family studies, case-parent trios,[28] and case-control-family design.[31]

In addition, efforts are now being made to integrate genomics into population-based epidemiological studies initiated in the pre-genomic era to study disease incidence and prevalence, natural history and risk factors. Examples of well known cohort studies include the Framingham study,[32] the Atherosclerosis Research in Communities (ARIC[33]), the European Prospective Investigation on Cancer (EPIC[34]), and the newly designed National Children Study,[35] a proposed US cohort study of 100 000 pregnant women and their offspring to be followed from before to birth to age 21 years.

An example of a population-based case-control study in the US is the ongoing CDC-sponsored National Birth Defects Prevention Study.[36] This study is conducted in 10 states to assess the role of genetic and environmental factors in the occurrence of major structural birth defects. Cases are ascertained from state-based birth defects surveillance systems. Controls are randomly selected from birth certificates or hospital medical records. This is the largest ongoing population-based collaborative study in the US that covers a base population of almost half a million births a year.[36] As of February 2004, the study included more than 12 000 cases and 4000 controls (P Honein, CDC, personal communication).

Epidemiological methods will also help estimate unbiased allele and genotype frequencies from cross-sectional population studies. Population prevalence data provide baseline estimates for guiding research and health policy. Unfortunately, most prevalence information comes from convenience samples or otherwise unrepresentative groups.[22] An example of a population-based prevalence study is the analysis of two common mutations in the haemochromatosis gene *(C282Y and H63D variants of HFE)* in the US population. Steinberg *et al.* genotyped 5171 samples from the CDC's Third National Health and Nutrition Examination Survey (NHANES III), a nationally representative survey conducted in the US from 1992 to 1994. Genotype and allele frequency data were cross-classified by sex, age, and race/ethnicity.[37]

Epidemiology and genomic tests

An epidemiological approach is fundamental to evaluating genomic tests, especially those intended for population screening and disease prevention.[38] Many tests that will emerge in the next decades will not be used for diagnosing rare genetic diseases but for predicting the risk of common diseases in otherwise healthy people in order to guide decisions about preventive interventions or therapies.[39] An example often quoted to illustrate this potential application of genetic tests is the hypothetical case scenario described by Collins in 1999.[40] He predicted that by 2010, a 23 year old man could undergo genetic testing and receive a report predicting his risk of several diseases based on analysis of genetic variants at multiple loci, a concept that has come to be called genomic profiling.[41] This example highlights the need to obtain epidemiological data that are required for developing risk estimates. In addition, discussion of genomic profiling raises a number of evidence-based practice issues. Targeting interventions on the basis of genetic information may not be more effective or economical than population-wide interventions. In this fictitious example, the patient was given medical advice on smoking cessation, treatment of hypercholesterolaemia, and initiation of colorectal cancer screening that currently represent best practices without the use of genetic tests. (ref 19, p. 5)

Although genomic profiling tests are commercially available, they are clearly not ready for prime time. Epidemiological studies and clinical trials to assess the clinical validity and utility of these tests have not been done.[41–43] An argument for the biological plausibility of genomic profiling could be based on understanding interaction among genes in well-defined pathways (e.g. folate metabolism or carcinogen metabolism[41]). Yang *et al.*[44] showed that bundling several variants from multiple loci could increase the predictive value of genetic testing for susceptibility to common diseases, especially in the presence of pertinent environmental exposures (e.g. dietary and supplemental folic acid intake). Khoury *et al.*[41] described a hypothetical example of a common disease (5% lifetime risk), for which three genetic variants at different loci and one environmental exposure are risk factors. Even for modest effects of each variant alone (risk ratios from 1.5 to 3.0) and modest interactions between the exposure and the genes, the predictive value of the test for people with two (and especially three) disease-associated variants can be quite high (50–100%) in the presence of a modifiable exposure. However, for rarer diseases (<1 per 10 000), the predictive value of multiple genotypes is much lower. Also, as increasing numbers of genes are added to a profile, the size of the highest risk group will decrease, thus diminishing the population impact of preventive interventions targeted to this group, especially in the setting of rare genotypes, weak associations, and weak interactions. We emphasize that these assessments of the potential impact of genomic profiling are based on hypothetical data. Even when real data are available from well-designed epidemiological studies, assessing the value of genomic profiling will require information on the costs, benefits, and risks of testing and interventions.

The epidemiological approach is relevant to pharmacogenomics, an emerging field at the intersection between pharmacology and genomics that promises customized treatment or chemoprevention on the basis of genetic variation and the use of genomics to better understand and improve target identification and drug delivery.[45] With the recent Food and Drug Administration[46] approval of the first DNA-based test for measuring variants in Factors V and II to provide management to people at increased risk of deep vein thrombosis, we expect of the pace of pharmacogenomics development to accelerate. Epidemiological parameters are important in cost-effectiveness analysis to determine the value added by pharmacogenomic testing. Veenstra[47] explored the conditions that would favour genetic testing in clinical practice to tailor individual treatment. He used the example of interaction between genetic variation in thiopurine S-methyltransferase (TPMT) and 6-mercaptopurine (6MP) in the treatment of childhood acute lymphoblasic leukaemia. TPMT is responsible for the inactivation of 6MP and TPMT deficiency is associated with severe haematopoietic toxicity when patients with deficient TPMT are given standard doses of 6MP. Veenstra described epidemiological parameters—such as prevalence of the deficient enzyme, and the strength of the drug–gene interaction in producing myelosuppression—that are needed to assess the cost-effectiveness of genetic testing before drug use in a decision analysis model (Table 3).[47,48]

Another example of the increasing potential of pharmacogenomics for clinical practice is the relationship between certain HLA alleles and abacavir hypersensitivity in the treatment of HIV1.[49] Similar to the TPMT example above, epidemiological data will be useful in decision analysis of cost-effectiveness.

Epidemiological parameters are also useful for assessing other tests that might be developed based on genomic technology. Examples of potentially emerging genomic tests include use of serum proteomic patterns for early detection of ovarian cancer,[50] and genomic analysis of stool samples for early detection of colorectal cancer.[51] However, initial research results are not immediately generalizable to a population setting and these two examples require further study. For example, the positive predictive value of the test for ovarian cancer[50] was overestimated because it was based on a study sample with 50% prevalence of ovarian cancer.[52] The stool-based genomic test for colorectal cancer[51] also requires further evaluation. For example, it is unknown whether stool-based genomic tests will be more sensitive and specific for detecting early colorectal cancer than traditional faecal occult blood testing (FOBT), which is reported to have 40% sensitivity and 96–98% specificity.[53] Although stool-based genomic tests may be more acceptable to patients because they do not require sedation and endoscopy, and preferable for physicians because they do not require specialized health care personnel, determining whether stool-based genomic tests offer any advantages over traditional screening will ultimately depend on randomized clinical trials using morbidity and mortality as endpoints.

The impact of genomics on epidemiology

'Genetic epidemiology is seen by many to be the only future for epidemiology.' (Davey Smith[54])

'The sequencing of the human genome offers the greatest opportunity for epidemiology since John Snow discovered the Broad Street pump.' (Shpilberg[55])

Just as epidemiology is crucial to the fulfilment of the promises of genomics, genomics will enhance the contributions of epidemiology in multidisciplinary scientific research. In response to general concerns about the value of epidemiological research, Butler commented that because of advances in genomics, 'epidemiology (is) set to get fast track treatment'.[56] Genomics will influence epidemiology throughout the continuum from epidemiological study design to analysis and inference (Table 4).

Table 3 The role of epidemiological data in assessing the cost-effectiveness of pharmacogenomic tests

Variable	Factors favouring cost-effectiveness of pharmacogenomic tests	Role of epidemiology
Genotype interacting with drug	Prevalent genotype	Need for population-based prevalence data
Health outcome	High burden of disease (in terms of morbidity and mortality)	Need for epidemiological data to measure morbidity, disability, and mortality
Gene–drug interaction	Strong association between gene variants and adverse health outcomes among people taking the drug (also in comparison with other therapies or no therapies)	Need for epidemiologic data to measure interactions

Adapted from Veenstra.[47]

Table 4 The impact of genomics on epidemiology: study design, analysis, and interpretation

From study design to inference	Impact of genomics on Epidemiology	Examples
Epidemiological study design	Emergence of large scale cohort & case-control studies	Decode Genetics[59] P3G collaboration[62] UK Biobank[56]
	Emergence of new study designs (case-only, family-based)	Population-based[67] registries for cancer and birth defects [66]
Epidemiological analysis	Peering into the 'black box' of risk factor epidemiology; Emerging methods for complex analyses	Hierarchical methods Recursive partitioning methods[77,80] Neural networks[79]
Epidemiological inference	Enhanced tools for causal inference & policy development with or without using genomics in practice	Concept of Mendelian randomization[54]

Table 5 Examples of large-scale population-based genomics studies in progress, 2004

Study	Sample size	Population	Study objectives
Decode Genetics[59]	>100 000	Iceland	'To identify genetic causes of common diseases and develop new drugs and diagnostic tools.' Measures genes, health outcomes, and link with genealogical database
UK Biobank[57]	500 000	Population sample of people 45–69 years:	'To study the role of genes, environment and lifestyle' Link with medical records
CartaGene[58] (Quebec)	>60 000	Population sample of people 25–74 years	'To study genetic variation in a modern population.' Link with health care records, and environmental and genealogical databases
Estonia Genome Project[60]	>1 000 000	Estonian population	'To find genes that cause and influence common diseases' Link with medical records
GenomeEUtwin[61]	~800 000 twin pairs	Twin cohorts from 7 European countries + Australia	'To characterize genetic, environmental and lifestyle components in the background of health problems

The last three projects are part of the global P3G collaboration (Public Population Project in Genomics).[62]

Emergence of large population cohort studies

Genomics is expanding the horizons of epidemiology by adding another dimension to the classic case-control, cohort, and cross-sectional studies. Indeed genomics is inspiring the development of very large longitudinal cohort studies and even studies of entire populations to establish repositories of biological materials ('biobanks') for discovery and characterization of genes associated with common diseases. Table 5 shows a partial listing of such studies in progress that range from large random samples of adult populations like the UK Biobank[57] (N = 500 000) and the CartaGene[58] project in Quebec (N = 60 000), to populations of entire countries such as Iceland[59] (N = 100 000) and Estonia[60] (N = 1 000 000), to a cohort of twins in multiple countries (GenomeEUtwin[61]). In addition to promoting gene discovery, these biobanks will help epidemiologists quantify the occurrence of diseases in various populations and to understand their natural histories and risk factors, including gene–environment interactions.

Longitudinal studies permit repeated phenotypic and outcome measures on individuals over time, including intermediate biochemical, physiological and other 'omic' precursors and sequels of disease (gene expression, protein patterns, etc). Large cohort studies could also be used for nested case-control studies or even case-only studies as an initial screening method (see below). These studies will produce a large amount of data on disease risk factors, lifestyles, and environmental exposures, and provide opportunities for data standardization, sharing, and joint analyses. An example of data standardization across international boundaries is the global P3G (Public Population Project in Genomics[62]), which includes so far three international studies from Europe and North America (Table 4). 'Harmonization' is crucial to create comparability across sites on measures of genetic variation, environmental exposures, questionnaire data, and long-term health outcomes. It is imperative to develop and agree on common epidemiological methods and approaches that can be used to generate and test hypotheses on genetic and environmental influences and gene–environment interactions, and that will allow pooling and synthesis of results across different population groups.

Emergence of novel epidemiological study designs: the case-only method

We are also seeing the emergence of new or otherwise infrequently used study designs. We would like to highlight the case-only study.[63] Although described before the genomics era,[64] the case-only approach has received renewed attention because of its ability under some circumstances to test for gene–environment interaction measured on a multiplicative scale.[65] The case-only approach may also have other more robust applications (see below). Population-based disease registries of incident cancers,[66] birth defects,[67,68]and other conditions offer a practical setting for case-only studies as an initial epidemiological mode of inquiry.[65] Although case-only studies will never replace traditional case-control studies, we believe they can be a useful adjunct to:

1. Scan for genotypes that potentially contribute most to disease occurrence in a population. By using the concept of population attributable fraction, case-only studies can provide an upper estimate of the contribution of complex risk factors, including multiple genetic variants at different loci, to disease occurrence. Genotypes comprising combinations of multiple genetic variants have a very low expected population frequency, even when the variants are individually common (for example, for four common alleles of 10% in the population, only 1 in 10 000 people are expected to have all four); therefore, the case-only approach could be useful in identifying combinations of genes to evaluate further for potential etiologic importance.[65]
2. Evaluate disease aetiological, diagnostic, and prognostic heterogeneity. Genotype–phenotype correlations can be examined among subsets of cases defined clinically, or by use of biological markers based on genotype, gene expression, protein products, or other features. For example, a recent study found that mutations of the CARD15 gene known to be associated with Crohn's disease were correlated with disease of the ileum but not the colon.[69] In another example, Le Marchand *et al.* conducted a population-based study to evaluate overall and stage-specific associations of the *CCND1 870A* allele of the *Cyclin D1 (CCND1)* gene with colorectal cancer. They found that the allele was associated with colorectal cancer, and particularly with more severe forms of the disease that result in higher morbidity and mortality.[70] We emphasize that results from both of the above studies are preliminary and need replication in additional studies. A third example is the analysis of genotype–phenotype correlation in cystic fibrosis, a common single gene disorder in people of Northern European descent. More than 1000 different mutations in the CFTR gene have been described,[71] and genotypic heterogeneity explains in part the highly variable clinical expression of cystic fibrosis. Because lung function varies even among patients with similar CFTR genotype, other genetic and environmental determinants also have a role.[72]
3. Detect gene–gene and gene–environment interaction on a multiplicative scale. Limitations of this approach have been raised, including the assumption of independence between factors, the inability to measure main effects of factors and the restriction of the analysis to detection of departure from multiplicative effects.[65]

Finally, it is important to note that case-only studies are susceptible to the same potential methodological limitations found in case-control studies, including selection bias, information bias, confounding, and sample size and power. These limitations require that results of epidemiological studies be replicated in multiple settings.

Epidemiology and the problem of complexity

The discovery of increasing numbers of genetic variants is confronting epidemiologists with immense analytical challenges. Integrating genomics into mainstream epidemiological research creates increased potential for type I and type II errors. The concept of the epidemiological 'fishing expedition' will become grander in the genomics era. Large scale data dredging will unavoidably lead to numerous positive associations that are not replicated,[73] risking backlash in the scientific community against the epidemiological approach. Although observational studies without a solid epidemiological foundation are particularly problematic, even well-designed epidemiological studies are susceptible to type I errors. One reason for type I errors is the custom of declaring statistical significance based on *P*-values. Wacholder[74] has developed an approach to assess the probability that—given a statistically significant finding—no true association exists between a genetic variant and disease. This approach incorporates not only the observed *P*-value but the prior probability of the gene–disease association and the statistical power of the test. The problem of false positivity is compounded by the obvious tendency of authors and journals to publish 'positive' or interesting findings (publication bias). We hope that the world-wide movement for open access scientific publishing[75] will be able to counter this bias so that both 'positive' and 'negative' results will be disseminated in a timely fashion.

The problem of type II errors or poor statistical power is equally challenging. Consider for a moment the staggering implication for epidemiology of too many genes. Imagine that for a common disease only 10 genes contribute a substantial population attributable fraction. Even if variation at each locus can be classified in a dichotomous fashion (e.g. susceptible genotype versus not), this classification will create 2 to the power 10, or over a 1000, possible strata. Dichotomous classification based on just 20 genes will produce over a million strata. This is methodologically untenable especially when one must consider the interactions of these genes with other genes and environmental factors. Emerging technology will allow us to study simultaneously hundreds and thousands of genome variations, gene expression profiles, and protein patterns. Our simple epidemiological analysis of 2×2 tables, stratified analysis and even logistic regression analysis, the work horse of case-control studies, will quickly face their limitations in an age when a large amount of data on each individual is the rule rather than the exception.[76]

Emerging statistical approaches to complexity

The problem of increasing complexity is generating enthusiastic responses from the statistical community. Novel methodologies have emerged, including hierarchical regression and Bayesian methods.[77,78] These methods may be suited to address the problem of false positive associations resulting from multiple comparisons. Bayesian methods integrate *a priori* expectations, which may be especially relevant for interaction analysis. Neural network analysis is another approach that can be viewed as generalization of logistic regression to non-linear relationships,[79] avoiding the issue of multiple dimensions. However, epidemiological analysis based on neural networks has not enjoyed much popularity yet.

Another approach to joint analysis of multiple genes for quantitative traits is the combinatorial partitioning method (CPM),[80] which represents an extension of traditional analysis of variance between and within genotypes at one gene locus. An excess of variability between the genotypes, relative to within genotypes, represents an association between the gene and the trait. The CPM extends this concept to many genes by genotypic partitioning based on multiple loci. The aim of CPM is to find genotypic partitions such that trait variability is much lower within than between the partitions. An extension of CPM is the multifactor dimensionality reduction (MDR) method.[81] Using this approach, genotypes at multiple loci are grouped into a few categories to create high-risk and low-risk groups. This reduces the number of genotypes from many dimensions to one. The new one-dimensional complex genotype is assessed for its ability to predict disease status.[82] Among analysis methods that have been proposed are recursive partitioning methods,[83] such as tree-based association analysis.[84] For example, these methods may be able to classify people into two or more distinct groups with respect to their propensity to 'bleed' or 'coagulate' based on the combination of genotypes at multiple loci involved in the delicate balance between bleeding and thrombosis (e.g. the cascade of factors I thru X). When based on understanding of underlying biology, it is possible that these composite complex genotypes may be useful for predicting disease outcomes or response to treatment. Although these methods continue to evolve, they may still be liable to data-driven findings, and so far have limited applications in epidemiological studies.

Emergence of Mendelian randomization as a tool for epidemiological inference on environmental risk factors

Genomics may also enhance the power of epidemiology by providing evidence that enhances causal inference on the association between environment (broadly defined to include chemical, biological, nutritional and social factors) and human diseases.[54] Associations between exposures and diseases seen in epidemiological studies are often confounded by unmeasured factors in spite of efforts to optimize the conduct of these studies. As reviewed by Davey Smith and Ebrahim, Mendelian randomization, the random assortment of genes from parents to offspring that occurs during meiosis, may provide an indirect method for assessing the causal nature of environmental exposures, as certain genotypes can be viewed as proxies for certain exposures.[54] The association between a disease and a polymorphism that mimics the biological relationship between a proposed exposure and disease is viewed as not susceptible to the effect of confounding that may plague observational studies.

The concept of Mendelian randomization can be illustrated using the example of the *C677T* variant of the methylene tetrahydrofolate reductase (*MTHFR*) gene, which results in reduced enzyme activity.[54] The enzyme is involved in the conversion of 5,10-methylene tetrahydrofolate (from dietary folate) to 5-methyl tetrahydrofolate, which is needed for the conversion of homocysteine to methionine. Thus, this genetic variant mimics low dietary folate intake leading to higher levels of homocysteine. The authors then discuss causal inference on the role of folate in neural tube defects (NTD). They found close agreement between the findings from observational studies showing protective effects of folic acid supplements; genetic association studies showing increased risk of NTD for maternal but not paternal TT genotypes (reflecting the low folate intrauterine environment); and randomized controlled clinical trials showing that folic acid supplements reduce risk of occurrence and recurrence of NTD. They suggest that epidemiological studies demonstrating the relationship between *MTHFR C677T* and NTD would have provided strong evidence of the beneficial effect of folic acid supplementation even before data became available from controlled clinical trials. Such evidence is important because traditional epidemiological studies of dietary folate and vitamin supplements in relation to NTD are subject to biased recall of diet and supplement use and to confounding of folate intake with other factors that may also influence NTD risk. Thus, examining the association of *MTHFR* genotype with NTD risk provides additional confidence in the causal nature of the protective effect of folates as a population-wide intervention without implying a need for genetic testing.

In another example of 'Mendelian randomization', Hines *et al.*. conducted a nested case–control study based on the Physicians' Health Study to investigate the relationship of myocardial infarction with alcohol consumption and a specific polymorphism in the gene for alcohol dehydrogenase type 3 (*ADH3*) that alters the rate of alcohol metabolism.[85] They found that moderate drinkers who are homozygous for the slow-oxidizing *ADH3* allele have higher high density lipoprotein levels and a substantially decreased risk of myocardial infarction. The authors commented that:

> observed associations between the risk of a disease and the presence of functional variants in genes that lead to the metabolism or transduction of the factor that underlies the disease add substantial support to the idea that the exposure to the factor is directly related to causation…Improving our ability to identify specific lifestyle and environmental factors as causes of a given disease may prove to be one of the main benefits of the study of common variants in metabolic genes and disease.[85]

While Mendelian randomization has the promise of helping epidemiologists derive better causal inference from environmental risk factor–disease association, there are some caveats.[86] Although appealing as a concept, Mendelian randomization still has to grapple with the common methodological issues that plague many association studies including small sample sizes, linkage disequilibrium, population stratification, and gene–gene and gene–environment interaction that may mask simple gene association effects. Currently, the utility of this approach is also limited by our incomplete understanding of gene functions and biological pathways important in the pathogenesis of common diseases. As our understanding improves in decades to come, the concept of Mendelian randomization may become increasingly useful to epidemiologists.

Of note, the February 2004 issue of the *International Journal of Epidemiology* has a series of articles on Mendelian randomization that address historical perspectives, applications, and methodological issues.[87–93] These papers are highly recommended reading for epidemiologists interested in the topic.

Concluding remarks

We have briefly considered how the advent of the genomics era could lead to a 'renaissance' of observational epidemiology, establishing its crucial role in 'translating' the human genome sequence into understanding health and preventing disease in populations. Genomics has the potential to enrich epidemiological methods along the continuum from study design and analysis to inference on 'environmental' as well as 'genetic' causes of human disease. Because epidemiology and genomics exist in largely separate worlds where different languages are spoken, there is a growing need for multidisciplinary dialogue, training, and collaboration. In the end, it is likely that although the 21st century epidemiologist will use genomic tools, the practice of 'omic' epidemiology will not be so different from that of 19th century Broad Street Pump epidemiology, as it continues to be concerned with calculation, communication, and intervention.[94]

References

[1] Millikan R. The changing face of epidemiology in the genomics era. *Epidemiology* 2003;**13:**472–80.

[2] Weed DL, Mink PJ. Roles and responsibilities of epidemiologists. *Ann Epidemiol* 2002;**12:**67–72.

[3] Taubes G. Epidemiology faces its limits. *Science* 1995;**269:**164–69.

[4] Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large scale biology. *Science* 2003;**300:**286–90.

[5] Collins FS, Guttmacher AE. Welcome to the genomics era. *New Engl J Med* 2003;**349:**996–98.

[6] Guttmacher AE, Collins FS. Genomic medicine—a primer. *New Engl J Med* 2002;**347:**1512–20.

[7] Khoury MJ, Burke W, Thomson E. *Genetics and Public Health in the 21st Century: Using Genetic Information to Improve Health and Prevent Disease*. New York: Oxford University Press, 2000, pp. 3–23.

[8] Sellers TA, Yates JR. Review of proteomics with applications to genetic epidemiology. *Genet Epidemiol* 2003;**24:**83–98.

[9] Kiechle FL, Holland-Staley CA. Genomics, transcriptomics, proteomics, and numbers. *Arch Pathol Lab Med* 2003;**127:**1089–97.

[10] Nicholson JK, Connelly J, Lindon JC, Holmes E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* 2002;**1:**153–61.

[11] van Ommen B, Stierum R. Nutrigenomics: exploiting systems biology in the nutrition and health arena. *Curr Opin Biotechnol* 2002;**13:**517–21.

[12] Cooper DN (ed.). *Nature Encyclopedia of the Human Genome*. London: Nature Publishing Group, 2003.

[13] Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* 2001;**291:**1224–29.

[14] Collins FS, Green ED, Guttmacher AE *et al.* A vision for the future of genomics research. *Nature* 2003;**422:**835–47.

[15] Zerhouni E. Medicine. The NIH Roadmap. *Science* 2003:**302:**63–72.

[16] Lenfant C. Shattuck lecture—clinical research to clinical practice—lost in translation? *N Engl J Med* 2003;**349:**868–74.

[17] Stafford RS, Radley DC. The underutilization of cardiac medications of proven benefit, 1990 to 2002. *J Am Coll Cardiol* 2003;**41:**56–61.

[18] Gwinn M, Khoury MJ. Research priorities for public health sciences in the postgenomic era. *Genet Med* 2002;**4:**410–11.

[19] Khoury MJ, Little J, Burke W (eds). *Human Genome Epidemiology: a Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. New York: Oxford University Press, 2004.

[20] The Human Genome Epidemiology Network-HuGE Net^{TM} Website accessed 20 February 2004 at http://www.cdc.gov/genomics/hugenet/default.htm

[21] Khoury MJ. Epidemiology and the continuum from genetic research to genetic testing. *Am J Epidemiol* 2002;**156:**297–99.

[22] CDC Genomics and disease prevention information system (GDP Info). Accessed on 18 February 2004 at: http://www2a.cdc.gov/genomics/GDPQueryTool/frmQueryBasicPage.asp

[23] Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. *JAMA* 1999;**281:**1919–26.

[24] Ioannidis JP. Genetic associations: false or true? *Trends Mol Med* 2003;**9:**135–38.

[25] Little J, Khoury MJ, Bradley L. The human genome project is complete. How do we develop a handle for the pump? *Am J Epidemiol* 2003;**157:**667–73.

[26] Editorial. In search of genetic precision. *Lancet* 2003;**361:**357.

[27] Potter JD. At the interfaces of epidemiology, genetics and genomics. *Nat Rev Genet* 2001;**2:**142–47.

[28] Thomas DC. Statistical issues in the design and analysis of gene-disease association studies. In: Khoury MJ, Little J, Burke W (eds). *Human Genome Epidemiology: a Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. New York: Oxford University Press, 2004, pp. 92–110.

[29] Peel DJ, Ziogas A, Fox EA *et al.* Characterization of hereditary nonpolyposis colorectal cancer families from a population-based series of cases. *J Natl Cancer Inst* 2000;**92:**1517–22.

[30] Daly MB, Offit K, Li F *et al.* Participation in the cooperative family registry for breast and ovaian cancer studies: issues of informed consent. *J Natl Cancer Inst* 2000;**92:**452–56.

[31] Hopper JL. Commentary: Case-control-family designs: a paradigm for future epidemiology research? *Int J Epidemiol* 2003;**32:**48–50.

[32] National Heart, Lung, and Blood Institute. The Framingham Heart Study: 50 years of research success. Website accessed 18 February 2004, http://www.nhlbi.nih.gov/about/framingham/

[33] The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am J Epidemiol* 1989; **129:**687–702.

[34] Riboli E, Hunt KJ, Slimani N *et al.*. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;**5:**1113–24.

[35] National Institutes of Health. National Children Study website accessed on 18 February 2004 at http://www.nationalchildrensstudy.gov/

[36] Yoon PW, Rasmussen SA, Lynberg MC *et al.* The National Birth Defects Prevention Study. *Public Health Rep* 2001;**116 (Suppl.1):**32–40.

[37] Steinberg KK, Cogswell ME, Chang JC *et al.* Prevalence of C282Y and H63D mutations in the hemochromatosis (HFE) gene in the United States. *JAMA* 2001;**285:**2216–22.

[38] Khoury MJ, McCabe ERB, McCabe L. Population screening in the age of genomic medicine. *N Engl J Med* 2003;**348:**50–58.

[39] Khoury MJ. Genetics and genomics in practice: the continuum from genetic disease to genetic information in health and disease. *Genet Med* 2003;**5:**261–68.

[40] Collins FS. Shattuck lecture—medical and societal consequences of the Human Genome Project. *N Engl J Med* 1999;**341:**28–37.

[41] Khoury MJ, Yang Q, Gwinn M *et al.* An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004;**6:**38–47.

[42] Vineis P, Christiani DC. Genetic testing for sale. *Epidemiology* 2004;**15:**3–5.

[43] Haga SB, Khoury MJ, Burke W. Genomic profiling to promote a healthy lifestyle: not ready for prime time. *Nat Genet* 2003;**34:**347–50.

[44] Yang Q, Khoury MJ, Botto L *et al.* Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am J Hum Genet* 2003;**72:**636–49.

[45] McCarthy JJ. Advances in pharmacogenomic research and development. *Mol Biotechnol* 2003;**25:**275–82.

[46] Food and Drug Administration (FDA). News release, 17 December 2003 http://www.fda.gov/bbs/topics/NEWS/2003/NEW00998.html

[47] Veenstra D. The interface between epidemiology and pharmacogenomics. In: Khoury MJ, Little J, Burke W (eds). *Human Genome Epidemiology: a Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. New York: Oxford University Press, 2004, pp. 234–46.

[48] Marshall E. Preventing toxicity with a gene test. *Science* 2003;**302:**588–90.

[49] Mallal S, Nolan D, Witt C *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* 2002;**359:**727–32.

[50] Petricoin EF, Ardekani AM, Hitt BA *et al.* Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002;**359:**572–75.

[51] Traverso G, Shuber A, Levin B *et al*. Detection of *APC* mutations in fecal DNA from patients with colorectal tumors. *New Engl J Med* 2002;**346:**311–20.

[52] Rockhill B. Proteomic patterns in serum and identification of ovarian cancer. *Lancet* 2002;**360:**169.

[53] Ransohoff DF, Sandler RS. Screening for colorectal cancer. *New Engl J Med* 2002;**346:**40–44.

[54] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32:**1–22.

[55] Shpilberg O, Dorman JS, Ferrell RE *et al.* The next stage: molecular epidemiology. *J Clin Epidemiol* 1997;**50:**633–38.

[56] Butler D. Epidemiology to get fast track treatment. *Nature* 2001;**414:**139.

[57] Wright AF, Carothers AD, Campbell H. Gene-environment interactions–the BioBank UK study. *Pharmacogenomics J* 2002;**2:**75–82.

[58] CARTaGENE project. Website accessed 19 February 2004 at http://www.cartagene.qc.ca/en/index.htm

[59] Hakonarson H, Gulcher JR, Stefansson K. *deCODE genetics, Inc.Pharmacogenomics* 2003;**4:**209–15.

[60] Estonian Genome Project. Website accessed 19 February 2004 http://www.geenivaramu.ee/index.php?show=main&lang=eng

[61] GenomEUtwin project. Website accessed 19 February 2004 at http://www.genomeutwin.org/

[62] Public Population Project in Genomics. Website accessed 19 February 2004 at http://www.p3gconsortium.org/index.cfm

[63] Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! *Am J Epidemiol* 1996;**144:**207–13.

[64] Prentice R, Vollmer W, Kalbfleicsh J. On the use of case series to identify disease risk factors, *Biometrics* 1984;**40:**445–58.

[65] Botto LD, Khoury MJ. Facing the challenge of complex genotypes and gene-environment interaction: the basic epidemiologic units in case-control and case-only designs. In: Khoury MJ, Little J, Burke W (eds). *Human Genome Epidemiology: a Scientific Foundation for Using Genetic Information to Improve Health and Prevent Disease*. New York: Oxford University Press, 2004, pp. 111–26.

[66] Weir HK, Thun MJ, Hankey BF *et al*. Annual report to the nation on the status of cancer, 1975–2000, featuring the uses of surveillance data for cancer prevention and control. *J Natl Cancer Inst* 2003;**95:**1276–99.

[67] Croen LA, Shaw GM, Jensvold NG *et al.* Birth defects monitoring in California: a resource for epidemiological research. *Paediatr Perinat Epidemiol* 1991;**5:**423–27.

[68] Correa-Villasenor A, Cragan J, Kucik J *et al*. The Metropolitan Atlanta Congenital Defects Program: 35 years of birth defects surveillance at the Centers for Disease Control and Prevention. *Birth Defects Res Part A Clin Mol Teratol* 2003;**67:**617–24.

[69] Mendoza JL, Murillo LS, Fernandez L *et al.* Prevalence of mutations of the NOD2/CARD15 gene and relation to phenotype in Spanish patients with Crohn disease. *Scand J Gastroenterol* 2003;**38:**1235–40.

[70] Le Marchand L, Seifried A, Lum-Jones A *et al.* Association of the cyclin D1 A870G polymorphism with advanced colorectal cancer. *JAMA* 2003;**290:**2843–48.

[71] Mickle JE, Cutting GR. Genotype-phenotype relationships in cystic fibrosis. *Med Clin North Am* 2000;**84:**597–607.

[72] Burke W. Genomics as a probe for disease biology. *N Engl J Med* 2003;**349:**969–74.

[73] Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;**361:** 865–72.

[74] Wacholder S, Chanock S, Garcia-Closas M *et al.* Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;**96:**434–42.

[75] Delamothe T, Smith R. Open access publishing takes off. *BMJ* 2004;**328:**1–3.

[76] Hoh J, Ott J. Mathematical multi-locus approaches to localizing complex human trait genes. *Natl Rev Genet* 2003;**4:**701–09.

[77] Greenland S. Principles of multilevel modelling. *Int J Epidemiol* 2000;**29:**158–67.

[78] Dunson D. Practical advantages of Bayesian analysis of epidemiologic data. *Am J Epidemiol* 2001;**153:**1222–26.

[79] Bhat A, Lucek PR, Ott J. Analysis of complex traits using neural networks. *Genet Epidemiol* 1999;**17(Suppl.1):**S503–07.

[80] Nelson MR, Kardia SL, Ferrell RE *et al.* A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res* 2001;**11:**458–70.

[81] Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 2003;**19:**376–82.

[82] Ritchie MD, Hahn LW, Roodi N *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; **69:**138–47.

[83] Zhang H, Yu CY, Singer B *et al.* Recursive partitioning for tumor classification with gene expression microarray data. *Proc Natl Acad Sci U S A* 2001;**98:**6730–35.

[84] Zhang H, Bonney G. Use of classification trees for association studies. *Genet Epidemiol* 2000;**19:**323–32.

[85] Hines LM, Stampfer MJ, Ma J *et al.* Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N Engl J Med* 2001;**344:**549–55.

[86] Little J, Khoury MJ. Mendelian randomisation: a new spin or real progress? *Lancet* 2003;**362:**930–31.

[87] Katan MB. Commentary: Mendelian randomization, 18 years on. *Int J Epidemiol* 2004;**33:**10–11.

[88] Keavney B. Commentary: Katan's remarkable foresight: genes and causality 18 years on. *Int J Epidemiol* 2004;**33:**11–14.

[89] Wheatley K, Gray R. Commentary: Mendelian randomization—an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *Int J Epidemiol* 2004;**33:**15–17.

[90] Brennan P. Commentary: Mendelian randomization and gene–environment interaction. *Int J Epidemiol* 2004;**33:**17–21.

[91] Thomas DC, Conti DV. Commentary: The concept of 'Mendelian Randomization'. *Int J Epidemiol* 2004;**33:**21–25.

[92] Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: Development of Mendelian randomization: from hypothesis test to 'Mendelian deconfounding'. *Int J Epidemiol* 2004;**33:**26–29.

[93] Davey Smith G, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol* 2004;**33:**30–42.

[94] Koplan JP, Thacker SB, Lezin NA. Epidemiology in the 21st century: calculation, communication, and intervention. *Am J Public Health* 1999;**89:**1153–5.