24. Horikawa, I. *et al.* Differential *cis*-regulation of human versus mouse *TERT* gene expression *in vivo*: identification of a human-specific repression element. *Proc. Natl Acad, Sci. USA* **102**, 18437–18442 (2005).
25. Ayabe, F. *et al.* A novel expression system for genomic DNA loci using a human artificial chromosome vector with transformation-associated recombination cloning. *J. Hum. Genet.* **50**, 592–599 (2005).
26. Noskov, V. N. *et al.* A novel strategy for analysis of gene homologs and segmental genome duplications. *J. Mol. Evol.* **56**, 702–710 (2003).
27. Pavlicek, A. *et al.* Evolution of the tumor suppressor *BRCA1* locus in primates: implications for cancer predisposition. *Hum. Mol. Genet.* **13**, 2737–2751 (2004).
28. Kouprina, N. *et al.* Accelerated evolution of the *ASPM* gene controlling brain size begins prior to human brain expansion. *PLoS Biol.* **2**, 653–663 (2004).
29. Kouprina, N. *et al.* The SPANX gene family of cancer-testis specific antigens: rapid evolution, an unusual case of positive selection and amplification in African great apes and hominids. *Proc. Natl Acad. Sci. USA* **101**, 3077–3082 (2004).
30. Kouprina, N. *et al.* The microcephaly *ASPM* gene is expressed in proliferating tissues and encodes for a mitotic spindle protein. *Hum. Mol. Genet.* **14**, 2155–2165 (2005).
31. Salem, R. M., Wessel, J. & Schork, N. J. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics* **2**, 39–66 (2005).
32. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
33. Kim, J. H., Leem, S. H., Sunwoo, Y. & Kouprina, N. Separation of long-range human *TERT* gene haplotypes by transformation-associated recombination cloning in yeast. *Oncogene* **22**, 2443–2456 (2003).
34. Westbrook, V. A. *et al.* Genomic organization, incidence, and localization of the SPANX family of cancer-testis antigens in melanoma tumors and cell lines. *Clin. Cancer Res.* **10**, 101–112 (2004).
35. Baffoe-Bonnie, A. B. *et al.* A major locus for hereditary prostate cancer in Finland: localization by linkage disequilibrium of a haplotype in the *HPCX* region. *Hum. Genet.* **117**, 307–316 (2005).
36. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).
37. Shaw, C. J. & Lupski, J. R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum. Mol. Genet.* **13**, R57–R64 (2004).
38. Lockwood, W. W., Chari, R., Chi, B. & Lam, W. L. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. *Eur. J. Hum. Genet.* **1**, 139–148 (2005).
39. Perry, G. H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proc. Natl Acad. Sci. USA* **103**, 8006–8011 (2006).
40. Feuk, L., Marshall, C. R., Wintle, R. F. & Scherer, S. W. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.* **15**, R57–R66 (2006).
41. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nature Rev. Genet.* **5**, 345–354 (2004).
42. Grimwood, J. *et. al.* The DNA sequence and biology of human chromosome 19. *Nature*, **428**, 529–535 (2004).
43. Grimes, B. R. & Monaco, Z. L. Artificial and engineered chromosomes: developments and prospects for gene therapy. *Chromosoma* **114**, 230–241 (2005).
44. Basu, J. & Willard, H. F. Artificial and engineered chromosomes: non-integrating vectors for gene therapy. *Trends Mol. Med.* **11**, 251–258 (2005).
45. Ikeno, M. *et al.* Generation of human artificial chromosomes expressing naturally controlled guanosine triphosphate cyclohydrolase I gene. *Genes Cells* **10**, 1021–1032 (2002).
46. Mejia, J. E. *et al.* Efficiency of *de novo* centromere formation in human artificial chromosomes. *Genomics* **79**, 297–304 (2002).
47. Kotzamanis, G. *et al.* Construction of human artificial chromosome vectors by recombineering. *Gene* **351**, 29–38 (2005).
48. Kakeda, M. *et al.* Human artificial chromosome (HAC) vector provides long-term therapeutic transgene expression in normal human primary fibroblasts. *Gene Ther.* **12**, 852–856 (2005).
49. Grimes, B. R., Rhoades, A. A. & Willard, H. F. α-Satellite DNA and vector composition influence rates of human artificial chromosome formation. *Mol. Ther.* **5**, 798–805 (2002).
50. Basu, J., Compitello, G., Stromberg, G., Willard, H. F. & Van Bokkelen, G. Efficient assembly of *de novo* human artificial chromosomes from large genomic loci. *BMC Biotechnol.* **5**, 21 (2005).
51. Kouprina, N. *et al.* Cloning of human centromeres by transformation-associated recombination in yeast and generation of functional human artificial chromosomes. *Nucleic Acids Res.* **31**, 922–934 (2003).
52. Ebersole, T. *et al.* Rapid generation of long synthetic tandem repeats and its application for analysis in human artificial chromosome formation. *Nucleic Acids Res.* **33**, e130 (2005).
53. Becker, M. *et al.* Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast. *Genome Res.* **14**, 2319–2329 (2004).

Competing interests statement
The authors declare no competing financial interests.

DATABASES
The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ASPM | *ATM* | *BRCA1* | *HPRT* | *KAI1* | *MUC2* | *NBS1* | *SCK1* | *SPANXA1* | *SPANXA2* | *SPANXB* | *SPANXC* | *SPANXD* | *TERT* |
UniProtKB: http://ca.expasy.org/sprot
CENPB

FURTHER INFORMATION
American Type Culture Collection: http://www.lgcpromochem-atcc.com
Access to this links box is available online.

OPINION

Genes, environment and the value of prospective cohort studies

Teri A. Manolio, Joan E. Bailey-Wilson and Francis S. Collins

Abstract | Case–control studies have many advantages for identifying disease-related genes, but are limited in their ability to detect gene–environment interactions. The prospective cohort design provides a valuable complement to case–control studies. Although it has disadvantages in duration and cost, it has important strengths in characterizing exposures and risk factors before disease onset, which reduces important biases that are common in case–control studies. This and other strengths of prospective cohort studies make them invaluable for understanding gene–environment interactions in complex human disease.

The sequencing of the human genome and increased investigation of its function are providing powerful research tools for identifying genetic variants that contribute to common diseases[1–3]. Recognition is growing, however, that genetic variants alone cannot account for most cases of chronic disease[4]. It is far more likely that environmental and behavioural changes, in interaction with a genetic predisposition, have produced most of the recent increases in chronic disease, and might therefore be the key to reversing this trend[5].

For these reasons, the search for gene–environment interactions — differences in the association of a genetic variant with disease in the presence of a particular environmental exposure, or vice versa — is gaining increased emphasis[6]. These interactions are important because they can mask the detection of a genetic (or environmental) effect if they are not identified and controlled for, and can also lead to inconsistencies in disease associations when populations are subject to different environmental exposures that modify the effect of a given genetic variant (or the reverse)[7–11] (FIG. 1). However, the most important implication of gene–environment interactions is that they can suggest approaches for modifying the effects of deleterious genes by avoiding the deleterious environmental exposure, as both the genetic variant and the exposure must be present to produce disease.

The most widely used method for investigating the genetic and environmental basis of complex disease is the case–control study. Case–control studies involve an investigation of all cases of disease, or a representative sample of cases, compared with a representative sample of disease-free controls. Cases and controls are typically investigated retrospectively for evidence of genetic and other risk factors along with environmental exposures that existed before disease onset, and so probably contributed

to disease development. However, because case–control studies typically begin with disease cases that have already occurred, they are subject to significant sources of bias, as described below.

By contrast, prospective cohort studies involve the investigation of a representative sample of the population before disease onset. This sample is then followed until the occurrence of specified endpoints (see FIGURE 2 for a comparison of this design with a case–control study[12]). The purpose of this design is to identify risk factors that predispose an individual to disease, or biomarkers for predicting disease development, in the population as a whole, not only among those individuals that come to medical attention. Prospective cohort studies are particularly valuable for detecting risk factors and risk markers that might be affected by disease, treatment or lifestyle changes[13], which are subject to imperfect or biased recall, and for identifying risk factors that might have early pathogenic effects[14]. Several large-scale prospective cohort studies of genes and environment are underway or in planning throughout the world, including the UK Biobank[15] and a proposed large-scale US cohort study[5]. However, the need for this design in genetic research has been questioned[16,17]. The high costs, large sample sizes and long durations that are typical of prospective cohort studies have been contrasted to the potentially more efficient case–control design[18].

Here we present the advantages of the prospective cohort design, which avoids or significantly reduces the important weaknesses of the case–control design, particularly with respect to identifying gene–environment interactions. We begin by discussing how bias can be introduced into studies of risk factors for disease, followed by an analysis of the extent to which each design is affected by such biases and other weaknesses, and the advantages that prospective cohort studies provide. We then outline the instances in which we believe that prospective cohort studies have important advantages, with a feasability analysis that includes the sample sizes needed to identify genetic and environmental risk factors and their interactions, and the challenges faced. On this basis, we argue that prospective cohort studies provide a valuable, feasible and, indeed, indispensable means of exploring the genetic basis of complex human diseases. We also put forward the case for carrying out new, large-scale studies of this type to determine the roles of genes and environment in diseases of major public health importance.

Potential sources of bias

The validity of the evidence from observational studies of the genetic and environmental influences on disease relies on the avoidance of bias, which is defined as: "Any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth".[19] Reduction of bias is the principal reason for preferring the prospective cohort design to the case–control design.

At least 35 types of bias have been described[19], but 8 are crucial in assessing the strengths and weaknesses of case–control and prospective cohort studies (BOX 1). Particularly important are biases in subject selection[20], especially prevalence–incidence bias, which occurs when a study of currently evident (prevalent) cases (which are often identified through medical records) overlooks fatal cases or other short episodes[21]. This is a particular problem if a sizeable subset of cases suffers a rapid and fatal course (as in coronary disease or some cancers), so that the 'aetiological' factors that are subsequently identified among the subset of survivors are actually more related to survival or a benign prognosis than to disease causation[22]. Another potentially important form of respondent bias in genetic studies is the tendency for people with a positive family history to be more likely to participate[23,24]. A critically important bias in the estimation of self-reported environmental exposures is recall bias. This type of bias occurs when disease status influences the reporting of exposures, for example, when questions about exposure to a putative cause might be asked many times of known cases (or they might repeatedly search their memories) but only once of those without disease.

Any of these forms of bias can severely affect the validity and generalizability of any observational study of disease aetiology. Although concerns about recall bias tend to be dismissed in genetic studies because determination of the key exposure (a genetic variant) does not rely on recall and the temporal nature of the genetic association is clear, the potential for bias in the selection of cases and controls and in the assessment of other exposures remains[25].

Case–control studies

The advantages of the case–control design are compared with those of the prospective cohort approach in TABLE 1. Although the case–control design is often preferred during initial efforts to identify putative risk factors for common diseases because

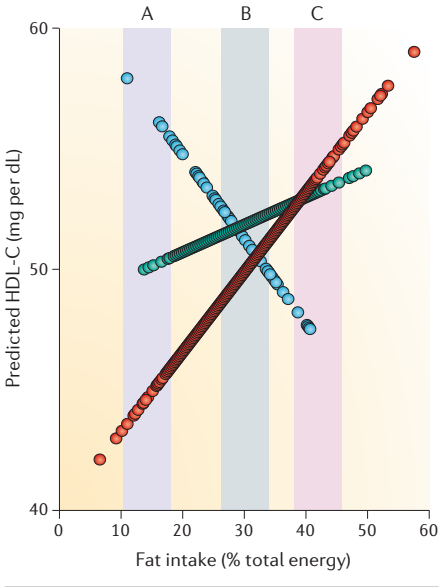

Figure 1 | **The importance of gene–environment interactions — an example.** Predicted values of high-density lipoprotein cholesterol (HDL-C) are shown for different hepatic lipase (*LIPC*) genotypes at different total levels of dietary fat intake (data from REF. 7). Low fat intake (band A) combined with the TT genotype results in the highest HDL-C level. For a moderate fat intake (band B), there is no relationship between genotype and HDL-C level. For a high fat intake (band C), the TT genotype has the lowest HDL-C level. Gene–environment interactions are therefore important in identifying genetic and environmental determinants of medically relevant phenotypes such as HDL-C levels; depending on the dietary fat intake, one could conclude that the TT genotype produces high (band A) or low (band C) HDL-C levels, or that it is not associated with HDL-C levels at all (band B).

of ease and cost, it actually has particularly important advantages in the study of rare diseases. This is because it starts with diagnosed cases of disease, often from specialized referral centres, making identification and recruitment relatively easy. By contrast, the prospective cohort design requires the follow-up of large numbers of people who will never develop a rare disease, in order to identify the few cases who do[14]. The case–control design also allows the assessment of multiple exposures in relation to disease outcome, provided that those exposures can be measured retrospectively, or after disease has occurred. It can also allow a more detailed assessment of a particular exposure (such as in occupational or recreational settings) if that exposure is known to be especially relevant to the disease under study.

Despite these advantages, case–control studies are prone to several of the sources of bias outlined in BOX 1. A key requirement for a bias-free case–control study is that cases be representative of all those who develop the disease that is being studied. However, because cases are often identified in the clinical setting, mild cases or those that cause early mortality are likely to be missed, leading to prevalence–incidence bias. Another requirement is that the controls be representative of all those at risk of developing the disease[26]. In this respect, the potential threats to the representativeness of cases are also relevant to controls, particularly non-response bias. Differential response rates that are related to an individual's genetic background are possible in cases and controls owing to sample stratification by ancestry or a positive family history of disease[24]. Findings from a biased group of cases or controls might not be generalizable to the population at large, and might actually be invalid. Selection of controls is one of the most difficult and most heavily criticized aspects of case–control studies; indeed, it has been suggested that the ideal control group probably does not exist[27].

A third requirement for a bias-free case–control study is that the collection of risk-factor and exposure information should be the same for cases and controls[20]. This can be difficult to ensure, particularly for information that has been collected in the course of clinical care, as invasive diagnostic approaches cannot be justified in healthy controls. Data collection methods must therefore be developed that can be applied equally to both groups. However, even this cannot control for the potential recall bias among the cases. Limiting the collection of risk-factor or biomarker information to the period before disease onset, if the time of onset can be clearly defined, will reduce biases in risk-factor ascertainment that are related to clinical care or awareness of disease status. Such use of pre-morbid risk-factor information will also strengthen inferences about the temporal nature of risk relationships, a key element in determining causality[28]. Unless extensive records exist before disease diagnosis, however, many key exposures, such as dietary patterns or medication use, cannot be collected retrospectively, and so pre-morbid risk factor information is often unavailable.

Another requirement for a valid case–control study is that the ancestral geographical origins and predominant environmental exposures of cases must not differ dramatically from those of controls. Fortunately, the collection of ancestry informative markers and information on potential environmental confounders allows adjustment for differences in genetic background and environmental exposures, as long as there is some commonality between cases and controls[29,30]. These must be applied carefully, however, to avoid over-adjusting for variants or exposures that might actually be causal[31].

Finally, case–control studies allow the investigation of only one primary outcome: the condition by which cases are defined. Because complex diseases rarely occur in isolation and often share risk factors, the ability to examine genetic and environmental risk factors for a number of conditions after costly genomic assays have been done is one of the main advantages of cohort studies.

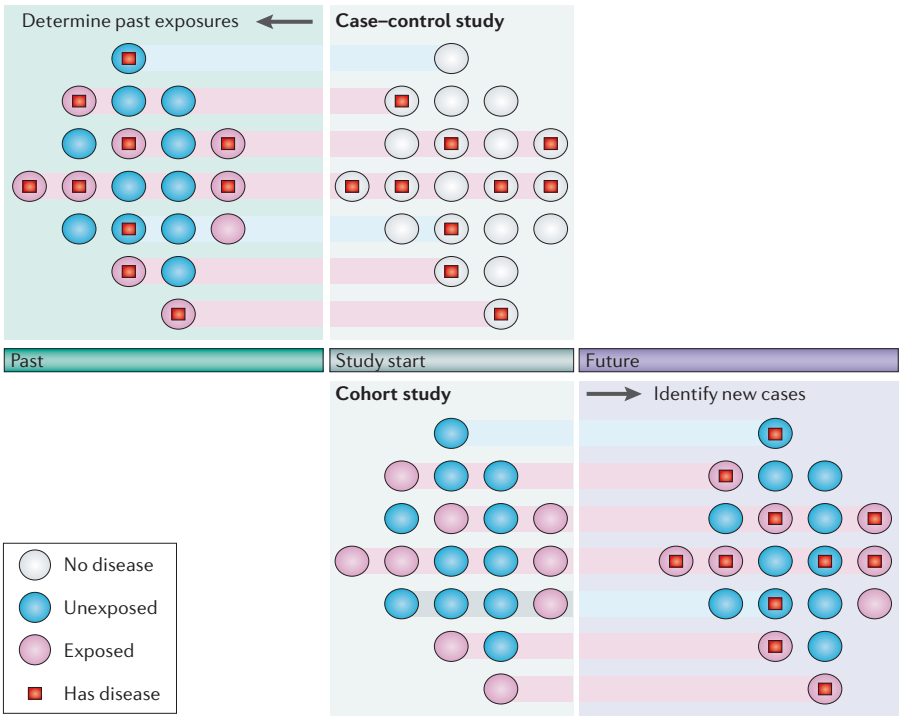

Figure 2 | **The case–control and prospective cohort study designs.** Case–control studies identify individuals with and without disease, determine the differences between them in past exposures or biological characteristics, and then examine those differences for potentially causative factors. Prospective cohort studies identify individuals with and without a given exposure, follow them through time to determine who develops disease, and then examine differences in the preceding exposures for potentially causative factors. Modified with permission from REF. 12 © (2003) Massachusetts Medical Society.

Prospective cohort studies

An important advantage of the prospective cohort design is that it allows standardized and detailed collection of pre-morbid exposure information, tailored to meet the goals of the study. The assessment of environmental risk factors, and therefore gene–environment interactions, is typically more extensive and less prone to bias in prospective cohort studies than in case–control studies, making the prospective cohort design much more suitable for studying environmental influences on disease risk. Recall bias in particular is avoided by collecting information before disease onset.

Another key aspect of the prospective cohort design is that all participants are followed in a systematic way, so that all cases of disease have an equal likelihood of being detected. This feature is important as it minimizes biases in case identification — particularly prevalence–incidence bias — that are typically encountered in clinical series. The time of disease onset can also be defined more clearly in prospective cohort studies than in case–control studies, and multiple disease outcomes can be studied.

The requirements for a generalizable prospective cohort study are that people recruited into the cohort have similar genetic and environmental exposures, and disease risk, to those who are not recruited, and that cohort members who are 'lost' to follow-up have similar exposures and disease risk to

Box 1 | **Major sources of bias that affect case–control and prospective cohort studies**

Biases that relate to subject selection

Prevalence–incidence or survival bias. Selection of existing cases that are currently available for study will miss fatal and short episodes, and might miss mild or silent cases[19].

Non-response (or respondent) bias. Differential rates of refusal or non-response to inquiries between cases and disease-free comparison subjects[19].

Diagnosis bias. Also known as diagnostic suspicion bias. Knowledge of a subject's exposure to a putative cause of disease can influence both the intensity and outcome of the diagnostic process[19].

Referral or admission-rate bias. Factors related to the probability of referral. Cases who are more likely to receive advanced care or to be hospitalized — such as those with greater access to health care or with co-existing illnesses — can distort associations with other risk factors in clinic-based studies, unless the same referral or admission biases are operative in disease-free comparison subjects[20].

Surveillance bias. If a condition is mild or likely to escape routine medical attention, cases are more likely to be detected in people who are under frequent medical surveillance[20].

Biases that relate to measuring exposures and outcomes

Recall bias. Questions about specific exposures might be asked more frequently of cases, or cases might search their memories more intensively for potential causative exposures.

Family information bias. The flow of family information about exposures or illnesses can be stimulated by, or directed to, a new case in its midst[19].

Exposure suspicion bias. Knowledge of a patient's disease status can influence the intensity and outcome of the search for exposure to a putative cause[19].

those remaining. A third requirement is that the likelihood of detection of disease is independent of the exposure of interest and potentially confounding factors such as age, other exposures and access to medical care. This ensures similarity of data collection (and avoidance of bias) between exposed and unexposed people.

Ascertainment methods and outcome definitions should be the same in all cohort members and should not differ in relation to the participants' genetic or environmental exposures. Changes in exposure history should be assessed by repeated collection of exposure information and analysed by appropriate longitudinal techniques[32].

Cohort studies that rely on outcomes that have been identified in the course of clinical care are prone to many of the biases discussed for case–control studies, so most prospective cohort studies implement a regular schedule of follow-up in which all participants are systematically investigated for the occurrence of disease and changes in exposure. The need for such ongoing follow-up has been one of the main criticisms of prospective cohort studies, as it is time-intensive and costly.

Other important limitations of the prospective cohort design include the large sample size needed to produce sufficient numbers of incident disease cases, which we discuss in more detail below, and the typically long duration needed for these cases to accrue. In addition, the need to identify and collect information on risk factors of interest before disease cases have accrued adds to the complexity and cost of prospective cohort studies, but is often the only way to obtain valid exposure information for the prediction of disease.

When should cohort studies be used?

Given the strengths and weaknesses of the two study designs, what are the areas of aetiological research for which the prospective cohort design is preferable? One such situation is the study of diseases for which case–control studies might miss the full range of disease manifestations, including those with high a mortality at onset, a short duration or a long preclinical phase. Such conditions include complex diseases that represent an important burden on health in the developed world, such as type 2 diabetes and pancreatic cancer (TABLE 2).

The prospective cohort design also allows the identification of predictive biomarkers that appear well before a disease is diagnosed clinically, and risk factors with a relationship to disease that is not constant over time, such as those that have a long latent period or a suggested early pathogenic effect. Prospective cohort studies are better suited to identifying risk factors that change after the onset of disease, such as those affected by disease, treatment or lifestyle changes, or those subject to imperfect or biased recall.

In addition, the prospective cohort design is preferable for studies of common diseases that seem to be genetically complex, that is, due to many genes of small effect rather than a single major gene. As discussed above,

Glossary

Exposure
A putative cause or characteristic determinant of a health outcome of interest.

Risk factor
An attribute or exposure that increases the probability of disease or other outcome; used by some to mean causal factor or 'determinant' and by others to mean 'risk marker'.

Cohort
Originally defined as a group of people born during a particular period (a 'birth cohort'); now broadened to include any designated group of people who are followed or traced over time.

Risk marker
An attribute or exposure that is associated with an increase in the probability of a specified outcome, but is not necessarily a causal factor.

Population stratification
The presence of different allele frequencies in cases and controls that is attributable to diversity in the background population and is unrelated to outcome status.

Ancestry informative (ancestral) marker
A locus with several polymorphisms that exhibit substantially different frequencies between ancestral populations. For example, the Duffy null allele has a frequency of almost 100% of sub-Saharan Africans, but occurs infrequently in other populations.

Incidence
The number of new cases of disease that develop during a period of time.

Odds ratio (or relative odds)
The odds of disease in the individuals exposed to an environmental factor or genetic variant divided by the odds in unexposed individuals; or the odds of exposure in the cases divided by the odds in the controls (they are algebraically equivalent). If the odds ratio is significantly greater than one, then the environmental factor or genetic variant is associated with the disease.

Study power
The probability of rejecting the null hypothesis of no association in a study if it is in fact false, or of detecting a difference between two groups if it does in fact exists.

Type I error rate
The probability of rejecting the null hypothesis of no association in a study if it is in fact true, or of detecting a difference between two groups when no difference exists.

Table 1 | **Comparison of case–control and prospective cohort studies**

Feature	Case–control Studies	Prospective cohort studies
Temporal relationship between exposure and disease	Can be hard to establish	Generally easy to establish
Types of association studied	Single disease in relation to multiple exposures	Multiple diseases in relation to multiple exposures
Duration of study	Relatively short	Typically long owing to the need for follow-up to disease occurrence
Cost of study	Low	High
Population size needed	Small	Large
Potential biases	Assessment of exposure (recall bias), prevalence–incidence bias	Assessment of outcome (exposure suspicion and diagnostic suspicion bias, referral bias)
Situation in which design is preferred	Disease is rare, exposure is frequent among diseased	Exposure is rare, disease is frequent among exposed
Characterization of cases	More complete clinical characterization at the time of diagnosis	More complete characterization of onset and progression following exposure
Characterization of exposures	Incomplete information on exposure, validation is difficult or impossible	Allows flexibility throughout the course in choosing the exposures to be measured, allows for ongoing quality control
Identification of predictive biomarkers	Rarely possible (requires specimens to be collected before disease onset)	Often possible through prospective collection of biospecimens
Comparison group	Selection of appropriate controls is often difficult	Selection of unexposed comparison group is often difficult

This table is adapted from REFS 14, 20.

this is because the breadth and reliability of the environmental exposure data that can be obtained prospectively allows the examination of key gene–environment interactions and, consequently, greater validity in estimates of genetic effects.

Prospective cohort studies are also particularly well suited to studying multiple disease outcomes, especially those that might share risk factors, such as cancer, heart disease and diabetes. This potential of prospective cohort studies is infrequently realized, with many studies still being designed to assess only one major disease or group of diseases[33,34]. However, several notable studies do include multiple endpoints[35–37]. Given that the lifetime risk of heart disease is estimated to be one in three men and one in four women[38], that of breast cancer is estimated to be one in eight women (as described in the SEER Cancer Statistics Review, 1975–2002), and that of prostate cancer is estimated to be one in six men[39], the assessment of multiple outcomes would dramatically increase the efficiency of these studies. Existing cohort studies might also be supplemented to expand their ascertainment methods to other disease endpoints[40,41], although this could require considerable additional funding, expertise and consent.

Last, prospective cohort studies are valuable for critically examining the potential risk factors that are initially identified through other approaches, including case–control studies. Many of the irremediable biases of case–control studies can be addressed only by confirming their findings in prospective cohort designs, so that a detailed and reliable estimation of environmental exposures can be included at the outset. Unfortunately, as important as such confirmatory studies are (for examples, see REFS 42–44), they also cause prospective cohort studies to be viewed as lacking original hypotheses and innovation[45–47]. Despite the negative way in which prospective cohort studies are sometimes viewed, however, their impact on public health is undeniable. This importance is highlighted by the fact that many clinical misperceptions, such as the ideas that isolated systolic hypertension is normal with ageing, that silent myocardial infarction does not carry an increased risk of mortality and that the risk of hypertension has a threshold rather than a continuous effect, have been dispelled by cohort studies[43,48].

The need for new studies

Although many prospective cohort studies are already in place[35,47,49], none is comprehensive enough to cover the main causes of morbidity and mortality that are relevant during an entire human lifetime, nor to provide sufficient diversity, in terms of racial, ethnic or socioeconomic groups, to be applicable to the general population in countries such as the United States. Although individual studies can address particular population segments, combining these existing studies into a single cohort carries the risk of significant between-study biases within the resulting large cohort.

This issue was highlighted in responses to a Request for Information issued by the US National Human Genome Research Institute (NHGRI) in 2004. In addition, the need for comparable and broad-based data collection in all cohort members would necessitate the collection of new exposure information, disease outcomes and informed consent, and would therefore be unlikely to produce appreciable cost savings.

These considerations led an NHGRI Expert Panel to conclude that although existing studies could provide valuable experience, previously obtained data and large numbers of potentially interested study participants, combining those data in a way that allows meaningful cross-study analyses would be almost impossible. It would also risk limiting the study to the lowest common denominator of exposure information collected. Far preferable, although more costly, would be to design a prospective cohort study with state-of-the-art measures of multiple exposures and diseases right from the start, which could recruit some of its participants from existing studies if desired.

In light of these considerations, the NHGRI Expert Panel has recommended establishing a new cohort that is broadly representative of the US population. The participants would be selected to represent the entire human lifespan at the time of their entry into the cohort, and would undergo periodic re-examinations and annual follow-up for major disease outcomes. Similar plans are proposed for the UK Biobank, although

Table 2 | **Situations for which prospective cohort studies are likely to be superior to case-control studies**

Situation	Example
Diseases with:	
High mortality at onset	Malignant ventricular arrhythmias, subarachnoid haemorrhage
Short duration	Pancreatic cancer, septicaemia
Long preclinical phase	Diabetes, chronic obstructive pulmonary disease
Risk factors with:	
Long latent period	Radiation exposure and cancer, smoking and chronic obstructive pulmonary disease
Predicted early pathogenic effect	Cholesterol and coronary disease, low education and cognitive decline
Risk factors affected by:	
Disease	Hypertension and myocardial infarction, social support and depression
Treatment	C-reactive protein and statins, obesity and diabetes
Lifestyle changes	Cholesterol levels and fat intake, blood pressure levels and salt intake
Other situations:	
Risk factors subject to imperfect recall	Maternal exposures during pregnancy, weight or physical activity levels in early life
Predictive biomarkers present long before the disease is clinically diagnosed	Various markers in cancer, C-reactive protein in coronary disease

that study has a more limited age range and periodic re-examinations of the entire cohort are not anticipated. Improved methods for exposure assessment have been highlighted as being crucial for such research to move forward[5], and are being actively pursued, for example by the US National Institute of Environmental Health Sciences[50] and the proposed Genes and Environment Initiative.

Feasibility of prospective cohort studies

Sample sizes and affordability. To examine the feasibility of carrying out successful large-scale prospective cohort studies, we estimated the sample sizes that would be needed to detect genetic and environmental effects, and gene–gene or gene–environment interactions. This was achieved by using incidence estimates from a common source (the Incidence and Prevalence Database Timely Data Resources, Capitola, California) for a range of diseases to determine the number of cases that would accrue over a 5-year period of follow-up in samples of varying sizes that would reflect the general US population. The samples that we used are representative of the full age (from birth), sex and ethnicity distributions of the 2000 US Census. The estimated numbers of cases that are expected to arise are shown in TABLE 3. These numbers were then used to determine the minimum odds ratios that could be detected for environmental, genetic, gene–environment and gene–gene effects. The QUANTO program[51] was used to calculate the minimum number of cases needed (assuming there are two matched controls for each case) for different frequencies of the risk allele, marginal genetic effect (odds ratio associated with the genetic variant alone), environmental exposure frequency and marginal environmental effect (odds ratio associated with the exposure alone) (FIG. 3).

According to our estimates, a prospective cohort study of 1,000,000 subjects (FIG. 3a) would have sufficient power to detect an environmental exposure odds ratio of ≥1.5 for diseases of ≥0.05% incidence per year, such as colorectal cancer, whereas a study of 200,000 people could only detect an environmental odds ratio of ≥2.3 for diseases with this incidence. The minimum detectable odds ratios for genetic factors were slightly lower (indicating the power of the study was higher), mainly because a single individual has two 'chances' of carrying a dominant risk allele (FIG. 3b). For interactions, however, the minimum detectable odds ratios were much higher (that is, the power was lower), as would be expected from the much smaller number of participants exposed to both

Table 3 | **Estimated disease incidence rates in prospective cohort studies**

Disease incidence per 100,000 per year (%)	Disease examples	Number of incident cases in 5 years for different cohort sizes		
		200,000	500,000	1,000,000
10 (0.01)	Parkinson disease, schizophrenia	91	228	457
50 (0.05)	Colorectal cancer, renal failure	456	1,141	2,282
100 (0.10)	Breast cancer, hip fracture	912	2,279	4,559
200 (0.20)	Diabetes, stroke, heart failure	1,820	4,550	9,100
500 (0.50)	Myocardial infarction, all cancers	4,524	11,309	22,618
3,000 (3.00)	Cataracts, hypertension	25,858	64,644	129,289

Estimated numbers of incident cases available after 5 years of follow-up across the entire age range in the US population are shown, assuming an attrition rate of 3% per year. Data are taken from the Incidence and Prevalence Database.

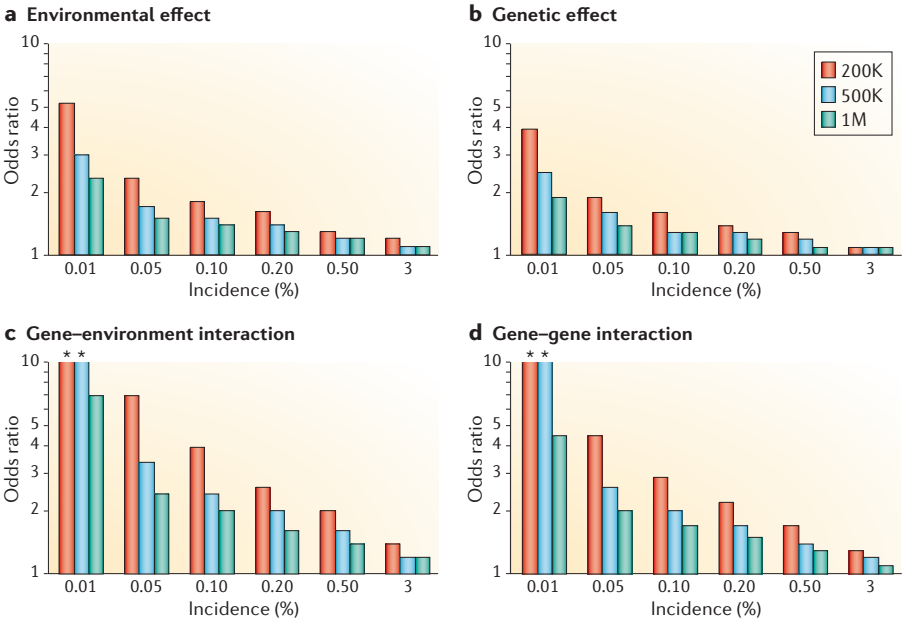

Figure 3 | **Sample-size requirements in prospective cohort studies.** The estimated minimum detectable odds ratios after 5 years of follow-up for various cohort sizes and disease incidences are shown, assuming: 10% allele frequency for a dominant risk allele, 10% environmental exposure frequency, no prevalent cases in the cohort at the start of the study, 3% annual loss to follow-up, 80% power, and a type I error rate of 0.0001. Minimum odds ratios are shown for: an environmental exposure effect (**a**); a genetic effect (for a dominant variant) (**b**); a gene–environment interaction, assuming genetic and environmental marginal effects of 1.5 (**c**); a gene–gene interaction, assuming genetic and environmental marginal effects of 1.5 (**d**). Asterisks indicate minimum detectable odds ratios in excess of 10.

genetic and environmental risk factors. Whereas a prospective cohort study of 1,000,000 had sufficient power to detect a gene–environment interaction odds ratio of ≥1.4 for diseases of ≥0.5% incidence a year, a study of 200,000 could only detect this gene–environment interaction odds ratio for diseases of ≥3% incidence (FIG. 3c). For a disease of 0.05% incidence, the minimum detectable odds ratio was about 2.4 in the 1,000,000-person study, and as much as 7.0 in the 200,000-person study. Minimum detectable gene–gene odds ratios were slightly lower than gene–environment odds ratios (FIG. 3d).

Genetic and environmental marginal odds ratios and interaction odds ratios of at least 1.5 are likely to be important to detect, as this is the magnitude of risk associated with genetic variants that is known to be important in complex diseases such as diabetes[52,53]. A cohort of 200,000 will provide adequate power within 5 years for only the most common diseases, such as cataracts and hypertension, and will miss these effects for important diseases such as myocardial infarction, diabetes and all cancers. By contrast, a cohort size of 500,000 — the number recommended by the NHGRI Expert Panel for a US cohort — will capture many more of these effects. For rarer diseases such as Parkinson disease or schizophrenia, gene–environment interactions would probably not be detectable within 5 years, even with 1,000,000 participants, but might be approached by continued follow-up and accrual of additional cases (or pooling with other cohort studies) over time. Conversely, gene–environment interactions for more common diseases, such as hypertension, could be examined early in follow-up and could be assessed for consistency in key subgroups. Of course, consideration of higher-order interactions (gene-by-gene-by-gene, or multiple interacting genetic and environmental factors) will require larger sample sizes and might not be approachable within a single study, even for the most common outcomes.

The recruitment of such large numbers of subjects will of course require substantial investment. The costs of the ongoing Women's Health Initiative Observational Study of 116,000 women, for example, have been estimated at US$128 per participant per year, with approximately $400 per participant for initial recruitment, or roughly $120 million for a 5-year study (J. Rossouw, personal communication).

Other factors that affect feasibility. Other challenges in conducting prospective cohort studies are well known, and include the difficulties in enrolling a generalizable population and maintaining high follow-up rates, assessing incident morbid events and classifying causes of death, and collecting detailed exposure information for the large number of exposures that are potentially relevant to multiple diseases. Monitoring incident diseases can also be difficult in settings that have no universal access to health care or electronic medical records. For example, this is the case in much of the United States, although electronic records do currently exist in large-scale health-maintenance organizations and military and veterans' health-care systems. Indeed, an electronic medical record for all US citizens is a high priority in the proposed National Health Infrastructure Initiative.

Although the size and complexity of a study addressing multiple diseases might seem daunting, complex diseases have many key risk factors in common. Data collection can therefore be prioritized to focus on the exposures with the greatest potential relevance to multiple diseases of public health importance, as described by the NHGRI Expert Panel and the Request for Information cited above. Challenges related to participant confidentiality and informed consent in large-scale genetic studies, and other difficult issues such as the return of genetic results, the costs of additional testing and clinical care, and the risks to insurance or employment status from research participation, are encountered in case–control as well as cohort studies and are being actively addressed in programmes such as the NHGRI Ethical, Legal and Social Issues programme[54] and the Ethics and Governance Framework of the UK Biobank. A dynamic consent process and the ongoing follow-up that is a feature of prospective cohort studies might make these studies uniquely suited to addressing the ethical issues and participant concerns that are emerging in relation to evolving scientific opportunities. This could help to ensure continued high rates of participation through frequent participant contact and updated consent.

Although the case–control design avoids some of these logistical challenges, the generalizability of the resulting information is limited considerably, as described above. More importantly, the difficulties in conducting good cohort

studies are far from insurmountable, as demonstrated by the many successful studies of this type. As discussed, added efficiency can be gained by expanding the number of disease outcomes ascertained, and by collecting expensive exposure measures on an informative subset using the 'case–control within a cohort' or 'nested case–control' design[55,56]. This design avoids many of the potential pitfalls of classic case–control studies by selecting incident cases and a sample of disease-free controls from within a prospective cohort study that was established earlier. The validity of the nested case–control design critically depends, however, on the ability to measure existing exposures before disease onset once cases have developed, as with biological samples collected and stored at study entry. Such an approach could also be used for limiting intensive assessment of outcomes to participants with a particular exposure, such as an environmental toxin, in a modification of the nested design.

Conclusion

As noted by Langholz *et al.* "…once the cohort study resource is established and a sufficient number of cases has occurred, a study of genetic factors can proceed much more quickly and efficiently than a population-based study."[13] Of course, the existence of such studies depends on researchers having the prescience, persistence and resources to establish the population-based cohort in the first place.

Despite the near universal preference for quick returns, complex diseases develop over decades and the reliable identification of their aetiological factors requires detailed examination and long-term follow-up of disease-free individuals in prospective cohort studies. Such studies are a necessary complement to case–control studies and other epidemiological designs. We might not need many in place, if they are comprehensive enough and provide wide access to data and samples[57] (with appropriate protections for participant confidentiality) and if they include the potential for adding new exposure or outcome assessments as science progresses. All of these characteristics have been recommended for the design of a possible large-scale US prospective cohort study[5], and are included to varying degrees in other similar efforts such as UK Biobank, Biobank Japan[58], and the Swedish National Biobank Program. The time to proceed with such studies is upon us.

Teri A. Manolio and Francis S. Collins are at the National Human Genome Research Institute, 31 Center Drive, Room 4B-09, Bethesda, Maryland 20892-2154, USA.

Joan E. Bailey-Wilson is at the Inherited Disease Research Branch, National Human Genome Research Institute, Baltimore, Maryland 21224, USA.

Correspondence to T.A.M.
e-mail: manolio@nih.gov

doi:10.1038/nrg1919

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921(2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**,1304–1351 (2001).
3. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
4. Chakravarti, A. & Little, P. Nature, nurture, and human disease. *Nature* **421**, 412–414 (2003).
5. Collins, F. S. The case for a US prospective cohort study of genes and environment. *Nature* **429**, 475–477 (2004).
6. Hunter, D. J. Gene–environment interactions in human diseases. *Nature Rev. Genet.* **6**, 287–298 (2005).
7. Ordovas, J. M. *et al.* Dietary fat intake determines the effect of a common polymorphism in the hepatic lipase gene promoter on high-density lipoprotein metabolism: evidence of a strong dose effect in this gene–nutrient interaction in the Framingham Study. *Circulation* **106**, 2315–2321 (2002).
8. Tai, E. S. *et al.* Singapore National Health Survey. Dietary fat interacts with the -514C > T polymorphism in the hepatic lipase gene promoter on plasma lipid profiles in a multiethnic Asian population: the 1998 Singapore National Health Survey. *J. Nutr.* **133**, 3399–3408 (2003).
9. Bos, G. *et al.* Interactions of dietary fat intake and the hepatic lipase -480C > T polymorphism in determining hepatic lipase activity: the Hoorn Study. *Am. J. Clin. Nutr.* **81**, 911–915 (2005).
10. Ko, Y. L., Hsu, L. A., Hsu, K. H., Ko, Y. H. & Lee, Y. S. The interactive effects of hepatic lipase gene promoter polymorphisms with sex and obesity on high-density-lipoprotein cholesterol levels in Taiwanese-Chinese. *Atherosclerosis* **172**, 135–142 (2004).
11. St-Pierre, J. *et al.* Visceral obesity attenuates the effect of the hepatic lipase -514C > T polymorphism on plasma HDL-cholesterol levels in French-Canadian men. *Mol. Genet. Metab.* **78**, 31–36 (2003).
12. Manolio, T. Novel risk markers and clinical practice. *N. Engl. J. Med.* **349**, 1587–1589 (2003).
13. Langholz, B., Rothman, N., Wacholder, S. & Thomas, D. C. Cohort studies for characterizing measured genes. *J. Natl Cancer Inst. Monogr.* **26**, 39–42 (1999).
14. Gordis, L. *Epidemiology* 2nd edn (W. B. Saunders, Philadelphia, 2000).
15. Foster, M. W. & Sharp, R. R. Will investments in large-scale prospective cohorts and biobanks limit our ability to discover weaker, less common genetic and environmental contributors to complex diseases? *Environ. Health Perspect.* **113**, 119–122 (2005).
16. Barbour, V. UK Biobank: a project in search of a protocol? *Lancet* **361**, 1734–1738 (2003).
17. Khoury, M. J. The case for a global human genome epidemiology initiative. *Nature Genet.* **36**, 1027–1028 (2004).
18. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356–1360 (2001).
19. Sackett, D. L. Bias in analytic research. *J. Chron. Dis.* **32**, 51–63 (1979).
20. Schlesselman, J. J. *Case–Control Studies: Design, Conduct, and Analysis* (Oxford Univ. Press, New York, 1982).
21. Neyman, J. Statistics: servant of all sciences. *Science* **122**, 401–406 (1955).
22. Taube, A. Matching in retrospective studies, sampling via the dependent variable. *Acta Soc. Med. Ups.* **73**, 187–196 (1968).
23. Wang, S. S., Fridinger, F., Sheedy, K. M. & Khoury, M. J. Public attitudes regarding the donation and storage of blood specimens for genetic research. *Community Genet.* **4**, 18–26 (2001).
24. Bhatti, P. *et al.* Genetic variation and willingness to participate in epidemiologic research: data from three studies. *Cancer Epidemiol. Biomarkers Prev.* **14**, 2449–2453 (2005).
25. Austin, H., Hill, H. A., Flanders, W. D. & Greenberg, R. S. Limitations in the application of case–control methodology. *Epidemiol. Rev.* **16**, 65–76 (1994).
26. Miettinen, O. S. The "case–control" study: valid selection of subjects. *J. Chronic Dis.* **38**, 543–548 (1985).
27. Wacholder, S., Silverman, D. T., McLaughlin, J. K. & Mandel, J. S. Selection of controls in case–control studies. III. Design options. *Am. J. Epidemiol.* **135**, 1042–1050 (1992).
28. Doll, R. Proof of causality. *Persp. Biol. Med.* **45**, 499–515 (2002).
29. Rosenberg, N. A., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).
30. Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J. & Stefansson, K. An Icelandic example of the impact of population structure on association studies. *Nature Genet.* **37**, 90–95 (2005).
31. Ben-Shlomo, Y., Smith, G. D., Shipley, M. & Marmot, M. G. Magnitude and causes of mortality differences between married and unmarried men. *J. Epidemiol. Community Health* **47**, 200–205 (1993).
32. Zeger, S. L., Liang, K. Y. & Albert, P. S. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060 (1998).
33. Kolonel, L. N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
34. The ARIC investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *Am. J. Epidemiol.* **129**, 687–702 (1989).
35. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control. Clin. Trials* **19**, 61–109 (1998).
36. Colditz, G. A., Manson, J. E. & Hankinson S. E. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *J. Womens Health* **6**, 49–62 (1997).
37. Newman, A. B. *et al.* Association of long-distance corridor walk performance with mortality, cardiovascular disease, mobility limitation, and disability. *JAMA* **295**, 2018–2026 (2006).
38. Lloyd-Jones, D. M., Larson, M. G., Beiser, A. & Levy, D. Lifetime risk of developing coronary heart disease. *Lancet* **353**, 89–92 (1999).
39. Troyer, D. A., Mubiru, J., Leach, R. J. & Naylor, S. L. Promise and challenge: markers of prostate cancer detection, diagnosis and prognosis. *Dis. Markers* **20**, 117–128 (2004).
40. Tsai, A. W. *et al.* Coagulation factors, inflammation markers, and venous thromboembolism: the longitudinal investigation of thromboembolism etiology (LITE). *Am. J. Med.* **113**, 636–642 (2002).
41. Leibowitz, H. M. *et al.* The Framingham Eye Study monograph: an ophthalmological and epidemiological study of cataract, glaucoma, diabetic retinopathy, macular degeneration, and visual acuity in a general population of 2631 adults, 1973–1975. *Surv. Ophthalmol.* **24** S335–S610 (1980).
42. Ellenberg, J. H. & Nelson, K. B. Sample selection and the natural history of disease. Studies of febrile seizures. *JAMA* **243**, 1337–1340 (1980).
43. Kannel, W. B. Clinical misconceptions dispelled by epidemiological research. *Circulation* **92**, 3350–3360 (1995).
44. Aleksic, N. *et al.* Factor XIIIA Val34Leu polymorphism does not predict risk of coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. *Arterioscler. Thromb. Vasc. Biol.* **22**, 348–352 (2002).
45. Taubes, G. Epidemiology faces its limits. *Science* **269**, 164–169 (1995).
46. Jamrozik, K., Weller, D. P. & Heller, R. F. Biobank: who'd bank on it? *Med. J. Aust.* **182**, 56–57 (2005).
47. Kannel, W. B. The Framingham Study: its 50-year legacy and future promise. *J. Atheroscler. Thromb.* **6**, 60–66 (2000).
48. Stamler, J. Blood pressure and high blood pressure. Aspects of risk. *Hypertension* **18**, I95–107 (1991).
49. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **26**, S6–S14 (1997).

50. Weis, B. K. *et al.* Personalized exposure assessment: promising approaches for human environmental health research. *Environ. Health Perspect.* **113**, 840–848 (2005).
51. Gauderman, W. J. Sample size requirements for matched case–control studies of gene–environment interaction. *Stat. Med.* **21**, 35–50 (2002).
52. Altshuler, D. *et al.* The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
53. Grant, S. F. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nature Genet.* **38**, 320–323 (2006).
54. Meslin, E. M., Thomson, E. J. & Boyer, J. T. The Ethical, Legal, and Social Implications Research Program at the National Human Genome Research Institute. *Kennedy Inst. Ethics J.* **7**, 291–298 (1997).
55. Prentice, R. L. On the design of synthetic case–control studies. *Biometrics* **42**, 301–310 (1986).
56. Mantel, N. Synthetic retrospective studies and related topics. *Biometrics* **29**, 479–486 (1973).
57. Marshall, E. Whose DNA is it, anyway? *Science* **278**, 564–567 (1997).
58. Triendl, R. Japan launches controversial Biobank project. *Nature Med.* **9**, 982 (2003).

Acknowledgements
The authors express appreciation to M. Boehnke, E. Boerwinkle, B. Foxman, M. Khoury, L. Kuller, J. Ordovas and B. Psaty for their critical review and comments on this manuscript.

Competing interests statement
The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:
OMIM:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Parkinson disease | Schizophrenia | Type 2 diabetes

FURTHER INFORMATION

Biobank Japan:
http: //www.src.riken.go.jp/eng/src/project/person.html
Ethics and Governance Framework of the UK Biobank:
http://www.wellcome.ac.uk/doc_wtd002848.html
Incidence and Prevalence Database:
http://library.dialog.com/bluesheets/html/bl0465.html
National Health Infrastructure Initiative:
http://aspe.hhs.gov/sp/NHII
National Institute of Environmental Health Sciences:
http://www.niehs.nih.gov
NHGRI Ethical, Legal and Social Issues:
http://www.genome.gov/10001618
NHGRI Expert Panel Recommendations for a population-based cohort: http://www.genome.gov/Pages/About/OD/ReportsPublications/PotentialUSCohort.pdf
NIH Genes and Environment Initiative:
http://www.genome.gov/17516707
Responses to NHGRI Request for Information:
http://www.genome.gov/Pages/About/OD/ReportsPublications/ResponsestoRFINot-OD-04-041.pdf
SEER Cancer Statistics Review, 1975-2002:
http://seer.cancer.gov/csr/1975_2002
Swedish National Biobank: http://www.biobanks.se
The TDR Data.com Incidence and Prevalence Database:
http://www.tdrdata.com/IPD/IPD_Init.asp
UK Biobank: http://www.ukbiobank.ac.uk
Women's Health Initiative: http://www.nhlbi.nih.gov/whi
Access to this links box is available online.

ONLINE CORRESPONDENCE

Nature Reviews Genetics publishes items of correspondence online. Such contributions are published at the discretion of the Editors and are subject to peer review. Correspondence should be a scholarly attempt to comment on a specific Review or Perspective article that has been published in the journal. To view the correspondence, please go to our home page at: http://www.nature.com/reviews/genetics and select the link to New correspondence, or, alternatively, go to the archived correspondence at: http://www.nature.com/nrg/archive/correspondence.

The following correspondence has been recently published:

Mining meiosis with genomic models
R. M. Ranganath & G. Venkatachalaiah

This correspondence relates to the article:

HOMOLOGOUS CHROMOSOME INTERACTIONS IN MEIOSIS: DIVERSITY AMIDST CONSERVATION
Jennifer L. Gerton & R. Scott Hawley
Nature Reviews Genetics **6**, 477–487 (2005)