The Pediatric Infectious Disease Journal Publish Ahead of Print

DOI: 10.1097/INF.0000000000001183

Genome-wide association studies in infectious diseases

Eleanor G. Seaby[1], Victoria J. Wright[1], Michael Levin[1]

[1]*Section for Paediatric Infectious Diseases, Division of Infectious Diseases, Department of Medicine, Imperial College London*

Email

Eleanor G. Seaby (egs1g09@soton.ac.uk)

Victoria J. Wright (v.wright@imperial.ac.uk): Corresponding author.

Michael Levin (m.levin@imperial.ac.uk)

Introduction

One of the most puzzling questions facing both parents and paediatricians is why some children suffer severe and life-threatening infections, while others who are exposed to the same pathogen remain uninfected, or suffer only mild illness. There is now growing evidence that differences in host genetics play a major role in determining susceptibility and outcome of childhood infections[1], in addition to environmental factors such as pathogen virulence and exposure dose.

The genetic contribution to infectious disease in children may be from rare, highly deleterious genetic variants (such as Mendelian complement pathway defects predisposing to meningococcal disease, and IFN-□/IL-12 pathway defects causing susceptibility to mycobacteria), or due to common genetic variants each contributing only a small proportion to total disease risk. This review will focus on the identification of common genetic variants in children using genome-wide approaches, and will not discuss Mendelian defects which have been reviewed elsewhere.[1]

Early attempts to identify genetic susceptibility to underlying childhood infections adopted a candidate gene approach, centred on disease immunopathogenesis, whereby genes suspected to play a role in a disease were examined in cohorts of cases and controls. Whilst this approach successfully identified many associations, the most relevant candidate gene may not have been studied, as selection was based around biological knowledge of the disease of interest at that time.

Our ability to search for the genes underlying infection using a non-candidate driven approach was greatly advanced once the human genome sequence was published in 2003. With subsequent human genetic variant mapping and improvements in technology permitting relatively inexpensive genotyping of large numbers of common genetic variants using high throughput micro-arrays in large cohorts, it became possible to search for genes influencing human disease without any prior biological knowledge – a process now known as a genome-wide association study (GWAS). Since publication of the first results in 2005, GWAS has emerged as a powerful tool to identify the genes underlying susceptibility and severity of many diseases, identifying approximately 2,000 robust associations with complex diseases.

GWAS approach, study design and analysis

A GWAS is based on comparison of the frequency of large numbers (thousands to millions) of single nucleotide polymorphisms (SNPs) across the genome in cohorts of patients with the disease of interest and unaffected controls. In some instances SNPs themselves are functional and determine changes in protein structure or function that might affect disease, resulting in greater frequency of the SNP in cases compared with controls. More commonly, an unknown genetic variant, which may be associated with disease susceptibility, is inherited together with SNPs in the neighbouring chromosomal regions. By detecting an association of the neighbouring SNPs with a disease, the unknown variant can be located, and then identified by fine-mapping or sequencing of the chromosomal region.

A number of steps are essential for a successful GWAS (Figure 1) which include a well-defined cohort of cases and appropriately matched controls, and quality control assessment of samples as well as genotyping data. GWAS data analysis should be corrected for multiple hypothesis testing with a P value required for statistical significance of associations generally

set at less than 5.00E-08. This value is calculated by dividing a P value of 0.05 by the number of SNPs assessed in the study, i.e. 0.05/1.00E-06 = 5.00E-08. As the P value cut-off is purely 'statistical', a number of alternative analytical approaches have been developed including pathway-based approaches and network analyses whereby the biological role of genes is assessed alongside its statistical association.

Regardless of the analysis approach, all GWAS results require replication and should be further validated by measurement of proteins, gene expression or function. Replication studies should have sufficient sample sizes to detect the effect; be tested on an independent dataset and use the same phenotype as the initial GWAS. Confidence in the result is increased if the effect is observed from the same SNP or another SNP in high linkage disequilibrium (LD) with the candidate SNP and is in the same direction (increased or decreased in the disease cohort relative to controls).

Unlike linkage studies (where genetic loci are mapped in related individuals with a given trait) and candidate gene analysis, GWAS provide greater resolution in the identification of modest effect alleles, and do not require *a priori* candidate gene selection; this has facilitated the identification of novel SNPs in genomic regions not previously implicated in disease. With LD data on neighbouring SNPs available from the HapMap project, it is now possible to impute untyped variants and assess their significance.

GWAS in paediatric Infectious diseases

Several GWAS have been conducted in childhood infections including meningococcal disease, Kawasaki disease and malaria.

Meningococcal disease

The first GWAS for meningococcal disease (MD) in 2010 by Davila et al.[2], used 475 disease cases from the UK and 4703 controls from the 1958 British Birth Cohort and the UK Blood Service Collection (genotyped by the Wellcome Trust Case Control Consortium). Primary analysis identified 79 SNPs with significance $P < 1 \times 10^{-4}$. The results were replicated first in 553 Western European cases and 839 matched controls, where two highly significant SNPs in complement factor H (*CFH*) were identified in a combined analysis, and these SNPs in *CFH* were further replicated in a second cohort from Spain of 415 cases and 537 controls. Individuals carrying the minor allele were protected against meningococcal disease with a relative risk of ~0.6 as compared with individuals carrying the wild type allele. Imputation analysis across all cohorts also revealed three SNPs reaching genome-wide significance in a combined analysis in the adjacent gene *CFHR3*. All SNPs decreased disease susceptibility for carriers of the minor allele.

CFH and CFHR3 are both biologically plausible candidates in the pathogenesis of MD. The discovery that *Neisseria meningitidis* express a FH binding protein (fHbp) with high affinity for human FH and FHR3 suggests that on entering the blood stream, *N. meningitidis* use a 'Trojan horse' strategy to evade complement mediated killing by coating themselves with host CFH. CFHR3 is postulated to compete with CFH for binding to meningococcal fHbp. However some form of genetic regulation between FHR3 and CFH may also explain the association.

This GWAS provides definitive evidence that genetic differences within the complement pathway underlie susceptibility to meningococcal disease in the general population, as well as in families with rare complement Mendelian defects.

Kawasaki disease

Kawasaki disease (KD) is a systemic vasculitis of unknown aetiology and has been a target for several GWAS. Burgner et al. (2009)[3] performed the first in a Dutch Caucasian population using a case-control design with validation using KD trios from further Caucasian populations. The study replicated, in a combined analysis, 8 significantly associated genes (of which five were linked in a biological network associated with cardiovascular pathology, inflammation, and apoptosis, and five had lower transcript abundance in the acute phase of illness. Two years later, Khor et al. (2011)[4] conducted a GWAS with replication in 2173 cases and 9383 controls using a case-control and family-based design in 5 independent sample collections. Validated results included a functional SNP in *FCGR2A*, a SNP near *MIA* and *RAB4B*, and a SNP in *ITPKC (ITPKC* had previously been identified by Onouchi et al (2008) as a KD susceptibility gene using a genome-wide linkage analysis). The SNP in *FCGR2A* association was replicated in a Japanese cohort of 428 cases and 3379 controls, together with two replication studies including 754 cases and 947 controls (Onouchi et al. (2012)[5]). This study also identified significant associations in the regions of *HLA*, *FAM167A-BLK*, and *CD40* in a combined analysis. Further GWAS in a Han Chinese population identified novel loci (*COPB2*, *ERAP1*, *IGHV*)[6] and a meta-analysis confirmed previous associations in Japanese, Taiwanese and Korean populations.[7]

The genes identified in these GWAS point to immunological differences in antibody production (CD40, BLK), clearance of immune complexes (FCGR2A), and T cell activation (HLA, ITPKC). These contribute to disease occurrence, and suggest that KD is triggered by, as yet unidentified, environmental factors or infectious agents in children whose immune system is genetically determined to respond differently compared with unaffected children.

Malaria

The initial GWAS in malaria only confirmed the already known association with *HBB*[8]. Later studies of severe malaria in large cohort sizes[9-11] replicated SNPs in *ATP2B4*, 16q22, *ABO* and *CD40LG*. A recent GWAS of severe malaria[12] in 5130 cases and 5291 controls, with replication in 13,946 individuals, identified novel loci with a highly significantly SNP (rs184895969) located between *FREM3* and genes encoding receptors known to be used by *P. falciparum* for invasion into the erythrocyte (*GYPA, GYPB*, *GYPE*). A haplotype at this locus showed protection against severe disease (OR = 0.67, 95% CI, 0.60–0.76, $P = 9.5 \times 10^{-11}$).

Overall, these data suggest that genetic susceptibility to severe malaria involves genes associated with blood group antigens, and the erythrocyte cell membrane.

Limitations and challenges

GWAS have successfully identified several validated genetic associations with disease susceptibility. However extricating precise causal mechanisms requires fine-mapping and functional studies; both of which are labour and cost intensive and particularly difficult for non-coding susceptibility loci.

GWAS are insufficiently powered to identify rare variants, and many common diseases are probably influenced by rare variants acting alone or in combination. GWAS also have a limited ability to identify copy number variants (CNVs) which have been shown to be an important source of genetic variation in infectious diseases.[13] Newer arrays and analytic approaches are now improving detection of CNVs.

Conclusions and future

GWAS for infectious diseases are producing valuable data with the potential to identify biological mechanisms that influence host defence to disease. As susceptibility to infection might be caused both by common variants each contributing a small amount to susceptibility (which can be identified by GWAS), and by rare Mendelian variants (which require other approaches such as exome or whole genome sequencing), future studies that combine both approaches with functional studies of the variants identified, are likely to improve our understanding of the genetic basis of childhood infection.

References

1. Chapman SJ, Hill AV. Human genetic susceptibility to infectious disease. Nat Rev Genet 2012;13:175-88.
2. Davila S, Wright VJ, Khor CC, et al. Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat Genet 2010;42:772-6.
3. Burgner D, Davila S, Breunis WB, et al. A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. PLoS Genet 2009;5:e1000319.
4. Khor CC, Davila S, Breunis WB, et al. Genome-wide association study identifies FCGR2A as a susceptibility locus for Kawasaki disease. Nature genetics 2011;43:1241-6.
5. Onouchi Y, Ozaki K, Burns JC, et al. A genome-wide association study identifies three new risk loci for Kawasaki disease. Nature genetics 2012;44:517-21.
6. Tsai FJ, Lee YC, Chang JS, et al. Identification of Novel Susceptibility Loci for Kawasaki Disease in a Han Chinese Population by a Genome-Wide Association Study. Plos One 2011;6.
7. Chang CJ, Kuo HC, Chang JS, et al. Replication and meta-analysis of GWAS identified susceptibility loci in Kawasaki disease confirm the importance of B lymphoid tyrosine kinase (BLK) in disease susceptibility. Plos One 2013;8:e72037.
8. Jallow M, Teo YY, Small KS, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 2009;41:657-65.
9. Band G, Le QS, Jostins L, et al. Imputation-based meta-analysis of severe malaria in three African populations. PLoS Genet 2013;9:e1003509.
10. Timmann C, Thye T, Vens M, et al. Genome-wide association study indicates two novel resistance loci for severe malaria. Nature 2012;489:443-6.
11. Malaria Genomic Epidemiology N, Malaria Genomic Epidemiology N. Reappraisal of known malaria resistance loci in a large multicenter study. Nat Genet 2014;46:1197-204.
12. Malaria Genomic Epidemiology N, Band G, Rockett KA, Spencer CC, Kwiatkowski DP. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature 2015;526:253-7.
13. Hollox EJ, Hoh BP. Human gene copy number variation and infectious disease. Hum Genet 2014;133:1217-33.

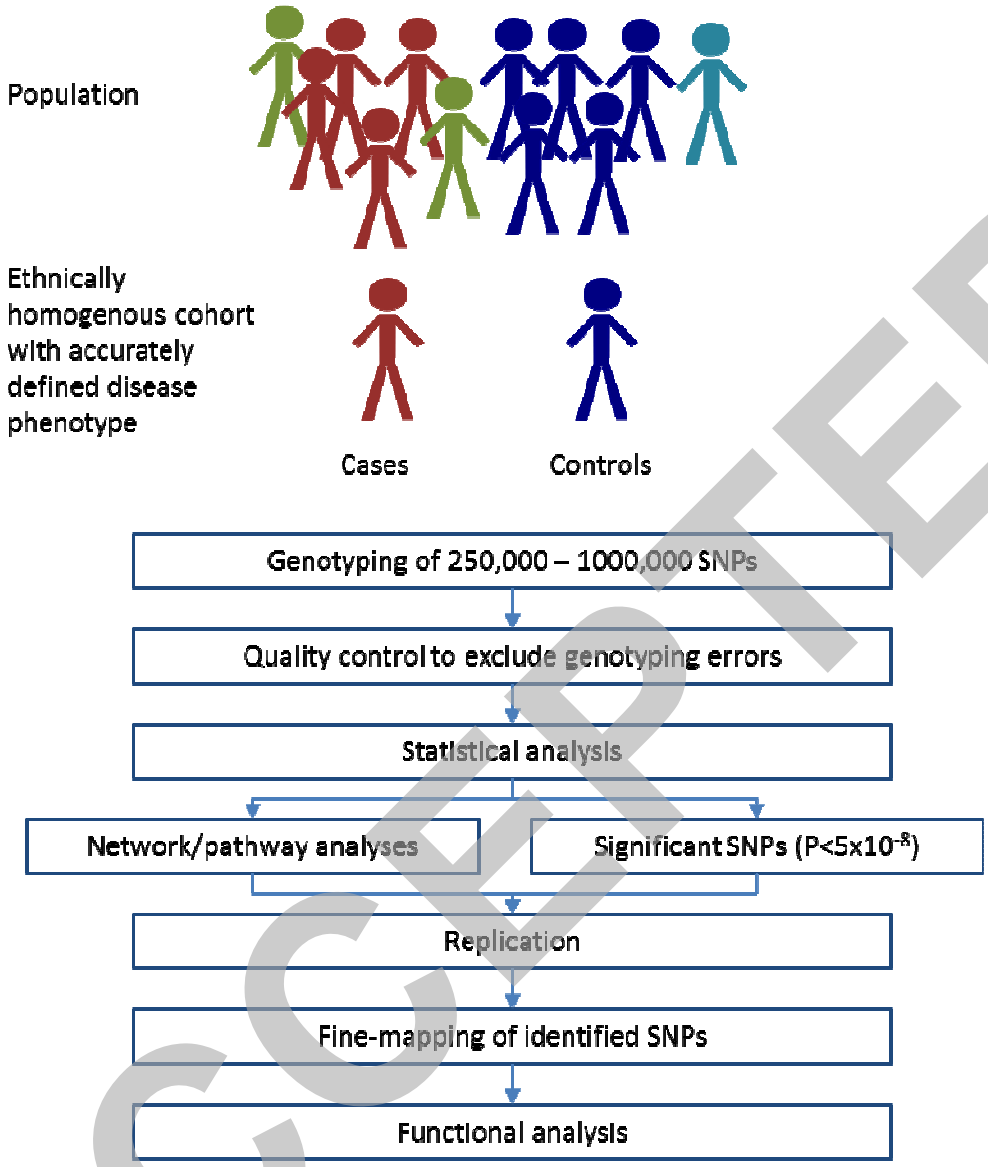

figure 1

Overview of genome-wide association study. Carefully selected cases and controls should reflect the population from which cases are drawn, and the phenotype of disease under study. Cases and controls are non-related individuals in many studies, but an alternative approach is to use parents of the index case as the controls and a family-based association test for analysis.