Phenome-wide HLA association landscape of 235,000 Finnish biobank participants

Jarmo Ritari[1]*, Satu Koskela[1], Kati Hyvärinen[1], FinnGen[2], Jukka Partanen[1]*

[1]Finnish Red Cross Blood Service, Helsinki, Finland

[2]Full list of participants and affiliations available as a supplemental file

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

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 the disease risk are 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 replication cohort of 89,340 Finns of whom SNP genotype data and 3,355 disease phenotypes were available as part of the FinnGen project.

In total, 3,649 statistically 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 modified by the other allele of the same gene. Furthermore, in many immunological diseases the strength of the top risk allele was significantly modified by an allele of another HLA gene.

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. Shared HLA alleles in autoimmune and infectious diseases support 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 loci within the major histocompatibility complex (MHC) on chromosome 6. Out of more than 200 genes harbored by the MHC region approximately half have known immune-related functions [1]. The HLA molecules play a key role in the initiation of immune response by binding internal (HLA class I molecules *A, B, C*) and external (HLA class II molecules *DR,*

DQ, DP) peptides and presenting them to T lymphocytes. While class I receptors present antigens directly to cytotoxic CD8+ T cells, the class II molecules are recognized by CD4+ T cells that polarize into different regulatory subtypes [2]. 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 [3], genetic variation in HLA has since been linked to a wide spectrum of immunological diseases [4]. In major multifactorial autoimmune diseases, HLA alleles and their protein-level motifs present the most important single genetic component in disease susceptibility [5], even though in most diseases the triggering peptide complexing with the implicated HLA protein polymorphism remains unknown [6]. On the other hand, varying degrees of protective allelic effects as distinguished from the absence of strong susceptibility alleles have been reported for major autoimmune disorders [7–9]. The effect towards the reduction of disease risk is presumably mediated through presentation a favourable selection of antigens in terms of specificity and self-regulation [10]. Accordingly, both susceptibility and resistance effects have been attributed to amino acid residues and their positions in the HLA protein sequence [11–13]. Different alleles sharing a similar structural motif also manifests 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 [14–16].

HLA allelic variance can cause differences in the strength of immune response against infectious agents such as HIV by differential preference of viral peptides [17]. However, in case of structural similarity between pathogen T cell epitope and a host peptide, immune reaction against the antigen may also increase the likelihood of developing autoimmunity [18]. Predisposition to infections before the onset of an autoimmune condition has been reported in several cases [19], and reaction of host T cell clones against the pathogen epitope mimicking host structures has been demonstrated experimentally [20]. Nevertheless, exposure to a rich microbial environment also contributes to achieving protective, tolerogenic setting through toll-like receptor, regulatory T cell and interleukin signaling [21].

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 [22], leading to increased power and novel discoveries in disease genetics [23–25]. Population-based approach for the analysis of phenotypic spectrum of HLA associations can give novel insights into the architecture of well-established autoimmune and immune disease associations and broaden the view toward other traits as well [25–27]. 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 [25]. 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. 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. These large single-population cohorts enabled us to perform HLA analysis in diseases not studied in detail before and to reveal cross-phenotype dependencies of allelic associations particularly between autoimmune and infectious diseases. Furthermore, as a systematic examination of risk-modifying effects have not, to our knowledge, been implemented at 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 a HLA locus of a different class.

Materials and Methods

Subjects and clinical endpoints

The discovery cohort of the study included all biobank participants in the FinnGen (www.finngen.fi) data release R3 (n_{total} = 146,630) while the independent replication cohort comprised the data release R5 (without R3; n_{total} = 89,340). Numbers of cases and controls for each phenotype are given in the Supplemental 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 an 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). All samples and individual-level data were pseudonymized and processed in accordance with the EU GDPR law.

Genotyping

Genotyping of FinnGen samples was performed on a customized ThermoFisher Axiom array at the Thermo Fisher genotyping service facility (San Diego, USA). Genotype calling and quality control steps are described in finngen.gitbook.io/documentation/methods/genotype-imputation. The array markes files can be downloaded from www.finngen.fi/en/researchers/genotyping. The protocol for genotype liftover to hg38/GRCh38 is described in detail in www.protocols.io/view/genotyping-chip-data-lift-over-to-reference-genome-xbhfij6?version_warning=no, and genotype imputation protocol is described in www.protocols.io/view/genotype-imputation-workflow-v3-0-xbgfijw.

HLA allele analysis

We implemented the PheWAS approach [22] 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 analysed

phenotype included at least five cases in both discovery and replication sets. HLA imputation at four-digit resolution (i.e. protein-level) was conducted as described previously [28]. Briefly, we used HIBAG v1.18.1 [29] 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 bi-allelic SNPs and assumed additive effects of allele dosages on the binary phenotype. Logistic regression models were run using SPAtest v3.0.2 [30] in R v3.6.3 [31] with 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 [32,33] procedure to the discovery cohort SPAtest saddlepoint approximated p-values using the R library mutoss v0.1-12 [34] 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 significantly 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 having a correlation >0.8 with another phenotype. Association for each HLA allele with a given phenotype was performed by including a different, non-correlating phenotype as a covariate along with age, sex and 10 genetic PCs using SPAtest as described above.

HLA diplotype analysis

To systematically study how the association effect of the primary risk allele was impacted by other alleles of the same HLA gene, we performed association analyses for HLA allele combinations (termed here as diplotypes). 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 all the diplotypes in the same model for a given

phenotype. With this approach our aim was to quantify actual allelic effects as distinguished from nonpredisposition to the risk allele. As described above, 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 association was found both in HLA class I and class II genes. To evaluate whether alleles in a class I gene affected the risk of an allele in a class II gene, or vice versa, we considered combinations of alleles from both class I and II. 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 combined with alleles of a HLA gene of a different class (termed here as haplotypes). Thus, the primary risk allele was studied in all available allele combinations of the secondary gene. HLAs were imputed on phased genotype data obtained from genotype imputation, and the combined loci under analysis were selected from the same phase. All haplotypes were included in the same regression model for a given phenotype. Two-tailed Z-test was used to evaluate the significance of the haplotoypic effects.

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 statistically significant HLA-allele-phenotype associations in 368 phenotypes (Supplemental Table 1). Supplemental Figure 1 summarises 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. 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 (Supplemental Figure 2).

1,620 of the 3,649 replicated HLA associations were in diabetes-related traits (Supplemental Table 1) with *DQB1*03:02* as the top risk allele. Celiac disease (CD) had the second highest number of HLA associations. The lowest p-values 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.

To validate our analysis we compared our results with previously published HLA PheWAS studies [25–27]. We observed a consistent relationship between the obtained odds ratios of associated HLA alleles or genes and those of the three other previously published HLA PheWAS studies (Figure 1a). Further, to evaluate the consistency of associations 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, diplotype and haplotype. Expectedly, we observed a strong correlation between the two independent cohorts (Pearson's correlation coefficient about 0.9; Figure 1b).

We discovered statistically 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 behavioural disorders due to cannabinoids (p-value = 10^{-5}; beta = 0.6). Moreover, drug-induced hypoglycaemia without coma, vitreous haemorrhage, otitis externa, acute sinusitis, and trigger finger were all associated with *DQB1*03:02* and scleritis and episcleritis was associated with *B*27:05*.

Cross-phenotype HLA allele associations

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 altogether 68 HLA alleles showed evidence of independent association with two or more phenotype categories (Supplemental Table 2). To study shared HLA associations in autoimmune and infectious diseases, we narrowed down the results for these phenotypes to include only alleles that in conditional analyses showed evidence of associating with infectious diseases independently of at least one autoimmune disease. The results are summarized by Figure 2, showing the alleles, p-values,

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 *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 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 associated with two or more infectious-autoimmune disease pairs.

HLA diplotype associations

To analyze the effect of HLA risk allele diplotypes on the level of disease susceptibility, we conducted conditional regression analyses with diploid allele combinations. We found 225 statistically significant (discovery FDR < 0.01, replication $p < 0.01$) phenotypes representing 21 different phenotype categories associated with at least one risk allele diplotype (Supplemental Table 3). 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 (Supplemental Table 4). Figure 3 shows significant modifying allelic effects in phenotypes that deviated the most from expectation (i.e. the sum of individual allele effects, Figure 3a). For example, in type 1 diabetes, the results replicated the well-estalished protective allele *DQB1*06:02* and showed that *DQB1*04:02* increased the *DQB1*03:02* mediated risk for insulin medication despite having negative effect direction (-0.16) in the allele-level association analysis (Figure 3b). In coeliac disease, alleles such as *DQB1*06:03* or *DQB1*04:02,* that showed negative beta values in the single allele association test, contributed towards increasing the *DQB1*02:01* mediated risk (Figure 3b). Potentially novel heterozygotic effects on the risk allele are listed by Table 1.

HLA haplotype associations

To test whether the effect of a primary risk allele was affected by alleles of a HLA gene of a different class, we conducted conditional regression analyses with

allele combinations from two HLA genes (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 statistically significant haplotype associations with 224 phenotypes representing 23 different phenotype categories (Supplemental Table 5). There was a statistically significant (discovery FDR < 0.01, replication p < 0.01) modifying effect on the risk allele in 56 phenotypes representing 10 phenotype categories (Supplemental Table 6). Figure 4a shows significant modifying allelic effects in phenotypes that deviated the most from expectation (i.e. the sum of individual allele effects). For example, in T1D, even though *B*44:27* by itself was not associated, together with *DQB1*03:02* the risk is increased (Figure 4b). Potentially novel haplotype modifier effects on the risk allele are listed by Table 1.

Discussion

The current study presents results of a systematic association analysis of imputed HLA alleles with over 3,000 clinical phenotypes in more than 235,000 individuals. In total, we report 3,649 statistically significant and successfully replicated allele-phenotype associations in 368 phenotypes distributed over 35 disease categories. Consistently with previous HLA PheWAS and other reports [6], our study uncovered well-established associations with major autoimmune disorders, and also found evidence of HLA pleiotropy [25,26] in particular between infectious and autoimmune diseases. Expectedly, 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, phenotype data and association analyses based on these. 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.

In a recent well-powered association study, MHC region was linked with multiple common infectious diseases, and fine-mapping revealed several independent signals among HLA-gene variants and alleles [35]. Moreover, in another study on MHC expression quantitative trait loci, protection from bacterial infections in cystic fibrosis by the common autoimmune risk haplotype AH 8.1 was found to be mediated by a non-HLA gene carried in the same haplotype [36]. Our finding that certain HLA alleles in common haplotypes were shared by infectious and

autoimmune diseases is intriguing in regard to the proposed triggering role of infections in autoimmunity [37]. The result on the *B*13:02 – DQB1*03:02* and C*07:01 – *DQB1*02:01* haplotypes showed that class I and II alleles exhibited different associated phenotypes, suggesting that these alleles may have effects that are not explained by linkage disequilibrium alone. As evidence of HLA pleiotropy was also reported by two previous MHC PheWASs [25,26], it will be of great interest to try to reveal the mechanistic background for these shared associations, especially between infections and autoimmunity [5,36].

The strong enrichment of HLA risk alleles in autoimmune diseases, e.g. DQ8 in T1D, DQ2 in coeliac 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 diplotype PheWAS replicated known protective allelic effects in e.g. in demyelinating diseases (*DRB1*07:01* and *01:01*) [38], arthropathic psoriasis (*C*07:01*) [39], diabetes (*DQB1*06:02*) [40], and seropositive RA (*DRB1:13:0*1 and *08:01*) [8] and provided estimates for risk-modifying effects of a range of alleles occurring together with the top risk allele in autoimmune disorders. Our results showed a risk-modifying effect of *DQB1*03:01* for *DQB1*05:01* in lichen planus (LP), helping resolve the somewhat contradictry results obtained by previous serotyping studies on frequencies of DQ1 and DQ3 in LP patients [41,42].

Population founder effect can lead to reduced genetic diversity and altered frequencies of genetic variants [43], including HLA alleles and haplotypes [44,45]. The current study was based on genetically defined cohort of Finns that constitute a Northern European genetic isolate. A characteristic genetic architecture is visible in the repertoire of HLA haplotypes where a number of Finnish enriched rare (FER) haplotypes are substantially more common than elsewhere in Europe [46]. Our HLA class I – class II analysis demonstrates how haplotype effects 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 [47]. This allele combination occurs in *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 effectively removed 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.

While HLA class I and II have been reported to be independently associated with T1D [48,49], the compund effect of allelic heterogeneity between HLA class I and II remains less comprehensively understood. We observed protective effects for HLA class I alleles that by themselves did not have association with T1D and its comorbidities in our analyses or elsewhere in the literature [50]. 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 due its infrequency or appearance in different class II haplotypes. Allele *B*44:27* is relatively rare also in Finland and occurs mostly with *C*07:04 – B*44:27 – DRB1*16:01 – DQB1*05:02, DRB1*08:01 – DQB1*04:02* and *DRB1*01:01 – DQB1*05:01*. 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 haplotypes carrying it are not widely studied, and also the risk factor associated with *B*44:27* may not be the same as in *DQB1*03:02* haplotypes.

Our study is also limited in some respects. First, analysis of HLA alleles alone cannot definitively attribute the observed associations directly to HLA owing to strong linkage disequilibrium within the MHC [4]. For example, the known associations between disorders of iron metabolism and *A*03:01*, and that between disorders of adrenal gland and *DRB1*04:04*, at least partially are a result from linkage disequilibrium with *HFE* gene and *CYP21* gene, respectively. Also, most of the rare HLA alleles were not covered by the used imputation panel and consequently the analysis did not cover their possible associations. 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 design of the study helps eliminate non-systematical 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, the FinnGen phenotypes, albeit carefully curated, were derived from health register which cannot be assumed to be totally accurate. Finally, haplotype analysis cannot prove that the alleles are encoded in

cis, but the effects between two HLA genes, or chromosomal regions between them, can also take place in *trans*.

In conclusion, the results of the present study illustrate the role of HLA alleles both separately and in combination in immune-mediated diseases, revealing potentially new HLA-linked disease phenotypes and providing a data resource for future HLA analyses in independent populations. The results expand the view of the complex genetic structure of HLA, motivating the consideration of allele and gene interactions in risk calculations. These results can serve as starting points for functional studies focusing on mechanistic molecular underpinnings of the discovered associations.

Acknowledgements

The study was supported by the Academy of Finland, the Finnish Cancer Association, VTR funding from the Finnish Government, and Business Finland. FinnGen funding statement is available as supplemental information. We are grateful to all FinnGen participants for their generous contribution to the project. 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

1. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401**, 921–923 (1999).
2. A. Barr, T., Gray, M. & Gray, D. B Cells: Programmers of CD4 T Cell Responses. *IDDT* **12**, 222–231 (2012).
3. Thorsby, E. A short history of HLA. *Tissue Antigens* **74**, 101–116 (2009).
4. Trowsdale, J. & Knight, J. C. Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genom. Hum. Genet.* **14**, 301–323 (2013).
5. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol* **18**, 76 (2017).
6. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat Rev Immunol* **18**, 325–339 (2018).
7. Bettencourt, A. *et al.* The Protective Role of HLA-DRB1 13 in Autoimmune Diseases. *Journal of Immunology Research* **2015**, 1–6 (2015).
8. van der Helm-van Mil, A. H. M. *et al.* An independent role of protective HLA class II alleles in rheumatoid arthritis severity and susceptibility. *Arthritis Rheum* **52**, 2637–2644 (2005).
9. van Lummel, M. *et al.* Epitope Stealing as a Mechanism of Dominant Protection by HLA-DQ6 in Type 1 Diabetes. *Diabetes* **68**, 787–795 (2019).

10. Tsai, S. & Santamaria, P. MHC Class II Polymorphisms, Autoreactive T-Cells, and Autoimmunity. *Front. Immunol.* **4**, (2013).
11. Gregersen, P. K., Silver, J. & Winchester, R. J. The shared epitope hypothesis. an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis & Rheumatism* **30**, 1205–1213 (1987).
12. Furukawa, H. *et al.* The role of common protective alleles HLA-DRB1*13 among systemic autoimmune diseases. *Genes Immun* **18**, 1–7 (2017).
13. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291–296 (2012).
14. Lenz, T. L. *et al.* Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat Genet* **47**, 1085–1090 (2015).
15. Hu, X. *et al.* Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk. *Nat Genet* **47**, 898–905 (2015).
16. Okada, Y. *et al.* Contribution of a Non-classical HLA Gene, HLA-DOA, to the Risk of Rheumatoid Arthritis. *The American Journal of Human Genetics* **99**, 366–374 (2016).
17. The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science* **330**, 1551–1557 (2010).
18. Oldstone, M. B. A. Molecular mimicry and immune-mediated diseases. *FASEB j.* **12**, 1255–1265 (1998).
19. Sfriso, P. *et al.* Infections and autoimmunity: the multifaceted relationship. *Journal of Leukocyte Biology* **87**, 385–395 (2010).
20. Wucherpfennig, K. W. & Strominger, J. L. Molecular mimicry in T cell-mediated autoimmunity: Viral peptides activate human T cell clones specific for myelin basic protein. *Cell* **80**, 695–705 (1995).

21. Bach, J.-F. The hygiene hypothesis in autoimmunity: the role of pathogens and commensals. *Nat Rev Immunol* **18**, 105–120 (2018).
22. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102–1111 (2013).
23. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat Commun* **9**, 4285 (2018).
24. Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *The American Journal of Human Genetics* **102**, 592–608 (2018).
25. Liu, J. *et al.* Phenome-wide association study maps new diseases to the human major histocompatibility complex region. *J Med Genet* **53**, 681–689 (2016).
26. Karnes, J. H. *et al.* Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci. Transl. Med.* **9**, eaai8708 (2017).
27. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat Genet* **51**, 470–480 (2019).
28. Ritari, J. *et al.* Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics and Bioinformatics* **2**, lqaa030 (2020).
29. Zheng, X. *et al.* HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J* **14**, 192–200 (2014).
30. Dey, R., Schmidt, E. M., Abecasis, G. R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *The American Journal of Human Genetics* **101**, 37–49 (2017).
31. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2020).

32. Kim, K. I. & van de Wiel, M. A. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* **9**, 114 (2008).
33. Benjamini, Y. & Hochberg, Y. On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83 (2000).
34. MuToss Coding Team *et al. mutoss: Unified Multiple Testing Procedures.* (2017).
35. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat Commun* **8**, 599 (2017).
36. D'Antonio, M. *et al.* Systematic genetic analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with disease. *eLife* **8**, e48476 (2019).
37. Ercolini, A. M. & Miller, S. D. The role of infections in autoimmune disease. *Clinical & Experimental Immunology* **155**, 1–15 (2009).
38. Wu, J.-S. *et al.* Influence of HLA-DRB1 allele heterogeneity on disease risk and clinical course in a West Australian MS cohort: a high-resolution genotyping study. *Mult Scler* **16**, 526–532 (2010).
39. Queiro, R. *et al.* HLA-C locus alleles may modulate the clinical expression of psoriatic arthritis. *Arthritis Res Ther* **8**, R185 (2006).
40. Pugliese, A. *et al.* HLA-DQB1*0602 Is Associated With Dominant Protection From Diabetes Even Among Islet Cell Antibody-Positive First-Degree Relatives of Patients with IDDM. *Diabetes* **44**, 608–613 (1995).
41. Porter, K., Klouda, P., Scully, C., Bidwell, J. & Porter, S. Class I and II HLA antigens in British patients with oral lichen planus. *Oral Surgery, Oral Medicine, Oral Pathology* **75**, 176–180 (1993).
42. Nasa, G. L. *et al.* HLA antigen distribution in different clinical subgroups demonstrates genetic heterogeneity in lichen planus. *British Journal of Dermatology* **132**, 897–900 (1995).

43. Chheda, H. *et al.* Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur J Hum Genet* **25**, 477–484 (2017).
44. Hurley, C. K. *et al.* Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA* **95**, 516–531 (2020).
45. Creary, L. E. *et al.* Next-generation sequencing reveals new information about HLA allele and haplotype diversity in a large European American population. *Human Immunology* **80**, 807–822 (2019).
46. Linjama, T., Eberhard, H.-P., Peräsaari, J., Müller, C. & Korhonen, M. A European HLA Isolate and Its Implications for Hematopoietic Stem Cell Transplant Donor Procurement. *Biology of Blood and Marrow Transplantation* **24**, 587–593 (2018).
47. Tuokko, J. *et al.* Increase of HLA-DRB1*0408 and -DQB1*0301 in HLA-B27 positive reactive arthritis. *Annals of the Rheumatic Diseases* **56**, 37–40 (1997).
48. Mikk, M.-L. *et al.* The association of the *HLA-A*24:02, B*39:01* and *B*39:06* alleles with type 1 diabetes is restricted to specific *HLA-DR/DQ* haplotypes in Finns. *HLA* **89**, 215–224 (2017).
49. Eike, M. C., Becker, T., Humphreys, K., Olsson, M. & Lie, B. A. Conditional analyses on the T1DGC MHC dataset: novel associations with type 1 diabetes around HLA-G and confirmation of HLA-B. *Genes Immun* **10**, 56–67 (2009).
50. Noble, J. A. & Valdes, A. M. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. *Curr Diab Rep* **11**, 533–542 (2011).

Tables

Table 1. Potentially novel HLA allele associations and modifying effects. Diplotype and haplotype analyses show effects of combinations of two alleles. Here, a strong protective effect on the risk allele can result in non-significant association.

Type of analysis	Phenotype	Primary HLA	Modifying HLA	Discovery			Replication	
				p-value	Beta	SE	Beta	SE
allele	Drug-induced hypoglycaemia without coma	DQB1*03:02		4.14E-05	0.667	0.155	0.917	0.173
		DQA1*03:01		4.83E-05	0.663	0.155	0.926	0.173
	Mental and behavioural disorders due to cannabinoids	DQA1*01:03		7.26E-06	0.603	0.128	0.439	0.127
		DQB1*06:03		1.22E-05	0.59	0.128	0.413	0.129
		DRB1*13:01		3.23E-05	0.54	0.125	0.378	0.125
	Vitreous haemorrhage	DQB1*03:02		1.36E-24	0.701	0.064	0.726	0.079
		DQA1*03:01		1.33E-23	0.688	0.065	0.707	0.079
		DRB1*04:01		4.61E-13	0.552	0.073	0.765	0.085
	Otitis externa	DQB1*03:02		1.39E-05	0.221	0.051	0.212	0.058
		B*18:01		2.46E-05	0.291	0.068	0.341	0.08
		DQA1*03:01		4.32E-05	0.209	0.051	0.208	0.059
	Acute sinusitis	DQA1*03:01		1.97E-07	0.147	0.028	0.145	0.033
		DQB1*03:02		2.66E-07	0.145	0.028	0.142	0.032
		DRB1*04:01		7.41E-06	0.142	0.032	0.134	0.036
	Trigger finger	DQB1*03:02		7.39E-08	0.333	0.061	0.261	0.074
		DQA1*03:01		7.67E-08	0.333	0.061	0.257	0.074
		DRB1*04:01		1.88E-06	0.328	0.068	0.357	0.079
	Scleritis and episcleritis	B*27:05		8.48E-08	0.579	0.102	0.574	0.122
diplotype	Lichen planus	DQB1*05:01	DQB1*05:01 [a]	2.46E-14	1.323	0.166	1.400	0.216
			DQB1*03:01	2.26E-02	0.385	0.165	0.621	0.201
	Seropositive rheumatoid arthritis	DRB1*04:08	DRB1*04:01 [a]	5.84E-35	1.739	0.146	2.490	0.204
			DRB1*08:01	2.50E-01	0.205	0.223	0.226	0.384
	Co-morbidites, CVD and metabolic diseases	DRB1*04:01	DRB1*03:01 [a]	1.27E-76	0.953	0.052	0.877	0.068
			DRB1*15:01	6.61E-01	0.020	0.045	-0.067	0.061
			DRB1*11:01	9.20E-01	0.009	0.089	0.204	0.124
			DRB1*14:54	3.38E-01	-0.167	0.177	-0.026	0.24
	Thyroiditis, ILD-related definition	DQB1*02:01	DQB1*03:02 [a]	4.71E-15	1.203	0.134	0.989	0.172
			DQB1*05:01	8.74E-03	0.387	0.144	-0.112	0.207
haplotype	Type1 diabetes, definitions combined	DQB1*03:02	B*44:27 [a]	7.24E-14	2.223	0.255	2.190	0.309
			B*40:01	6.04E-08	0.885	0.15	1.129	0.165
			B*27:05	3.50E-06	0.823	0.164	0.713	0.214
	Diabetes, kidney failure	DQB1*03:02	B*56:01 [a]	8.34E-19	1.333	0.129	1.240	0.177
			B*27:05	3.86E-03	0.518	0.173	-0.355	0.366
			B*18:01	1.58E-02	0.371	0.149	0.486	0.19
	Diabetic maculopathy	DQB1*03:02	B*56:01 [a]	1.34E-17	1.353	0.136	1.502	0.176
			B*18:01	4.83E-04	0.548	0.15	0.656	0.191
			B*27:05	3.70E-01	0.217	0.215	-0.171	0.366
	ILD Co-morbidites, CVD and metabolic diseases	DRB1*04:01	B*44:27 [a]	1.51E-06	0.695	0.147	0.586	0.18
			B*44:02	8.63E-03	0.100	0.038	0.042	0.052
	Other (seronegative) rheumatoid arthritis, wide	B*27:05	DRB1*04:08 [a]	2.04E-19	1.046	0.102	0.978	0.132
			DRB1*04:04	8.75E-01	0.047	0.191	0.090	0.249

[a] Protective effect was determined relative to this allele combination in diplotype and haplotype analyses.

Figures

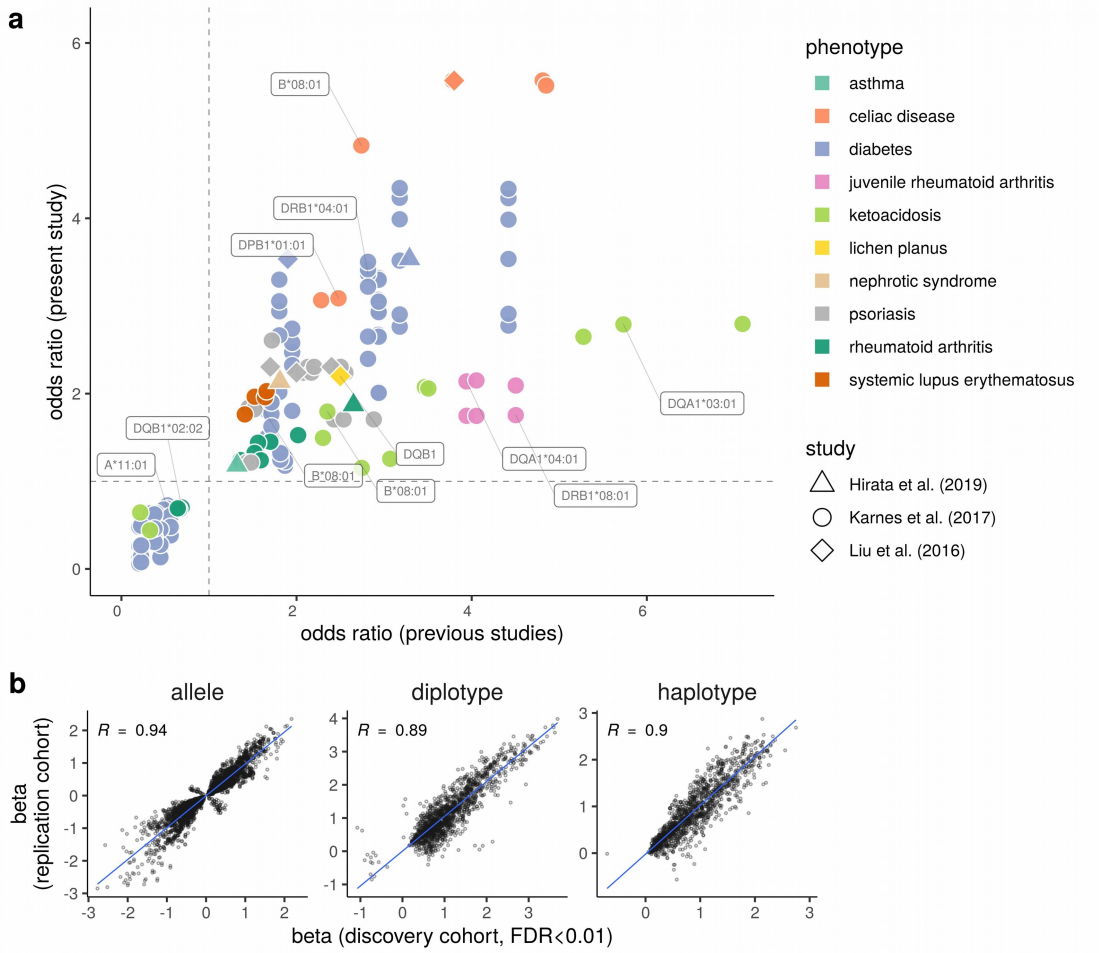

Figure 1. Comparison of HLA association effects. **a)** Odds ratios of previously reported HLA PheWAS associations (x-axis) vs. the discovery cohort of the present study (y-axis). Depending on the study, associations are shown either at the level of four-digit alleles (Karnes *et al*.) or at gene-level 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.

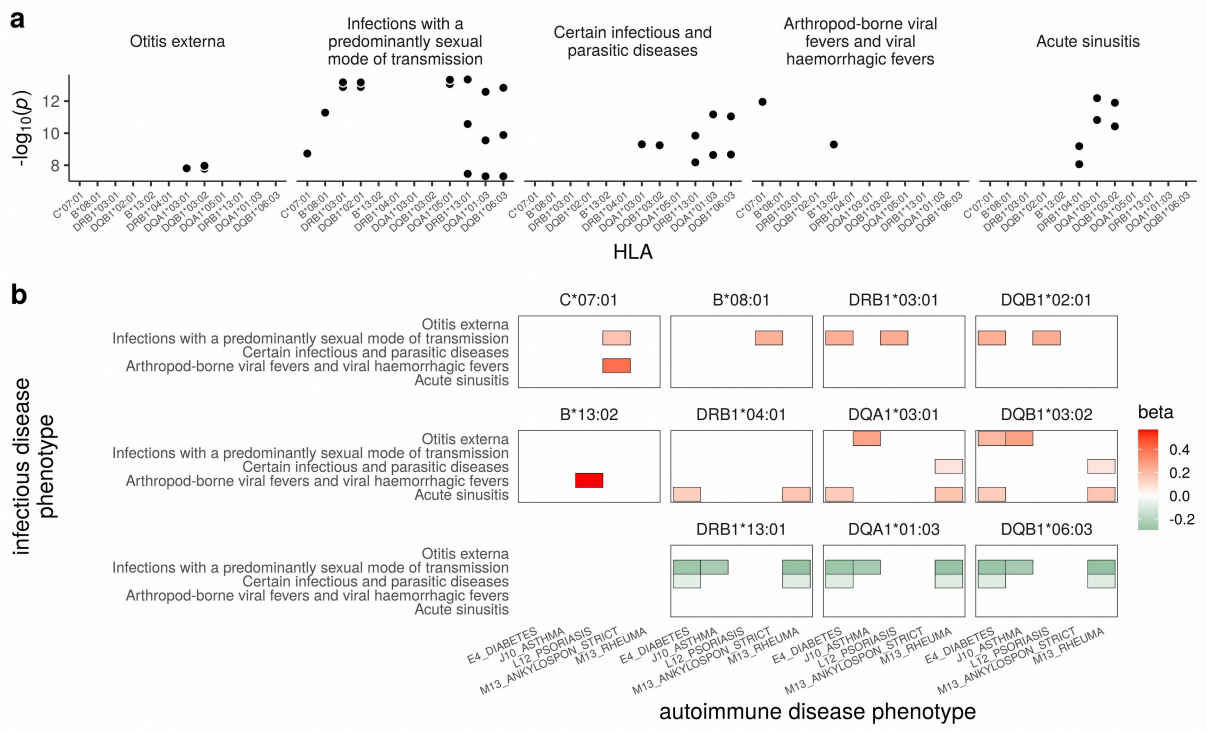

Figure 2. HLA alleles associated with an infectious disease independently of autoimmune disease. **a)** P-values of significant (< 5 x 10^{-8}) alleles. Each panel show an infectious disease phenotype. Within panels, one allele can associate with more than one autoimmune disease. **b)** HLA alleles associated with infectious diseases (y-axis) independently from 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 as its profile is indentical to the other two shown class II alleles. The data are based on conditional regression analyses with p < 5 x 10^{-8} threshold in the full dataset (discovery+replication), where selected phenotypes were analysed by adding a different phenotype as an additional covariate in the model one at a time.

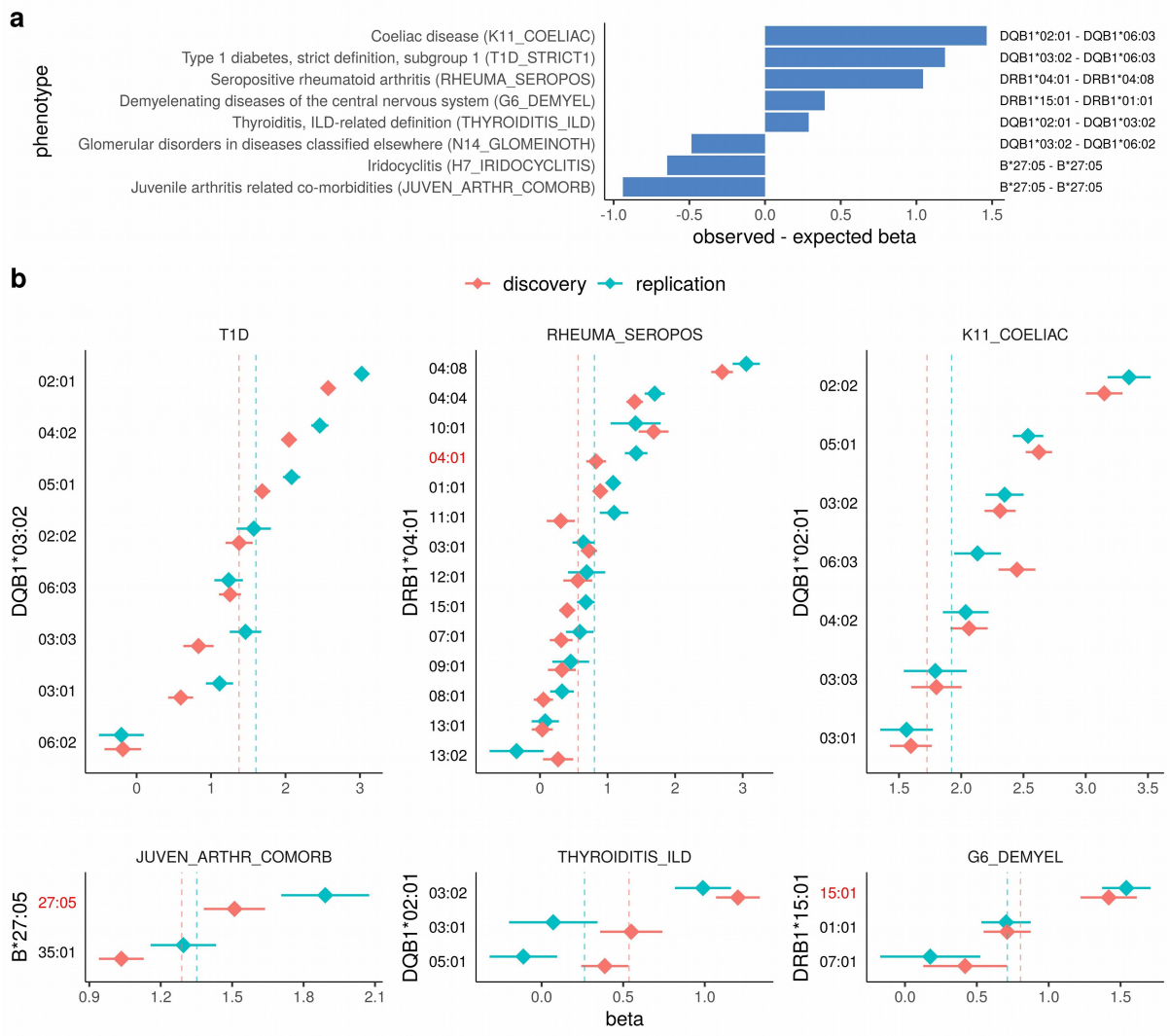

Figure 3. HLA diplotype effects. Risk allele in combination with the second allele of the same HLA gene. **a)** Phenotypes in which the risk allele diplotype association effects deviate from expected (i.e. sum of individual allelic effects). **b)** The x-axis shows log-odds ratios (betas) for different diplotypes 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.

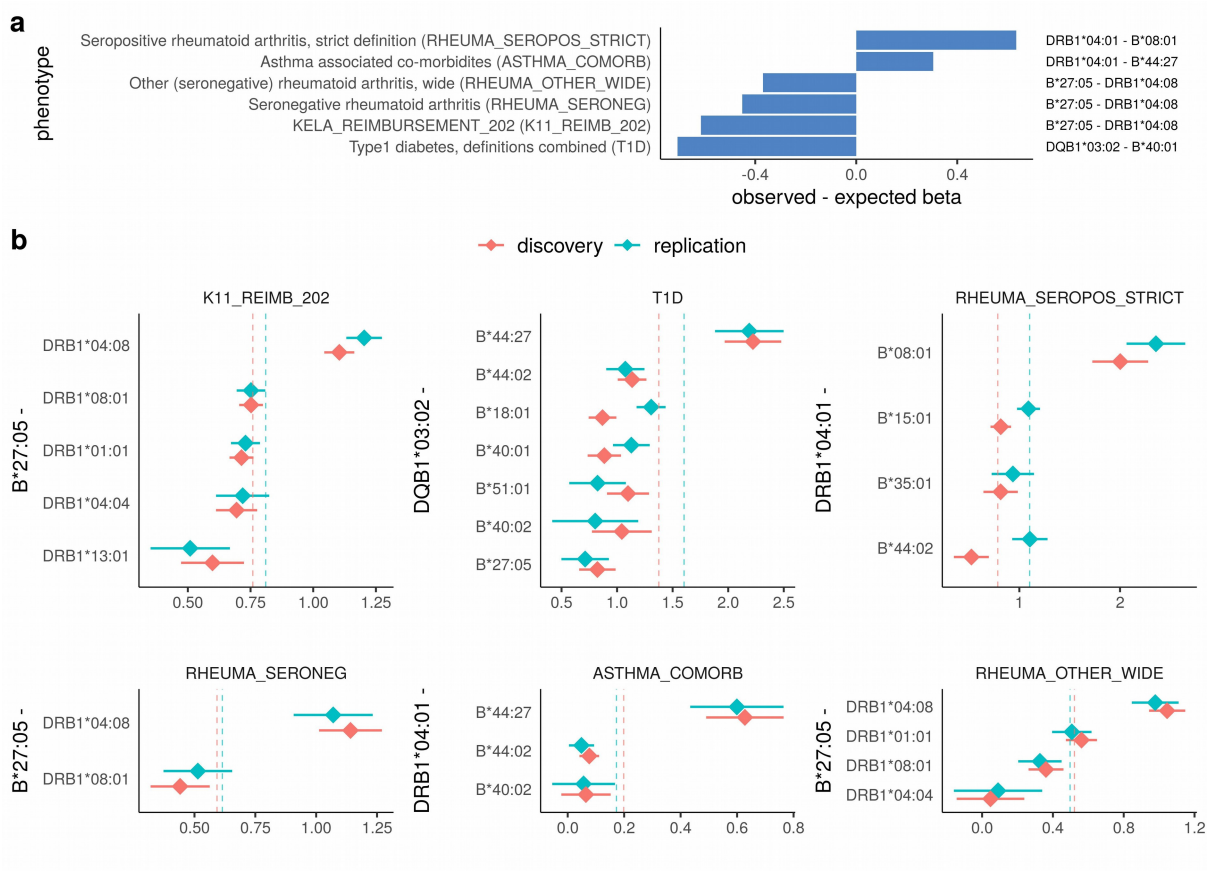

Figure 4. HLA haplotype effects. Risk allele in combination with another allele of a different HLA gene. **a)** Phenotypes in which the risk allele haplotype association effects deviate from expected (i.e. sum of individual allelic effects). **b)** The x-axis shows log-odds ratios (betas) for different 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.