Genetic Epidemiology 7

Genetic epidemiology and public health: hope, hype, and future prospects

George Davey Smith, Shah Ebrahim, Sarah Lewis, Anna L Hansell, Lyle J Palmer, Paul R Burton

Lancet 2005; 366: 1484–98

This is the seventh, and final, paper in a Series on genetic epidemiology.

Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK (Prof G Davey Smith FRCP, Prof S Ebrahim FRCP, S Lewis PhD); Department of Epidemiology, Imperial College, London, UK (A L Hansell MB); Laboratory for Genetic Epidemiology, School of Population Health and Western Australian Institute for Medical Research and Centre for Medical Research, University of Western Australia, Perth, Australia (Prof L J Palmer PhD); and Department of Health Sciences and Genetics, University of Leicester, Leicester, UK (Prof P R Burton MD)

Correspondence to: Prof George Davey Smith zetkin@bristol.ac.uk

Genetic epidemiology is a rapidly expanding research field, but the implications of findings from such studies for individual or population health are unclear. The use of molecular genetic screening currently has some legitimacy in certain monogenic conditions, but no established value with respect to common complex diseases. Personalised medical care based on molecular genetic testing is also as yet undeveloped for common diseases. Genetic epidemiology can contribute to establishing the causal nature of environmentally modifiable risk factors, throught the application of mendelian randomisation approaches and thus contribute to appropriate preventive strategies. Technological and other advances will allow the potential of genetic epidemiology to be revealed over the next few years, and the establishment of large population-based resources for such studies (biobanks) should contribute to this endeavour.

The recent advances covered in this series have equipped genetic epidemiologists with powerful methods for studying the genetic architecture of complex diseases, but direct contributions to public health have been restricted so far. The major current focus is on attempts to use genetic variants to identify individuals who are at high risk of disease, coupled with appropriate management to reduce their risk.[1] The potential of pharmacogenomic studies to contribute to personalised medicine has also been widely heralded.[2–4] Major contributions to either health care or public health are only just beginning to be made. More encouragingly, findings from association studies of well-characterised functional genetic variants are being used by epidemiologists to strengthen causal inferences about modifiable environmental exposures—a strategy sometimes referred to as mendelian randomisation[5–14]—offering a powerful method for observational epidemiology. The continuing integration of genetics into mainstream epidemiology offers enormous potential for both fields; the adoption of well-planned and adequately powered study designs will be essential for future progress. If genetic epidemiology is to make robust contributions to understanding the causes, prevention, and treatment of disease within populations, new ways of thinking and appropriately designed studies are needed. In this article, we discuss the current and potential effects of the genomic revolution on public health science and mainstream epidemiology, especially in the context of the very large-scale population resources (Biobanks) that are being established internationally.

	Genes involved*	Relative risk (current estimate)	Lifetime risk
Reduced risk			
Prostate cancer	*HPC1, HPC2, HPC3*	0·4 (1)	7%
Alzheimer's disease	*APOE, FAD3, XAD*	0·3 (0·3 [for *APOE*])	10%
Increased risk			
Coronary heart disease	*APOB, CETP*	2·5 (1)	70%
Colon cancer	*FCC4, APC*	4·0 (1 [for *APC*])	23%
Lung cancer	*NAT2*	6·0 (1)	40%

**HPC1, HPC2,* and *HPC3*=the three genes for hereditary prostate cancer. *APOE*=gene for apolipoprotein E. *FAD3* and *XAD*=hypothetical genes for familial Alzheimer's dementia. *APOB*=gene for apolipoprotein B. *CETP*=gene for cholesteryl ester transfer protein. *FCC4*=hypothetical gene for familial colon cancer. APC=gene for adenomatous polyposis coli. *NAT2*=gene for N-acetyltransferase 2.

Table 1: **Results of genetic testing in a hypothetical patient in 2010 (from Collins,[18]), and updated estimates of relative risks**

Genomic profiling in the prevention and treatment of common diseases

Since the launch of the human genome project the potential of increased genetic knowledge to improve human health has been widely championed.[15–17] In a striking image from his 1999 Shattuck lecture, Francis Collins, of the US National Human Genome Research Institute, described a hypothetical consultation in 2010 in which a 23-year-old man has a high concentration of cholesterol identified during screening and undergoes extensive genetic testing.[18] Table 1 shows the genotypes that are identified and the relative risks of various diseases with which they might be associated. These numbers are very worrying: 2·5 and 6 times the risk of coronary heart disease and lung cancer, respectively. Unsurprisingly, in this future scenario, three of the 11 variants were fictional names for unknown variants of a type that, Collins predicted, would be identified by 2010. But what of the eight variants that were already known in 1999? With few exceptions, later evidence suggests that these variants are related to much smaller increased risks of disease, if any, and would not be of value within a routine battery of genetic tests applied during medical consultations (panel 1).

Do we merely have to wait a bit longer to achieve Collins' vision of "genetically based, individualised preventive medicine"? For genomic profiling to have a role in public health, the technology should be evaluated on criteria established for screening programmes generally (panel 2), and on these criteria most of the proposed genetic screening tests would fail, either because the excess risk borne by a carrier of the variant is too low or because identification would not point to use of an acceptable

Panel 1: Fate of genetic variants advocated for screening in 2010

How has evidence about the genetic variants advocated for screening in Francis Collins' Shattuck lecture[18] (table 1) fared over the past 6 years?

Coronary heart disease

The patient was said to be at elevated risk of coronary artery disease because of variants in his cholesterol ester transfer protein (*CETB*) and apolipoprotein B (*APOB*) genes. Variants in *CETP* and *APOB* were said to identify a relative risk of coronary heart disease of 2·5, based on small studies up to that date. The CETP TaqIb variant has been extensively investigated, and in a case-control study with 8145 participants—much larger than previous studies—the odds ratio (OR) for coronary heart disease associated with the *B2/B2* genotype was 0·94 (95% CI 0·83–1·06).[19] In the same study, *APOB Asn4311Ser* and *APOBThr71Ile* genotypes were also investigated yielding similarly unimpressive results; the OR for *Ser/Ser* versus *Asn/Asn* was 1·15 (0·91–1·46) and for *Ile/Ile* versus *Thr/Thr* 0·95 (0·82–1·11).

Lung cancer

A large relative risk of 6 was given in Collins' table for lung cancer risk in smokers for variants in the *NAT2* gene, again on the basis of small studies. A case-control study of more than 2000 participants, most of whom were or had been smokers, reported an OR of 0·96 (0·79–1·16) comparing slow versus fast acetylators.[20] It is well recognised that small genetic association studies can show large effect sizes that are not replicated in larger studies.[21,22] Similarly, the association between a genetic polymorphism and disease has been shown to be stronger in the first study than in subsequent research.[23]

Alzheimer's disease

One gene for which the risk of complex disease does seem sizeable is the association of the *APOE* gene with Alzheimer's disease, although a recent cohort study of individuals found that the *APOEϵ4* allele acts as a risk factor for Alzheimer's disease by accelerating onset, but has a more modest effect on lifetime susceptibility.[24] In a further study, the relative risk of Alzheimer's disease for individuals who were heterozygous for ϵ4 was 1·4 (1·0–2·0), and for those were homozygotes it was 3·1(1·6–5·9).[25] Although the *APOE* gene is one of the few genes for which the risk of complex disease risk seems established, in 1995 the American College of Medical Genetics and American Society of Human genetics did not recommend that the gene be used for routine diagnosis or predictive testing for Alzheimer's disease, because this genotype did not provide sufficient sensitivity and specificity to allow it to be used as a test.[26]

Other genes

As for the other genes listed in Collins' table, rare mutations in these genes confer very high risk for a small number of families, but little or no increased risk in the general population. For example, the *ELAC* gene, which lies in the *HPC2* region, is linked to prostate cancer in family studies. A recent meta-analysis of two common polymorphisms in ELAC[27] reported ORs of 1·04 (0·50–1·09) for Leu217 homozygotes and 1·18 (0·98–1·42) for Thr541 homozygotes and heterozygotes combined. The meta-analysis also showed that the largest and most recent study showed no effect associated with either polymorphism.

Colon cancer

Rare mutations in *APC* are related to colon cancer risk, but at the time of Collins' Shattuck lecture, a large risk had been reported that was associated with a common variant, E1317Q.[28] However, a later case-control study of colorectal cancers reported no association between this variant in an analysis comparing cases with spouse controls (OR 0·83, 0·31–2·26). The investigators concluded that E1317Q "does not appear to confer an increased risk for colorectal neoplasia in the general population. Genetic screening for E1317Q is not indicated".[23]

treatment or would not influence an already supported management strategy. However, although we are at the beginning of our ability to map complex disease genes and the promise has not yet been realised for most diseases, a small but increasing number of genes associated with complex diseases have been discovered.[30–34] Concomitantly, there are a growing number of examples of potential clinical importance for diseases and pharmacogenetic reponses in areas such as oncology,[35] inflammatory bowel disease,[36] and infectious disease.[37]

Genetic screening

Although this series focuses principally on the genetics of common complex diseases, it is genetic screening, mainly for monogenic disorders, that has provided most opportunities for potential interaction between genetics and public health. Genetic screening can take several forms—recessive carrier screening, recessive disease screening, autosomal dominant disease screening, pharmacogenetic risk screening, employment risk screening, and complex genetic disease screening (panel 3).[38]

The aim of population genetic screening is to detect individuals who are at high risk of developing a particular disease or of responding badly to a particular treatment. Crucially, such screening is only worthwhile if the early identification of enhanced risk improves the ultimate clinical outcome (panel 2). Furthermore, like all screening tests, there are costs as well as benefits of genetic screening. In particular, there could well be adverse psychological effects associated with knowing that one is at enhanced risk of developing a particular disease, and how these might best be dealt with can be unclear, especially when the risk model is difficult to interpret—(eg, when dealing with mutations of incomplete penetrance (see paper 1 in this series[39]). This consideration raises important and unresolved ethical issues.

Recessive carrier screening

This form of screening aims to identify couples who are at risk of having children affected by a recessive disease and therefore facilitates prenatal diagnosis and informed choice about conception and termination of pregnancy.

Cystic fibrosis has been a major focus, especially since US authorities recommended that couples seeking prenatal advice should be offered such a service.[40–42] A panel of mutations that lead to defective cystic fibrosis transmembrane conductance regulator protein are screened for, with the most common of the 1000 or so identified mutations being used.[43] Such testing is now very widespread in the USA[44] but doubts remain about the accuracy and selection of the tests,[44,45] and unexpected complexity has emerged in the genotype-phenotype relations.[46] There are also concerns about the cost-effectiveness of such screening[45] and the potential psychosocial consequences of a widespread counselling and screening service.[47–51]

Screening strategies have been developed for particular population-of-origin groups, such as Ashkenazi-Jewish populations,[52,53] with predisposition to rare diseases due to rare recessive variants. Populations derived from small founder groups, such as the Finnish population,[54] are at increased risk of rare hereditary diseases and characterised rare mutations could be used for screening.[55]

Recessive disease screening

This form of screening aims to detect homozygotes or compound heterozygotes at an increased but modifiable risk of disease. For example hereditary haemochromatosis is a common autosomal recessive disorder, present in about 1 in 300 people in populations of European origin.[56] In this condition, increased iron absorption results in excessive accumulation.[56] The disease meets many of the guidelines for population genetic screening (panel 2). It is mainly caused by a single mutation in the *HFE* gene;[57,58] around 80% of patients are homozygous for the C282Y mutation,[57] which can be easily, accurately, and inexpensively detected. A second variant, H63D, might increase the risk of haemochromatosis for individuals with a single copy of the C282Y mutation, although the consequences of the H63D mutation are less well understood.[59] The symptoms of the condition are severe and non-specific, but early diagnosis and treatment improves prognosis.[59] The current clinical approach is to search for haemochromatosis in the presence of clinical disease such as kidney failure; however, such patients will already have irreversible complications. The College of American Pathologists has stated that screening is warranted for all people older than 20 years. However, the recommendation is for phenotypic (iron overload) testing rather than genotypic screening,[60] because a negative test for C282Y homozygosity does not rule out disease due to other mutations and there is growing uncertainty about the penetrance of the common mutation. Thus, even for haemochromatosis—regarded as the paradigm of a

Panel 2: **UK National Screening Committee guidelines for appraising viability, effectiveness, and appropriateness of screening programme,[29] adapted for genetic screening tests**

Disorder

- Important health problem
- Epidemiology and natural history adequately understood and genetic risk factors detectable
- Limited number of mutations in responsible gene(s) within target population responsible for high proportion of genetic risk
- Detectable genetic mutations or polymorphisms with high penetrance
- All cost-effective primary prevention interventions implemented as far as practicable

Test

- Simple, safe, precise, and validated genetic screening test
- Acceptable to the population
- Agreed policy on further diagnostic investigation of individuals with positive test result and on choices available to those individuals

Treatment

- Effective treatment or intervention for patients identified as being at risk through genetic testing, with evidence of early treatment consequent on results of genetic testing leading to better outcomes than late treatment initiated after risk becomes evident for other reasons, such as development of symptoms
- Agreed evidence-based policies covering which individuals should be offered treatment and appropriate treatment offered
- Clinical management of the condition and patient outcomes optimised by all health-care providers before participation in a screening programme

Screening programme

- Evidence from high quality randomised controlled trials that genetic screening programme is effective in reducing mortality or morbidity
- Evidence that complete genetic screening programme is clinically, socially, and ethically acceptable to health professionals and the public
- Benefit from the genetic screening programme outweighs physical and psychological harm
- The opportunity cost of the screening programme economically balanced in relation to expenditure on medical care as a whole (ie, value for money)
- Plan for managing and monitoring screening programme and agreed set of quality assurance standards
- Adequate staffing and facilities for testing, diagnosis, treatment and programme management available before start of screening programme
- All other options for managing the condition considered
- Evidence-based information, explaining the consequences of testing, investigation and treatment, available to potential participants to assist them in making an informed choice
- Public pressure for widening the eligibility criteria for the genetic screening test anticipated

Panel 3: **Categories of molecular genetic risks screening (with examples)[38]**

Recessive carrier screening
- Cystic fibrosis
- Fragile X syndrome
- Ashkenazi-Jewish screening panel

Recessive disease screening
- Hereditary haemochromatosis

Autosomal dominant disease screening
- *BRCA1/BRCA2*
- Hereditary non-polyposis colon cancer

Pharmacogenetic risk screening
- Malignant hyperthermia

Employment risk screening
- N-acetyl-transferase and occupational exposure to arylamines

Complex genetic disease screening
- Methylene tetrahydrofolate reductase
- Angiotensin-1-converting enzyme

disease for which genetic screening had major potential[61]—considerable doubts remain. For other widely discussed conditions (eg, risk of venous thromboembolism in relation to factor V Leiden and prothrombin variants), the case for molecular genetic screening is weaker.[62]

Autosomal dominant disease screening

This form of screening identifies individuals who have inherited one or two copies of a dominant disease allele and are therefore at high risk of developing the disease. Around 5% of women who develop breast cancer have a strong hereditary predisposition to the disease, with multiple family members affected, often at an early age.[63] Mutations in one of two genes, *BRCA1* and *BRCA2*, are responsible for susceptibility in most of these so-called breast cancer families, and women with at least one copy of these mutations are also at increased risk for several other cancers, especially ovarian cancer. *BRCA* gene mutations are highly penetrant, although the extent seems to vary from family to family. This variability is reflected in the pattern of familial recurrence[39] and is related to the inherent severity of the pathophysiological consequences of the mutation.[64] It is therefore difficult to predict lifetime risk, and hence to offer appropriate advice. Among carriers opting for prophylactic mastectomy, the number of life-years gained compared with those who opted for surveillance alone was just 2·9 years for those with a low-penetrance mutation.[65] Population-wide screening with *BRCA1* and *BRCA2* mutations is technically difficult and expensive given the many family-specific mutations. In the UK, many clinical genetics services only screen individuals with a strong family history.[66] As in all autosomal dominant diseases of high penetrance (other than those cases caused by a new mutation), family members will very probably be affected, and screening by family history is likely to be more cost-effective than genotyping as an initial step.

Pharmacogenetic screening

Some individuals have severe drug reactions as a result of mutations in genes involved in drug metabolism or drug receptors. Identification of such mutations before treatment can avoid adverse effects. For some rare conditions there are developments that could be of benefit—for example, a rare variant in the *TPMT* gene identifies children with acute leukaemia who are at increased risk of severe side-effects from mercaptopurine.[67] For more widely used drugs, few clear examples indicate a role for pharmacogenetic screening. For instance, malignant hyperthermia, a rare but life threatening condition that is as a result of exposure to anaesthetics and depolarising muscle relaxants in genetically susceptible individuals,[68] has a known genetic cause in some cases. However, although a high proportion of disease cases are autosomal dominant and are the result of variation in the ryanodine receptor gene,[68] many different disease-causing mutations have been identified, which makes screening difficult and expensive at present.

Employment risk screening

This form of screening aims to detect people who are susceptible to specific workplace exposures, allowing them to avoid exposure and reduce their risk. For example, the enzyme N-acetyltransferase is involved in the metabolism of various chemical substances including arylamines, which are used in dry cleaning and other industries. These compounds cause cancers, especially bladder cancer.[69] However, in common with most mutations, the relative risk of cancer is small in genetically susceptible individuals, and all individuals will generally benefit from a reduced exposure to carcinogens. The risk of disease associated with exposure tends to be greater than the risk associated with genotype;[69] at present, primary prevention through limiting exposure is the preferred public-health intervention. There are fears that employers might use genetic screening in preference to primary prevention, and such testing could lead to employment discrimination according to genotype.

Genomic profiling and susceptibility to common diseases

The contribution of genetic epidemiology to an evidence base for conventional genetic screening has not yet delivered an appreciable number of targets that can be implemented within programmes. However, as illustrated by Francis Collins' hypothetical patient, the greatest potential benefit from genomic profiling would be in improving common disease prevention. Currently,

though, prospects for such prevention are few.[70] The main problem is that very few common genetic variants are known to increase risk of common diseases substantially (panel 1).

Where polymorphisms that increase disease risk have been identified, interventions are generally in terms of advice to reduce exposure to lifestyle factors that are, in any case, the target of interventions. Thus, Francis Collins' hypothetical patient was told he had six times the risk of lung cancer compared with other smokers because of his genotype. However, even non-susceptible individuals who smoke are at a high and preventable risk of lung cancer (and other common diseases), so it would be most effective to apply smoking cessation programmes to the whole population. There is little evidence that genetic test results can motivate behavioural change.[70] Indeed a potential problem is that people identified as being at lower risk of lung cancer than other smokers if they continue smoking might be less inclined to quit after such genomic profiling.

Even when confirmed associations between genetic variants and increased disease risk are identified, this new knowledge might not affect prevention strategies. For example, the *MTHFR 677TT* genotype has been suggested to increase coronary artery disease risk by around 20% relative to the *677CC* genotype,[71] and testing for this variant has been advocated to identify individuals at heightened risk.[72] However, if indeed this genotype does increase coronary artery disease risk (the evidence is uncertain[73]), then it does so by increasing homocysteine levels, which can be lowered by folate supplementation. Within each genotype group there will be wide variations in homocysteine levels because there are several polymorphisms in other genes, together with environmental factors, that determine blood homocysteine. Less than 2% of the variability in homocysteine is explained by the *C667T* polymorphism of the *MTHFR* gene.[74] Genotyping only one variant will give a less sensitive indication of risk than simply measuring homocysteine levels. As Humphries and others state, "for a genetic test to be useful in the management of CVD [cardiovascular disease] it must have predictive powers over and above accepted risk factors which can easily be measured, usually inexpensively, and with high reproducibility."[62] If folate supplementation and homocysteine lowering do reduce CVD risk then it would be preferable to recommend folate supplementation to all individuals with high homocysteine levels, irrespective of genotype.

Behavioural and physiological risk factors show substantial clustering, such that people with one adverse factor (elevated cholesterol) are more likely to have others (obesity, insulin resistance or smoking), generating high risk of disease.[75] This clustering arises because social processes or underlying states such as obesity, generate such inter-relationships. By contrast, possession of one risk-increasing genetic variant will generally be independent of others, and consequently the proportion of the population bearing several variants associated with moderate risk, that together could produce substantial elevated risk, would be very small. Overall reduction in disease burden based on population intervention irrespective of genotype will generally be more substantial than intervention targeted according to genotype.[76]

Currently, asking about family history is probably as good a method as any other we have for genetic screening of the general population, though family history does not necessarily imply a genetic cause.[39] So perhaps we should concentrate on family history rather than pursuing a list of currently favoured (though often not confirmed) polymorphisms that might or might not have modest effects on disease risk.[77,78] However, this argument has not deterred the apparently lucrative business of offering genetic profiling for complex disease by many internet-based companies.

Personalised medicine: hype or hope?

The use of genetic variants as screening tests overlaps with, but is distinct from, the notion of personalised medicine, in which precise treatment protocols are envisaged that depend upon genotypic information. Our concern is with common variants that might influence choice of therapeutic regimen for disease prevention or treatment of common disease, rather than the identification of genes expressed in rare diseases that aid selection of specific treatments,[79] such as imatinib for *Bcr-Abl* positive chronic myeloid leukaemia or screening for *TPMT* mutations before treatment of inflammatory bowel disease with azathioprine.[36]

The main developments in this field relate to pharmacogenomics—tailoring of pharmacological treatment of disease or predisease states to possession of genetic variants that influence response. Pharmacogenomic researchers anticipate that developments will "improve the chances of choosing the right drug for a patient by categorizing patients into genetically definable classes that have similar drug effects".[80] In addition to tailoring treatment, pharmacogenetics in common diseases seeks to optimise treatment response, reduce side-effects, and contribute to appropriate scheduling and dosage of pharmacological treatment.[2–4,81,82] As in other areas of genetic epidemiology, pharmacogenetics has been characterised by persistent optimism in the face of failure to replicate initial claims of common variants being related to drug responsiveness, which is likely to be at least partly due to inadequate sample sizes.[83] Other potential reasons for non-replication include poor or inappropriate statistical analysis, poor study design, indirect assessment of causal pathways, complexity of the phenotypes studied, and the complexity of allelic or genotypic contributions to phenotype. One recent adequately powered study, however, reported an apparently robust difference in response to statin therapy; two variant forms of the 3-hydroxymethyl-3-methyl-

Panel 4: Why "mendelian randomisation?"

In his study of peas, Gregor Mendel concluded that "the behaviour of each pair of differentiating characteristics [such as shape of the seeds and colour of the seeds] in hybrid union is independent of the other differences between the two original plants".[96] Karl Correns referred to this "law of independent assortment" in 1900,[97] and suggested that the inheritance of one trait is independent of (ie, randomised with respect to) other traits. The analogy with a randomised controlled trial will be most applicable to parent-offspring designs where the frequency with which one of two alleles from a heterozygous parent is transmitted to offspring with a particular disease is investigated.[5] However, at a population level, traits influenced by genetic variants are generally not associated with the social, behavioural, and environmental factors that confound relationships in conventional epidemiological studies; thus although the so-called randomisation is approximate, rather than absolute, in genetic association studies empirical observations suggest that it applies in most circumstances (table 2 and table 3). The term mendelian randomisation itself was first introduced in a somewhat different context, in which the random assortment of genetic variants at conception is used to provide an unconfounded study design for estimating treatment effects for childhood malignancies.[98,99] The term has recently become widely used with the meaning we ascribe in this article.

The notion that genetic variants can serve as an indicator of the action of environmentally modifiable exposures has been expressed in many contexts. For example, since the mid-1960s, various investigators have pointed out that the autosomal dominant condition of lactase persistence is associated with drinking milk, and thus associations of lactase persistence with osteoporosis, bone mineral density, or fracture risk provide evidence that milk drinking protects against these conditions.[100,101] Similarly, it was proposed in 1979 that as N-acetyltransferase pathways are involved in the detoxification of arylamine, a potential bladder carcinogen, then increased bladder cancer risk in people with genetically determined slow acetylator phenotype suggested that arylamines are involved in the cause of the disease.[102]

Various commentators have since pointed out that the associations of genetic variants of known function with disease outcomes provides evidence about aetiological factors,[103–107] but the key strengths of mendelian randomisation—avoidance of confounding, bias due to reverse causation or reporting tendency, and the underestimation of risk associations due to variability in behaviours and phenotypes—were not emphasised. These key concepts have appeared in scattered sources over the past two decades, most notably in Martijn Katan's suggestion that genetic variants related to cholesterol level could be used to investigate whether the association between low cholesterol and increased cancer risk was real,[108] and by Honkenen and colleagues'[109] understanding of how lactase persistence could better characterise the difficult-to-measure environmental exposure of calcium intake than could direct dietary reports. From 2000[5,12–14] a series of reports have appeared that use the term mendelian randomisation in the way it is used here, and its use is now widespread. The fact that mendelian randomisation is one of a family of techniques referred to as instrumental variable approaches for obtaining robust causal inferences from observational data has also been recognised.[110]

glutaryl coenzyme A reductase gene were associated with falls in total cholesterol of 42 mg/dL (1·1 mmol/L) and 33 mg/dL (0·9 mmol/L).[84] Such a difference would be unlikely to change clinical practice, since the reductions in LDL-cholesterol in both groups would be expected to give substantial cardioprotection and statin dose can be modified in response to observed cholesterol reductions, rather than genotype. It is important to recognise that possession of these variants does not constitute statin responder versus non-responder status, as suggested by some commentators.[3]

Pharmacogenomics is developing rapidly and evidence needs to be gathered to determine whether genetic testing has clinical benefit and is cost effective. The health economics aspects of genetic epidemiology have only recently begun to be explored. Pragmatic randomised controlled trials need to be done to assess the outcomes and costs of drug treatment with and without information from genetic testing.[3]

Mendelian randomisation: strengthening causal inference in observational epidemiology

The basic aim of aetiological observational epidemiology is to identify modifiable causes of disease and through this to contribute to strategies for prevention. This enterprise has, however, had several setbacks. Observational studies have apparently identified robust associations, which are interpreted as probably or possibly causal, but when they are tested in randomised controlled trials, they have proved illusory.[85,86] Examples include hormone replacement therapy and coronary heart disease,[87,88] beta-carotene and lung cancer,[89,90] vitamin C and coronary heart disease,[91,92] dietary fibre and colon cancer,[93] and vitamin E and coronary heart disease.[94,95] It is obvious that the candidate causes receiving the strongest support from observational and mechanistic studies will be the first ones to be assessed in randomised controlled trials, and the many associations reported from observational studies that have not been tested in controlled trials are probably even less likely to be truly causal (figure).

Why have observational studies and randomised controlled trials produced different findings? The most plausible reason for the examples discussed is confounding. Controlling for confounding has proved difficult when the exposure under study is related to many other factors influencing disease risk. In such situations, appropriately designed genetic epidemiological studies can contribute to drawing robust inferences, using an approach that has come to be termed mendelian randomisation (panel 4).

The basic principle in such studies is that if genetic variants either alter the level of or mirror the biological effects of a modifiable environmental exposure that itself alters disease risk, then these genetic variants should be related to disease risk to the extent predicted by their effect on exposure to the risk factor. Common genetic polymorphisms that have a well characterised biological function (or are markers for such variants) can therefore

	C-reactive protein quartile*				p trend across categories
	1	2	3	4	
Hypertension	45·8%	49·7%	57·5%	60·7%	<0·001
BMI (kg/m²)	25·2	27·0	28·5	29·7	<0·001
HDL cholesterol (mmol/L)	1·80	1·69	1·63	1·53	<0·001
Lifecourse socioeconomic position score	4·08	4·37	4·46	4·75	<0·001
Doctor diagnosis of diabetes	3·5%	2·8%	4·1%	8·4%	<0·001
Current smoker	7·9%	9·6%	10·9%	15·4%	<0·001
Physically inactive	11·3%	14·9%	20·1%	29·6%	<0·001
Daily moderate alcohol consumption	22·2%	19·6%	18·8%	14·0%	<0·001

Data are % or mean. n=3529. *Geometric means for quartiles 1, 2, 3, and 4 are 0·42, 1·23, 2·55, and 7·32 mg/L, respectively.

Table 2: **Potential confounders by quartiles of C-reactive protein**[112]

be used to study the effect of a suspected environmental exposure on disease risk.[5,10]

Use of functional genetic variants (or their markers) has several advantages in this respect. First, unlike environmental exposures, genetic variants are not generally associated with the wide range of behavioural, social, and physiological factors that, for example, confound the association between vitamin C and coronary heart disease. Further, aside from the effects of population structure (see paper 4 in this series),[111] such variants will not be associated with other genetic variants, apart from those with which they are in linkage disequilibrium. This latter assumption follows from the law of independent assortment (sometimes referred to as Mendel's second law): hence the term mendelian randomisation. We illustrate these powerful aspects of mendelian randomisation in table 2 and table 3, showing the strong associations between a wide range of variables and blood C-reactive protein (CRP) concentrations, but no association of the same factors with genetic variants in the gene for this protein. The only factor related to genotype is the expected biological influence of the genetic variant on CRP concentrations. Thus, the genetic variant defines groups that differ according to concentrations of CRP but do not differ with respect to the wide range of potential factors that would confound direct associations of such concentrations with an outcome.

	Means or proportions by genotype		P
	GG	GC or CC	
C-reactive protein (mg/L log scale)*	1·81	1·39	<0·001
Hypertension	53·3%	53·1%	0·95
BMI (kg/m²)	27·5	27·8	0·29
HDL cholesterol (mmol/L)	1·67	1·65	0·38
Lifecourse socioeconomic position score	4·35	4·42	0·53
Doctor diagnosed diabetes	4·7%	4·5%	0·80
Current smoker	11·2%	9·3%	0·24
Physically inactive	18·9%	18·9%	1·0
Daily moderate alcohol consumption	18·6%	19·8%	0·56

Data are % or mean. *Geometric mean.

Table 3: **C-reactive protein, and potential confounders by 1059G/C genotype of the *CRP* gene**[112]

Second, inferences from observational studies are prone to bias due to reverse causation. Disease processes can influence exposure levels; for example ill people might start drinking less alcohol, or illness can affect measures of intermediate phenotypes such as cholesterol, CRP, and fibrinogen.

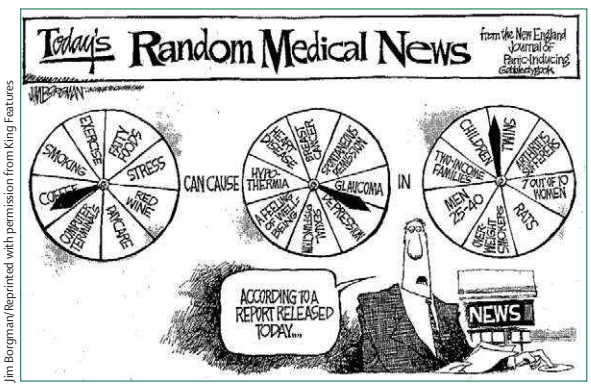
Third, many environmental exposures might be prone to reporting bias. For example, alcohol intake is often poorly reported, with a tendency for heavy drinkers to underestimate their intake.[113,114] A genetic variant related to exposure will not, of course, be altered by knowledge of disease status. Thus, the strong association between the null variant of the aldehyde dehydrogenase 2 gene and usual amount of alcohol consumption would mean that this genotype can serve as a useful—unbiased and unconfounded—marker of usual amount of alcohol consumption before the development of disease.[115] Finally, a genetic variant will relate to long-term levels of an exposure (sometimes from before birth), and if the variant is taken as a proxy for such exposure it will not be affected by the measurement error inherent in phenotypes that have high within-individual variability.

Categories of mendelian randomisation

Several categories of inference can be drawn from studies using mendelian randomisation.[10,11,116] In the most direct forms, genetic variants can be related to the probability or level of exposure (exposure propensity) or to intermediate phenotypes believed to affect disease risk. Less direct evidence can come from genetic variant-disease associations that indicate that a particular biological pathway could be of importance, perhaps because the variants modify the effects of environmental exposures.

Implications of mendelian randomisation study findings

Establishing the causal influence of environmentally modifiable risk factors from mendelian randomisation designs informs policies for improving population health through population-level interventions, not through genetic screening to identify those at high risk. For example, the implications of studies on maternal *MTHFR* genotype and risk of neural-tube defect (NTD) in offspring



Jim Borgman/Reprinted with permission from King Features

Figure: **Observational epidemiology—confusing the public?**

is that population risk for NTDs can be reduced through increased folate intake periconceptually and in early pregnancy. It does not suggest that women should be screened for *MTHFR* genotype; women without the *TT* genotype but with low folate intake experience a preventable risk of having babies with NTDs. Similarly, establishing the association between genetic variants (such as familial defective *ApoB*) and raised cholesterol and coronary heart disease risk strengthens causal evidence that increased blood cholesterol is a modifiable risk factor for coronary heart disease across the whole population. Even though the population attributable risk for coronary heart disease for this specific variant is tiny it usefully informs public-health approaches to improving population health. This aspect of the approach illustrates its distinction from the conventional risk identification and genetic screening outcomes of genetic epidemiology.

Limitations of mendelian randomisation

The limitations of mendelian randomisation are important, and have been discussed previously (panel 5).[5,7,8,111,117] The major limitation is common to all genetic association studies; the failure to establish reliable associations between genotype and intermediate phenotype, or between genotype and disease. This limitation is largely related to study design issues, especially small sample size.[118] A lack of suitable polymorphisms for studying modifiable exposures of interest will limit the application of mendelian randomisation. For example, despite extensive work on genetic factors that could be related to concentrations of vitamin C,[119] functional genetic variants suitable for mendelian randomisation approaches have not yet been identified. As with the other limitations of mendelian randomisation, increased understanding in basic biology should strengthen the usefulness of the approach. Other limitations include confounding by linkage disequilibrium, pleiotropy, canalisation (developmental compensation),[5,120–122] and the difficulty of properly interpreting the complex biology that might underlie a particular trait. Interpretation of findings from studies that seem to fall within the remit of mendelian randomisation can often be complex, as has been discussed with respect to *MTHFR* and folate intake.[5] A second example is the association of extracellular superoxide dismutase (EC-SOD) and coronary heart disease. EC-SOD is an extracellular scavenger of superoxide anions and thus genetic variants associated with higher circulating EC-SOD levels might be expected to mimic higher levels of antioxidants. However, findings are actually the opposite—carriers of such variants have increased risk of coronary heart disease.[123] A possible explanation of this apparent paradox is that the higher concentrations of circulating EC-SOD associated with the variant might arise from movement of EC-SOD from arterial walls and into the circulation; thus the in-situ antioxidative properties of these arterial walls is lower in individuals with the variant associated with higher circulating EC-SOD. The complexity, and sometimes speculative nature, of the appropriate interpretation of findings such as these detracts from the transparency that otherwise makes mendelian randomisation attractive.

Biobanks

Several articles have begun to address the features of a good genetic association study.[31,124–127] Focus on study design has increased because genetic association studies of complex phenotypes have typically failed to discover susceptibility loci or have failed to replicate those findings that initially seemed positive.[118,125,127–132]Despite the widespread use of genetic case-control studies, their lack of consistency is a well recognised limitation.[118,128,129] This lack of reproducibility is often ascribed to small samples with inadequate statistical power, biological and phenotypic complexity, population-specific linkage disequilibrium, effect-size bias, and population substructure.[118,129,130,133,134] Other possible reasons for the non-replication of true-positive results include inter-investigator and interpopulation heterogeneity in study design, analytical methods, phenotype definition, genetic structure, environmental exposures, and the choice of genetic markers that are genotyped. Large sample sizes (thousands rather than hundreds, generally), rigorous p-value thresholds, and replication in multiple inde-

***Panel 5:* Limitations of mendelian randomisation[5,11]**

- Failure to establish reliable association between genotype and intermediate phenotype or genotype and disease
- Confounding of associations between genotype, intermediate phenotype, and disease through linkage disequilibrium or population stratification
- Pleiotropy and the multifunctionality of genes
- Canalisation and developmental stability
- Complexity of interpretations of association between genotype, intermediate phenotype, and disease
- Lack of suitable polymorphisms for studying modifiable exposures of interest

pendent datasets are necessary for reliable results.[43,118,126,128,135]

These concerns have driven the development of a number of large-scale biobanking projects. Biobanks are so-called because they involve the systematic storage of biological material (eg, blood or extracted DNA) and information from a large number of people. Biobanks may be either disease oriented or population-based.[136] The former gather material and information from people after they have developed a specific disease (or any one of a broader class of diseases). Tumour banks are a good example and represent a resource that could be used not only for research but also to guide the clinical management of individual patients. Population-based biobanks store material and information from people recruited from the general population, often on the basis of location of residence. Such biobanks are mainly used for bioclinical research. UK Biobank is a population-based research biobank.

See http://www.ukbiobank.ac.uk/

From a more conventional perspective, most such biobanks, including UK Biobank itself, are traditional cohort studies. Fundamentally, therefore, there is nothing very new about the modern concept of biobanks—the differences are the sheer size of the largest initiatives being proposed and that a particularly strong emphasis is placed on obtaining biological material. However, although the fundamentals might not really have changed, the most ambitious biobank projects are now so large that they present a number of unique scientific, logistical, and political challenges to national funding organisations and to the scientific communities that must work together to design, manage, and exploit them. The biobanks in table 4 have the stated aim of enrolling a minimum of 100 000 people with collection not only of conventional epidemiological information but also of blood with the explicit intention to undertake genetic analysis.

See http://www.ukdnabank.mrc.ac.uk/

How large is large?

Given the obvious emphasis that has been placed on sample size by recent biobanking initiatives (table 4), just how large do infrastructural projects of this nature really need to be? Power calculations undertaken for UK Biobank[137] indicate that, even under ideal circumstances, if 80% power is required to detect small direct effects, such as an odds ratio around 1·15–1·30 associated with a binary exposure (genetic or environmental) with a population prevalence between 10% and 25%, at least 5000 cases of the disease of interest are needed. Here, ideal circumstances means, for example, that there is minimum misclassification error and that one is interested in candidate genes and can therefore work with significance no more rigorous than $p<0·0001$. A minimum of 5000 cases, and ideally 10 000, is also required to provide 80% power to detect a moderately sized interaction effect (eg, an interaction odds ratio around 1·5–2·0 between two binary exposures each with a population prevalence between 10% and 25%). 5000 cases would ideally be needed for each disease of interest in a national biobank.[138]

Such a large number of cases can be amassed in several ways: (1) a coordinated infrastructure for large genetic case-control studies might be constructed; (2) a very large prospective cohort study could be set up, and incident cases would inevitably accumulate; (3) international groups might work together to harmonise study designs and then pool information across a number of biobanks or large pre-existing cohort studies. Excellent examples already exist in the UK of initiatives of the first type (ie, those set up with the primary intention of providing an infrastructure for large case-control studies). The UK Medical Research Council (MRC) DNA Network includes 13 extensive series of cases of various important complex diseases and, following on from this, the Wellcome Trust Case Control Consortium is taking a large number of

	Sample Size (n)	Recruitment	Age at recruitment (years)	URL
Cohort studies				
EPIC Europe	>500 000	1993–97	45–74	http://www.iarc.fr/epic/centers/iarc.html
ProtecT Study	120 000	1999–2006	50–69	http://www.epi.bris.ac.uk/protect/index.htm
Kadoorie Study China	500 000	Ongoing	35–74	http://www.ctsu.ox.ac.uk
Prospective Study				
Mexico	200 000	Ongoing	>40	http://www.ctsu.ox.ac.uk/projects/mexicoblood.shtml
UK Biobank	500 000	2006–10	40–69	http://www.ukbiobank.ac.uk/
Birth Cohorts				
Mother and Child Cohort Study (Norway)	100 000 babies*	2001–05	At birth	http://www.fhi.no/
Danish National Birth Cohort	100 000 babies†	1997–2002	At birth	http://www.serum.dk/sw9314.asp
Twin Cohorts				
GenomEUtwin	>600 000 twin pairs	Various	Various	http://www.genomeutwin.org/
Total populations				
Decode Genetics	>100 000	Iceland	Various	http://www.decode.com/
Estonian Genome Project	>100 000‡	Estonia	Various	http://www.geenivaramu.ee
Western Australian Genome Project	About 2 000 000§	Australia	Various	http://www.genepi.com.au/wagp

*Blood also from mothers and as many fathers as possible.†Blood taken from mothers and umbilical cord blood. ‡Recruitment aim for end of 2007, if further funding obtained. §Study being piloted in 2006.

Table 4: **Large population-based research biobanks (planned and current)**

cases from each of eight complex diseases (some from the MRC network) and will compare them to geographically representative controls sampled from the 1958 Birth Cohort[139] and from a national sample of blood transfusion donors. In this article, however, we focus on infrastructural initiatives of the second and third type.

UK Biobank

UK Biobank is a good example of a population-based research biobank. It is to be a multipurpose research platform that will take the form of a very large cohort study that will recruit 500 000 middle-aged volunteers (40–69 years) from across Britain. The first participants are now being enrolled into pilot evaluations and the aim is to commence the main study in the first half of 2006. In keeping with a number of other large initiatives underway internationally (table 4), recruitment will be population-based rather than disease-based or exposure-based, and will be undertaken in six large collection regions. UK Biobank aims to encompass all elements of British society, and most of the inferences that it generates should be generalisable to the community as a whole. However, the study is not intended to be precisely representative of the general population in Britain. Social, demographic, and health data will be obtained via questionnaire and from a physical examination. Blood will be taken and stored as a source of DNA and for biomarker-based exposure and phenotype assessment. Once recruited, the state of health of individual participants will be monitored via the health-care information systems and, in particular, new cases of important complex disease will be identified. A substantial component of the research involving UK Biobank will be nested case-control studies. Because incident disease is to be identified from routine information systems, investment will have to be made in the validation and classification of the cases that are to be used in the nested studies.

Why are governments investing in large prospective cohort studies of this sort? Some distinguished scientists argue that biobanks are not necessary and that the money would be better spent on case-control studies or on other forms of research in population genetics.[13,140] In our view, we must maintain a full range of complementary study types if science is to advance in the best way in the biomedical arena. So, the key question is whether a large cohort of the type exemplified by UK Biobank contributes something that is not available via other study types. The answer is that only a cohort study enables a full range of exposure and outcome information to be gathered prospectively. If assessment is undertaken retrospectively, both systematic and random errors are more likely to distort the measurement of premorbid lifestyle and environment and measured relations with disease. These errors[141] not only subsume various types of information bias and selection bias but also include biologically mediated reverse causality. Such biases make it extremely difficult, if not impossible, to tease out the subtle effects that could be the only measurable effect of the determinants of a complex disease. Prospective cohort studies are not immune from errors and biases, but in many settings they are much less susceptible than retrospective studies. We acknowledge that retrospective assessment at the time of disease occurrence can sometimes be preferable. For example, in studying the effect of oral contraceptive use on venous thrombosis, interest might focus on exposure in the 3 months immediately before the thrombosis rather than historical exposure on recruitment to a cohort. However, these are not arguments against cohort studies. They confirm the need for a range of studies with complementary epidemiological designs.

Table 5 presents the expected rate of accrual of cases for selected complex diseases in UK Biobank.[142] The expected times to each threshold take account of the fact that participants in cohort studies tend to be unusually healthy. They also assume that the proportion of participants that are prepared to remain in active contact with the study over time will be similar to the loss-to-follow-up profile of the 1958 Birth Cohort Study[143] and the Whitehall Study.[144] Rigorous investigation of the joint effects of a genetic and an environmental determinant within a nested case-control study requires at least 5000, and ideally at least 10 000, cases of a complex disease.[142,145] Any cohort design that is much smaller than 500 000 participants will not generate enough cases fast enough for many conditions, particularly if scientific interest centres on a homogeneous subset of cases of the disease of interest (eg, haemorrhagic stroke rather than any stroke).

Role of large genetic cohort studies

Large genetic cohort studies offer several important scientific opportunities. Nested case-control studies based within a cohort permit study of the joint effect of genes and premorbid environment and lifestyle on a disease of interest using prospectively obtained measures of the non-genetic exposures in a population-based sample of cases. Cohort studies also support exposure-based studies, in which subsets of the cohort are investigated intensively by use of comparison groups that are defined not by disease status, as in a nested case-control study, but by exposure status. Such studies can investigate and compare the function of an intermediate biological pathway in participants with a genotypic or environmental exposure of interest with that in people who are unexposed. Genotype-based studies will become commonplace when extensive genotyping of whole cohorts becomes economically feasible.[146] A cohort study allows for repeated assessments of key phenotypes and exposures and supports studies of the genetic and environmental determinants of disease progression. It also enables repeated assessments of exposure measures, reducing random measurement error and allowing analysis of the structure of exposure variation over time. Additionally, large cohort studies provide a solid foundation for

research based on mendelian randomisation.[13] They enable the simultaneous assessment of disease-gene and intermediate phenotype-gene associations in large numbers of people. Furthermore, intermediate phenotypes assessed at recruitment will usually be pre-morbid.

See http://www.p3gconsortium.org

Other uses of large genetic cohort studies include the provision of a sound platform for other types of research that, because of cost, would not usually be based on a cohort study. If a biobank is being set up for other purposes, the marginal cost of these additional types of research can be low. For example, UK Biobank will provide a cost-effective source of population-based cases and controls for nested case-control studies that are aimed at investigating simple disease-gene associations. Furthermore, it will enable case-control studies based on prevalent cases at recruitment as well as those based on incident cases. A national population-based study that contains participants that have been well characterised both in terms of exposure and outcome also provides a source of common controls, for comparison (in terms of genotype frequencies) with case groups either within, or external to, the study. In the UK, this would potentially be of particular value for research based on cases from ethnic minority populations.

See http://www.genomeutwin.org

See http://www.cdc.gov/genomics/hugenet/default.htm.

International harmonisation of biobanks

The pooling of data between biobanks nationally and internationally offer several benefits. First, it will allow investigation of diseases such as stomach cancer (table 5) and ovarian cancer that are not rare but are not common enough to generate 5000 cases even in a cohort study including 500 000 participants. Second, they will support the study of very modest associations between causal determinants and common diseases. Third, they will enable powerful analyses based on homogeneous subgroups within disease categories, or on cases in particular strata, possibly defined by age, sex, or ethnic origin. Fourth, data pooling will enable analyses to be undertaken earlier than they could be within a single prospective study. Fifth, the synthesis of information from cohort studies from around the world provides the potential to investigate the effect of a broader range of lifestyles.

	Years to achieve cases numbering				
	1000	2500	5000	10 000	20 000
Non-cancers					
Myocardial infarction and coronary death	2	4	5	8	14
Diabetes mellitus	2	3	5	7	11
Chronic obstructive pulmonary disease	4	6	9	14	27
Rheumatoid arthritis	7	15	36	..	..
Cancers					
Breast cancer (female)	4	7	11	19	..
Colorectal cancer	6	10	15	25	..
Lung cancer	7	13	22	..	..
Stomach cancer	17	36	..	..	..

Table 5: **The expected rate of accrual of incident cases of selected complex diseases in UK Biobank[141]**

Several international groups are already working on the harmonisation of large population-based research biobanks. The Public Population Project in Genomics[148] (P3G) is increasingly seen as the global umbrella organisation for biobanks of this type. The European Union funded Population Biobanks project, led from Norway, involves collaboration between P3G researchers and population scientists involved in Co-ordination of Genome Research Across Europe (COGENE). Population Biobanks represents a natural evolution of the GenomEUtwin initiative, led from Finland, and will scope out opportunities and difficulties inherent to the harmonisation of biobanks. The European Prospective Investigation into Cancer and Nutrition study[147] (table 4) was designed as a harmonised group of cohort studies running in ten European countries. The Human Genome Epidemiology Network (HuGENet) provides an international network that is "committed to the assessment of the impact of human genome variation on population health and how genetic information can be used to improve health and prevent disease".

Biobanks: conclusions

Ultimately, genetic knowledge will only be useful in the clinical arena if it can be placed in an epidemiological and medical or public-health context.[107–152] We think that the main purpose of the science to be underpinned by biobanks (and other large population-based genetic epidemiological studies) will be to inform our knowledge of the mechanisms linking causal determinants to disease and disease progression. Knowledge of these mechanisms will ultimately lead to new diagnostic, preventative, and therapeutic interventions that will have an important effect on clinical medicine and the health of the public. An implication of this view is that, as in mendelian randomisation, associations with genetic determinants that have a small population attributable risk or a small relative risk can nevertheless provide important information about causal pathways.

The major payoffs from biobanks based on a prospective cohort design are unlikely to be realised before the second or third decades after recruitment (table 5). That said, there could be some quick wins. For example, cross-sectional analyses based on common binary or quantitative phenotypes at recruitment (eg, diabetes mellitus or FEV_1) will allow powerful population-based confirmation, or non-replication, of genetic associations previously identified in subpopulations. Similar opportunities will exist for certain questions in pharmacogenetics.[153] But implementation of research findings in clinical practice is slow and can take a decade or more.[154] Biomedical and public-health scientists must not overstate the pace at which returns can be expected.

Although not without controversy,[13,140] it is now widely accepted[138,155] that large genetic cohort studies have an important role in furthering our understanding of complex human disease. In the face of overwhelming

uncertainty about how best to proceed with the discovery of genes for complex human disease and how best to make use of the discoveries that we do make, it is important to emphasise that everything in human genetics is context-specific: no one study design or analytic approach will be the best for all circumstances. For example, some complex phenotypes can be modulated by many rare alleles, whereas others can be modulated by fewer common alleles, and this variability has profound implications for the best way to identify such genes. A flexible, mixed approach is desirable and an intensive period of hypothesis-free information collection is necessary. For optimum progress, genetic cohort studies, case-control studies, and family studies will all be needed, and all will have to be large.

The future

The past decade has been an important time for human genetics. Growth in technical capacities and genomic knowledge has been tempered by initial failures to find genes for complex phenotypes with any strategy—linkage or association. Our statistical capacities and our ability to process and interpret data still lag behind the technical capability to produce very large amounts of genomic data. An unfortunate feature of the genomics revolution has been a tendency to hyperbole, leading to unrealistic expectations about the scope and timing of the integration of disease-gene discovery into clinical medicine and epidemiology and, in turn, to scepticism within the academic community. Those investigating the pathogenesis of complex diseases do well by not adding to the hyperbole surrounding genetic epidemiology and by communicating realistic expectations.

Where do we stand with regard to the discovery of genes for complex human disease? Most such diseases will almost certainly involve multiple disease-predisposing genes of modest individual effect, gene-gene interactions, gene-environment interactions, and interpopulation heterogeneity of both genetic and environmental determinants of disease. These all impair statistical power, sometimes seriously, and both the initial detection of genes and the subsequent replication of positive results are very difficult.[118,125,128,156] However, we now have much greater insight into the difficulty of the task. Part of the purpose of this series has been to describe some of the ways in which genetic epidemiologists, working with biological and clinical scientists, have contributed to overcoming the difficulties. The successful localisations of some genes for complex diseases[32–36] suggests that we can be at least cautiously optimistic about the future. A recent trend towards the amalgamation of genetic epidemiology with mainstream epidemiology provides additional grounds for optimism. Traditional epidemiologists, genetic epidemiologists and statisticians, bioinformaticians, geneticists, and clinical and public-health scientists have much to learn from one another. We hope that, working together, they can solve problems in study design, conduct, analysis, and interpretation that make gene discovery and replication of findings so difficult.

Traditional epidemiology has already started to benefit from the gene-based approach to causal inference known as mendelian randomisation.[5] The use of genetic data in epidemiological investigations offers fresh hope for a discipline beleaguered by the difficulty of identifying small causal associations against a background of bias, confounding, reverse causality, and aetiological heterogeneity.[85,86] Our focus emphasises the extent to which traditional thinking, based on mainstream epidemiology and medical statistics, has influenced our view of the statistical and explanatory power of genetic studies. Accepted sample size requirements have increased by an order of magnitude over the past decade[18,31,126,127,129,131] and are likely to go higher yet.[142]

The series has emphasised the parallels between traditional epidemiology and genetic epidemiology. For many years genetic epidemiology has had a reputation for being mysterious and difficult, and some of its methods are indeed hard to understand in detail. But the same is true of traditional epidemiology, and much of the mystery of genetic epidemiology has arisen from the use of different terminology to describe basic concepts that arise elsewhere. We hope that this final paper and the six previous ones have helped to demystify genetic epidemiology, and that some readers who have deliberately avoided the area will now feel more confident to explore the published work and undertake their own research.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

We are grateful for the constructive suggestions of a number of anonymous referees, several of which we incorporated directly into our manuscript. The research programme in genetic epidemiology at the University of Leicester is, and has been, supported in part by MRC Cooperative Grant#G9806740, Program Grant # 00\3209 from the National Health and Medical Research Council of Australia and by Leverhulme Research Interchange Grant # F/07134/K. George Davey Smith's work on mendelian randomisation is supported by the World Cancer Research Fund (ref no: 2004/18). Anna Hansell is funded by UK Biobank.

References

1 Bell J. Predicting disease using genomics. *Nat Rev* 2004; **429:** 453–56.
2 Johnson JA. Pharmacogenetics: potential for individualized drug therapy through genetics. *Trends Genet* 2003: **19:** 660–66.
3 Haga SB, Burke W. Using pharmocogenetics to improve drug safety and efficacy. *JAMA* 2004; **291:** 2869–71.
4 Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. *Nature* 2004; **429:** 464–68.
5 Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003; **32:** 1–22.
6 Davey Smith G, Harbord R, Ebrahim S. Fibrinogen, C-reactive protein and coronary heart disease: does Mendelian randomization suggest the associations are non-causal? *QJM* 2004; **97:** 163–66.
7 Brennan P. Commentary: Mendelian randomization and gene-environment interaction. *Int J Epidemiol* 2004; **33:** 17–21.
8 Tobin MD, Minelli C, Burton PR. Commentary: Development of Mendelian randomisation: from hypothesis test to 'Mendelian deconfounding'. *Int J Epidemiol* 2004; **33:** 21–25.
9 Keavney B. Commentary: Katan's remarkable foresight: genes and causality 18 years on. *Int J Epidemiol* 2004; **33:** 11–14.

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

11 Davey Smith G. Randomised by (your) god: unbiased effect estimates from an observational study design. *J Epidemiol Community Health* (in press).

12 Youngman LD, Keavney BD, Palmer A, et al. Plasma fibrinogen and fibrinogen genotypes in 4685 cases of myocardial infarction and in 6002 controls: test of causality by "Mendelian randomization". *Circulation* 2000; **102** (suppl 2): 31–32.

13 Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358:** 1356–60.

14 Fallon UB, Ben-Shlomo Y, Davey Smith G. Homocysteine and coronary heart disease. http://heart.bmjjournals.com/cgi/eletters/85/2/153. *Heart Online*, March 14, 2001 (accessed Aug 30, 2005).

15 Caskey CT. DNA-Based Medicine: Prevention and Therapy. In: Kevles DJ, Hood L, eds. The Code of Codes: Scientific and Social Issues in the Human Genome Project. London: Harvard University Press, 1993: 112–35.

16 Watson JD. A Personal View of the project. In: Kevles DJ, Hood L, eds. The Code of Codes: Scientific and Social Issues in the Human Genome Project. London: Harvard University Press, 1993: 164–73.

17 Bell J. The new genetics in clinical practice. *BMJ* 1998; **31:** 618–20.

18 Collins FS. Medical and societal consequences of the human genome project. *NEJM* 1999; **341:** 28–37.

19 Keavney B, Palmer A, Parish S, et al. Lipid related genes and myocardial infarction in 4685 cases and 3460 controls: discrepancies between genotype, blood lipid concentrations, and coronary disease risk. *Int J Epidemiol* 2004; **33:** 1002–13.

20 Zhou W, Liu G, Thurston SW, et al. Genetic polymorphisms in N-acetyltransferase-2 and Microsomal Expoxide Hydrolase, cumulative cigarette smoking, and lung cancer. *Cancer Epidemiol Biomarkers Prev* 2002; **11:** 15–21.

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

22 Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003; **361:** 567–71.

23 Hahnloser D, Petersen GM, Rabe K, et al. The APC E1317Q variant in adenomatous polyps and colorectal cancers. *Cancer Epidemiol Biomark Prev* 2003; **12:** 1023–28.

24 Khachaturian AS, Corcoran CD, Mayer LS, Zandi PP, Breitner JC, Cache County Study Investigators. Apolipoprotein E epsilon4 count affects age at onset of Alzheimer disease, but not lifetime susceptibility: The Cache County Study. *Arch Gen Psychiatry* 2004; **61:** 518–24.

25 Qiu C, Kivipelto M, Aguero-Torres H, Winblad B, Fratiglioni L. Risk and protective effects of the APOE gene towards Alzheimer's disease in the Kungsholmen project: variation by age and sex. *J Neurol Neurosurg Psychiatry* 2004; **75:** 828–33.

26 Statement on use of apolipoprotein E testing for Alzheimer disease. American College of Medical Genetics/American Society of Human Genetics Working Group on ApoE and Alzheimer disease. *JAMA* 1995; **274:** 1627–29.

27 Severi G, Giles GG, Southey MC, et al. ELAC2/HPC2 polymorphisms, prostate-specific antigen levels, and prostate cancer. *J Natl Cancer Inst* 2003; **95:** 818–24.

28 Frayling IM, Beck NE, Ilyas M, et al. The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. *Proc Natl Acad Sci USA* 1998; **95:** 10722–27.

29 UK National Screening Committee. The UK National Screening Committee's Criteria for appraising the viability, effectiveness and appropriateness of a screening programme. http://www.nsc.nhs.uk/pdfs/criteria.pdf (accessed Oct 4, 2005).

30 The International HapMap Consortium. The International HapMap Project. *Nature* 2003; **426:** 789–96.

31 Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004; **5:** 89–100.

32 Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411:** 599–603.

33 Begovich AB, Carlton VE, Honigberg LA, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet* 2004; **75:** 330–37.

34 Arnold SE, Talbot K, Hahn CG. Neurodevelopment, neuroplasticity, and new genes for schizophrenia. *Prog Brain Res* 2005; **147:** 319–45.

35 Ren R Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat Rev Cancer* 2005; **5:** 172–83.

36 Kelleher D, Farrell R, McManus R. Pharmacogenetics of inflammatory bowel disease. *Novartis Found Symp* 2004; **263:** 41–53; discussion 53–56, 211–18.

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

38 Grody WW. Molecular genetic risk screening. *Annu Rev Med* 2003; **54:** 473–90.

39 Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. *Lancet* 2005; **366:** 941–51.

40 Genetic testing for cystic fibrosis. National Institutes of Health Consensus Development Conference Statement on genetic testing for cystic fibrosis. *Arch Intern Med* 1999; **159:** 1529–39.

41 Grody WW, Cutting GR, Klinger KW, Richards CS, Watson M, Desnick RJ. Laboratory standards and guidelines for population-based cystic fibrosis carrier screening. *Genet Med* 2001; **3:** 149–54.

42 American College of Obstetrics and Gynecology and American College of Medical Genetics. Preconception and prenatal carrier screening for cystic fibrosis, clinical and laboratory guidelines, 2001. Washington DC: American College of Obstetrics and Gynecology publication, 2001.

43 Strom CM, Crossley B, Redman JB, et al. Cystic fibrosis screening: lessons learned from the first 320,000 patients. *Genet Med* 2004; **6:** 136–40.

44 Vastag B. Cystic fibrosis gene testing a challenge. *JAMA* 2003; **289:** 2923–24.

45 Palomaki GE. Prenatal screening for cystic fibrosis: an early report card. *Genet Med* 2004; **6:** 115–16.

46 Richards CS, Grody WW. Prenatal screening for cystic fibrosis: past, present and future. *Expert Rev Mol Diagn* 2004; **4:** 49–62.

47 Gordon C, Walpole I, Zubrick SR, Bower C. Population screening for cystic fibrosis: knowledge and emotional consequences 18 months later. *Am J Med Genet* 2003; **120:** 199–208.

48 Marteau TM, Dundas R, Axworthy D. Long-term cognitive and emotional impact of genetic testing for carriers of cystic fibrosis: the effects of test result and gender. *Health Psychol* 1997; **16:** 51–62.

49 Marteau TM, Michie S, Miedzybrodzka ZH, Allanson A. Incorrect recall of residual risk three years after carrier screening for cystic fibrosis: a comparison of two-step and couple screening. *Am J Obstet Gynecol* 1999; **181:** 165–69.

50 Honnor M, Zubrick SR, Walpole I, Bower C, Goldblatt J. Population screening for cystic fibrosis in Western Australia: community response. *Am J Med Genet* 2000; **93:** 198–204.

51 Clausen H, Brandt NJ, Schwartz M, Skovby F. Psychological and social impact of carrier screening for cystic fibrosis among pregnant women: a pilot study. *Clin Genet* 1996; **49:** 200–05.

52 Eng CM, Schechter C, Robinowitz J, et al. Prenatal genetic carrier screening using triple disease screening. *JAMA* 1997; **278:** 1268–72.

53 Eng CM, Desnick RJ. Experiences in molecular-based prenatal screening for Ashkenazi Jewish genetic diseases. *Adv Genet* 2001; **44:** 275–96.

54 Norio R. Finnish disease Heritage I: characteristics, causes, background. *Hum Genet* 2003; **112:** 441–56.

55 Pastinen T, Perola M, Ignatius J, et al. Dissecting a population genome for targeted screening of disease mutations. *Hum Mol Genet* 2001; **10:** 2961–72.

56 Godard B, ten Kate L, Evers-Kiebooms G, Ayme S. Population genetic screening programmes: principles, techniques, practices, and policies. *Eur J Hum Genet*. 2003; **11** (suppl 2): S49–87.

57 Feder JN, Gnirke A, Thomas W, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nat Genet* 1996; **13:** 399–408.

58 Byrnes V, Ryan E, Barrett S, Kenny P, Mayne P, Crowe J. Genetic hemochromatosis, a Celtic disease: is it now time for population screening? *Genet Test* 2001; **5:** 127–30.

59 Niederau C, Strohmeyer G. Strategies for early diagnosis of haemochromatosis. *Eur J Gastroenterol Hepatol* 2002; **14:** 217–21.

60 Witte DL, Crosby WH, Edwards CQ, Fairbanks VF, Mitros FA. Practice guideline development task force of the College of American Pathologists. Hereditary hemochromatosis. *Clin Chim Acta*. 1996; **245:** 139–200.

61 Beaudlet AL. Making genomic medicine a reality. *Am J Hum Genet* 1999; **64:** 1–13.

62 Humphries SE, Ridker PM, Talmud PJ. Genetic testing for cardiovascular disease susceptibility: a useful clinical management tool or possible misinformation? *Arterioscler Thromb Vasc Biol* 2004; **24:** 628–36.

63 Eby N, Chang-Claude J, Bishop DT. Familial risk and genetic susceptibility for breast cancer. *Cancer Causes Control* 1994; **5:** 458–70.

64 Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 2003; **72:** 1117–30.

65 Griffith GL, Edwards RT, Gray J, et al. Estimating the survival benefits gained from providing national cancer genetic services to women with a family history of breast cancer. *Br J Cancer* 2004; **90:** 1912–19.

66 Eccles DM. Hereditary cancer: guidelines in clinical practice. Breast and ovarian cancer genetics. *Ann Oncol*. 2004; **15** (suppl 4)**:** 133—38.

67 McLeod HL, Krynetski EY, Relling MV, Evans WE. Genetic polymorphisms of TPMT and its clinical relevance for childhood acute lymphoblastic leukaemia. *Leukemia* 2000; **14:** 567–62.

68 McCarthy TV, Quane KA, Lynch PJ. Ryanodine receptor mutations in malignant hyperthermia and central core disease. *Hum Mutat* 2000; **15:** 410–17.

69 Vineis P, Schulte PA. Scientific and ethical aspects of genetic screening of workers for cancer risk: the case of the N-acetyltransferase phenotype. *J Clin Epidemiol* 1995; **48:** 189–97.

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

71 Klerk M, Verhoef P, Clarke R, et al. MTHFR 677C→T polymorphism and risk of coronary heart disease. A meta-analysis. *JAMA* 2002; **288:** 2023–31.

72 Scheuner MT. Genetic evaluation for coronary artery disease. *Genet Med* 2003; *5:* 269–85.

73 Lewis, SJ, Ebrahim S, Davey Smith G. Meta-analysis of MTHFR 677C→T polymorphism and coronary heart disease: does the totality of evidence support causal role for homocysteine and preventive potential of folate? *BMJ* (in press).

74 Dekou V, Whincup P, Papacosta O, et al. The effect of the C677T and A1298C polymorphisms in the methylenetetrahydrofolate reductase gene on homocysteine levels in elderly men and women from the British Regional Heart Study. *Atheroslcerosis* 2001; **154:** 659–66.

75 Ebrahim S, Montaner D, Lawlor DA. Clustering of risk factors and social class in childhood and adulthood in British women's heart and health study: cross sectional analysis. *BMJ* 2004; **328:** 861.

76 Khoury MJ, Yang Q, Gwinn M, Little J, Dana Flanders W. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004; **6:** 38–47.

77 Yoon PW, Scheuner MT, Peterson-Oehlke KL, Gwinn M, Faucett A, Khoury MJ. Can family history be used as a tool for public health and preventive medicine? *Genet Med* 2002; **4:** 304–10.

78 Guttmacher AE, Collins FS, Carmona RH. The family history: more important than ever. *N Engl J Med* 2004; **351:** 2333–36.

79 Workman P. The opportunities and challenges of personalized genome-based molecular therapies for cancer: targets, technologies, and molecular chaperones. *Cancer Chemother Pharmacol*. 2003; **52** (suppl 1)**:** S45–56

80 Kalow W. Pharmacogenetics and personalised medicine. *Fundam Clin Pharmacol* 2002; **16:** 337–42.

81 Oscarson M. Pharmacogenetics of drug metabolising enzymes: importance for personalised medicine. *Clin Chem Lab Med* 2003; **41:** 573–80.

82 Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet* 2003; **4:** 937–47.

83 Kajinami K, Takekoshi N, Brousseau ME, Schaefer EJ. Pharmacogenetics of HMG-CoA reductase inhibitors: exploring the potential for genotype-based individualization of coronary heart disease management. *Atherosclerosis* 2004; **177:** 219–234.

84 Chasman DI, Posada D, Subrahmanyan L, Cook NR, Stanton VP Jr, Ridker PM. Pharmacogenetic study of statin therapy and cholesterol reduction. *JAMA* 2004; **291:** 2821–27.

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

86 Davey Smith G. Reflections on the limitations to epidemiology. *J Clin Epidemiol* 2001; **54:** 325–31.

87 Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Prev Med* 1991; **20:** 47–63.

88 Petitti D. Commentary: Hormone replacement therapy and coronary heart disease—four lessons. *Int J Epidemiol* 2004; **33:** 461–63.

89 Willett WC. Vitamin A and lung cancer. *Nutrition Rev* 1990; **48:** 201–11.

90 Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 1994; **330:** 1029–35.

91 Khaw K-T, Bingham S, Welch A, et al. Relation between plasma ascorbic acid and mortality in men and women in EPIC-Norfolk prospective study: a prospective population study. *Lancet* 2001; **357:** 657–63.

92 Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of antioxidant vitamin supplementation in 20 536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002; **360:** 23–33.

93 Lawlor DA, Ness AR. Commentary: The rough world of nutritional epidemiology—does dietary fibre prevent large bowel cancer? *Int J Epidemiol* 2003; **32:** 239–43.

94 Rimm EB, Stampfer MJ, Ascherio A, Giovannucci E, Colditz GA, Willett WC. Vitamin E consumption and the risk of coronary heart disease in men. *N Engl J Med* 1993; **328:** 1450–56.

95 Shekelle PG, Morton SC, Jungvig LK, et al. Effect of supplemental vitamin E for the prevention and treatment of cardiovascular disease. *J Gen Intern Med* 2004; **19:** 380–89.

96 Mendel G. Experiments in plant hybridization. http://www.mendelweb.org/archive/Mendel.Experiments.txt. (accessed Aug 31, 2005).

97 Correns C. G. Mendel's Regel über das Verhalten der Nachkommenschaft der Bastarde. *Berich Deutsch Botan Gesells* 1900; **8:** 158–68.

98 Gray R, Wheatley K. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant* 1991; **7** (suppl 3)**:** 9–12

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

100 Birge SJ, Keutmann HT, Cuatrecasas P, Whedon GD. Osteoporosis, intestinal lactase deficiency and low dietary calcium intake. *N Engl J Med* 1967; **276:** 445–48.

101 Newcomer AD, Hodgson SF, Douglas MD, Thomas PJ. Lactase deficiency: prevalence in osteoporosis. *Ann Intern Med* 1978; **89:** 218–20.

102 Lower GM, Nilsson T, Nelson CE, et al. N-acetylransferase phenotype and risk in urinary bladder cancer: approaches in molecular epidemiology. *Env Health Perspec* 1979; **29:** 71–79.

103 McGrath J. Hypothesis: is low prenatal vitamin D a risk-modifying factor for schizophrenia? *Schiz Res* 1999; **40:** 173–77.

104 Ames BN. Cancer prevention and diet: help from single nucleotide polymorphisms. *Proc Natl Acad Sci USA* 1999; **96:** 12216–18.

105 Rothman N, Wacholder S, Caporaso NE, Garcia-Closas M, Buetow K, Fraumeni JF. The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta* 2001; **1471:** C1–C10.

106 Brennan P. Gene environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis* 2002; **23:** 381–87.

107 Kelada SN, Eaton DL, Wang SS, Rothman NR, Khoury MJ. The role of genetic polymorphisms in environmental health. *Env Health Perspect* 2003; **111:** 1055–64.

108 Katan MB. Apoliopoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1986; **327:** 507–08.

109 Honkanen R, Pulkkinen P, Järvinen R, et al. Does lactose intolerance predispose to low bone density? A population-based study of perimenopausal Finnish women. *Bone* 1996; **19:** 23–28.

110 Thomas DC, Conti DV. Commentary on the concept of "Mendelian Randomization". *Int J Epidemiol* 2004; **33:** 17–21.

111 Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* 2005; **366:** 1223–34.

112 Davey Smith G, Lawlor D, Harbord R, et al. Association of C-reactive protein with blood pressure and hypertension: lifecourse confounding and Mendelian randomisation tests of causality. *Arterioscler Thromb Vasc Biol* 2005; **25:** 1051–56.

113 Buchsbaum DG, Welsh J, Buchanan RG, Elswick RK Jr. Screening for drinking problems by patient self-report. Even 'safe' levels may indicate a problem. *Arch Intern Med* 1995; **155:** 104–08.

114 Fuller RK, Lee KK, Gordis E. Validity of self-report in alcoholism research: results of a Veterans Administration Cooperative Study. *Alcohol Clin Exp Res* 1998; **12:** 201–05.

115 Lewis S, Davey Smith G. Alcohol, ALDH2 and esophageal cancer: a meta-analysis which illustrates the potentials and limitations of a Mendelian randomisation approach. *Cancer Epidemiol Biomarkers Prev* 2005; **14:** 1967–71.

116 Davey Smith G, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* 2005; **330:** 1076–79.

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

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

119 Eck P, Erichsen HC, Taylor JG, et al. Comparison of the genomic structure and variation in the two human sodium-dependent vitamin C transporters, SLC23A1 and SLC23A2. *Hum Genet* 2004; **115:** 285–94.

120 Waddington CH. Canalization of development and the inheritance of acquired characteristics. *Nature* 1942; **150:** 563–65.

121 Wilkins AS. Canalization: a molecular genetic perspective. *BioEssays* 1997; **19:** 257–62.

122 Hartman JL, Garvik B, Hartwell L. Principles for the buffering of genetic variation. *Science* 2001; **291:** 1001–04.

123 Juul K, Tybjaerg-Hansen A, Marklund S, et al. Genetically reduced antioxidative protection and increased ischaemic heart disease risk: the Copenhagen city heart study. *Circulation* 2004; **109:** 59–65.

124 Silverman EK, Palmer LJ. Case-control association studies for the genetics of complex respiratory diseases. *Am J Respir Cell Mol Biol* 2000; **22:** 645–48.

125 Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; **2:** 91–99.

126 Dahlman I, Eaves IA, Kosoy R, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet* 2002; **30:** 149–50.

127 Goldstein DB, Ahmadi KR, Weale ME, Wood NW. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet* 2003; **19:** 615–22.

128 Terwilliger JD, Goring HH. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000; **72:** 63–132.

129 Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000; **26:** 151–57.

130 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioaniddis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:** 306–09.

131 Tabor HK, Risch NJ, Myers RM. Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; **3:** 391–97.

132 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33:** 177–82.

133 Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405:** 847–56.

134 Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001; **69:** 1357–69.

135 Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. *Nat Rev Genet* 2003; **4:** 937–47.

136 Husebekk A, Iversen O-J, Langmark F, Laerum OD, Ottersen OP, Stoltenberg C. Biobanks for Health - Report and Recommendations from an EU workshop. Oslo: Technical report to EU Commission, 2003.

137 Burton PR, Hansell A. UK Biobank: the expected distribution of incident and prevalent cases of chronic disease and the statistical power of nested case control studies. Manchester, UK: UK Biobank Technical Reports, 2005.

138 Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004; **429:** 475–77.

139 Ferri E, Bynner J, Wadsworth M. Changing Britain, Changing Lives: Three Generations at the End of the Century. London: Institute of Education Bedford Way Papers, 2003.

140 Highfield R. Daily Telegraph (London), Sept 22, 2004: 8.

141 Rothman K, Greenland S. Case-control studies. In: Rothman K, Greenland S, eds. Modern Epidemiology. Second Edition. Philadelphia: Lippincott-Raven, 1998: 93–114.

142 Burton PR, Hansell A. UK Biobank: the expected distribution of incident and prevalent cases of chronic disease and the statistical power of nested casecontrol studies. Manchester, UK: UK Biobank Technical Reports, 2005.

143 Centre for Longitudinal Studies. National Child Development Study; http://www.cls.ioe.ac.uk/studies.asp?section=000100020003 (accessed Aug 30, 2005).

144 Clarke R, Breeze E, Sherliker P, et al. Design, objectives, and lessons from a pilot 25 year follow up re-survey of survivors in the Whitehall study of London Civil Servants. *J Epidemiol Community Health* 1998; **52:** 364–69.

145 Collins FS. The case for a US prospective cohort study of genes and environment. *Nature* 2004; **429:** 475–77.

146 Roses AD. Pharmacogenetics. *Hum Mol Genet* 2001; **10:** 2261–67.

147 Gormally E, Hainaut P, Caboux E, et al. Amount of DNA in plasma and cancer risk: a prospective study. *Int J Cancer* 2004; **111:** 746–49.

148 Khoury MJ, Millikan R, Little J, Gwinn M. The emergence of epidemiology in the genomics age. *Int J Epidemiol* 2004; **33:** 936–44.

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

150 Khoury MJ, McCabe LL, McCabe ER. Population screening in the age of genomic medicine. *N Engl J Med* 2003; **348:** 50–58.

151 Merikangas KR, Risch N. Genomic priorities and public health. *Science* 2003; **302:** 599–601.

152 Shostak S. Locating gene-environment interaction: at the intersections of genetics and public health. *Soc Sci Med* 2003; **56:** 2327–42.

153 Israel E, Chinchilli VM, Ford JG, et al. Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet* 2004; **364:** 1505–12.

154 Rosenberg RN. Translating biomedical research to the bedside. *JAMA* 2003; **289:** 1305–06.

155 Khoury MJ, Millikan R, Little J, Gwinn M. The emergence of epidemiology in the genomics age. *Int J Epidemiol* 2004; **33:** 936–44.

156 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273:** 1516–17.