Phenome-wide HLA association study of Finnish biobank participants reveals infection-autoimmune disease links and modifier effects between HLA genes and alleles

Jarmo Ritari¹*, Satu Koskela¹, Kati Hyvärinen¹, FinnGen², Jukka Partanen¹*

¹Finnish Red Cross Blood Service, Helsinki, Finland
²Full list of participants and affiliations available as a supplementary file

*Correspondence: jarmo.ritari@bloodservice.fi; jukka.partanen@bloodservice.fi

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

The human leukocyte antigen (HLA) system is the single most important genetic susceptibility factor for many autoimmune diseases and immunological traits. However, in a range of clinical phenotypes the impact of HLA alleles or their combinations on disease risk is not comprehensively understood.

For systematic population-level analysis of HLA-phenotype associations we imputed the alleles of classical HLA genes in a discovery cohort of 146,630 and a replication cohort of 89,340 Finnish individuals for whom SNP genotype data and 3,355 disease phenotypes were available as part of the FinnGen project.

In total, 3,649 significant single HLA allele associations in 368 phenotypes were found in both cohorts. In addition to known susceptibility alleles, we discovered a number of previously poorly established HLA associations. For example, DRB1*04:01-DQB1*03:02, a frequent high-risk haplotype for many autoimmune diseases, was also independently associated with infectious diseases. Conditional analyses to distinguish protective effects from nonpredisposition showed that in 21 disease categories the effect of the high-risk allele was significantly decreased by a heterozygous allele in the same locus. Furthermore, in many immunological diseases the strength of the top risk-allele was significantly modified by an HLA gene of a different class.

The results highlight the complex structure of HLA-disease associations and suggest that the entire HLA composition should be considered in genetic risk estimation and functional studies. Independent cross-phenotype HLA class II associations imply pleiotropic effects particularly with autoimmune and infectious diseases, supporting a link between environmental exposure and immunogenetics in these diseases.

Introduction

Regulation of adaptive immune system function is based on recognition of foreign antigens and infectious agents by human leukocyte antigen (HLA) receptors encoded by highly polymorphic genes within the major histocompatibility complex (MHC) on chromosome 6. Of more than 200 genes harbored by the MHC region approximately half have known immune-related
functions (The MHC sequencing consortium 1999). HLA molecules play a key role in the initiation of the immune response by binding internal (HLA class I molecules A, B, and C) and external (HLA class II molecules DR, DQ, and DP) peptides and presenting them to T lymphocytes. While class I receptors present antigens directly to cytotoxic CD8+ T cells, class II molecules are recognized by CD4+ T cells that polarize into different regulatory subtypes (A. Barr, Gray, and Gray 2012). The extremely high genetic polymorphism of HLA genes results in structural variation in the peptide binding pockets between HLA alleles, consequently leading to different peptide-binding preferences and varying antigen repertoires presented to T cells.

Originally discovered over 50 years ago as the major determinant of organ and hematopoietic graft rejection (Thorsby 2009), genetic variation in HLA has since been linked to a wide spectrum of immunological diseases (Trowsdale and Knight 2013). In major multifactorial autoimmune diseases, HLA alleles and their protein-level motifs present the most important single genetic component in disease susceptibility (Matzaraki et al. 2017), even though in most diseases, the triggering peptide complexing with the implicated HLA protein polymorphism remains unknown (Dendrou et al. 2018). On the other hand, varying degrees of protective allelic effects distinguished from the absence of strong susceptibility alleles have been reported for major autoimmune disorders (Bettencourt et al. 2015; van der Helm-van Mil et al. 2005; van Lummel et al. 2019). The effect toward the reduction of disease risk is presumably mediated through presentation of a favorable selection of antigens in terms of specificity and self-regulation (Tsai and Santamaria 2013). Accordingly, both susceptibility and resistance effects have been attributed to amino acid residues and their positions in the HLA protein sequence (Gregersen, Silver, and Winchester 1987; Furukawa et al. 2017; Raychaudhuri et al. 2012). Different alleles sharing a similar structural motif also manifest in local epistasis. Detailed analyses of large cohorts of patients with rheumatoid arthritis or type 1 diabetes have demonstrated that the MHC-mediated risk can be pinpointed to specific amino acid positions, and the effect is being modified non-additively by amino acid polymorphisms in a few other positions in the same or different class II gene (Lenz et al. 2015; Hu et al. 2015; Okada et al. 2016).

HLA allelic variance can cause differences in the strength of the immune response against infectious agents such as HIV by differential preference of viral peptides (The International HIV Controllers Study 2010). However, in the case of
structural similarity between the pathogen T cell epitope and a host peptide, an immune reaction against the antigen may also increase the likelihood of developing autoimmunity (Oldstone 1998). Predisposition to infections before the onset of an autoimmune condition has been reported in several cases (Sfriso et al. 2010), and reaction of host T cell clones against the pathogen epitope mimicking host structures has been demonstrated experimentally (Wucherpfennig and Strominger 1995). Nevertheless, exposure to a rich microbial environment also contributes to achieving a protective, tolerogenic setting through Toll-like receptor, regulatory T cell and interleukin signaling (Bach 2018). Immunological regulation and its perturbation are therefore dependent on both environmental and host genetic factors that are mediated by individually varying HLA presentation.

Large biobank genome data collections combined with electronic health records have made phenome-wide association studies (PheWAS) feasible (Denny et al. 2013), leading to increased power and novel discoveries in disease genetics (Diogo et al. 2018; Verma et al. 2018; Liu et al. 2016). A population-based approach for the analysis of the phenotypic spectrum of HLA associations can provide novel insights into the architecture of well-established autoimmune and immune disease associations and broaden the view toward other traits as well (Karnes et al. 2017; Liu et al. 2016; Hirata et al. 2019). The first reported HLA PheWAS analysis with over 11,000 individuals found eight novel phenotypes linked with MHC SNPs as well as five previously unknown associations across multiple phenotypes (Liu et al. 2016). Karnes and coworkers (2017) imputed HLA alleles from cohorts of 28,839 and 8,431 individuals of European origin and tested HLA associations with 1,368 phenotypes. A total of 104 significant associations were observed with 29 phenotypes and 29 HLA alleles. In addition to well-established HLA associations, four novel phenotypes were reported. Hirata and coworkers (2019) analyzed 106 clinical phenotypes for association with MHC variation in a cohort of 166,190 individuals from Japan. They reported significant genotype-phenotype associations in 52 phenotypes, and their fine-mapping showed multiple different patterns of HLA associations, some of which were independent from classical HLA genes.

Here, we report a systematic, population-based association study of imputed HLA alleles in 3,355 phenotypes in discovery and replication cohorts of 146,630 and 89,340 individuals, respectively. Our aim was to perform HLA analysis in diseases not studied in detail before and to define cross-phenotype
dependencies of allelic associations, particularly between autoimmune and infectious diseases. Furthermore, as a systematic examination of risk-modifying effects has not, to our knowledge, been implemented at the biobank scale to date, we sought to define protective allelic effects as opposed to nonpredisposition to the top risk alleles. To this end, we studied heterozygous risk allele genotypes and hypothesized that a risk allele effect could also be modified by an HLA locus of a different class.

Materials and Methods

Subjects and clinical endpoints

The FinnGen project (www.finngen.fi) aims to collect and to analyze genome data together with detailed electronic health records of 500,000 Finnish individuals, i.e., 10% of the entire Finnish population. The FinnGen participants are listed in Supplementary Text. The discovery cohort of the study included all biobank participants in the FinnGen data release R3 (ntotal = 146,630), while the independent replication cohort comprised the data release R5 (without R3; ntotal = 89,340). The numbers of cases and controls for each phenotype are given in Supplementary Table 1. The clinical disease endpoint definitions were curated from ICD 9-10, ICD-O-3, the Social Insurance Institute (KELA) drug reimbursement codes and ATC-codes as a part of the FinnGen project (finngen.gitbook.io/documentation/methods/endpoints). For clarity, the FinnGen phenotypes include many partially overlapping diseases or traits, particularly in diabetes and its comorbidities. Thus, the included phenotypes are not necessarily independent.

All patients and control subjects provided informed consent for biobank research in accordance with the Finnish Biobank Act, with the exception of FinnGen legacy samples which were approved by the National Supervisory Authority for Welfare and Health (Valvira). The FinnGen study protocol was approved by the Ethical Review Board of the Hospital District of Helsinki and Uusimaa (Nr HUS/990/2017). The FinnGen study is approved by the Finnish Institute for Health and Welfare (THL), approval number THL/2031/6.02.00/2017, amendments THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, Digital and population data service agency VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3, the


Genotyping


HLA allele analysis

We implemented the PheWAS approach (Denny et al. 2013) for imputed alleles of HLA-A, -B, -C, -DRB1, -DQA1, -DQB1 and -DPB1 genes to analyze their correlation with 3,355 clinical case-control endpoints in 37 broad disease categories. Each analyzed phenotype included at least five cases in both the discovery and replication sets. HLA imputation at four-digit resolution (i.e., protein-level) was conducted as described previously (Ritari et al. 2020). Briefly, we used the HIBAG v1.18.1 (Zheng et al. 2014) R library with a Finnish population-specific HLA reference panel (n = 1,150) based on ~4,500 SNPs within the MHC region (chr6:28.51-33.48 Mb; hg38/GRCh38) and considered
imputation posterior probabilities >0.5 as acceptable. For association analyses, we defined the imputed HLA alleles as biallelic SNPs and assumed additive effects of allele dosages on the binary phenotype. Logistic regression models were run using SPAtest v3.0.2 (Dey et al. 2017) in R v3.6.3 (R Core Team 2020) with the top 10 genetic principal components (PCs), age and sex as covariates. To correct for multiple testing under dependency and to identify associations for validation in the replication cohort, we applied adaptive Benjamini-Hochberg (Kim and van de Wiel 2008; Benjamini and Hochberg 2000) procedure to the discovery cohort SPAtest saddlepoint approximated p-values using the R library mutoss v0.1-12 (MuToss Coding Team et al. 2017) at FDR <0.01 threshold. We considered an association valid if the replication p-value was <0.01 and the effect direction was consistent with the discovery cohort.

To evaluate independent contributions of HLA alleles associated with multiple disease categories, we performed conditional analyses that systematically included a phenotype from a different disease category as an additional covariate. In this analysis, we used the whole dataset (data release R5) and genome-wide p-value threshold of 5x10^-8. To exclude phenotypes in strong correlation with each other from the analysis, we first computed an all-vs-all Pearson’s correlation matrix between the phenotypes and removed those with a correlation of >0.8 with another phenotype. With this step, our purpose was to exclude redundant phenotypes containing a significant proportion of overlapping subjects. Association for each HLA allele with a given phenotype was performed by including a different, noncorrelating phenotype as a covariate along with age, sex and 10 genetic PCs using SPAtest as described above.

**HLA diplotype analysis**

To study systematically whether the primary risk-allele association effect was impacted by other alleles in the same locus, we defined HLA diplotypes as top risk-allele heterozygotes for each phenotype in a given locus. The top risk alleles were identified based on the lowest significant single-allele p-value for each phenotype in the discovery cohort. We performed conditional regression analyses by including additive terms for all the risk-allele genotypes in the target locus as covariates in the model. With this approach, our aim was to quantify actual allelic effects as distinguished from nonpredisposition to the risk allele.
Similar to single-allele analysis, the top 10 genetic PCs, age and sex were included as other covariates.

To identify significant effects relative to the top risk genotype for a given phenotype, we performed a two-tailed Z-test on the obtained conditional logistic regression coefficients (betas) and their standard errors.

**HLA haplotype analysis**

The haplotype analysis was based on the observation that in some phenotypes, a significant allelic association was found both in HLA class I and class II. To evaluate whether alleles in a class I locus affected the observed risk of an allele in a class II locus, or vice versa, we considered alleles from both class I and II loci to examine how different allelic combinations contributed to the primary risk-allele effect. The top risk-allele for each phenotype was first identified based on the lowest significant single-allele p-value in the discovery cohort and then joined with alleles of a locus of a different class into haplotypic allele combinations. Thus, the primary risk allele was studied in all available allelic combinations of the secondary locus. HLAs were imputed on phased genotype data obtained from genotype imputation, and the combined loci under analysis were selected from the same phase. Additive terms for all haplotypes were included as covariates in the regression models. Similar to the genotype analysis, two-tailed Z-test was used to evaluate the significance of the haplotypic effects.

**SNP analysis**

MHC region (chr6:28.51-33.48 Mb; hg38/GRCh38) SNP analysis was produced as a part of the FinnGen PheWAS pipeline for the R5 data release. The association tests and other procedures, described in detail in https://finngen.gitbook.io/documentation/methods/phewas, were run using the SAIGE R library (Zhou et al. 2018).

**Results**
Associations of imputed HLA alleles

Altogether, 155 four-digit HLA alleles were imputed with posterior probability >0.5, and of these, 84 alleles had at least one confirmed association in both cohorts. In total, we found 3,649 significant HLA-allele-phenotype associations in 368 phenotypes (Supplementary Table 1). Of these, 50 phenotypes had only one significantly associated HLA allele while the rest had two or more. Figure 1 summarizes the distribution of allele associations across the main phenotype categories for each HLA gene. HLA class II genes harboured both the largest number of associations and the strongest associations, as indicated by their effect sizes. As expected, the highest numbers of associations were found with DRB1 (857 associations), DQA1 (772) and DQB1 (751), followed by B (549) and HLA-C (426). The chromosomally most distal HLA genes, A and DPB1, had substantially lower numbers of associations (91 and 203, respectively). The top disease categories in terms of number of associations were type 1 diabetes and rheumatic diseases. We did not find a relationship between the number of significant associations and the number of available cases in a phenotype (Supplementary Figure 1).

A total of 1,620 of the 3,649 replicated HLA associations were in diabetes-related traits (Supplementary Table 1). The highest number of associated HLA alleles, 56, was for type 1 diabetes, wide definition (T1D_WIDE); the lowest p-values were below 10^{-200} for DQA1*03:01 and DQB1*03:02, followed by DRB1*04:01 and DQB1*02:01 with p-values below 10^{-70}. It is of interest that the strongest negative associations in practically all subtypes of diabetes-related diseases and comorbidities were not only by the established protective DQB1*06:02 allele, but also by the DQB1*03:01 allele. There were no clear differences in associated HLA alleles between the different subgroups of diabetes (Supplementary Table 1).

Celiac disease (CD) had the second highest number of HLA associations, 41 alleles; the lowest p-values, below 10^{-150}, were for DRB1*03:01, DQA1*05:01 and DQB1*02:01 followed by other alleles known to be in a strong linkage disequilibrium with this HLA class II haplotype. There appeared to be strong negative associations with DRB1*13:01, DQB1*06:03 and DQB1*06:02.

When ranked according to effect size, there were only a few disease groups among the top 50 HLA associations. The highest beta in the discovery cohort, 2.18, was for B*27:05 and ankylosing spondylitis, followed by two other subgroups of ankylosing spondylitis (B*27:05; 1.92 and 2.06) and guttate
psoriasis (B*37:01; 1.78). In fact, multiple HLA alleles and various subgroups of ankylosing spondylitis, psoriasis, celiac disease and type I diabetes strongly dominated the list of the top 50 effect sizes. The only other phenotypes were microscopic polyangiitis (DRB1*04:04) and ulcerative colitis (DRB1*01:01).

To validate our analysis, we compared our results with previously published HLA PheWAS studies (Karnes et al. 2017; Liu et al. 2016; Hirata et al. 2019). We observed a consistent relationship between the obtained odds ratios of associated HLA alleles or genes and those of three other previously published HLA PheWAS studies (Figure 2A). Furthermore, to evaluate the consistency between the discovery and replication cohorts, we correlated the logistic regression log-odds ratios (betas) for the three types of analysis implemented here: HLA allele, genotype and haplotype. As expected, we observed a strong correlation between the two independent cohorts (Pearson’s correlation coefficient approximately 0.9; Figure 2B).

Fifty phenotypes had merely one significantly associated HLA allele (Supplementary Table 2). Thirteen of these phenotypes were related to infections; in particular, all seven pneumonia-related phenotypes were associated with DRB1*04:04 and viral and bacterial infections with DQA1*03:01. It is of note that even though the effect sizes of these associations were low (typically within the range of 0.1 – 0.5), the numbers of cases amounted to thousands, both in the discovery and replication cohorts. Unspecified parasitic disease (B*27:05, beta = 1.15), unspecified thyroiditis (B*35:01, beta = 1.16), ulcerative colitis with primary sclerosing cholangitis (DRB1*01:01, beta = 1.81), and microscopic polyangiitis (DRB1*04:04, beta = 1.58) had an association beta value over 1.0.

Potentially novel HLA allele associations

We discovered significant (discovery FDR <0.01, replication p <0.01) HLA allele associations in seven phenotypes for which we found scarce prior evidence of HLA association in the literature (Table 1). For example, we observed an association for DQA1*01:03 and DQB1*06:03 in mental and behavioral disorders due to cannabinoids (p-value = 10^-5; beta = 0.6). The DQB1*03:02 allele was a common feature in these associations: drug-induced hypoglycemia without coma, vitreous hemorrhage, otitis externa, acute sinusitis, and trigger finger all
included the DQB1*03:02 association. Scleritis and episcleritis endpoint was associated with B*27:05. The effect sizes in all these associations were relatively low, with beta values between 0.1 and 0.7.

**Cross-phenotype HLA allele associations**

HLA alleles associated independently with multiple phenotypes suggest pleiotropic effects for these alleles. To evaluate possible independence of an HLA association between two phenotypes, we conducted analyses by including a phenotype as an additional covariate in the regression models. We observed that 68 HLA alleles showed evidence of independent association with two or more phenotype categories (Figure 3A). The most prominent of these were DQA1*03:01 and DQB1*03:02, both of which were associated with altogether 13 phenotype categories independently of at least one other category. For these alleles, the 13 significant phenotype categories were associated independently of on average 18 other phenotype categories (Supplementary Table 3).

To study the pleiotropy of HLA susceptibility between autoimmune and infectious diseases in more detail, we narrowed down the results for these phenotypes to include only alleles that in conditional analyses showed evidence of association with both diseases independently of each other. The results are summarized in Figure 3B, showing the alleles, phenotypes and effect sizes of the associations. We found 12 alleles in five infectious and five autoimmune diseases that fulfilled the above criteria of association. Nine HLA alleles, eight of which appeared to be parts of the C*07:01 – B*08:01 – DRB1*03:01 – DQA1*05:01 – DQB1*02:01 and DRB1*04:01 – DQA1*03:01 – DQB1*03:02 haplotypes, as well as B*13:02, predisposed individuals to both autoimmune diseases and infections. Three alleles, all part of the DRB1*13:01 – DQA1*01:03 – DQB1*06:03 haplotype, showed a lower frequency in cases. Altogether, ten alleles were associated with two or more infectious-autoimmune disease pairs. The DQB1*03:02 allele was associated with as many as five infectious-autoimmune disease pairs.

**HLA diploptype associations**

To analyze the effect of HLA risk-allele diplotypes on the level of disease susceptibility, we conducted conditional regression analyses by including the
available diploid allele combinations in the model. We found 225 (discovery FDR <0.01, replication p <0.01) phenotypes representing 21 different phenotype categories associated with at least one risk-allele diplotype (Supplementary Table 4). In 91 phenotypes representing 13 different phenotype categories, the other HLA allele in the same locus exerted a statistically significant (discovery FDR <0.01, replication p <0.01) modifying effect on the risk allele (Supplementary Table 5). Figure 4 shows the significant modifying allelic effects in a representative selection of phenotypes. The results showed a clear genotype-dependent spectrum of risk effects, for example, in type I diabetes, insulin medication and celiac disease. In type I diabetes, as assumed, DQB1*03:02 occurring with DQB1*02:01 conferred the highest risk, followed by DQB1*04:02 and DQB1*05:01 heterozygosities. The profile was similar in insulin medication, except for a high risk conferred by DQB1*03:02 homozygosity. When occurring with the high-risk DQB1*03:02 allele, DQB1*04:02 increased the risk for insulin medication despite having a negative effect direction (-0.16) in the allele-level association test (Supplementary Table 1). In celiac disease, DQB1*02:01 had the single-allele association test effect of 1.72, but when occurring together with DQB1*02:02, the risk increased to near 3.5. The genotype effects in celiac disease in most cases increased the risk: even alleles such as DQB1*06:03 or DQB1*04:02 that showed negative beta values in the single allele association test, contributed toward increasing the DQB1*02:01 mediated risk. The protective effect of non-risk alleles was in some cases sufficient even to remove the entire risk effect; for example, in type 1 diabetes, the effect of the well-estalished protective allele, DQB1*06:02, removed the risk conferred by DQB1*03:02, and in demyelinating diseases of the central nervous system, DRB1*07:01 almost removed the DRB1*15:01 mediated risk effect. Furthermore, in lichen planus, DQB1*03:01 significantly reduced the risk of DQB1*05:01. In seropositive rheumatoid arthritis, DRB1*08:01 and *13:01, which both showed negative beta values in single-allele association tests, showed nearly dominant protection over the high-risk DRB1*04:01 allele while DRB1*04:08 increased the risk even more than DRB1*04:01 homozygosity. Potentially novel heterozygotic effects on the risk allele are listed in Table 1.

**Haplotype associations**
To test whether the risk of a primary risk-allele was affected by alleles of an HLA gene of a different class, we conducted conditional regression analyses by including allelic combinations from two HLA genes. These analyses are termed here as haplotype associations. The analysis was performed using phased data, but we cannot prove that they genuinely formed haplotypes. We found a total of 16 significant haplotype associations (i.e., risk allele haplotypes associated) with 224 phenotypes representing 23 different phenotype categories (Supplementary Table 6). There was a significant (discovery FDR <0.01, replication p <0.01) modifying effect on the risk-allele in 56 phenotypes representing 10 phenotype categories (Supplementary Table 7). Figure 5 shows the results for a representative selection of phenotypes. In type I diabetes, the effect of high-risk DQB1*03:02 was strongly modified by the HLA-B alleles: B*44:27 substantially increased the DQB1*03:02-associated risk, whereas some alleles, in particular B*27:05 and B*40:02, decreased the risk. A similar profile was seen in diabetes-related retinopathy and insulin medication. HLA-B*44:27 is a relatively uncommon allele in the Finnish population and occurs mostly in a DRB1*16:01 – DQB1*05:02 haplotype with a frequency of approximately 0.4% and with DRB1*08:01 – DQB1*04:02 (0.026%) or DRB1*01:01 – DQB1*05:01 (0.018%). In seropositive rheumatoid arthritis, B*08:01 significantly increased the risk conferred by DRB1*04:01. As the frequency of the HLA-B*08:01 – DQB1*03:02 haplotype is very low in Finland (0.08%), it is unlikely that the effect can be attributed to a single haplotype. In B*27:05-associated diseases, the DRB1 alleles modulated the risk. It seems that the most frequent B*27:05 haplotype in Finland, with DRB1*08:01 – DQB1*04:02 (frequency 2.0%) lowered the B*27:05 associated risk whereas the B*27:05 – DRB1*04:08 combination (frequency 0.9%), which in Finland occurs with DQB1*03:01, clearly increased the risk. Potentially novel haplotype modifier effects on the risk allele are listed in Table 1.

**SNP-level associations**

To compare the discovered significant HLA allele associations with SNP association profiles in the MHC region, we collected variant summary statistics data of FinnGen data release R5 encompassing both our discovery and replication cohorts. We found 1,128,450 significant (p <5x10^-8) MHC region SNPs in 264 phenotypes with a significant HLA allele association. The MHC region
contained 58,887 SNPs of which 4,004 were significant on average. In most cases, however, the top MHC SNP did not fall within the gene implicated by the top HLA allele association. This relationship between SNP peaks and HLA allele associations as well as between different phenotype categories is illustrated by Figure 6, showing both infectious (Figure 6A) and rheumatic autoimmune diseases (Figure 6B).

In unspecific bacterial infections (Figure 6A top) the HLA allele association pointed to DQA1*03:01 but the top SNP association mapped to the DRB1 gene. Notably, phenotype Viral and other specified intestinal infections (Figure 6A bottom), which also had DQA1*03:01 as the top HLA allele association, showed a broader SNP association profile from DRA1 to TAP1, obviously due to the known high linkage disequilibrium in the class II segment. Unspecific parasite infections (Figure 6A middle) had the strongest allele association with HLA B*27:05; however, in the SNP analysis the top association was with an HLA-C marker, and genome-wide significant SNP associations included, e.g., the MICA gene. The B*27:05 association of seronegative RA also included SNPs from HLA-C to MICA (Figure 6B top) and the MICA marker rs145575084 showed the strongest association. Other associations included markers near genes DDR1, MUC22 and SLC44A4, as well as class I genes HLA-A, -G and -F. In seropositive RA (Figure 6B bottom), the top HLA allele association was with DRB1*04:01, with the HLA-B gene showing a significant associations as well.

Finally, we observed a consistent relationship between the effect size of the best HLA allele and the effect size of the best MHC SNP (Supplementary Figure 2). However, in a few phenotypes, the relationship deviated from expected due to the top SNP having a substantially higher effect than the top HLA allele (Supplementary Table 8).

**Discussion**

The current study presents the results of a systematic association analysis of imputed HLA alleles with over 3,000 clinical phenotypes in more than 230,000 individuals. In total, we report 3,649 significant and successfully replicated allele-phenotype associations in 368 phenotypes distributed over 35 disease categories. Consistent with previous HLA PheWAS and other reports (Dendrou et al. 2018), our study uncovered well-established associations with major
autoimmune disorders, and also found evidence of HLA pleiotropy (Liu et al. 2016; Karnes et al. 2017), particularly between infectious and autoimmune diseases. As expected, the effect size estimates between the previous studies and our discovery and replication data sets showed overall high concordance, validating the accuracy of HLA imputation and association analyses based on it. The results from conditional analyses focusing on selected combinations of HLA alleles and cross-phenotype associations further add to the existing knowledge by including risk-modifying effects not studied before in a phenome-wide context.

Seven of the associations found in the present study have not, to the best of our knowledge, been linked to HLA before. Five of these were associations with DQA1*03:01 – DQB1*03:02 haplotype and in most cases also with DRB1*04:01. Associations of these alleles with vitreous hemorrhage and drug-induced hypoglycemia may reflect the strong role of DQB1*03:02 in T1D. Susceptibility for mental and behavioral disorders, which were due to cannabinoids, was found to be associated with the DRB1*13:01 – DQB1*06:03 haplotype. The finding, if replicated in independent populations, may be of interest for risk assessment and as a candidate for in-depth mechanistic studies.

The strong enrichment of HLA risk alleles in autoimmune diseases, e.g. DQ8 in T1D, DQ2 in celiac disease, or B27 in arthropathies, automatically leads to lower frequencies of other alleles in the risk locus and consequently to risk-reducing effect estimates irrespective of actual association. Conditional analyses adjusted for allelic variation can reveal genuine effects of the risk-gene HLA genotypes. In line with previous analyses, our HLA genotype PheWAS replicated previously known protective allelic effects, e.g., in demyelinating diseases (DRB1*07:01 and 01:01) (Wu et al. 2010), arthropathic psoriasis (C*07:01) (Queiro et al. 2006), diabetes (DQB1*06:02) (Pugliese et al. 1995), and seropositive RA (DRB1:13:01 and 08:01) (van der Helm-van Mil et al. 2005), and provided estimates for risk-modifying effects of a range of alleles occurring together with the top risk allele in autoimmune disorders. As a novel finding, our results showed a risk-modifying effect of DQB1*03:01 for DQB1*05:01 in lichen planus, helping resolve the somewhat contradictory results obtained by previous serotyping studies on the frequencies of DQ1 and DQ3 in lichen planus patients (Porter et al. 1993; Nasa et al. 1995). Moreover, our genotype results demonstrated that B*51:01 was the only allele that contributed to mitigating the risk for sulfasalazine medication
relative to $B^*27:05$ while not modulating other $B^*27$ associations, suggesting that $B^*51:01$ can have a genuine medication-dependent protective effect.

The population founder effect can lead to reduced genetic diversity and altered frequencies of genetic variants (Chheda et al. 2017), including HLA alleles and haplotypes (Hurley et al. 2020; Creary et al. 2019). The current study was based on a genetically defined cohort of Finnish individuals that constitutes a Northern European genetic isolate. A characteristic genetic architecture is visible in the repertoire of HLA haplotypes in Finland, where a number of Finnish enriched rare (FER) haplotypes are substantially more common than elsewhere in Europe (Lijnama et al. 2018). Our HLA class I – class II haplotype analysis demonstrates both the modifying effect and how its significance can be estimated in a genetically characteristic population. We found that $B^*27:05$ occurring together with $DRB1^*04:08$ carried the highest risk for seronegative rheumatic diseases, confirming an association that has been previously described in the Finnish population (Tuokko et al. 1997). This allele combination occurs exclusively in the $C^*01:02 – HLA-B^*27:05 – DRB1^*04:08 – DQB1^*03:01$ haplotype that belongs to the FER group and is 3300 times more frequent in the Finnish population than in other European populations. Our study further demonstrated that the predisposing effect of $B^*27:05$ was nearly vanquished by $DRB1^*04:04$. This allele pair is known to occur in the $C^*01:02/02:02 – B^*27:05 – DRB1^*04:04 – DQB1^*03:02$ FER haplotype with a population frequency of 0.24% in Finland but is very rare elsewhere. This example strongly suggests that the HLA class II haplotype modifies the predisposing effect of $B^*27:05$.

Furthermore, multiple $HLA-B$ alleles also showed a modifying effect on the susceptibility to T1D by $DQB1^*03:02$. While HLA classes I and II have been reported to be independently associated with T1D (Mikk et al. 2017; Eike et al. 2009), the compound effect of allelic heterogeneity between HLA classes I and II remains less comprehensively understood. We observed protective effects for HLA class I alleles that by themselves did not have an association with T1D and its comorbidities in our analyses or elsewhere in the literature (Noble and Valdes 2011). For example, $B^*27:05$ and $B^*40:01$ occurring together with $DQB1^*03:02$ reduced the risk conferred by $DQB1^*03:02$, while $B^*44:27$ substantially increased it. The predisposing effect of the uncommon $B^*44:27$ allele in diabetes-related conditions can go unnoticed in mixed populations because of its infrequency or appearance in different class II haplotypes. Allele $B^*44:27$ is also relatively rare in Finland and occurs mostly in the $C^*07:04 – B^*44:27 – DRB1^*16:01$ –
DQB1*05:02 haplotype, with a population frequency of 0.4%, and with DRB1*08:01 – DQB1*04:02 (0.026%) or DRB1*01:01 – DQB1*05:01 (0.018%). As these haplotypes lack known risk alleles, the causative variant remains unknown but suggests a potential role for B*44:27. Obviously, rare alleles such as B*44:27 and its haplotypes, are not widely studied and the risk factor associated with B*44:27 may not lie in the same haplotype as DQB1*03:02.

In a recent well-powered association study, the MHC region was linked with multiple common infectious diseases, and fine-mapping revealed several independent signals among HLA-gene variants and alleles (Tian et al. 2017). Moreover, in another study on MHC expression quantitative trait loci, protection from bacterial infections in cystic fibrosis by the common autoimmune risk haplotype 8.1 was found to be mediated by a non-HLA gene carried in the same haplotype (D’Antonio et al. 2019). Our finding that certain HLA alleles were shared independently by infectious and autoimmune diseases is intriguing in regard to the proposed triggering role of infections in autoimmunity (Ercolini and Miller 2009). However, the HLA alleles showing pleiotropy are carried in known haplotypes in the population, making it difficult to separate class I and class II involvement. The results on the alleles of the B*13:02 – DQB1*03:02 haplotype showed that class I and II alleles exhibited different associated phenotypes, and thus the alleles can have effects that are not explained by linkage disequilibrium alone. In the Finnish population, DQB1*03:02 occurs most frequently with B*15:01 (haplotype frequency ~2.5%), yet this allele had no clear role in DQB1*03:02-associated infections, supporting independence of class I and II. In contrast, C*07:01 – B*08:01 – DQB1*02:01 were associated with the same infectious phenotype, namely, sexually transmitted diseases, but class I and class II showed different autoimmune susceptibilities. This could indicate a common factor behind the infection that triggers autoimmunity manifesting in different ways depending on the haplotypic context. In many infections, the strong inflammatory response rather than the infection as such causes tissue damage. This tissue damage and subsequent unveiling of triggering autoantigens may in genetically predisposed individuals lead to more severe immunological imbalance, ultimately precipitating into autoimmune disease (Fujinami et al. 2006). For example, in Puumala hantavirus infection, most of the population presents with only mild symptoms, but individuals in need of hospital care due to kidney damage or shock were reported to have the common autoimmune risk B8 - DR3 haplotype (Mustonen et al. 1996). Other explanations,
including pathogen-driven selection, molecular mimicry, viral persistence and bystander activation, have been suggested as links between infections and autoimmune diseases (Fujinami et al. 2006). As evidence of pleiotropy was also reported in two previous MHC PheWASs (Karnes et al. 2017; Liu et al. 2016), it will be of great interest to try to reveal the mechanistic background for these shared associations, especially between infections and autoimmunity (Matzaraki et al. 2017; D’Antonio et al. 2019).

Our study is also limited in several respects. First, analysis of HLA alleles cannot definitively attribute the observed associations directly to HLA owing to strong linkage disequilibrium within the MHC. Despite the recognized role of HLA in immune-mediated traits, our observations could in principle be explained by nearby loci, because in many cases the association profile extended beyond the HLA genes (Trowsdale and Knight 2013). For example, the known associations between disorders of iron metabolism and A*03:01 and between disorders of the adrenal gland and DRB1*04:04 at least partially are a result of linkage disequilibrium with the HFE gene and CYP21 gene, respectively. Whether the primary associations pinpoint to HLA alleles, individual SNPs or particular amino acid residues shared by HLA molecules requires further research. Second, our study is restricted by statistical power particularly in conditional analyses with many covariates and in endpoints having a low number of cases. While the independent replication data subset employed in the study helps eliminate nonsystematic false positives arising from, e.g., relatedness, batch and other chance factors, it cannot categorically rule them out or remove sampling uncertainty in low-powered endpoints. Third, haplotype analysis in this study cannot prove that the alleles are encoded in cis but rather is based on likelihoods of preanalyzed haplotypes in the population. The effects between two HLA genes can also take place in trans. Finally, the biobank cohorts analyzed in the present study have been collected over a period of several years and consist of subcohorts of different types of patient groups, as well as control groups such as blood or bone marrow donors who may have lower disease burden than the general population. As such, individual data releases or the data collection as a whole do not represent an unbiased sampling from the population in terms of clinical phenotypic variation.

In conclusion, the results of the present study illustrate the role of HLA alleles separately and in tandem in immune-mediated diseases, providing a data resource for future HLA analyses in independent populations. The complex
genetic structure of HLAs and MHC in general motivates the consideration of several linked genes in risk calculations and as a starting point of functional studies focusing on mechanistic molecular underpinnings of these diseases.

Acknowledgements

The study was supported by the Academy of Finland, the Finnish Cancer Association, VTR funding from the Finnish Government, and Business Finland.

We are grateful to all FinnGen participants for their generous contribution to the project. FinnGen is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and twelve industry partners (AbbVie Inc, AstraZeneca UK Ltd, Biogen MA Inc, Celgene Corporation, Celgene International Il Sàrl, Genentech Inc, GlaxoSmithKline, Janssen Biotech Inc. Maze Therapeutics Inc., Merck Sharp & Dohme Corp, Novartis, Pfizer Inc., Sanofi).


The funders and biobanks had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare no conflicts of interest.
Author contributions

JP supervised the study. JR conceived of the study design with contributions from JP. JR performed the data analysis and drafted the manuscript. SK provided expertise on genetics of HLA. All authors contributed to interpretation of the results and editing of the manuscript.

Data availability

The FinnGen summary statistics data can be accessed through the Finnish Biobanks’ FinnBB portal (www.finbb.fi).

Code availability

The analysis code is available at https://github.com/FRCBS/HLA_PheWAS. The FinnGen genotyping and imputation protocol is described at https://doi.org.libproxy.helsinki.fi/10.17504/protocols.io.nmndc5e.

References


Mustonen, Jukka, Jukka Partanen, Mari Kanerva, Kari Pietilä, Olli Vapalahti, Amos Pasterнак, and Antti Vaheri. 1996. “Genetic Susceptibility to Severe
Course of Nephropathia Epidemica Caused by Puumala Hantavirus.”


**Table 1.** Potentially novel HLA allele associations and modifying effects.

Diplotype and haplotype analyses show the effects of combinations of two alleles. Here, a strong protective effect on the risk allele can result in a nonsignificant association.

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Phenotype</th>
<th>Primary HLA</th>
<th>Modifying HLA</th>
<th>Discovery</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Beta</td>
<td>SE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DQB1*03:02</td>
<td>DQA1*03:01</td>
<td>4.14E-05</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>Drug-induced hypoglycaemia without coma</td>
<td>DQA1*03:01</td>
<td>DQA1*06:03</td>
<td>4.83E-05</td>
<td>0.663</td>
</tr>
<tr>
<td></td>
<td>Mental and behavioural disorders due to cannabinoids</td>
<td>DQB1*01:03</td>
<td>DQB1*13:01</td>
<td>7.26E-06</td>
<td>0.603</td>
</tr>
<tr>
<td></td>
<td>Vitreous haemorrhage</td>
<td>DQB1*03:02</td>
<td>DQA1*03:01</td>
<td>1.22E-05</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DQB1*06:03</td>
<td>DRB1*13:01</td>
<td>3.23E-05</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>Otitis externa</td>
<td>DQB1*03:02</td>
<td>B*18:01</td>
<td>1.36E-24</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DQA1*03:01</td>
<td>DRB1*14:01</td>
<td>4.83E-05</td>
<td>0.663</td>
</tr>
<tr>
<td></td>
<td>Acute sinusitis</td>
<td>DQB1*03:02</td>
<td>DQA1*01:03</td>
<td>1.97E-07</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRB1*04:01</td>
<td>B*18:01</td>
<td>2.46E-05</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DQA1*03:01</td>
<td>4.32E-05</td>
<td>0.209</td>
</tr>
<tr>
<td></td>
<td>Trigger finger</td>
<td>DQB1*03:02</td>
<td>DQB1*03:01</td>
<td>7.39E-08</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DRB1*04:01</td>
<td>DRB1*04:01</td>
<td>7.67E-08</td>
<td>0.333</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*18:01</td>
<td>DRB1*04:01</td>
<td>1.88E-06</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>Scleritis and episcleritis</td>
<td>DQB1*03:02</td>
<td>B*27:05</td>
<td>8.48E-08</td>
<td>0.579</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DQB1*05:01</td>
<td>DQB1*03:02</td>
<td>2.46E-14</td>
<td>1.323</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DQB1*03:01</td>
<td>DQA1*03:03</td>
<td>2.26E-02</td>
<td>0.385</td>
</tr>
<tr>
<td></td>
<td>Lichen planus</td>
<td>DRB1*04:08</td>
<td>DRB1*04:01</td>
<td>5.84E-35</td>
<td>1.739</td>
</tr>
<tr>
<td></td>
<td>Rheumatoid arthritis</td>
<td>DRB1*10:02</td>
<td>DRB1*10:01</td>
<td>2.50E-05</td>
<td>0.205</td>
</tr>
<tr>
<td></td>
<td>Co-morbidities, CVD and metabolic diseases</td>
<td>B*18:01</td>
<td>DRB1*11:01</td>
<td>6.61E-01</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>Thyroiditis, ILD-related definition</td>
<td>DQB1*03:02</td>
<td>DRB1*14:54</td>
<td>9.20E-01</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>Type 1 diabetes, definitions combined</td>
<td>DQA1*03:01</td>
<td>DRB1*14:01</td>
<td>3.38E-01</td>
<td>-0.167</td>
</tr>
<tr>
<td></td>
<td>Diabetes, kidney failure</td>
<td>B*44:27 a</td>
<td>B*44:02 a</td>
<td>7.41E-06</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>Diabetic maculopathy</td>
<td>B*18:01 a</td>
<td>B*18:01</td>
<td>1.27E-76</td>
<td>0.953</td>
</tr>
<tr>
<td></td>
<td>ILD Co-morbidities, CVD and metabolic diseases</td>
<td>DQB1*03:01 a</td>
<td>DRB1*11:01 a</td>
<td>1.27E-76</td>
<td>0.953</td>
</tr>
<tr>
<td></td>
<td>Other (seronegative) rheumatoid arthritis, wide</td>
<td>B*27:05 a</td>
<td>DRB1*04:01 a</td>
<td>4.71E-15</td>
<td>1.203</td>
</tr>
</tbody>
</table>

* Protective effect was determined relative to this allele combination in diplotype and haplotype analyses.
Figure 1. Overview of HLA allele associations in top phenotype categories. The distribution of association p-values for each analyzed HLA gene in the discovery cohort is shown on a double-logarithmic scale (x-axis). The top disease phenotype categories (y-axis) are shown in descending order of significant associations. The dashed vertical lines from left to right indicate FDR <0.01 and genome-wide significance (p <5x10^{-8}) thresholds, respectively. The results reflect both the number of associated alleles and the number of phenotypes within the disease categories.
**Figure 2.** Comparison of HLA association effects between datasets. **A)** Odds ratios of previously reported HLA PheWAS associations (x-axis) vs. the discovery cohort of the present study (y-axis). Depending on the study, associations are shown either at the level of four-digit alleles (Karnes et al.) or at the gene-level tagged by the highest ranking variant (Liu et al. & Hirata et al.). **B)** Correlation of HLA association FDR <0.01 log-odds ratios (betas) between the discovery cohort (x-axis) and the replication cohort (y-axis) of the present study. Panels from left to right show the data for HLA allele, genotype, and two-locus haplotype association analyses.
Figure 3. Cross-phenotype associations of significant HLA alleles. A) HLA alleles (x-axis) associated independently with two or more phenotype categories (y-axis). B) HLA alleles associated independently with infectious diseases (y-axis) and autoimmune diseases (x-axis). The color-filled squares indicate the effect size and direction. The results are grouped by known haplotypes in each row. DQA1*05:01 is omitted from the first row because its profile is identical to the other two shown class II alleles. The data are based on conditional regression analyses with $p < 5 \times 10^{-8}$ threshold in the full dataset (discovery+replication), where selected phenotypes were analyzed by adding a different phenotype as an additional covariate in the model one at a time.
Figure 4. Effect of the second allele on the top risk allele within the same HLA gene. The results for the risk-allele genotypes are shown in a representative selection of disease phenotypes. The x-axis shows log-odds ratios (betas) for different genotypes depicted on the y-axis. The y-axis label indicates the primary risk allele, and the tick mark labels indicate the other alleles in the same locus. The vertical dashed lines indicate the risk allele’s effect estimates based on allele-level analysis. Only significant (discovery FDR <0.01, replication p <0.01) effects on the risk allele are shown. The error bars indicate standard errors for the beta values.
**Figure 5.** Effects of HLA alleles in a locus other than the top risk allele. The two-locus allelic combinations, termed here haplotypes, are shown in a representative selection of disease phenotypes. The x-axis shows log-odds ratios (betas) for different two-locus allele combinations depicted on the y-axis. The y-axis label indicates the primary risk allele and the tick marks indicate alleles of a different HLA gene. The vertical dashed lines indicate the risk allele’s effect estimates based on allele-level analysis. Only significant (discovery FDR <0.01, replication p <0.01) effects on the risk allele are shown. The error bars indicate standard errors for the beta values.
Figure 6. Manhattan association profiles of MHC region SNPs of selected disease phenotypes. **A**) Bacterial, parasitic and viral infectious diseases with only one significantly associated HLA allele. **B**) Seropositive and seronegative rheumatic disorders harboring both HLA class I and II allele associations and a significant risk-modifying effect between the genes. The position on chromosome 6 is shown by the top x-axis while the bottom x-axis shows gene symbols. SNPs located within genes are shown with a darker shade of blue. The dashed horizontal line marks the genome-wide threshold of significance (p <5e-8). The shown SNP data represent the whole dataset (discovery+replication).