Tuberculosis is a multifactorial disease in which the environment interacts with host-related factors to contribute to the overall clinical spectrum (1). In a companion paper (2), we described the design of a large and complex multicenter study currently being carried out in three countries in West Africa (The Gambia, the Republic of Guinea, and Guinea-Bissau) for investigation of these risk factors. We addressed a number of methodological issues arising in the design of the study.

There is an increasing body of evidence that genetic factors contribute to differences in host response to mycobacteria (such as *Mycobacterium tuberculosis*), affecting both susceptibility to infection and, more clearly, the pattern of clinical disease (3). Evidence for the general principle of a heritable response to tuberculosis comes from studies of twins. Because twins theoretically share the same environment, higher rates of concordance for monozygous (identical) twins than for dizygous (fraternal) twins suggest that genetic factors are important in susceptibility to tuberculosis (4, 5) and provide an estimate of the magnitude of this effect. Recently, numerous studies on the genetics of susceptibility to mycobacterial diseases other than tuberculosis have been carried out. A mutation in the interferon-γ receptor gene has been identified as the cause of increased susceptibility to disseminated nontuberculous mycobacterial disease in a Maltese kindred (6). Mutations in the interleukin-12 receptor gene have been found to be associated with impaired immune defense against mycobacteria and *Salmonella* in humans (7, 8). A survey of cases of disseminated Bacillus Calmette-Guérin infection following immunization in France revealed that one quarter of all cases were observed among offspring of consanguineous parents (9). There have been more numerous studies of susceptibility to *Mycobacterium leprae*, the causative agent of leprosy. Evidence of an association with human leukocyte antigen (HLA) class II gene variation, particularly the HLA-DR2 types, is well supported, and recently a major locus for susceptibility to leprosy in India has been mapped to some 10p13 (10).

For the past 10 years, investigations of tuberculosis susceptibility have focused mainly on the risks associated with
alleles of individual “candidate” genes, selected on the basis of their function. Thus, a possible association of the HLA-DR2 allele with susceptibility to tuberculosis was proposed (11–14). In mice, investigators have identified a gene (nramp) affecting resistance to some strains of Bacillus Calmette-Guérin that also controls resistance to leishmaniasis and Salmonella infection (15, 16) but probably not M. tuberculosis (17). Several polymorphisms (variants) have been described within the human homologue gene, NRAMP1, and it has been suggested that they could influence its function (18). A case-control study carried out in The Gambia showed that polymorphisms in the NRAMP1 gene were significantly associated with susceptibility to tuberculosis (19), although it was not formally possible to distinguish between susceptibility to infection with M. tuberculosis and susceptibility to disease progression. The effects of such genes may be modified by environmental (i.e., nongenetic) factors. For example, in a recent case-control study of Gujarati Asians in London, United Kingdom (20), it was proposed that a polymorphism in the vitamin D receptor gene combined with vitamin D deficiency could contribute to susceptibility to tuberculosis. However, when the data are correctly analyzed, they do not provide evidence of a significant interaction (21).

A further approach used to address genetic susceptibility to tuberculosis has been to screen the human genome with regularly spaced highly polymorphic genetic markers to search for chromosome regions that demonstrate linkage to the disease. If found, such linkage indicates that there is a disease susceptibility gene or genes in the neighborhood of the marker, and detailed investigation of genes in the region is indicated. Such a genome-wide scan of affected sibling pairs from The Gambia and South Africa identified potential susceptibility loci on chromosomes 15q and Xq (22).

Thus, genetic factors might play a role in susceptibility to tuberculosis, although their level of action and the physiologic pathways remain to be fully understood, as does their relative importance, given the large role environmental factors play in the incidence of the disease (1). In this paper, we outline the components of our study that are designed to investigate the effect of host genetic factors on tuberculosis development, and we address a number of somewhat neglected methodological issues that arise, such as the effects of consanguinity, half-siblings, and nonpaternity, as well as assessment of gene-environment interaction. Consideration of these issues may be useful in the design of studies of genetic susceptibility to other diseases, particularly studies to be carried out in developing countries.

OVERVIEW OF THE STUDY’S DESIGNS

The objectives of the genetic studies are to determine the risks associated with allelic variants of candidate genes, to map more finely regions of the genome for which there is evidence of linkage to disease susceptibility (“candidate regions”), and to contribute cases to a genome-wide search for susceptibility genes. Within the overall project described in the companion paper (2), three studies are being conducted in pursuit of these objectives: a case-control study, a family-based association study, and a linkage study of affected relatives.

Case-control study

In the case-control study, the frequencies of specific alleles at a polymorphic candidate locus are being compared between affected individuals and unrelated unaffected controls, using a community case-control study design. Candidate genes of interest include various cytokines and other mediators known to be important in the pathogenesis of tuberculosis, such as NRAMP1, vitamin D3 receptor, interferon-γ, interleukin-1β, interleukin-12, tumor necrosis factor-α, interleukin-4, and interleukin-10, their receptors, and the transcription factors that regulate their expression. Details on the design of the case-control study and associated methodological issues are provided in the companion paper (2).

Family-based association study

In the family-based association study, parents or siblings of cases will be genotyped, and an extension of the transmission disequilibrium test (23) will be used to determine whether an allele of interest is transmitted from each parent to an affected offspring more often than would be expected by chance. Such a family-based study tests for an association between an allelic variant of the candidate gene or marker and the disease, while avoiding the risk of population stratification or confounding by ethnic group, to which a standard case-control study is vulnerable.

Linkage study of affected relatives

In the linkage study, data from families with more than one affected member will contribute, along with similar data from other studies, to a genome-wide screen designed to identify chromosomal regions linked to tuberculosis. This study will use the nonparametric form of linkage analysis based on the frequency with which affected relatives share alleles of genetic markers (24).

METHODOLOGICAL ISSUES

Abel and Dessein (25) have reviewed the use of genetic linkage and association studies in the investigation of genetic susceptibility to infectious diseases. Implementation of such studies in developing countries presents some particular challenges.

Family-based association study

If an allelic variant of a candidate gene or of a genetic marker were associated with increased risk of disease, one would expect that variant to be transmitted from a heterozygous parent to an affected offspring more often than the 50 percent frequency expected by chance. In the simplest case of a biallelic locus, let $b$ be the number of times the A1 allele
is transmitted from a heterozygous $A_1A_2$ parent to an affected offspring and $c$ the number of times the $A_2$ allele is transmitted; then $(b - c)^2/(b + c)$ is distributed as chi-squared with 1 df. This is known as the transmission disequilibrium test (22), and it will be recognized as McNemar’s test for a pair-matched case-control study. In this situation, the “matched pair” is the parental alleles: The “case” is the transmitted allele and the “control” is the non-transmitted allele. The transmission disequilibrium test tests for an association (disease risk is increased among persons carrying a particular allele of the candidate gene or marker) and for linkage (the “disease gene” is near the candidate gene or marker on the same chromosome) at the same time; hence, it avoids the possible risk of population stratification or confounding by ethnic group or other endogamous subgroup, to which a standard case-control study may be vulnerable despite precautions in the design.

Since tuberculosis occurs mainly in adults, parents of a case are frequently unavailable for genotyping. Alternative approaches using unaffected siblings as controls are possible (26). Such comparisons protect against population stratification, but in comparison with the same number of unrelated controls, sibling controls have reduced statistical power: A case-control pair in which the case and the control carry the same allele provides no information for determination of the risk associated with that allele, and this is much more likely to occur for siblings than for unrelated controls.

The use of these tests in West Africa is complicated by three cultural features that are more common in Africa than in “Northern” societies, where the tests have generally been used up to the present: consanguinity, the prevalence of half-siblings, and nonpaternity.

**Consanguinity.** Consanguineous marriages are common in West Africa. Preliminary data from our study indicate that approximately 30 percent of marriages in The Gambia take place between first cousins. This affects the distribution of parental genotypes and the probability of transmission from heterozygous parents to an affected offspring. Bennett and Curnow (27) have shown that the transmission disequilibrium test is robust under these circumstances: The type I error probability is essentially unchanged, except for a purely recessive disease allele. For a candidate gene, the power of the test is increased for a recessive allele and decreased for a dominant allele, whereas for a genetic marker the effects are likely to be negligible.

**Half-siblings.** The practice of polygamy in West Africa means that paternal half-siblings are more common than full siblings. Genetically speaking, half-siblings are midway between full siblings and unrelated controls. Thus, unaffected half-siblings may be compared with cases in the same way as unaffected full siblings, but they will generally have greater statistical power, since they will share fewer genes with the case. Furthermore, because of cultural habits and social mobility, children in West Africa frequently live not with their genetic parents but rather with relatives (the maternal uncle, for instance), which complicates the identification of “real” siblings and the drawing of pedigrees. The study of such foster-relationships may help researchers to differentiate genetic from environmental sources of household clustering of disease.

**Apparent nonpaternity.** As noted above, it is relatively common in West Africa for children to live with another member of the family, not the biologic parents. There is also the possibility of a second marriage in which the second husband adopts the children from the first marriage. The effect of such apparent nonpaternity (or more conventional nonpaternity in which the social father is not the genetic father) on tests of transmission from parents has been shown to be small (28). The effect of nonpaternity on sibling-based comparisons is simply that some unaffected “full siblings” of cases will in fact be unrecognized half-siblings, which will increase the power of the comparisons, as described above.

We adopted the strategy of genotyping the parents of cases whenever possible. In situations where both parents are not available, two unaffected full siblings are typed, and if full siblings are not available, then half-siblings are typed. Since the planning of this study, there have been rapid new developments in the statistical analysis of association studies, and we intend to implement new likelihood-based methods of analysis which offer possibilities for the straightforward study of gene-environment interactions and the determination of multilocus haplotypes (29).

The validity of these tests relies heavily on the accuracy of reported genealogy. The drawing of pedigrees is complicated considerably by consanguinity, polygamy, and the wider understanding of the concept of “sibling” in Africa than in Western society. Thus, field-workers will be carefully trained in how to inquire about genetic relatedness, with special emphasis on probing for determination of biologic parentage and for whether relationships are “full” or “half.” Field-workers will be provided with a coded list of possible relationships. The need to respect confidentiality will be also emphasized. Addresses of relatives not living in the household will be collected so we can trace them for genetic sampling if required.

**Linkage study**

Association studies permit the study of risk associated with specific alleles of a candidate gene, and they also may be used for the fine mapping of a small candidate region by linkage disequilibrium. Because of genetic recombination, allelic association between genetic markers and disease breaks down over larger distances. The use of such studies for mapping large regions of the human genome or the entire genome is now becoming possible using information on tens or hundreds of thousands of biallelic single nucleotide polymorphisms (30), although the number that will be needed is still a subject of debate (31–33).

At present, genetic mapping of the entire genome is possible only through linkage studies, which determine whether a hypothetical disease gene is on the same chromosome as a given genetic marker (“linked”) and which, in some cases, estimate the distance between the marker and the gene. Such studies are popularly carried out by examining the proportion of marker alleles shared by pairs of affected siblings.
Gene-environment interaction

In the study of complex diseases, consideration of a possible interaction between genotype and environment is of growing importance, and recent technological developments allow researchers to address more complex questions. Gene-environment interactions may be explored not only in conventional epidemiologic study designs, such as cross-sectional, cohort, and case-control studies, but also in family-based designs, such as case-parent, affected sibling-pair, and twin studies (34, 35). Khoury and Flanders (36) have proposed the use of a case-only design to study interaction effects, but interpretation of the results from such a design rests crucially on the assumption of independence between environmental exposure and genotype in the population, and this design does not allow one to estimate the main effects of genotype or environmental exposure.

Gene-environment interaction may be characterized as the genetic control of sensitivity to environmental exposure. In the present case-control study, environmental risk factors and genetic markers are being examined together as independent predictors of disease and as interacting factors. To avoid population stratification due to variability in either disease genotype or marker frequencies in population subgroups, it is advisable either to match carefully on ethnic or geographic background or to adopt internal family controls. In the present study, household-level variables (such as crowding, hygiene, and socioeconomic status) are measured in cases and community controls, while data on individual-level variables (such as medical history, smoking status, and intercurrent disease) are recorded in cases and within-household controls. In both situations, the interaction of these variables with candidate genes will be investigated, and any interaction found may be explored further in the family-based association study.

Household contacts of each case will be followed up for 2 years to determine the occurrence of secondary cases of tuberculosis. Baseline measures of genetic and geographic proximity to the case will be used to determine the relative importance of each factor in the risk of secondary disease. In principle, it would also have been possible to investigate gene-environment interaction in this longitudinal aspect of the combined study, by measuring both genetic markers and environmental risk factors in all members of the case’s household (i.e., the exposed) and all members of the community control’s household (i.e., those unexposed to tuberculous infection). However, for ethical and practical reasons, we did not include this element in the study. Although some household contacts will be genotyped, we do not expect the number of cases observed to provide adequate power for detection of any interaction within the cohort study.

SAMPLE SIZE AND POWER

An assessment of Tuberculosis Control Programme activities in the three study countries indicates that it is feasible to recruit 800 cases over a period of 18 months. This will give us 80 percent power to detect an odds ratio of at least 1.6 for any genotype with a prevalence between 10 percent and 85 percent in the controls and to detect an odds ratio of at least 1.4 for any genotype with a prevalence between 20 percent and 70 percent.

In calculating the power of the family-based association study, we must consider the mode of inheritance of the putative disease gene, which is best explained in terms of the genotype relative risk. Considering, as before, the simplest case of a locus with two alleles ($A_1, A_2$), there are three possible genotypes: $A_1A_1$, $A_1A_2$, and $A_2A_2$. If the $A_2$ allele is associated with a genotype relative risk of 2, this may be expressed in terms of the risk for the three ordered genotypes in four commonly used models of inheritance: recessive (1:1:2), dominant (1:2:3), additive (1:2:4), and multiplicative (1:2:4). Sample-size calculation for the family-based association study must also take into account the high rate of consanguinity in this population and the problems involved in tracing the biologic parents. We assume that both parents can be traced for only 50 percent of cases and that 25 percent of parental marriages are marriages between first cousins. The 800 cases recruited will then give us more than 90 percent power to detect an association at the 5 percent level of disease with an allele of a candidate gene with an allele frequency of at least 0.1 and a genotype relative risk of 2 or more, under a dominant, additive, or multiplicative model (27). Under a recessive model, a genotype relative risk of 5 will be detectable.

In epidemiologic studies with a given sample size, power to detect statistical interaction is usually lower than the power to detect main effects, and the variance of the interaction estimate will be greater than the variance of the
main-effects estimate under a no-interaction model (37, 38). In case-control studies, in the presence of a relatively rare susceptibility genotype and in a situation where exposure has a larger effect on disease risk than does genotype alone, a large number of cases and controls is needed to detect gene-environment interaction (39). In our situation, sample size calculation shows that 800 cases and controls will be sufficient to obtain (for example) 80 percent power to detect at the 5 percent significance level an interaction odds ratio of 3 between an environmental risk factor with a prevalence of 30 percent and a gene with a prevalence between 0.05 and 0.85, where the environmental factor and the gene each have an odds ratio of 1.5 in the absence of the other (40).

The linkage study depends on recruiting people with secondary cases of tuberculosis during the 2-year follow-up of the case and control households. Even in such a large prospective study, the number of secondary cases to be expected is not great. The incidence of active tuberculosis in the general population is estimated to be approximately 0.1 percent per year (41). In the control households, with an average of 10 members per household, follow-up of 800 households for 2 years will be expected to yield only 13 cases, assuming 20 percent loss to follow-up. Preliminary results indicate an incidence among members of case households of 1.4 percent per annum, such that we can expect to detect approximately 160 secondary cases in these households. Since most of the secondary cases will be relatives of the index cases, they will contribute one affected pair to the allele-sharing study.

CONCLUSION

In this paper, we have described the range of methods we intend to use in a multicenter study of relations between environment, genes, and tuberculosis in West African countries. Although such studies are usually conducted in developed countries for investigation of non-communicable diseases, we have addressed theoretical aspects of their use in the study of infectious disease, as well as the methodological implications of carrying out such studies in developing countries in the presence of consanguinity, polygamy, and apparent nonpaternity. We believe that studies such as these can contribute usefully to a better understanding of the natural history of infectious diseases in humans and to the search for new methods of prevention and treatment.

ACKNOWLEDGMENTS

This study is funded by the UK Medical Research Council and the European Commission (DGXII, contract IC18CT980375).

The authors thank Drs. Robin Bailey, Richard Bellamy, Arnaud Marchant, Jane Rowley, and Peter Smith for their contribution to the study’s development.

REFERENCES

24. Kruglyak L, Daly MJ, Reeve-Daly MP, et al. Parametric and


