Genetic susceptibility to infectious disease

Shelley Segal and Adrian V.S. Hill

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

Our understanding of the variation in individual clinical responses to pathogens has become increasingly relevant, particularly in the face of new emerging epidemics as well as the increasing number of multi-drug-resistant organisms. An effective immune response to infection has contributed to the development of host genetic diversity through selective pressure, with an increasing number of studies characterizing the role that host genetics plays in disease susceptibility. Knowledge of the role host mechanisms play in the pathogenesis of infectious disease can contribute to the design of new therapeutic strategies. Rapid advances in the field of human genomics offer great opportunities for adopting this approach to find new insights into pathogenesis.

Microorganisms are continually interacting with and responding to their environment and are able to alter their pattern of gene expression rapidly in response to several environmental signals. This continual adaptation leads to enormous variation in the host–microbe interaction. The vast majority of such interactions have evolved to be of benefit to both microbe and host so that symbiosis develops. The development of disease is an exception to this rule and results from a complex interaction between the microbe, the host and the environment. Interest in the role that genetic variation within the host plays in susceptibility to infection has arisen from both the individual severity of disease and its progression, as well as infection rates on a population level. Therefore, genetic associations might therefore provide clues to fundamental questions about the pathogenesis of an organism and the host response.

Probably, the best-known example of such genetically determined factors is found in malaria and the haemoglobinopathies [1,2]. Heterozygosity for sickle haemoglobin is strongly associated with protection against death and severe disease, probably owing to decreased growth of the parasite in infected erythrocytes. Both the thalassaemias and G6PD deficiencies have been shown to protect the host against malaria. The absence of the Duffy antigen on the surface of red blood cells is known to prevent the entry of *Plasmodium vivax* into erythrocytes. The outcome of infection with tuberculosis, and also human immunodeficiency virus (HIV), was subsequently correlated with polymorphisms in other genes [3]. Further evidence of the selective pressure of infectious agents can be demonstrated by the higher level of polymorphisms seen in the human leukocyte antigen (HLA) region when compared with other regions in the human genome. This region encodes several proteins involved in the immune response, including complement and tumor necrosis factor (TNF)-α. The HLA complex consists of Class 1 and Class 11 alleles, which present foreign antigens to CD4 helper and CD8 cytotoxic T cells, respectively. HLA variation has been associated with several viral diseases, including delayed progression of HIV infection and resistance to Hepatitis B carriage [4,5]. These effects are associated with HLA heterozygosity [6] and possibly contribute to the maintenance of HLA diversity.

Measuring the genetic effect on infection

In order for an infectious disease to cause selective genetic pressure on host allele frequencies it would need to have a significant effect on morbidity and mortality, before reproductive age, over a long period of time. Genes encoding many monogenic diseases have been mapped and identified using techniques such as positional cloning. Similar techniques are now also used in the mapping of multifactorial diseases where there is a complex genetic component. These include infectious as well as non-infectious diseases where inheritance is not Mendelian because several factors could lead to disease manifestation. Recent genetic tools, in particular the newly available entire human genome sequence (www.ensembl.org), have provided detailed genetic and physical maps that have increased our ability to find variations in the human genome and correlate them with disease.

Estimates of the part of the host's genetic contribution to disease manifestation or progression that is relatively independent of environmental effects can be obtained from well-designed studies in adoptees or twins. Numerous comparisons have been made based on concordance for particular infectious diseases between identical and non-identical twins. If a disease has a genetic component, monozygotic twins who share 100% of their genes will show higher concordance. This concordance has been shown in several infectious diseases including tuberculosis, leprosy and chronic hepatitis B infection [2,7]. Various approaches can be taken to estimate the extent of the host genetic contribution to variation in susceptibility to infectious disease. A sibling risk ratio (λ_S) can be estimated. This describes the increase in the risk of developing disease in a sibling of a diseased individual

Corresponding author: Shelley Segal (shelley.segal@paediatrics.ox.ac.uk).

 0966-842X/$ - see front matter doi:10.1016/S0966-842X(03)00207-5

compared with that of the general population. In autoimmune diseases such as Type 1 diabetes and multiple sclerosis this is in the order of 15–20. The λ_S in infectious diseases appear to be lower: in the order of 1.5–5. These values might overestimate the genetic component of susceptibility because of the increased risk of exposure to infection within family members compared with the general population. Such λ_S estimates can be made from epidemiological surveys or from twin studies and are of value in providing an upper limit to the likely genetic contribution, but are population dependent.

Adoption studies are another method of controlling for environmental variation in genetic studies. Sorensen *et al.* reported a large study of causes of premature death in 960 families, where children were adopted early in life [8]. The risk of these adoptees dying from specific causes was assessed. Adoptees with biological parents who had died before the age of 50 from an infectious disease had an increase in the relative risk of dying from the same cause of 5.8. By contrast, the death of an adoptive parent from an infectious disease had no significant effect on the adoptees' risk of such a death. There was also a strong genetic effect on death resulting from cardiovascular disease, but a much smaller genetic effect on deaths from cancer [8].

Another approach in examining possible host genetic effects on infectious agents is to search for inter-ethnic differences in disease prevalence or severity among different population groups that have been exposed to the same infectious agent. The Fulani population are more resistant to malaria parasitaemia than other West African ethnic groups despite similar exposure, which might be related to increased specific antibody responses [9].

Identifying the genetic component

Susceptibility to infectious disease seldom follows a simple pattern of Mendelian inheritance, except in a few rare familial cases involving single gene defects. Variations in the genes coding for interleukin (IL)-12, its receptor and interferon gamma receptor 1 (IFNGR1) have been associated with increased susceptibility to atypical mycobacterial infection and Salmonella [10]. However, most infectious diseases follow a complex mode of inheritance. There are two main approaches to mapping and identifying genetic associations in complex disease. These involve either linkage or association studies (Fig. 1) or both.

Association studies are statistically powerful in testing for genes that have small contributory effects. They are more sensitive than linkage studies as the collection of case–control series is often easier than multi-case families. However, association studies require assessment of polymorphisms in or close to the relevant causal change in a candidate gene.

Candidate genes encode products that are hypothesized to influence susceptibility to, or the course of, a particular disease. To determine if a gene contributes to the disease phenotype, its polymorphisms are examined for association in case–control or family studies. A candidate gene may be proposed if biological information suggests that its products can influence the known pathogenesis. Such an approach was used in the study of HIV. The chemokine receptor CCR5 was shown to act as a co-receptor for HIV entry into the CD4 cell and the gene was screened for polymorphisms. Individuals that were homozygous for a 32-base-pair deletion CCR5Δ32 were also resistant to HIV infection, and individuals heterozygous for the deletion had a delayed progression to AIDS [11]. Other methods for proposing candidate genes include studying homologous genes identified in animal models, or genes that have been implicated in a similar disease. For example natural-resistance-associated macrophage protein (NRAMP1 or SLC11A1) was initially examined for tuberculosis susceptibility because it is the human homologue of the mouse gene that is partly responsible for susceptibility to intracellular pathogens [12–14]. Limitations to these

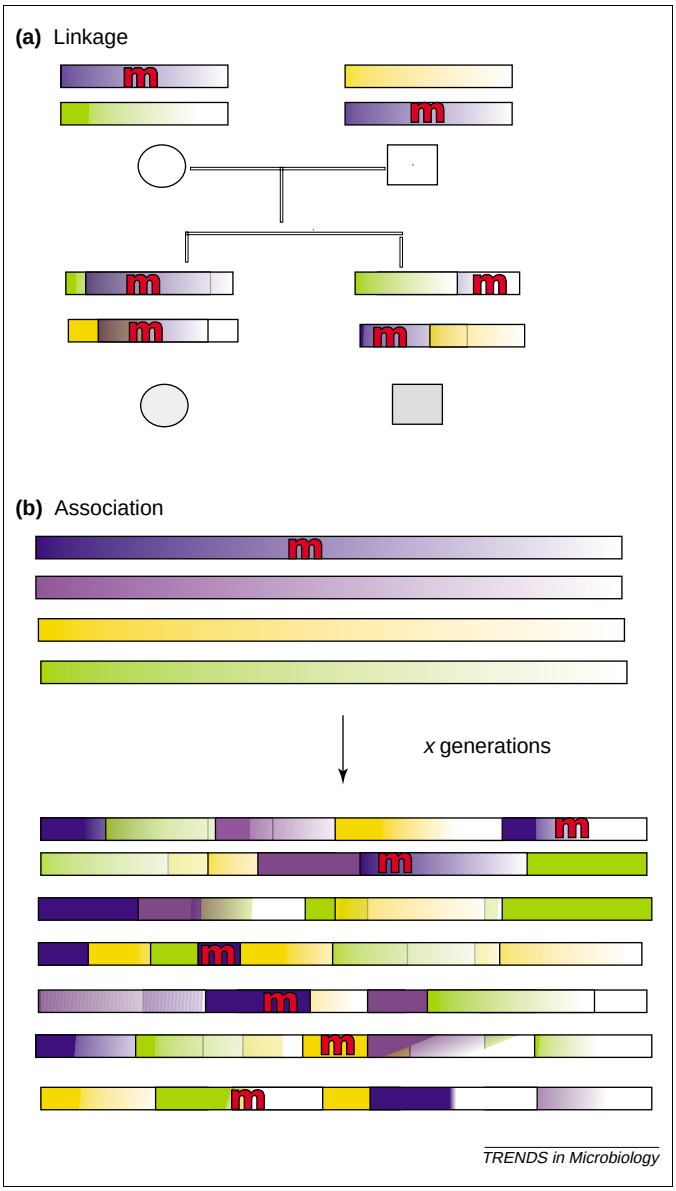

Fig. 1. Linkage versus association studies. Flanking marker diversity in linkage and association studies is shown. The functional mutation m occurs on a particular haplotypic background. In linkage or family studies there is little time for recombination to occur so the range of flanking markers that can be used to track the disease gene region will be large. Association studies concentrate on populations that have already undergone random mating over a long period of time, therefore the resulting disease gene region will be small. Over time the disease gene or mutation becomes separated from genes with its original haplotypic background, although some can remain together; this non-random association of alleles forms the basis for linkage disequilibrium. Adapted from [16].

methods include the need for prior identification of a gene, and the need to identify polymorphisms, preferably ones of functional significance.

Association studies can be designed between cases and unrelated controls, or within families. The most commonly used within-family association test uses an affected offspring and both parents to test if a particular allele is transmitted more frequently than would be expected under Mendelian inheritance. This is known as the transmission disequilibrium test (TDT).

Case–control differences in disease prevalence might suggest a role for genetic factors in disease but they rely on the recruitment of large numbers of both cases and controls. They are also prone to error because of inadequate matching and population stratification and need to be well matched in terms of environment and ethnicity; matching for age and sex are less important, as allele frequencies seldom differ between sexes or change much with age. Large sample sizes are usually required for these studies so that sufficient power is obtained to detect effects in the order of a two-fold change in risk. Case–control association studies have become increasingly popular as more polymorphisms have been identified in genes that are considered to have important roles in pathogenesis or protection. Several studies have shown an association for infectious disease susceptibility with loci in both the major histocompatibility complex (MHC) and other regions of the genome. A small selection of these are listed in Tables 1 and 2 and all are available through OMIM (www.ncbi.nlm.nih.gov) and Genecard (http://bioinformatics.weizmann.ac.il/cards/).

With recent advances in the human genome project and high throughput genotyping techniques, including mass spectrometry, genome-wide association studies are becoming more feasible, although still daunting in terms of cost and scale. These studies rely on the presence of linkage disequilibrium (LD) between markers and the disease-causing polymorphism. LD describes the non-random association of alleles that occur together at neighbouring loci to form haplotypes with a frequency greater than would be expected by chance. If a disease-associated polymorphism and another marker are in LD then the marker should be found more frequently in those with disease [15,16]. New single nucleotide polymorphisms (SNPs) are being identified at an increasingly rapid rate owing to the advances in bioinformatics and sequencing technology and the continual updating of the human genome sequence. These, together with advances in genotyping technology, mean that genome screens by association are becoming increasingly feasible. The possibility of identifying novel genes and gene families, as well as finding associations in genes that might not have been identified previously using the candidate gene approach, becomes more practicable. The number of SNPs that are needed to scan the genome by association ranges from thousands to millions. With SNP markers now being identified so rapidly through the comparison of sequences in EST and genome sequence databases, and compiled through the SNP consortium (http://snp.cshl.org/) and NCBI (http://www.ncbi.nlm.nih.gov:80/entrez/SNP), the challenge is to recruit large enough sample sizes to correct for multiple comparisons. Multi-centre cooperation could prove most productive. However, which SNPs to genotype initially for these large association studies remains an issue; the selection of genome-rich areas has been suggested together with the identification of phase-known genotypic data, which might identify recombination hotspots. The Haplotype Map (http://www.genome.gov/10001688) project has been anticipated; this would involve a collective effort for the identification of defined haplotypes within different population groups. The questions then posed by multiple genetic effects and their epistatic interactions would remain a constant challenge in this ever-evolving field.

Table 1. Association between HLA alleles and infectious disease. Adapted from [20]

MHC allele	Disease	Effect	Refs
Class1			
B8	Pulmonary TB	Susceptibility	[21]
B35	AIDS	Susceptibility	[4]
B53	Severe malaria	Resistance	[22]
B57	AIDS	Resistance	[23]
Class11			
DRB1 * 1302	Hepatitis B	Resistance	[5]
DRB1 * 1352	Malarial anaemia	Resistance	[22]
DRB1 * 1101	Hepatitis C	Resistance	[24]
DRB180dr24	Typhoid fever	Susceptibility	[25]
DR2	Leprosy	Susceptibility	[26]

Another approach to identify candidate genes is a genome-wide linkage analysis, which uses approximately 300 to 400 microsatellite markers spaced evenly across the genome, to identify regions that are shared by affected relatives and involves the recruitment of families with more than one affected child. The method involves a systematic search through the genome looking for the non-random segregation of microsatellite markers from parent to child. If the genetic marker and disease are linked with a disease phenotype, the affected children will have a greater likelihood of sharing markers than would be expected by chance. If this is observed in a sufficient number of families that are affected by disease, the observed segregation is non-random and deviates from what is expected under Mendel's law of independent assortment. A statistical analysis is then used to estimate the likelihood that the observed linkage has occurred by chance. Because the region of linkage is often very large and can span many megabases, fine mapping using more closely spaced markers decreases the identified interval further. Polymorphic genes within these regions of linkage are then screened for association. One advantage of this approach is that previously undefined genes, or genes that might not have been implicated in the specific disease biology, are investigated. This method is relatively insensitive at detecting small genetic effects whereas association studies are more powerful. However, significant linkage has been found in several infectious diseases, including leprosy and schistosomiasis [17–19].

Conclusions

The study of host factors and their genetic contribution to infectious diseases addresses fundamental issues in our understanding of infectious disease pathogenesis. Twin-, adoptee-, family- and case–control studies have indicated a

Table 2. Association between non-HLA regions and infectious disease. Adapted from [20]

Gene	Variant	Disease	Population	Effect	Refs
α-globin	Thalassaemia	Malaria	SW pacific	resistance	[27]
β-globin	Sickle cell	Malaria	Liberia	resistance	[28]
G6PD	Deficiency	Malaria	Gambia	resistance	[29]
TNF	Promoter-308	Malaria	Gambia	susceptibility	[30]
TNF	Promoter-238	Leprosy	India	susceptibility	[31]
NRAMP1	5′, 3′ variants	TB	Gambia	susceptibility	[13]
IFNGR	Various	Disseminated BCG	Malta	susceptibility	[32]
CCR5	32 bp deletion	HIV	Caucasian	resistance	[33]
CCR2	Codon 64	HIV	Caucasian	resistance	[34]

genetic basis for inter-individual differences in infectious disease manifestations. This allows for further functional and immunological analyses based on the knowledge gained through such genetic studies. The association of candidate gene polymorphisms in several diseases, including HIV, tuberculosis and malaria, has stimulated research for possible new drug and vaccine interventions. Individuals who present with either undefined sepsis or known specific infection, or who are at particular risk for disease, might be identified using rapid genotyping techniques and targeted treatment used accordingly. For example, IFN-γ and mannose-binding lectin could be considered in individuals with known deficiencies. Infection is still responsible for a considerable degree of worldwide mortality and morbidity despite the widespread use of antimicrobials and the availability of intensive care support. Improved understanding of the mechanisms of disease that result from complex interactions between the genetically variable host and microbe has been made possible through advances in high throughput genotyping techniques and bioinformatics. In this post-genomic era much more should be learnt about the basic principles governing microbe–host interactions, and new opportunities for therapeutic intervention will be realized.

References

1 Weatherall, D.J. *et al.* (1988) Genetic factors as determinants of infectious disease transmission in human communities. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 321, 327–348

2 Hill, A.V. (2001) Immunogenetics and genomics. *Lancet* 357, 2037–2041

3 Hill, A.V. (1996) Genetic susceptibility to malaria and other infectious diseases: from the MHC to the whole genome. *Parasitology* 112(Suppl), S75–S84

4 Carrington, M. *et al.* (1999) HLA and HIV-1: heterozygote advantage and B * 35-Cw * 04 disadvantage. *Science* 283, 1748–1752

5 Thursz, M.R. *et al.* (1995) Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia. *N. Engl. J. Med.* 332, 1065–1069

6 Thursz, M.R. *et al.* (1997) Heterozygote advantage for HLA class-II type in hepatitis B virus infection. *Nat. Genet.* 17, 11–12

7 Hill, A.V. (1996) Genetics of infectious disease resistance. *Curr. Opin. Genet. Dev.* 6, 348–353

8 Sorensen, T.I. *et al.* (1988) Genetic and environmental influences on premature death in adult adoptees. *N. Engl. J. Med.* 318, 727–732

9 Modiano, D. *et al.* (1996) Different response to *Plasmodium falciparum* malaria in West African sympatric ethnic groups. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13206–13211

10 Ottenhoff, T.H. *et al.* (2002) Genetics, cytokines and human infectious disease: lessons from weakly pathogenic mycobacteria and salmonellae. *Nat. Genet.* 32, 97–105

11 McNicholl, J.M. *et al.* (1997) Host genes and HIV: the role of the chemokine receptor gene CCR5 and its allele. *Emerg. Infect. Dis.* 3, 261–271

12 Bellamy, R. (1999) The natural resistance-associated macrophage protein and susceptibility to intracellular pathogens. *Microbes Infect.* 1, 23–27

13 Bellamy, R. *et al.* (1998) Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *N. Engl. J. Med.* 338, 640–644

14 Awomoyi, A.A. *et al.* (2002) Interleukin-10, polymorphism in SLC11A1 (formerly NRAMP1), and susceptibility to tuberculosis. *J. Infect. Dis.* 186, 1808–1814

15 Abecasis, G.R. *et al.* (2001) Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* 68, 191–197

16 Cardon, L.R. and Bell, J.I. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99

17 Marquet, S. *et al.* (1996) Genetic localization of a locus controlling the intensity of infection by *Schistosoma mansoni* on chromosome 5q31-q33. *Nat. Genet.* 14, 181–184

18 Siddiqui, M.R. *et al.* (2001) A major susceptibility locus for leprosy in India maps to chromosome 10p13. *Nat. Genet.* 27, 439–441

19 Mira, M.T. *et al.* (2003) Chromosome 6q25 is linked to susceptibility to leprosy in a Vietnamese population. *Nat. Genet.* 33, 412–415

20 Cooke, G.S. and Hill, A.V. (2001) Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* 2, 967–977

21 Selby, R. *et al.* (1978) Tuberculosis associated with HLA–B8, BfS in a Newfoundland community study. *Tissue Antigens* 11, 403–408

22 Hill, A.V. *et al.* (1991) Common West African HLA antigens are associated with protection from severe malaria. *Nature* 352, 595–600

23 Kaslow, R.A. *et al.* (1996) Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat. Med.* 2, 405–411

24 Thursz, M. *et al.* (1999) Influence of MHC class II genotype on outcome of infection with hepatitis C virus. The HENCORE group. Hepatitis C European Network for Cooperative Research. *Lancet* 354, 2119–2124

25 Dunstan, S.J. *et al.* (2001) Genes of the class II and class III major histocompatibility complex are associated with typhoid fever in Vietnam. *J. Infect. Dis.* 183, 261–268

26 Visentainer, J.E. *et al.* (1997) Association of leprosy with HLA-DR2 in a Southern Brazilian population. *Braz. J. Med. Biol. Res.* 30, 51–59

27 Allen, S.J. *et al.* (1997) alpha + -Thalassemia protects children against disease caused by other infections as well as malaria. *Proc. Natl. Acad. Sci. U. S. A.* 94, 14736–14741

28 Willcox, M. *et al.* (1983) Falciparum malaria and beta-thalassaemia trait in northern Liberia. *Ann. Trop. Med. Parasitol.* 77, 335–347

29 Ruwende, C. *et al.* (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* 376, 246–249

30 McGuire, W. *et al.* (1994) Variation in the TNF-alpha promoter region associated with susceptibility to cerebral malaria. *Nature* 371, 508–510

31 Roy, S. *et al.* (1997) Tumor necrosis factor promoter polymorphism and susceptibility to lepromatous leprosy. *J. Infect. Dis.* 176, 530–532

32 Newport, M. *et al.* (1995) Evidence for exclusion of a mutation in NRAMP as the cause of familial disseminated atypical mycobacterial infection in a Maltese kindred. *J. Med. Genet.* 32, 904–906

33 Dean, M. *et al.* (1996) Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* 273, 1856–1862

34 Easterbrook, P.J. *et al.* (1999) Chemokine receptor polymorphisms and human immunodeficiency virus disease progression. *J. Infect. Dis.* 180, 1096–1105