GENE–ENVIRONMENT INTERACTIONS IN HUMAN DISEASES

David J. Hunter

Abstract | Studies of gene–environment interactions aim to describe how genetic and environmental factors jointly influence the risk of developing a human disease. Gene–environment interactions can be described by using several models, which take into account the various ways in which genetic effects can be modified by environmental exposures, the number of levels of these exposures and the model on which the genetic effects are based. Choice of study design, sample size and genotyping technology influence the analysis and interpretation of observed gene–environment interactions. Current systems for reporting epidemiological studies make it difficult to assess whether the observed interactions are reproducible, so suggestions are made for improvements in this area.

PHARMACOGENETICS
The study of drug responses that are related to inherited genetic differences.

EPIDEMIOLOGY
A discipline that seeks to explain the extent to which factors that people are exposed to (environmental or genetic) influence their risk of disease, by means of population-based investigations.

Program in Molecular and Genetic Epidemiology, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA, and the Channing Laboratory, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.
e-mail: dhunter@hsph.harvard.edu
doi:10.1038/nrg1578

Since the days of Archibold Garrod[1], it has been increasingly accepted that the aetiology of most common diseases involves not only discrete genetic and environmental causes, but also interactions between the two. Garrod suggested that "the influences of diet and diseases" might "mask" some of the "inborn errors of metabolism" that he proposed, and that "idiosyncracies as regards drugs" were presumably due to inherited differences — thereby presaging the field of PHARMACOGENETICS by more than half a century.

In the context of medical genetics and EPIDEMIOLOGY, the study of gene–environment interactions is useful for several reasons (BOX 1). If we estimate only the separate contributions of genes and environment to a disease, and ignore their interactions, we will incorrectly estimate the proportion of the disease (the 'population attributable risk') that is explained by genes, the environment, and their joint effect. Restricting analysis of environmental factors in epidemiological studies to individuals who are genetically susceptible to the exposure should increase the magnitude of relative risks, increasing our confidence that the observed associations are not due to chance. The identification of susceptibility and/or resistance alleles in CANDIDATE-GENE STUDIES provides direct evidence that these genes and their biological pathways are relevant to specific diseases in humans. Understanding these pathways might help to determine which compounds in a complex mixture cause disease. Ultimately, understanding gene–environment interactions might allow us to give individualized preventive advice before disease diagnosis, in addition to offering personalized treatment after a disease, or disease susceptibility, has been diagnosed.

Some gene–environment interactions can be identified without any molecular analysis; one example is the much stronger effect of sunlight exposure on skin cancer risk in fair-skinned humans than in individuals with darker skin[2]. Others can be observed as a reproducible effect of an environmental exposure on a susceptible individual; for example, the flushing response seen after alcohol ingestion in individuals with low-activity polymorphisms in the aldehyde dehydrogenase gene[3]. However, our rapidly expanding ability, particularly after the completion of the Human Genome Project, to define genetic differences at the DNA-sequence level is opening up a vast new terrain in the search for gene–environment interactions.

Although the phrase 'gene–environment interaction' is frequently used to imply a specific relationship between genes and the environment, the many existing disease models differ with respect to the statistical association between genes and the environment. At least in part because of the many potential models of interaction, a gene–environment interaction will only be

CANDIDATE-GENE STUDIES
Studies of specific genes in which variation might influence the risk of a specific disease, usually because the gene is part of a biological pathway that is plausibly related to the disease.

BIOMARKER
A molecular marker of a biological function or external exposure.

ASSOCIATION STUDY
An approach to gene mapping that looks for associations between a particular phenotype and allelic variation in a population.

PRIOR PROBABILITY
An attempt to distinguish between more likely and less likely interactions on the basis of knowledge of biological mechanisms, before an interaction is observed.

accepted if it can be reproduced in two or more studies and also seems plausible at the biological level. Obtaining this evidence will necessitate a high degree of coordination between studies, and will require mechanisms for pooling unpublished data across studies, to prevent publication bias. High-throughput genotyping technologies such as whole-genome SNP scans hold the promise of finding the major genetic variants that contribute to the risk of common diseases over the next 5 to 10 years. Obtaining high-quality information on environment and lifestyle in conjunction with biological samples to assess these genetic variants will be crucial in the assessment of gene–environment interactions. Planning for these studies is needed now if reliable data on gene–environment interactions are to keep pace with rapidly emerging genetic knowledge.

This article reviews some of the challenges in the design and analysis of studies that are intended to uncover and confirm gene–environment interactions, and proposes some means by which data on the reproducibility of these interactions can be assessed before they are incorporated into public-health and clinical practice.

Describing gene–environment interactions

The study of gene–environment interactions requires information on both elements of the relationship. Genetic predisposition can be inferred from family history, phenotype (for example, skin colour), or direct analysis of DNA sequence. Environmental and lifestyle factors are measured in epidemiological studies using self-reported information; this can be obtained by interview or questionnaire, from records or direct measures in participants (for example, anthropometry), or BIOMARKER-based inference on environmental exposures. Until recently, many studies of genetic predisposition (for example, pedigree-based studies) obtained little information on environment and lifestyle. Similarly, many typical epidemiological studies of unrelated individuals (ASSOCIATION STUDIES) did not obtain blood samples or other sources of DNA that would allow direct assessment of genetic variation. More recently, in both family-based and association studies, collection of both genetic and environmental data, so that the interaction between the two can be examined, is becoming more common, and greater use of population-based designs in genetic epidemiology has been advocated[4]. As studies of genetic susceptibility and environmental exposures have been largely pursued by different groups of investigators, multidisciplinary collaboration is necessary to generate the best studies in the field.

However, even with well-designed studies, there are many ways of declaring 'success' in the search for interactions. This is largely because of variability in the qualitative and statistical models of interaction, and the difficulty of assessing biological plausibility (either *a priori* — that is, when trying to prioritize PRIOR PROBABILITIES in these analyses — or once an interaction has been observed). These factors are described below.

Qualitative models. In the simplest case of dichotomous genotype (such as carriers versus non-carriers of a gene variant) and dichotomous exposure (for example, exposed versus non-exposed), the four possible combinations of genotype and exposure can be displayed in a 2×4 table[5], and the relative risks can be shown in a graph such as that shown in FIG. 1A. However, even in this simplest case there are several models for describing interactions between genetic susceptibility and environmental exposures in different diseases[6,7] (BOX 2). The possibilities are more numerous if there are many categories of environmental exposure (for example, three or more categories of exposure; see FIG. 1B) and/or many genetic categories (as in the case of three genotypes for a biallelic system) or different genetic models (recessive, co-dominant and dominant) (BOX 2). So, many departures from the null result — where the risk of disease is the same in all cross-classified categories of exposure and genotype — might be compatible with the overall hypothesis of gene–environment interaction.

The 'multiple comparisons' problem that is inherent in examining thousands or even hundreds of thousands of SNPs in association studies is relatively familiar. For gene–environment interactions, however, we face a comparison problem that arises from a model involving multiple genes, multiple exposures and multiple interactions. In addition to using statistical approaches to control the false-positive rate, the reproducibility of gene–environment interactions across two or more studies will be crucial. Current models of publication of individual studies favour suppression of 'negative' results, leading to publication bias[8]. This is particularly problematic for interactions owing to the large number of comparisons being made and to the limited space in conventional publications for the main effects of genotype, let alone gene–environment interactions. Databases of results, and/or prior coordination of large studies, will be required to assess the reproducibility of gene–environment interactions.

Statistical models. In addition to the many qualitative models of interaction described, there are several methods of assessing the statistical significance of interactions[9]. In the simplest case of dichotomous environmental exposure and genotype, perhaps the most commonly

Box 1 | Rationales for the study of gene–environment interactions

- Obtain a better estimate of the population-attributable risk for genetic and environmental risk factors by accounting for their joint interactions.
- Strengthen the associations between environmental factors and diseases by examining these factors in genetically susceptible individuals.
- Help to dissect disease mechanisms in humans by using information on susceptibility (and resistance) genes to focus on the biological pathways that are most relevant to that disease, and the environmental factors that are most relevant to the pathways.
- Determine which specific compounds in the complex mixtures of compounds that humans are exposed to (such as diet or air pollution) cause disease.
- Use the information on biological pathways to design new preventive and therapeutic strategies.
- Offer tailored preventive advice that is based on the knowledge that an individual carries susceptibility or resistance alleles.

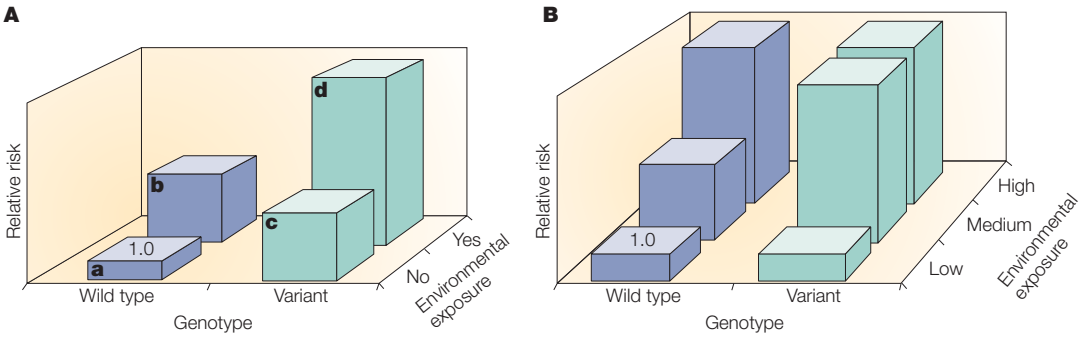


Figure 1 | **Models of gene–environment interactions. A** | In the most simplified example of a dichotomous genotype (for example, carriers versus non-carriers of an allele corresponding to a dominant trait), and dichotomous exposure (for example, 'exposed' versus 'non-exposed'), three categories of joint exposure can be compared with a reference category (for which the relative risk is, by definition, 1.0). Using this simple scheme, BOX 2 shows the different patterns of risk that are observed in some diseases in which inherited susceptibility clearly interacts with environmental exposures to jointly determine disease risk. In the example shown here, the relative risk of developing a disease is much greater in individuals who are both genetically susceptible to the condition and have been exposed to the environmental variable (cell **d**), than in individuals who carry the wild-type genotype and are not exposed to the environmental variable (cell **a**), or who are either only exposed to the environment or genetically susceptible (cells **b** and **c**, respectively). **B** | In the slightly more complex situation in which there are three categories of exposure, it has been proposed that genetically susceptible individuals could be at risk of disease at lower levels of exposure; in this model, the difference in risk between genotypes among individuals at the medium level of exposure is the only indication of an interaction.

used procedure is to test departure from the multiplicative model of interaction. This involves testing whether the relative risk for joint exposure (cell d in FIG. 1A) is statistically significantly greater ('supermultiplicative') or smaller ('submultiplicative') than would be expected by multiplying the relative risks for environmental exposure or genetic predisposition alone (that is, multiplying the relative risks of cells b and c in FIG. 1A). Another commonly used test for interaction uses rate differences rather than relative risks, and proposes that the joint effect of genes and the environment is different from the expectation that the incidence rate in cell d in FIG. 1A is described by adding the rates in cells b and c in FIG. 1A. This additive model is often said to be of greater relevance to assessing the public-health impact of an interaction[9]. Again, the option to use either of these two models adds further potential for multiple comparisons to the statistical analysis of gene–environment interactions.

Box 2 | **Some patterns of relative risk in gene–environment interactions**

The table shows just three examples of different patterns of relative risk for three classical genetic diseases that have an environmental component, assuming dichotomous genetic susceptibility and environmental exposure (the data are from REFS 5,6).

In the first example, xeroderma pigmentosum (XP), exposure to ultraviolet light increases the risk of developing skin cancer in non-carriers of XP mutations, but the combination of these mutations and exposure to ultraviolet light vastly increases the risk of skin cancer. In theory, if individuals with XP mutations completely avoid ultraviolet light their risk of skin cancer becomes close to the background risk.

The example in the second column is that of phenylketonuria (PKU); only individuals with recessive mutations in the causative gene (phenylalanine hydroxylase) that are exposed to phenylalanine in the diet are susceptible to PKU.

In the third column, exemplified by a deficiency in the α-1 antitrysin gene, both non-smokers that are at genetic risk and smokers that are not at genetic risk have an increased risk of developing emphysema, and the combination (smokers that are at genetic risk) is associated with the highest risk.

There are many other patterns of gene–environment interactions, including 'protective' alleles and exposures.

Gene variant	Environmental exposure	Relative risk (XP)	Relative risk (PKU)	Relative risk (emphysema)
Absent	Absent	1.0	1.0	1.0
Present	Absent	~1.0	1.0	Modest
Absent	Present	Modest	1.0	Modest
Present	Present	Very high	Very high	High

Biological plausibility. Screening a large number of potential gene–environment interactions in datasets with a large number of genotypes and many variables of exposure greatly increases the chance of finding false-positive results at conventional levels of statistical significance. As most studies are not powerful enough to detect modest interactions, demanding small *p* values to counteract this problem will result in a lower probability of declaring true-positive interactions as 'significant'. Restricting the search for gene–environment interactions to those that involve gene products and exposures that plausibly interact in the same biological pathways is an attractive option. Furthermore, restricting analysis to gene variants that plausibly alter gene function is also attractive, although for variants that affect gene regulation this science is in its infancy[10]. However, defining plausibility *a priori* has a large subjective component, and one person's 'plausible candidate' might be another person's 'low-probability hypothesis'.

Study designs for gene–environment interactions

Genetic epidemiology has been dominated by the use of family-based designs from which inherited susceptibility can be inferred. However, with the advent of methods for assessing DNA-sequence variability directly, association studies using unrelated individuals are increasingly being used.

Family-based studies. By comparing disease concordance rates between monozygotic and dizygotic twins, twin studies can be used to partition components of VARIANCE between genetic and shared and non-shared environmental factors[11]. Most reports of studies from twin registries do not include information on environmental exposures that could be shared (or different) between the twins, precluding any inferences about specific gene–environment interactions.

Analyses of multigenerational pedigrees might provide a preliminary assessment of the hypothesis that the PENETRANCE of a mutation has changed over chronological time, which would indicate that changes in lifestyle and environment have influenced gene penetrance. For example, in an analysis of 333 North American women who were carriers of *BRCA1* (breast cancer 1, early onset) mutations, mainly from high-risk breast- and ovarian-cancer families, penetrance increased with more recent birth cohorts[12]. This indicates the influence of environmental and lifestyle factors that are more prevalent in recent birth cohorts, although it does not provide direct clues about specific factors. A further limitation of this approach is that assessments of this nature can only be made for relatively highly penetrant gene mutations (that is, where the penetrance is sufficiently high to cause clear familial aggregation).

Incorporation of environmental data into pedigree or other family-based designs (for example, studies that use sib-pairs or case–parent designs) allows direct estimates of specific gene–environment interactions. In some cases, fewer matched sets might be required for these designs than for case–control studies using unrelated controls[13]. Collection of adequate numbers of sib-pairs, however, might take more effort than the use of unrelated controls and, for late-onset diseases, the availability of living parents might limit case–parent accrual.

Association studies in unrelated individuals. Epidemiology has been remarkably successful at identifying the main risk factors for many common diseases; use of the best available study designs and data-collection methods has been important in this success. The relative merits of population-based epidemiological studies are well established. However, the search for gene–environment interactions imposes some further constraints on the use of these designs (outlined in TABLE 1). In retrospective case–control studies, data on environmental and lifestyle factors, and samples for DNA and biomarker studies, are obtained after diagnosis of disease in the cases. In prospective cohort studies, environmental and lifestyle data are obtained at baseline (the start of the study), and ideally at other points before diagnosis. Samples for DNA and biomarker studies are also ideally obtained at baseline, although in prospective studies that do not have banked samples, DNA can be obtained after diagnosis from living cohort members. It should also be noted that, under certain assumptions, gene–environment interactions can be estimated from case–case studies without controls[14].

Association studies: retrospective design. The main limitations of retrospective studies, particularly case–control studies, are well described[15]; for example, selection bias (in particular, the use of controls who do not represent the population in which the cases occurred) is an important potential problem. If the race or ethnicity of the controls is substantially different from that of the cases, then spurious associations with gene variants that differ by race or ethnicity (that is, POPULATION STRATIFICATION[16,17]) will occur. The potential influence of population stratification is still controversial, with some authorities in the field pointing out that it can be substantially eliminated with attention to appropriate choice of controls and by controlling for self-reported ethnicity[16,18]. Methods to assess the population substructure of cases and controls by genotyping non-causal gene variants ('genomic control') have been proposed, and can be used to correct for this phenomenon[19,20].

Even if the optimal population of controls can be identified, obtaining a high participation rate can be challenging, particularly as genetic studies also require a blood or buccal sample for DNA analysis. Although estimates of interaction parameters might be unbiased even if the main effects of genotype and environment are distorted by selection bias[21], this is of little comfort as we are usually interested in obtaining estimates of both the main effects and their interaction. In a rapidly fatal disease, only a fraction of the original cases might be available for interview, leading to 'survivor bias' if genotype or exposures differ between those who succumb quickly compared with longer-term survivors. With respect to gene–environment interactions, the principal problem is likely to consist of misclassified ('noisy') or biased information on environmental exposures. Bias can arise if cases report their pre-diagnosis exposure histories differently once they are diagnosed with the disease compared with what they would have reported before diagnosis (recall bias). Interestingly, however, the presence of a biased main effect for the environmental factor does not automatically imply a biased estimate of gene–environment interaction[22].

Of more concern in retrospective studies is the possibility that poor recall (misclassification) of past exposures among both cases and controls might attenuate the estimates of risk to the point where any difference in risk according to genotype cannot be reliably detected. The potential for these biases and misclassification can be reduced, but rarely eliminated, by paying careful attention to best practices in enrollment and exposure assessment. The chief advantage of case–control studies is the potential for the sample size to be limited only by cost and by the number of cases of the disease that are available in the study area. Given the need for large sample sizes (see below) in gene–environment studies, this is a great potential advantage, which, when combined with the potential for increased detail of exposure assessment and disease phenotype (see below), might make the case–control study the design of choice for some diseases.

VARIANCE
A statistic that quantifies the dispersion of data about the mean. In quantitative genetics, the phenotypic variance (*V*p) is the observed variation of a trait in a population. *V*p can be partitioned into components, owing to genetic variance (*V*g), environmental variance (*V*e) and gene–environment correlations and interactions.

PENETRANCE
The frequency with which individuals that carry a given gene variant will show the manifestations associated with that variant. If penetrance of a disease allele is 100% then all individuals carrying that allele will express the associated disorder.

POPULATION STRATIFICATION
The presence of multiple subgroups with different allele frequencies within a population. The different underlying allele frequencies in sampled subgroups might be independent of the disease within each group, and they can lead to erroneous conclusions of linkage disequilibrium or disease relevance.

Association studies: prospective design. The problems of selection and recall bias in case–control studies can be minimized in prospective studies. Here, DNA samples and exposure information are obtained from participants in a longitudinal cohort who are followed up, usually for years or decades. If follow-up rates are high, then a virtually complete set of cases can be assembled and compared with a sample of individuals who did not develop the disease. The use of this 'nested' case–control study minimizes selection bias because the population that gave rise to the cases is defined. Because information on exposures is collected before diagnosis (in most cases, years to decades before), recall bias is eliminated as knowledge of diagnosis cannot influence the reporting of exposures. However, particularly in the many cohort studies that only have a baseline assessment and do not involve repeated measurements during follow up, a single measure of an exposure might not be a good reflection of the pattern of exposure over time.

A variant of the true nested case–control approach might be useful in studies that include prospectively collected environmental and lifestyle data, but that lack a source of DNA for genetic analyses. In this design, an attempt is made to obtain a DNA sample from cases arising in the cohort, and from matched or unmatched non-cases. Although the environmental data should be secure from recall bias in this design, failure to obtain DNA samples from a high proportion of cases and controls can result in selection bias in the same manner as in a conventional case–control study. Differential participation by cases and controls according to ethnicity could give rise to population stratification, particularly in populations whose ancestors have been recently mixed by intercontinental migration[17]. However, again this bias can be minimized by collecting information on and controlling for ethnic background, or by using genomic control methods.

The principal problem with prospective studies is that adequate sample sizes of cases will only be obtained for common conditions, such as hypertension, myocardial infarction and stroke, and common cancers, in the population that is being followed. Rare diseases, such as sarcomas, will not occur at sufficient frequency to provide statistical power, so retrospective studies are essentially the only option. A typical cohort study might only accrue several hundred cases of a disease of moderate incidence (for example, Parkinson disease) over many years, and because most cohort studies enroll men and women in middle life, diseases with relatively early onset (for example, multiple sclerosis) will be under-represented. In addition, some of the special requirements for genomic analyses might only be met in case–control studies. For certain diseases, particularly cancers, expression-array analyses indicate that cancer types that look histologically similar might represent more than one disease process[23]. Obtaining the fresh-frozen tissue or tumour blocks necessary to subtype these outcomes might be difficult in prospective studies, but more feasible in cases that are studied in a limited number of institutions.

Phenotypic assays, such as assays that measure the activity of an enzyme or biochemical pathway by giving a test dose of a compound and measuring metabolites in blood, might be possible in a limited number of cases and controls, but are unlikely to be feasible in large prospective studies. Information on disease diagnosis and subtype from non-genomic tests, such as histology

Table 1 | **Observational study designs for the analysis of gene–environment interactions**

Characteristics	Nested case–control 1*	Nested case–control 2‡	Case–control§
Potential for selection bias	Low if follow-up rate is high	Moderate if DNA is not obtained from almost all cases	Moderate to high
Population stratification	Minimized by sampling from a defined cohort	Possible if DNA is not obtained from all cases and controls; sampling from a defined cohort should reduce the risk	Risk potentially increases if source population for controls is hard to define; controlling for ethnicity should reduce the risk; genomic control methods might also mitigate the risk
Survivor bias	Nil	Moderate potential if DNA is not obtained from all cases and controls	Moderate to high potential
Recall bias	Nil	Nil	Moderate to high potential
Environmental exposure detail	Usually limited	Usually limited	Potentially higher than alternative designs
Ability to use plasma phenotypes	Yes	No	No
Disease-phenotype subtyping	Might be possible from medical records; tissues are often difficult to obtain	Might be possible from medical records; tissues are often difficult to obtain	Medical records and tissues might be easier to obtain; diagnostic procedures might be more uniform
Achievable sample sizes			
Common diseases	Adequate if the follow up is sufficiently long	Adequate if the follow up is sufficiently long	Depends mainly on the funds that are available
Rare diseases	Inadequate unless data are pooled across many studies	Inadequate unless data are pooled across many studies	Might require enrollment at many centres

*Prospective collection of both environmental and lifestyle data and DNA. ‡Prospective collection of environmental and lifestyle data, retrospective DNA collection. §Retrospective collection of both environmental and lifestyle data and DNA.

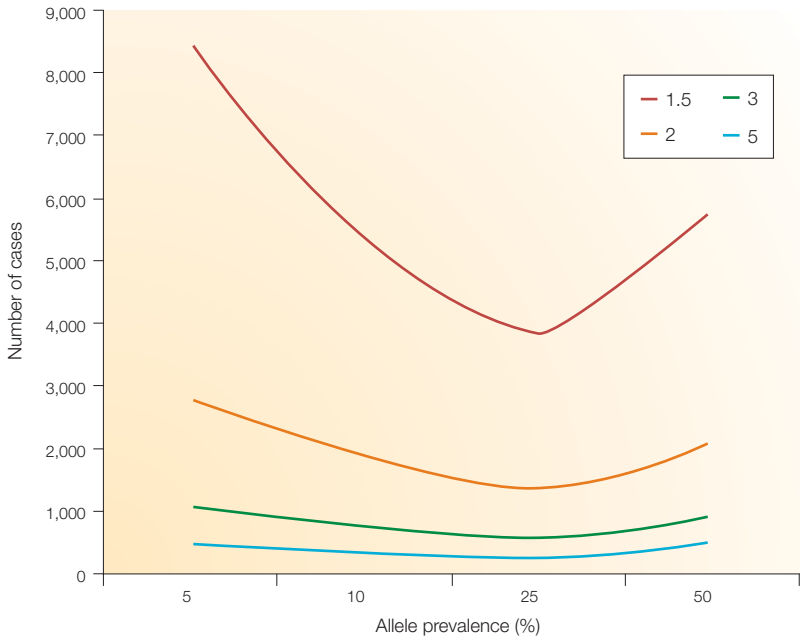


Figure 2 | **Number of cases needed to detect a range of multiplicative interactions, according to allele prevalence.** The model assumes the following: a dominant genetic model, a dichotomous exposure prevalence of 10%, a relative risk for a genotype of 1.5, a relative risk for exposure of 1.5 and a 1:1 case:control ratio. As the graph shows, thousands of cases and controls are needed to detect interactions with relative risks of 1.5 and 2. Calculations were carried out using Quanto Beta version 0.5 (REF. 13).

or imaging, might also be hard to obtain in a uniform manner in a prospective study, in which almost all cases might be diagnosed at different institutions, as opposed to a case–control study that operates in a limited number of hospitals.

Association studies: case-only designs. It has been shown that when a genotype is not correlated with an environmental factor and a disease is rare (few 'controls' are likely to have undiagnosed or incipient disease), then departure from multiplicative interaction can be tested by examining information from the cases only[14,24]. In this case–case design, the prevalence of the exposure in the genotype-positive cases would be expected to be the same as the prevalence of the exposure in the genotype-negative cases. Statistically significant departures from this expectation of equal prevalence indicate an interaction between genotype and exposure.

The idea of dispensing with the need to identify appropriate controls and enroll them is attractive. However, we are usually interested in obtaining an estimate of the main effect of genotype (particularly given the explosion of potential candidate genotypes as our knowledge of gene function and common gene variants increases), and we are frequently interested in new, or as yet unproven, environmental hypotheses. To estimate these an appropriate control group is needed. In addition, for high-penetrance genes, the assumption that the disease is rare among exposed individuals is violated, leading to a distortion of the interaction estimates[25]. Furthermore, relatively modest violations of the assumption of independence between genotype and exposure can have a substantial impact on bias relating to the interaction parameters[26]. It is possible that, as we identify the main genetic influences on common diseases, case–case methods might become more popular in assessing the interaction of established causal genotypes with environmental factors, particularly as these studies can form the baseline cohort for finding prognostic markers of disease outcomes.

HAPLOTYPE-TAGGING SNP
One of a small subset of SNPs that is needed to uniquely identify a complete haplotype.

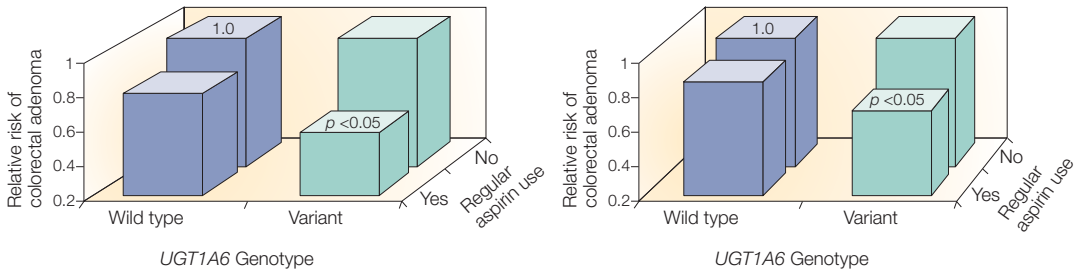
Technical challenges

Sample size. Some of the distinctions between retrospective and prospective studies can seem like methodological niceties compared with the biggest problem in the field — the need for large sample sizes. A long-standing rule of thumb for calculating sample sizes has been that the sample size required to detect a departure from the expectation that the joint effect of two variables is multiplicative is at least four times the sample size that is needed to evaluate the main effect of each of the variables[27]. FIGURE 2 illustrates some estimates of the sample sizes that are needed for alleles with different frequencies, with optimistic assumptions about the frequency of genotypes, exposure and interaction effects. If information on exposures is misclassified, then the power to detect interactions is attenuated, and even larger sample sizes are required[28,29]. As many epidemiological studies are underpowered for main effects, this predicts that they will be seriously underpowered to detect interactions. Therefore, a substantial problem for the foreseeable future is the occurrence of false-negative findings for interactions in individual studies, unless the interactions are strong.

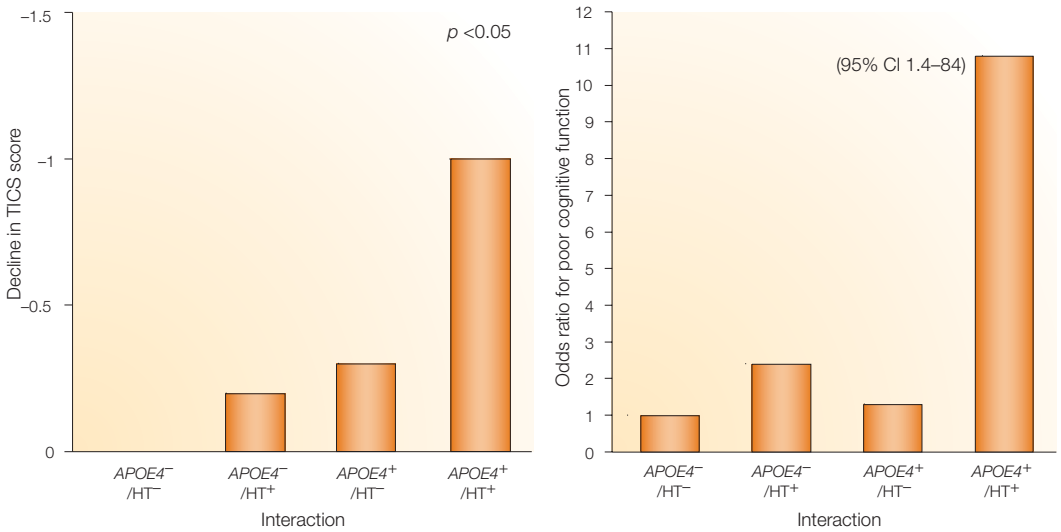
It has been pointed out that most cohort studies will not accrue sufficient numbers of cases for rare diseases and might have only marginal power for common diseases[30]. Therefore, rigorously designed case–control studies will remain the only option for assessing gene–environment interactions for rare diseases, and should be considered as a cost-effective approach for common diseases. A potentially effective means of mitigating the lack of power in prospective studies is to pool data across these studies. For example, the US National Cancer Institute (NCI) Breast and Prostate Cancer and Hormone-Related Gene Variants Cohort Consortium is examining gene–environment interactions in over 6,000 cases of breast cancer and 8,000 cases of prostate cancer, pooled across 10 prospective studies; over 800,000 people are being followed up and over 7 million years of life of follow up have already been accrued. A further advantage of coordinated pooling of data is that genotyping can be standardized to ensure inter-study comparability of the genetic variants that are ascertained. The emerging interest in assessing HAPLOTYPE-TAGGING SNPs makes prior coordination particularly desirable, as it might not be possible to pool data from studies that have used different SNPs to capture haplotypes. Maximizing the power of ongoing prospective studies in this manner can mitigate the main weakness of prospective studies (that is, the limited number of incident cases) while capitalizing on the methodological strengths of the prospective design.

Box 3 | Examples of gene–environment interactions from epidemiological studies

a Inverse association of aspirin on colorectal adenomas by *UGT1A6* genotype



b Untreated hypertension, *APOE4* carriers and risk of cogitive impairment



A growing number of studies have reported that polymorphisms in drug-metabolism genes alter responses to therapy (for a review see REF. 40). There are fewer examples of polymorphisms that are known to affect the response to drugs that are being given for chemoprevention. It could be argued that these will be even more important to find, as the potential for harm owing to toxicity is greater in healthy individuals who are being given a drug for prophylaxis rather than for therapeutic purposes. Furthermore, without a parameter for measuring short-term therapeutic response (for example, blood pressure for an antihypertensive drug) to allow dose titration or to indicate a switch from an ineffective drug, inter-person differences in response to long-term preventive drugs cannot be clinically measured. A functional genotypic variant at the *UGT1A6* (UDP glycosyltransferase 1 family, polypeptide A6) locus modifies the protective association of aspirin on colorectal polyp formation. In two studies (see panel **a**; data for the figure on the left are from REF. 57 and on the right are from REF. 58), the protective association was essentially limited to the group with the 'slow' allele of *UGT1A6*, which takes longer to metabolize aspirin[57,58]. Further studies are needed to see whether this variant also modifies the effects of aspirin on gastrointestinal bleeding, haemorrhagic stroke and prevention of heart attack. The risk–benefit balance of prophylactic aspirin use could depend on aspirin-metabolism genotype.

The reproducible association of common polymorphic gene variants with risk of disease does not automatically imply that there will be any benefit to assessing these variants in clinical or public-health practice. For example, perhaps the most robust association of a common variant with a serious disease is the association of the apolipoprotein E4 (*APOE4*) allele with risk of cognitive decline and Alzheimer disease. Expert panels that were asked to assess whether it is worth screening for carriers of the *APOE4* allele have, however, almost uniformly recommended against this. Their argument is that, in the absence of an intervention that reduces risk in *APOE4* carriers, screening would constitute 'prediction without promise'; it would expose people to the potential psychological and social risks of learning they have a high risk of disease without offering them any means to reduce this risk[59]. However, recent studies suggest that the risk of cognitive decline is particularly high in *APOE4* carriers who have untreated hypertension (the $APOE4^+/HT^+$ group in panel **b**; the panel shows the interaction observed in two studies of the *APOE4* allele and untreated hypertension with cognitive function; TICS is a test of cognitive function; data on the right are from REF. 60 and data on the left are from REF. 61). Although it could be argued that treatment of high blood pressure is a uniform recommendation that should not depend on any specific genotype, only a proportion of people with high blood pressure are treated. Motivation for blood-pressure control might be higher among people who are at elevated risk for dementia, and would therefore justify screening for an allele that does not otherwise meet the criteria for a useful screening measure. CI, confidence interval; *p*, the probability that the observed interaction is due to chance.

Some applications

Assessing inference in complex mixtures. An enduring problem in environmental epidemiology is deciding which components of 'complex mixtures' — such as air pollution, diet or cigarette smoke — cause disease. This is difficult to study observationally as most components of complex mixtures are highly correlated, so that their effects cannot be statistically separated. If the effect of the environmental factor differs according to variation in one or more specific genes, then the function of the gene might help to isolate the causal components in the complex mixture. For example, heterocyclic amines formed by the high-temperature cooking of proteins are carcinogenic in animal models, and are sometimes found in grilled and pan-fried meats[31]. Red meat intake has been consistently associated with risk of colorectal cancer. However, the chemical components of red meat that are responsible for this risk — such as fatty acids, haem iron or protein — are unclear. Exposure to heterocyclic amines is one hypothesis, but obtaining information on the degree to which meats are usually cooked is problematic in epidemiological studies. Some studies[32–35], but not all[36], have found that the association of red meat intake with colorectal neoplasia is stronger in carriers of the 'rapid' *NAT2* (*N*-acetyltransferase 2) alleles, which have been shown to be associated with a faster metabolism of various substrates, including heterocyclic amines. As the heterocyclic amines are metabolized by NAT2, and are probably therefore the red-meat-specific substrate, this indicates that heterocyclic amines are the causal carcinogens in red meat. In this manner, the finding of an interaction between exposure to a complex mixture and a specific variant of a metabolic gene 'points the finger' at the substrates of the gene as the causal components of the complex mixture. This might also lead to the identification of unsuspected disease-causing exposures[37].

The related concept of 'Mendelian randomization' has been used to argue that a reproducible effect on disease risk of a genotype that alters the level of an intermediate biomarker indicates that the relation of the biomarker to disease risk is unlikely to be confounded by other lifestyle variables, because in most cases, these other lifestyle variables would not be expected to correlate with genetic variation[38,39].

Table 2 | **Selected examples of gene–environment interactions observed in at least two studies**

Gene symbol	Variant(s)	Environmental exposure	Outcome and nature of interaction	References
Genes for skin pigmentation (for example, *MC1R*)	Variants for fair skin colour	Sunlight or ultraviolet light B	Risk of skin cancer is higher in people with fair skin colour that are exposed to higher amounts of sunlight	62
CCR5	Δ-32 deletion	HIV	Carriers of the receptor deletion have lower rates of HIV infection and disease progression	41
MTHFR	Ala222Val polymorphism	Folic acid intake	Homozygotes for the low activity Ala222Val variant are at different risk of colorectal cancer and adenomas if nutritional folate status is low	63
NAT2	Rapid versus slow acetylator SNPs	Heterocyclic amines in cooked meat	Red meat intake is more strongly associated with colorectal cancer among rapid acetylators	33
F5	Leiden prothrombotic variant	Hormone replacement	Venous thromboembolism risk is increased in factor V Leiden carriers who take exogenous steroid hormones	64
UGT1A6	Slow-metabolism SNPs	Aspirin	Increased benefit of prophylactic aspirin use in carriers of the slow metabolism variants	58
APOE	*E4* allele	Cholesterol intake	Exaggerated changes in serum cholesterol in response to dietary cholesterol changes in *APOE4* carriers	65
ADH1C	γ-2 alleles	Alcohol intake	Inverse association between ethanol intake and myocardial infarction; risk is stronger in carriers of slow-oxidizing γ-2 alleles	66
PPARG2	Pro12Ala	Dietary fat intake	Stronger relation between dietary fat intake and obesity in carriers of the Pro12Ala allele	67
HLA-DPB1	Glu69	Occupational beryllium	Exposed workers who are carriers of the Glu69 allele are more likely to develop chronic beryllium lung disease	68
TPMT	Ala154Thr and Tyr240Cys	Thiopurine drugs	Homozygotes for the low-activity alleles of *TPMT* are likely to experience severe toxicity when exposed to thiopurine drugs	69
ADRB2	Arg16Gly	Asthma drugs	Arg16Gly homozygotes have a greater response in the airway to albuterol	70

ADH1C, alcohol dehydrogenase 1C (class I), γ-polypeptide; *ADRB2*, adrenergic, β-2-, receptor, surface; *CCR5*, chemokine (C–C motif) receptor 5; *APOE*, apolipoprotein E; *F5*, coagulation factor V; HIV, human immunodeficiency virus; *HLA-DPB1*, major histocompatibility complex, class II, DP β-1; *MC1R*, melanocortin receptor 1; *MTHFR*, 5,10-methylenetetrahydrofolate reductase (NADPH); *NAT2*, *N*-acetyltransferase 2; *PPARG2*, peroxisome proliferative activated receptor-γ; *TPMT*, thiopurine *S*-methyltransferase; *UGT1A6*, UDP glycosyltransferase 1 family, polypeptide A6.

LINKAGE DISEQUILIBRIUM (LD). A measure of whether alleles at two loci co-exist in a population in a non-random fashion. Alleles that are in LD are found together on the same haplotype more often than would be expected by chance.

Pharmacogenetics of chemoprevention ('personalized prevention'). The field of pharmacogenetics is a special case of gene–environment interaction in which the environmental exposure (a drug) is usually well measured, or even randomly assigned (in the context of a randomized clinical trial). This area has been extensively reviewed[40], and has the potential to identify individuals who are at risk for adverse drug reactions or treatment failure; these individuals can then either avoid exposure to the drug or have their dose modified. It has been argued that, with the exception of life-threatening illnesses or avoidance of severe toxic reactions, the ability to predict drug response in therapeutic situations might not have compelling advantages over the standard algorithms that titrate drug type and dose to the clinical response.

However, in chemoprevention applications, in which drugs are given to large numbers or whole populations of healthy people to prevent a future disease, even a few adverse reactions could tip the risk–benefit balance away from benefit. Similarly, if there is a group in which the benefit is not substantial (for example, wild-type homozygotes at the *UGT1A6* (UDP glycosyltransferase 1 family, polypeptide A6) gene obtain a lesser reduction of colorectal adenoma risk from taking aspirin), then the intervention risk–benefit might not favour these people (BOX 3A). Therefore, tailoring drug dose to genetic profile might be necessary to maximize the risk–benefit ratio in chemoprevention. Even for interventions that would routinely be recommended across the board (for example, the treatment of hypertension), there might be some benefit in identifying people in whom this treatment might confer further benefits (for example, among carriers of the apolipoprotein E4 (*APOE4*) allele; see BOX 3 figure part b).

Infectious diseases. Among the better-understood disease-causing components of the environment are infectious-disease agents, which we often assume to be the sole determinants of the corresponding infectious diseases. There is increasing evidence that some inter-population and inter-individual differences in the attack rate and prognosis of specific infectious organisms are due to inherited genetic variants. Perhaps the best recent example is the role of a 32-bp deletion in the HIV co-receptor chemokine (C–C motif) receptor 5 (*CCR5*) gene in blocking HIV infection in the homozygous state and slowing disease progression in heterozygotes[41]. The recognition of this gene–environment interaction confirmed the crucial importance of the CCR5 receptor in humans that are exposed to HIV. In addition, the existence of healthy individuals who carried the homozygous deletion that abrogates CCR5 function implied that drugs that block CCR5 would not cause side effects related to immune deficiency. A new class of chemokine receptor antagonist drugs designed to mimic the inherited deletions in these receptors is being studied for activity in slowing HIV disease progression[42]. Therefore, understanding the biological interaction of an inherited polymorphism with an infectious organism can suggest new therapeutic strategies.

Box 4 | **Future needs for the study of gene–environment interactions**

- Increase the power of analyses for common diseases by studying larger cohorts over a longer time.
- Add DNA collection and informed consent to ongoing prospective studies that do not have biobanks, particularly studies among minority groups that are under-represented in current studies.
- Start new prospective studies to replace the current generation of studies.
- Coordinate continuing and future studies to ensure maximum compatibility of the genetic and environmental information obtained.
- Encourage mechanisms for presenting unpublished and unpublishable results that are an inevitable result of the large amounts of data on interactions made available from these large, long-term studies.
- Ensure that well-designed case–control studies of less common diseases collect DNA samples and obtain appropriate informed consent.
- Refine and validate methods for whole-genome amplification, and associated informed consent, to ensure the maximum benefit from current and future studies.
- Develop new statistical approaches to integrate epidemiological data with information on biological processes that are obtained from applying systems-biology approaches.

How will we get the interactions right?

The need for coordination. Despite much information on both genetic and environmental disease-risk factors, there are relatively few examples of robust, replicated gene–environment interactions in the epidemiological literature (some examples are given in TABLE 2). The main reason is that many individual studies have been designed to examine the main effects of individual factors and do not have adequate power to examine interactions.

Even then, to convincingly show the main effect of a single factor might require a meta-analysis of many studies, and it is uncommon for this level of detail to be available for interactions. As replication of results will be even more important for interpreting interactions — the large number of comparisons that will be required will increase the inherent potential for false positives — it will be necessary to obtain data that will allow pooling of results across many studies. These data are unlikely to be routinely available through the published literature, as only a small subset of interaction analyses, if any, are usually included in a published article. There are two approaches to mitigate this problem: to facilitate web-based presentation of unpublished results in supplementary tables[43], and to pre-plan analyses across many studies so that the data are analysed and displayed in as uniform a format as possible. The latter approach is a prospective variant of the 'meta-analysis of individual participants' data' approach that has been reviewed in the context of genetic epidemiology studies, with the extra advantage that pre-planned analyses allow more consistent treatment of LINKAGE DISEQUILIBRIUM (LD) and haplotype definition[44]. This necessity for coordination and other future needs are presented in BOX 4.

The study of gene–environment interactions has at least one advantage over that of conventional two-way environmental interactions because it should be possible to measure a defined functional genetic polymorphism almost without error. However, when several

polymorphisms in a gene contribute to altered function, measuring a subset will result in misclassification and so will increase the sample size that is required to detect interactions[45]. Furthermore, if we do not know the functional gene variants — for example, if we are trying to detect genetic association through LD — it is likely that there will be substantial misclassification of the genetic variable, leading to dilution of the relative risk for the interaction.

With a plethora of imminent association studies owing to the rapid expansion in genotyping capacity, different researchers will probably genotype different sites in the same gene; this will lead to difficulties in assessing replication of genetic main effects and of gene–environment interactions. With the increasing use of haplotype-tagging or LD-tagging SNPs to explore genetic associations in candidate genes and regions, there is even more potential for incompatible information. Some degree of coordination of the main studies in each disease area would at least reduce the potential for incompatibility of information, and could hasten the confirmation of replication of both genetic effects and gene–environment interactions. The NCI Breast and Prostate Cancer and Hormone-Related Cohort Consortium, for example, is a planned assessment of the same genetic variants in 53 candidate genes across 10 studies that collectively contribute more than 6,000 cases of breast cancer and 8,000 cases of prostate cancer. Consortia such as this have the potential to provide much more uniform data and analyses than are available through *post-hoc* or literature analyses.

The need for new studies. Although a substantial number of people around the world have already given DNA samples as part of long-term prospective studies, many high-quality continuing prospective studies do not have a source of DNA from most, or any, participants. A highly cost-effective way of increasing the number of studies that can contribute to future gene–environment interaction analyses would be to collect samples for future prospective analyses in those already established studies in which it can be demonstrated that this can be done efficiently. The development of relatively simple methods for obtaining buccal-cell DNA through a mouthwash method has greatly expanded the potential for obtaining DNA in these studies[46,47], although a blood sample is usually preferred because of the potential to measure exposures using plasma or serum biomarkers.

Collecting specimens in ongoing studies offers the opportunity of capitalizing on many years of follow up and environmental data collection and phenotyping. New prospective studies that are being planned, such as the UK BioBank and the US National Institutes of Health AGES study[48], will also be needed to replace the current generation of studies. However, feasibility and informed-consent issues will be important challenges to these large centralized studies. For diseases that will be simply too rare for study through prospective mechanisms, population-based case–control studies that involve many research centres will need to be carried out.

New technologies. The ability to carry out large-scale SNP analysis and to pool data across studies has been hampered by concerns about 'running out of DNA'. Few studies, and none of the large studies, have the resources to establish cell lines from cohort participants. However, recent results from new whole-genome amplification techniques indicate that nanogram amounts of DNA can be amplified to microgram amounts, without altering the genotypes obtained before and after amplification[49–51]. This technique has the potential to revolutionize our ability to ask questions about gene–environment interactions across many studies. However, it will be necessary to establish mechanisms for collaboration across studies that operate within the boundaries of the original informed consent given by study participants, and to keep the environmental and lifestyle data confidential.

New statistical methods will be needed to extract meaning from large data sets and to incorporate knowledge from other branches of science. SYSTEMS BIOLOGY approaches to integrating 'omics' information from many sources is predicted to lead to new insights about cellular and whole-organism function[52]. This information will have to be integrated into the interpretation of studies of genes and the environment. Incorporating pharmacokinetic knowledge of specific biochemical pathways has been proposed as the first step in this direction for carrying out epidemiological studies of gene variants in these pathways[53]. Predicting the probability that a SNP alters function on the basis of phylogenetic or biochemical data, or from predicted effects on protein structure, might help to determine which SNPs to genotype, as well as the interpretation of subsequent results[54]. Reporting of false-positive probability by incorporating the prior probability of an interaction might also be helpful for reducing false positives in the literature[55]. As we move from a field that is accustomed to hypothesis-testing to a more neutral data-mining approach, large changes in the philosophies and methods of statistical analysis that are applied to epidemiological data will be required to cope with these issues of scale and disparate data sources.

Future prospects

A common model for future preventive health care proposes that, initially, physicians will test their patients for hundreds or thousands of genetic variants, and that ultimately we will all have our entire genome sequence on a card or chip. Advice on disease prevention will be based on this information, implying that the relevant gene–environment interactions will have been proposed, replicated and validated, so that this advice is evidence-based. We face the prospect that affordable individual genome sequencing will be the easy part; developing a credible database on replicable gene–environment interactions will be the challenge.

In addition, the concept of 'personalized prevention' might also seem to conflict with the view, articulated by Geoffrey Rose[56], and others, that population-wide interventions are usually more effective in reducing the incidence of common diseases than interventions that target

SYSTEMS BIOLOGY
The study of the complex interactions that occur at all levels of biological information — from whole-genome sequence interactions to developmental and biochemical networks — and their functional relationship to the phenotypes of organisms.

high-risk individuals. We could imagine that the idea that inherited susceptibility as a chief determinant of disease risk could increase latent feelings of genetic determinism and undermine support for 'broad-brush' and 'one-size-fits-all' preventive recommendations that are the cornerstone of many public-health campaigns. Past experiences, such as screening programmes for sickle-cell anaemia, warn of the complexity of extending genetic testing beyond the counselling-intensive, high-penetrance disorders. The financial costs, and potential psychological consequences, of genotyping individuals to make preventive recommendations are still uncertain and need to be established. It seems inevitable that some DNA-based tests will become part of adult risk-factor and susceptibility screening, but the uptake of these tests should be highly dependent on the development of proven interventions to take advantage of this knowledge. However, the scope of new knowledge that is likely to be uncovered by incorporating information on genetic variation among individuals into epidemiological studies of disease risk is likely to be vast. An attempt to distinguish between more likely and less likely interactions on the basis of knowledge of biological mechanisms, before an interaction is observed, but whether the interventions are lifestyle changes or drugs, modification of inherited susceptibility by altering environmental exposures is likely to become an accepted part of future public-health and clinical practice.

1. Garrod, A. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* **2**, 1616–1620 (1902).
A classical work of intuition, in which Garrod inferred the existence of inherited biochemical characteristics from family histories of alkaptonuria at the turn of the last century.
2. Green, A. & Trichopoulos, D. in *Textbook of Cancer Epidemiology* (eds Adami, H., Hunter, D. & Trichopoulos, D.) 281–300 (Oxford Univ. Press, Oxford, 2002).
3. Takeshita, T., Mao, X. Q. & Morimoto, K. The contribution of polymorphism in the alcohol dehydrogenase-β subunit to alcohol sensitivity in a Japanese population. *Hum. Genet.* **97**, 409–413 (1996).
4. Thomas, D. C. Genetic epidemiology with a capital 'E'. *Genet. Epidemiol.* **19**, 289–300 (2000).
5. Botto, L. D. & Khoury, M. J. Commentary. Facing the challenge of gene–environment interaction: the two-by-four table and beyond. *Am. J. Epidemiol.* **153**, 1016–1020 (2001).
6. Khoury, M. J., Adams, M. J. & Flanders, W. D. An epidemiologic approach to ecogenetics. *Am. J. Hum. Genet.* **42**, 89–95 (1988).
The patterns by which genes and environment jointly determine risk were laid out using commonly accepted clinical models such as xeroderma pigmentosum and phenylketonuria.
7. Ottman, R. An epidemiologic approach to gene–environment interaction. *Genet. Epidemiol.* **7**, 177–185 (1990).
8. Begg, C. B. & Berlin, J. A. Publication bias and dissemination of clinical research. *J. Natl Cancer Inst.* **81**, 107–115 (1989).
9. Rothman, K. J. & Greenland, S. *Modern Epidemiology* 2nd edn (Lippincott–Raven, Philadelphia, 1998).
10. Rebbeck, T. R., Spitz, M. & Wu, X. Assessing the function of genetic variants in candidate gene association studies. *Nature Rev. Genet.* **5**, 589–597 (2004).
11. Armitage, P. & Colton, T. (eds) *Biostatistical Genetics and Genetic Epidemiology* (John Wiley & Sons, West Sussex, 2002).
12. Narod, S. A. *et al.* Risk modifiers in carriers of *BRCA1* mutations. *Int. J. Cancer* **64**, 394–398 (1995).
13. Gauderman, W. J. Sample size requirements for matched case–control studies of gene–environment interaction. *Stat. Med.* **21**, 35–50 (2002).
14. Khoury, M. J. & Flanders, W. D. Nontraditional epidemiologic approaches in the analysis of gene–environment interaction: case–control studies with no controls! *Am. J. Epidemiol.* **144**, 207–213 (1996).
15. Adami, H. & Trichopoulos, D. in *Textbook of Cancer Epidemiology* (eds Adami, H., Hunter, D. & Trichopoulos, D.) 87–109 (Oxford Univ. Press, New York, 2002).
16. Wacholder, S., Rothman, N. & Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl Cancer Inst.* **92**, 1151–1158 (2000).
Using simulation studies, Wacholder and colleagues showed that the fear of population stratification in well-designed associated studies was exaggerated.
17. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature Genet.* **36**, 388–393 (2004).
18. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
19. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
20. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
21. Morimoto, L. M., White, E. & Newcomb, P. A. Selection bias in the assessment of gene–environment interaction in case–control studies. *Am. J. Epidemiol.* **158**, 259–263 (2003).
22. Garcia-Closas, M., Thompson, W. D. & Robins, J. M. Differential misclassification and the assessment of gene–environment interactions in case–control studies. *Am. J. Epidemiol.* **147**, 426–433 (1998).
23. Liotta, L. & Petricoin, E. Molecular profiling of human cancer. *Nature Rev. Genet.* **1**, 48–56 (2000).
24. Umbach, D. M. & Weinberg, C. R. Designing and analysing case–control studies to exploit independence of genotype and exposure. *Stat. Med.* **16**, 1731–1743 (1997).
25. Schmidt, S. & Schaid, D. J. Potential misinterpretation of the case-only study to assess gene–environment interaction. *Am. J. Epidemiol.* **150**, 878–885 (1999).
26. Albert, P. S., Ratnasinghe, D., Tangrea, J. & Wacholder, S. Limitations of the case-only design for identifying gene–environment interactions. *Am. J. Epidemiol.* **154**, 687–693 (2001).
27. Smith, P. G. & Day, N. E. The design of case–control studies: the influence of confounding and interaction effects. *Int. J. Epidemiol.* **13**, 356–365 (1984).
28. Garcia-Closas, M., Rothman, N. & Lubin, J. Misclassification in case–control studies of gene–environment interactions: assessment of bias and sample size. *Cancer Epidemiol. Biomarkers Prev.* **8**, 1043–1050 (1999).
Using standard techniques, these authors pointed out the inadequate size of contemporary studies of gene–environment interactions, particularly once measurement error of the environmental variables was considered.
29. Wong, M. Y., Day, N. E., Luan, J. A. & Wareham, N. J. Estimation of magnitude in gene–environment interactions in the presence of measurement error. *Stat. Med.* **23**, 987–998 (2004).
30. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356–1360 (2001).
A careful reconsideration of the advantages of the retrospective case–control design compared with prospective studies.
31. Sinha, R. *et al.* Heterocyclic amine content in beef cooked by different methods to varying degrees of doneness and gravy made from meat drippings. *Food Chem. Toxicol.* **36**, 279–287 (1998).
32. Roberts-Thomson, I. C. *et al.* Diet, acetylator phenotype, and risk of colorectal neoplasia. *Lancet* **347**, 1372–1374 (1996).
33. Chen, J. *et al.* A prospective study of *N*-acetyltransferase genotype, red meat intake, and risk of colorectal cancer. *Cancer Res.* **58**, 3307–3311 (1998).
34. Kampman, E. *et al.* Meat consumption, genetic susceptibility, and colon cancer risk: a United States multicenter case–control study. *Cancer Epidemiol. Biomarkers Prev.* **8**, 15–24 (1999).
35. Le Marchand, L. *et al.* Well-done red meat, metabolic phenotypes and colorectal cancer in Hawaii. *Mutat. Res.* **506–507**, 205–214 (2002).
36. Barrett, J. H. *et al.* Investigation of interaction between *N*-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectal cancer. *Carcinogenesis* **24**, 275–282 (2003).
37. Rothman, N. *et al.* The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim. Biophys. Acta* **1471**, C1–C10 (2001).
38. Ames, B. N. Cancer prevention and diet: help from single nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 12216–12218 (1999).
39. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
40. Goldstein, D. B., Tate, S. K. & Sisodiya, S. M. Pharmacogenetics goes genomic. *Nature Rev. Genet.* **4**, 937–947 (2003).
41. Smith, M. W. *et al.* Contrasting genetic influence of *CCR2* and *CCR5* variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study. *Science* **277**, 959–965 (1997).
42. Shaheen, F. & Collman, R. G. Co-receptor antagonists as HIV-1 entry inhibitors. *Curr. Opin. Infect. Dis.* **17**, 7–16 (2004).
43. Colhoun, H. M., McKeigue, P. M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865–872 (2003).
44. Ioannidis, J. P., Rosenberg, P. S., Goedert, J. J. & O'Brien, T. R. Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am. J. Epidemiol.* **156**, 204–210 (2002).
45. Deitz, A. C. *et al.* Impact of misclassification in genotype-exposure interaction studies: example of *N*-acetyltransferase 2 (NAT2), smoking, and bladder cancer. *Cancer Epidemiol. Biomarkers Prev.* **13**, 1543–1546 (2004).
46. Le Marchand, L. *et al.* Feasibility of collecting buccal cell DNA by mail in a cohort study. *Cancer Epidemiol. Biomarkers Prev.* **10**, 701–703 (2001).
47. Garcia-Closas, M. *et al.* Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiol. Biomarkers Prev.* **10**, 687–696 (2001).
48. Collins, F. S. The case for a US prospective cohort study of genes and environment. *Nature* **429**, 475–477 (2004).
49. Tranah, G. J., Lescault, P. J., Hunter, D. J. & De Vivo, I. Multiple displacement amplification prior to single nucleotide polymorphism genotyping in epidemiologic studies. *Biotechnol. Lett.* **25**, 1031–1036 (2003).
50. Paez, J. G. *et al.* Genome coverage and sequence fidelity of ϕ29 polymerase–based multiple strand displacement whole genome amplification. *Nucleic Acids Res.* **32**, e71 (2004).
51. Barker, D. L. *et al.* Two methods of whole-genome amplification enable accurate genotyping across a 2320-SNP linkage panel. *Genome Res.* **14**, 901–907 (2004).
52. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* **19**, 551–560 (2003).

53. Conti, D. V., Cortessis, V., Molitor, J. & Thomas, D. C. Bayesian modeling of complex metabolic pathways. *Hum. Hered.* **56**, 83–93 (2003).
An example of an emerging interest in new statistical techniques for addressing perturbations in complex biological pathways.
54. Cai, Z. *et al.* Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum. Mutat.* **24**, 178–184 (2004).
55. Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst.* **96**, 434–442 (2004).
56. Rose, G. Sick individuals and sick populations. *Int. J. Epidemiol.* **14**, 32–38 (1985).
A leader in epidemiology contended that public health could be bettered by addressing health problems at the population, rather than the individual, level.
57. Bigler, J. *et al. CYP2C9* and *UGT1A6* genotypes modulate the protective effect of aspirin on colon adenoma risk. *Cancer Res.* **61**, 3566–3569 (2001).
58. Chan, A. T. *et al.* Genetic variants in the UGT1A6 enzyme, aspirin use, and the risk of colorectal adenoma. *J. Natl Cancer Inst.* (in the press).
59. National Institute on Aging/Alzheimer's Association Working Group. Apolipoprotein E genotyping in Alzheimer's disease. *Lancet* **347**, 1091–1095 (1996).
60. Peila, R. *et al.* Joint effect of the *APOE* gene and midlife systolic blood pressure on late-life cognitive impairment: the Honolulu-Asia aging study. *Stroke* **32**, 2882–2889 (2001).
61. Kang, J. H., Logroscino, G., De Vivo, I., Hunter, D. & Grodstein, F. Apolipoprotein E, cardiovascular disease and cognitive function in aging women. *Neurobiol. Aging* **26**, 475–484 (2005).
62. Rees, J. L. The genetics of sun sensitivity in humans. *Am. J. Hum. Genet*. **75**, 739–751 (2004).
63. Chen, J., Giovannucci, E. & Hunter, D. J. *MTHFR* polymorphism, methyl-replete diets and the risk of colorectal carcinoma and adenoma among US men and women: an example of gene–environment interactions in colorectal tumorigenesis. *J. Nutr*. **129**, S560–S564 (1999).
64. Bloemenkamp, K. W., Rosendaal, F. R., Helmerhorst, F. M., Buller, H. R. & Vandenbroucke, J. P. Enhancement by factor V Leiden mutation of risk of deep-vein thrombosis associated with oral contraceptives containing a third-generation progestagen. *Lancet* **346**,1593–1596 (1995).
65. Lehtimaki, T. *et al.* Association between serum lipids and apolipoprotein E phenotype is influenced by diet in a population-based sample of free-living children and young adults: the Cardiovascular Risk in Young Finns Study. *J. Lipid Res*. **36**, 653–661 (1995).
66. Hines, L. M. *et al.* Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N. Engl. J. Med.* **344**, 549–555 (2001).
67. Memisoglu, A. *et al.* Interaction between a peroxisome proliferator-activated receptor-γ gene polymorphism and dietary fat intake in relation to body mass. *Hum. Mol. Genet.* **12**, 2923–2929 (2003).
68. Maier, L. A. Genetic and exposure risks for chronic beryllium disease. *Clin. Chest Med.* **23**, 827–839 (2002).
69. Weinshilboum, R. Thiopurine pharmacogenetics: clinical and molecular studies of thiopurine methyltransferase. *Drug Metab. Dispos.* **29**, 601–605 (2001).
70. Israel, E. *et al.* The effect of polymorphisms of the β2-adrenergic receptor on the response to regular use of albuterol in asthma. *Am. J. Respir. Crit. Care Med.* **162**, 75–80 (2000).

Acknowledgements
I thank my colleagues at the Harvard School of Public Health and the Channing Laboratory for helpful discussions, and P. Kraft for reviewing the manuscript.

Competing interests statement
The author declares no competing financial interests.

Online links

DATABASES
The following terms in this article are linked online to:
Entrez: http://www.ncbi.nih.gov/Entrez/
APOE4 | *BRCA1* | *CCR5* | *NAT2*
OMIM: http://www.ncbi.nlm.nih.gov/Omim/
Alzheimer disease | multiple sclerosis | Parkinson disease | phenylketonuria | sickle-cell anaemia | xeroderma pigmentosum

FURTHER INFORMATION
GeneSNPs: http://www.genome.utah.edu/genesnps
Human Genome Project:
http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
National Cancer Institute Breast and Prostate Cancer and Hormone-Related Gene Variants Cohort Consortium:
http://epi.grants.cancer.gov/BPC3
National Cancer Institute Cancer Genome Anatomy Project SNP500Cancer Database: http://snp500cancer.nci.nih.gov
UK Biobank: http://www.ukbiobank.ac.uk
US National Institute of Environmental Health Sciences Environmental Genome Project: http://egp.gs.washington.edu
Access to this interactive links box is free online.