Genetics of infectious diseases

Angela J. Frodsham and Adrian V.S. Hill*

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

Received June 22, 2004; Revised and Accepted July 19, 2004

Infectious diseases represent a major health problem worldwide, both in terms of morbidity and mortality. A complex combination of environmental, pathogen and host genetic factors plays a role in determining both susceptibility to particular microbes and the course of infection. Numerous studies have now mapped and identified relevant genes using a variety of both family-based and population-based approaches. Much interest has been focused on susceptibility to malaria, HIV/AIDS and mycobacterial infection, but other bacterial, viral and parasitic diseases are receiving increasing attention. Some major genes have been identified by genome scans of multi-case families, and mouse genetics has contributed to mapping and identification of a few genes. However, the great majority of known susceptibility loci emerged from screening of likely candidate genes. The emerging picture is of highly polygenic diseases, with occasional major genes, along with significant inter-population heterogeneity. This genetic architecture likely reflects the role that evolutionary selection has played in generating and maintaining a diverse repertoire of susceptibility/resistance loci, most with individually small effects. Genome-wide association studies with large sample sizes will be required to define the majority of the relevant polygenes.

Infectious diseases account for a major part of the global health problem, with most of the burden falling in developing countries. Around 14.5 million deaths in the year 2001 were attributable to the effects of infectious diseases. Increasing evidence is becoming available to help define the role of host genetics in susceptibility to, or outcome of infectious diseases. Owing to the nature of infectious diseases and the complex immune response that ensues after exposure, it is likely that many host genes will play a role in determining differential susceptibility and that only a small fraction of these have been identified thus far.

Haldane’s 1949 proposal that genetic variation in globin genes might be driven by providing malaria resistance, and that similar forces from other pathogens could maintain great biochemical diversity, is being increasingly supported (1). This view has now been supported by studies on twins and adoptees (2–7), in addition to considerable data on specific genes and diseases.

Several different, yet complementary approaches to the identification of genetic variation important in the course of infectious disease progression have been taken. By far the most common approach has been to look for association in candidate genes using case–control studies. In general, large sample sets are needed to detect even moderate genetic effects in order to eliminate the possibility of false positive associations. Often, this has not been the case, leading to examples of inconsistent findings in genetic association studies, a problem not specific to the study of infectious diseases (8). However, publication bias may have resulted in many more such findings going unreported. A major problem of this candidate gene approach is in the problem of selection of appropriate candidate genes, but the record of successful guessing in this specific field is reasonably good, perhaps because genes related to infection resistance have higher levels of variation than most. More recently, the use of microarray technology has identified novel candidate genes on the basis of differential expression (9). Finally, clues as to what genes affect response to infection in humans are being found in the analysis of outcome of infection in model organisms such as mice or drosophila (10–12).

More recently, family-based approaches have also become more widely reported. Utilization of large numbers of families to look for linkage to infectious disease has been reported in a relatively small number of infectious diseases thus far (13–19) (Fig. 1). Other linkage studies to identify genes causing rare, monogenic susceptibility phenotypes have been reported, though the causative variation identified are rare mutations rather than polymorphisms and the
MALARIA

Approximately 40% of the world’s population is at risk from malaria, and around 1 million deaths each year, predominantly in children, can be attributed to this protozoan parasite (22). Probably the greatest number of genes conferring differential susceptibility to any disease has been reported for various manifestations of malaria. The classic example is the selective advantage conferred by the sickle haemoglobin heterozygous genotype (23) which is associated with a 90% reduction in the risk of severe malaria (24). Haemoglobin E is associated with a reduction in disease severity in southeast Asia whereas haemoglobin C, like sickle haemoglobin, is also associated with reduced malaria susceptibility in West Africa, but is clinically less severe (25–27). Additionally, the geographical distribution of the common α- and β-thalassaemias, in which globin synthesis is imbalanced, supports a protective effect against Plasmodium falciparum malaria, but the exact mechanism of this remains unclear (28). There are conflicting data on haptoglobin associations with malaria. Ahaptoglobinemia (i.e. the absence of haptoglobin) is a common phenotype in malaria infected individuals, but this is not entirely dependent on genotype and can be affected by the disease status of the individual (29–31). Although studies of the relative frequencies of the protein isoforms of haptoglobin have shown an association with malaria infection, no evidence of this was seen in a more recent, larger study in West Africa (32–34). By far the strongest effect that a genetic variant has on protection from Plasmodium infection is conferred by the Duffy negative phenotype. A single nuclear polymorphism (SNP) in the promoter of the Duffy antigen/chemokine receptor (DARC) gene alters the binding of the haematopoietic cell specific transcription factor GATA-1, thus inhibiting DARC expression (35). The absence of this receptor from the red blood cell surface results in complete protection from P. vivax malaria infection (36) and this variant has reached fixation in most sub-Saharan Africans.

Extended haplotypes around two genes on the X chromosome have recently been demonstrated to have signatures of positive selection, probably as a result of malaria (37). Both have been associated with differential malaria susceptibility, compellingly in the case of the first, glucose 6-phosphate dehydrogenase (G6PD) (38). The other gene encodes CD40 ligand, which interacts with the key lymphocyte and dendritic cell receptor CD40 (39).

Immunologically important molecules such as cytokines and receptors have been popular candidates for analysis. Variants in tumour necrosis factor alpha (TNFA), CD32 (FcγRIIa) and subunits of the interferon alpha and gamma receptors (IFNARI/IFNGRI) have all been reported to influence susceptibility to clinical malaria (40–45). Previous analyses of the differences in ethnic groups in West Africa revealed that the Fulani are less susceptible to clinical malaria than their neighbours in Burkina Faso (46). Analysis of polymorphisms in the ILA cytokine gene promoter demonstrated an association that could partially explain this difference (47).

The importance of the HLA locus in initiation and regulation of the immune response, together with their well-documented variation, has led to numerous studies of their influence on disease susceptibility and progression. Some of this research is summarized in Table 1. However, because HLA molecules interact with polymorphic parts of the parasite (in turn due to immune selection on these pathogen sequences), and parasite allele frequencies differ geographically (at times very strikingly), it is perhaps not surprising that HLA association with malaria shows interpopulation heterogeneity. Both HLA class I and class II alleles have been found to influence malaria susceptibility in Africa (24,51). HLA has also been observed to influence the parasite strain associated with clinical malaria and the complex interactions between these factors may lead to further variability in HLA allele associations (59,60). A major genetic effect of the HLA region was observed even in studies of a mild malaria phenotype (rather than severe malaria) in a linkage study of Gambian dizygotic twins (14). Other, more thorough, linkage studies have yet to be completed, but could reveal further major as yet unidentified genes that influence the course of malaria infection.

Figure 1. Diagrammatic representation of the location of some major infectious disease susceptibility loci identified via genomewide linkage analysis. Text colour represents the disease area: red, schistosomiasis parasite burden; aqua, malaria; blue, leprosy; green, tuberculosis; purple, tuberculosis and leprosy combined; orange, kala azar. Loci: A, Marquet et al. (15); B, Jepson et al. (14); C, Mira et al. (16); D, Siddiqui et al. (17); E, Blackwell et al. (69); F and J, Bellamy et al. (13); G, Miller et al. (70); H, Tosh et al. (18); I, Bucheton et al. (19).

identified genes rarely, if ever, relevant to common variation in susceptibility (20,21).
MYCOBACTERIAL DISEASES

Familial clustering data, twin studies and complex segregation analysis have all suggested a strong genetic component to the human chronic mycobacterial diseases, leprosy and tuberculosis. Early segregation analysis suggested that single or few major genes might be implicated in some populations (3,4,61,62). Several family studies to search for major genes have been undertaken and locations of the loci identified are shown in Figure 1.

There are around 700 000 new cases each year of clinical leprosy which can be divided into two polar types. Lepromatous leprosy is associated with a high bacterial load, a strong type 1 cytokine response and the presence of skin nodules or plaques. Tuberculoid leprosy is associated with the presence of few bacteria (paucibacillar) and hypopigmented, desensitized skin lesions and often the thickening of local nerves. Host genes that have been identified thus far have been implicated in both types of leprosy or with leprosy per se. Family studies have revealed the presence of three loci that are linked to leprosy. Studies on a South Indian population have revealed that two loci are involved in susceptibility to tuberculoid leprosy on the short arms of both chromosomes 10 and 20 (17,18), and the chromosome 10 gene has recently been identified (Tosh et al., submitted for publication).

Analysis of a cohort of Vietnamese families confirmed the linkage to chromosome 10 and reported a further locus on chromosome 6 (16). Fine mapping of the chromosome 6 locus revealed that a cluster of non-coding polymorphisms within the shared promoter region of two genes; PARK2 and PACRG (63,64). The association within this region was replicated in a large set of Brazilian leprosy cases and controls. Which of these two genes is the most important is, as yet, unknown, as neither represents an obvious candidate for a leprosy susceptibility gene. Both are expressed in the host cells for M. leprae; the Schwann cells and macrophages (63). PARK2, a gene previously associated with juvenile Parkinsonism, is a ubiquitin E3 ligase, whereas PACRG forms part of a molecular chaperone complex (65,66).

Pulmonary tuberculosis (TB) is characterized clinically by fever, cough, weight loss and an abnormal chest X-ray, whereas the symptoms of non-pulmonary TB differ depending on which organ(s) are affected. A study of African families from both West and South Africa found suggestive evidence of susceptibility loci, one on chromosome 15 and the other on chromosome X. Subsequent fine mapping of the linkage on chromosome 15 identified another ubiquitin ligase gene, UBE3A, as a positional candidate in this region (67), indicating again the possible importance of ubiquitin ligases in susceptibility to mycobacterial infection. Further analysis of the chromosome X locus excluded the CD40L gene as the susceptibility gene for this gene (68). Additional studies in Brazilian multi-case families have found some evidence of linkage on both chromosomes 11 and 17, the latter requiring a combined analysis of both leprosy and TB families (69–71).

In addition to family studies, candidate gene studies have been relatively successful in the identification of genes implicated in host susceptibility to leprosy and TB. SLCA11A1 (or NRAMP1) was identified originally after work in a mouse model identified Nramp1 as a susceptibility locus for infection with some strains of Mycobacterium bovis BCG, Leishmania donovani and Salmonella typhimurium (72,73). Studies of the effect of variation in this gene on susceptibility to TB and leprosy have produced conflicting results. The first study of TB susceptibility found that four variants in the gene were associated with a higher risk of infection in the Gambia (74). Subsequent studies have replicated this association not only in the initial population, but also in Conakry, Korea, Japan, Cambodia and the USA (75–79). However, not all reports confirm the original findings (Fitness et al., unpublished data), and there are several examples where no association was found, but many of the studies looking at NRAMP1 as a susceptibility gene for TB are small and underpowered for the magnitude of the genetic effect observed (80–82). The findings for leprosy are less clear, one study found evidence of NRAMP1 variation affecting susceptibility to leprosy per se, but several other studies have failed to replicate this (83–86).

Another gene that has shown evidence for involvement in mycobacterial disease is the vitamin D receptor (VDR). Vitamin D is an immunomodulatory molecule and via its receptor, can modulate cytokine responses in T cells (87,88). Several small-scale studies have suggested that variants of this receptor may alter both leprosy and TB risk (85,89,90).

Toll-like receptor (TLR) genes were first identified in drosophila gene knockout studies that produced severely immunocompromised flies that are fatally susceptible to fungal infections (91). An extension of this work in mice resulted in the positional cloning of a gene responsible for lipopolysaccharide (LPS) sensitivity in mice. This gene was homologous to the drosophila Toll gene, and was subsequently identified as the murine homologue of the TLR4 gene (92–94). Of the 10 TLR family genes in humans, TLR2 has been most strongly implicated in analysis of mycobacterial ligands (95). Significant association has been published with respect to TB infection, finding that an amino acid substitution causing variant in TLR2 is associated with increased susceptibility to TB in Turkey and Tunisia (96,97). A second amino acid variant is reported to be found only in lepromatous leprosy cases in Korea (98). Whether these changes are themselves affecting

<table>
<thead>
<tr>
<th>HLA allele</th>
<th>Disease</th>
<th>Effect</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-B53</td>
<td>Malaria</td>
<td>Resistance</td>
<td>24</td>
</tr>
<tr>
<td>HLA-B35</td>
<td>HIV-1 progression</td>
<td>Susceptibility</td>
<td>48</td>
</tr>
<tr>
<td>HLA-B27</td>
<td>HIV-1 progression</td>
<td>Resistance</td>
<td>49</td>
</tr>
<tr>
<td>HLA-B57</td>
<td>HIV-1 progression</td>
<td>Resistance</td>
<td>50</td>
</tr>
<tr>
<td>HLA-DR.B1*1302</td>
<td>Malaria</td>
<td>Resistance</td>
<td>51</td>
</tr>
<tr>
<td>HLA-DPB1</td>
<td>Tuberculosis</td>
<td>Resistance</td>
<td>52</td>
</tr>
<tr>
<td>HLA-DQBI</td>
<td>Tuberculosis</td>
<td>Susceptibility</td>
<td>52</td>
</tr>
<tr>
<td>HLA-DQBI</td>
<td>Leprosy</td>
<td>Resistance</td>
<td>53</td>
</tr>
<tr>
<td>HLA-DR2</td>
<td>Tuberculosis</td>
<td>Susceptibility</td>
<td>54</td>
</tr>
<tr>
<td>HLA-DR2</td>
<td>Leprosy</td>
<td>Susceptibility</td>
<td>55</td>
</tr>
<tr>
<td>HLA-DRB1*1302</td>
<td>HBV persistence</td>
<td>Resistance</td>
<td>56</td>
</tr>
<tr>
<td>HLA-DR2</td>
<td>HBV persistence</td>
<td>Resistance</td>
<td>57</td>
</tr>
<tr>
<td>HLA-DR7</td>
<td>HBV persistence</td>
<td>Susceptibility</td>
<td>57</td>
</tr>
<tr>
<td>HLA-DRB1*11</td>
<td>HCV persistence</td>
<td>Resistance</td>
<td>58</td>
</tr>
</tbody>
</table>
TLR2 function or are in linkage disequilibrium with the intronic functional microsatellite is thus far unclear (99). As yet, no evidence has been published linking TLR4 with mycobacterial infection (100,101).

In addition to numerous HLA studies of mycobacterial infection and disease (Table 1) a large number of other immunological candidates have been assessed. Putatively functional variation in genes such as IL1, IFNγ and MBL have been associated with TB susceptibility (102–106), but subsequent analyses of other populations have failed to replicate these associations (106–109). Larger, better designed studies will be required to distinguish real population heterogeneity from chance differences.

**VIRAL DISEASES**

Studies on the host genetics of viral infections are not limited to analyses of susceptibility or severity of infection. In the case of HIV/AIDS, the most frequently studied phenotypes are time to AIDS or death following HIV infection. Studies of persistent hepatitis have been published using multiple phenotypes such as outcome of infection, treatment response and major complications such as hepatocellular carcinoma or fibrosis.

HIV-1 infection is characterized by an extended clinically asymptomatic phase that leads, at a much later stage, to immune deficiency, opportunistic infections, neurological problems and malignancies that characterize AIDS *per se*. Studies on cohorts of HIV-infected individuals identified a small proportion of a cohort of commercial sex workers in Nairobi who remain HIV seronegative despite repeated exposure (110), though to date, no genetic factors have been identified that can adequately explain this apparent resistance to infection.

Since the discovery of the chemokine receptors as co-receptors with CD4 for viral entry into the host cells, many studies have investigated the role of chemokines and their ligands. The role of variation in the CCR5 gene in resistance to infection and slower disease progression is now well established (111,112). Analysis of the flanking CCR2 gene also showed that an amino acid change in this gene is linked to delayed progression to AIDS, although the mechanism of action of this variant in a rarely used co-receptor is unclear (113). Several studies have looked at the role of the CCR5 ligand RANTES. The presence of an intronic SNP that differentially binds regulatory proteins was associated with increased disease progression to AIDS in both European and African Americans (114), suggesting an evolutionarily important role for this new SNP in immunomodulation, which may be unlinked to HIV-1 disease progression. A second smaller study in the Chinese Han population has replicated this result (115). Another HIV-1 co-receptor ligand in HIV infection, stromal derived factor 1 (SDF1) is also associated with HIV-1 disease progression in some populations (116,117). Many other genes have been identified as HIV-1/AIDS host genetic determinants. These include genes such as IL4, IL10 and NRAMP1 that have been linked with other infectious diseases (118–123).

Infection with either the hepatitis B or hepatitis C virus results in either an acute, self-limited disease or, in a minority, in persistent infection. Persistent carriage rates, which confer an increased risk of liver complications, failure or end stage carcinoma, are ~10–20% in hepatitis B when compared with ~80–90% of hepatitis C infections. Studies of the host genetic factors implicated in disease chronicity have found that non-HLA genes that were mentioned earlier in this review, such as TNFA, MBL and VDR, are all associated with persistence of hepatitis B infection (89,124,125). The unusual immunological dichotomy of the outcome of disease makes HBV viral persistence an ideal candidate for family linkage-based studies enabling the identification of novel major genes that determine this fate. Recently a genomewide scan in Gambian families has mapped a major susceptibility locus to chromosome 21 and two neighbouring genes appear to be involved (Frodsham *et al.*, submitted for publication).

In HCV viral persistence, the high numbers that are treated (with either interferon alpha alone or in combination with antiretroviral therapy) allow for the analysis of genetic influence on treatment response and complications of infection in addition to the outcome of infection. The role of the CCR5 receptor, already acknowledged to be important in viral infection, has been investigated with respect to HCV infection. An initial study among haemophiliacs suggested that CCR5 may play a role in determining HCV persistence (126) but this has been disputed (127–129), though a recent paper suggests that CCR5 may be important in the risk of liver complications such as portal inflammation and risk of fibrosis (130) but not to the outcome of infection. Studies of the interferon induced genes such as *MxA* and *PKR* show that variation in these genes is a factor in the hosts’ response to therapy (131). The picture for IL10 however, is not so clear, with contradictory results from different studies (132–134). Even though persistence rates are higher in HCV than in HBV, the lower prevalence makes family identification difficult, therefore limiting the possibility of family studies the genetics of HCV persistence.

**CONCLUDING REMARKS**

Available information supports the view that for the majority of infectious diseases host susceptibility is likely to be highly polygenic. Indeed, the relatively few major genes that have been identified in several genomewide linkage scans for bacterial, parasitic and viral infectious diseases support the view that the genetic susceptibility in these diseases is widely distributed among numerous polygenes.

The International HapMap project (135), an entire human genome sequence and increasing amounts of resequencing data for chromosomal segments, means that there are now millions of SNPs available for association studies (136). With high throughput technologies it is now feasible, but still very expensive, to perform a genomewide association screen to identify novel genes involved in response to infection, even in the diseases such as HIV and HCV persistence, where family studies have thus far, been impossible. Such data should provide far greater insights into disease pathogenesis than has been possible with the few dozen candidate genes that have been evaluated to date. The impact of these new insights on the control of these major global diseases could be considerable in identifying targets for new therapies, in improved vaccination strategies and eventually in disease elimination. In time, it should also be possible to define...
each individual a personalized risk profile that could predict their own unique susceptibility to various infectious organisms. It is likely that many of the SNPs contained in such a screen will be relevant to multiple infectious pathologies, as there is already considerable evidence that the immunomodulatory role of certain polymorphisms extends across disease boundaries. Extension of this approach into pharmacogenetics should result in individually tailored immunotherapy according on the host's genetic profile.

ACKNOWLEDGEMENTS

The authors would like to thank Graham Cooke and Andrew Walley for their help in compiling this paper. A.J.F. is a Wellcome Trust Research Training Fellow and A.V.S.H. is a Wellcome Trust Principal Research Fellow.

REFERENCES


