Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections

Chao Tian,[1]* Bethann S Hromatka,[1] Amy K Kiefer,[1] Nicholas Eriksson,[1] Joyce Y Tung, [1] David A. Hinds,[1]**

[1]23andMe, Mountain View, California, United States of America.

*Corresponding authors

Email address:

*CT: ctian@23andme.com
**DAH: dhinds@23andme.com

ABSTRACT

We performed 23 genome-wide association studies for common infections, including chickenpox, shingles, cold sores, mononucleosis, mumps, hepatitis B, plantar warts, positive tuberculosis test results, strep throat, scarlet fever, pneumonia, bacterial meningitis, yeast infections, urinary tract infections, tonsillectomy, childhood ear infections, myringotomy, measles, hepatitis A, rheumatic fever, common colds, rubella and chronic sinus infection, in more than 200,000 individuals of European ancestry. For the first time, genome-wide significant associations ($P < 5 \times 10^{-8}$) were identified for many common infections. The associations were mapped to genes with key roles in acquired and innate immunity (*HLA*, *IFNA21, FUT2*, *ST3GAL4*, *ABO*, *IFNL4, LCE3E, DSG1, LTBR*, *MTMR3*, *TNFRSF13B*, *TNFSF13B*, *NFKB1*, *CD40*) and in regulation of embryonic developmental process (*TBX1*, *FGF, FOXA1 and FOXN1*). Several missense mutations were also identified (in *LCE5A, DSG1, FUT2, TBX1, CDHR3, PLG, TNFRSF13B, FOXA1, SH2B3, ST5* and *FOXN1*). Missense mutations in *FUT2* and *TBX1* were implicated in multiple infections. We applied fine-mapping analysis to dissect associations in the human leukocyte antigen region, which suggested important roles of specific amino acid polymorphisms in the antigen-binding clefts. Our findings provide an important step toward dissecting the host genetic architecture of response to common infections.

Introduction

Infectious diseases, the second leading cause of death worldwide, represent persistent challenges to human health due to increasing resistance to established treatments, lack of life-saving vaccines and medications in developing countries, and increasing distribution[1]. Studies have linked susceptibility to infectious agents to cancers, autoimmune diseases, and drug hypersensitivity. Human papillomaviruses are associated with multiple cancers[2]; rubella and mumps infection have been linked to development of type 1 diabetes (T1D) in children[3]; and risk of developing multiple sclerosis (MS) was greater among patients with herpes zoster (shingles) than in matched controls [4]. Reactivation of chronic persistent human herpesviruses has been linked to drug-induced hypersensitivity [5]. Thus, infectious diseases have a profound impact on our health, both directly as well as through connections with other diseases. Nevertheless, genetic studies of common infectious diseases lag somewhat behind those of other major complex diseases. Few genome-wide association studies (GWAS) have been undertaken for common infectious diseases of lower mortality or for infectious diseases for which vaccines are available [6].

In this study, we conducted 23 imputed GWASs with more than 200,000 European research participants who were genotyped on the 23andMe platform and were asked to report on their history of infections (Table 1 and Supplementary note). Classical human leukocyte antigen (HLA) loci have been prototypical candidate genetic susceptibility to multiple infectious diseases[7]. We further imputed and tested the HLA classical alleles and amino acid polymorphisms to dissect independent HLA signals for multiple common infections that have significant HLA associations in our GWASs.

Results

In total, 59 genome-wide significant (GWS) regions ($P < 5 \times 10^{-8}$) were discovered (Table 2, see also Supplementary Figure 1 and Supplementary Figure 2 for GWAS Manhattan plots and regional plots). The HLA region is significantly associated with 13 of the infections that we studied and we dissected independent signals in each association (Figure 1 and Table 3, see also Supplementary Figure 3 for HLA regional plots). In the later sections, we describe the result for each phenotype in detail.

Genetic correlations

The pairwise LD score correlation on GWAS statistics (Figure 2, Supplementary Table 2) showed significant positive genetic correlations between strep throat and tonsillectomy (r_g = 0.85, se = 0.09, P = 1.89×10^{-21}), between childhood ear infection and myringotomy (r_g = 0.88, se = 0.14, P = 1.27×10^{-9}), and between chickenpox and shingles (r_g = 0.56, se = 0.16, P = 4×10^{-4}), which are pairs of related phenotypes. A highly significant genetic correlation was found between tonsillectomy and childhood ear infections (r_g = 0.42, se = 0.05, P=8.19×10^{-15}). Tonsillitis and ear infections are sometimes comorbid symptoms. A previous twin study suggested a substantial overlap in genetic factors influencing variations in liability to ear infection and tonsillitis[8]. In our GWASs, we identified the missense mutation (N397H) in *TBX1* as a significant susceptibility locus for both. Significant positive correlations were also observed for pneumonia with chronic sinus infections (rg = 0.72, se = 0.19, P = 1.0×10^{-4}), childhood ear infections (rg = 0.39, se = 0.07, P = 1.47×10^{-8}), colds last year (rg = 0.57, se = 0.12, P = 2.66×10^{-6}) and strep throats (rg = 0.69, se = 0.13, P=1.0×10^{-7}). These are all upper respiratory infections. The same viruses that cause colds and sore throat (if these infect the throat, sinuses and upper respiratory tract) can also cause pneumonia (if these reach the lungs). We also saw a high genetic correlation between urinary tract

infection (UTI) and yeast infections (rg = 0.68, se = 0.08, P = 4.30 x 10^{-16}). There was a phenotypic correlation (Pearson r = 0.41, P < 2.2 x 10^{-16}) among women who reported both UTI and yeast infections in our cohort. These two infections can have similar symptoms and both cause discomfort in vaginal area, but different infectious agents cause them. The genetic correlation from LD score regression removes the correlation in GWAS summary statistics due to the correlation in phenotypes and thus suggests a shared host genetic susceptibility to UTI and yeast infections in females.

Chickenpox (Varicella-zoster virus, Herpesviruses family)

Chickenpox, characterized by red, itchy bumps on the skin, is a highly contagious disease caused by primary infection with varicella-zoster virus (VZV). After the initial chickenpox infection, the virus remains dormant in the nervous system, though in approximately 20% of people it reactivates and manifests as shingles (see below). We identified independent associations in HLA-A and HLA–B in the class I region. On conditional analysis within HLA-A, the amino acid polymorphism HLA-A-Gly107 (P = 3.77 x 10^{-10}) accounted for the HLA allele association (HLA-A*02:01, P = 1.08 x 10^{-9}, conditional p-value is 0.90) in this interval. Although HLA-A-Gly107 is not located in the peptide-binding cleft, it is in high LD with HLA-A-Asp74, HLA-A-Gly62 and HLA-A-V95 (r^2 > 0.95), which are all in the peptide-binding cleft. In HLA-B, only rs9266089 (P = 1.00 x 10^{-10}) met the threshold for GWS.

Shingles (Varicella-zoster virus, Herpesviruses family)

Shingles, also known as herpes zoster, is characterized by a painful band-shaped rash [9] and is caused by reactivation of the VZV (see above) . We identified multiple independent HLA signals for shingles. Within HLA-A, the amino acid polymorphism HLA-A-Arg97 (P = 2.75 x 10^{-22}), located in the peptide-binding groove, accounted for most of the SNP effect (rs2523815, P = 1.83 x 10^{-22}, conditional p-value is 0.04) and the HLA allele effect (HLA-

A*02:01, P = 8.91 x 10^{-20}, conditional p-value is 0.06). HLA-A*02:01 is implicated in both our chickenpox and shingles GWASs. The amino acid associations are different, but HLA-A-Arg97 is in high LD with HLA-A-Gly107 (r^2 = 0.73). Within HLA-B, rs2523591 (P = 1.74 x 10^{-27}) had a stronger association than HLA classical variants. There was a significant residual effect for rs2523591 after conditioning on HLA allele (conditional p-value is 1.71 x 10^{-9}) or amino acid (conditional p-value is 3.13 x 10^{-5}) associations in HLA-B region. The SNP rs2523591 is in LD with rs9266089 (r^2 = 0.24, D' = 0.98), which is a GWS signal in chickenpox GWAS, suggesting potential shared genetic factors for chickenpox and shingles. After adjusting for the signals in HLA-A and HLA-B regions, we observed independent secondary signals in the class III region (rs41316748, P=1.44 x 10^{-12}) and in the class II region index by DRB1-PheSerHis13 (P= 5.9 x 10^{-10}), which is in the peptide-binding cleft of HLA-DRB1.

We identified a variant upstream of *IFNA21* (rs7047299, P = 1.67 x 10^{-8}) as a GWS association with shingles. None of the variants within 500kb and in moderate LD (r^2 > 0.6) with the index SNP rs7047299 were coding, nor were they reported as expression quantitative trait loci (eQTL). *IFNA21*, a member of the alpha interferon gene cluster at band 9p21, encoded type I interferon and is mainly involved in innate immune response against viral infection. It have been shown to be involved in the pathogenesis of rubella[10], and may also influence susceptibility to asthma and atopy[11].

Cold sores (type 1 herpes simplex virus, Herpesviruses family)

Over half of the US population suffers from cold sores (herpes simplex labialis), which are most commonly caused by the herpes simplex virus type I (HSV-1) [12]. Once someone has been infected, the virus usually cannot be eliminated and lies dormant in nerve cells where it may reactivate years later. Although most people are infected with HSV at some

point in their lives, it is not clear why only some people suffer from cold sore outbreaks and reactivation [13]. We identified two independent GWS SNP associations in the HLA class I region. One is 2kb upstream of *POU5F1* indexed by rs885950 ($P = 7.47 \times 10^{-13}$) and the other is upstream of *HCP5* indexed by rs4360170 ($P = 3.41 \times 10^{-9}$). We also found a GWS association with HLA-B-ThrGly45 in the peptide-binding cleft of the HLA-B protein ($P = 4.91 \times 10^{-12}$). Upon conditioning on HLA-B-ThrGly45, the two SNPs had significant residual effects (conditional p-values are 6.28×10^{-9} and 1.32×10^{-5}). However, the effect of HLA-B-ThrGly45 was largely removed (conditional p-value is 0.001) after conditioning on the two SNPs.

Mononucleosis (Epstein-Barr virus, Herpesvirus family)

Over 90% of the world's adult population is chronically infected with Epstein-Barr virus (EBV). Primary infection at childhood is usually asymptomatic or with only mild symptoms. Primary infection later in life is often accompanied by infectious mononucleosis (IM) [14], commonly referred to as "mono" and characterized by fever, tonsillitis, and fatigue. A longitudinal study of IM in a cohort of 2,823,583 Danish children found evidence of familial aggregation of IM that warrants GWAS on IM disease etiology [14]. We identified the HLA class I region upstream of *HCP5* as a GWS association (indexed by rs2596465, $P = 2.48 \times 10^{-9}$). No GWS HLA allele or amino acid variant was identified.

Mumps (mumps virus)

Mumps is an illness caused by the mumps virus. There is no disease modifying treatment for mumps infection. It is a rare disease in developed countries since the introduction of routine vaccination; however, outbreaks can still occur and mumps remains a significant health threat in developing countries [15]. We identified four independent GWS associations with mumps.

Variation in HLA-A, in the class I region, was significantly associated with mumps. The amino acid polymorphism HLA-A-Gln43 ($P = 2.51 \times 10^{-17}$) accounted for most of the SNP association (rs114193679, $P = 2.23 \times 10^{-17}$, conditional p-value is 0.19) and the HLA allele association (HLA-A*02:05, $P = 1.73 \times 10^{-17}$, conditional p-value is 0.26) in the HLA-A region.

We identified *FUT2* (indexed by rs516316, $P = 9.63 \times 10^{-72}$) and *ST3GAL4* (indexed by rs3862630, $P = 1.21 \times 10^{-8}$) as highly significant associations with mumps. Both belong to the glycosphingolipid (GSL) biosynthesis pathway, which was the most significant MAGENTA-analyzed pathway (Table 4: $P = 1 \times 10^{-4}$ and FDR = 0.003) identified using mumps GWAS data. This pathway also includes the *ABO* gene (indexed by rs8176643, $P = 5.81 \times 10^{-5}$). Glycosphingolipids (GSLs), found in the cell membranes of bacteria all the way to humans, play important biological roles in membrane structure, host-pathogen interactions, cell-cell recognition, and modulation of membrane protein function. Genetic variation in the human *FUT2* gene determines whether ABO blood group antigens are secreted into body fluids[16]. Detailed functional annotations of the genes identified in our mumps GWAS are provided in Supplementary table 3. The high-risk allele rs516316 (C) is in complete LD with rs601338 (A), a nonsense variant in *FUT2* that encodes the "non-secretor" (se) allele, which has been reported to provide resistance to Norovirus [17] and susceptibility to Crohn's disease [18,19] and T1D [20]. The surface glycoprotein of mumps virus, hemagglutinin-neuraminidase, attaches to sialic acid receptors and promotes fusion and viral entry into host cells [21]. ABO antigens can modulate the interaction between pathogens and cell surface sialic acid receptors [22]; this presents a plausible mechanism by which secretor status could modify susceptibility to mumps infections. *ST3GAL4* encodes a member of the glycosyltransferase 29 family, a group of enzymes involved in protein glycosylation.

We also identified a significant association between mumps and a variant on chromosome 14 (rs11160318, $P = 4.56 \times 10^{-12}$) that is located in the intergenic region upstream of *BDKRB2*. The role of *BDKRB2* in mumps susceptibility is not clear, but the *BDKRB2* product has high affinity for intact kinins, which mediate a wide spectrum of biological effects including inflammation, pain, vasodilation, and smooth muscle contraction and relaxation [23].

Hepatitis B (hepatitis B virus)

Chronic hepatitis B virus (HBV) infection is a challenging global health problem. Several GWASs have been carried out to identify genetic loci involved in HBV susceptibility and immune response to hepatitis B vaccine, and the HLA-DRB1- HLA-DQB1 region has been consistently shown to be GWS[24–26]. In our GWAS of HBV infection we identified a GWS association with rs9268652 ($P = 3.14 \times 10^{-9}$, in HLA-DRA), consistent with the association of the HLA class II region with HBV susceptibility. Fine mapping analysis failed to identify GWS associations with HLA classical variants, but DQB1*06:02 was moderately associated with hepatitis B ($P = 5.47 \times 10^{-6}$). HLA-DQB1*06:02 is protective against HBV in our data, which is consistent with previous studies that the DRB1*15:01-DQB1*06:02 haplotype is associated with protection from type 1 autoimmune hepatitis and development of hepatocellular carcinoma[27]. Although chronic infection with the HBV has been linked to the development of hepatocellular carcinoma for more than 30 years[28], the mechanism by which HBV infection leads to hepatocellular carcinoma is unclear.

Hepatitis A (hepatitis A virus)

Infection with hepatitis A virus (HAV) can be asymptomatic or result in an acute illness characterized by flu-like symptoms and jaundice. Serious complications are rare and the hepatitis A vaccine is effective for prevention[29]. No large cohort study has been done for

HAV infection. Our GWAS showed a suggestive association with *IFNL4*, indexed by rs66531907 ($P = 5.7 \times 10^{-8}$, OR = 1.23, 1kb upstream of *IFNL4*). *IFNL4* is located upstream of *IFNL3* (also known as *IL28B*). The high-risk allele rs66531907 (C) in our hepatitis A GWAS is in almost complete LD with rs8099917 (T) ($r^2 = 0.992$), which has been associated with progression to chronic hepatitis C infection (HCV) and poor response to HCV therapy in multiple studies[30].

Plantar warts (human papillomavirus)

The human papillomavirus (HPV) causes plantar warts, which occur on the soles of the feet. Some strains of HPV are also associated with certain forms of cancer [2]. We identified two GWS associations with plantar wart. Our first association was with the HLA-DRB1 in class II region. The amino acid polymorphism DRB1_PheSerTyr13 (P = 1.84 x 10-39), in the peptide binding clefts of HLA-DRB1, was highly significant. Conditioning on it, the effects of HLA-DRB1*15:01 (P=4.48e-14, conditional p-value is 0.36), HLA- DRB1*04:01 ($P = 4.88 \times 10^{-18}$, conditional p-value is 0.001), and rs9272050 ($P = 9.66 \times 10^{-31}$, conditional p-value is 0.06) were largely removed, but a small residual effect remained for rs28752534 ($P = 1.50 \times 10^{-30}$, conditional p-value is 2.15×10^{-8}).

We also identified a GWS association with rs6692209 ($P = 5.25 \times 10^{-9}$) near *LCE3E* in the Epidermal Differentiation Complex (EDC) on 1q21. The EDC comprises a remarkable density of gene families that determine differentiation of the human epidermis[31]. The high-risk allele rs6692209 (T) is in high LD ($r^2 = 0.86$) with rs2105117 (A) ($P = 2.98 \times 10^{-8}$ in our GWAS), a missense variant in *LCE5A*. Variants in late cornified envelope (LCE) genes have been implicated in GWASs of atopic dermatitis (rs3126085-A, $P = 6 \times 10^{-12}$, $r^2 = 0.047$ with rs6692209) [32] and psoriasis (rs4085613-A, $P = 7 \times 10^{-30}$, $r^2 = 0.002$ with rs6692209) [33], although the index SNPs identified in those GWASs were in low LD with our identified

region. The skin is the primary interface with the external environment and provides a critical first barrier of the innate immune defense to infections[34]. While studies have linked the epidermal differentiation genes to various skin diseases, the role of the LCE proteins in epidermal biology has not been extensively studied. One study suggested that LCE proteins of groups 1, 2, 5 and 6 are involved in normal skin barrier function, while *LCE3* genes encode proteins that are involved in barrier repair after injury or inflammation[35].

Positive TB test (Mycobacterium tuberculosis)

Tuberculosis is a common and, in many cases, lethal infectious disease caused by various strains of mycobacteria, usually Mycobacterium tuberculosis. A chest x-ray and a sample of sputum are needed for a diagnosis of active TB disease [36]. The phenotype used in our GWAS was defined by asking customers whether they had had a positive response (red bumps) to a tuberculosis (TB) skin test. A positive TB test indicates that a person has at some time been exposed to and infected with TB bacteria. A negative TB test indicates that latent TB infection or TB disease is unlikely. In our GWAS of positive TB tests, the cases have either active TB disease, latent tuberculosis infections, or were vaccinated against TB, while controls may or may not have been exposed to TB bacteria and have never been infected.

We identified multiple independent HLA associations. In the most significant HLA-DRA-DQA1 region, conditioning on HLA-DRB1-Leu67 ($P = 1.14 \times 10^{-29}$) and HLA-DQA1-His129 ($P = 1.56 \times 10^{-13}$) removed the effect of DQA1*01:02 ($P = 7.04 \times 10^{-22}$, conditional p-value is 0.54), DQA1*03:01 ($P = 3.99 \times 10^{-10}$, conditional p-value is 0.36), and rs3135359 ($P = 8.82 \times 10^{-21}$, conditional p-value is 0.06); however, the effect of rs2894257 ($P = 8.16 \times 10^{-36}$, conditional p-value is 1.2×10^{-7}) remained significant. Upon conditioning on the signals in the HLA-DRA-DQA1 region, secondary signals were detected in the class I (HLA-B*08:01, $P=8.43 \times 10^{-14}$) and class III regions (rs148844907 in C6orf47, $P=2.21 \times 10^{-12}$).

Strep throat (group A streptococcus bacteria)

Strep throat (also called streptococcal tonsillitis or streptococcal pharyngitis) is characterized by fever, headache, severe sore throat, swollen tonsils and lymph nodes, and is caused by group A streptococcus (GAS) bacteria [37]. It is routinely treated with antibiotics. The symptoms of untreated strep throat typically resolve within a few days, but can sometimes cause complications such as rheumatic fever and kidney inflammation. We found a HLA-B class I region (index SNP rs1055821, P = 7.69 x 10^{-11}) as a GWS association with strep throat. HLA-B*57:01 was almost GWS (*P* = 5.06 x 10^{-8}). HLA-DRB1_CysTyr30 in class II region also showed a significant effect (P = 2.14 x 10^{-9}). Conditioning on rs1055821, the effect of HLA-B*57:01 was removed (conditional p-value is 0.84) and the effect of HLA-DRB1_CysTyr30 was substantially decreased (conditional p-value is 0.005).

We also found an GWS association with a small insertion, rs35395352, in 1p36.23 (P = 3.90 x 10^{-8}, effect = 0.03). None of the variants within 500kb and in moderate LD ($r^2 > 0.6$) with this SNP were coding. However, two variants have been reported as cis-eQTLs: rs7548511 (P = 7.84 x 10^{-8}, effect = 0.03, risk allele A, $r^2 = 0.96$ with rs35395352) is an eQTL for *ENO1* (eQTL p-value is 8 x 10^{-5}) in lymphoblastoid[38] and rs11121129 (P = 9.62 x 10^{-8}, effect = 0.03, risk allele G, $r^2 = 0.96$ with rs35395352) is an eQTL for *TNFRSF9* (eQTL p-value is 4.4 x 10^{-4}) in lymphoblastoid[39]. The product of ENO1, Alpha enolase, has been identified as an autoantigen in Hashimoto's encephalopathy [40]and a putative autoantigen in severe asthma[41] and Behcet's disease[42]. The *TNFRSF9*-encoded protein is a member of the TNF-receptor superfamily, which contributes to the clonal expansion, survival, and development of T cells. The same 1p36.23 region has been implicated in GWASs of psoriasis (rs11121129, risk allele A, P = 1.7 x 10^{-8}, $r^2 = 0.96$ with rs35395352)[43]. Interestingly, the strep throat risk allele rs11121129-A is protective against psoriasis.

Studies found that exacerbation of chronic psoriasis can be associated with streptococcal throat infections, and the T-cells generated by palatine tonsils can recognize skin keratin determinants in patients' blood[44].

Scarlet fever (group A streptococcus bacteria)

Scarlet fever (also known as scarlatina) is caused by group A streptococcus (GAS) and is characterized by a bright red rash that covers most of the body, sore throat, and fever. It was a major cause of death before the discovery of effective antibiotics. We found a GWS association between scarlet fever and the HLA class II region. The amino acid polymorphism HLA-DQB1-Gly45 ($P = 6.35 \times 10^{-14}$), located in the peptide-binding cleft of HLA-DQB1, accounted for the SNP association (rs36205178, $P = 9.49 \times 10^{-14}$, conditional p-values is 0.83) and the HLA allele association (HLA-DQB1*03:01, $P = 2.98 \times 10^{-14}$, conditional p-values is 0.14) in this region. Although scarlet fever develops in some people with strep throat, we did not observe any cross disease effect between 'scarlet fever' and 'strep throat' in our analysis.

Pneumonia (multiple origins)

Pneumonia is characterized by inflammation of the alveoli in the lungs and can be caused by a variety of organisms, including bacteria, viruses, and fungi. The most common cause of bacterial pneumonia in adults – isolated in nearly 50% of community-acquired cases -- is Streptococcus pneumonia [45]. We found a GWAS association between pneumonia and a HLA class I region indexed by rs3131623 ($P = 1.99 \times 10^{-15}$). Conditioning on rs3131623, the residual effects of HLA-B-SerTrpAsn97 ($P = 2.37 \times 10^{-12}$, conditional p-value is 9.33×10^{-4}) and HLA-DRB1-LS11 ($P = 4.10 \times 10^{-11}$, conditional p-value is 3.89×10^{-6}), both in the peptide-binding cleft, remained significant. There was also a residual effect left for

rs3131623 (conditional p-value is 6.94×10^{-5}) after conditioning on the two amino acid associations.

Bacterial meningitis (Streptococcus pneumoniae, Group B Streptococcus bacteria and many other bacteria)

Bacterial meningitis is a severe infection causing substantial neurological morbidity and mortality worldwide. Streptococcus pneumoniae is the leading cause of bacterial meningitis and is associated with a 30% mortality rate [46]. We identified a GWS association with *CA10* (carbonic anhydrase X) indexed by rs1392935 ($P = 1.19 \times 10^{-8}$, intronic of *CA10*). None of the variants within 500kb and in moderate LD ($r^2 > 0.6$) with rs1392935 were coding, nor were they reported as an eQTL. However, rs1392935 falls within an enhancer that is defined in the fetal brain. The protein encoded by *CA10* is an acatalytic member of the alpha-carbonic anhydrase subgroup and it is thought to play a role in the central nervous system, especially in brain development[47].

Yeast infection (Candida spp.)

A yeast infection, also called candidiasis, is usually a localized fungal infection of the skin or mucosal membranes, typically affecting the oral cavity, esophagus, gastrointestinal tract, urinary bladder, toenails, or genitalia[48]. Most yeast infections are caused by species of *Candida*, often *Candida albicans*. We found GWS associations between yeast infections and variants in *DSG1* (rs200520431, $P = 1.87 \times 10^{-16}$). The index SNP rs200520431 is intronic and in high LD ($r^2 > 0.8$) with multiple missense mutations (rs8091003, rs8091117, rs16961689, rs61730306, rs34302455) in *DSG1*. The *DSG1* gene product is a calcium-binding transmembrane glycoprotein component of desmosomes in vertebrate epithelial cells. It connects the cell surface to the keratin cytoskeleton and plays a key role in maintaining epidermal integrity and barrier function[49]. This glycoprotein has been

identified as the autoantigen of the autoimmune skin blistering disease pemphigus foliaceus[50] and homozygous mutations in *DSG1* have been showed to result in severe dermatitis, multiple allergies, and metabolic wasting syndrome[49]. A variant downstream of *PRKCH* (indexed by rs2251260, P = 3.46 x 10^{-10}) was also significantly associated with yeast infections. None of the variants within 500kb and in moderate LD ($r^2 > 0.6$) with the index SNP rs2251260 were coding, nor were they reported as an eQTL. However, rs2251260 falls within strong enhancers defined in hepatocellular carcinoma. *PRKCH* is a member of a family of serine- and threonine-specific protein kinase and is predominantly expressed in epithelial tissues. The *PRKCH* protein kinase can regulate keratinocyte differentiation [51]. We also found a significant association with a variant in the 14q32.2 gene desert (rs7161578-T, p = 4.04 x 10^{-8}). The index SNP is in high LD with many enhancer sequences defined in epidermal keratinocytes. A variant in the same region (rs7152623-A, p = 3 x 10^{-15}, r^2 = (+) 0.98 with rs7161578-T) was implicated in aortic stiffness[52].

Urinary Tract Infection (Escherichia coli, Staphylococcus saprophyticus, rarely viral or fungal)

About 80%-85% of urinary tract infections (UTIs) are caused by Escherichia coli (E. coli), and 5-10% are caused by Staphylococcus saprophyticus [53]. Very rarely, viruses or fungi may cause UTIs [54]. UTIs occur more commonly in women than men [55] and in our dataset, significantly more female than male customers reported having two or more UTIs. We found an association between UTIs and variants in *PSCA* (indexed by rs2976388, *P* = 3.27 x 10^{-10}). The index SNP rs2976388 falls within strong enhancers defined in epidermal keratinocytes. It is also in high LD (r2 = 0.95) with rs2294008 (P = 1.01 x 10^{-9}), which is in the 5'-UTR promoter sequence of *PSCA* and can cause changes in transcriptional repressor CTCF motif. The PSCA gene variants conferred risk of UTI only in female when we further

tested the effect in female and male cohort separately (P < 1e-11 in female, and P >0.1 in male). Prior GWASs have found associations between the *PSCA* gene and duodenal ulcer (rs2294008-C, p = 2e-33) and bladder cancer (rs2294008-T, p = 4 x 10^{-11}) [56,57]. *PSCA* encodes a glycosylphosphatidylinositol-anchored cell membrane glycoprotein of unknown function. This glycoprotein was initially identified as a prostate-specific cell-surface marker [58] and is overexpressed in a large proportion of prostate cancers, but also detected in bladder and pancreatic cancers [57].

We also identified a variant intronic of *FRMD5* (rs146906133, P = 2.02 x 10^{-8}) that is in LD with rs138763871 (r^2 = 0.8, p = 1.75 x 10^{-6} in our GWAS), a missense variant in *STRC* (R1521W). The role of FRMD5 or STRC in UTIs is unclear.

Tonsillectomy (multiple origins)

Tonsils are considered the first line of defense against respiratory infections. Previous twin studies have suggested a substantial genetic predisposition for recurrent tonsillitis[59]. Tonsillectomy is a surgery to remove the tonsils typically due to repeated occurrence of tonsillitis or tonsil hypertrophy. Our tonsillectomy GWAS revealed 35 independent chromosome regions reaching GWS. We observed strong HLA associations at HLA-B in the class I region. The amino acid polymorphism HLA-B-Val97 (P = 1.44 x 10^{-21}), in the peptide-binding cleft of the HLA-B protein, accounted for most of the SNP effect (rs41543314, P = 5.36 x 10^{-21}, conditional p value is 0.06) and HLA allele effect (HLA-B*57:01, P = 6.29 x 10^{-22}, conditional p value is 0.05) in this interval. We also noted additional independent signals in the class II and class III regions. Within HLA-DRB1 in the class II region, the effect of rs140177540 (P = 5.02 x 10^{-15}, conditional p-value is 0.001) was largely removed after conditioning on DRB1_LeuValGly11 (P = 1.02 x 10^{-12}) in the peptide-binding cleft of the HLA-DRB1 protein.

We detected 34 additional GWS associations (Table 2 and Supplementary Table 4), including signals from genes encoding tumor necrosis factors/receptors superfamily members (*LTBR, TNFRSF13B, TNFSF13B, CD40, TNFAIP3*), myotubularin (MTMR3), chemokines (*CXCL13*), mitogen-activated protein kinases/regulators (*IGFBP3, SPRED2, DUSP10, ST5, MAPK3*), SRC homology domain binding proteins (*SBK1, ST5, SH2B3*), disintegrins (*ADAM23, ADAMTS10*), various transcription factors (*NFKB1, HOXA2, FOXA1, IKZF1, MDFIC, TFEC, RERE, TBX1, FOXN1*), and signaling molecules (*WNT2, GNA12, ITSN1*). Many of these genes map to pathways involved in immune and inflammatory processes, while others are important regulators of embryonic development (e.g. most of the transcription factors), and a few are potentially involved in platelet production or hemostasis (*GNA12, SH2B3, RAP1GAP2, ADAM23, ADAMTS10*). When we tested the significant loci for enrichment in canonical pathways, the top over-represented pathway was 'Intestinal immune network for IgA production' (Table5, $P = 2 \times 10^{-6}$ and FDR = 3.5e-3). The genes that contributed significantly in this pathway include *LTBR* (rs10849448, $P = 2.35 \times 10^{-35}$), *TNFRSF13B* (rs34557412, missense variant, $P = 2.66 \times 10^{-21}$), *TNFSF13B* (rs200748895, $P = 1.20 \times 10^{-17}$) and *CD40* (rs6032664, $P = 7.31 \times 10^{-12}$). IgA is an antibody that plays a critical role in mucosal immunity. Tonsils belong to nasopharyngeal-associated lymphoid tissues and the generation of B cells is considered one of the major tonsillar functions; secretory dimeric IgA produced by the B cells is capable of preventing absorption and penetration of bacteria and/or viruses into the upper respiratory tract mucosa[60]. Tonsillectomy has been shown to significantly decrease levels of serum IgA and salivary secretory IgA levels [61]. Other identified significant signaling pathways include 'TACI and BCMA stimulation of B cell immune system' ($P = 9 \times 10^{-4}$ and FDR = 0.078), 'ceramide signaling pathway' ($P = 3 \times 10^{-4}$ and FDR = 0.08) and 'Glucocorticoid receptor

regulatory network' ($P = 5 \times 10^{-4}$ and FDR = 0.099). The contributing genes are *TNFSF13B* (rs200748895, $P = 1.2 \times 10^{-17}$), *TNFRSF13B* (rs34557412, $P = 2.66 \times 10^{-21}$), *NFKB1* (rs23052, $P = 4.54 \times 10^{-14}$) and *MAPK3* (rs12931792, $P = 4.06 \times 10^{-8}$).

Although tonsillectomy has been performed for over 100 years, possible immunological effects of this procedure remain controversial. Many reports have demonstrated that in certain patients tonsillectomy is an effective therapy for psoriasis[44] and rheumatoid arthritis (RA)[62]; the rationale for this effect is unknown. The 'Intestinal immune network for IgA production' pathway was also identified as the most significant pathway in a recent GWAS of IgAN[63]. In that study, they also linked the intestinal mucosal inflammatory disorders and inflammatory bowel disease (IBD) with risk of IgAN. Our tonsillectomy GWAS showed significant overlap of the identified loci with those autoimmune and inflammatory disorders[64]. We found five risk loci that are also associated with RA (*HLA, TNFIP3, CD40, SPRED2 and SH2B3*); four risk loci that are also associated with IBD (*HLA, MTMR3, TNFAIP3 and CD40*) and ulcerative colitis (*HLA, GNA12, MAPK3 and NFKB1*); three risk loci that are also associated with psoriasis (*HLA, SLC12AB and 1p36.23*); and two risk loci (*HLA* and *MTMR3*) that are shared with IgAN. We observed both concordant and opposing effects compared to these immune-mediated diseases. Our results may help to elucidate the connection between tonsillectomy and these diseases and may provide insight into clinical markers that could be used as indicators of tonsillectomy as a therapy for these diseases.

Childhood ear infection (multiple origins)

Although ear infections can have a bacterial origin, they are often secondary to an upper respiratory infection such as a cold or sore throat, which can be caused by viruses or bacteria[65]. Children are more likely than adults to have ear infections because their

Eustachian tubes are less effective at draining fluid and their immune systems are not fully developed. In our GWAS, we identified 14 regions that were significantly associated with childhood ear infection. Signals in the HLA region mapped to HLA-DRB1 in class II region. Our strongest association within HLA-DRB1 was with HLA-DRB1-Gln96 in the peptide-binding groove. After conditioning on HLA-DRB1-Gln96 ($P = 2.11 \times 10^{-12}$), the effects of rs4329147 ($P = 9.55 \times 10^{-12}$, conditional p value is 0.27) and DQB1*06:02 ($P = 5.79 \times 10^{-10}$, conditional p value is 0.85) were eliminated.

We found 13 additional GWS signals (Table 2 and Supplementary Table 5). Although MAGENTA pathway analysis did not identify significant pathways (Table 6), variants in *FUT2* and *ABO*, which were involved in the glycosphingolipid biosynthesis pathway and were implicated in our mumps GWAS, were also significantly associated with childhood ear infection. The risk allele rs681343(C) ($P = 3.51 \times 10^{-30}$) is a synonymous mutation in *FUT2* and is in almost complete LD with rs601338(G) ($r^2 = 0.9993$), the 'secretor' (se) allele is associated with higher risk of childhood ear infection in our data. This is consistent with a previous report in which secretors were over-represented among patients with upper respiratory infections [66]. Our second most significant association was with a variant in *TBX1* (rs1978060, $P = 1.17 \times 10^{-19}$). The low-risk allele rs1978060 (A) is in LD ($r^2 = 0.45$) with rs72646967 (C), a missense (N397H) mutation in *TBX1* that was also implicated in our tonsillectomy GWAS. T-box genes encode transcription factors that play important roles in tissue and organ formation during embryonic development. In mice, *TBX1* haplo-insufficiency in the DiGeorge syndrome (22q11.2 deletion syndrome) region has been showed to disrupt the development of the pharyngeal arch arteries[67] and the middle and outer ear [68,69]. We also identified other genes involved in developmental processes, including *MKX* (rs2808290, $P = 5.09 \times 10^{-16}$), *FGF3* (rs72931768, $P = 2.63 \times 10^{-9}$) and *AUTS2*

(rs35213789, P = 3.75 x 10^{-9}). The *MKX* product is an IRX family-related homeobox transcription factor and has been shown to play a critical role in tendon differentiation by regulating type I collagen production in tendon cells[70]. *FGF3* is a member of the fibroblast growth factor family, which is involved in a variety of biological processes including embryonic development, cell growth, morphogenesis, tissue repair and invasion. Study of similar genes in mouse and chicken suggested the role in inner ear formation[71].. *AUTS2* has been implicated in neurodevelopment and is a candidate gene for autism spectrum disorders and developmental delay[72,73]. Finally, we identified missense mutations in *CDHR3* (rs114947103, P = 5.40 x 10^{-9}) and *PLG* (rs73015965, P = 3.78 x 10^{-8}). The index SNP *rs114947103 is in almost complete LD with rs6967330 (r2=0.97), a missense variant in CDHR3, that* was recently identified in a GWAS as a susceptibility locus for asthma[74]. The biological role of *CDHR3* is not clear, but it is highly expressed in airway epithelium and belongs to the cadherin family that is involved in cell adhesion and epithelial polarity[74]. Mutations in the *PLG* gene could cause congenital plasminogen deficiency, which results in inflamed growths on the mucous membranes. Studies in mice have showed that *PLG* plays an essential role in protecting against the spontaneous development of chronic otitis media (middle ear infection) and have also suggested the possibility of using *PLG* for clinical therapy of certain types of otitis media[75].

Myringotomy (multiple causes)

Myringotomy is a surgical procedure in which a tiny incision is created in the eardrum to allow the release of middle-ear fluid. Before the invention of antibiotics, myringotomy without tube placement was a common treatment for severe or frequent acute otitis media[76]. The same variant in *TBX1* that was associated with tonsillectomy and childhood

ear infection (see earlier) was also found to be significantly associated with myringotomy (rs1978060, $P = 3.34 \times 10^{-10}$).

Discussion

We identified the greatest number of independent associations for tonsillectomy (n=35) and childhood ear infection (n=14), two relatively nonspecific phenotypes. For many of the infections, we only identified HLA associations. These findings may be due to our relatively larger sample sizes for the two traits and may also be due to the fact that tonsillectomy and childhood ear infections are heterogeneous phenotypes, influenced by a variety of pathogens and also by anatomical abnormalities. People undergo tonsillectomy for different reasons, including recurrent tonsillitis, obstructive sleep apnea, and nasal airway obstruction[77]. The causes of childhood ear infections are also diverse, involving different types of bacteria and viruses, and ear developmental defects[78]. Variants in the *TBX1* gene, which is essential for inner ear development[68,69], were associated with both tonsillectomy and inner ear infections. Many genes that are involved in embryotic development were also identified in our tonsillectomy and childhood ear infection GWASs. The vulnerability to tonsillectomy and childhood ear infections may also be largely inherited. There were earlier studies calculated that heredity account for over 70% of susceptibility to recurrent ear infections in children[79,80]. Previous twin studies had suggested a substantial genetic predisposition for recurrent tonsillitis[59] and sleep-disordered breathing[81]. According to the LD score regression on our GWAS data, tonsillectomy and childhood ear infections have relatively higher heritability on the observed scale than most of the other infections we studies. We did not calculate the liability scale heritability due to the lack of population prevalence information.

In the HLA region, we found that viral diseases — e.g., chicken pox, shingles, cold sores, mononucleosis (all caused by herpesvirus) as well as mumps (caused by mumps virus) — were mainly associated with variation in class I molecules (Figure 1). The bacterial

diseases — specifically having a positive TB test (caused by Mycobacterium tuberculosis), scarlet fever (caused by GAS), and childhood ear infection (mostly caused by bacteria) — were mainly associated with variation in class II molecules. Tonsillectomy and pneumonia, caused by either bacteria or viruses, were associated with both class I and class II molecules. These observations are consistent with previous knowledge about antigen presentation; viruses mostly replicate within nucleated cells in the cytosol and produce endogenous antigens that are presented by class I MHC molecules, while bacteria grow extracellularly and are taken up by endosomal compartments where they are processed for presentation by class II MHC molecules [82]. These two intracellular pathways of protein processing may not be completely separate. Activation of CD8+, HLA class I-restricted T-cells by exogenous antigens has been reported [83] and HLA class II-restricted CTLs that recognize endogenously synthesized antigen, such as HBV envelope antigens, have also been described [84,85].

Some diseases, however, do not follow the common principles of antigen presentation. Plantar warts (caused by HPV) are different from other virus-causing diseases. HPV infections are exclusively intraepithelial and there is no viraemia or cytolysis in the infection cycle [86]. The primary response to HPV antigens is more likely to be initiated by the antigen-presenting cells (APCs) of squamous epithelia, the Langerhans cells (LCs). LCs capture antigens by macropinocytosis and receptor-mediated endocytosis, and then initiate class II processing of exogenous antigens [87], which explains why our plantar warts GWAS mainly identified associations with MHC class II molecules. In other cases, our results suggest even more complex interactions. Strep throat (caused by GAS bacteria) has both MHC class I and class II associations. Some of our reported strep throat cases may represent secondary bacterial infections in response to a viral upper respiratory infection,

meaning that the discovered MHC class I associations are actually with the virus infection that initiated the process [88]. Since most sore throats are caused by viruses [89], another possibility is that some of the reported strep throat cases actually had a misdiagnosed or misreported viral infection. A final possibility is that HLA class I molecules can bind bacterial peptides derived from exogenous proteins that are internalized by endocytosis or phagocytosis, a process called cross-presentation [85,90].

We failed to detect any GWS association for certain infectious diseases, including rubella, measles, chronic sinus infection, the common cold, or rheumatic fever. It is known that HLA affects responses to measles and rubella vaccinations [91,92] and may thus also be associated with susceptibility. One limitation of our study that might affect findings with respect to measles and rubella is our lack of data on vaccination status. Since there is now a common vaccination for measles and rubella, misclassifying vaccinated people as controls would theoretically reduce our power to detect associations. This limitation might also affect the power of our shingles GWAS. The controls used for the shingles GWAS were not filtered for a positive history of chickenpox, though having chickenpox is a prerequisite for shingles. Another reason might be the heterogeneity of the phenotype. Colds are caused by several different viruses (mostly rhinoviruses and coronaviruses), and larger sample sizes may be required to identify alleles only associated with specific pathogens.

Although additional studies will be required to validate our associations, our findings are an important step towards dissecting the host genetic contribution to variation in response to infections. Research insights into infectious diseases will help to derive new diagnostic approaches and perhaps new therapies and preventions. This study may also help us to understand some of the autoimmune disorders that are associated with various infection triggers. Many of the identified associations were also found in autoimmune

diseases. One postulated connection between infectious diseases and immunological disorders is that the immune cells that are activated in response to a pathogen epitope are also cross-reactive to self and lead to direct damage and further activation of other arms of the immune system[93].

Methods

Subjects

We conducted GWASs of 23 infectious diseases. All participants were drawn from the customer base of 23andMe, Inc., a personal genetics company. All individuals included in the analyses provided informed consent and answered surveys online according to our human subjects protocol, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). All participants were of primarily European ancestry (>97% European ancestry) as determined through an analysis of local ancestry[94]. A maximum set of unrelated individuals was chosen using a segmental identity-by-descent (IBD) estimation algorithm [95]. Individuals were defined as related if they shared more than 700 cM of IBD, including regions where the two individuals share either one or both genomic segments IBD. This level of relatedness (involving ~20% of the human genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

Table 1 shows our discovery sample sizes, drawn from more than 650,000 genotyped customers who reported via web-based questionnaires whether they had been diagnosed with one or more infectious diseases. (For more detail on survey questions used to collect the phenotype data, see the Supplementary Notes).

Genotyping and SNP imputation

DNA extraction and genotyping were performed on saliva samples by National Genetics Institute (NGI), a CLIA-certified laboratory that is a subsidiary of Laboratory Corporation of America. Samples were genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550+ BeadChip and contained a total of

about 560,000 SNPs, including about 25,000 custom SNPs selected by 23andMe. The V3 platform was based on the Illumina OmniExpress+ BeadChip and contained a total of about 950,000 SNPs and custom content to improve the overlap with our V2 array. The V4 platform in current use is a fully custom array of about 950,000 SNPs and includes a lower redundancy subset of V2 and V3 SNPs with additional coverage of lower-frequency coding variation. Samples that failed to reach 98.5% call rate were re-analyzed. Individuals whose analyses failed repeatedly were re-contacted by 23andMe customer service to provide additional samples, as is done for all 23andMe customers.

Participant genotype data were imputed using the March 2012 "v3" release of the 1000 Genomes Project reference haplotypes [96]. We phased and imputed data for each genotyping platform separately. First, we used BEAGLE [97] (version 3.3.1) to phase batches of 8,000 to 9,000 individuals across chromosomal segments of no more than 10,000 genotyped SNPs, with overlaps of 200 SNPs. We excluded SNPs with minor allele frequency (MAF) < 0.001, Hardy-Weinberg equilibrium $P < 1 \times 10^{-20}$, call rate < 95%, or large allele frequency discrepancies compared to the European 1000 Genomes Project reference data. Frequency discrepancies were identified by computing a 2x2 table of allele counts for European 1000 Genomes samples and 2000 randomly sampled 23andMe customers with European ancestry, and identifying SNPs with a chi squared $P < 10^{-15}$. We imputed each phased segment against all-ethnicity 1000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac2 [98] with five rounds and 200 states for parameter estimation.

For the non-pseudoautosomal region of the X chromosome, males and females were phased together in segments, treating the males as already phased; the pseudoautosomal regions were phased separately. We then imputed males and females together using

Minimac as with the autosomes, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

GWAS analysis

For case control comparisons, we tested for association using logistic regression, assuming additive allelic effects. For quantitative traits, association tests were performed using linear regression. For tests using imputed data, we use the imputed dosages rather than best-guess genotypes. We included covariates for age, gender, and the top five principal components to account for residual population structure. The association test p-value was computed using a likelihood ratio test, which in our experience is better behaved than a Wald test on the regression coefficient. Results for the X chromosome were computed similarly, with men coded as if they were homozygous diploid for the observed allele.

Genetic correlations

We computed LD scores as previously described[99] using the samples in the 1000 Genomes Project reference panel (phase 3 verion 5a) with a MAF cutoff of 5%. We calculated the pairwise genetic correlation (r_g) on the GWAS summary statistics (including MHC region) using LD score regression[99]. The cluster map of genetic correlation, base on the lower bound of the 95% confident interval of calculated r_g, was plotted using 'Nearest Point Algorithm' implemented in python "Searborn" package (https://stanford.edu/~mwaskom/software/seaborn/index.html). We found no significant negative correlations in our data, thus we set all negative values to zero.

SNP Function Annotation

To explore whether any of the significant SNPs identified might link to functional mutations or have potential regulatory functions, we used the online tools HaploReg

(http://www.broadinstitute.org/mammals/haploreg/haploreg.php) to confirm the location of each SNP in relation to annotated protein-coding genes and/or non-coding regulatory elements. We queried only the variants that were within 500kb of and in moderate LD (r^2>0.6) with the index SNP.

Imputation of HLA classical alleles and respective amino acid variations

HLA imputation was performed with HIBAG [100], an attribute bagging based statistical method that comes as a freely available R package and includes a pre-fit classifier. This classifier was trained from a large database of individuals with known HLA alleles and SNP variation within the HLA region. Over 98% of the tagging SNPs used in HIBAG were genotyped and passed quality control on 23andMe's platform. We imputed allelic dosage for HLA-A, B, C, DPB1, DQA1, DQB1, and DRB1 loci at four-digit resolution. We used the default settings of HIBAG and the run time for 100,000 samples was about 10 hours on our cluster.

Using an approach suggested by P. de Bakker [101], we downloaded the files that map HLA alleles to amino acid sequences from https://www.broadinstitute.org/mpg/snp2hla/ and mapped our imputed HLA alleles at four-digit resolution to the corresponding amino acid sequences; in this way we translated the imputed HLA allelic dosages directly to amino acid dosages. We encoded all amino acid variants in the 23andMe European samples as biallelic markers, which facilitated downstream analysis. For example, position 45 of HLA-B protein had five different alleles (E: Glu, M: Met, T: Thr, K: Lys, G: Gly), we first encoded the position using five binary markers, each corresponding to the presence or absence of each allele (e.g. HLA-B-Gly45 tags Gly at position 45 of HLA-B protein). For positions having three or more alleles, we also created markers that tag multiple alleles, each corresponding to the presence or absence of the multiple alleles (e.g. HLA-B-ThrGly45 tags Thr and Gly at

position 45 of HLA-B protein). Thus, we created binary indicators for all possible combinations of amino acid variants. We use this naming convention for amino acid polymorphisms throughout this paper.

We imputed 292 classical HLA alleles at four-digit resolution and 2395 bi-allelic amino acid polymorphisms. Similar to the SNP imputation, we measured imputation quality using r^2, which is the ratio of the empirically observed variance of the allele dosage to the expected variance assuming Hardy-Weinberg equilibrium. The imputation quality (r^2) of the top associated HLA alleles and amino acids are in Supplementary Table 6.

HLA region fine mapping

To test associations between imputed HLA allele/amino acid dosages and phenotypes, we performed logistic or linear regression using the same set of covariates used in the SNP-based GWAS for that phenotype. We applied a forward stepwise strategy, within each type of variant, to establish statistically independent signals in the HLA region. Within each variant type(e.g. SNP, HLA allele, and HLA amino acid), we first identified the most strongly associated signals (lowest p-value) for each disease and performed forward iterative conditional regression to identify other independent signals if the conditional p-value was $< 5 \times 10^{-8}$. All analyses controlled for sex and five principal components of genetic ancestry. The p-values were calculated using a likelihood ratio test. The iterative conditional regression dissected HLA signal into independent HLA associations. Within each identified HLA locus (HLA-A, B, C, DPB1, DQA1, DQB1, and DRB1), We further carried out reciprocal analyses, which are the conditional analyses across variants types, to see if the association can be attributed to the amino acid polymorphism within each HLA locus.

Pathway analysis

In order to better understand how multiple genes in the same pathway may contribute to certain infections, we performed pathway analysis using Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA)[102], which tests for enrichment of genetic associations in predefined biological processes or sets of functionally related genes, using GWAS results as input. We used gene sets of 1320 canonical pathways from the Molecular Signatures Database (MsigDB) compiled by domain experts[103] and default settings of MAGENTA tool.

Acknowledgements:

We thank the customers of 23andMe for participating in this research and the employees of 23andMe for their contributions to this work.

Conflict of interest:

All authors listed were/are employed by 23andMe, and own stock or stock options in 23andMe.

REFERENCE:

1. Becker, K., Hu, Y. & Biller-Andorno, N. Infectious diseases - a global challenge. *Int. J. Med. Microbiol. IJMM* **296,** 179–185 (2006).
2. Mighty, K. K. & Laimins, L. A. The role of human papillomaviruses in oncogenesis. *Recent Results Cancer Res. Fortschritte Krebsforsch. Prog. Dans Rech. Sur Cancer* **193,** 135–148 (2014).
3. Filippi, C. M. & von Herrath, M. G. Viral trigger for type 1 diabetes: pros and cons. *Diabetes* **57,** 2863–2871 (2008).
4. Kang, J.-H., Sheu, J.-J., Kao, S. & Lin, H.-C. Increased risk of multiple sclerosis following herpes zoster: a nationwide, population-based study. *J. Infect. Dis.* **204,** 188–192 (2011).
5. Pavlos, R., Mallal, S., Ostrov, D., Pompeu, Y. & Phillips, E. Fever, rash, and systemic symptoms: understanding the role of virus and HLA in severe cutaneous drug allergy. *J. Allergy Clin. Immunol. Pract.* **2,** 21–33 (2014).
6. Blackwell, J. M., Jamieson, S. E. & Burgner, D. HLA and infectious diseases. *Clin. Microbiol. Rev.* **22,** 370–385, Table of Contents (2009).
7. Cooke, G. S. & Hill, A. V. S. Genetics of susceptibitlity to human infectious disease. *Nat. Rev. Genet.* **2,** 967–977 (2001).
8. Kvestad, E. *et al.* Recurrent otitis media and tonsillitis: common disease predisposition. *Int. J. Pediatr. Otorhinolaryngol.* **70,** 1561–1568 (2006).
9. Weinberg, J. M. Herpes zoster: epidemiology, natural history, and common complications. *J. Am. Acad. Dermatol.* **57,** S130–135 (2007).

10. Mo, X. *et al.* Microarray analyses of differentially expressed human genes and biological processes in ECV304 cells infected with rubella virus. *J. Med. Virol.* **79,** 1783–1791 (2007).
11. Chan, A. *et al.* Variation in the type I interferon gene cluster on 9p21 influences susceptibility to asthma and atopy. *Genes Immun.* **7,** 169–178 (2006).
12. Xu, F. *et al.* Trends in herpes simplex virus type 1 and type 2 seroprevalence in the United States. *JAMA J. Am. Med. Assoc.* **296,** 964–973 (2006).
13. Kriesel, J. D. *et al.* C21orf91 genotypes correlate with herpes simplex labialis (cold sore) frequency: description of a cold sore susceptibility gene. *J. Infect. Dis.* **204,** 1654–1662 (2011).
14. Rostgaard, K., Wohlfahrt, J. & Hjalgrim, H. A genetic basis for infectious mononucleosis: evidence from a family study of hospitalized cases in Denmark. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **58,** 1684–1689 (2014).
15. *Harrison's principles of internal medicine / editors, Anthony S. Fauci ... [et al.]*. (McGraw-Hill Medical, 2008).
16. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G. & Lowe, J. B. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. *J. Biol. Chem.* **270,** 4640–4649 (1995).
17. Larsson, M. M. *et al.* Antibody prevalence and titer to norovirus (genogroup II) correlate with secretor (FUT2) but not with ABO phenotype or Lewis (FUT3) genotype. *J. Infect. Dis.* **194,** 1422–1427 (2006).
18. McGovern, D. P. B. *et al.* Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum. Mol. Genet.* **19,** 3468–3476 (2010).

19. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–1125 (2010).

20. Smyth, D. J. *et al.* FUT2 nonsecretor status links type 1 diabetes susceptibility and resistance to infection. *Diabetes* **60,** 3081–3084 (2011).

21. Zaitsev, V. *et al.* Second sialic acid binding site in Newcastle disease virus hemagglutinin-neuraminidase: implications for fusion. *J. Virol.* **78,** 3733–3741 (2004).

22. Cohen, M., Hurtado-Ziola, N. & Varki, A. ABO blood group glycans modulate sialic acid recognition on erythrocytes. *Blood* **114,** 3668–3676 (2009).

23. Ma, J. X. *et al.* Structure and chromosomal localization of the gene (BDKRB2) encoding human bradykinin B2 receptor. *Genomics* **23,** 362–369 (1994).

24. Pan, L. *et al.* A genome-wide association study identifies polymorphisms in the HLA-DR region associated with non-response to hepatitis B vaccination in Chinese Han populations. *Hum. Mol. Genet.* **23,** 2210–2219 (2014).

25. Hu, Z. *et al.* New loci associated with chronic hepatitis B virus infection in Han Chinese. *Nat. Genet.* **45,** 1499–1503 (2013).

26. Kamatani, Y. *et al.* A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat. Genet.* **41,** 591–595 (2009).

27. Umemura, T. *et al.* Human leukocyte antigen class II haplotypes affect clinical characteristics and progression of type 1 autoimmune hepatitis in Japan. *PloS One* **9,** e100565 (2014).

28. Sherlock, S., Fox, R. A., Niazi, S. P. & Scheuer, P. J. Chronic liver disease and primary liver-cell cancer with hepatitis-associated (Australia) antigen in serum. *Lancet* **1,** 1243–1247 (1970).

29. Vogt, T. M., Wise, M. E., Bell, B. P. & Finelli, L. Declining hepatitis A mortality in the United States during the era of hepatitis A vaccination. *J. Infect. Dis.* **197,** 1282–1288 (2008).

30. Rauch, A. *et al.* Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology* **138,** 1338–1345, 1345.e1–7 (2010).

31. Mischke, D., Korge, B. P., Marenholz, I., Volz, A. & Ziegler, A. Genes encoding structural proteins of epidermal cornification and S100 calcium-binding proteins form a gene complex ('epidermal differentiation complex') on human chromosome 1q21. *J. Invest. Dermatol.* **106,** 989–992 (1996).

32. Sun, L.-D. *et al.* Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population. *Nat. Genet.* **43,** 690–694 (2011).

33. Zhang, X.-J. *et al.* Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. *Nat. Genet.* **41,** 205–210 (2009).

34. Sugita, K. *et al.* Innate immunity mediated by epidermal keratinocytes promotes acquired immunity involving Langerhans cells and T cells in the skin. *Clin. Exp. Immunol.* **147,** 176–183 (2007).

35. Bergboer, J. G. M. *et al.* Psoriasis risk genes of the late cornified envelope-3 group are distinctly expressed compared with genes of other LCE groups. *Am. J. Pathol.* **178,** 1470–1477 (2011).

36. Learn the Signs and Symptoms of TB Disease.

37. Tart, A. H., Walker, M. J. & Musser, J. M. New understanding of the group A Streptococcus pathogenesis cycle. *Trends Microbiol.* **15,** 318–325 (2007).

38. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).

39. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–1250 (2009).

40. Yoneda, M. *et al.* High prevalence of serum autoantibodies against the amino terminal of alpha-enolase in Hashimoto's encephalopathy. *J. Neuroimmunol.* **185,** 195–200 (2007).

41. Nahm, D.-H. *et al.* Identification of alpha-enolase as an autoantigen associated with severe asthma. *J. Allergy Clin. Immunol.* **118,** 376–381 (2006).

42. Lee, K. H. *et al.* Human alpha-enolase from endothelial cells as a target antigen of anti-endothelial cell antibody in Behçet's disease. *Arthritis Rheum.* **48,** 2025–2035 (2003).

43. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44,** 1341–1348 (2012).

44. Thorleifsdottir, R. H. *et al.* Improvement of psoriasis after tonsillectomy is associated with a decrease in the frequency of circulating T cells that recognize streptococcal determinants and homologous skin determinants. *J. Immunol. Baltim. Md 1950* **188,** 5160–5165 (2012).

45. Pneumococcal Disease.

46. Castelblanco, R. L., Lee, M. & Hasbun, R. Epidemiology of bacterial meningitis in the USA from 1997 to 2010: a population-based observational study. *Lancet Infect. Dis.* **14,** 813–819 (2014).

47. Taniuchi, K. *et al.* Developmental expression of carbonic anhydrase-related proteins VIII, X, and XI in the human brain. *Neuroscience* **112,** 93–99 (2002).

48. *Medical microbiology*. (University of Texas Medical Branch at Galveston, 1996).

49. Samuelov, L. *et al.* Desmoglein 1 deficiency results in severe dermatitis, multiple allergies and metabolic wasting. *Nat. Genet.* **45,** 1244–1248 (2013).

50. Amagai, M. & Stanley, J. R. Desmoglein as a target in skin disease and beyond. *J. Invest. Dermatol.* **132,** 776–784 (2012).

51. Kashiwagi, M., Ohba, M., Chida, K. & Kuroki, T. Protein kinase C eta (PKC eta): its involvement in keratinocyte differentiation. *J. Biochem. (Tokyo)* **132,** 853–857 (2002).

52. Mitchell, G. F. *et al.* Common genetic variation in the 3'-BCL11B gene desert is associated with carotid-femoral pulse wave velocity and excess cardiovascular disease risk: the AortaGen Consortium. *Circ. Cardiovasc. Genet.* **5,** 81–90 (2012).

53. Nicolle, L. E. Uncomplicated urinary tract infection in adults including uncomplicated pyelonephritis. *Urol. Clin. North Am.* **35,** 1–12, v (2008).

54. Amdekar, S., Singh, V. & Singh, D. D. Probiotic therapy: immunomodulating approach toward urinary tract infection. *Curr. Microbiol.* **63,** 484–490 (2011).

55. Valiquette, L. Urinary tract infections in women. *Can. J. Urol.* **8 Suppl 1,** 6–12 (2001).

56. Tanikawa, C. *et al.* A genome-wide association study identifies two susceptibility loci for duodenal ulcer in the Japanese population. *Nat. Genet.* **44,** 430–434, S1–2 (2012).

57. Wu, X. *et al.* Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.* **41,** 991–995 (2009).

58. Reiter, R. E. *et al.* Prostate stem cell antigen: a cell surface marker overexpressed in prostate cancer. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 1735–1740 (1998).

59. Kvestad, E. *et al.* Heritability of recurrent tonsillitis. *Arch. Otolaryngol. Head Neck Surg.* **131,** 383–387 (2005).

60. Bernstein, J. M. Mucosal immunology of the upper respiratory tract. *Respir. Int. Rev. Thorac. Dis.* **59 Suppl 3,** 3–13 (1992).

61. D'Amelio, R., Palmisano, L., Le Moli, S., Seminara, R. & Aiuti, F. Serum and salivary IgA levels in normal subjects: comparison between tonsillectomized and non-tonsillectomized subjects. *Int. Arch. Allergy Appl. Immunol.* **68,** 256–259 (1982).

62. Kawano, M. *et al.* Simultaneous, clonally identical T cell expansion in tonsil and synovium in a patient with rheumatoid arthritis and chronic tonsillitis. *Arthritis Rheum.* **48,** 2483–2488 (2003).

63. Gharavi, A. G. *et al.* Genome-wide association study identifies susceptibility loci for IgA nephropathy. *Nat. Genet.* **43,** 321–327 (2011).

64. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–1006 (2014).

65. Heikkinen, T. & Chonmaitree, T. Importance of respiratory viruses in acute otitis media. *Clin. Microbiol. Rev.* **16,** 230–241 (2003).

66. Raza, M. W. *et al.* Association between secretor status and respiratory viral illness. *BMJ* **303,** 815–818 (1991).

67. Lindsay, E. A. *et al.* Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. *Nature* **410,** 97–101 (2001).

68. Raft, S., Nowotschin, S., Liao, J. & Morrow, B. E. Suppression of neural fate and control of inner ear morphogenesis by Tbx1. *Dev. Camb. Engl.* **131,** 1801–1812 (2004).

69. Vitelli, F. *et al.* TBX1 is required for inner ear morphogenesis. *Hum. Mol. Genet.* **12,** 2041–2048 (2003).

70. Ito, Y. *et al.* The Mohawk homeobox gene is a critical regulator of tendon differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10538–10542 (2010).

71. Sensi, A. *et al.* LAMM syndrome with middle ear dysplasia associated with compound heterozygosity for FGF3 mutations. *Am. J. Med. Genet. A.* **155A,** 1096–1101 (2011).

72. Oksenberg, N. & Ahituv, N. The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet. TIG* **29,** 600–608 (2013).

73. Jolley, A. *et al.* De novo intragenic deletion of the autism susceptibility candidate 2 (AUTS2) gene in a patient with developmental delay: a case report and literature review. *Am. J. Med. Genet. A.* **161A,** 1508–1512 (2013).

74. Bønnelykke, K. *et al.* A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* **46,** 51–55 (2014).

75. Eriksson, P.-O., Li, J., Ny, T. & Hellström, S. Spontaneous development of otitis media in plasminogen-deficient mice. *Int. J. Med. Microbiol. IJMM* **296,** 501–509 (2006).

76. Smith, N. & Greinwald, J. To tube or not to tube: indications for myringotomy with tube placement. *Curr. Opin. Otolaryngol. Head Neck Surg.* **19,** 363–366 (2011).

77. Baugh, R. F. *et al.* Clinical practice guideline: tonsillectomy in children. *Otolaryngol.--Head Neck Surg. Off. J. Am. Acad. Otolaryngol.-Head Neck Surg.* **144,** S1–30 (2011).

78. Ear Infections in Children.

79. Casselbrant, M. L. *et al.* The heritability of otitis media: a twin and triplet study. *JAMA* **282,** 2125–2130 (1999).

80. Kvaerner, K. J., Tambs, K., Harris, J. R. & Magnus, P. Distribution and heritability of recurrent ear infections. *Ann. Otol. Rhinol. Laryngol.* **106,** 624–632 (1997).

81. Carmelli, D., Colrain, I. M., Swan, G. E. & Bliwise, D. L. Genetic and environmental influences in sleep-disordered breathing in older male twins. *Sleep* **27,** 917–922 (2004).

82. Blum, J. S., Wearsch, P. A. & Cresswell, P. Pathways of antigen processing. *Annu. Rev. Immunol.* **31,** 443–473 (2013).

83. Rock, K. L., Gamble, S. & Rothstein, L. Presentation of exogenous antigen with class I major histocompatibility complex molecules. *Science* **249,** 918–921 (1990).

84. Penna, A. *et al.* Hepatitis B virus (HBV)-specific cytotoxic T-cell (CTL) response in humans: characterization of HLA class II-restricted CTLs that recognize endogenously synthesized HBV envelope antigens. *J. Virol.* **66,** 1193–1198 (1992).

85. Nuchtern, J. G., Biddison, W. E. & Klausner, R. D. Class II MHC molecules can use the endogenous pathway of antigen presentation. *Nature* **343,** 74–76 (1990).

86. Stanley, M. HPV - immune response to infection and vaccination. *Infect. Agent. Cancer* **5,** 19 (2010).

87. Stanley, M. Immune responses to human papillomavirus. *Vaccine* **24 Suppl 1,** S16–22 (2006).

88. Metzger, D. W. & Sun, K. Immune dysfunction and bacterial coinfections following influenza. *J. Immunol. Baltim. Md 1950* **191,** 2047–2052 (2013).

89. Sore Throat.

90. Rock, K. L. & Shen, L. Cross-presentation: underlying mechanisms and role in immune surveillance. *Immunol. Rev.* **207,** 166–183 (2005).

91. Haralambieva, I. H. *et al.* Genome-wide characterization of transcriptional patterns in high and low antibody responders to rubella vaccination. *PloS One* **8,** e62149 (2013).

92. Ovsyannikova, I. G., Pankratz, V. S., Vierkant, R. A., Jacobson, R. M. & Poland, G. A. Consistency of HLA associations between two independent measles vaccine cohorts: a replication study. *Vaccine* **30,** 2146–2152 (2012).

93. Ercolini, A. M. & Miller, S. D. The role of infections in autoimmune disease. *Clin. Exp. Immunol.* **155,** 1–15 (2009).

94. Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. *Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution*. (2014).

95. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PloS One* **7,** e34267 (2012).

96. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

97. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).

98. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinforma. Oxf. Engl.* (2014). doi:10.1093/bioinformatics/btu704

99. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

100. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14,** 192–200 (2014).

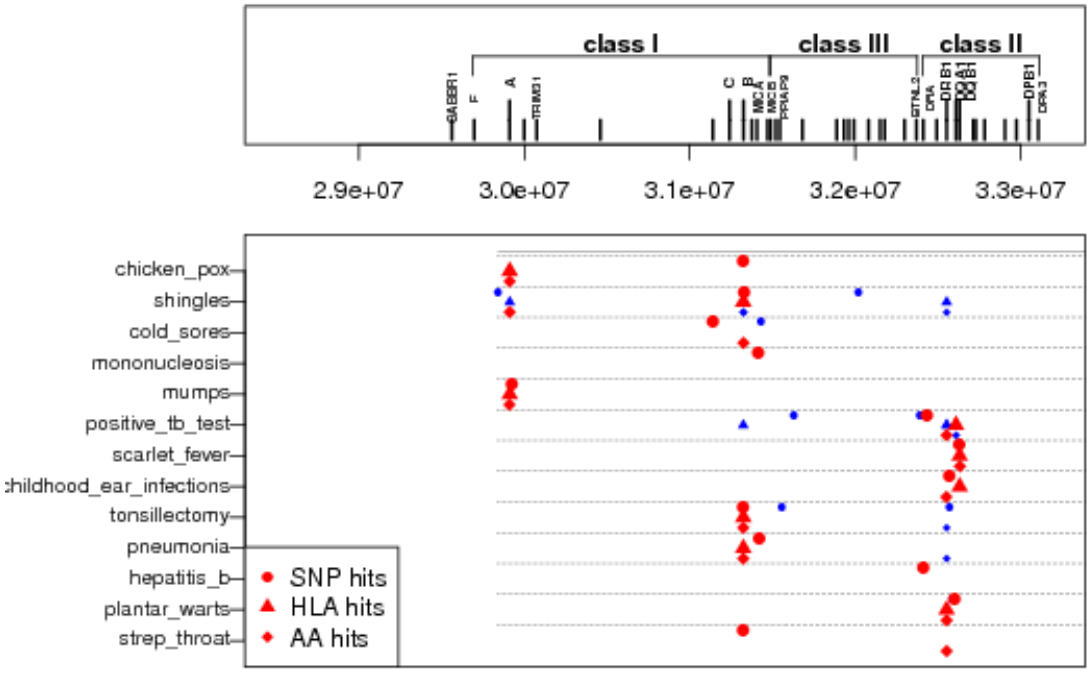
101. Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PloS One* **8,** e64683 (2013).

102. Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** (2010).

103. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

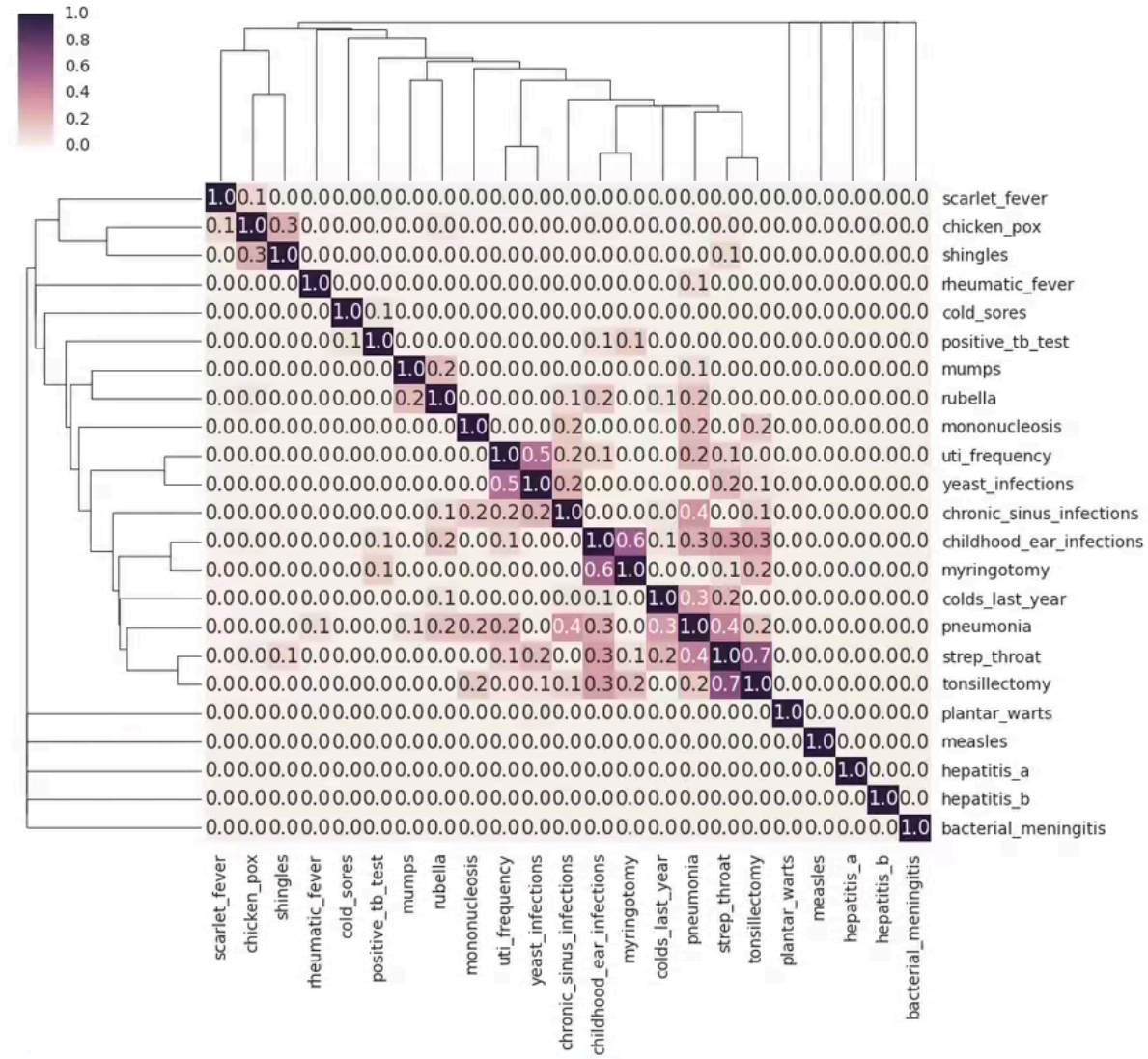
Figure Titles and Legends

Figure 1. Summary of Independent HLA Signals



The strongest associated signals (red) and putative independent secondary signals (blue) are shown for each disease along with their location within the region. Independent secondary signals were defined as those with residual conditional association using the same significant threshold as the primary association signal mentioned in the main text.

Figure 2. Heat map showing the genetic correlations



Each square [i, j] shows the lower bound of the 95% confidence interval of estimated genetic correlation (rg) between trait i and trait j, where i indexes row and j indexes columns. Darker colors represent larger genetic correlations. All negative values to zero. Phenotypes were clustered by 'Nearest Point Algorithm' implemented in python "Seaborn" package.

Tables

Table 1 Discovery cohort characteristics

Phenotypes with significant GWAS findings						
	Chicken pox	**Shingles**	**Cold sores**	**Mononucleosis**	**Mumps**	**Hepatitis B**
Cases	107,769	16,711	25,108	17,457	31,227	1,425
Control	15,982	118,152	63,332	68,446	54,153	218,180
	Plantar warts	**Positive TB**	**Strep throat[a]**	**Scarlet fever**	**Pneumonia**	**Bacterial meningitis**
Cases	24,994	4,426	[qt] 52,487	6,812	40,600	842
Control	37,451	84,290	22,017	113,837	90,039	82,778
	Yeast infections[a]	**UTI frequency[a]**	**Tonsillectomy**	**Childhood ear infection**	**Myringotomy**	
Cases	[qt] 52,218	[qt] 35,000	60,098	46,936	4,138	
Control	10,235	33,478	113,323	74,874	85,089	
Phenotypes with no significant GWAS findings						
	Measles	**Hepatitis A**	**Rheumatic fever**	**# of Colds last year[a]**	**Rubella**	**Chronic sinus infection**
Cases	38,219	2,442	1,115	[qt] 43,826	12,000	5,291
Control	47,279	217,137	88,076	15,720	71,597	79,622

[a]. Strep throat, yeast infections, urinary tract infection (UTI) frequency and number of colds last year were quantitative traits ([qt]).

Table 2: Genome-wide Significant Associations for Each Disease

Phenotype	Cytoband	Gene context	Variants	Position	Alleles[a]	Freq	OR/effect[b]	(95% CI)	P-value
Chickenpox	6p21.33	[HLA]	rs9266089	31320509	G/A	0.85	1.12	(1.1-1.14)	1.00E-10
Shingles	6p21.33	[HLA]	rs2523591	31326960	G/A	0.58	1.14	(1.13-1.16)	1.74E-27
	9p21.3	[IFNA21]	rs7047299	21167465	A/G	0.56	1.07	(1.06-1.09)	1.67E-08
Cold sores	6p21.33	[HLA]	rs885950	31140152	C/A	0.42	1.08	(1.07-1.09)	7.47E-13
Mononucleosis	6p21.33	[HLA]	rs2596465	31412948	T/C	0.47	1.08	(1.06-1.09)	2.48E-09
Mumps	19q13.33	[FUT2]	rs516316	49206145	C/G	0.48	1.25	(1.24-1.27)	9.63E-72
	6p22.1	[HLA]	rs114193679	29925103	G/C	0.99	1.72	(1.62-1.84)	2.23E-17
	14q32.2	C14orf132--[]--BDKRB2	rs11160318	96659919	G/A	0.68	1.1	(1.08-1.11)	4.56E-12
	11q24.2	[ST3GAL4]	rs3862630	126243649	T/C	0.12	1.13	(1.1-1.15)	1.21E-08
Hepatitis A	19q13.2	[IFNL4]	rs66531907	39740675	A/C	0.21	1.23	(1.14-1.33)	5.70E-08
Hepatitis B	6p21.32	[HLA]	rs9268652	32409056	G/A	0.74	1.32	(1.25-1.38)	3.14E-09
Plantar warts	6p21.32	[HLA]	rs9272050	32599071	G/A	0.38	1.19	(1.17-1.21)	9.66E-31
	1q21.3	CRCT1--[]--LCE3E	rs6692209	152506444	T/C	0.62	1.08	(1.06-1.09)	5.25E-09
Positive TB Test	6p21.32	[HLA]	rs2894257	32433276	C/G	0.53	1.36	(1.33-1.39)	8.16E-36
Strep throat	6p21.33	[HLA]	rs1055821	31321919	T/G	0.06	0.08	(0.06-0.09)	7.69E-11
	1p36.23	ERRFI1---[]---SLC45A1	rs35395352	8266829	D/I	0.72	0.03	(0.03-0.04)	3.90E-08
Scarlet fever	6p21.32	[HLA]	rs36205178	32627938	C/T	0.81	1.24	(1.21-1.28)	9.49E-14
Pneumonia	6p21.33	[HLA]	rs3131623	31418810	T/A	0.85	1.1	(1.09-1.12)	1.99E-15
Bacterial meningitis	17q21.33	[CA10]	rs1392935	49939129	G/A	0.91	1.78	(1.6-1.99)	1.19E-08
Yeast infections	18q12.1	[DSG1]	rs200520431	28918197	D/I	0.07	0.11	(0.09-0.12)	1.87E-16
	14q23.1	PRKCH	rs2251260	62023767	T/C	0.17	0.05	(0.04-0.06)	3.46E-10
	14q32.2	[]---C14orf177	rs7161578	98586162	T/C	0.39	0.04	(0.03-0.04)	4.04E-08
UTI frequency	8q24.3	JRK-[]-PSCA	rs2976388	143760256	G/A	0.56	0.04	(0.04-0.05)	3.27E-10
	15q15.3	[FRMD5]	rs146906133	44262951	T/C	1.00	0.38	(0.32-0.45)	2.02E-08
Tonsillectomy	12p13.31	[LTBR]	rs10849448	6493351	A/G	0.25	1.13	(1.12-1.14)	2.35E-35
	22q12.2	[MTMR3]	rs201112509	30324654	I/D	0.69	1.12	(1.11-1.13)	3.39E-34

17p11.2	[TNFRSF13B]	rs34557412	16852187	G/A	0.01	1.59	(1.52-1.67)	2.66E-21
6p21.33	[HLA]	rs41543314	31322690	G/A	0.03	1.23	(1.2-1.26)	5.36E-21
13q33.3	[TNFSF13B]	rs200748895	108960380	D/I	0.03	1.26	(1.23-1.29)	1.20E-17
7p12.3	IGFBP3---[]	rs80077929	46094089	T/C	0.11	1.11	(1.1-1.13)	1.74E-15
4q24	[NFKB1]	rs230523	103462038	T/C	0.67	1.07	(1.06-1.08)	4.54E-14
7p15.2	HOXA1-[]-HOXA2	rs6668	27138183	T/C	0.37	1.06	(1.06-1.07)	1.71E-13
14q21.1	MIPOL1-[]--FOXA1	rs148131694	38025098	T/C	0.44	1.07	(1.06-1.08)	7.86E-13
4q24	[TET2]	rs1391439	106151642	G/A	0.40	1.06	(1.05-1.07)	2.76E-12
2p14	[SPRED2]	rs201473667	65562862	D/I	0.47	1.06	(1.05-1.07)	5.39E-12
20q13.12	NCOA5--[]-CD40	rs6032664	44739419	A/T	0.26	1.07	(1.06-1.08)	7.31E-12
7p12.2	C7orf72---[]--IKZF1	rs876037	50308692	T/A	0.68	1.06	(1.05-1.07)	1.12E-11
13q21.33	[KLHL1]	rs9542155	70579527	T/C	0.32	1.06	(1.05-1.07)	1.92E-11
7p22.2	[GNA12]	rs2644312	2841164	G/A	0.71	1.06	(1.05-1.07)	2.73E-11
3q21.2	[SLC12A8]	rs1980080	124911841	C/T	0.33	1.06	(1.05-1.07)	3.16E-11
16p11.2	[SBK1]	rs141876325	28330968	I/D	0.28	1.06	(1.05-1.07)	4.72E-11
1q41	DUSP10---[]---HHIPL2	rs12126292	222165909	G/T	0.12	1.09	(1.07-1.1)	7.21E-11
4p15.2	LGI2--[]-SEPSECS	rs10939037	25114601	A/G	0.56	1.05	(1.05-1.06)	1.42E-10
12q24.12	[SH2B3]	rs3184504	111884608	T/C	0.50	1.05	(1.04-1.06)	3.07E-10
4q21.1	[CXCL13]	rs7685785	78440425	C/T	0.89	1.09	(1.07-1.1)	4.52E-10
7q31.2	MDFIC---[]---TFEC	rs2023703	114944919	C/A	0.33	1.05	(1.05-1.06)	7.79E-10
19p13.2	ADAMTS10---[]--ACTL9	rs2918308	8789912	A/C	0.84	1.07	(1.06-1.08)	1.10E-09
17p13.3	[RAP1GAP2]	rs67968065	2729247	D/I	0.52	1.05	(1.04-1.06)	2.08E-09
21q22.11	[ITSN1]	rs200746495	35052995	D/I	0.87	1.07	(1.06-1.09)	4.45E-09
2q33.3	[ADAM23]	rs1448903	207308961	G/A	0.09	1.08	(1.07-1.1)	5.13E-09
1p36.23	RERE--[]--ENO1	rs12068123	8894346	G/A	0.45	1.05	(1.04-1.06)	5.60E-09
11p15.4	[ST5]	rs11042055	8756856	A/G	0.43	1.05	(1.04-1.06)	6.96E-09
20q13.33	[SAMD10]	rs41278232	62606617	G/A	0.89	1.08	(1.06-1.09)	7.54E-09
9q34.2	ABO-[]--SURF6	rs635634	136155000	T/C	0.20	1.06	(1.05-1.07)	8.47E-09
17q11.2	[FOXN1]	rs62066768	26858395	A/G	0.08	1.09	(1.07-1.11)	1.09E-08
6q23.3	OLIG3---[]---TNFAIP3	rs11757201	138003822	C/G	0.20	1.06	(1.05-1.07)	2.97E-08

	16p11.2	MAPK3--[]--CORO1A	rs12931792	30156398	G/A	0.54	1.06	(1.05-1.07)	4.06E-08
	7q31.2	[WNT2]	rs200608253	116917244	I/D	0.52	1.05	(1.04-1.05)	4.07E-08
	22q11.21	[TBX1]	rs41298830	19752943	A/C	0.77	1.06	(1.05-1.07)	4.95E-08
Childhood Ear Infections	19q13.33	[FUT2]	rs681343	49206462	C/T	0.52	1.11	(1.1-1.12)	3.51E-30
	22q11.21	[TBX1]	rs1978060	19749525	G/A	0.60	1.09	(1.08-1.1)	1.17E-19
	10p12.1	RAB18--[]--MKX	rs2808290	27900882	C/T	0.51	1.07	(1.07-1.08)	5.09E-16
	15q26.2	SPATA8---[]---LINC00923	rs7174062	97452829	G/A	0.73	1.08	(1.07-1.09)	3.49E-14
	6p21.32	[HLA]	rs4329147	32566577	T/C	0.83	1.11	(1.09-1.13)	9.55E-12
	9q34.2	[ABO]	rs8176643	136149095	D/I	0.36	1.06	(1.05-1.07)	3.67E-11
	2p16.1	[EFEMP1]	rs1802575	56093204	G/C	0.87	1.09	(1.07-1.1)	1.47E-10
	2p24.1	NT5C1B-RDH14---[]---OSR1	rs5829676	19274784	D/I	0.40	1.06	(1.05-1.07)	1.78E-10
	11q13.3	FGF3---[]--ANO1	rs72931768	69829717	G/C	0.88	1.09	(1.07-1.1)	2.63E-09
	7q11.22	[AUTS2]	rs35213789	69268012	C/T	0.74	1.06	(1.05-1.07)	3.75E-09
	7q22.3	[CDHR3]	rs114947103	105658927	C/T	0.18	1.07	(1.06-1.08)	5.40E-09
	8q22.2	NIPAL2--[]--KCNS2	rs13281988	99372329	C/G	0.31	1.06	(1.05-1.07)	9.84E-09
	3p21.31	[BSN]	rs67035515	49656530	I/D	0.18	1.07	(1.05-1.08)	1.56E-08
	6q26	[PLG]	rs73015965	161127501	G/A	0.01	1.43	(1.34-1.53)	3.78E-08
Myringotomy	22q11.21	[TBX1]	rs1978060	19749525	G/A	0.60	1.17	(1.14-1.2)	3.34E-10

a. The alleles for SNP are represented as high-risk allele/low-risk allele;

b. Strep throat, plantar warts and UTI frequency are quantitative traits, for which we reported the effect size instead of the odds ratio.

Table 3: HLA Fine-mapping Results for Each Disease

Phenotype	Gene	Cytoband	Variants[b]	Position	Alleles[c]	Freq	OR/ Effect[d]	(95% CI)	Con. P-value	P-value
Chickenpox	[HLA-A]	6p22.1	A_Gly107	29911093	GW	0.71	1.09	(1.07-1.1)	3.77E-10	3.77E-10
		6p22.1	A*02:01	29911953	NA	0.27	0.92	(0.91-0.93)	1.08E-09	1.08E-09
	[HLA-B]	6p21.33	rs9266089	31320509	G/A	0.85	1.12	(1.1-1.14)	1.00E-10	1.00E-10
Shingles	[HLA-A]	6p22.1	rs2523815	29840243	A/G	0.64	1.12	(1.11-1.14)	6.17E-20	1.83E-22
		6p22.1	A_Arg97	29911063	IRM	0.36	0.88	(0.87-0.9)	2.75E-22	2.75E-22
		6p22.1	A*02:01	29911953	NA	0.27	0.88	(0.87-0.89)	1.72E-12	8.91E-20
	[HLA-B]	6p21.33	B*44:02	31323318	NA	0.09	0.81	(0.79-0.82)	3.54E-23	3.54E-23
		6p21.33	B*57:01	31323318	NA	0.04	1.22	(1.18-1.25)	1.03E-08	1.81E-10
		6p21.33	B_Trp147	31324051	WL	0.95	0.92	(0.89-0.94)	6.67E-10	1.24E-03
		6p21.33	B_GluMet45	31324602	EMTKG	0.52	1.11	(1.1-1.13)	9.51E-16	1.65E-18
		6p21.33	rs2523591	31326960	G/A	0.58	1.14	(1.13-1.16)	1.74E-27	1.74E-27
	[TNXB]	6p21.33	rs41316748	32019512	C/T	0.04	1.19	(1.15-1.23)	1.02E-08	1.44E-12
		6p21.32	DRB1*11:01	32552079	NA	0.06	0.79	(0.77-0.82)	9.84E-10	2.96E-11
		6p21.32	DRB1_PheSerHis13	32552131	FSHYGR	0.66	0.92	(0.91-0.94)	1.22E-09	5.90E-10
Cold sores	POU5F1-[]--HCG27	6p21.33	rs885950	31140152	C/A	0.42	1.08	(1.07-1.09)	7.47E-13	7.47E-13
	[HLA-B]	6p21.33	B_ThrGly45	31324602	EMTKG	0.23	0.91	(0.9-0.92)	4.91E-12	4.91E-12
	MICA--[]--MICB	6p21.33	rs4360170	31430359	A/G	0.91	1.13	(1.11-1.16)	3.38E-09	3.41E-12
Mononucleosis	MICA--[]--MICB	6p21.33	rs2596465	31412948	T/C	0.47	1.08	(1.06-1.09)	2.48E-09	2.48E-09
Mumps	[HLA-A]	6p22.1	A_Gln43	29910660	QR	0.99	1.71	(1.6-1.82)	2.51E-17	2.51E-17
		6p22.1	A*02:05	29911953	NA	0.01	0.56	(0.52-0.6)	1.73E-17	1.73E-17
		6p22.1	rs114193679	29925103	G/C	0.99	1.72	(1.62-1.84)	2.23E-17	2.23E-17
Hepatitis B	[HLA-DRA]	6p21.32	rs9268652	32409056	G/A	0.74	1.32	(1.25-1.38)	3.14E-09	3.14E-09
Plantar warts	[HLA-DRB1]	6p21.32	DRB1*04:01	32552079	NA	0.08	1.21	(1.19-1.24)	4.88E-18	4.88E-18
		6p21.32	DRB1*15:01	32552079	NA	0.13	1.14	(1.12-1.15)	3.50E-19	4.48E-14
		6p21.32	DRB1_PheSerTyr13	32552131	FSHYGR	0.63	0.85	(0.84-0.86)	1.84E-39	1.84E-39

		6p21.32	rs28752534	32585068	T/C	0.64	1.11	(1.1-1.13)	9.28E-09	1.50E-30
		6p21.32	rs9272050	32599071	G/A	0.38	1.19	(1.17-1.21)	9.65E-31	9.66E-31
Positive TB Test	[HLA-B]	6p21.33	B*08:01	31323318	NA	0.11	1.29	(1.24-1.33)	1.73E-11	8.43E-14
	[C6orf47]	6p21.33	rs148844907	31628397	A/T	0.01	1.92	(1.74-2.12)	4.95E-10	2.21E-12
	[HLA-DRA-DQA1]	6p21.32	rs3135359	32390578	C/T	0.71	1.21	(1.18-1.24)	2.63E-13	8.82E-21
		6p21.32	rs2894257	32433276	C/G	0.53	1.36	(1.33-1.39)	8.16E-36	8.16E-36
		6p21.32	DRB1_Leu67	32551969	LIF	0.41	1.3	(1.27-1.33)	1.14E-29	1.14E-29
		6p21.32	DQA1*01:02	32608660	NA	0.20	0.75	(0.73-0.78)	7.04E-22	7.04E-22
		6p21.32	DQA1*03:01	32608660	NA	0.10	1.27	(1.22-1.32)	1.59E-09	3.99E-10
		6p21.32	DQA1_His129	32609872	QHx	0.66	1.19	(1.16-1.22)	2.40E-13	1.56E-13
Strep throat	[HLA-B]	6p21.33	rs1055821	31321919	T/G	0.06	0.08	(0.06-0.09)	7.69E-11	7.69E-11
	[HLA-DRB1]	6p21.32	DRB1_CysTyr30	32552080	CYLGRH	0.83	-0.04	(-0.05--0.04)	2.14E-09	2.14E-09
Scarlet fever	[HLA-DQB1]	6p21.32	rs36205178	32627938	C/T	0.81	1.24	(1.21-1.28)	9.49E-14	9.49E-14
		6p21.32	DQB1*03:01	32630853	NA	0.20	0.83	(0.81-0.86)	2.98E-14	2.98E-14
		6p21.32	DQB1_Gly45	32632724	GE	0.80	1.19	(1.16-1.22)	6.35E-14	6.35E-14
Pneumonia	[HLA-B]	6p21.33	B*08:01	31323318	NA	0.11	0.91	(0.9-0.92)	2.87E-11	2.87E-11
		6p21.33	B_SerTrpAsn97	31324201	STWRNV	0.34	0.94	(0.93-0.95)	2.37E-12	2.37E-12
	MICA--[]--MICB	6p21.33	rs3131623	31418810	T/A	0.85	1.1	(1.09-1.12)	1.99E-15	1.99E-15
	[HLA-DRB1]	6p21.32	DRB1_LeuSer11	32552137	LSVGDP	0.52	0.94	(0.94-0.95)	1.42E-09	4.10E-11
Tonsillectomy	[HLA-B]	6p21.33	rs41543314	31322690	G/A	0.03	1.23	(1.2-1.26)	5.36E-21	5.36E-21
		6p21.33	B*57:01	31323318	NA	0.04	1.22	(1.19-1.24)	6.29E-22	6.29E-22
		6p21.33	B_Val97	31324201	STWRNV	0.04	1.21	(1.19-1.24)	1.44E-21	1.44E-21
	[LST1,LST1]	6p21.33	rs1052248	31556581	A/T	0.28	1.05	(1.04-1.06)	3.69E-08	3.15E-17
	[HLA-DRB1]	6p21.32	DRB1_LeuValGly11	32552137	LSVGDP	0.42	1.06	(1.05-1.07)	1.67E-09	1.02E-12
		6p21.32	rs140177540	32569453	D/I	0.47	1.06	(1.05-1.07)	2.92E-12	5.02E-15
	GRM4--[]-HMGA1	6p21.31	rs201220830	34195176	D/I	0.05	1.11	(1.09-1.14)	1.32E-08	3.65E-08
Childhood Ear Infections	[HLA-DRB1]	6p21.32	DRB1_Gln96	32549612	EHYxQ	0.17	0.92	(0.91-0.93)	2.11E-12	2.11E-12
		6p21.32	rs4329147	32566577	T/C	0.83	1.11	(1.09-1.13)	9.55E-12	9.55E-12

	6p21.32	DQB1*06:02	32630853	NA	0.13	0.92	(0.91-0.94)	5.79E-10	5.79E-10

a. We identify independent associations for each category of variants: SNP, HLA (classical allele) and AA (HLA amino acid), separately using iterative conditional regression. For amino acid polymorphism, the label specifies the amino acid and the position. For example, 'A-Val95' means amino acid Val at position 95 of the HLA-A protein. We use this naming convention throughout the paper, which is composed from the name of the HLA protein (A), followed by the three-letter code for the amino acid (Val) and its location in the protein (95). The amino acid polymorphism is highlighted as 'red' if the signal in that gene region can be attribute to it by reciprocal analysis.

b. The alleles for SNP are represented as risk allele/non-risk allele; the allele for HLA classical allele is recorded as 'NA', which means it either has or does not has the tested allele; the alleles for HLA amino acid polymorphisms are all the possible amino acids (one-letter code) at that position.

c. Strep throat, plantar warts and UTI frequency are quantitative traits, for which we reported the effect size instead of the odds ratio.

d. The 'Cond. P-value' column contains conditional p-values that are after iterative conditional regression within each category of variants (SNP, imputed HLA alleles and imputed HLA amino acid polymorphisms). The 'P-value' column contains the unconditional association test p-values.

Table 4: MEGANTA Pathway Analysis on Mumps GWAS

Pathway	Pathway. Pvalue	Pathway .FDR	Gene	Gene.Pvalue	SNP	SNP.Pvalue
Glycosphingolipid biosynthesis	1.00E-04	0.0029	FUT1	0.00E+00	rs584768	2.38E-62
			FUT2	0.00E+00	rs516316	9.63E-72
			ST3GAL4	1.12E-08	rs3862630	1.21E-08
			B3GNT1	4.20E-02	rs7947391	2.42E-04
			GCNT2	4.34E-02	rs147839826	7.16E-05
			B3GNT2	4.46E-02	rs116508040	2.60E-04
			ABO	4.97E-02	9:136149709:AC_A	5.81E-05

The most significant pathway (gene sets) with a gene set enrichment FDR <0.01 are presented. The genes that contain genome-wide significant SNPs in the GWAS analysis are highlighted red.

Table 5: MEGANTA Pathway Analysis on Tonsillectomy GWAS

Pathway	Pathway. Pvalue	Pathway .FDR	Gene	Gene.Pvalue	SNP	SNP.Pvalue
Intestinal immune network for IgA production	2.00E-06	3.50E-03	LTBR	0.00E+00	rs10849448	2.35E-35
			TNFSF13B	0.00E+00	rs200748895	1.20E-17
			TNFRSF13B	0.00E+00	rs34557412	2.66E-21
			CD40	3.19E-10	rs6032664	7.31E-12
			ICOS	1.40E-03	rs231779	1.52E-06
			CXCR4	4.75E-03	rs2680880	6.91E-06
			CCL27	7.23E-03	rs12350154	6.71E-06
			TNFRSF17	8.69E-03	rs456178	1.13E-05
			IL6	9.25E-03	rs57561814	7.45E-06
			IL15RA	1.63E-02	rs11256464	6.99E-06
TACI and BCMA stimulation of B	9.00E-04	7.80E-03	TNFSF13B	0.00E+00	rs200748895	1.20E-17
			TNFRSF13B	0.00E+00	rs34557412	2.66E-21

cell immune responses.			NFKB1	9.08E-14	rs230523	4.54E-14
			TNFRSF17	8.69E-03	rs456178	1.13E-05
			TRAF2	1.21E-02	rs17250694	1.56E-05
Ceramide signaling pathway	3.00E-04	7.95E-03	TNFRSF1A	0.00E+00	rs10849448	2.35E-35
			NFKB1	9.08E-14	rs230523	4.54E-14
			MAPK3	3.78E-06	rs12931792	4.06E-08
			MAP2K2	4.75E-04	rs350906	5.34E-07
			MYC	1.13E-03	rs13279820	1.44E-06
			TRAF2	1.21E-02	rs17250694	1.56E-05
			MADD	3.01E-02	rs2856661	3.76E-05
			CTSD	3.13E-02	rs7929314	3.17E-05
			CASP8	3.15E-02	rs190209406	3.72E-05
Glucocorticoid receptor regulatory network	5.00E-04	9.93E-03	NFKB1	9.08E-14	rs230523	4.54E-14
			MAPK3	3.78E-06	rs12931792	4.06E-08
			NCOA2	9.50E-04	rs1481046	5.65E-07
			SELE	1.66E-03	rs6691453	5.38E-07
			BGLAP	1.94E-03	rs2075165	2.52E-06
			IL6	9.25E-03	rs57561814	7.45E-06
			MAPK11	1.20E-02	rs4076566	1.23E-05
			SMARCA4	1.41E-02	rs9797720	1.58E-05
			SPI1	2.48E-02	rs11604680	3.38E-05
			FKBP5	3.06E-02	rs74853132	1.68E-05
			ICAM1	3.12E-02	rs74956615	4.07E-05
			PCK2	3.34E-02	14:24604399:AAC_A	5.33E-05

The most significant pathway (gene sets) with a gene set enrichment FDR <0.01 are presented. The genes that contain genome-wide significant SNPs in the GWAS analysis are highlighted red.

Table 6: MEGANTA Pathway Analysis on Childhood Ear Infections GWAS

Pathway	Pathway. Pvalue	Pathway .FDR	Gene	Gene.Pvalue	SNP	SNP.Pvalue
Catalysis of the transfer of organic anions from one side of a membrane to the other	9.00E-04	0.0062	SLCO1A2	1.44E-03	rs10841784	2.76E-06
			SLCO1B1	3.45E-03	rs2417971	5.67E-06
			ABCC3	4.74E-03	rs186049592	1.82E-05
			SLC22A6	3.43E-02	rs67754977	1.20E-04
			SLC22A8	3.80E-02	rs67754977	1.20E-04

The most significant pathway (gene sets) with a gene set enrichment FDR <0.01 are presented. The genes that contain genome-wide significant SNPs in the GWAS analysis are highlighted red.