Infectious diseases not immune to genome-wide association

Paul I W de Bakker & Amalio Telenti

Two genome-wide association studies for meningococcal disease and tuberculosis identify new loci associated with susceptibility to these infectious diseases. They highlight a role for the acquired and innate immune systems in host control of several human pathogens and demonstrate that denser genotyping platforms and population-specific reference panels are necessary for genetic studies in African populations.

Progress in identifying susceptibility genes for infectious disease has been slow in comparison to other common diseases, despite the widely held view that exposure to pathogens during human evolution has exerted evolutionary pressure influencing host susceptibility to infectious disease. In the past several years, genome-wide association studies (GWAS) have been reported for susceptibility to infectious disease including HIV[1], hepatitis B[2] and C[3,4] viral infection, leprosy[5] and malaria[6]. On page 772 of this issue, Sonia Davila and colleagues report a GWAS for host susceptibility to meningococcal disease[7], and on page 739 Adrian Hill and colleagues report a GWAS for tuberculosis conducted in African populations[8]. They each identify new susceptibility loci and, together with other recent studies, highlight a role for acquired and innate immunity in host control of these infectious diseases.

Host susceptibility

Meningococcal disease is a bacterial infection caused by *Neisseria meningitidis*, and when untreated, is associated with high mortality rates worldwide. Although asymptomatic colonization by this bacterium is common, infected individuals have varying risk in developing sepsis, widespread blood infection, or meningitis, the often life-threatening inflammation of the protective membrane that covers the brain and spinal cord. Davila *et al.*[7] now report a GWAS of 475 cases and 4,703 population-based controls from the UK, with replication in Western European and South European cohorts, identifying a set of strongly correlated SNPs at a single locus associated with meningococcal disease. These SNPs span the genes encoding complement factor H (*CFH*) and CFH-related protein 3 (*CFHR3*), and they have a relatively strong effect size (odds ratio = 0.6). Worldwide, the protective alleles of these SNPs appear to be least common in Africa, where the incidence of meningococcal disease is the highest, suggesting that these variants may explain (to some degree) population differences in disease rates. A role for the alternative complement pathway is not unexpected as inherited complement deficiencies have been associated with meningococcal disease[9]. *Neisseria meningitidis* is also thought to have evolved the ability to recruit human factor H to protect from complement activation, as a mechanism for evading host immune control[9]. *CFH* was the first locus identified as being associated with susceptibility to age-related macular degeneration (AMD)[10]. Although fine-mapping studies are needed to pinpoint the causal variant(s) for meningococcal disease at the *CFH* locus, a near-perfect proxy for the SNPs associated in Davila *et al.*[7] has not been associated with AMD risk (rs742855, $P = 0.28$ from ref. 11), suggesting that they are independent.

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis*. Although numerous genetic association studies have implicated candidate genes in tuberculosis susceptibility, limited sample sizes as well as complications in validating the findings across populations have meant that the statistical evidence for association has often not been convincing[12]. Thye *et al.*[8] now report a combined GWAS for host susceptibility to pulmonary tuberculosis, including 2,100 cases and 3,000 controls from Ghana and The Gambia, with replication in additional cohorts from Ghana and Malawi. They identified a single locus with a modest effect size (OR = 1.2) on chromosome 18q11, with no obvious functional candidate genes nearby. Given the poor track record of candidate-gene studies of complex traits in general, and the limited coverage for an African population of 333,754 tested SNPs, it is not surprising that they failed to confirm previously reported associations (for example, *NRAMP1* or *CISH*).

Lessons for infectious diseases

The new findings from Davila *et al.*[7] on meningococcal disease, as well as recently published GWAS for hepatitis C[3,4] and leprosy[5], point strongly to a role for genetic variants of the innate immune system in human susceptibility to infectious disease (**Table 1**). Variants at *IL28B*—encoding interferon-λ—have been associated with response to interferon-ribavirin combination therapy for hepatitis C[3] as well as with natural clearance

Paul I.W. de Bakker is at the Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. Amalio Telenti is at the Institute of Microbiology, University Hospital Center, University of Lausanne, Lausanne, Switzerland.
e-mail: pdebakker@rics.bwh.harvard.edu or amalio.telenti@chuv.ch

Table 1 Genetic loci identified by genome-wide association studies for host susceptibility to infectious diseases

Disease	Pathogen	Gene or locus	Biological mechanism
AIDS[1]	Human immunodeficiency virus-1	Major histocompatibility complex, class I (*HLA-B–HLA-C*), *CCR5*	Acquired immunity, deletion of viral co-receptor
Hepatitis B[2]	Hepatitis B virus (HBV)	Major histocompatibility complex, class II (*HLA-DP*)	Acquired immunity
Hepatitis C[3,4]	Hepatitis C virus (HCV)	*IL28B*	Innate immunity
Leprosy[5]	*Mycobacterium leprae*	Major histocompatibility complex, class II (*HLA-DR–DQ*), *NOD2*, *TNFSF15*, *RIPK2*, *CCDC122* and *C13orf31*)	Acquired and innate immunity, and unknown mechanisms
Tuberculosis[8]	*Mycobacterium tuberculosis*	18q11.2 (*GATA6*, *CTAGE1*, *RBBP8*, *CABLES1*)	Unknown
Meningococcal disease[7]	*Neisseria meningitidis*	*CFH*, *CFHR3*, *CFHR1*	Innate immunity

of the virus[4]. In the case of leprosy[5], a single GWAS identified several candidate susceptibility genes (*NOD2*, *TNFSF15* and *RIPK2*), all encoding proteins involved in innate immune response. The apparent overlap between the hits for leprosy[5] and Crohn's disease[13] might also motivate a more targeted analysis of all variants involved in autoimmunity, leveraging the recent wave of new loci associated with autoimmune disorders identified through GWAS.

The human leukocyte antigen (HLA) region also has an important role in infectious diseases. GWAS have implicated variants of major histocompatibility (MHC) class I genes as having strong associations with effective control of HIV-1 replication[1]. A recent GWAS of chronic hepatitis B infection highlighted variants in *HLA-DP*, an MHC class II gene[2]. There was no overlap of genetic signals between the GWAS for tuberculosis[8] and leprosy[5] despite the various aspects of histopathology shared between the two diseases. However, Thye *et al.* did report a weak association in the MHC class II (*HLA-DQ* region) in tuberculosis[8], a region in which Zhang *et al.* also identified variants associated to leprosy[5].

African genetics

With most of the GWAS reported over the past few years focusing on European populations, there remain technical challenges with genome-wide association analysis conducted in non-European populations[14]. Imputation based on linkage disequilibrium between genotyped markers and SNPs typed in a reference panel (such as HapMap) can improve coverage and power[15] and has been a key tool for meta-analyses across different genotyping platforms. In African populations, the coverage of common variation in current platforms is also limited because of the lesser extent of linkage disequilibrium (LD) between variants in African populations. For example, the SNPs on the Affymetrix 500K platform (used by Thye *et al.*[8]) capture at most 46% of all SNPs in HapMap-YRI (Yoruba from Ibidan, Nigeria), with pairwise $r^2 > 0.8$ (ref. 15).

Moreover, imputation can only be expected to work well if the HapMap panel is adequately representative of the haplotype structure in the population samples studied. Indeed, Thye *et al.* noted that the quality of the imputations in their Ghanaian and Gambian samples was rather poor (with a mean imputation error of 8%) when they used HapMap-YRI as the reference panel and Affymetrix 500K SNPs from the GWAS as input genotypes. Although this is by no means a systematic evaluation of the problem, it is consistent with findings from a recent malaria GWAS, which found substantial population substructure in their samples from The Gambia[6]. Jallow *et al.*[6] identified a rather weak association signal for malaria susceptibility at the previously identified β-hemoglobin gene *HBB* ($P \approx 10^{-7}$). After sequencing this locus in a custom reference panel of 96 Gambians, they were able to impute and detect the sickle-cell hemoglobin causal amino acid change that is known to protect against malaria with much stronger statistical significance ($P \approx 10^{-14}$). These are compelling demonstrations that the genotyping platforms routinely used for GWAS in European populations do not provide sufficient coverage in African populations and that a more diverse set of reference panels are needed for effective imputations. HapMap Phase 3 has already added two African continental populations (Luhya in Webuye from Kenya, and Maasai in Kinyawa from Kenya). The 1000 Genomes Project is well underway building a resource to document all variants down to 1% frequency across the genome in 2,500 individuals sampled from >20 populations (including five from the African continent). With these new tools emerging, we should be better armed to conduct genetic association studies in non-European populations, which contain most of the global burden of infectious disease.

Future studies

The identification of variants in genes of the acquired and innate immunity in GWAS for infectious disease calls for a systematic dissection of the entire MHC region, going beyond the classical HLA genes. The complex nature of the MHC region, with its broad LD, high diversity and structural variation, makes it one of the most difficult regions of the genome to study, but doing so should be a worthwhile effort, as it is the genomic region that shows the highest density of associations to complex diseases. A complementary, hypothesis-driven approach would be to test specifically the set of about 1,000 genes that have been characterized as involved in innate immunity. Another challenge in identifying host susceptibility factors is potential confounding by the infecting strain, as most studies do not take into account the variability of the pathogen itself. Genetic studies for host susceptibility to infectious disease will need to incorporate such considerations for a more complete picture of how the pathogen evades host immune responses.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Fellay, J. *et al. Science* **317**, 944–947 (2007).
2. Kamatani, Y. *et al. Nat. Genet.* **41**, 591–595 (2009).
3. Ge, D. *et al. Nature* **461**, 399–401 (2009).
4. Rauch, A. *et al. Gastroenterology* **138**, 1338–45, 1345e1–e7 (2010).
5. Zhang, F.R. *et al. N. Engl. J. Med.* **361**, 2609–2618 (2009).
6. Jallow, M. *et al. Nat. Genet.* **41**, 657–665 (2009).
7. Davila, S. *et al. Nat. Genet.* **42**, 772–776 (2010).
8. Thye, T. *et al. Nat. Genet.* **42**, 739–741 (2010).
9. Serruto, D., Rappuoli, R., Scarselli, M., Gros, P. & van Strijp, J.A. *Nat. Rev. Microbiol.* **8**, 393–399 (2010).
10. Klein, R.J. *et al. Science* **308**, 385–389 (2005).
11. Raychaudhuri, S. *et al. Nat. Genet.* **42**, 553–555, author reply 555–556 (2010).
12. Möller, M., de Wit, E. & Hoal, E.G. *FEMS Immunol. Med. Microbiol.* **58**, 3–26 (2010).
13. Barrett, J.C. *et al. Nat. Genet.* **40**, 955–962 (2008).
14. Teo, Y.Y., Small, K.S. & Kwiatkowski, D.P. *Nat. Rev. Genet.* **11**, 149–160 (2010).
15. The International HapMap Consortium. *Nature* **449**, 851–861 (2007).