Genetic Epidemiology 1

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.

Lancet 2005; 366: 941–51

See Comment page 880

This is the first in a Series of seven papers on genetic epidemiology.

Department of Health Sciences and Department of Genetics, University of Leicester, Leicester, UK
(Prof P R Burton MD, M D Tobin PhD); **and Centre for Genetic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia**
(Prof J L Hopper PhD)

Correspondence to:
Prof Paul R Burton, Department of Health Sciences, University of Leicester, 22–28 Princess Road West, Leicester LE1 6TP, UK
pb51@le.ac.uk

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.[2–7] 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[8–11] implied a biological model for the pattern of sharing of genes between close relatives. If knowledge of this pattern could be supplemented by an assumed 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 (segregation 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 series[12]), 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,[13] 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,[13,15–20] which are characteristically caused by several interacting genetic and environmental determinants.[14,21] 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

Genetic marker
A genetic marker is a variable DNA sequence that has a non-variable component that is sufficiently specific to localise it to a single genomic locus and a variable component that is sufficiently heterogeneous to identify differences between individuals and between homologous chromosomes in an individual. Genetic markers have a pivotal role in gene mapping. Sequence variations at genetic markers are not usually functional.

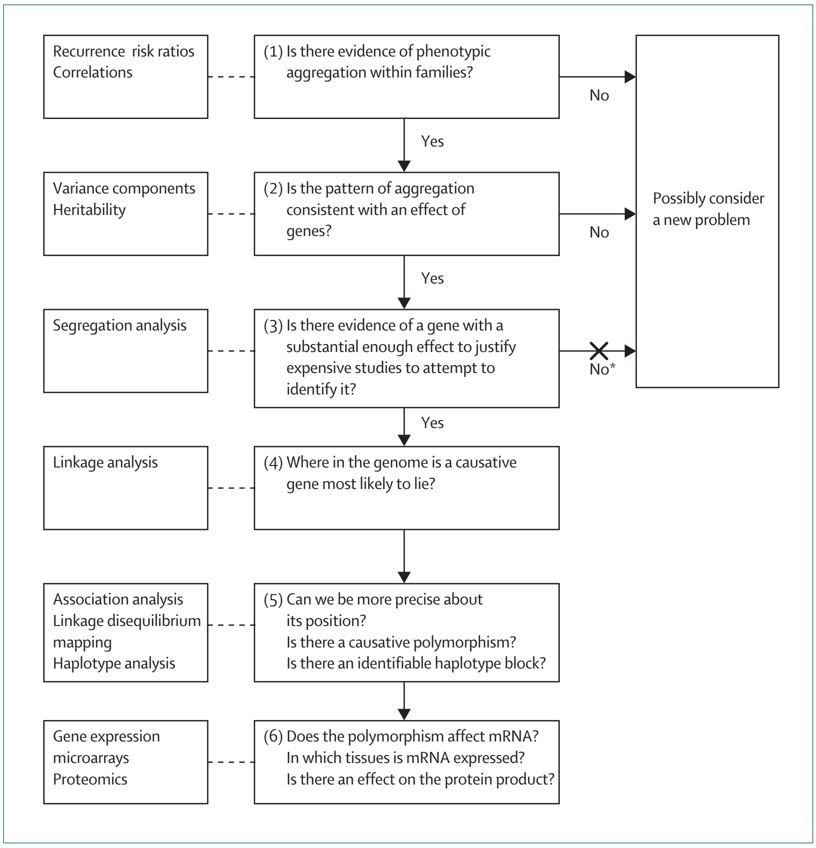
Recurrence risk ratios
Correlations
(1) Is there evidence of phenotypic aggregation within families?
No
Yes
Variance components
Heritability
(2) Is the pattern of aggregation consistent with an effect of genes?
No
Yes
Segregation analysis
(3) Is there evidence of a gene with a substantial enough effect to justify expensive studies to attempt to identify it?
No*
Yes
Possibly consider a new problem
Linkage analysis
(4) Where in the genome is a causative gene most likely to lie?
Association analysis
Linkage disequilibrium mapping
Haplotype analysis
(5) Can we be more precise about its position?
Is there a causative polymorphism?
Is there an identifiable haplotype block?
Gene expression microarrays
Proteomics
(6) Does the polymorphism affect mRNA?
In which tissues is mRNA expressed?
Is there an effect on the protein product?

Figure 1: **Framework outlining systematic approach to identification and characterisation of genetic determinants of complex disease**

*It is probably illogical to stop trying to identify genetic determinants of disease simply because segregation analysis fails to provide significant evidence of major gene.

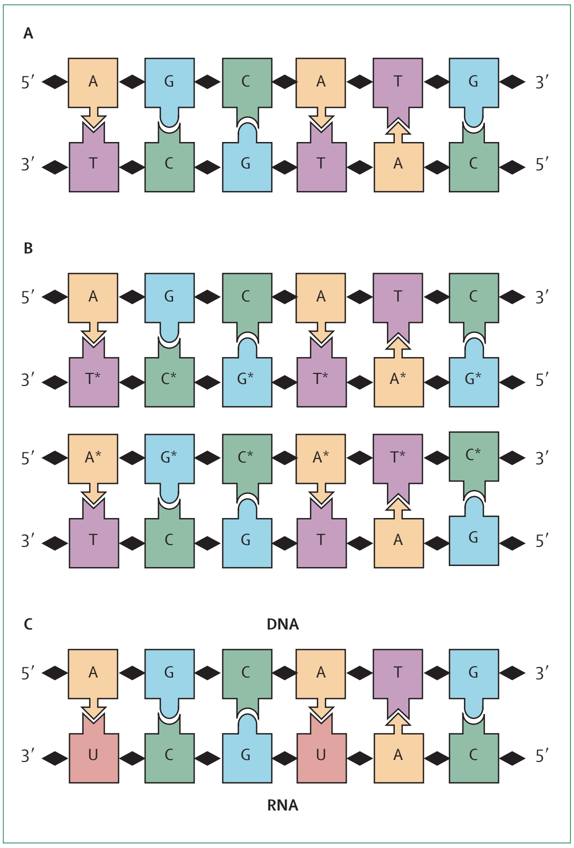

Figure 2: **DNA structure (A), replication (B), and transcription (C)**

A=base on newly synthesised strand.

Exon
A segment of a gene that is represented in the mature RNA product. Individual exons typically include protein-coding sequences.

Intron
Non-coding DNA that separates neighbouring exons in a gene.

mRNA (messenger RNA)
RNA transcribed from genes undergoes posttranscriptional processing and the resultant mature mRNA is used as the template for the translation process that results in synthesis of a protein.

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.[22,23] 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 called 5′ and 3′, oriented in opposite directions. 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

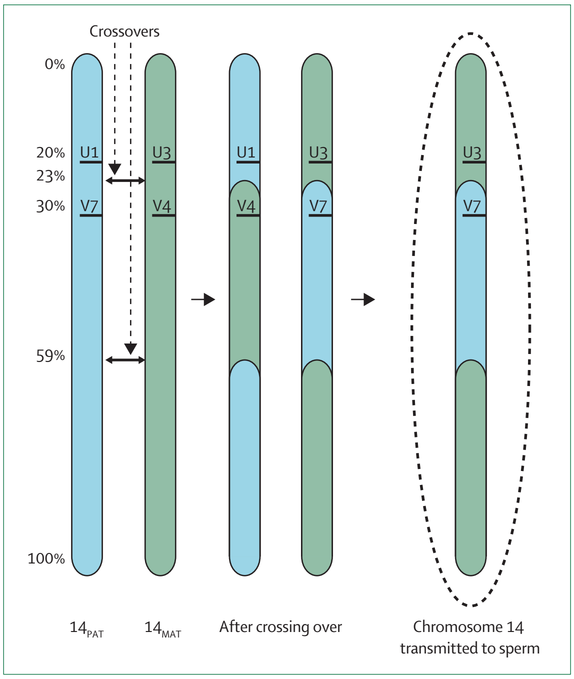

***Figure 3:* Crossing over and recombination**

Two hypothetical loci, U and V, are sited 20% and 30%, respectively, along the length of chromosome 14. They existed as alleles U$_1$ and V$_7$ on chromosome 14$_{PAT}$ (the chromosome derived originally from the man's father) and alleles U$_3$ and V$_4$ on chromosome 14$_{MAT}$ (the chromosome derived originally from the man's mother). Crossovers at 23% and 59% along the chromosome produce two mixed chromosomes. In this example, the right-hand chromosome is transmitted to the gamete, containing alleles U$_3$ and V$_7$. These two alleles were independently derived from the man's mother and the man's father, respectively.

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,[23] and there are 30 000–40 000 protein-coding genes.[24–26] 99·9% of the genome of any two unrelated individuals is identical, but the DNA sequence may vary between two versions of the same chromosome in several ways.

Many different types of DNA sequence variant exist, and they can be classified in different ways[23]—eg, by the physical nature of the sequence variation, by the effect on protein formation, and by the associated susceptibility to a disease. The two most important structural classes are **microsatellites** and **single nucleotide polymorphisms (SNPs)**. **Alleles** are differentiated by the number of repeats (eg, CA$_{12}$ indicates 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.[26]

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.[26] 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.[26] 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.[26]

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 cell division and accompanying replication and partitioning of DNA that leads to the formation of sperm and ova is meiosis.[23] 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 the 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

See http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi

Haploid

Gametes (sperm and ova) are haploid. They contain only one member of each homologous chromosomal pair (for example, only one version of chromosome 14). All ova have chromosomal complement 23,X and sperm are either 23,X or 23,Y. When sperm and ova fuse to form a zygote, the diploid chromosomal complement is restored.

Microsatellite

Microsatellites consist of multiple repeats of a short sequence (typically 2–8 bp) such as: CACACA . . . . The alleles of a microsatellite are differentiated by the number of repeats they involve (eg, CA$_{12}$ would denote 12 CA repeats in a row).

Polymorphism

Implies genetic variation at a designated locus. A locus that is polymorphic has at least two alternative alleles. Unfortunately, polymorphism has alternative, more specific definitions (none universally accepted), an important example being "the existence of two or more genetic variants (alleles, [other] sequence variants, chromosomal structure variants) at significant frequencies in the population."[22] In this series, polymorphism is used either as a component of the term single nucleotide polymorphism (see below) or it refers simply to a locus at which genetic variation is evident. Unless stated otherwise, its usage implies nothing about the type of variation observed or its frequency.

Single nucleotide polymorphism (SNP)

A DNA variant that represents variation in a single base. A common SNP can be defined as a locus at which two SNP alleles are present, both at a frequency of 1% or more.[109] Across the human genome there could be 10 million common SNPs.[109]

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.

transformed into an expected number of crossover events. Distance along a chromosome can be expressed in **centimorgans**. The relation between the length of DNA as measured in bp or centimorgans varies between men and women and from place to place in the genome, but a rule of thumb is that 1 centimorgan corresponds to about 1 billion bases.[27]

Genotypes, haplotypes, and phenotypes

Although the genotype is sometimes used to refer to the overall genetic constitution of an individual,[23] genetic epidemiologists use the term to refer to a particular locus. If three loci—U, V, and W—lie on a given chromosome and we take alleles U_3, V_2, W_2 along one homologous chromosome and U_1, V_2, W_1 along the other, the genotypes of the individual at the three loci are U_1U_3, V_2V_2, and W_1W_2. Expressed in this manner, a genotype has no natural order and the genotypes would have been the same if the two chromosomes had carried U_1,V_2,W_2, and U_3,V_2,W_1. The allelic configuration 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.[27]

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.[23] This definition includes binary disease states (presence or absence of asthma) and quantitative characteristics (systolic blood pressure). Some simple binary phenotypes are only 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[22,28,29] (figure 1). 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[30] or allied measure.[31] The pattern of recurrence risk ratios across different types of relatives can provide valuable information about the origin of a binary trait,[30] and can inform the statistical power of linkage studies.[15] 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".[32] The recurrence risk ratio (λ_R) 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, λ_S and P_S would be used. P and P_R will almost always be estimates, so λ 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.[33] 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 λ_R, 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.[34] The consequences of non-random sampling must be considered carefully, and any ascertainment bias should be dealt with in the analysis.[34–44] 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.[36–39,43,44]

Three interpretational issues warrant emphasis. First, the prevalence of many complex diseases increases

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 aminoacid coding is changed. The new code specifies an alternative aminoacid or changes the code for an aminoacid 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 codes can specify the same aminoacid. 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.

Centimorgan
1 centimorgan (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 U_1U_3–V_1V_2 is consistent with two possible phases: (1) U_1–V_1 on one chromosome and U_3–V_2 on the other; or (2) U_1–V_2 on one chromosome and U_3–V_1 on the other. If a child receives U_1–V_1 on a paternally derived chromosome from a father with diplotype U_1U_3–V_1V_2 it either implies that the father was in phase (1) and no recombination has occurred, or he was in phase (2) and there has been recombination.

steeply with age, whereas λ often declines.[45] 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[46]), λ_R cannot be greater than 2·0, even if every available relative is affected. Comparisons of λ_R across different diseases or different settings thus require care. Third, λ_R 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[47]), the term familial relative risk is used instead of λ.[48]

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[49,50] and Pearson.[51] The ICC indicates the proportion of the total variability in a phenotype that can reasonably be attributed to real variability between families.[52] Thus, the assessment of aggregation of a continuous measure in genetic epidemiology is fundamentally no different from, and could be viewed as predating,[10,50] analogous problems in traditional epidemiology and social science.[52,53] Consequently, techniques such as linear regression and mulitilevel modelling analysis of variance[52–58] can be imported directly into genetic epidemiology. As with the binary phenotype, non-random ascertainment can seriously bias an ICC.[42]

Interpretation

For many complex diseases, the average λ_R in first-degree relatives is around 2.[45] It tends to be greater the younger the age at onset in the affected individual,[45] to fall as the familial relationship becomes more distant,[30] and to increase as the number of affected relatives of the at-risk individual rises. Although a λ_R of 2 might appear modest, it does suggest that uncovering all sources of familial aggregation might well be worthwhile. A moderate λ_R generally implies the presence of underlying familial risk factors (genetic or non-genetic) that are at least an order of magnitude stronger than λ_R itself.[59,60] 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 (λ_R is about 1·1). A value of λ_R of around 2 would also be consistent with a number of common alleles each associated with a more modest relative risk. Knowing λ_R 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.[60] 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.

Penetrance
The probability that a particular phenotype is expressed in a person with a particular genotype.

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).[10,61,62] 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,[55,57,61,62] 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.[10,29,61,62,63,65] 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 a constant 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_2 but subtracts 3 if paired with G_3) the additive model will usually capture much of the aetiological information that can reasonably be explained by genes.

Panel 2: Identity by descent and identity by state

If two parents both of genotype G_1G_2 have two children who are also G_1G_2, these offspring could have received their G_1 from the same parent (case A) or one from either (case B). If the G_1 alleles are from different parents then so are the G_2 alleles. Any two individuals with genotypes G_1G_2 are said to share two alleles that are identical by state (IBS), irrespective of the origin of the two alleles and irrespective of whether the two individuals are related. An allele is identical by descent (IBD) only if it has been inherited directly from a common ancestor (which could be one of the two individuals themselves). Thus, the siblings in case A share 2 alleles IBD, and those in case B share no alleles IBD. Excess sharing of IBD alleles differentiates relatives from non-relatives, and this sharing is generally most important in genetic epidemiology. The table illustrates the patterns of IBD sharing between relatives.

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[64]) in addition to additive genetic effects.[10,55,61–63,65] 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.[29,55,57,61–63] 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.[55,66,–68]

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_G, 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[64]). Heritability in the broad sense is defined as S^2_G 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.[10,29,77] Fisher[78] pointed out that although the numerator has a simple genetic meaning, the "hotch-potch of a denominator" does not.[78] S^2_T 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.[7,77]

Heritability is formally defined for quantitative traits.[77] 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.[62,74,77,79] Unfortunately, with a binary phenotype, the heritability of the liability does not have a clear meaning and is prone to confused interpretation.[45,80–83]

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

	Parents	Parent–child	Full siblings	Grandparent–grandchild	Uncle–niece	First cousins	Half siblings	Identical twins
IBD sharing at a single locus								
Expected probability 2 alleles shared IBD	0	0	0·25	0	0	0	0	1
Expected probability 1 allele shared IBD	0	1	0·5	0·5	0·5	0·25	0·5	0
Expected probability 0 alleles shared IBD	1	0	0·25	0·5	0·5	0·75	0·5	0
Proportion of alleles shared IBD	Exactly 0	Exactly 0·5	On average 0·5	On average 0·25	On average 0·25	On average 0·125	On average 0·25	Exactly 1

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^2_T. Interpretation also depends on which covariates are included. For example, including an important environmental covariate might well decrease $S^2_{T,}$ but leave S^2_A 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_A) individually.[68,78,84]

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.[29,85,86] 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[88] (see a later paper in this series[12]).

Elston[89] 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[90]). 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.[38]

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.[91] 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).

Panel 3: Fitting of variance components models

Variance components analysis can be undertaken with conventional statistical models such as maximum likelihood[65] and generalised least squares,[55] or Markov chain Monte Carlo based approaches.[57] Genetic epidemiologists use various approaches to aid the specification of such models, including path analysis, which was invented by Sewall Wright nearly 100 years ago[69] and the fitting is achieved by various programs;[54,55,61,70–73] the details are beyond the scope of this article but a key feature is flexibility. So, if information is available about characterised genotypes, measured environmental determinants, and known demographics, it can enter the analysis. Equivalent approaches can also be used for binary phenotypes[55,57,74] and for traits that can best be expressed as a survival time,[75,76] such as age at onset or age at death.

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

Panel 4: A simple association analysis

The simplest class of association analysis involves a binary disease trait and a functional gene with two alleles, and requires an adequate number of unrelated individuals to have been typed for the gene of interest and classed as having, or not having, the disease. The simplest approach is to construct a 2×3 table:

	G_1G_1	G_1G_2	G_2G_2
Disease	109	118	26
No disease	138	88	21

We focus on analyses based on the distribution of genotypes by disease status . A conventional χ^2 test (with 2 degrees of freedom) takes the value 8·23 (p=0·016) implying significant heterogeneity in the risk of disease associated with the three genotypes. χ^2 test for linear trend is 6·23 (p=0·013). Logistic regression suggests that, on average, each additional copy of G_2 increases the odds of disease by a factor of 1·41 (95% CI 1·07–1·85). How these results are interpreted depends critically upon whether this is a one-off test on a single candidate gene (when the analysis can be interpreted at face value), or whether this is merely one marker gene among many tested, so demanding adjustment for multiple testing and the very low a-priori probability that a given locus is truly associated with the disease (see a later paper in this series64).[16–19]

Panel 5: Linkage disequilibrium *vs* simple linkage

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 M_{17} 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 M_{17} even though M_{17} has nothing to do with the cause of the disease. This is linkage disequilibrium (see Cordell and Clayton[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 M_3 in one family but equally closely with M_8 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[12]). Furthermore, linkage analyses for complex diseases are now often based on non-parametric methods (see a later paper in this series[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,[13] but there have been far fewer successes with complex diseases.[92,93] This is mainly because of limitations to statistical power (see a later paper in this series[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[12] 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[88,94–96] 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.[97–99] This disease was historically referred to as hereditary non-polyposis colorectal cancer or Lynch syndrome.[100] Other examples are familial breast-ovary syndrome (*BRCA1*)[101] and the female and male breast cancer syndrome (*BRCA2*).[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.[58,103,104] 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).[54,55,75]

Association analysis is covered in later papers in this series,[64,105,106] 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.[20,107,108] An association analysis based on a putative functional genetic variant can be called direct and one based on linkage disequilibrium with a marker indirect..[64,105,106] 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.[26]

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.[91,111,116–119] 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.[20,26] This move is not without its critics,[120] and a later paper in this series[47] will look at the future role of population-based family studies and the need to ensure that important opportunities are not missed.[121]

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.[122,123] 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.[91]

See http://www.hapmap.org

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.[126,129] This important debate will be covered in a later paper in this series.[91]

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 cooperative grant # G9806740, programme grant # 00\3209 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 #299955, 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

1 Last J. A dictionary of epidemiology. New York: Oxford University Press, 2001.

2 Neel JV, Schull WJ. Human Heredity. Chicago: University of Chicago Press, 1954.

3 Morton NE, Chung CS. Genetic Epidemiology. New York: Academic Press, 1978.

4 King MC, Lee GM, Spinner NB, Thomson G, Wrensch MR. Genetic Epidemiology. *Annu Rev Public Health* 1984; **5**: 1–52.

5 Morton NE. Outline of Genetic Epidemiology. London: Karger, 1982.

6 Roberts DF. A definition of genetic epidemiology. In: Chakraborty R, Szathmary EJE, eds. Diseases of Complex Etiology in Small Populations: Ethnic Differences and Research Approaches. New York: Alan R Liss, 1985: 9–20.

7 Hopper JL. The epidemiology of genetic epidemiology. *Acta Genet Med Gemellol* 1992; **41**: 261–73.

8 Mendel JG. The origins of genetics: a Mendel source book (translation). In Stern C, Sherwood E, eds. San Francisco: Freeman, 1966: 1–48.

9 Galton F. Hereditary talent and character. *MacMillan's Magazine* 1865; **12**: 157–66.

10 Fisher R. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinburgh* 1918; **52**: 399–433.

11 Wijsman EM. Mendel's laws. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 527–29.

12 Teare MD, Barrett JH. Genetic linkage studies. *Lancet* (in press).

13 Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* 2003; **33** (suppl): 228–37.

14 Palmer LJ. Complex diseases. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 141–43.

15 Risch N. Linkage strategies for genetically complex traits II. The power of affected relative pairs. *Am J Hum Genet* 1990; **46**: 229–41.

16 Lander ES, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11**: 241–47.

17 Todd JA. Interpretation of results from genetic studies of multifactorial diseases. *Lancet* 1999; **354**(suppl): S15–16.

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

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

20 Zondervan KT , Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004; **5**: 89–101.

21 Elston RC. The genetic dissection of multifactorial traits. *Clin Exp Allergy* 1995; **25** (suppl 2): 103–06.

22 Elston R, Olsen J, Palmer L. Biostatistical genetics and Genetic Epidemiology. Chichester , Wiley, Wiley Reference Series in Biostatistics, 2002.

23 Strachan T, Read AP. Human Molecular Genetics 3. Oxford: Garland Science Publishers, 2003.

24 Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.

25 Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001; **291**: 1304–51.

26 Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004; **429**: 446–52.

27 Sham P. Statistics in Human Genetics. London: Arnold, 1998.

28 Balding DJ, Bishop M, Cannings C. Handbook of Statistical Genetics. Chichester: Wiley, 2003.

29 Burton PR, Tobin MD. Epidemiology and Genetic Epidemiology. In Balding DJ, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Chichester: Wiley, 2003.

30 Risch N. Linkage strategies for genetically complex traits I. Multilocus models. *Am J Hum Genet* 1990; **46**: 222–28.

31 Kopciuk KA, Bull SB. Risk Ratios. In Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 687–91.

32 Rothman K, Greenland S. Measures of disease frequency. In Rothman K, Greenland S, eds. Modern Epidemiology, 2nd edition. Philadelphia: Lippincott-Raven, 1998: 29–46.

33 Rothman K, Greenland S. Types of Epidemiological Studies. In Rothman K, Greenland S, eds. Modern Epidemiology, 2nd edition. Philadelphia: Lippincott-Raven, 1998: 67–78.

34 Guo S-W. Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. *Am J Hum Genet* 1998; **63**: 252–58.

35 Weinberg W. Mathematische Grundlagen der Probandenmethode. *Z Indukt Abstamm Vererbungsl* 1928; **48**: 179–228.

36 Fisher RA. The effect of methods of ascertainment upon the estimation of frequencies. *Ann Eugen* 1934; **6**: 13–25.

37 Morton NE. Genetic tests under incomplete ascertainment. *Am J Hum Genet* 1959; **11**: 1–16.

38 Elston RC, Sobel E. Sampling considerations in the gathering and analysis of pedigree data. *Am J Hum Genet* 1979; **31**: 62–69.

39 Ewens WJ, Shute NC. The limits of ascertainment. *Ann Hum Genet* 1986; **50**: 399–402.

40 Kraft P, Thomas DC. Bias and efficiency in family-based gene-characterisation studies: conditional, prospective, retrospective and joint likelihoods. *Am J Hum Genet* 2000; **66**: 1119–31.

41 Hodge SE. Ascertainment. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 20–28.

42 Burton PR, Palmer LJ, Jacobs K, Keen KJ, Olsen JM, Elston RC . Ascertainment adjustment: where does it take us? *Am J Hum Genet* 2000; **67**: 1505–14.

43 Burton PR. Erratum: Ascertainment adjustment: where does it take us? *Am J Hum Genet* 2001; **69**: 692.

44 Burton PR. Correcting for non-random ascertainment in generalized linear mixed models (GLMMs) fitted using Gibbs sampling. *Genet Epidemiol* 2003; **24**: 24–35.

45 Risch N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol Biomarkers Prev* 2001; **10**: 733–41.

46 Cookson W, Palmer L. Investigating the Asthma Phenotype. *Clin Experiment Allergy* 1998; **28**: 88–89.

47 Hopper JL, Bishop DT, Easton DF. Population-based family studies in genetic epidemiology. *Lancet* (in press).

48 Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *J Natl Cancer Inst* 1994; **86**: 1600–08.

49 Galton F. Typical laws of heredity. *Proc R Inst* 1877; **8**: 282–301.

50 Galton F. Family likeness in stature. *Proc R Soc* 1886; **40**: 42–73.

51 Pearson K. Mathematical contributions to the theory of evolution: III. Regression, heredity and panmixia. *Phil Trans R Soc A* 1896; **187**: 253–318.

52 Burton PR, Gurrin L, Sly P. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Stat Med* 1998; **17**: 1261–91.

53 Goldstein H. Multilevel Models in Educational and Social research. London: Charles Griffin and Company Ltd, 1987.

54 Zeger SL, Liang KY. An overview of methods for the analysis of longitudinal data. *Stat Med* 1992; **11**: 1825–39.

55 Neale MC, Cardon LR. Methodology for Genetic Studies of Twins and Families. Boston: Kluwer, 1992.

56 Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; **88**: 9–25.

57 Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genet Epidemiol* 1999; **17**: 118–40.

58 Armitage P, Berry G, Matthews JNS. Oxford, Blackwell Scientific Publications, 2002.

59 Peto J. Genetic predisposition to cancer. In Cairns J, Lyon JL, Skolnick M, eds. Banbury Report 4: Cancer incidence in defined populations. Cold Spring Harbour Laboratory, 1980: 203–13.

60 Hopper JL, Carlin JC. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *Am J Epidemiol* 1992; **136**: 1138–47.

61 Hopper J. Variance components for statistical genetics: applications in medical research to characteristics related to human disease and health. *Stat Methods Med Res* 1993; **2**:199–223.

62 Khoury MJ, Beaty TH, Cohen BH. Fundamentals of Genetic Epidemiology. Oxford: Oxford University Press, 1993.

63 Hopper JL, Visscher PM. Variance Component Analysis. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 778–88.

64 Cordell HJ, Clayton DG. Genetic association studies. *Lancet* (in press).

65 Hopper JL, Mathews JD. Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 1982; **46**: 373–83.

66 Jinks JL, Fulker DW. Comparison of the biometrical, genetical, MAVA, and classical approaches to the analysis of human behavior. *Psychol Bull* 1970; **73:** 311–49.

67 Duffy DL, Martin NG. Inferring the direction of causality in cross-sectional twin data: theoretical and empirical considerations. *Genet Epidemiol* 1994; **11:** 483–502.

68 Neale MC. Twin analysis. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 743–56.

69 Wright S. Correlation and causation. *J Agric Res* 1921; **20**: 557–85.

70 Neale MC, Boker SM, Xie G, Maes HH. Mx: Statistical Modeling . Virginia, USA: Richmond, 2002.

71 Lange ES, Boehnke M, Weeks D. Programs for Pedigree Analysis. Los Angeles: Department of Biomathematics, UCLA, 1987.

72 Rasbash J, Browne W, Goldstein H, et al. A User's Guide to MLwiN. London: Institute of Education, 1999.

73 Spiegelhalter D, Thomas A, Best N. WinBUGS Version 1.3: User Manual. Cambridge: MRC Biostatistics Unit, 2000.

74 Falconer DS. The inheritance of liability to certain disease, estimated from the incidence among relatives. *Ann Hum Genet* 1965; **29**: 51–71.

75 Scurrah KJ, Palmer LJ, Burton PR. Variance components analysis for pedigree-based censored survival data using generalized linear mixed models (GLMMs) and Gibbs sampling in BUGS. *Genet Epidemiol* 2000; **19**: 127–48.

76 Gauderman WJ, Thomas DC. Censored survival models for genetic epidemiology: a Gibbs sampling approach. *Genet Epidemiol* 1994; **11**: 171–88.

77 Hopper JL. Heritability. In Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 371–72.

78 Fisher RA. Limits to intensive production in animals. *Br Agric Bull* 1951; **4**: 217–18.

79 Hopper J. Variance components for statistical genetics: applications in medical research to characteristics related to human disease and health. *Stat Methods Med Res* 1993; **2**: 199–223.

80 Burton PR, Tobin MD. Epidemiology and Genetic Epidemiology. In: Balding DJ, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Chichester: Wiley, 2003.

81 Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer - analyses of cohorts and twins from Sweden, Denmark and Finland. *N Engl J Med* 2000; **343**: 78–85.

82 Hoover RN. Cancer: nature, nurture or both. *N Engl J Med* 2000; **343**: 135–36.

83 Spector N, Shapiro BL, Peto R, et al. Cancer, genes and environment (correspondence). *N Engl J Med* 2000; **343**: 1494–96.

84 Hopper JL, Macaskill G, Powles JG, Ktenas D. Pedigree analysis of blood pressure in subjects from rural Greece and relatives who migrated to Melbourne, Australia. *Genet Epidemiol* 1992; **9**: 225–38.

85 Majumder PP. Segregation analysis, Classical. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 693–96.

86 Blangero J. Segregation Analysis, Complex. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 696–708.

87 Palmer LJ, Cookson WO, James AL, Musk AW, Burton PR. Gibbs sampling-based segregation analysis of asthma-associated quantitative traits in a population based sample of nuclear families. *Genet Epidemiol* 2001; **20**: 356–72.

88 Terwilliger JD. Linkage analysis model based. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 448–60.

89 Elston RC. Segregation analysis. *Adv Hum Genet* 1981; **11**: 372–73.

90 Cui J, Antoniou AC, Dite GS, et al. After BRCA1 and BRCA1: what next? Multifactorial analyses of three-generational, population-based Australian female breast cancer families. *Am J Hum Genet* 2001; **68**: 420–31.

91 Davey Smith G, Burton PR. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet* (in press).

92 Weiss ST, Raby BA. Asthma Genetics 2003. *Hum Mol Genet Adv Access* 2004; **13**: 83R–89R.

93 Mathew CG, Lewis CM. Genetics of inflammatory bowel disease: progress and prospects. *Hum Mol Genet* 2004; **13**: 161R–68R.

94 Thompson EA. Linkage analysis. In: Balding DJ, Bishop M, Cannings C, eds. Handbook of Statistical Genetics Chichester: Wiley, 2001: 541–63.

95 Holmans P. Non-parametric linkage. In: Balding DJ, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Chichester: Wiley, 2001: 487–505.

96 Olson JM. Linkage analysis model-based. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 461–72.

97 Fishel R, Lescoe MK, Rao MR, et al. The human mutator gene homolog *MSH2* and its association with hereditary nonpolyposis colon cancer. *Cell* 1993; **75**: 1027–38.

98 Kolodner RD, Hall NR, Lipford J, et al. Structure of the human MSH2 locus and analysis of two Muir-Torre kindreds for msh2 mutations. *Genomics* 1994; **24**: 516–26.

99 Kolodner RD, Hall NR, Lipford J, et al. Structure of the human MLH1 locus and analysis of a large hereditary nonpolyposis colorectal carcinoma kindred for mlh1 mutations. *Cancer Research* 1995; **55**: 242–48.

100 Lynch HT, Lynch J. Lynch syndrome: genetics, natural history, genetic counseling, and prevention. *J Clin Oncol* 2000; **18**: 19S–31S.

101 Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994; **266**: 66–71.

102 Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 1995; **378**: 789–92.

103 Breslow NE, Day NE. Statistical Methods in Cancer Research. Volume 1: the analysis of case-control studies. Lyon: International Agency for research on Cancer, 1980.

104 Breslow NE, Day NE. Statistical Methods in Cancer research. Volume 2: the design and analysis of cohort studies. Lyon: International Agency for Research on Cancer, 1987.

105 Palmer LJ, Cardon LR. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* (in press).

106 Hattersley AJ, McCarthy MI. What makes a good genetic association study? *Lancet* (in press).

107 Clayton DG. Population association. In: Balding DJ, Bishop M, Cannings C, eds. Handbook of Statistical Genetics. Chichester: Wiley, 2001.

108 Chakravarti A. Linkage disequilibrium. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 472–75.

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

110 Schaid DJ. Disease Marker Association. In: Elston R, Olsen J, Palmer L, eds. Biostatistical Genetics and Genetic Epidemiology. Chichester: Wiley, 2002: 206–17.

111 Marchini J, Cardon LC, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36**: 512–17.

112 Thomas DC, Witte JS. Point: Population stratification: a problem for case-control studies of candidate-gene associations. *Canc Epidemiol Biomarkers Prev* 2002; **11**: 505–12.

113 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361**: 598–604.

114 Wacholder S , Rothman N, Caporaso N. Counterpoint: Bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Canc Epidemiol Biomarkers Prev* 2002; **11**: 513–20.

115 Freedman LF , Reich D, Penney K, et al. Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36:** 388–93.

116 Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependant diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–16.

117 Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet* 1995; **57**: 455–64.

118 Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.

119 Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–59.

120 Terwilliger JD, Weiss KM. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* 1998; **9**: 578–94.

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

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

123 Davey Smith G, Ebrahim S. Epidemiology: is it time to call it a day? *Int J Epidemiol* 2001; **30**: 1–11.

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

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

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

127 Tobin MD, Minelli C, Burtin PR, Thompson JR. The development of Mendelian randomisation: from hypothesis testing to “Mendelian deconfounding”. *Int J Epidemiol* 2004; **33**: 26–29.

128 Minelli C, Thompson JR, Tobin MD, Abrams KR. An integrated approach to the Meta-Analysis of Genetic Association Studies using Mendelian Randomisation. *Am J Epidemiol* 2004; **160**: 445–52.

129 Burton PR, McCarthy M, Elliott P. Study of genes and environmental factors in complex diseases. *Lancet* 2002; **359**: 1155–56.