Key concepts in genetic epidemiology

Paul R Burton, Martin D Tobin, John L Hopper

This article is the first in a series of seven that will provide an overview of central concepts and topical issues in modern genetic epidemiology. In this article, we provide an overall framework for investigating the role of familial factors, especially genetic determinants, in the causation of complex diseases such as diabetes. The discrete steps of the framework to be outlined integrate the biological science underlying modern genetics and the population science underpinning mainstream epidemiology. In keeping with the broad readership of *The Lancet* and the diverse background of today’s genetic epidemiologists, we provide introductory sections to equip readers with basic concepts and vocabulary. We anticipate that, depending on their professional background and specialist knowledge, some readers will wish to skip some of this article.

**What is genetic epidemiology?**
Epidemiology is usually defined as “the study of the distribution, determinants [and control] of health-related states and events in populations”,1 By contrast, genetic epidemiology means different things to different people.2,3 We regard it as a discipline closely allied to traditional epidemiology that focuses on the familial, and in particular genetic, determinants of disease and the joint effects of genes and non-genetic determinants. Crucially, appropriate account is taken of the biology that underlies the action of genes and the known mechanisms of inheritance. The word “appropriate” is crucial because the manner in which biology is taken into account varies from setting to setting and depends on the genetic information available. With advances in technology and biological knowledge, the work undertaken by those who investigate the health consequences of genetic variants continues to evolve.

Before information about DNA became available, scientists trying to relate genetic variation to disease relied on the fact that the mendelian laws of inheritance implied a biological model for the way in which a putatively causative genetic variant might lead to disease (eg, two abnormal copies of gene G are required to cause disease D). Aetiological inferences could be drawn from the distribution of disease and phenotypic aggregation within large families or across groups of families (segmentation analysis; see below). In time, more became known about the human genome, and especially about genetic markers, although they are not necessarily considered responsible for determining health or disease. By incorporating the biology of gamete formation and chromosomal recombination into a mathematical model of the extent to which a given marker tends to be transmitted through a family in conjunction with a disease, we can estimate whether a causative genetic variant is likely to lie close to that marker and, if so, how close. The marker and the causative variant need not be within the same gene. This principle is the basis of genetic linkage analysis (see a later paper in this series12), which has achieved many of the breakthroughs in the genetics of disease causation. Many such breakthroughs involve conditions caused by variants in a single gene and have been achieved by geneticists and clinical geneticists who would not view themselves as genetic epidemiologists. Nevertheless, linkage analysis is one of the most important tools available to the genetic epidemiologist.

Extensive information about the human genome can now be included in genetic epidemiology studies. Once it is known which two versions of a potentially causative gene an individual possesses, looking for an association between variants in that gene and the disease of interest is fundamentally no different from an exploration of a disease-exposure association in traditional epidemiology. There is often no need to take particular note of the underlying biological model, but this does not mean that genetic epidemiologists can ignore biology. A recurring theme of this series is that knowledge about the underlying biology, coupled with the inferential tools of modern epidemiology and biostatistics, allows important aetiological questions to be answered in ways that are more rigorous, and often more powerful, than approaches that fail to make best use of both the epidemiology and the genetics.

Although many of the greatest successes have been with monogenic disorders,11 where familial recurrence seems to follow the laws of mendelian inheritance,11,14 genetic epidemiology today is increasingly focusing on complex diseases such as diabetes mellitus, ischaemic heart disease, asthma, and cancer,11,15–20 which are characteristically caused by several interacting genetic and environmental determinants.11,15–20 This series aims to illustrate the challenges that genetic epidemiologists face and the methods they use in their collaborative work with other scientists.

We provide a framework for investigating the role of genetic variation in complex diseases. Such a daunting
investigation can be broken down into manageable steps (figure 1). Figure 1 represents the template around which the discussion in this article has logically been structured. It is not a prescriptive statement about how such research should be conducted. Genetic epidemiological research does not have to be done this way: historical evidence, ease of recruiting study populations, and decreasing cost of genotyping are just some of the reasons why one or more steps may be omitted or taken in a different order. However, a proper understanding of the logical basis of each step helps to decide when short cuts are reasonable.

**Genetics for genetic epidemiology**

The role of the underlying biological model in our definition of genetic epidemiology means that some understanding of basic genetics is required. Those familiar with human genetics may wish to skip this section.

**DNA, RNA, and proteins**

The human genome is made up of DNA, which consists of a long sequence of nucleotide bases of four types: adenine (A), cytosine (C), guanine (G), and thymine (T). Strong covalent bonds bind bases together along a single strand, and weaker hydrogen bonds pair A with T and C with G between the two strands. Each single strand has two different ends. Under native conditions, in the nucleus of a cell, DNA is double stranded (figure 2). Double-stranded DNA is replicated by breakage of the two strands and construction of a new complementary strand for each, resulting in two identical copies of the original. A single strand of DNA can also act as a template for a complementary strand of RNA. This transcription RNA is similar to DNA, but T is replaced by U (uracil). Crucially, in certain regions of the DNA, which can be called genes, transcribed RNA encodes instructions that tell the cell how to assemble aminoacids to make proteins. Most genes contain alternating regions called exons and introns. The RNA that is transcribed is complementary to the whole gene (exons and introns). Mature mRNA is then created by post-transcriptional processing, which cuts out the introns and splices the exonic elements to produce mRNA, which codes for a protein. The production of protein via mRNA is called translation. It is mainly through altered protein
function that changes in the DNA sequence affect health and disease.

Human genome and variation in DNA sequence
The complete DNA sequence is the human genome, and the repertoire of proteins is the proteome. The haploid genome is about 3·3 billion bp. Some 3% of the genome consists of coding sequences, and there are 30 000–40 000 protein-coding genes. 99·9% of the genome is non-coding, consisting of microsatellites and single nucleotide polymorphisms (SNPs). Alleles are differentiated by the number of repeats (eg, CA repeats are 12 CA repeats in a row). Microsatellites are highly variable and most people are heterozygous at any given locus. Coding regions tend not to contain microsatellite sequences. SNPs, by contrast, represent variation in a single nucleotide. As of July, 2005, the number of known SNPs (with a unique position) in the human genome exceeded 10 million, and more than half these had been independently validated. Although individual SNPs might carry limited information, their ease of typing and large number means that they are widely used in genetic epidemiology.

SNPs in protein-coding regions are non-synonymous or synonymous, depending on whether they do or do not modify the aminoacid sequence in the gene product. Non-synonymous SNPs can also be called coding SNPs. Intronic and intergenic SNPs lie in the non-coding regions. A non-synonymous SNP in a coding sequence is generally more likely than other classes of SNP to affect the function or availability of a protein. However, all types of SNP can cause disease, for example by altering the regulation of transcription of a critical protein. The true distribution of disease-associated variants between non-coding and coding sequences is unknown.

Chromosomes, gamete formation, and recombination
The human genome is distributed among 46 chromosomes, 22 homologous pairs of autosomes, and one pair of sex chromosomes. The complete set is the diploid complement. One chromosome in each of the 22 homologous pairs is derived from the mother and one from the father, and the two homologues will have the same sequence of genes in the same positions, but they will usually exhibit sequence variations at several loci and can therefore be distinguished.

The division and accompanying replication and partitioning of DNA that leads to the formation of sperm and ova is meiosis. Each gamete receives (at random) one member of each homologous chromosomal pair.

It might seem that there is a 50% probability that any given gamete receives one chromosome rather than the other from a particular homologous pair, and that there are 2 to the power of 23 distinct gametes that any given individual might produce. Crucially, however, this is not the whole story. At gamete formation, the choice is not between the whole of one chromosome or the whole of the other. Instead, the gamete receives a mixture of two homologous chromosomes because of crossover events (figure 3). Crossovers can split alleles that lie together on a common parental chromosome and can result in alleles that originally came from different grandparents being on the same chromosome.

Gene mapping makes use of recombination. The further apart two genes are, the higher the probability of an odd number of crossovers (odd numbers cause recombination), to a maximum of 50%. The recombination fraction (the proportion of meioses that result in a recombination) is an indication of how far apart two genes are. This fraction can be mathematically

Across the human genome there could be 10 million common SNPs.

Allele
If the DNA sequence at a given locus (often a gene or a marker) varies between different chromosomes in the population, each different version is an allele. If there are two alleles at a given locus, the allele that is less common in the population is the minor allele.
**Locus**
A locus is a unique chromosomal location defining the position of an individual gene or DNA sequence. In genetic linkage studies, the term can also refer to a region involving one or more genes, perhaps including non-coding parts of the DNA.

**Non-synonymous SNPs**
An SNP that alters the DNA sequence in a coding region such that the amino acid coding is changed. The new code specifies an alternative amino acid or changes the code for an amino acid to that for a stop translation signal or vice versa. Synonymous SNPs alter the DNA sequence but do not change the protein coding sequence as interpreted at translation, because of redundancy in the genetic code; several different codons can specify the same amino acid. Non-synonymous SNPs can also be called coding SNPs.

**Diploid**
Most human cells are diploid, containing all 46 chromosomes: one copy of both members of each homologous pair (eg, two versions of chromosome 14). The full diploid human chromosome complement can be expressed as 46,XX in a woman and 46,XY in a man.

**Centimorgans**
1 centimorgon (cm) corresponds to a region within which a crossover is expected once every 100 meioses. This implies a 1% chance of a single crossover at a single meiosis, and because the probability of a double crossover is exceedingly small (about 0.01%), this also corresponds to a chance of roughly 1% of recombination at each meiosis.

**Haplotype**
A series of alleles at linked loci along a single chromosome.

**Phase**
Denotes the haplotypic configuration of linked loci. The diplotype U1,U3–V1,V2 is consistent with two possible phases: (1) U1,V1–U3,V2, on one chromosome and U1,V2–U3,V1, on the other; or (2) U1,V2–U3,V1, on one chromosome and U1,V1–U3,V2, on the other. If a child receives U1,V1 from one parent and U3,V2 from the other, phase information is unobtainable.

**Genotypes, haplotypes, and phenotypes**
Although the genotype is sometimes used to refer to the overall genetic constitution of an individual, genetic epidemiologists use the term to refer to a particular locus. If three loci—U, V, and W—are given, a genotype is a named set of alleles at those loci. The full haplotype along a single chromosome is called a haplotype and the haplotypes do differ between these two scenarios. The haplotype information in a parent is also known as the phase of that parent’s meioses. Throughout this series, phenotype will be used interchangeably with trait to refer to a measurable characteristic of an individual that is not itself a genotype. This definition includes binary disease states (presence or absence of asthma) and quantitative characteristics (systolic blood pressure). Some simple binary phenotypes are present (or expressed) if there are two copies of an abnormal allele, in which case the allele is recessive. If an abnormal phenotype can be expressed in full with just one copy, the abnormal allele is dominant. An intermediate state often exists (penetrance). If penetrance in a heterozygote lies between the penetrance of the two corresponding homozygotes, this gene is codominant. If expression depends on age, penetrance can be modelled in terms of differing distributions of the age-at-onset by genotype. These concepts all extend to traits defined on fully quantitative or ordinal scales.

**Fusion of genetics and epidemiology**
The fusion of epidemiology and genetics provides the foundation for genetic epidemiology. We focus on assessment of indirect evidence for a genetic contribution to disease causation through the study of familial aggregation and segregation analysis, because these topics are not covered in detail elsewhere in the series.

**Phenotypic aggregation within families**
It is important to distinguish between the clinical sense of familial clustering (extended families that happen to have multiple cases of a disease or syndrome of interest) and the epidemiological sense of familial aggregation (there is, on average, a greater frequency of disease in close relatives of individuals with the disease than in relatives of individuals without the disease). Simple analyses of familial aggregation treat the family like any other unit of clustering. In addressing whether there is phenotypic aggregation within families, no attempt is made to determine the cause of any aggregation.

**Binary traits**
If the phenotype is a binary trait, familial aggregation is often assessed by the recurrence risk ratio or allied measure. The pattern of recurrence risk ratios across different types of relatives can provide valuable information about the origin of a binary trait, and can inform the statistical power of linkage studies. The recurrence risk ratio is a ratio of prevalences—the proportion of a population that has a [particular] disease at a specific point in time. The recurrence risk ratio (λ1) in relatives of type R is the prevalence of the disease in relatives of type R of affected cases (P(R)) divided by the prevalence in the general population (P). If the relatives are siblings, λ1 and P(R) would be used. P and P(R) will almost always be estimates, so χ will be an estimate too.

Prevalence is difficult to estimate. First, the disease (phenotype) must be assessed carefully, taking into account issues such as disease definition, age at onset and duration. Second, the study sample must be representative of the target population, to avoid systematic overrecruitment or underrecruitment of those with disease. It can be necessary to invest substantial resources to ensure a high response rate to guard against such biases.

In genetic epidemiology, as in mainstream epidemiology, it is often difficult to obtain a representative or random sample of the general population that is large enough to ensure adequate statistical power. Consequently, families are often recruited precisely because they have affected members. This outcome-based sampling is often more informative and increases power. Furthermore, it has obvious benefits for a study aimed at estimating λ1, the prevalence of disease in a particular subgroup of relatives. However, because the familial determinants of the trait of interest are usually unobserved in a study of familial aggregation, this sampling method can lead to severe ascertainment bias. Furthermore, the data to estimate P(R) can come in many different forms. The consequences of non-random sampling must be considered carefully, and any ascertainment bias should be dealt with in the analysis. If necessary, expert advice should be sought. These are not trivial issues. The same concerns apply equally well to other measures of familial aggregation and to the investigation of the pattern of aggregation within families.

Three interpretational issues warrant emphasis. First, the prevalence of many complex diseases increases...
steady with age, whereas $\lambda$ often declines.\textsuperscript{46} Careful attention must therefore be paid to the age distributions of both the general population sample and the relatives and, at the very least, adjustments must be made for any differences between the two. Second, if a phenotype is common (eg, $P=0.5$, as it roughly is for some measures of skin-prick sensitivity to common allergens\textsuperscript{46}), $\lambda_s$ cannot be greater than 2-0, even if every available relative is affected. Comparisons of $\lambda_s$ across different diseases or different settings thus require care. Third, $\lambda_s$ measures the combined effect of all causes of familial aggregation, not just the effect of genes. In some settings (and in a later paper in this series\textsuperscript{48}), the term familial relative risk is used instead of $\lambda$\textsuperscript{46}.  

**Quantitative traits** 
Assessment of familial aggregation of a continuous trait, such as (untreated) blood pressure, is most commonly undertaken with a correlation or covariance-based measure such as the intrafamily correlation coefficient (ICC). This approach dates back more than a century to Galton\textsuperscript{49,50} and Pearson.\textsuperscript{51} The ICC indicates the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families.\textsuperscript{52} Thus, the assessment of aggregation of a continuous measure in genetic epidemiology is fundamentally no different from, and could be viewed as predating\textsuperscript{10,50} analogous problems in traditional epidemiology and social science.\textsuperscript{52,53} Consequently, techniques such as linear regression and multilevel modelling analysis of variance\textsuperscript{25-29} can be converted directly into genetic epidemiology. As with the binary phenotype, non-random ascertainment can seriously bias an ICC.\textsuperscript{42}

**Interpretation** 
For many complex diseases, the average $\lambda_s$ in first-degree relatives is around 2.\textsuperscript{65} It tends to be greater the younger the age at onset in the affected individual,\textsuperscript{66} to fall as the familial relationship becomes more distant,\textsuperscript{20} and to increase as the number of affected relatives of the at-risk individual rises. Although a $\lambda_s$ of 2 might appear modest, it does suggest that uncovering all sources of familial aggregation might well be worthwhile. A moderate $\lambda_s$ generally implies the presence of underlying familial risk factors (genetic or non-genetic) that are at least an order of magnitude stronger than $\lambda_s$ itself.\textsuperscript{25,60} This effect strengthens with the rarity of the determinant in question. For example, dominant alleles in the BRCA1 gene affect about 1 in 500 women, and result in a ten to 20-fold increase in the risk of breast cancer. But this increase only slightly raises the risk of disease in first-degree relatives across the population ($\lambda_s$ is about 1-1). A value of $\lambda_s$ of around 2 would also be consistent with a number of common alleles each associated with a more modest relative risk. Knowing $\lambda_s$ alone does not tell us which genetic or familial model is most likely. Because a simple assessment of familial aggregation takes no account of the underlying biology, one should not assume that evidence of familial aggregation implies genetic effects. For many complex diseases, the non-genetic risk factors identified to date have a modest effect and are weakly correlated in relatives. They therefore seem to explain little familial aggregation. For example, known risk factors such as parity, age at menarche, age at menopause, and body-mass index explain less than 5% of the enhanced risk of breast cancer in first-degree relatives of affected people.\textsuperscript{65} But such determinants are probably just surrogates for aetiologically stronger factors that are as yet beyond the reach of epidemiology. They are typically measured by questionnaire and can be subject to substantial measurement error. Such errors attenuate both their apparent effect on risk and their estimated correlation between relatives. Consequently, the non-genetic contribution to familial aggregation might be greatly understated: this point is often overlooked.

**Explanation for the pattern of aggregation** 

**Variance components modelling** 
To estimate the extent to which any familial aggregation identified is caused by genes, we need a biologically rational model that specifies how a phenotype of interest might be modulated by the effect of one or more genes. One of the most common is the additive genetic effects model (panel 1).\textsuperscript{10,41-43} The model needs to incorporate some measure of the extent to which different classes of relatives have different probabilities of sharing alleles that are identical by descent (panel 2). With both these elements it is possible to quantify, by hierarchical variance components modelling for example,\textsuperscript{15,77,64} the extent to which genetic variability might be consistent

---

**Panel 1: Additive genetic effects** 
One of the simplest paradigms for the effect of genes on a continuous complex trait is the additive genetic effects model.\textsuperscript{30,34,41,42,43,65} There are assumed to be an unspecified number of genes that influence the trait, each with an unspecified number of alleles. The model implies that a given allele at a given locus adds a constant to, or subtracts from, the expected value of the trait. The amount added or subtracted varies in an unknown way from allele to allele and from locus to locus. For example, suppose gene G had four alleles: $G_1$ adds 3 to the trait; $G_2$ adds 6; $G_3$ subtracts 2; and $G_4$ adds 1. The contribution of $G$ to the expected value of the trait in an individual who is, for example, $G_1G_2$ is +9. The effect that any one allele exerts is assumed to be the same regardless of which allele it is paired with. Unless there is a marked departure from this assumption (eg, $G_1$ adds 6 if paired with $G_1$ but subtracts 3 if paired with $G_2$) the additive model will usually capture much of the aetiological information that can reasonably be explained by genes.
with the familial patterns of variability in the phenotype. Other genetic and non-genetic models might also be consistent with the data, so a good fit of any one model does not prove that that model is right.

This approach can be extended to include the covariance or correlation patterns (or both) that would be expected for other more complex models of genetic determination; for example, by including genetic dominance (see a later paper in this series) in addition to additive genetic effects. One can also allow for correlation or covariance patterns due to unmeasured environmental determinants that are shared by a whole family, those that are shared just by siblings, and those which wax and wane as individuals spend time living together or living apart.

Finally, many environmental and lifestyle exposures are unique to an individual. These unshared determinants contribute nothing to the tendency for relatives to be more similar than non-relatives (ie, they do not contribute to the covariance between relatives), but they do affect the total variability of a quantitative trait. Many methodological developments in this area come from work on the analysis of twin studies.

Crucially—and this point is often misunderstood—variance components analyses require no information about genotypes or measured environmental determinants. No blood needs to be taken for DNA analysis. However, if information is available about specific genes and environmental determinants, it can be added to the analysis. Panel 3 gives pointers to types of variance component modelling most commonly used in genetic epidemiology.

## Heritability

One of the principal reasons for fitting a variance components model is to estimate the variance attributable to additive genetic effects. This quantity \( S^2_A \) represents that component of the total phenotypic variance \( S^2_T \), usually after adjustment for measured genetic and non-genetic determinants, that can be attributed to unmeasured additive genetic effects (panel 1). Heritability in the narrow sense is defined as \( S^2_A \) divided by \( S^2_T \). Particular family studies, especially those including monozygous twins, also allow estimation of \( S^2_C \), the phenotypic variance attributable to all genetic effects, including non-additive effects at individual loci and between loci (see a later paper in this series). Heritability in the broad sense is defined as \( S^2_e \) divided by \( S^2_T \).

Heritability is a beguiling concept but is open to misinterpretation. It is not about cause in itself, but about the causes of variation in a particular trait in a particular population at a particular time. Fisher pointed out that although the numerator has a simple genetic meaning, the “hotch-potch of a denominator” does not. Fisher conflates the variance attributable to genes and to shared environment and residual variance attributable to unshared and unmeasured determinants and to measurement error. In consequence, heritability for a given phenotype can vary quite substantially from setting to setting, and even within a given setting.

Heritability is formally defined for quantitative traits. For binary traits, it is usually calculated by invoking a hypothetical construct known as liability, and applying a version of variance components modelling. Liability is an underlying, unobservable, normally-distributed trait that is assumed to determine the probability that an individual develops the disease of interest. Unfortunately, with a binary phenotype, the heritability of the liability does not have a clear meaning and is prone to confused interpretation.

Some scientists and the media treat heritability as meaning the extent to which a trait is caused by genetic factors. This view is incorrect. If a trait is dependent upon a particular allele for which everybody is homozygous, variation at that locus will play no part in determining the

<table>
<thead>
<tr>
<th>IBD sharing at a single locus</th>
<th>Parents</th>
<th>Parent–child</th>
<th>Full siblings</th>
<th>Grandparent–grandchild</th>
<th>Uncle–niece</th>
<th>First cousins</th>
<th>Half siblings</th>
<th>Identical twins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected probability 2 alleles shared IBD</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Expected probability 1 allele shared IBD</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Expected probability 0 alleles shared IBD</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table: Characteristic IBD sharing for different categories of relative on the assumption that parents are unrelated
variance of the trait, and will not contribute to heritability. A near-ubiquitous environmental exposure will also make little or no contribution to the denominator, \( S_1 \). Interpretation also depends on which covariates are included. For example, including an important environmental covariate might well decrease \( S_1 \), but leave \( S_2 \) unchanged, which will apparently increase the heritability in the narrow sense. For these reasons, it is often preferable to quote the magnitude of the variance components (such as \( S_2 \)) individually. 60,78,84

If there are so many pitfalls in the interpretation of heritability, why calculate it? The power of most studies to discover genes is positively associated with the heritability of the trait of interest; so, all else being equal and if the option exists, analytical efficiency can be enhanced by selecting a study population in which the heritability of the trait of interest is believed to be high. Furthermore, subject to all the caveats, knowledge that a trait of interest has high heritability can support a study that proposes to investigate the genetic determinants of that trait. Equally, if heritability is low, those contemplating doing or funding the study are forewarned that genetic effects might be difficult to find. In either case, interpretation demands expert understanding of the nature of the trait.

Justification for expensive studies

Is there evidence of one or a few genes with substantial enough effect to justify expensive studies? This question falls under the scope of segregation analysis. 73,83,86 Are there one or more major genes (ie, genetic variants that have a strong effect on susceptibility, however rare they may be) whose mendelian segregation within families explains all or part of the observed familial aggregation of the trait of interest? This information may be useful in its own right, 87 and it could also be used to generate estimates for a parametric linkage analysis 87 (see a later paper in this series). Elston\textsuperscript{88} defines segregation analysis as: “the statistical methodology used to determine from family data the mode of inheritance of a particular phenotype, especially with a view to elucidating [major] gene effects”. Although computationally demanding, it is now possible to fit models (to estimate allele frequencies and risk functions) that include more than one mode of inheritance, providing the family structures have sufficient information (eg, Cui and colleagues’ work on breast cancer genetics\textsuperscript{89}). Like variance components analysis, classical segregation analysis has no requirement for observed genotypes. It can be viewed as a special case of the investigation of familial aggregation, often focusing on the pattern of aggregation within individual families rather than averaging across the population. The results of a segregation analysis can be very sensitive to inappropriate adjustment for ascertainment. 89

How substantial the effect of major genes must be before they are deemed worthy of biological investigation depends on many factors. These include the prevalence of the deleterious variant(s), the prevalence and natural history of the disease they might cause, and the strengths of other genetic and environmental influences on the same disease. Furthermore, account can also be taken of the potential usefulness of information about the cause of disease that might come from identifying a particular genetic variant as being related to the disease. These important issues will be discussed in a later paper in this series. 90 Whether a particular segregation analysis can detect a major gene effect or not also depends on other factors, including the quantity and quality of the family data that are available. In light of all of these uncertainties, it seems irrational not to progress with further investigation of a putative gene effect simply because a segregation analysis has failed to provide evidence for a major gene (figure 1). Segregation analyses have been used less often since the revolution in DNA technology. This decline is partly due to concurrent increases in computational power so
A functional gene (D) which affects a binary disease trait lies 0-01 cM away from a known marker. Suppose that, 2000 generations ago, a new, deleterious mutation (D\*) appeared in a single individual on a chromosome that happened to carry the allele M₁ at the marker. Any individual who carries D\* today will have inherited the relevant part of the original disease-bearing chromosome via an inheritance pathway that will have involved 2000 meioses. For the given distance between marker and disease gene, the probability of a crossover at any one meiosis will be 0·0001, and the probability of no crossovers in any of the 2000 meioses will be (1−0·0001)^2000 (ie, 0·82). This could well allow detection of a population-wide association between the disease and M₁, even though M₁ has nothing to do with the cause of the disease. This is linkage disequilibrium (see Cordell and Clayton\textsuperscript{64}).

Linkage disequilibrium implies linkage that is so tight that it leads to an association at the population level, unlike simple linkage where the two loci tend to be further apart and the chance of recombination at any single meiosis is greater. Here, a disease-causing variant might be closely associated with marker allele M₁ in one family but equally closely with M₈ in another. The within-family associations over a few generations are strong and consistent, but there is no systematic association across the population as a whole.

that one can now handle complex parametric linkage models (see a later paper in this series\textsuperscript{12}). Furthermore, linkage analyses for complex diseases are now often based on non-parametric methods (see a later paper in this series\textsuperscript{12}) so that parameter estimates from segregation analyses are no longer needed. Segregation analysis might come back into favour when the more common major genes are identified, to inform strategies for detection of secondary genetic determinants of disease.

**Location of a causative gene**

Having obtained evidence of a likely genetic component in the cause of a complex disease (without genotyping genes), the next step is to locate and identify any causative genes. One option is to move straight to the obvious candidates (see section on association analysis), but for most complex diseases there are so many candidates and so many genes whose usual effects are completely unknown (let alone their effects when they carry sequence variants) that candidate gene work is often preceded or accompanied by an attempt to localise regions of the genome that are aetiologically relevant.

Major genes for monogenic conditions have been located by linkage analysis,\textsuperscript{13} but there have been far fewer successes with complex diseases.\textsuperscript{12,13} This is mainly because of limitations to statistical power (see a later paper in this series\textsuperscript{12}). For example, most true effect sizes tend to be small when averaged across the population, complex phenotypes are often multidimensional and subject to substantial measurement error, there is marked aetiological heterogeneity, and the measurable predictor variables might not be strongly associated with the actual causative agent(s).

Genetic linkage analysis\textsuperscript{42} is perhaps the best example of a common investigative approach in genetic epidemiology that derives almost entirely from a consideration of the underlying genetics. There is no precise analogue in traditional behavioural and environmental epidemiology, although there are parallels in other specialised fields of epidemiology that must also incorporate a biological model, for example in infectious disease epidemiology.

Genetic linkage analysis\textsuperscript{83,94–96} relies entirely on the tendency for shorter haplotypes to be passed on to the next generation intact, without recombination events at meiosis. If a marker can be identified that is passed down through a family such that it consistently accompanies the disease of interest, this suggests a gene with a functional effect that is located close to that marker.

This focus on the underlying biology should not obscure the importance of clinical knowledge accrued over many years: identification of familial syndromes has been crucial in the success of linkage studies of complex disease. The reason is that one is often attempting to reduce the complex disease to one of its monogenic forms. An example is the syndrome of bowel cancer that led to the identification of the cancer-predisposing role of mutations in DNA mismatch repair genes.\textsuperscript{97–99} This disease was historically referred to as hereditary non-polyposis colorectal cancer or Lynch syndrome.\textsuperscript{100} Other examples are familial breast-ovary syndrome (BRCA1)\textsuperscript{101} and the female and male breast cancer syndrome (BRCA2).\textsuperscript{102}

**Association analysis**

Traditional epidemiology often asks whether it can be proved that, across a study population as a whole, measured environmental exposure E is consistently associated with observed disease D. Association analysis in genetic epidemiology asks the same question of genetic exposures. This approach can be seen as traditional epidemiology applied to genotypes or alleles across a population (panel 4), and many of the analytical approaches used in epidemiology and medical statistics can be applied directly to association analyses in genetic epidemiology. These include univariate methods and regression analysis.\textsuperscript{13,111–113} Furthermore, the approaches outlined above and in panel 3 can be extended to deal with data that have a complex correlation structure including: family data; longitudinal data; data naturally subject to geographical or temporal clustering; and/or data collected under a multistage sampling scheme and applied to phenotypes in various classes, including binary traits, continuous normally distributed traits, and time to event (survival time).\textsuperscript{14,15,75}

Association analysis is covered in later papers in this series,\textsuperscript{14,102,103} so we will limit ourselves to a few comments. A test of association can be informative even when based on genetic variants that are not functional. It can also be useful to detect linkage disequilibrium (panel 5) between a disease and a non-functional marker.\textsuperscript{2,97} An association analysis based on a putative functional genetic variant can be called direct and one based on linkage disequilibrium with a marker indirect.\textsuperscript{14,102,103} Indirect association analysis allows finer mapping than conventional linkage analysis.
The International HapMap Project seeks to map out regions of linkage disequilibrium and “develop a haplotype map of the human genome.”109 One exciting opportunity is the potential for whole genome scans based on indirect association rather than linkage analysis; however, there are still many challenges.20

A potential problem for association studies using unrelated cases and controls is ethnic stratification, which can mimic the signal of association and lead to more false positive results or to missed real effects.107,110,111 This problem has been put forward as one explanation for the repeated failure to replicate positive findings in genetic epidemiology.112,113 The effect of population stratification on the results of association analyses are potentially more severe when small effects are studied in very large studies.111 This result has important implications for national biobanks and large case-control initiatives. This concern is the subject of much debate and study at a national level in the UK.111–115

Addressing population stratification demands an understanding of both the underlying biology and the relevant epidemiology.99,111,116–119 Approaches to dealing with such stratification will be discussed in detail later in the series.64

Gene expression and gene product function
The identification of genes that might be implicated in complex diseases only partly explains the biological pathways that lead to disease. The fuller picture requires knowledge of gene expression and gene product function, and the place of DNA, RNA, and proteins in the living environment of an integrated organism. Such research is underway, but the issues are very different from those that form the primary focus of this series. Its importance is acknowledged by step 6 in figure 1, and some of the issues will be touched upon in a later paper.106

Where do we go from here?
For reasons mainly of statistical power and recruitment of large samples, genetic epidemiology is moving away from linkage studies based on families to allelic association studies based on unrelated individuals.20,21 This move is not without its critics,120 and a later paper in this series121 will look at the future role of population-based family studies and the need to ensure that important opportunities are not missed.115

One serious problem facing mainstream epidemiology is that residual confounding by unobserved covariates could be strong enough to swamp the small aetiological effects now being sought.112,113 The distribution of alleles at any given locus tends not to be correlated either with environmental exposures or with the distribution of alleles at other loci (except those few in tight linkage disequilibrium). Therefore, the biology underpinning genetic epidemiology offers a potentially useful way to study environmental determinants in disease without residual confounding. This approach, often called mendelian randomisation,124–128 will be considered in a later paper.115

Studies involving at least 5000 cases are now being discussed as an essential element of biomedical research. Such research will involve huge national and international investment and incur important opportunity costs. As a result, scientific debate, particularly about study design, can be heated. Even within the contributors to this Lancet series there is disagreement about key issues such as the role of large national cohort studies.126,127 This important debate will be covered in a later paper in this series.95

Contributors
P R Burton and M D Tobin are joint first authors. M Tobin was funded by a Medical Research Council (MRC) research training fellowship in health services and health of the public research. The methodological research programme in genetic epidemiology at the University of Leicester is supported in part by MRC collaborative grant # C980678, programme grant # 00/1209 from the National Health and Medical Research Council (NHMRC) of Australia, and by Leverhulme research interchange grant # F/07134/K. J Hopper is supported by fellowship grant #299953, and is a group leader of the Victorian Breast Cancer Research Consortium. None of these funding bodies had any input into the writing of this review.

Conflict of interest statement
We declare that we have no conflict of interest.

References