European Journal of Epidemiology **18**: 607–616, 2003.

REVIEW

Prospects of genetic epidemiology in the 21st century

Marieke C.J. Dekker & Cornelia M. van Duijn
Department of Epidemiology and Biostatistics, Erasmus MC, Rotterdam, The Netherlands

Accepted in revised form 14 April 2003

Abstract. Genetic epidemiology is a young but rapidly developing discipline. Although its early years were largely dedicated to family-based research in monogenic disorders, now genetic–epidemiologic research increasingly focuses on complex, multifactorial disorders. Along with the development of the human-genome map and advances in molecular technology grows the importance of genetic–epidemiologic applications. Large-scale population-based studies, requiring close integration of genetic and epidemiologic research, determine future research in the field. In this paper, we review the basic principles underlying genetic–epidemiologic research, such as molecular genetics and familial aggregation of disease, as well as the typical study approaches of genome screening and candidate-gene studies.

Key words: Familial aggregation, Genetics, Genetic epidemiology, Polymorphisms, Study design

Abbreviations: APOE = apolipoprotein-E gene; CYP2D6 = cytochrome P450 debrisoquine-4-hydroxylase gene; DNA = deoxyribose nucleic acid; LOD score = logarithm-of-odds score; PSEN = presenilin gene; RNA = ribonucleic acid; SNP = single-nucleotide polymorphism; STR = short tandem-repeat; TDT = transmission-disequilibrium test; UCH-L1 = ubiquitin carboxy-terminal hydroxylase L1

Introduction

Sprung from genetics and epidemiology, genetic epidemiology is a new and rapidly growing field of epidemiology. Genetic–epidemiologic studies focus upon the role of inherited factors in disease aetiology. In recent years, research interest has shifted from genetic disorders that are caused by a single gene (e.g., Huntington's disease) to common multifactorial disorders or complex genetic disorders, which are likely to be the outcome of an interaction of genes and environment. Genetic–epidemiologic research distinguishes itself from general epidemiology not only by the study objective of genetic factors as disease determinants, but also by study design and statistical analysis. The study design concerns examination of individuals as well as data of relatives. Depending on the degree of relationship, relatives share their genetic and environmental background to a great extent. In the statistical analysis, primarily these relationships have to be taken into account. A second statistical issue to be dealt with is the fact that genes are organised into linear structures (chromosomes).

Genetic epidemiology is a discipline that covers a broad spectrum of research, ranging from familial aggregation of disease to the molecular origin of a disorder. This review discusses the principles and methods of research. The two major task forces of genetic–epidemiologic research are addressed, i.e., the identification of genetic risk factors involved in a condition, and quantification of their impact on occurrence of the disease in the general population. Finally, new developments in the field are discussed.

Genetic transmission of disease

Genetic–epidemiologic research of a disorder starts off with the question whether there is evidence for transmission of the trait in families. The epidemiologic approach to this question is the case–control study design, comparing disease prevalence in relatives of patients to that in relatives of controls. In such studies, the strength of familial aggregation may be expressed as an odds ratio or relative risk. Alternatively, 'classical' genetics address a familial trait by estimating heritability. The heritability of a trait is defined as the proportion of the total variance in a trait that can be explained by (additive effects of) genes.

The next step to be taken in genetic–epidemiologic research is the identification of the genetic basis of disease. A distinction commonly made is that between Mendelian and non-Mendelian traits. Mendelian disorders are caused by either a dominant or recessive mutation. For a dominant mutation, disease is already established when an individual receives one mutant form of a gene from either parent. Huntington's disease is an example of an autosomal dominant disorder. Autosomal refers to the location of the disease gene on one of the 22 autosomes, i.e. chromosomes that are identical in men and women. When two defect copies of a gene are needed to develop disease, a mutation is referred to as recessive. A classical example of an autosomal recessive disease is cystic fibrosis, which the most common recessive disorder among Caucasians. A special situation concerns the sex chromosomes, X and Y. As males only possess one X-chromosome, a recessive mutation on this chromosome can be pathogenic when only one copy is present, whereas in women two defective copies are needed to establish disease.

Examination of patterns of familial clustering may yield clues to the nature of the mutation involved in the disease. For a dominant mutation, the disease is expected to be present in multiple generations, because presence of one copy of the mutation is sufficient to lead to pathology. Since subjects carrying two dominant mutations (homozygotes) are rare, the condition often being incompatible with life, most patients carry only one copy of the mutation (heterozygotes). Since the probability that the mutant or normal gene is transmitted to offspring is equal, patients will pass on a dominant disease to approximately 50% of the offspring. Parents of patients with a recessive disorder are (most) often heterozygotes and therefore not affected themselves. Typically, recessive disorders often emerge in consanguineous matings, since this increases the probability that both parents are carriers of the same mutation. In a recessive disorder, offspring of carriers of a recessive mutation have a 25% chance of developing the disease. Duchenne's muscular dystrophy shows a typical clustering of disease in males, with no male-to-male transmission. This pattern is typical for X-linked disorders.

A problem in examining Mendelian transmission of disease is that mutations may not always lead to disease. The cumulative incidence of disease (penetrance) may depend on age, sex and other factors. Sometimes there is an obvious explanation for non-penetrance, e.g., testicular cancer in women or breast cancer in men. However, in most cases the reason for a variable penetrance is unknown, and may be determined by other genes or environmental agents.

In contrast to the Mendelian inheritance patterns discussed thus far, inheritance of non-Mendelian, multifactorial disorders is more complex and difficult to define. In non-Mendelian disorders, the disease may be the resultant of the interaction of multiple genes, each of which has a minor contribution to pathogenesis. However, also the presence of a (large) number of dominant or recessive mutations with reduced age- or sex-related penetrance may result in apparently non-Mendelian transmission of disease. Most common complex disorders (Alzheimer's disease; Parkinson's disease; hypertension; diabetes) exhibit complex inheritance patterns, which may be the result of expression of multiple genes. In non-Mendelian disorders, the relationship between genotype and phenotype (the observable trait) is not straightforward.

Assessment of familial aggregation is a first step in genetic–epidemiologic research. Knowledge of the transmission of disease (Mendelian or non-Mendelian) is crucial in order to determine the most powerful strategy for a study of the molecular basis of disease, as will be dealt with later. First, we briefly review the molecular basis of genetic disorders.

Molecular basis of disease

Genetic information is stored by deoxyribose nucleic acid (DNA). At the molecular level, DNA is made up of a sugar, a phosphate and a base [1]. The DNA sequence is described by the order of bases (adenine, guanine, cytosine, thymine), represented by their initials A, G, C, and T. Three-base units, together with the sugar and phosphate component (referred to as codons) translate into amino acids. In a process called transcription, DNA is copied into single-stranded ribonucleic acid (RNA), which is subsequently translated into protein.

In the 19th century, without knowledge of underlying molecular biology, the Augustinian clergyman Mendel introduced the term gene as the fundamental unit that transmits traits from parents to offspring [2]. The knowledge of underlying molecular biology, however, provides different ways of defining a gene. The most straightforward definition of a gene is that part of DNA encoding for a protein. Not all DNA codes for protein – 50% of the genome is made up of repetitive and non-coding sequences. At present, the number of genes coded by the human genome is estimated to be less than 40,000 [3, 4]. Within a gene, non-coding sequences occur (introns). Exons are the parts of a gene translated into protein whereas introns are the parts that are removed (spliced out) upon translation from RNA to protein. Mutations in exons leading to a change in amino acid (order) in a protein may be pathogenic due to loss or gain in protein function. Although introns do not encode for proteins, mutations can affect intron splicing, subsequently changing the protein's structure or synthesis.

Furthermore, the genetic code is degenerate, which implies that several triplets can code for the same amino acid. Hence point mutations do not always result in changes at the level of the amino acid. Such silent mutations may be dispersed throughout the population. This also applies to mutations in the widely spread non-coding sequences. Thus, at one particular locus in the human genome, several forms of the same gene may exist. These are called polymorphisms. At a molecular level, the difference between mutations and polymorphisms is not clear-cut, leaving frequency and clinical penetrance as the two distinctive factors. Mutations have a low frequency (<5%) but are thought to be highly pathogenic. In the presenilin-1 (PSEN1) gene, various mutations are known which in virtually all carriers lead to Alzheimer's disease with an onset before age 55 years [5]. Polymorphisms are common (⩾5%) in the general population. Polymorphisms may be associated with only a modest increase in disease risk, or even be functionally unrelated to the disease. An example of a disease-related polymorphism is the apolipoprotein ε4-allele (APOE*4), which has an allele frequency of 17% in Caucasians. Its association with Alzheimer's disease has been studied widely. The risk of developing Alzheimer's disease for carriers of the APOE*4 allele is 1.5–2.5-fold increased [6]. Nevertheless, APOE*4 is neither necessary nor sufficient to develop Alzheimer's disease [7].

Methods in genetic epidemiology

Within genetic epidemiology, two main lines of research can be distinguished. The first line of research aims to identify new genes involved in disorders. The second line of research aims to quantify the risk of disease for carriers of a known mutation or polymorphism. The first line of research only distantly relates to classical epidemiology. The work in the second line is closely linked to epidemiologic and clinical research. Basically, risk estimation for genes follows the classical approach of epidemiologic studies. Risks of disease may be derived from follow-up studies comparing risk of disease in carriers to that in non-carriers. Relative risks may be estimated from case–control studies by the odds ratio [8]. Different strategies followed in order to identify genes involved in disease will be described.

Genomic screens

There are two different strategies to identify genes involved in a disorder. The first and hitherto most successful approach to identify new genes is genome screening, which involves a complete genome search for genes involved in a disorder. This approach starts by genotyping a set of STR of SNP polymorphic markers, of which the genomic location is known. Usually, these markers are more or less equally distributed across the genome, covering all chromosomes. These markers are not necessarily located in a gene, but often are located in non-coding areas not known to be involved in any biological process.

The rationale of genome screening is, that a causally related mutation should be found more often in patients with a particular disorder than in controls. However, given the size of the genome of about 2.9 billion base pairs [3], the probability that a random marker is located at a disease mutation by chance next to zero. Our genetic information is linearly arranged in chromosomes. Loci physically close together on a chromosome are likely to be transmitted together from parent to offspring. Therefore, patients who inherit a disease gene from a common ancestor not only receive the disease mutation, but also adjacent parts of the chromosome. Any marker located physically nearby a causal mutation should at least be present more often in cases than in unaffected relatives or unrelated controls, merely flagging the mutation. Consequently, disease genes can be identified by genome screening with a limited number of markers.

Candidate-gene studies

An alternative strategy is that of candidate-gene studies. Based on the gene product, the protein, or homology to a gene that is known to be involved in the disease, genes can be candidate to cause a certain disease. To determine whether mutations in the gene are involved, the gene can be screened for mutations or polymorphisms. A major drawback of this approach is that *a priori* knowledge of the pathogenesis of the disease is required: Proteins or genes involved in the disease should be known. For a large number of disorders, there is limited knowledge of proteins involved in the etiology. For instance, before cloning of the presenilin (PSEN) genes involved in early-onset Alzheimer's disease, the presenilin protein and its function were unknown, [9] as were various proteins involved in the ubiquitin pathway (alpha-synuclein; parkin; UCH-L1) [10–12].

Another problem in studies of candidate genes is that the candidacy of a gene for a disorder is sometimes ill-defined. A great number of proteins may be hypothetically involved in diseases such as Parkinson's disease, based on proteins detected post-mortem in brain tissue, or based on the neurotransmitter pathway involved. Multiple testing becomes an important issue in studies in which candidate genes are addressed *ad libidum*. Given the large number of genes that can be tested for, an important issue to resolve in candidate-gene studies is adjusting the significance level for multiple testing. Performing thousands of tests will yield a large number of false-positive findings at a significance level of 0.05. For genome screening in families with multiple affected generations, established criteria are available adjusting for multiple testing based on the number of tests that can be made given the size of the genome and the linkage between regions. For candidate-gene studies, the debate on how to adjust for multiple testing is ongoing [13]. Due to the nature of the problem, it is questionable whether consensus will be reached [8, 13]. Although adjustment for multiple testing is necessary, the need for replication of findings in genetically different populations is perhaps more important.

Genetic–epidemiologic research in families

Family studies

Traditionally, family-based study design has been the backbone of genetic–epidemiologic research. Family studies have been of great importance to the identification of new genes. Using studies in extended pedigrees has lead to the unravelment of several genetic disorders including Huntington's disease and cystic fibrosis. This particularly applies to monogenic disorders, in which there is a clear-cut relation between genetic factor and occurrence of disease.

The understanding of linkage analysis requires some basic knowledge of meiotic divisions during gametogenesis. At meiosis, two homologous chromosomes are separated. Subsequently, either of the two chromosomes is transmitted randomly to a gamete (see Figure 1). During this process, homologous parts of chromosome pairs may cross over. Crossing-over is an exchange or recombination of the genetic information encoded on the two homologous chromosomes. As shown in Figure 1, two loci close together on a chromosome will more frequently be transmitted together. The closer two loci are, the less likely it is that recombination occurs in between two loci: they are linked to one another. The more distantly two loci are situated on the chromosome, the more closely the probability of recombination approaches the probability of non-recombination.

Linkage analysis uses the principle of recombination to localise a disease mutation transmitted in a family. Relatives who develop a disease due to the same mutation are expected to share alleles on DNA markers flanking the disease mutation (Figure 2). The objective of a linkage study is to find markers of which a particular allele is preferentially transmitted to patients. Detailed statistical aspects of linkage studies are beyond the scope of this paper. Basically, in a linkage analysis the number of recombinations between disease status and a marker allele (observed in a family) is compared to the expected number of recombinations under the null hypothesis. The test statistic for linkage is the LOD score. The LOD score is the log of the likelihood ratio of linkage of the disease to the studied marker, versus the likelyhood of no linkage. Significant evidence for linkage is found when LOD scores exceed 3, whereas LOD scores below −2 imply definite exclusion of the region [14].

Linkage studies have been extremely successful in disclosing the aetiology of monogenic disorders. At present, genetic research focuses on chronic disorders with a complex etiology. Several genetic and environmental factors may be implicated in these disorders, the mutation-associated risk heavily depending on presence of other genetic or environmental risk factors. As it is often impossible to clinically distinguish between patients who developed the disease due to a specific mutation and those who have a different aetiology (called phenocopies), recombination between the disease and marker may be falsely inferred. The power of linkage analysis in complex disorders is therefore low. An alternative approach to analysis of complex disorders is the use of affected sibling-pairs.

Sib-pair studies

Siblings share a high proportion (50%) of their genetic material including large parts of DNA (Figure 3). The *a priori* probability of a patient sharing no alleles with any other sib is 25%; one allele is 50%; two alleles is again 25%. For markers located close to the disease mutation, affected sibs are expected to share more alleles than the average of one allele. The test statistic for the analysis is based on counting alleles shared by a pair of affected siblings. Counts exceeding the expected value under the null hypothesis (one allele shared) are compatible with a disease locus nearby the marker examined [15].

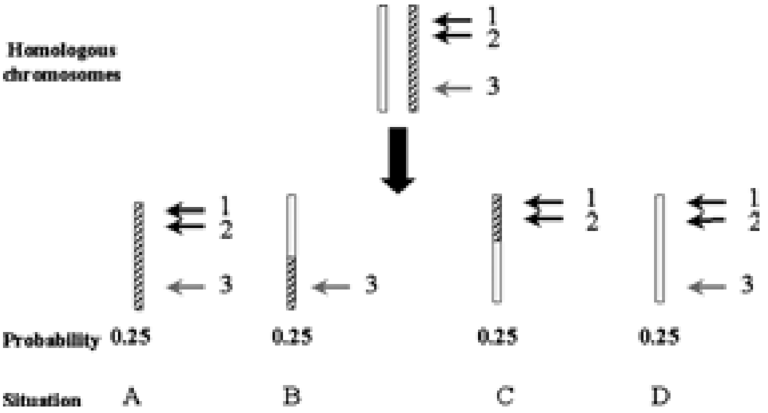

Figure 1. Crossing-over between homologous chromosomes during meiosis. Loci close together (indicated by arrows 1 and 2) are likely to be transmitted together (probability >50%). When loci are physically well separated on the chromosome, such as those indicated by arrows 1 and 3, separation may occur by recombination (shown in situation B and C). For two distant loci, the probability of recombination (50%; B and C) equals the probability of non-recombination (50%; A and D).

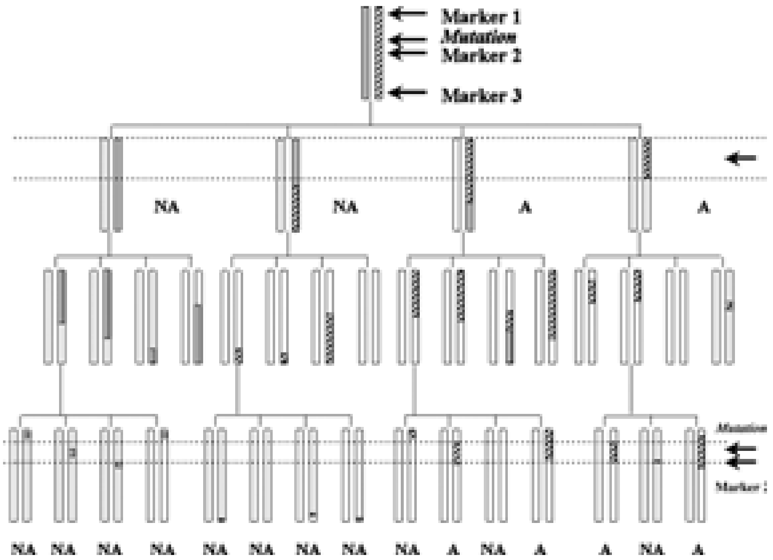

Figure 2. Linkage analysis. This figure depicts two homologous chromosomes of a founding parent (dashed and grey). In one of them (dashed chromosome), a mutation occurred, which will be passed down to 50% of the offspring. Each carrier receiving the mutation may pass it to offspring with a 50% probability. Along with the mutation, an amount of flanking DNA is transmitted. Due to recombination, the piece of DNA shared by patients consecutively becomes smaller over generations. In the figure three markers (1,2,3) flanking the mutation are shown. In a genomic search, patients in the third generation may no longer share markers 1 and 2 of the mutated chromosome, but marker 2 still flags the mutation.

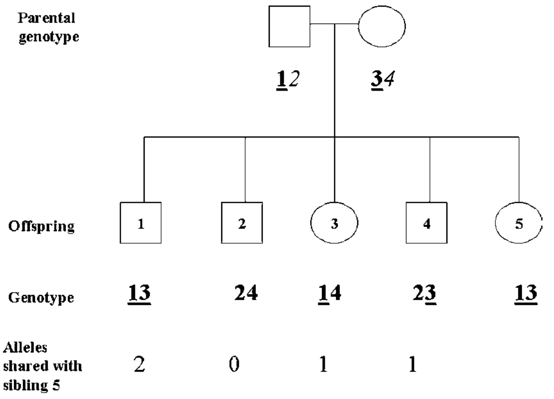

Figure 3. Expected allele-sharing in affected and unaffected siblings. This figure shows a pedigree of two parents and five children. Squares indicate males and circles females. Since parents pass 1 of 2 alleles to offspring with equal probability, siblings share 50% of DNA on average. On average, two siblings share 1 allele.

An advantage of the sib-pair design is that two siblings with the same common disease are more likely to have developed the disease due to the same mutation than two distantly related subjects. Furthermore, siblings do not only share a high proportion of DNA but also large chromosomal regions. In principle, the disease gene may thus be detected with a limited number of markers. However, the statistical power of sib-pair studies is limited, particularly if multiple genes are involved [13]. Detecting significant linkage in such disorders requires several thousands of (affected) sibling pairs.

Choice of study design

The choice of the most powerful study design (linkage vs. sib-pair approach) requires an impression of the extent of familial aggregation, in order to determine whether a disorder is segregated as Mendelian or non-Mendelian trait. The most powerful design for a Mendelian disorder is linkage analysis, whereas for a non-Mendelian trait sib-pair analysis is more suitable. Among patients with a common non-Mendelian disorder, however, subgroups with distinctly Mendelian segregation may be identified. For instance within Alzheimer's disease and Parkinson's disease, early-onset Mendelian forms have been recognised for which single-gene mutations have been identified successfully through linkage analysis [11, 16–18]. Although these traits are rare and may only explain a minor fraction of disease in the population, knowledge of the molecular-genetic origin of the disease in these early-onset cases may yield clues toward key proteins involved in pathogenesis. These proteins may serve as targets in the improvement of therapeutic strategies.

When examining familial aggregation, clustering of a disease may not only be due to genetic factors, but also to environmental factors. For instance, the familial clustering of nutritional habits may explain familial aggregation of disease. This may also apply to occupational factors and intoxication. Hypertension is a complex condition showing strong familial aggregation and high degree of heritability. Salt sensitivity is known to be a genetically determined risk factor for hypertension [19]. However, in a population with a uniformly low salt consumption, genetic contribution to the incidence of hypertension will be low, while in populations with larger variation in salt intake its contribution may be considerable. For complex disorders, familial aggregation is therefore not a fixed property of a trait.

Estimation of the risk of a disease mutation in families requires ascertainment of a random group of unrelated families. Such studies are expensive and time-consuming. Therefore, this study design is rarely used to estimate gene-associated risk. For extremely rare mutations on a population level, families ascertained for the presence of rare mutations are useful to study the risk of disease in carriers. Risk-assessment studies on mutations involved in Huntington's disease, early-onset Alzheimer's or Parkinson's disease requires an astronomically large sample size. Furthermore, studies of genetic and environmental risk-modifying factors associated with these mutations may only be feasible in those few families segregating the rare diseases.

Population-based studies

Family-based studies constitute the classic approach to determine the genetic aetiology of a trait. Only in monogenic and oligogenic (which indicates involvement of only a few genes) Mendelian disorders, however, is this approach feasible. At present, the mainstay of genetic research is its focus on common disease such as Alzheimer's disease, Parkinson's disease, hypertension and diabetes. A family-based approach rarely yields sufficient power to detect a genetic cause for common disorders, although rare and conspicuous variants of a disorder can give clues about its etiology. An alternative for linkage analysis is a study of affected sib-pairs. Unfortunately, in order to detect genes involved, very large numbers of affected sib-pairs are required to gain sufficient statistical power. For disorders with a high mortality or with late onset, affected sib-pairs are difficult to trace, which further limits the feasibility.

There is increasing interest in population-based studies of individuals in order to overcome the limitations of such family-based research. As in family studies, the basic principle of molecular studies in the population is that besides a disease gene, DNA flanking this gene is also passed on to the next generation and is thus dispersed throughout the population. Hence, a mutation related to disease can be ascertained in a genomic screen by identifying chromosomal regions shared by patients (Figure 4). This alignment of genes along the chromosome, called a haplotype, is unique for each individual.

In the general population, it is not yet statistically feasible to perform a genomic screen for markers in linkage disequilibrium with a disease. Firstly, there is only a small probability that any two patients from the general population with a common complex disorder have inherited a gene from a common ancestor. Secondly, people with a common trait, randomly derived from the general population, will on average be related only (very) distantly. Any two individuals hence only share a small amount of DNA. As shown in Figure 2, the amount of DNA shared progressively diminishes over generations. This requires marker and disease locus to be very close together in order to

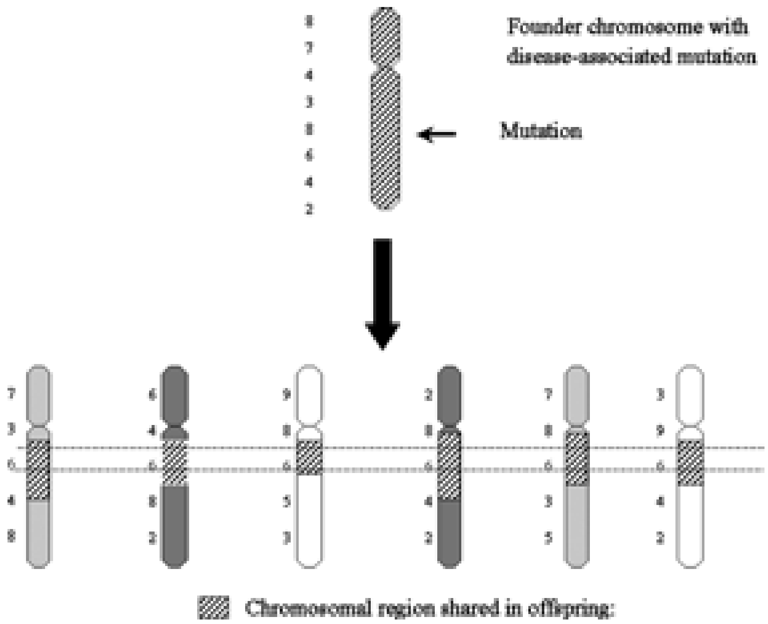
8
7
4
3
8
6
4
2
Founder chromosome with disease-associated mutation
Mutation
7
3
6
4
8
6
4
6
8
2
9
8
6
5
3
2
8
6
4
2
7
8
6
3
5
3
9
6
4
2
Chromosomal region shared in offspring:

Figure 4. Allele sharing in a genetically isolated population. Above: Founder chromosome with disease-associated mutation. Below: Region surrounding the disease locus, shared by patients with the same phenotype. Although they are possibly unaware of this, these affected individuals all descend from a common ancestor.

localise the gene in a genome screen. A large number of markers with dense spacing and extensive numbers of patient are therefore needed [20]. The use of dense single-nucleotide polymorphism or SNP maps to scan the genome in large numbers of people is quickly gaining efficiency, using the human-genome map and high-throughput apparatus.

Candidate-gene studies in the general population

Nevertheless are samples derived from the general population also very suitable to study candidate genes. Candidate-gene studies have been widely criticised because of the repeated failure to replicate results. An example of an extensively studied disorder in terms of candidate genes is Parkinson's disease. An impaired enzyme detoxification-capacity has long been thought to account for an increased susceptibility to Parkinson's disease [21], possibly due to an impaired ability to handle environmental neurotoxins. CYP2D6 (encoding for debrisoquine-4-hydroxylase, a cytochrome-p450 enzyme) is the most frequently studied candidate gene for Parkinson's disease. Due to their function in detoxification, CYP2D6 was studied as an obvious candidate. Initial findings were positive, but could neither be replicated in individual studies nor in a several meta-analyses [22].

Candidate-gene studies have proved to be more successful when used as a follow-up of linkage or sib-pair studies. APOE*4, the most common genetic factor implicated in Alzheimer's disease, was primarily discovered by means of the candidate-gene approach following linkage analysis suggesting an Alzheimer's disease gene on chromosome 19 [23]. APOE was considered as a 'positional candidate', because its gene product, apolipoprotein E, was found to be associated with senile plaques in brains of patients with Alzheimer's disease. On the contrary, polymorphisms in genes for familial Parkinson's disease have not shown a consistent association with sporadic Parkinson's disease [22].

Bias

As mentioned in the section 'Methods in genetic epidemiology' on candidate-gene study design, false-positive findings may to a great extent be accounted for by multiple testing. Another problem leading to population bias is the phenomenon of population admixture. Whenever a distinct population comprises different subgroups with respect to genetic make-up, bias due to population admixture may occur both in follow-up and case–control studies. In a follow-up study, bias will occur if the population studied consists of subpopulations, which differ in risk of disease as well as genetic make-up. In a case–control study, cases and controls may have been drawn from different subpopulations. Bias due to population stratification can occur in any population study in which cases and controls are not matched for their genealogical history.

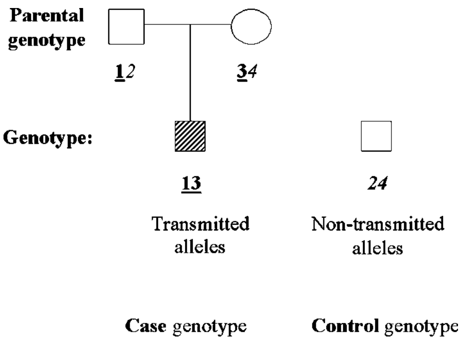

Figure 5. Principles of the TDT. Non-transmitted alleles of a patient are used to construct a 'virtual' control genotype. If associated with disease, an allele will be more frequently transmitted to the patient (dashed symbol) than to the control.

In order to overcome the problem of population admixture, the transmission–disequilibrium test (TDT) can be used. Originally, the TDT was used in family-based studies, but design is also applicable to patients derived from the general population [24]. Principles of the TDT are shown in Figure 5. Rather than ascertaining a control group, alleles of parents not transmitted to the patients can be used to construct a 'virtual' control genotype [25]. A disease-associated allele will be transmitted to the patient more frequently. The TDT approach requires ascertainment of DNA from the parental generation. Due to the fact that parents are often deceased, this approach is of limited value for late-onset disorders. A variant on the TDT using siblings rather than parents was developed [26], but this sib-TDT possessed considerably less power than the original TDT. By means of alternative strategies [27], the problem of admixture in classical epidemiologic study designs – such as case–control and follow-up studies can be overcome. Presently, it is only possible to test for presence of admixture, whilst new developments will make it possible to effectively adjust for distortions caused by admixture.

Another way to minimize the possibility of admixture is the use of genetically homogeneous populations. The population of a genetically isolated community originates from a limited number of ancestors (founders). Such a founder population limits the degree of genetic diversity introduced, leading to a more homogeneous population. Genetic drift, a random process occurring in small populations, further reduces the number of putative susceptibility genes in these populations.

In recent years, there is growing interest for genetic studies in homogeneous, genetically isolated, populations with the aim to identify new genes. In these isolates, there is a higher probability that patients have developed the disease due to a mutation inherited from a common ancestor. In contrast to studies in the general population, genome screens have proven to be useful in long-standing genetic isolates [13]. A population widely studied because of its genetic isolation is Finland. Finland has experienced isolation for over 100 generations, and expanded from a small group of founders into the 5 million inhabitants of today, resulting in a genetically homogeneous population [28]. Another example of a genetic isolate is Iceland, in which to date genes and susceptibility loci for several common disorders have been mapped [29–31].

In addition to studies of populations of prolonged isolation, studies in more recently isolated populations have been equally successful. In these populations, the founder effect is the major determinant of the limited genetic variation [32]. Using populations isolated up to 20 generations (approximately 300–400 years), loci associated with genetically complex disease have been identified including manic depression in Costa Rica [33] and susceptibility to mycobacterial infection in Malta [34]. A method applied in these studies successfully is that of haplotype sharing [32], as depicted in Figure 5.

The drawback of studies in isolated populations is the limited value when study results from genetic isolates are extrapolated to other populations. Isolation over 100 generations may have caused a population like Finland to have a more or less private make-up of the genome [28]. An advantage of studies in populations of more recent isolation is that genetic make-up of the isolated population is expected to more closely resemble that of the general population. However, it remains to be determined whether disease-related mutations or polymorphisms, even when detected in a recent isolate are also present in the general population.

Risk quantification

Population studies play a pivotal role in the assessment of risks associated with genetic factors. Such studies are needed to assess frequencies of mutations and polymorphisms, and their contribution to disease, this whilst regarding other, disease-modifying, risk factors. Basically, risk estimation for genetic factors follows the classical approach of epidemiologic studies. To estimate absolute risk, follow-up studies of carriers are needed [8]. Relative risks of disease may be derived from studies comparing risk of disease in carriers to that in non-carriers. Alternatively, relative risks may be estimated from case–control studies using incident patients derived from a single, strictly defined, study base [8].

For complex genetic disorders in which multiple genetic and environmental factors are involved, it is unlikely that a gene effect is independent of other risk factors. A key issue to resolve is interaction of different genetic and environmental factors implicated in the disease. Large-scale (clinical) epidemiologic studies are therefore needed to ascertain both data on genes and environmental factors, in order to determine the role of a mutation in the occurrence of disease in the population.

Population admixture is a problem to overcome when assessing risk. Although epidemiologic effect parameters can be deducted in a parent-control approach using the TDT, it will be more efficient to embed these studies in large ongoing epidemiologic studies. To evaluate risk estimates in these studies may possibly be prone to bias by population admixture, the test developed by Pritchard can be used [27].

Studies aiming to quantify risks associated with common polymorphisms are lagging behind developments in molecular genetics. Few studies have addressed the risk of Alzheimer's disease associated with the APOE*4 allele. After initial findings in 1991, most studies up to date have been based on prevalent, often clinic-based patients [7]. For clinical as well as public health purposes, unbiased risk estimates are essential. Therefore, follow-up studies in a cohort derived from the general population are more representative of absolute and relative risk of disease in carriers and non-carriers [35].

New developments

Genetic epidemiology

Major progress has been made in identifying genes involved in the genetic epidemiology of a large number of familial disorders including Huntington's disease and Mendelian forms of Alzheimer's and Parkinson's disease. Genetic research has made a significant contribution to the understanding of the molecular etiology of these diseases. Albeit outside the scope of genetic epidemiology, the uncovering of hitherto unknown proteins involved in pathogenesis has yielded important targets for drug development and further molecular-biological research.

On a population level, studies of these rare monogenic forms of common disease have only made a limited contribution to our understanding of the occurrence of disease in the general population. Mendelian genes can explain only a minor fraction of disease in the general population. The challenge for the near future for genetic–epidemiologic research will be the identification of genes involved in the aetiology of common late-onset disorders. With the shift in genetic–epidemiologic research from monogenic to complex disorders, design of its studies will change dramatically. As discussed earlier, studies of complex disorders will move from extended families towards affected sib pairs and the general population. This demands a shift of data collection towards ascertainment of large series of patients and affected siblings in order to reach sufficient statistical power in a study [20]. In order to cope with the requirement of large-scale genotyping, developments in genetic epidemiology heavily depend on those in molecular biology.

Molecular genetics

One of the most important developments in molecular genetics boosting genetic–epidemiologic research has been the publication of the sequence of the human genome [3, 4]. Consequently, an ever-increasing amount of information on genes and genetic variation in man is becoming available. A possibility created by availability of the human-genome sequence is that of addressing all human genes in a genome screen. This approach differs from the classical genome screen in addressing only genes and not untranslated sequences, which are considered to be mostly non-functional (95% of DNA) [20].

A technical development important for the feasibility of such large-scale genetic–epidemiologic research has been the introduction of microarrays, which include (binary) information on presence or absence of polymorphism in a gene. These technical devices create the opportunity to rapidly screen for DNA mutations or variations in large series of affected individuals [36, 37]. In this respect, the identification of single-nucleotide polymorphism maps (SNPs) throughout the human genome is crucial. Major advances in this field will mark the next decade.

Along with growing insight into the biology of disease, the availability of the human-genome map creates vast opportunities for candidate-genes studies. Although up to now, these studies have not yielded major breakthroughs, lack of success can be contributed at least partly to flaws in study design. An issue still much neglected is unbiased control selection. Major improvement of the design of candidate-gene studies can be achieved by embedding those studies into large epidemiologic cohort studies. In combination with functional studies of the polymorphism, candidate-gene studies may then re-capture a prominent position.

Public health

Perhaps the most dramatic change in genetic–epidemiologic research arising from the switch from studies of monogenic disorders to complex disorders are the implications in terms of clinical and public health. In contrast to the limited number of subjects at risk for monogenic disorders such as Huntington's disease and familial forms of Alzheimer's and Parkinson's disease, the clinical and public health implications in studies of complex genetic disorders are relevant for a large number of subjects. An important task for genetic epidemiologists will be to provide unbiased estimates. In order to reach unbiased risk assessment, follow-up studies comparing risks of disease in carriers and non-carriers are needed. Absolute and relative risks associated with a disease-associated mutation or polymorphism may largely depend on interaction with other genetic and environmental factors. To provide carriers with a valid estimate of their risk, it will be necessary to study the gene-associated risk in interaction with other genetic and environmental risk factors. To study these interactions with sufficient statistical power requires large series of patients and controls [38]. Given the need of information on a broad spectrum of exposures, there is a strong argument to include such studies in ongoing epidemiologic follow-up studies in which data on various genetic and environmental factors are ascertained simultaneously.

Acknowledgements

Drs Dekker and van Duijn are both supported by a grant of The Netherlands Organisation for Scientific Research (NWO).

References

1. Watson JD, Crick FHC. Molecular structure of nucleic acid. A structure for deoxyribose nucleic acid. Nature 1953; 171: 737–738.
2. Mendel G. Versuche uber pflanzenhybriden. In: Verhandl. Naturf. Verein. Brunn 4 (Abhandl.) 1866: 3–47.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860–921.
4. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291: 1304–1351.
5. Cruts M, Van Broeckhoven C. Molecular genetics of Alzheimer's disease. Ann Med 1998; 30: 560–565.
6. Slooter AJ, van Duijn CM. Genetic Epidemiology of Alzheimer's disease. Epidemiol Rev 1997; 1: 107–119.
7. Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: A meta-analysis. JAMA 1997; 349: 1349–1356.
8. Rothman KJ, Greenland S. Modern Epidemiology. Philadelphia, PA: Lippincott-Raven, 1998.
9. Van Broeckhoven C, Backhovens H, Cruts M, et al. Mapping of a gene predisposing to early-onset Alzheimer's disease to chromosome 14q24.3. Nat Genet 1992; 2: 335–339.
10. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 1997; 276: 2045–2047.
11. Kitada T, Asakawa S, Hattori N, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature 1998; 392: 605–608.

12. Leroy E, Boyer R, Auburger G, et al. The ubiquitin pathway in Parkinson's disease. Nature 1998; 395: 451–452.
13. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265: 2037–2048.
14. Lander E, Kruglyak L. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat Genet 1995; 11: 241–247.
15. Risch N. Linkage strategies for genetically complex traits. Am J Hum Genet 1990; 46: 222–253.
16. Sherrington R, Rogaev EI, Liang Y, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 1995; 375: 754–760.
17. Rogaev EI, Sherrington R, Rogaeva EA, et al. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. Nature 1995; 376: 775–778.
18. Bonifati V, Rizzu P, van Baren MJ, et al. Mutations in the DJ-1 Gene Associated with Autosomal Recessive Early-Onset Parkinsonism. Science 2002; 299: 256–259.
19. Svetkey LP, McKeown SP, Wilson AF. Heritability of salt sensitivity in black Americans. Hypertension 1996; 28(5): 854–858.
20. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996; 273: 1516–1517.
21. McCann SJ, Pond SM, James KM, Le Couteur DG. The association between polymorphisms in the cytochrome P-450 2D6 gene and Parkinson's disease: A case–control study and meta-analysis. J Neurol Sci 1997; 153: 50–53.
22. Tan EK, Khajavi M, Thornby JI, Nagamitsu S, Jankovic J, Ashizawa T. Variability and validity of polymorphism association studies in Parkinson's disease. Neurology 2000; 55: 533–538.
23. Pericak-Vance MA, Bebout JL, Gascell PC, et al. Linkage studies in familial Alzheimer's disease: Evidence for chromosome 19 linkage. Am J Hum Genet 1991; 48: 1034–1050.
24. Pericak-Vance MA. Linkage disequilibrium and genetic association. In: Haines J, Pericak-Vance MA, et al. Approaches to Gene Mapping in Complex Human Diseases. New York: Wiley-Liss, 1998; 323–333.
25. Ewens WJ, Spielman RS. The transmission/disequilibrium test: History, subdivision, admixture. Am J Hum Genet 1995; 57: 455–464.
26. Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. Am J Hum Genet 1998; 62: 450–458.
27. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999; 65: 220–228.
28. Peltonen L. Positional cloning of disease genes: Advantages of genetic isolates. Hum Hered 2000; 50: 66–75.
29. Stefansson H, Sigurdsson E, Steinthorsdottir V, et al. Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet 2002; 71: 877–892.
30. Kristjansson K, Manolescu A, Kristinsson A, et al. Linkage of essential hypertension to chromosome 18q. Hypertension 2002; 39: 1044–1049.
31. Hicks AA, Petursson H, Jonsson T, et al. A susceptibility gene for late-onset idiopathic Parkinson's disease. Ann Neurol 2002; 52: 549–555.
32. Houwen RH, Baharloo S, Blankenship K, et al. Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 1994; 8: 380–386.
33. Freimer NB, Reus VI, Escamilla M, et al. An approach to investigating linkage for bipolar disorder using large Costa Rican pedigrees. Am J Med Genet 1996; 67: 254–263.
34. Newport MJ, Huxley CM, Huston S, et al. A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection. N Engl J Med 1996; 335: 1941–1949.
35. Slooter AJ, Cruts M, Kalmijn S, Breteler MM, Van Broeckhoven C, van Duijn CM. Risk estimates of dementia by apolipoprotein E genotypes from a population-based incidence study: The Rotterdam Study. Arch Neur 1998; 55: 964–968.
36. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998; 280: 1077–1082.
37. Cheung VG, Gregg JP, Gogolin-Ewens KJ, et al. Linkage-disequilibrium mapping without genotyping. Nat Genet 1998; 18: 225–230.
38. Smith PG, Day NE. The design of case–control studies: The influence of confounding and interaction effects. Int J Epidemiol 1984; 13: 356–365.

Address for correspondence: C.M. van Duijn, Department of Epidemiology & Biostatistics, Erasmus MC, P.O. Box 1738, Rotterdam, The Netherlands
Phone: +31-10-4087394; Fax: +31-10-4089406
E-mail: c.vanduijn@erasmusmc.nl