ASSOCIATION STUDIES ARTICLE

Genome-wide association study identifies SNPs in the MHC class II loci that are associated with self-reported history of whooping cough

George McMahon1, Susan M. Ring1,2, George Davey-Smith1,2 and Nicholas J. Timpson1,2,*

1School of Social and Community Medicine and 2MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

*To whom correspondence should be addressed at: Oakfield House, Oakfield Grove, Bristol, BS8 2BN, UK. Tel: +44 1179287279; Fax: +44 1179287325; Email: n.j.timpson@bristol.ac.uk

Abstract

Whooping cough is currently seeing resurgence in countries despite high vaccine coverage. There is considerable variation in subject-specific response to infection and vaccine efficacy, but little is known about the role of human genetics. We carried out a case–control genome-wide association study of adult or parent-reported history of whooping cough in two cohorts from the UK: the ALSPAC cohort and the 1958 British Birth Cohort (815/758 cases and 6341/4308 controls, respectively). We also imputed HLA alleles using dense SNP data in the MHC region and carried out gene-based and gene-set tests of association and estimated the amount of additive genetic variation explained by common SNPs. We observed a novel association at SNPs in the MHC class II region in both cohorts [lead SNP rs9271768 after meta-analysis, odds ratio [95% confidence intervals (CIs)] 1.47 (1.35, 1.6), P-value 1.21E −18]. Multiple strong associations were also observed at alleles at the HLA class II loci. The majority of these associations were explained by the lead SNP rs9271768. Gene-based and gene-set tests and estimates of explainable common genetic variation could not establish the presence of additional associations in our sample. Genetic variation at the MHC class II region plays a role in susceptibility to whooping cough. These findings provide additional perspective on mechanisms of whooping cough infection and vaccine efficacy.

Introduction

Whooping cough, also known as pertussis, is a highly contagious infection of the respiratory tract. It is transmitted from person to person by coughing or sneezing and caused primarily by the gram-negative bacteria Bordetella pertussis. Without intervention it is a highly morbid infection in infants, and in the pre-vaccine era, it was responsible for 10,000 deaths per year in the UK (1). Where vaccination is not routine, whooping cough is still a major source of mortality leading to 195,000 deaths in 2008 (2).

In developed countries, the incidence of whooping cough has been effectively controlled by vaccination. However, it is endemic with periodic epidemics every 3–4 years. Improving vaccination efficacy represents a sustained goal to improve the control of whooping cough (3,4) and thereby reducing associated mortality, morbidity and economic costs (5,6). Current vaccination does not provide complete protection (7) and immunity wanes with time. Furthermore, there is a requirement to combine multiple strategies to improve coverage among the most vulnerable, including an accelerated schedule and preschool booster (8), cocooning (9) and immunization of expectant mothers (10). There is also growing evidence of a recent increasing numbers of cases, particularly in infants and among adolescents and adults (8,11,12). Among

Received: February 27, 2015. Revised: June 9, 2015. Accepted: July 17, 2015
© The Author 2015. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
the latter group this increase may be partly due to better surveillance and reporting (11,13) but among very young infants notifications are already well established and less likely to explain an increase (14). These factors have led to re-evaluation of the effectiveness of current vaccination efforts (15,16).

Studies suggest a genetic component exists for susceptibility to whooping cough. A twin study in The Gambia indicated interferon-gamma and interleukin-13 responses to B. pertussis antigens are heritable (17). Murine models of pertussis infection show that mice with a functional mutation in the TLR4 gene are more susceptible to pertussis (18,19). A candidate gene approach has indicated that variation in the human TLR-4 gene influences antibody response to pertussis vaccine (20–22). Although these studies are promising they have been limited in replicating initial associations by small sample sizes. For other infectious diseases, the genome-wide association study (GWAS) has been successful in discovering and replicating several genetic associations (23,24). A difficulty in examining a role for common genetic variants in causing whooping cough is the accurate identification of a large enough numbers of cases to robustly identify associations of small to moderate effect expected from common variant analysis (23,24). This leads to the dilemma of using measures of whooping cough collected in large prospective cohorts, which may rely on historic data or retrospectively self-reported history of whooping cough from a time when whooping cough was more common. Self-reported measures, in particular, can suffer from recall bias on the exact nature of the outcome introducing unwanted heterogeneity into the analysis leading to a loss of power and difficulties in the interpretation of the nature of an observational association.

Despite these potential difficulties, genome-wide analysis is a potentially powerful approach to identifying associations between whooping cough and common genetic variation, and this approach has yet to be tested in this context. Therefore, this study aimed to carry out a GWAS using data from two independent cohorts from the UK (1573 cases and 10 649 controls combined) to attempt to identify associations at individual loci and to determine the degree to which common genetic variation influences reported history of whooping cough.

### Results

Overall the proportion of participants who reported a history of whooping cough was 0.11 (95% CIs 0.10–0.12) in ALSPAC and 0.15 (95% CIs 0.14–0.16) in the 1958 birth cohort (Table 1). In ALSPAC, all participants were female and 72% were born between 1958 and 1968 whereas in the 1958 birth cohort, 50.6% were female and all were born in 1958. In ALSPAC, the proportion of participants who reported a positive history varied across year of birth, being highest for participants born pre-1958 and after 1968 and lowest between 1958 and 1967. In the 1958 birth cohort, history of whooping cough was more common in girls than boys. In ALSPAC, there was good agreement between self-reported history of whooping cough reported 8 years after initial data collection. 81 and 89% of subjects with repeat data on positive or negative self-reported history of whooping cough reported consistently at both time periods (Table 2, Cohen’s kappa 0.56).

We carried out a genome-wide association analysis in the ALSPAC cohort and observed a single strong association peak on chromosome 6, with a complementary excess at the extremes of the test statistic distribution and an absence of systematic genome-wide inflation (Fig. 1). The lead SNP, located in the MHC class II region, was rs2760994 [odds ratio (95% CIs) 1.54 (1.37, 1.72), P-value 2.1E−14]. A strong association signal was also observed using subjects with repeated measures collected 8 years later [odds ratio (95% CIs) 1.38 (1.21,1.57), P-value 1.2E−06, n = 4,265]. We followed this up by conducting a genome-wide association analysis in the 1958 birth cohort and observed a complementary association signal [odds ratio (95% CIs) 1.27 (1.13, 1.44), P-value 9.6E−05, Supplementary Material, Figs S1–S3]. After meta-analysis, the lead SNP was located 20 KB away at rs9271768 [effect

### Table 1. History of whooping cough by sex, year of birth and region of birth

<table>
<thead>
<tr>
<th></th>
<th>ALSPAC Overall</th>
<th>ALSPAC Cases</th>
<th>ALSPAC Controls</th>
<th>P-value</th>
<th>1958 Birth Cohort Overall</th>
<th>1958 Birth Cohort Cases</th>
<th>1958 Birth Cohort Controls</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>0.49 (2499)</td>
<td>0.14 (342)</td>
<td>0.86 (2157)</td>
<td>0.013</td>
</tr>
<tr>
<td>Female</td>
<td>1 (7156)</td>
<td>0.11 (815)</td>
<td>0.89 (6341)</td>
<td>—</td>
<td>0.51 (2567)</td>
<td>0.16 (416)</td>
<td>0.84 (2151)</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>1 (7156)</td>
<td>0.11 (815)</td>
<td>0.89 (6341)</td>
<td>—</td>
<td>1 (5066)</td>
<td>0.15 (758)</td>
<td>0.85 (4308)</td>
<td>—</td>
</tr>
</tbody>
</table>

The proportions (counts) of participants within those designated as ‘cases’ and ‘controls’ and overall are shown within strata with a P-value for a chi-square test of association where possible. The symbol ‘—’ indicates no data. A u-shaped association with year of birth was confirmed by second-order polynomial regression (P = 2E−10 for addition of the squared term).

### Table 2. Agreement between self-reported history of whooping cough in the ALSPAC cohort

<table>
<thead>
<tr>
<th>Time</th>
<th>Response</th>
<th>+ 8 years</th>
<th>Yes</th>
<th>No</th>
<th>Unsure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Yes</td>
<td>444 (0.81)</td>
<td>75</td>
<td>28</td>
<td>547</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>138</td>
<td>3608 (0.89)</td>
<td>330</td>
<td>4076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>582</td>
<td>3683</td>
<td>358</td>
<td>4623</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

64% (4623/7156) subjects included in the main analysis had a repeat measure available. Proportions indicate subjects with consistent measures across time points. Similar levels of agreement were observed using all subjects with repeat measures (n = 7084).
allele G, odds ratio (95% CIs) 1.47 (1.35, 1.6), P-value 1.21E−18] in an intergenic region between HLA-DRB1 and HLA-DQA1 (Table 3 and Fig. 2).

We split participants in each data set into groups based on region of birth and re-examined an association between SNP rs9271768 and risk of history of whooping cough because alleles at the HLA region exhibit geographical variation (26), which could correlate with risk of exposure to pertussis (Table 4). While in ALSPAC there was evidence of a weak association between history of whooping cough and region of birth (P = 0.06), no strong evidence of stratification by rs9271768 genotype was found (P = 0.175) and adjustment for region of birth had little or no effect on the strength of the original principal component-corrected association between rs9271768 and history of whooping cough (P = 4.30E−06).

Table 3. Lead SNP associations

<table>
<thead>
<tr>
<th>SNP</th>
<th>Position (build 37)</th>
<th>Effect allele</th>
<th>Frequency</th>
<th>Imputation quality</th>
<th>OR (95% CIs)</th>
<th>P-value</th>
<th>Direction</th>
<th>OR (95% CIs)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2760994</td>
<td>chr1: 257,430</td>
<td>T/C</td>
<td>0.60</td>
<td>0.98</td>
<td>1.54 (1.37, 1.72)</td>
<td>2.11E−14</td>
<td>++</td>
<td>1.41 (1.3, 1.53)</td>
<td>0.092</td>
</tr>
<tr>
<td>rs9271768</td>
<td>chr1: 259,418</td>
<td>G/A</td>
<td>0.52</td>
<td>0.88</td>
<td>1.54 (1.37, 1.72)</td>
<td>1.05E−13</td>
<td>++</td>
<td>1.47 (1.35, 1.6)</td>
<td>0.272</td>
</tr>
</tbody>
</table>

Odds ratios (OR) and confidence intervals (CI) are shown for a one unit increase in the number of effect alleles at each SNP within each cohort and after meta-analysis.
We also conditioned on SNP rs9271768 in both cohorts and meta-analysed the results for SNPs within the xMHC but did not obtain clear evidence of further association (\(P > 1 \times 10^{-6}\) after meta-analysis, Supplementary Material, Table S3). The lead SNPs within each cohort while located within 20 KB of each other differed slightly in their location and were in weak-to-moderate linkage disequilibrium (Supplementary Material, Table S4). Therefore, we examined whether they represented distinct signals. Conditional analysis indicated that correcting for the lead SNP in the ALSPAC cohort could account for any additional signal at the lead 1958 BC cohort SNP and vice versa (Supplementary Material, Table S5). We also examined epistasis among the lead SNPs. The inclusion of two-way interaction terms did not result in a better fitting model than a model including two SNPs alone (\(P\)-value >0.2). At these SNPs, conditional analysis in the ALSPAC cohort, stratifying on year of birth as an instrument for vaccination status to test the hypothesis of environmental mediation by this factor, did not indicate evidence of effect modification (Supplementary Material, Table S6).

Classical HLA alleles imputed to four-digit resolution using either HLA-IMP*02 or SNP2HLA in ALSPAC showed concordant effects sizes (Supplementary Material, Fig. S4). Haplotype analysis of imputed HLA alleles resulted in a strong association peak over the HLA class II alleles in both cohorts (Supplementary Material, Tables S7 and S8) with evidence of association at HLA-DQA, HLA-DQB, HLA-DRB1 and HLA-DRB3/4/5. The most strongly associated locus was HLA-DQB1 \((P\)-value 7.8E\(−16\)) with strongest individual effects for alleles HLA-DQB1*0602, HLA-DQB1*0501 and HLA-DRB1*0301 \((P\)-value <1.E\(−5\)). Addition of the HLA class II alleles to a model already including the lead SNP rs9271768 indicated the majority of signal at these alleles could be attributed to the lead SNP association (Supplementary Material, Table S7). Strong association signals for amino acids within the HLA class II alleles were also observed. The strongest evidence of association was for a set of amino acids at position 13 in HLA-DRB1 (presence of either serine or histidine versus all others, \(P\)-value 7.86.E\(−17\)). Conditioning on lead SNP rs9271768 resulted in most of the signal being explained (\(P\)-value 1.03E\(−03\), Supplementary Material, Table S9).

Estimates of the variance explained of all SNPs within the extended MHC were 0.007 (SE 0.003, \(P\)-value 9.E\(−5\)) in the ALSPAC cohort and 0.007 (SE 0.004, \(P\)-value 0.01) in the 1958 birth cohort, respectively, whereas the variability explained by all SNPs genome-wide were 0.11 (SE 0.05, \(P\)-value 0.01) and 0 (SE 0.07, \(P\)-value 0.5). Genome-wide gene-based tests of association confirmed the single SNP analyses giving a strong association signal at the MHC class II loci HLA-DQB1, HLA-DRB1, HLA-DQA1 and TAP2 whereas outside the MHC region, gene-based and gene-set...
### Table 4. Test of confounding of lead association by region of birth

<table>
<thead>
<tr>
<th>Study</th>
<th>Region of birth</th>
<th>History of whooping cough</th>
<th>rs9271768 genotype</th>
<th>rs9271768 G counts (positive history of whooping cough/total)</th>
<th>Regression coefficient*</th>
<th>Overall P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>cases</td>
<td>controls</td>
<td>P-value*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AA</td>
<td>AG</td>
<td>GG</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALSPAC</td>
<td>East/West Ridings</td>
<td>0.02 (155)</td>
<td>0.11 (17)</td>
<td>0.89 (138)</td>
<td>0.06</td>
<td>0.1751</td>
</tr>
<tr>
<td></td>
<td>Eastern</td>
<td>0.04 (280)</td>
<td>0.12 (27)</td>
<td>0.9 (253)</td>
<td>0.257 (72)</td>
<td>0.265 (41)</td>
</tr>
<tr>
<td></td>
<td>London</td>
<td>0.06 (403)</td>
<td>0.12 (50)</td>
<td>0.88 (353)</td>
<td>0.273 (110)</td>
<td>0.281 (80)</td>
</tr>
<tr>
<td></td>
<td>Midlands</td>
<td>0.04 (270)</td>
<td>0.14 (37)</td>
<td>0.86 (233)</td>
<td>0.270 (110)</td>
<td>0.258 (40)</td>
</tr>
<tr>
<td></td>
<td>North Midlands</td>
<td>0.01 (68)</td>
<td>0.06 (4)</td>
<td>0.94 (64)</td>
<td>0.25 (17)</td>
<td>0.213 (86)</td>
</tr>
<tr>
<td></td>
<td>Northern</td>
<td>0.02 (120)</td>
<td>0.19 (23)</td>
<td>0.81 (97)</td>
<td>0.183 (22)</td>
<td>0.204 (56)</td>
</tr>
<tr>
<td></td>
<td>Northwest</td>
<td>0.04 (234)</td>
<td>0.12 (29)</td>
<td>0.88 (205)</td>
<td>0.244 (57)</td>
<td>0.231 (94)</td>
</tr>
<tr>
<td></td>
<td>Scotland</td>
<td>0.02 (103)</td>
<td>0.17 (18)</td>
<td>0.83 (85)</td>
<td>0.243 (25)</td>
<td>0.282 (29)</td>
</tr>
<tr>
<td></td>
<td>Southeast</td>
<td>0.01 (93)</td>
<td>0.11 (10)</td>
<td>0.89 (83)</td>
<td>0.376 (35)</td>
<td>0.183 (25)</td>
</tr>
<tr>
<td></td>
<td>Northern</td>
<td>0.04 (244)</td>
<td>0.09 (21)</td>
<td>0.91 (223)</td>
<td>0.225 (55)</td>
<td>0.213 (52)</td>
</tr>
<tr>
<td></td>
<td>Southwest</td>
<td>0.66 (4193)</td>
<td>0.11 (458)</td>
<td>0.89 (3715)</td>
<td>0.275 (1152)</td>
<td>0.229 (29)</td>
</tr>
<tr>
<td></td>
<td>Wales</td>
<td>0.03 (205)</td>
<td>0.1 (23)</td>
<td>0.89 (184)</td>
<td>0.351 (72)</td>
<td>0.195 (40)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1 (6368)</td>
<td>0.11 (715)</td>
<td>0.89 (5653)</td>
<td>0.273 (1738)</td>
<td>0.223 (1417)</td>
</tr>
</tbody>
</table>

Three tests of association were carried out: between region of birth and history of whooping cough, region of birth and rs9271768 genotype and between rs9271768 genotype and history of whooping cough stratified by region of birth.

The distribution of proportions and their corresponding counts are shown within cells.

*Rate of change of proportion with number of G alleles of rs9271768.

**Test of trend between counts of the G allele of rs9271768 and history of whooping cough.

*Overall test of trend between counts of the G allele of rs9271768 and history of whooping cough after adjusting for region of birth.

**Proportion of participants in each strata.

*Chi-square test of the association between region of birth and counts of cases and controls.

*Chi-square test of the association between region of birth and counts of rs9271768 genotype.
analysis did not result in identifiable additional associations (Supplementary Material, Tables S10 and S11). Candidate SNP analysis at SNPs at the TLR4 gene (rs2770150, rs4986790), which have been previously shown to be associated with antibody response to pertussis toxin, did not show evidence of association with reported history of whooping cough in this sample (Supplementary Material, Table S12).

Discussion

This study applied genome-wide association analysis to reported history of whooping cough. We observed evidence of association within the MHC class II region with the strongest signal at SNP rs9271768. The genomic location of the lead association signal suggests a role for antigen binding and presentation in its etiology as extracellular proteins from invading pathogenic bacteria are primarily processed via class II molecules. Biochemical data indicate molecules encoded by the HLA class II loci respond biologically to pertussis. Cell surface expression of HLA-D molecules on human monocytes is modulated by pertussis infection (27), and B. pertussis peptides require HLA-DR1 to elicit a response (28). Furthermore, genetic variation at HLA-DQ1 has been shown to modify CD4+ T-cell response, upon stimulation with pertactin in vitro (29).

We observed strong overlapping signals in both cohorts where lead SNPs within each cohort and after meta-analysis were located within 20 KB of each other in the MHC class II region. Pinpointing the causal mutation definitively may have been made more challenging owing to the complex patterns of linkage disequilibrium across the MHC. Imputed alleles at the MHC class II loci were found to be strongly associated with our outcome although no more strongly associated than the lead SNP. Identifying the nature of the genetic effect in the region and its potential relationship with HLA antigen loading represents an important next step in understanding this association.

Our attempts to confirm whether additional variation was present in our sample by testing all SNPs simultaneously and using genome-wide gene-based tests and gene-set tests were inconclusive. This sample may have been underpowered to detect additional associations. There are alternative explanations. Historical accounts indicate that pertussis emerged in Europe only in the last 500 years (30), and sequence diversity measures indicate that B. pertussis has a relatively recent evolutionary origin (31). A possible explanation for the lack of additional genetic associations might be that too little time has passed for B. pertussis to exert an appreciable selective pressure. Another possible explanation is that host adaptation by B. pertussis is thought to have been mediated by large-scale gene loss and inactivation (32), which may have led to a small number of pathogenic loci that are responsible for determining host susceptibility. Either way the estimates of variance explained by all SNPs represent a lower bound for the total additive genetic variability explained owing to incomplete LD between directly typed SNPs and causative mutations (33), and by examining all common SNPs jointly, power to detect the influence of a small number of moderate to weak effects is less compared with the joint effect of true causal loci alone. Furthermore, our results are not well powered to inform on non-additive and/or rare variations. Nevertheless, our results suggest that effect sizes at the MHC class II region are relatively larger than those from other genomic regions in these population-based samples.

Genome-wide association studies can be biased by population stratification and errors in phenotype definition (34). We applied two methods of correction for potential bias owing to population substructure; principal component analysis of genome-wide data, a standard approach in GWAS that adjusts cases and controls for differences in ancestry genetic variation (35) and traditional epidemiological tests of confounding. Both methods indicated a similar strength of association between SNP rs9271768 and self-reported history of whooping cough, which indicates population stratification is unlikely to have led to a spurious association in these samples. We have been careful to describe our phenotype as self-reported history of whooping cough to acknowledge its potential heterogeneity. Obtaining a large enough number of cases to robustly determine a common genetic association was complicated by the widespread and effective control of pertussis in developed countries. In practice, this meant we were limited to cohorts with historical data. Furthermore, data were collected by either self-retrospective report or parent reported questionnaire. While a proportion of these would have been doctor-diagnosed cases, some cases of whooping cough involve Bordetella parapertussis (36–38) and some viruses such as adenovirus and parainfluenza virus can cause pertussis-like syndrome in children leading to differential diagnosis (39-40).

The above limitations may mean that, while we believe this is a replicable association, robust to confounding, in theory it is possible the association is partly driven by a whooping cough like syndrome caused by agents different to B. pertussis or possibly the observed association acts mechanistically through vaccine-driven susceptibility (by effect modification).

In summary, our data indicate that common genetic variation in the MHC class II region influences susceptibility to self-reported history of whooping cough. Examination of this genetic variant in laboratory-confirmed cases and its possible correlation with pertussis endophenotypes or response to vaccination is an important next step to understand this association.

Material and Methods

Study cohorts

ALSPAC

The Avon Longitudinal Study of Children and Parents (ALSPAC) is a longitudinal birth cohort with a sampling frame of all pregnant women with an expected delivery date between April 1991 and December 1992 within a well-defined geographic area in the Southwest UK including a major city (Bristol). Initially, 14 541 pregnant women (and their children) were recruited and over the 20 years of follow-up, the majority are still participating in questionnaires that are sent out at regular intervals and by attending study clinics (41). The study website contains details of all the data that are available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). Mothers completed a self-report questionnaire at approximately the twelfth week of pregnancy, which included a question on history of whooping cough infection. Mothers were asked ‘Have you ever had any of the following infections’ and were presented with a list of common infections including whooping cough to which they could respond either ‘Yes’, ‘No never’ or ‘I don’t know’. The same question was asked again in a separate questionnaire 8 years later. Participants’ region of birth was identified from the question ‘Where were your parents living at the time you were born?’ with spaces left blank for town, county and country. Postal area was extracted from this information and placed in categories matching the 1958 birth cohort definitions. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.
1958 birth cohort
The 1958 British Birth cohort (1958 birth cohort) also known as the National Child Development Study is a longitudinal birth cohort with a sampling frame of all those born in a particular week in March in 1958 in England, Scotland and Wales. Initially, 17,000 people were recruited, the majority of whom have participated in regular follow-up surveys and a biomedical survey (42). Detailed documentation about the 1958 Birth Cohort, including a fully searchable data dictionary, can be found online (http://www.cls.ioe.ac.uk/). When participants were aged between 7 and 8 years, their parent completed a questionnaire by interview, which included the question ‘What infectious diseases has the child had, and at what ages’. Parents were asked to select from a list of common childhood diseases including whooping cough and could indicate ‘Yes’, ‘No’ and ‘Don’t know’. Region of residence of Great Britain was defined by postcode (26).

Genotype quality control and imputation
PLINK (43) (v1.07) was used to carry out quality control filtering on directly typed data from both cohorts. SNPs were removed if they had >5% missingness or Hardy–Weinberg equilibrium P < 1.0E − 6. Additionally, SNPs with MAF of <1% were removed. Samples were excluded if they had >5% missingness, indeterminate X-Chromosome heterozygosity or extreme autosomal heterozygosity. Analysis was restricted to individuals with European ancestry. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity-by-state pairwise distances using the CEU, YRI, JPT and CHB HapMap populations as a reference and were then excluded. Cryptic relatedness was defined as a Pi hat value of >0.125, which is expected to correspond to roughly 12.5% of alleles being shared IBD or related at a first-cousin level. Phasing was carried out in Shapeit2 (44) (v2.r644), and data were imputed to the 1000 Genomes Phase I Integrated Release (version 3) in IMPUTE2(v2.3) (45).

ALSPAC
Genome-wide data of common genetic variation were generated at The Centre National de Génotyage (CNG) on the Illumina Human 660W-Quad array, and genotypes were called with Illumina GenomeStudio. Out of an initial set of 10,015 subjects and 557,124 directly genotyped SNPs, 9048 subjects and 526,688 SNPs passed all other quality control thresholds. We combined 477,482 SNPs of these subjects with genotype data of a sample of 9115 children of the ALSPAC and 1958 Birth Cohort data, respectively. Alleles were imputed to a four-digit level, apart from HLA-DRB1 paralogs, which were to a two-digit level. We removed genotypes with a posterior probability under 0.70. Additionally, SNPs with MAF of <1% were removed. Samples were excluded if they had >5% missingness, indeterminate X-Chromosome heterozygosity or extreme autosomal heterozygosity. Analysis was restricted to individuals with European ancestry. Samples showing evidence of population stratification were identified by multidimensional scaling of genome-wide identity-by-state pairwise distances using the CEU, YRI, JPT and CHB HapMap populations as a reference and were then excluded. Cryptic relatedness was defined as a Pi hat value of >0.125, which is expected to correspond to roughly 12.5% of alleles being shared IBD or related at a first-cousin level. Phasing was carried out in Shapeit2 (44) (v2.r644), and data were imputed to the 1000 Genomes Phase I Integrated Release (version 3) in IMPUTE2(v2.3) (45).

Statistical analysis
A genome-wide association analysis was carried out in each set of directly genotyped and imputed bi-allelic markers within each cohort using an additive logistic model, log(odds) = a + bX + c where X is the allelic dosage of the effect allele (0, 1 or 2 copies of b) at the nth SNP and c are study-specific covariates. These were year of birth (first and second order) and the top ten principal components of population ancestry (PCs) in ALSPAC and sex, a study-specific indicator and the ten PCs in the 1958 Birth Cohort. Within-cohort association analyses were carried out in SNPTEST (v2.4.0) (48). Effects were matched by marker name and were combined using inverse variance-based meta-analysis in METAL (49). Regional plots were drawn in LocusZoom (50).

Analysis of imputed HLA alleles
In haplotype analysis to examine whether associations with SNP data could be mapped to variation in the classical HLA alleles, we
used multiple logistical regression that would allow the addition of covariates and assumed additivity in the effects at a haplotype at a locus (53). We used the same within-cohort covariates as described earlier. Statistical details of this analysis are given in Supplementary Material, Section S5. Additional variation in the region was analysed using imputed allele dosage in PLINK and a logistic regression model for the presence/absence of a marker plus the same within-cohort covariates as described earlier.

GCTA, gene-based and gene-set analysis

We set out to assess whether additional regions of the genome were associated with our phenotype, which could only be observed by assessing SNPs jointly. The variance explained by all directly genotyped SNPs and SNPs within the extended MHC region correcting for the same covariates as single SNP analysis was estimated using GCTA (54). Gene associations were calculated in a Versatile Gene-Based Test for Genome-wide Association Studies (VEGAS, v0.8.27) (55). An empirical P-value of P < 2.8⋅6 was taken as a conservative threshold for identifying sites of interest (a Bonferroni-corrected threshold of 0.05/17 787 tested genes). Gene-set analysis, to identify sets of variants affecting susceptibility across genes, was carried out in Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA, v2.4) using all available gene sets (56). A Bonferroni correction was used to correct for multiple testing owing to the number of gene sets being analysed (P < 1.55E – 5, n = 3216). SNP P-values after meta-analysis were used for both gene-based and gene-set analyses. For gene-based and pathway analysis, we filtered out SNPs with an INFO score of <0.5 and MAF of <0.01.

Supplementary Material

Supplementary Material is available at HMG online.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. This publication is the work of the authors and George McMahon and Nic Timpson will serve as guarantors for the contents of this paper.

Conflict of Interest statement. None declared.

Funding

The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. G.D.S., N.J.T. and S.M.R. work in the Medical Research Council Integrative Epidemiology Unit at the University of Bristol, which is supported by the Medical Research Council (grant numbers MC_UU_12013/1 and MC_UU_12013/3) and the University of Bristol. This work made use of data and samples generated by the 1958 Birth Cohort (NCDS). Access to these resources was enabled via the 58READIE Project funded by Wellcome Trust and Medical Research Council (grant numbers WT095219MA and G1001799). A full list of the financial, institutional and personal contributions to the development of the 1958 Birth Cohort Biomedical resource is available at https://www2.le.ac.uk/projects/birthcohort. Genotyping was undertaken as part of the Wellcome Trust Case–Control Consortium (WTCCC) under Wellcome Trust award 076113, and a full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

References


