

Haplotype Analysis Reveals Pleiotropic Disease Associations in the HLA Region

Courtney J. Smith^{1,2}, Satu Strausz^{1,2,3,4}, FinnGen, Jeffrey P. Spence¹
Hanna M. Ollila^{2,5,6,7}, Jonathan K. Pritchard^{1,8}

- 1 1. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
- 2 2. Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki,
- 3 Helsinki, Finland
- 4 3. Department of Oral and Maxillofacial Surgery, Helsinki University Hospital and University of
- 5 Helsinki, Helsinki, Finland
- 6 4. Department of Plastic Surgery, Cleft Palate and Craniofacial Center, Helsinki University Hospital
- 7 and University of Helsinki, Helsinki, Finland
- 8 5. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, MA, USA
- 9 6. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
- 10 7. Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Boston, MA,
- 11 USA
- 12 8. Department of Biology, Stanford University, Stanford, CA, USA

13 Address correspondence to: Courtney J. Smith (courtrun@stanford.edu), Satu Strausz
14 (satu.strausz@helsinki.fi), Jeffrey P. Spence (jspence@stanford.edu), Hanna M. Ollila
15 (hanna.m.ollila@helsinki.fi), Jonathan K. Pritchard (pritch@stanford.edu)

16 July 26, 2024.

17 Abstract

18 The human leukocyte antigen (HLA) region plays an important role in human health through in-
19 volvement in immune cell recognition and maturation. While genetic variation in the HLA region is
20 associated with many diseases, the pleiotropic patterns of these associations have not been system-
21 atically investigated. Here, we developed a haplotype approach to investigate disease associations
22 phenome-wide for 412,181 Finnish individuals and 2,459 traits. Across the 1,035 diseases with a
23 GWAS association, we found a 17-fold average per-SNP enrichment of hits in the HLA region.
24 Altogether, we identified 7,649 HLA associations across 647 traits, including 1,750 associations un-
25 covered by haplotype analysis. We find some haplotypes show trade-offs between diseases, while
26 others consistently increase risk across traits, indicating a complex pleiotropic landscape involving
27 a range of diseases. This study highlights the extensive impact of HLA variation on disease risk,
28 and underscores the importance of classical and non-classical genes, as well as non-coding variation.

29 Introduction

30 The major histocompatibility complex (MHC) plays a crucial role in mediating tissue graft com-
31 patibility and immune system recognition of pathogens and self [1–3]. The human MHC, referred
32 to as the human leukocyte antigen (HLA) region, has been found to be associated with numerous
33 diseases [2–8]. The tension between being able to recognize a diverse array of pathogens while
34 avoiding autoimmunity suggests that variants within the HLA region may affect multiple distinct
35 phenotypes simultaneously. Yet, little work has been done to characterize the patterns of pleiotropy
36 and the trade-offs across diseases within the region.

37 The HLA region is approximately 5 megabases in length, and contains hundreds of genes, but
38 is most known for the classical HLA genes, which are involved in response to infection and autoim-
39 munity [9]. The classical HLA genes, which include class I genes (*HLA-A*, *-B*, *-C*) and class II
40 genes (*HLA-DR*, *-DQ*, and *-DP*), encode cell surface proteins that present peptides to immune cells
41 resulting in activation and maturation [10].

42 The classical HLA genes are highly polymorphic, with each gene having multiple distinct alleles.
43 These alleles are functionally diverse: some act as generalists, and others are specific to particular
44 types of peptides [11–13]. Different HLA alleles vary in their ability to recognize certain pathogens,
45 thus genetic variation modulating this ability can result in a variety of disease associations [9, 14].
46 Meanwhile, some pathogens have evolved to avoid common HLA alleles in a host-pathogen arms
47 race [15, 16]. This arms race has resulted in long-term balancing selection at classical HLA genes,
48 leading to trans-species polymorphisms and extreme nucleotide diversity—more than 70-times the
49 genome-wide average [17–19].

50 At the individual level, this genetic variation in the classical HLA genes affects the ability of
51 the immune system to detect pathogens, fight infections, and attack cancerous cells, as well as the
52 ability to limit inappropriate immune responses, such as autoimmune diseases [2–5]. Furthermore,
53 genetic variation in the HLA region can influence the balance between these conflicting goals of
54 pathogen response and the prevention of autoimmunity, resulting in potential risk trade-offs [20–
55 22]. On the other hand, the risk trade-offs between autoimmunity, infection, and other traits can be
56 more complicated, as demonstrated by Epstein-Barr virus (EBV) infection. Chronic EBV infection
57 is known to cause various cancers, including nasopharyngeal carcinoma and Hodgkin lymphoma [23–
58 25], and it has also been shown to play a role in the development of multiple sclerosis, a degenerative
59 demyelinating disease of the central nervous system caused by immune-mediated inflammation [26,
60 27]. Although there is clinical evidence of the complex interplay between infection, autoimmunity,
61 cancer, and other diseases, the genetic contribution to these disease trade-offs and risks has not
62 been well-characterized at the biobank level [2, 20, 21, 23].

63 Association studies have implicated particular HLA alleles in many diseases [6, 7]. These canon-
64 ical HLA association studies have provided countless biologically and clinically informative associ-
65 ations, for example, seronegative spondyloarthritis has been associated with the *HLA-B27* allele
66 family, Type 1 Diabetes with the *HLA-DR3* allele family, and Rheumatoid arthritis with the *HLA-*
67 *DR4* allele family [28, 29]. In addition to providing biological insight into disease mechanisms, these
68 studies have resulted in the use of HLA allele associations in the clinical setting [30–32].

69 While there has been much focus on protein-coding variation within the classical HLA genes,
70 there has been less work characterizing the majority of the genetic variation in the region, which
71 falls outside of the coding regions of the classical HLA genes. Disease-associated variants are
72 typically presumed to be protein-coding, affecting the peptide-binding groove of a classical HLA
73 gene, but variation in regulatory regions may also be a major risk factor in a subset of diseases

74 by influencing gene expression [33–35]. Recent experimental studies have demonstrated that for
75 some traits, regulatory variation in the region confers more risk than HLA coding variation [36].
76 There is also evidence for disease associations with variation in non-HLA genes within the locus,
77 including *C4A* [37], *SLC44A4* [38], and *NOTCH4* [39]. Therefore, investigation of genetic variation
78 throughout the entire HLA region has the potential to reveal additional contributions beyond those
79 found by HLA allele analysis alone.

80 Analyses of the HLA region in genome-wide association studies (GWAS) in large cohorts such
81 as FinnGen [40], UK Biobank [41], and Japan Biobank [42] have identified many trait associations
82 with single nucleotide polymorphisms (SNPs) in the HLA region [43]. These traits span a variety
83 of systems, including infections such as HIV [44] and Hepatitis B [45], and autoimmune conditions
84 ranging from neurological conditions (such as multiple sclerosis [46]), gastrointestinal disorders
85 (such as Celiac disease and inflammatory bowel disease [47]), and rheumatic disorders (such as
86 systemic lupus erythematosus [48]). These studies typically either investigate associations with
87 many traits across the entire genome [8, 39, 49], treating the HLA region as just another locus, or
88 they specifically focus on the HLA region but consider only a small number of traits at a time [50,
89 51]. However, in order to understand how genetic variation in the HLA region contributes to the
90 complicated interplay between different disease risks, it is crucial to study associations for many
91 traits simultaneously. This motivates the need for investigating the role of HLA loci in modulating
92 trade-offs in these disease associations at the phenome-wide scale.

93 In this study, we quantified how genetic variation and pleiotropy at the HLA region contribute to
94 disease risk across a broad range of diseases. We analyzed data from 412,181 Finnish individuals for
95 2,459 traits. We focused on understanding the spatial distribution of disease associations throughout
96 the HLA region and the nature of pleiotropy between different traits. We developed a haplotype-
97 based approach to robustly characterize patterns of disease associations throughout the entire HLA
98 region, including non-coding variation and variation outside of classical HLA genes. We applied our
99 approach at a phenome-wide scale and evaluated the role of HLA in modulating risk and trade-offs
100 across a broad range of diseases in the context of the full complexity and breadth of HLA genetic
101 variation.

102 Results

103 Enrichment of significant trait associations in the HLA region

104 To identify disease associations with genetic loci throughout the entire HLA region, we analyzed
 105 data from 412,181 Finnish individuals and 2,459 traits (Figure 1). We used fine-mapped GWAS
 106 summary statistics released by FinnGen, as well as new association data we generated at the level of
 107 individual phased haplotypes and HLA alleles. We corrected for sex, age, and the first ten principal
 108 components of the genome-wide genotype matrix (see Methods). Results from these association
 109 tests were used in subsequent analyses.

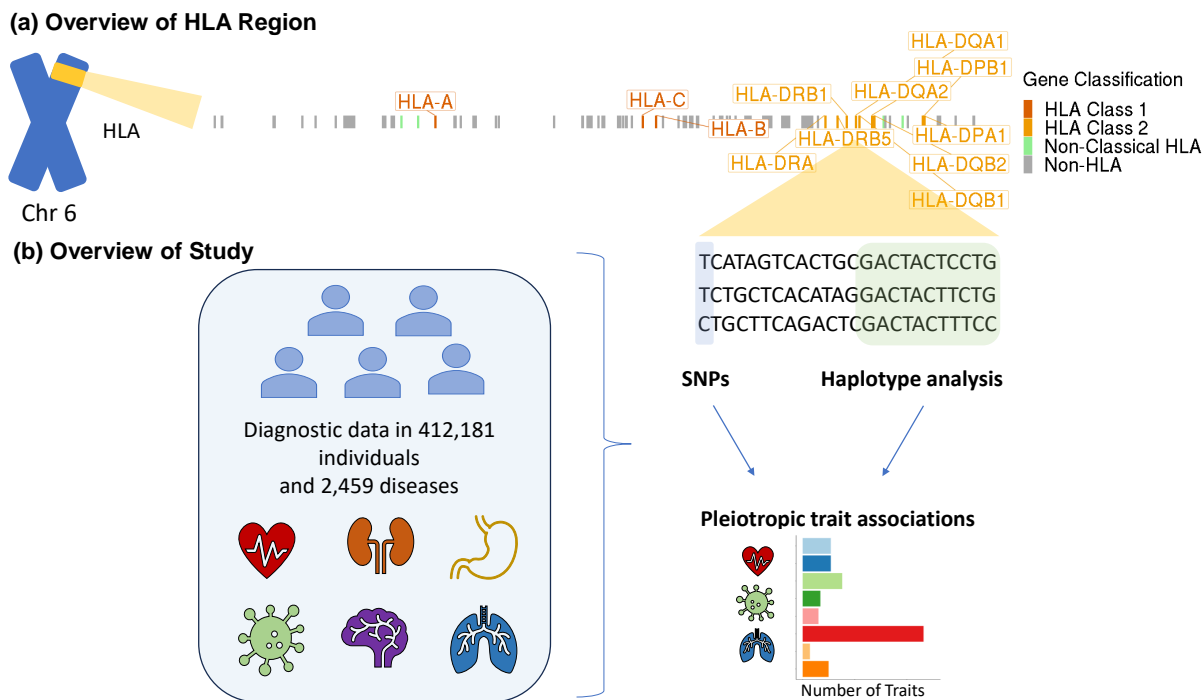


Figure 1: Study Overview. *A.* An overview of the HLA region showing the nearest genes to trait-associated SNPs, colored by HLA class, spanning approximately 5 megabases. *B.* An overview of the study data and design.

110 While the importance of HLA variation in disease has been well-established, we first sought to
 111 systematically quantify the enrichment of association signals across diseases, focusing on how en-
 112 richment varies by disease type. We considered the 1,035 disease traits in FinnGen that had at least
 113 one genome-wide significant association anywhere in the genome. We then identified independent
 114 genome-wide significant SNP associations for each trait, and binned these SNPs into 100 kb bins
 115 (Figure 2a). We found the mean number of significant associations per bin was 2.75, with a median
 116 of 1. One of the bins on chromosome 6 that overlaps the class II region of the HLA region had
 117 the highest number of associations in a single bin with 282 associations. Five of the six bins with
 118 the most associations overlapped the HLA region. The remaining bin is on chromosome 19 and
 119 has 101 associations. This bin contains an apolipoprotein gene cluster including *APOE*, *APOC1*,
 120 *APOC2*, *APOC4*, which are involved in lipid metabolism and affect Alzheimer's disease risk. These
 121 results show that the HLA region harbors a higher density of disease associations than the rest of
 122 the genome.

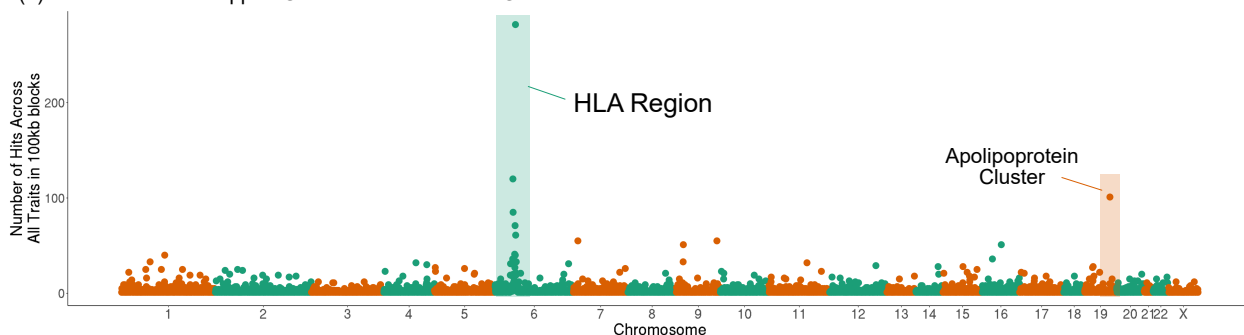
123 While the role of HLA in infectious disease and autoimmunity is well-established, its role in
124 other disease types is less clear. As such, we sought to quantify the enrichment of association signal
125 stratified by disease groups. We classified the 1,035 diseases that had at least one GWAS association
126 into 45 trait categories based on ICD codes. We then calculated the average per SNP enrichment of
127 association signals for each disease category by comparing the number of independent associations
128 inside the HLA region to the number in the rest of the genome (Supplementary Table 1).

129 Overall, we found a 17x enrichment in the HLA region relative to the rest of the genome averaged
130 across all 1,035 diseases that had at least one GWAS association anywhere in the genome. The
131 individual disease category with the highest enrichment was the Infectious trait group, with a 396x
132 enrichment relative to the rest of the genome (Figure 2b). The overall enrichment across all diseases
133 remained relatively unchanged (16.6x) even after excluding all infectious traits. In addition, the
134 majority of other trait groups, including groups such as Dental traits (71x), Dermatologic traits
135 (63x enriched), Rheumatic traits (53x enriched), Hematologic (50x enriched), and Ear traits (45x
136 enriched) also showed a major enrichment in the HLA region. In contrast, the Congenital group was
137 the only group not enriched in the HLA region. This could be because the traits in the Congenital
138 group are oligogenic, with an average of 2.2 hits outside the HLA region and none within the HLA
139 locus. The most enriched trait groups showed enrichment for primarily two reasons (Supplementary
140 Figure S1). First, some traits had high enrichment because they had many associations across the
141 genome, with proportionately even more associations in the HLA region, such as the Rheumatic
142 traits. Alternatively a subset of the enriched traits did not have many associations overall, but the
143 few associations they had were in the HLA region, such as the Infectious traits.

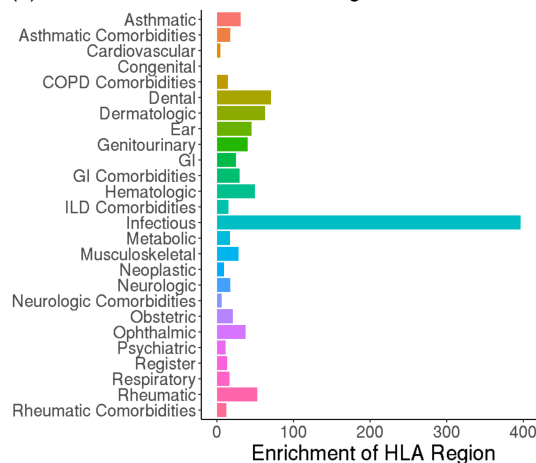
144 To ensure that our results were robust and not driven by the unusually high gene density or by
145 differences in genotype array coverage of the HLA region, we repeated our analyses to identify per-
146 gene and per-base pair enrichments. The results were qualitatively consistent, differing by factors of
147 0.48x and 2.4x respectively. Overall, these results emphasize the involvement of the HLA region in
148 a broad range of disease groups, including those from a variety of different pathologic mechanisms
149 and organ systems.

150 In order to understand how the HLA region contributes to disease mechanisms, we next examined
151 traits that had associations within the locus ($N = 572$ diseases). To remove essentially redundant
152 traits, we focused on the subset of these traits that had LDSC genetic correlation ≤ 0.95 . This
153 included 269 diseases and 3 continuous traits (height, weight, body mass index). We then used
154 forward stepwise regression to identify conditionally independent SNP associations for each trait.
155 This resulted in 428 associations ($MAF > 1\%$, $P < 10^{-6}$) across all traits.

(a) Number of Fine-mapped GWAS Hits Across FinnGen Traits



(b) HLA Enrichment Relative to Background



(c) Overview of Hits by Traits

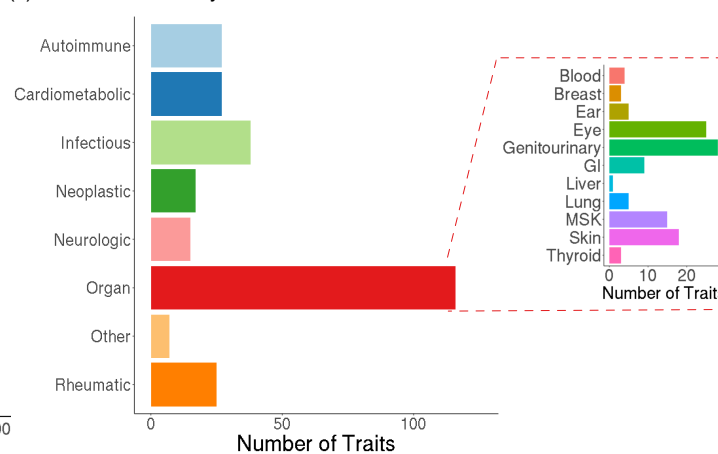


Figure 2: Distribution of GWAS hits across the genome and trait group enrichment. A. Distribution of fine-mapped GWAS hits throughout the genome across 1,035 FinnGen disease traits, binned into 100 kb bins. **B.** Enrichment of association signal in the HLA region by disease group. The 1,035 diseases were categorized into 45 disease groups based on ICD codes and the average per SNP enrichment in the HLA region was calculated by comparing the number of independent associations in the HLA region relative to that in the rest of the genome. **C.** Classification of traits with at least one significant association in the HLA region by shared pathophysiology.

156 Classifying disease categories by ICD code, as was done in the enrichment analysis above, pri-
 157 marily results in anatomical groups as opposed to groups based on shared pathophysiology. To
 158 understand the contribution of HLA to biological disease mechanisms, we manually classified the
 159 269 HLA-associated diseases based on pathophysiology (Figure 2c, Supplementary Table 2). For
 160 traits where the underlying mechanism is unknown or ambiguous, we classified by the organ system
 161 affected.

162 We calculated the number of traits in each of these trait categories that had at least one sig-
 163 nificant HLA association in the HLA region (Figure 2c). Two of the top disease categories were
 164 Rheumatic (40 traits) and Infectious (38 traits). In contrast to the enrichment analysis, addi-
 165 tional multi-system disease groups beyond Rheumatic and Infectious traits were well-represented,
 166 including Autoimmune (27 traits) and Cardiometabolic (27 traits).

167 Pleiotropy and spatial structure of significant SNP association signal within the 168 HLA region

169 We aimed to evaluate the spatial distribution of the significant SNP association signal across the
170 HLA region. We first categorized the associations by assigning each variant to its nearest gene
171 (Figure 3a). We observed association signals throughout the extended HLA region with the highest
172 density of associations near the twelve classical HLA genes, particularly the class II genes. However,
173 associations were spread broadly across the region, with a total of 75 genes that were the nearest
174 gene for at least one association, 59 of which were non-HLA genes. Overall, the associations were
175 spread relatively consistently across trait groups, although the autoimmune and rheumatic traits had
176 slightly higher signal near the class I genes than the other trait groups did, likely driven at least in
177 part by the well-known associations of *HLA-B* alleles with rheumatic traits [52, 53] (Supplementary
178 Figure S2).

179 We next evaluated the role of genetic variation in the HLA region in modulating disease risk
180 trade-offs. We calculated normalized Z-scores for each association discovered in the forward stepwise
181 analysis ($\text{sign}(Z) * Z^2 / (\max Z^2 \text{ of trait})$; See Methods), and visualized how these association signals
182 were spread across the locus (Figure 3b). We found that 99% of the associations were also significant
183 ($P < 10^{-6}$) for one or more diseases beyond the trait for which they were identified as a conditionally
184 independent significant association. Moreover, we found variants that significantly increased the
185 risk for one disease while significantly decreasing risk for another disease suggesting a possible risk
186 trade-off between traits.

187 The normalized Z-scores visually clustered around three main genomic regions within the HLA
188 locus. The first cluster spanned two non-classical and one class I HLA gene (*HLA-F*, *HLA-G*, *HLA-*
189 *A*). The second spanned two class I HLA genes and one non-HLA gene (*HLA-C*, *HLA-B*, *MICA*).
190 The third spanned one non-HLA gene and two sets of class II HLA genes (*NOTCH4*, *HLA-DR*,
191 *HLA-DQ*).

192 The overall pleiotropic structure revealed large blocks of SNPs spanning hundreds of kilobases
193 that have similar effects across traits. These encompass multiple genes, and likely arise due to the
194 high gene density and the extensive linkage disequilibrium (LD) in the region (Figure 3; Supple-
195 mentary Figure S3).

196 Pleiotropic disease associations at the haplotype level

197 The HLA region is particularly challenging for standard association studies because of its strong LD,
198 multiallelic sites, and large effect coding variants within the classical HLA genes. Motivated by the
199 block-like structure of the HLA locus (Figure 3b), we developed an approach to explore pleiotropy
200 at the haplotype level, with haplotype blocks spanning multiple genes and including non-classical
201 HLA, non-HLA, and non-coding regions.

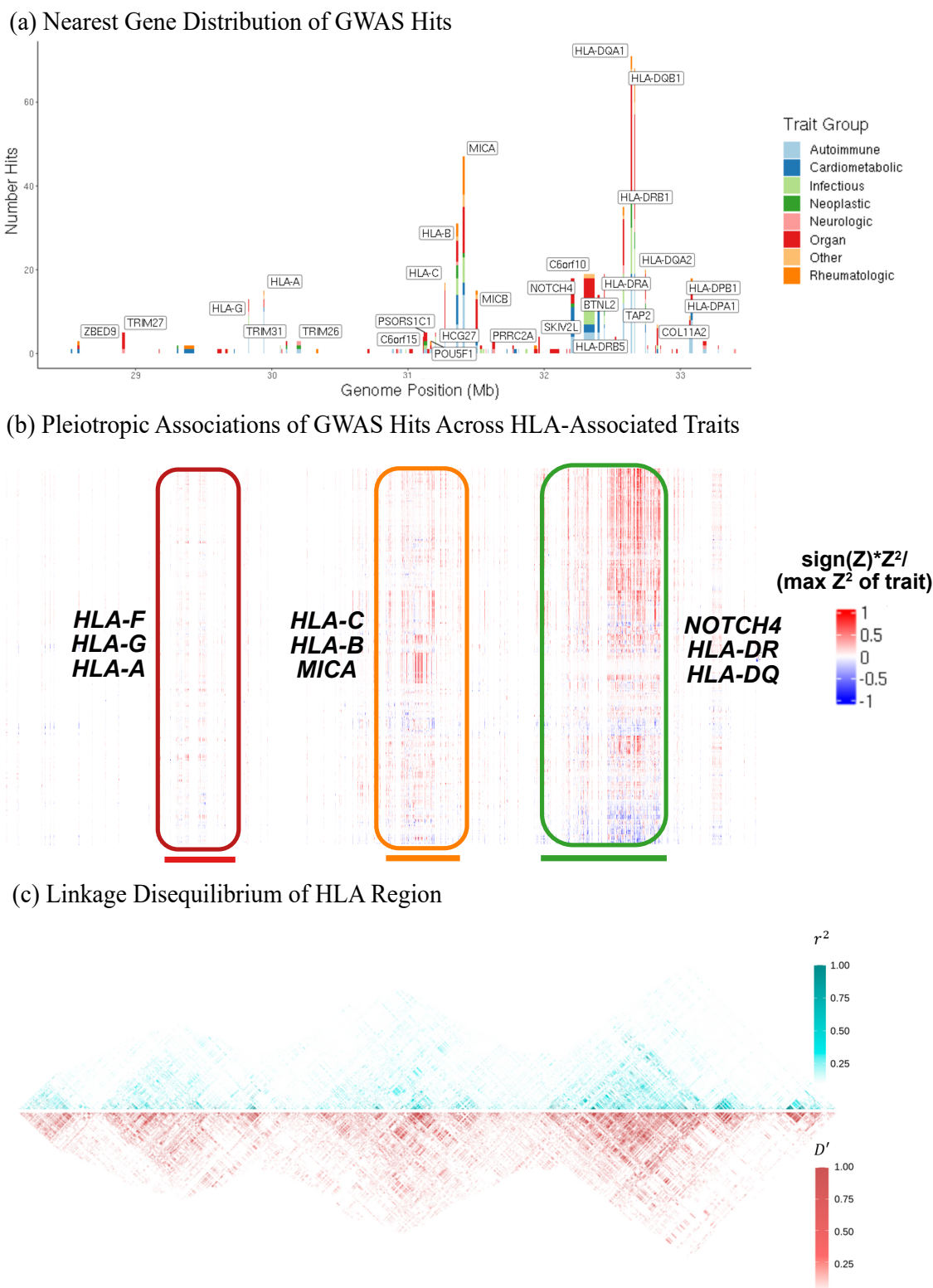


Figure 3: Pleiotropic structure of the HLA region. **A.** Distribution of significant SNP associations across the HLA region, binned by nearest gene. Each bar represents a different gene and the width corresponds to the length of the gene boundaries. **B.** Heatmap of normalized Z-scores for the 428 variants in the HLA region significantly associated with at least one trait. The x-axis corresponds to the genome position of the variant, the y-axis corresponds to the HLA-associated traits. Associations with all HLA-associated traits are shown for all variants that had an independent significant association with at least one trait. The three blocks used in subsequent analysis are circled, underlined, and labeled by well-known genes within each block. **C.** Linkage disequilibrium as measured by r^2 and D' of the approximately 40,000 SNPs covering the HLA region ($MAF > 1\%$).

202 The three main regions ("blocks") described above were selected based on the density of signal
 203 from the significant SNP associations, overlapping LD patterns, and functional relevance. We
 204 defined haplotypes for each of the three regions by the unique combination of phased nucleotides
 205 at 1,000 randomly selected biallelic SNPs with MAF > 1% (Figure 4; see Methods for additional
 206 details). We then clustered related haplotypes into groups (Supplementary Table 3; Supplementary
 207 Information 1), and for each block performed association analyses between the haplotype groups and
 208 the 269 HLA-associated diseases. We discovered 469 significant trait-haplotype group associations
 209 ($|Z| > 4$) across blocks (Figure 5; Supplementary Table 4), representing 64 traits. Of these traits,
 210 25 had significant associations with all three blocks. Celiac disease had the most trait-haplotype
 211 group associations with 36 total (8 in Block 1, 16 in Block 2, 12 in Block 3), followed by rheumatic
 212 disease prescriptions with 34, spondylopathies with 32, and iridocyclitis and type 1 diabetes with
 213 25 each.

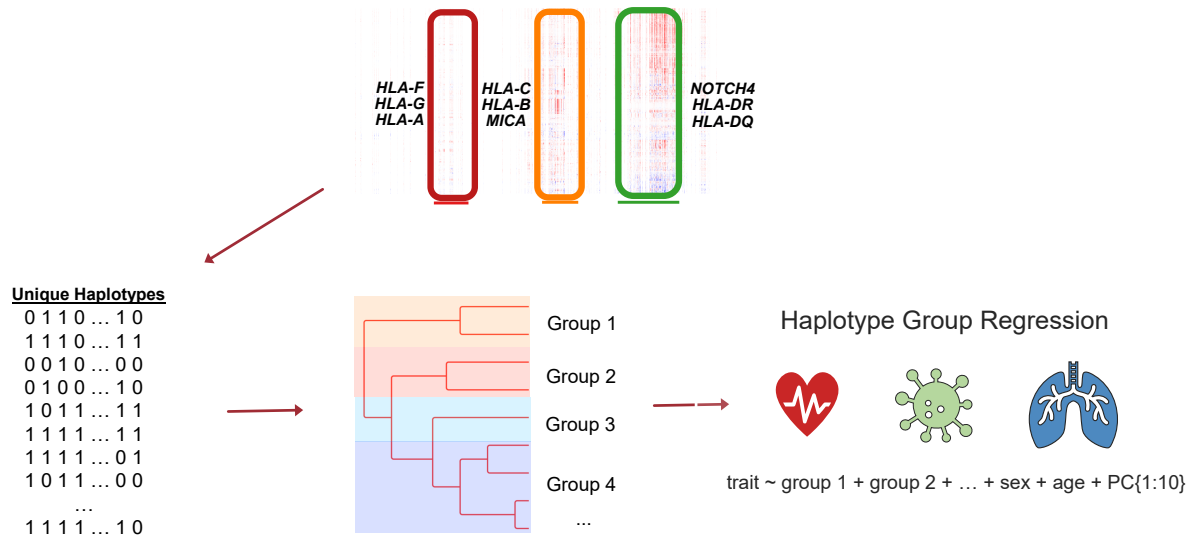


Figure 4: Haplotype group regression analysis pipeline. Overview of the pipeline for identifying the haplotype groups for each of the three blocks in the HLA region and performing trait associations. For each block, all unique phased combinations of nucleotides at 1,000 randomly selected SNPs were considered as haplotypes. We then clustered related haplotypes into groups by recursively splitting the dendrogram at each branch point (see Methods). Finally, for each of the three blocks, we performed association analyses between the haplotype groups and the 269 HLA-associated diseases, including all haplotype groups for a given block except the most frequent in each regression, as well as sex, age, and the first ten principal components of the genome-wide genotype matrix as covariates.

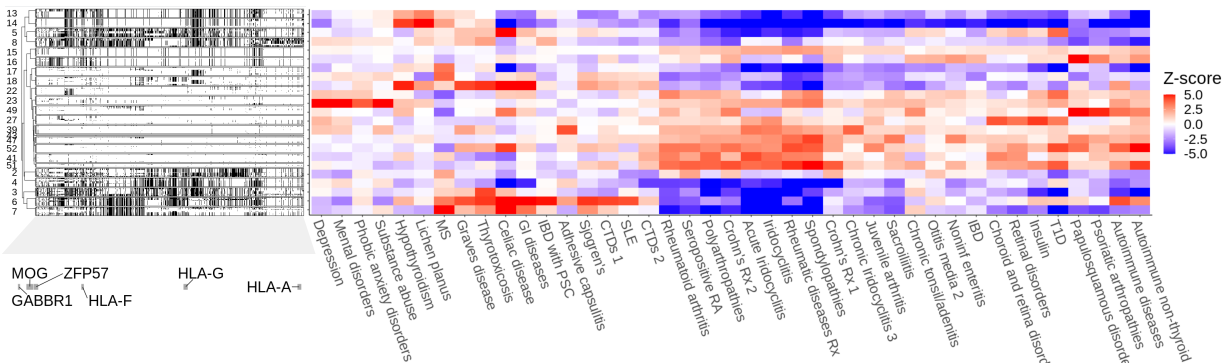
214 We sought to explore the patterns of pleiotropy within these blocks. For each block, we consid-
 215 ered all traits with at least one association ($|Z| > 4$) in that block, and all haplotype groups with at
 216 least one trait association or total copies greater than the minimum cutoff of 20,000 copies (Figure
 217 5). This resulted in 41 traits and 23 haplotype groups for Block 1, 46 traits and 25 haplotype groups
 218 for Block 2, and 36 traits and 21 haplotype groups for Block 3.

219 The majority of the haplotype groups were significantly associated with multiple traits. A subset
 220 of haplotype groups were associated with increased risk for some diseases, but decreased risk for oth-
 221 ers, consistent with disease risk trade-offs. For example, in Block 1, haplotype group 6 is associated
 222 with increased risk ($Z > 3$) for 10 traits, including GI autoimmune disorders, thyroid conditions,
 223 and connective tissue and rheumatic disorders. However, this haplotype group is also associated
 224 with decreased risk for 8 traits, mostly other rheumatic and inflammatory traits (Supplementary

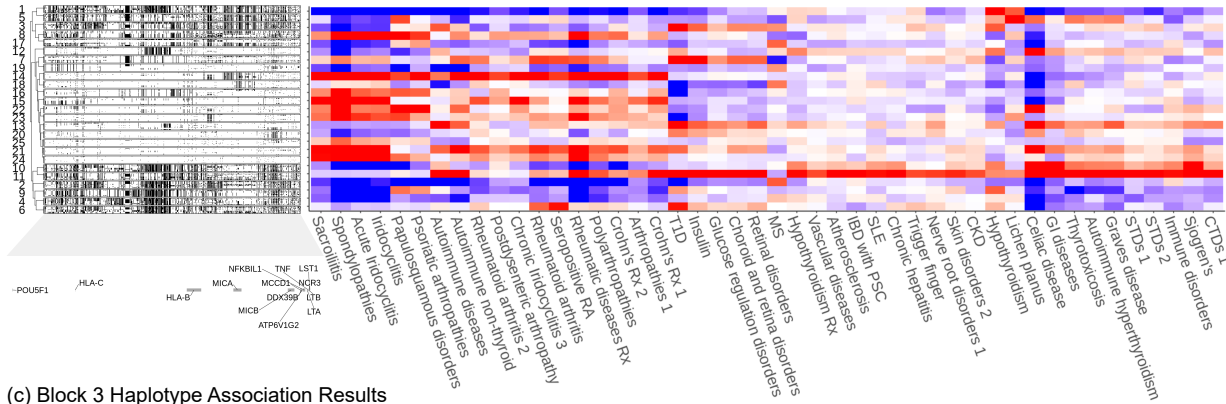
225 Table 4). Overall, of the 58 haplotype groups that showed a significant disease association ($|Z| > 4$),
226 the mean number of associations ($|Z| > 3$) per haplotype group was 5 risk-increasing associations,
227 and 7 risk decreasing associations (Supplementary Figure S4; Supplementary Information 2).

228 In contrast to these haplotype groups showing disease risk trade-offs, we also observed that some
229 haplotype groups had the same direction of effect across the majority of associated traits (Figure
230 5). For example, haplotype group 49 in Block 1 was one of the rarest haplotype groups (0.09%
231 frequency), but all 6 of the diseases with which it was significantly ($|Z| > 3$) associated were in the
232 risk increasing direction, including depression and phobic anxiety disorders. This finding motivated
233 us to calculate overall disease burden proportions for each haplotype group (Supplementary Fig-
234 ure S5). We defined the set of relevant diseases for each block as any disease that was significantly
235 associated with at least one of the haplotype groups in that block. Then for each haplotype group
236 in a given block, we identified the proportion of individuals in the haplotype group that had a
237 diagnosis of at least one of the block's relevant diseases. To identify the overall disease proportion
238 as a baseline comparison, for each block we identified the proportion of all 412,181 individuals that
239 had a diagnosis of at least one of the block's relevant diseases. We then compared the haplotype
240 group disease proportion to the overall disease proportion (Supplementary Figure S6). For example,
241 compared to the baseline prevalence in FinnGen of 67.5%, we found that haplotype group 49 in
242 Block 1 had one of the highest block-relevant disease burdens with 73% of carriers having at least
243 one of the block's significantly associated ($|Z| > 4$) diseases ($P = 0.001$). Our findings indicate that
244 while some haplotypes had trade-offs in which diseases they increased and decreased the risk of,
245 other haplotypes had an overall net positive or net negative impact across traits.

(a) Block 1 Haplotype Association Results



(b) Block 2 Haplotype Association Results



(c) Block 3 Haplotype Association Results

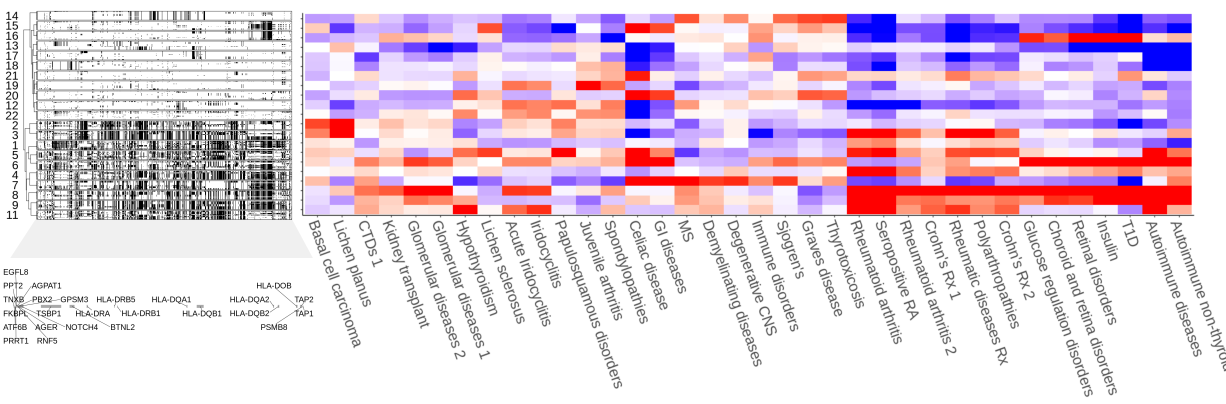


Figure 5: Haplotype group regression results. A dendrogram showing the clustering of the 40 most frequent haplotypes per haplotype group, with white representing the reference allele and black representing the effect allele. Genes are labeled below the corresponding SNPs overlapping their genome position, indicating which are within gene boundaries and which are intergenic. Heatmap showing the Z-scores from the haplotype group regression analysis across associated traits for **A. Block 1**, **B. Block 2**, and **C. Block 3**, including all traits with at least one association $|Z| > 4$ in that block, and all haplotype groups with at least one trait association or total copies greater than the minimum cutoff of 20,000 copies. For visualization purposes, traits are clustered and Z-scores were set to a maximum of $|Z|$ of 5.

246 **Comparison of effects on trait pairs across haplotype groups**

247 Many diseases have shared underlying pathology resulting in comorbidity. As a result, we expected
 248 to see sharing of associations across these diseases for the HLA haplotype groups. Indeed, our anal-
 249 ysis recapitulated shared pathology for many traits, such as rheumatoid arthritis and seropositive
 250 rheumatoid arthritis, with similar associations across haplotype groups. More broadly we found that

251 the inflammatory and rheumatic traits, such as spondylopathies, iridocyclitis, polyarthropathies,
 252 and rheumatoid arthritis clustered together throughout the three blocks (Figure 5). This could
 253 result from phenotypic correlations, caused, for example, by being co-morbid. An alternative ex-
 254 planation is that these traits have a shared biological mechanism modulated by genetic variation
 255 in the HLA region. Finally, it is possible that these correlations are an artifact of long-range LD
 256 extending beyond the haplotypes.

257 In contrast, we observed a surprising lack of concordance for a subset of seemingly similar traits,
 258 such as IBD and "IBD with primary sclerosing cholangitis" (IBD with PSC) (Figure 5). IBD with
 259 PSC is an idiopathic chronic liver disease complication developed by a subset of IBD patients, in
 260 which the bile ducts become inflamed and scarred, causing liver damage. IBD and IBD with PSC
 261 have a genome-wide genetic correlation of 0.45 and have similar effects across haplotypes in Block 3
 262 (Pearson's correlation of 0.57, SE = 0.12, P = 0.005), suggesting a shared etiology (Supplementary
 263 Figure S7). However, the haplotype groups have essentially uncorrelated effects on the two diseases
 264 in Block 1 (Pearson's correlation of 0.10, SE = 0.13, P = 0.47). In fact, some haplotype groups
 265 in Block 1, such as group 6, are associated with increased risk for IBD with PSC, but not IBD
 266 (Supplementary Figure S7).

267 The difference in haplotype group effects on IBD and IBD with PSC is particularly interesting
 268 because it is difficult for clinicians to predict which IBD patients will develop liver damage and
 269 the mechanism leading to this damage is unknown [54]. Thus, understanding which parts of the
 270 genome are associated with increased risk for both a disease and its complications—as opposed to
 271 loci that differentially affect a disease and its complications—may help us better understand the
 272 factors that modulate the risk of certain disease complications. Understanding these differences
 273 may help explain why individuals with the same disease can present with a wide range of symptoms
 274 and outcomes.

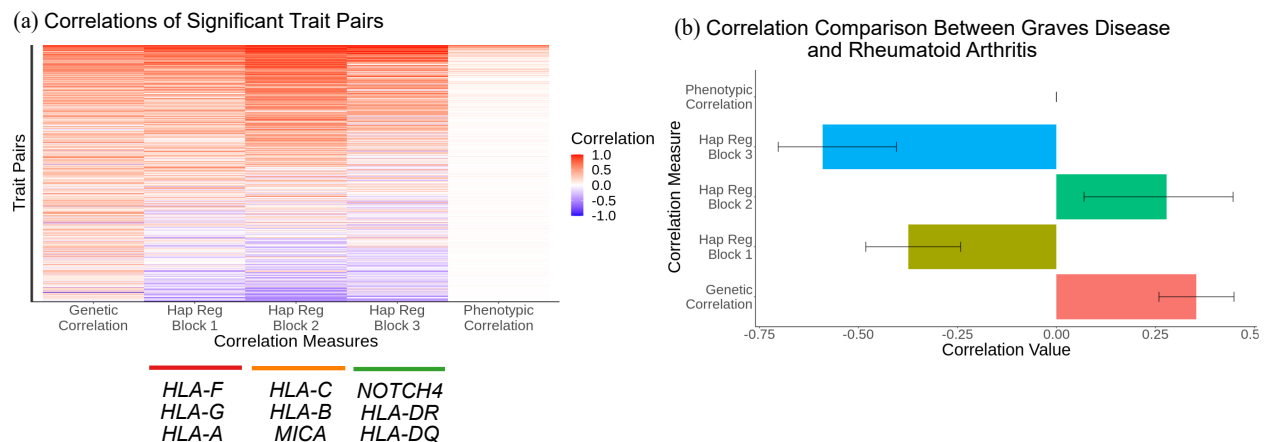


Figure 6: Correlation of haplotype associated traits. *A.* Overview and comparison of the pairwise relationships between traits that were significantly associated with the haplotype group regression analysis, comparing genome-wide LDSC genetic correlations, Pearson's correlation across haplotype groups in each block, and phenotypic correlations. *B.* Comparison of correlation measures between Graves Disease and Rheumatoid Arthritis.

275 To better disentangle whether these pleiotropic associations were due to LD, comorbidity, or
 276 shared biological pathways, we quantified the genome-wide LDSC genetic correlation, phenotypic
 277 correlation, and Pearson's correlation across haplotype group trait associations for all pairwise com-
 278 binations of haplotype group-associated traits for each block (Figure 6a). We discovered 1,520 pairs
 279 of traits with genome-wide genetic correlations greater than 0.3 where both traits are significantly

280 associated ($|Z| > 4$) with at least one block. Of these trait pairs, 408 have a correlation across
281 haplotype group effects > 0.3 for all three blocks, and surprisingly 256 had a discordant correlation
282 of less than -0.3 in at least one block. We also observed discordant association signals for diseases
283 with previously well-defined genetic associations and with clinical impact [52, 55, 56], such as Graves
284 Disease and Rheumatoid Arthritis (Figure 6b).

285 Graves disease is a condition where autoantibodies against the TSH receptor lead to overstimula-
286 tion of the thyroid gland resulting in hyperthyroidism. Rheumatoid Arthritis is an idiopathic chronic
287 inflammatory autoimmune disorder, primarily affecting the joints. Graves Disease and Rheumatoid
288 Arthritis have a genome-wide genetic correlation of 0.35 ($P = 0.0002$), despite a phenotypic correla-
289 tion of approximately 0 ($P = 0.6$). The correlation of effects within Block 2 is concordant—although
290 not significantly so (Pearson’s correlation of 0.28, $P = 0.18$)—with this genome-wide genetic cor-
291 relation. However, the effects in Blocks 1 and 3 are significantly negatively correlated (Pearson’s
292 correlations of -0.37 and -0.59 , $P = 0.006$ and 0.004 respectively). A potential explanation of this
293 discordance between the genome-wide genetic correlation and the correlation within HLA regions is
294 that these discordant regions affect a biochemical mechanism that breaks shared pathology, resulting
295 in an increased risk in one trait while decreasing risk in another, when relevant variants elsewhere
296 in the genome typically cause a shared increase or decrease risk in both traits. In previous work, we
297 showed that such mechanisms can result in associations with opposite signs on the traits, in spite
298 of a positive genome-wide genetic correlation, driven by variants acting at the shared biochemical
299 pathways between both diseases [57].

300 Evaluation of haplotype group signal independent of HLA alleles

301 While protein-coding variation within the HLA genes likely contributes significantly to the disease
302 associations at the haplotype level, a feature of the haplotype analysis is that it includes genetic
303 variation beyond coding variants in classical HLA genes, including non-classical HLA genes, non-
304 HLA genes, and non-coding variation. Therefore, we sought to determine if the haplotype analysis
305 was able to capture signal beyond the HLA alleles. To be conservative, we only considered signal
306 entirely independent (directly, or indirectly due to LD) of the HLA alleles by performing the hap-
307 lotype group regressions while including all classical HLA alleles (frequency $> 1\%$) in each block as
308 covariates. Overall, we found that 129 haplotype associations remained significant ($|Z| > 4$) after
309 accounting for HLA allelic variation (Supplementary Figure S8; Supplementary Table 4), particu-
310 larly for Block 1. Specifically, Block 1 had 171 significant associations across 48 unique traits in our
311 original analysis, and 50 significant associations ($|Z| > 4$) across 18 unique traits after adjusting for
312 the alleles.

313 This indicates that many associations cannot be explained by HLA allele variation or signal
314 tagged by it, and demonstrates that the haplotype group analysis was able to pick up on disease
315 associations that would have been missed in traditional allele association analysis. Block 1 over-
316 lapped only one classical HLA gene, *HLA-A*, suggesting that our haplotype regression approach
317 may be particularly beneficial for regions of the HLA that cover non-classical HLA genes. More-
318 over, including the HLA alleles as covariates *increased* the strength of 42 significant haplotype-trait
319 associations, indicating that the haplotypes explain some variation independent of that explained
320 by the HLA alleles.

321 To further disentangle the information provided by haplotypes, SNPs, and HLA alleles, we
322 performed association analyses at each of these levels separately. For the allele associations, we
323 performed regressions using two approaches. The first approach used the standard method of

324 including one allele per regression, while the second performed a multivariable regression of all
325 alleles (variance inflation factor < 5) within a given block. The results of our association analyses
326 at the haplotype, SNP, and HLA allele level on the full cohort across all 2,459 traits are available in
327 Supplementary Tables 4-6. In total, we identified 7,649 associations and 647 HLA-associated traits
328 across the combined association analyses. In particular, we identified 1,750 significant associations
329 within the HLA locus in the haplotype analysis, including 27 traits not identified in the SNP or
330 HLA allele analyses. These traits included non-organic psychotic disorders, otorrhagia, vascular
331 dementia, and rectal cancers. This emphasizes that analyzing variation at the haplotype level
332 provides orthogonal information about the role of the HLA region in disease.

333 Discussion

334 In this work, we investigated how genetic variation throughout the HLA region associates with
335 disease with a focus on broad pleiotropic patterns. We quantified the enrichment of association
336 signal in the HLA region relative to the rest of the genome. We found a strong enrichment of disease
337 associations across a broad range of disease groups and organ systems. Unsurprisingly, infectious
338 traits were almost 400-fold enriched in the HLA region compared to the rest of the genome, in
339 spite of infections making up a minority of the HLA-associated traits. We also found enrichment
340 across multiple disease categories and organ systems including cardiovascular and neuropsychiatric
341 diseases. Overall, these findings indicate HLA is a major locus for disease risk, not only for infectious
342 diseases, but for diseases across many organ systems and etiologies.

343 Even with the extreme enrichment for infection-related associations, we expect that there is still
344 substantially more information to be gleaned about the role of HLA in mediating infection. Our
345 enrichment analysis controls for how well-powered a trait is by using the number of associations in
346 the rest of the genome as a baseline. However, while we find a huge enrichment, the absolute number
347 of total associations is small. Infectious traits are often under-reported in large biobank cohorts:
348 identifying cases requires patients to seek care for the infection, followed by testing to confirm
349 the specific pathogen. The infectious traits that we identified with the clearest signal tended to
350 be those with more consistent reporting such as sexually transmitted infections. Therefore, our
351 findings indicate that there is likely more signal for infectious traits that will be discovered with
352 larger samples or more systematic reporting.

353 We performed disease association testing with SNPs, HLA alleles, and haplotypes to capture
354 disease associations throughout the entire HLA region, including non-classical HLA genes and non-
355 coding regions. We developed a haplotype analysis approach that includes genetic variation outside
356 of the classical HLA alleles. While many diseases strongly associate with canonical HLA alleles,
357 the HLA region harbours hundreds of genes, many of which also play an important role in immune
358 response and other biological processes. Our haplotype approach discovered disease associations in
359 the HLA region that remained after adjusting for classical HLA alleles, particularly in the region
360 that overlaps more non-HLA and non-classical HLA genes.

361 Furthermore, we found some haplotype groups that displayed disease risk trade-offs, being pro-
362 tective for some diseases and risk-increasing for others. Meanwhile, we found some haplotype
363 groups that were more consistently associated with increased disease burden across tested diseases.
364 In addition, our haplotype analysis discovered that local genetic correlation, genome-wide genetic
365 correlation and phenotypic correlation between trait pairs are not always concordant. This discor-
366 dance suggests that the HLA region plays not only an important, but also a distinct role relative to
367 the rest of the genome in contributing to the shared biology underlying these diseases.

368 In total we identified 7,649 significant trait associations across 647 unique diseases in the HLA
369 region. Here, we highlight interesting patterns across these traits and example associations, but
370 we have only begun to explore the thousands of disease associations generated by these analyses.
371 Therefore we are releasing the association test results as a resource for future studies of the HLA
372 region (Supplementary Tables 4-6). For example, our haplotype association results identify multiple
373 traits or disease complications of previously unknown pathology that cluster with traits with known
374 mechanism. It could be fruitful to use these clusters to generate hypotheses about the biology
375 underlying idiopathic traits. In addition, the haplotypes present in FinnGen represent only a fraction
376 of the genetic diversity present in the world. As more large cohort data continue to become available
377 from regions around the world, future studies will benefit from application of these methods in other
378 cohorts to study the HLA region as the haplotype level.

379 In conclusion, this work offers insights into the role of the HLA region in modulating the complex
380 interplay between hundreds of diseases. Our findings highlight haplotype regression analysis as an
381 additional approach for studying genetic variation in the region beyond the classical HLA alleles.
382 Our results also provide insight into the nature of pleiotropy in the region and highlight novel
383 pathological processes for not only infectious and autoimmune diseases typically associated with
384 HLA, but also across a broad range of diseases.

385 **Methods**

386 **Biobank samples and participants**

387 The FinnGen study (see Supplementary Table 7 for full list of FinnGen contributors) is a large-scale
388 genomics initiative that has analyzed over 500,000 Finnish biobank samples and correlated genetic
389 variation with health data to understand disease mechanisms and predispositions. The project is a
390 collaboration between research organisations, biobanks within Finland, and international industry
391 partners. Here, we used data from FinnGen Data Freeze 10, which is comprised of samples from
392 412,181 Finnish individuals, 21,311,942 variants, and 2,459 traits.

393 **FinnGen Identification of SNP Associations**

394 The summary statistics used in this study were generated using Regenie v2.2.4 and the FinnGen
395 Regenie pipeline [58]. Current age or age at death, sex, genotyping chip, genetic relationship, and
396 the first 10 principal components of the genome-wide genotype matrix were included as covariates
397 [59]. Fine-mapping was performed using the SuSiE "Sum of Single Effects" model [60], excluding
398 the HLA region. Further details are available at <https://www.finnngen.fi/en>.

399 **Defining the HLA region**

400 The HLA region was defined as 28,510,120-33,480,577 based on the Genome Reference Consortium
401 Assembly Grch38.p14 (hg38) (<https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC>).
402 Protein coding genes were identified by overlapping FinnGen annotated genes with the protein
403 coding gene file from HGNC (<https://www.genenames.org/download/statistics-and-files>).
404 The LD plot represents linkage disequilibrium as measured by r^2 and D' for 41,183 SNPs covering
405 the HLA region. This set of SNPs corresponds to the subset of the 41,234 SNPs ($MAF > 1\%$)
406 within the HLA boundaries remaining after pruning with "plink -ld-window 999999 -ld-window-kb
407 1000 -ld-window-r2 0.1".

408 **GWAS hit processing**

409 GWAS results were filtered to include all traits with at least one hit in the HLA region with P
410 $< 10^{-6}$. LD score regression [61] was used to generate genetic correlation estimates, with relevant
411 `eur*_ld_chr` files downloaded from <https://data.broadinstitute.org/alkesgroup/LDSCORE/>.
412 To remove essentially redundant traits, we further filtered to traits with LDSC genetic correlation
413 < 0.95 with all remaining traits. We filtered to the most significant SNP ($MAF > 1\%$) in the
414 HLA region for each of the remaining traits. We then used stepwise forward conditional analysis
415 with Plink2 (<https://www.cog-genomics.org/plink/2.0/>) for each trait to identify additional
416 independent significant SNPs ($MAF > 1\%$) in the HLA region with $P < 10^{-6}$. A significance
417 threshold of $P < 10^{-6}$ was selected modified from the genome-wide significance threshold of 5×10^{-8}
418 because here we are only considering SNPs in the HLA region.

419 In the conditional analysis, we considered only unrelated individuals, reducing the sample size
420 to 259,802. We adjusted for age, sex and 10 principal components of the genome-wide genotype
421 matrix. Z-scores were calculated from the GWAS results for the associations of the 428 hits in
422 the HLA region with all the 272 HLA-associated traits. For visualizing effects across traits, we

423 normalized squared Z-scores for each trait by the maximum Z^2 for that trait. The sign of each
424 SNP's effects were assigned such that the SNP had a positive median Z-score across traits.

425 Enrichment analysis

426 All traits with at least one associated SNP ($MAF > 1\%$ and $P < 10^{-6}$) anywhere in the genome were
427 included, and binned into trait groups. A threshold of $P < 10^{-6}$ was chosen for ascertaining SNP
428 associations in the genome outside HLA to conservatively match the significance threshold used to
429 identify significant associations in the HLA region via the method described above. Enrichment
430 was calculated for each trait group by dividing the number of independent hits per SNP in the HLA
431 region by the number of independent hits per SNP outside the HLA region. For verification that
432 this enrichment was not driven by SNP density, this process was also repeated using enrichment
433 per genes and per base pair.

434 Defining Haplotype Groups

435 Three regions ("blocks") in the HLA region were selected based on the density of signal from the
436 significant SNP associations, overlapping LD patterns, and functional relevance. The first block
437 was defined as 100kb below the start of the gene boundary of *HLA-F* to 100kb past the end of the
438 gene boundary of *HLA-A*, 29,622,820 to 30,045,616 (Grch38.p14) and contained 5,022 SNPs. The
439 second block was defined as 100kb below the start of the gene boundary of *HLA-C* to 100kb past
440 the end of the gene boundary of *MICB*, 31,168,798 to 31,611,071 (Grch38.p14) and contained 8,073
441 SNPs. The third block was defined as 100kb below the start of the gene boundary of *NOTCH4*
442 to 100kb past the end of the gene boundary of *HLA-DQA2*, 32,094,910 to 32,847,125 (Grch38.p14)
443 and contained 11,027 SNPs.

444 Each block was then subset down to 1,000 randomly selected biallelic SNPs with $MAF > 1\%$ due
445 to computational constraints of the clustering process. Each individual's two phased haplotypes at
446 these 1,000 positions were identified. Haplotypes were clustered by first removing rare haplotypes
447 (defined as < 10 total copies across all participants), generating a dendrogram, and recursively
448 splitting the dendrogram at each branch point from the root toward the tips until the total number
449 of haplotypes below each node was less than the maximum threshold (defined as 80,000 copies or the
450 maximum in a single haplotype, whichever was greater). Once the haplotype groups were identified,
451 the rare haplotypes were then added to the group with which they clustered.

452 Performing haplotype regression analysis

453 Logistic regression was then performed separately for each block for each of the 269 diseases with
454 at least one SNP association in the HLA region for all haplotype groups, leaving out the haplotype
455 group with the highest frequency. Sex, age, and the first ten principal components of the genome-
456 wide genotype matrix were included as covariates. The left out haplotype group was then set to 0
457 and the Z-scores of the regression results were then rescaled for each trait to have a mean of 0. A
458 significance threshold of $|Z| > 4$ was chosen based approximately on the Bonferroni correction for
459 the number of regressions (one for each of the 269 diseases) for each block at a significance level of
460 0.05.

461 In a follow-up analysis, we additionally performed haplotype regression analysis for all traits
462 regardless of whether there was a GWAS hit in the HLA region for that trait, and for these regressions

463 we applied a more stringent significance threshold of $P < 6.7 \times 10^{-6}$ to account for the additional
464 traits tested (2459 traits * 3 blocks).

465 Analysis of haplotype regression results

466 Subsequent analyses investigating patterns of pleiotropy of these haplotype groups focused on only
467 the subset of diseases with at least one association $|Z| > 4$ in that block and the subset of haplotype
468 groups with at least one trait association or total copies greater than the minimum cutoff of 20,000
469 copies. For these analyses, a significant threshold of $|Z| > 3$ was chosen based on the Bonferroni
470 correction for the number of regressions (41 traits for Block 1, 46 for Block 2, and 36 for Block 3)
471 for each block at a significance level of 0.05.

472 To calculate the overall disease burden proportion for each haplotype group, we defined the set
473 of relevant diseases for each block as any disease that was significantly associated with at least one
474 of the haplotype groups in that block. Then for each haplotype group in a given block, we identified
475 the proportion of individuals in the haplotype group that had a diagnosis of at least one of the
476 block's relevant diseases. An individual was considered to be in a haplotype group if they were a
477 carrier for at least one haplotype in the haplotype group. To identify the overall disease proportion
478 as a baseline comparison, for each block we identified the proportion of all 412,181 individuals that
479 had a diagnosis of at least one of the block's relevant diseases. We performed an exact binomial test
480 to determine the significance of the disease burden for haplotype group 49 in Block 1 to the block's
481 baseline disease prevalence of 67.5%.

482 Allele regression analysis

483 To determine the extent to which the haplotype group signal remained after adjusting for the
484 classical HLA alleles, we reran the haplotype group regressions while adjusting for the HLA alleles
485 in each block (frequency $> 1\%$ and variance inflation factor < 5). We performed Firth's Bias-
486 Reduced Logistic Regression for all haplotype groups and alleles for each block and each trait using
487 `logistf` (<https://cran.r-project.org/web/packages/logistf/index.html>). We then compared
488 the Z-scores from the regression before and after adjusting for the alleles, using $|Z| > 4$ for the
489 significance threshold. A significance threshold of $|Z| > 4$ was chosen based approximately on the
490 Bonferroni correction for the number of regressions (one for each of the 269 diseases) for each block
491 at a significance level of 0.05.

492 We performed the allele associations on all traits, regardless of whether there was a GWAS hit
493 in the HLA region, using two approaches with sex, age, and 10 PCs included as covariates. For the
494 first approach, we performed logistic regression separately for each block and each trait with one
495 allele included in each regression, with a significance threshold of $P < 2 \times 10^{-7}$. This threshold
496 was chosen to account for the additional traits tested (2459 traits * 98 alleles). For the second
497 approach, we modeled all alleles within a block together jointly after we iteratively removed one
498 regression variable at a time until all remaining had variance inflation factor < 5 to minimize issues
499 of multi-collinearity, and applied a significance threshold of $P < 6.7 \times 10^{-6}$. This threshold was
500 chosen to account for the additional traits tested (2459 traits * 3 blocks).

501 Ethics statement

502 Participants in FinnGen provided informed consent for biobank research based on the Finnish
503 Biobank Act. Alternatively, separate research cohorts, collected before the Finnish Biobank Act
504 came into effect (in September 2013) and the start of FinnGen (August 2017), were collected based
505 on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea
506 (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Re-
507 cruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics
508 Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study
509 protocol (number HUS/990/2017).

510 The FinnGen study is approved by the Finnish Institute for Health and Welfare (permit numbers:
511 THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018,
512 THL/283/6.02.00/2019, THL/1721/5.05.00/2019 and THL/1524/5.05.00/2020), the Digital and
513 population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-
514 3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018,
515 KELA 70/522/2019, KELA 98/522/2019, KELA 134/522/2019, KELA 138/522/2019, KELA 2/522/2020,
516 KELA 16/522/2020), Findata permit numbers (THL/2364/14.02/2020, THL/4055/14.06.00/2020,
517 THL/3433/14.06.00/2020, THL/4432/14.06/2020, THL/5189/14.06/2020, THL/5894/14.06.00/2020,
518 THL/6619/14.06.00/2020, THL/209/14.06.00/2021, THL/688/14.06.00/2021, THL/1284/14.06.00/2021,
519 THL/1965/14.06.00/2021, THL/5546/14.02.00/2020, THL/2658/14.06.00/2021, THL/4235/14.06.00/2021),
520 Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-
521 20) TK/1735/07.03.00/2021, TK/3112/07.03.00/2021) and the Finnish Registry for Kidney Diseases
522 permission/extract from the meeting minutes on 4th July 2019.

523 The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze
524 10 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67,
525 BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, BB2021_65, Finnish Red Cross
526 Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, HUS/248/2020, HUS/430/2021
527 §28, §29, HUS/150/2022 §12, §13, §14, §15, §16, §17, §18, §23, §58, §59, HUS/128/2023 §18, Au-
528 ria Biobank AB17-5154 and amendment #1 (August 17 2020) and amendments BB_2021-0140,
529 BB_2021-0156 (August 26 2021, Feb 2 2022), BB_2021-0169, BB_2021-0179, BB_2021-0161,
530 AB20-5926 and amendment #1 (April 23 2020) and its modifications (Sep 22 2021), BB_2022-
531 0262, BB_2022-0256, Biobank Borealis of Northern Finland (2017_1013, 2021_5010, 2021_5010
532 Amendment, 2021_5018, 2021_5018 Amendment, 2021_5015, 2021_5015 Amendment, 2021_5015
533 Amendment_2, 2021_5023, 2021_5023 Amendment, 2021_5023 Amendment_2, 2021_5017, 2021_5017
534 Amendment, 2022_6001, 2022_6001 Amendment, 2022_6006 Amendment, 2022_6006 Amend-
535 ment, 2022_6006 Amendment_2, BB22-0067, 2022_0262, 2022_0262 Amendment), Biobank of
536 Eastern Finland (1186/2018 and amendment 22§/2020, 53§/2021, 13§/2022, 14§/2022, 15§/2022,
537 27§/2022, 28§/2022, 29§/2022, 33§/2022, 35§/2022, 36§/2022, 37§/2022, 39§/2022, 7§/2023, 32§/2023,
538 33§/2023, 34§/2023, 35§/2023, 36§/2023, 37§/2023, 38§/2023, 39§/2023, 40§/2023, 41§/2023),
539 Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), BB2021-
540 0140 8§/2021, 9§/2021, §9/2022, §10/2022, §12/2022, 13§/2022, §20/2022, §21/2022, §22/2022,
541 §23/2022, 28§/2022, 29§/2022, 30§/2022, 31§/2022, 32§/2022, 38§/2022, 40§/2022, 42§/2022, 1§/2023,
542 Central Finland Biobank 1-2017, BB_2021-0161, BB_2021-0169, BB_2021-0179, BB_2021-0170,
543 BB_2022-0256, BB_2022-0262, BB22-0067, Decision allowing to continue data processing until 31st
544 Aug 2024 for projects: BB_2021-0179, BB22-0067, BB_2022-0262, BB_2021-0170, BB_2021-0164,
545 BB_2021-0161, and BB_2021-0169, and Terveystalo Biobank STB 2018001 and amendment 25th
546 Aug 2020, Finnish Hematological Registry and Clinical Biobank decision 18th June 2021, Arctic

547 biobank P0844: ARC_2021_1001.

548 Acknowledgements

549 We want to acknowledge the participants and investigators of the FinnGen study. We thank Alyssa
550 Lyn Fortier, Mineto Ota, Roshni Patel, Matthew Aguirre, Tami Gjorgjieva, and other members of
551 the Pritchard lab for helpful discussions. This work has been supported by the National Science
552 Foundation Graduate Research Fellowship, Stanford's Knight-Hennessy Scholars Program, and the
553 Stanford Center for Computational, Evolutionary and Human Genomics (C.J.S), the Finnish Med-
554 ical Foundation (S.S.), and Instrumentarium Science Foundation and Academy of Finland #340539
555 (H.M.O.). The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016
556 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd,
557 Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl),
558 Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property
559 Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novar-
560 tis AG, and Boehringer Ingelheim International GmbH. The following biobanks are acknowledged
561 for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL
562 Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank
563 Borealis of Northern Finland ([https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/
564 Pages/Biobank-Borealis-briefly-in-English.aspx](https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx)), Finnish Clinical Biobank Tampere
565 (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere),
566 Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Biobank of East-
567 ern Finland (www.ita-suomenbiopankki.fi/en), Finnish Red Cross Blood Service Biobank
568 (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank ([www.
569 terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/](http://www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/)). All Finnish
570 Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooper-
571 ative - FINBB (<https://finbb.fi/>) is the coordinator of BBMRI-ERIC operations in Fin-
572 land. The Finnish biobank data can be accessed through the Fingenious® services ([https:
573 //site.fingenious.fi/en/](https://site.fingenious.fi/en/)) managed by FINBB. This work was supported by NIH grants
574 RO1HG008140 and R01AG066490 (to J.K.P.).

575 Competing interests

576 No competing interests to declare.

577 Data and Code availability

578 Code for project data analysis, processing and visualization is available at https://github.com/courtrun/HLA_finnngen. Data generated from this study are available at <https://doi.org/10.5281/zenodo.12763469>.
579
580

581

Supplement

582 Supplementary Figures

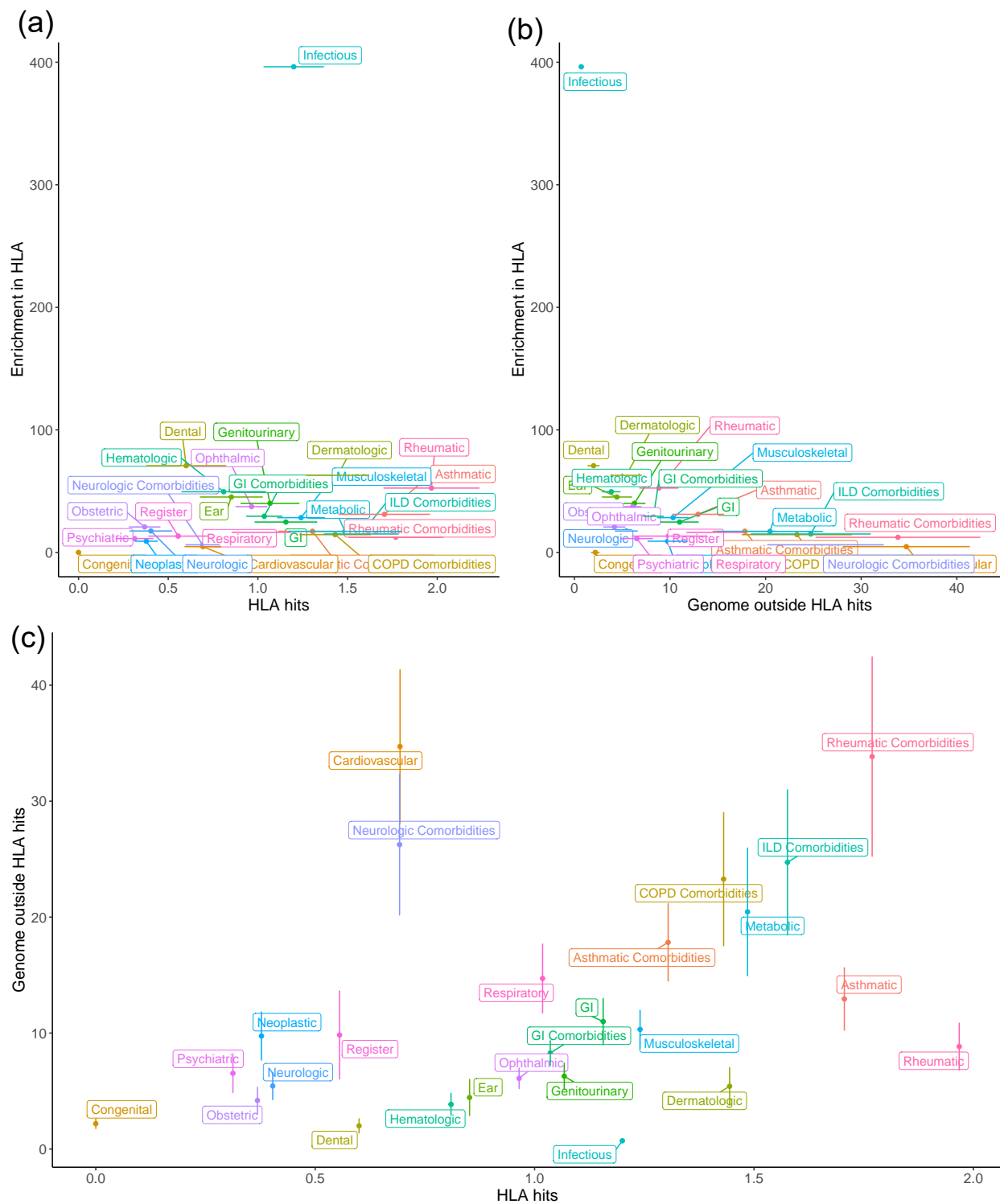


Figure S1: Enrichment by hits. Enrichment of fine-mapped GWAS hits in the HLA region relative to the number of hits throughout the genome outside the HLA region by trait group, compared to **A.** the number of HLA hits and **B.** the number of genome hits. **C.** Number of HLA hits versus the number of genome hits for each trait group.

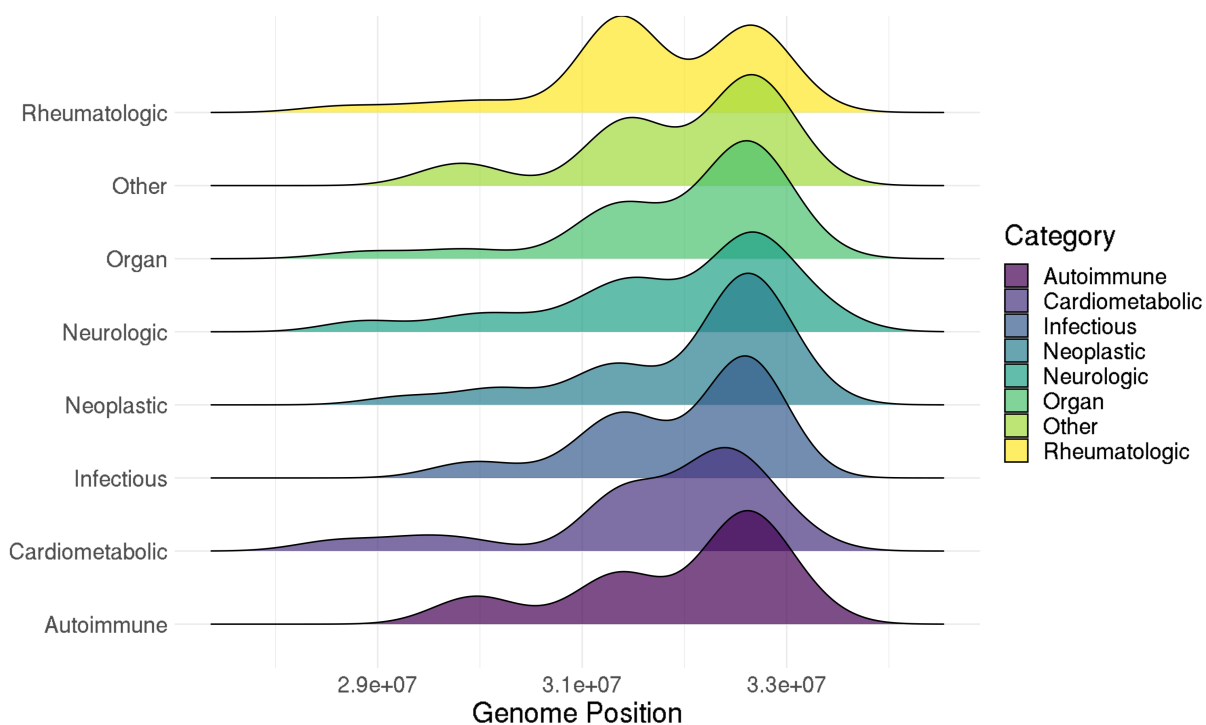


Figure S2: HLA distribution of significant SNP associations. Distribution of the significant SNP associations throughout the HLA region by trait group.

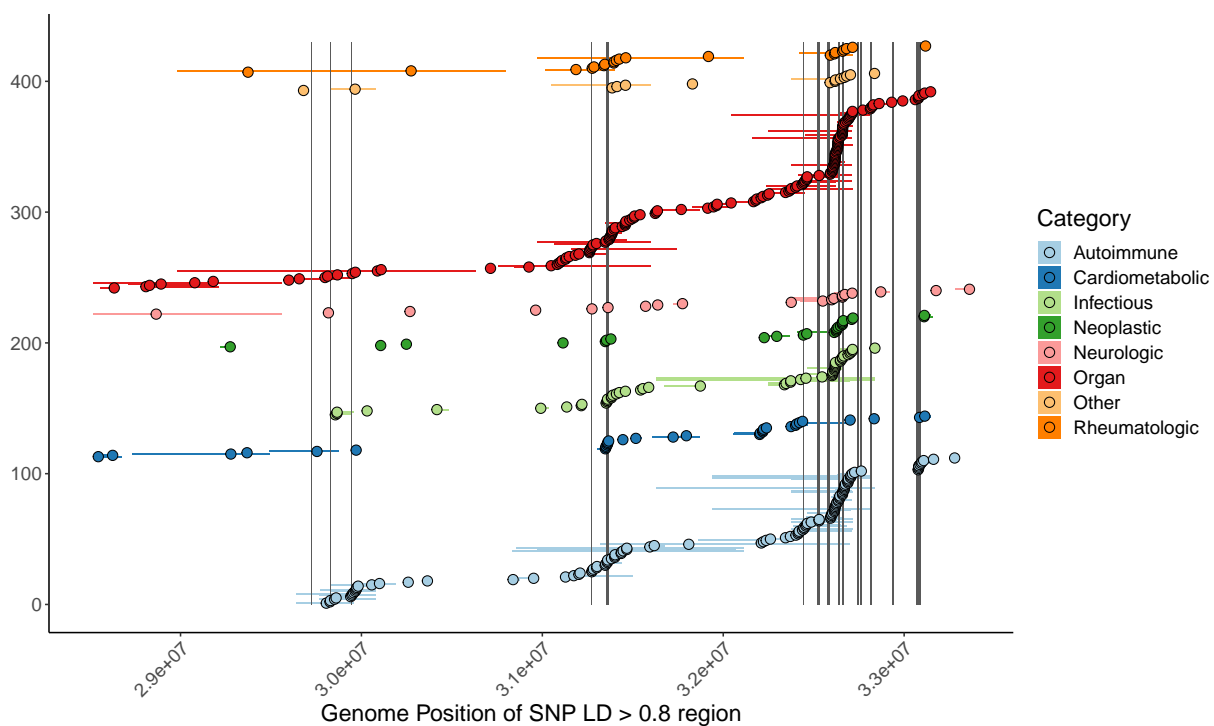


Figure S3: LD boundaries of hits. Genome position of the significant SNP associations in the HLA region with points corresponding to the SNP position and horizontal lines with the bounds corresponding to the lowest and highest genome position of SNPs in LD $r^2 > 0.8$ with each hit.

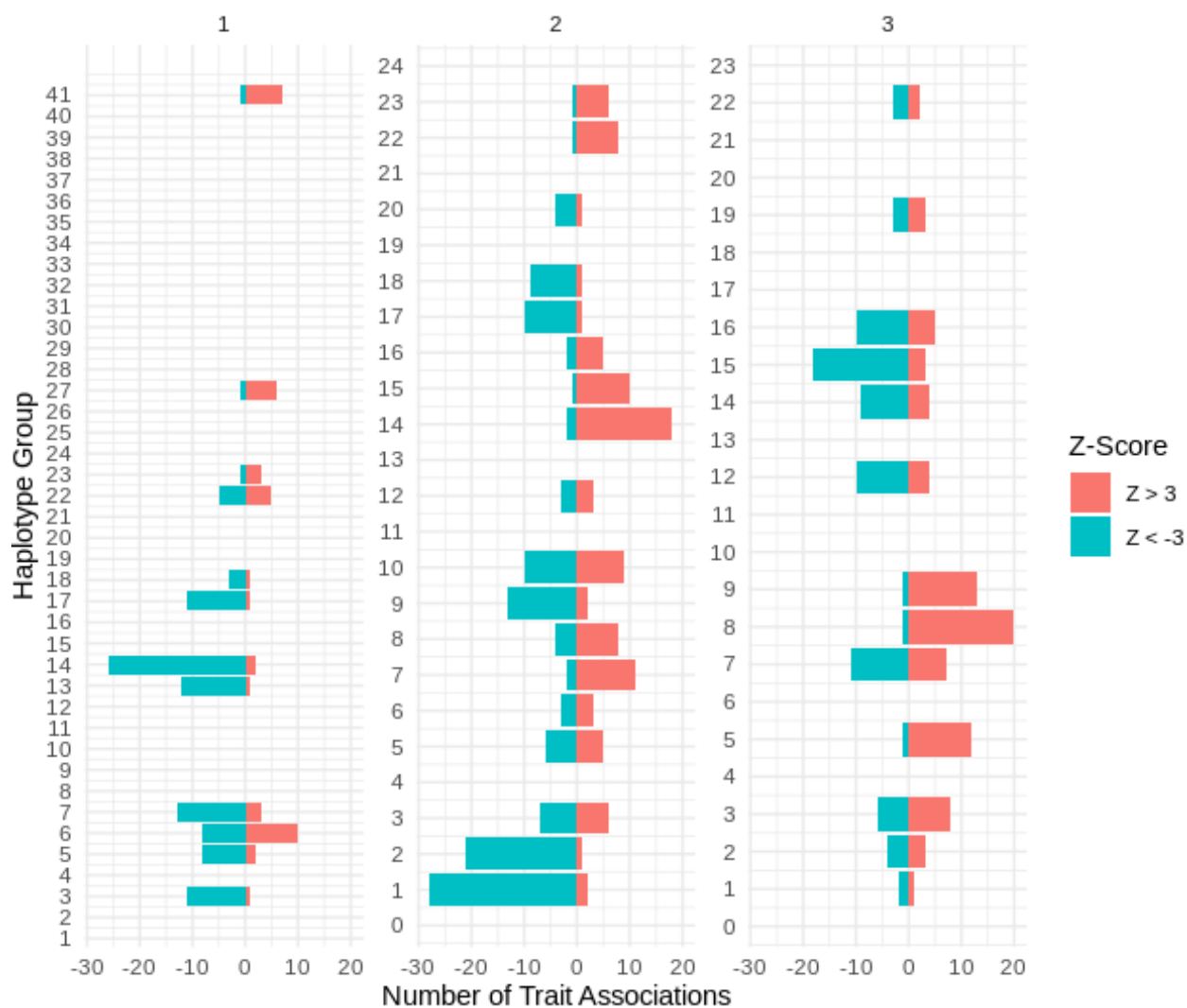


Figure S4: Haplotype group trait associations. Number of traits positively and negatively associated ($|Z| > 3$) with each haplotype group for all three blocks.

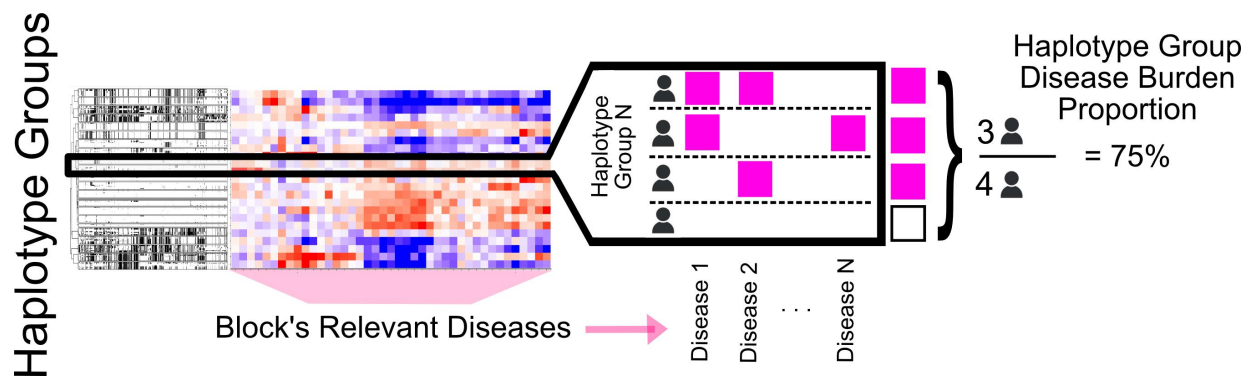


Figure S5: Disease burden analysis overview. Schematic of disease burden proportion analysis. For each haplotype group in a given block, the haplotype disease burden was defined as the proportion of individuals who were a carrier of at least one copy of a haplotype in the haplotype group that had a diagnosis of at least one of the block's relevant diseases (example shown). For each block, the overall disease proportion across all individuals was calculated as the proportion of all individuals that had a diagnosis of at least one of the block's relevant diseases.

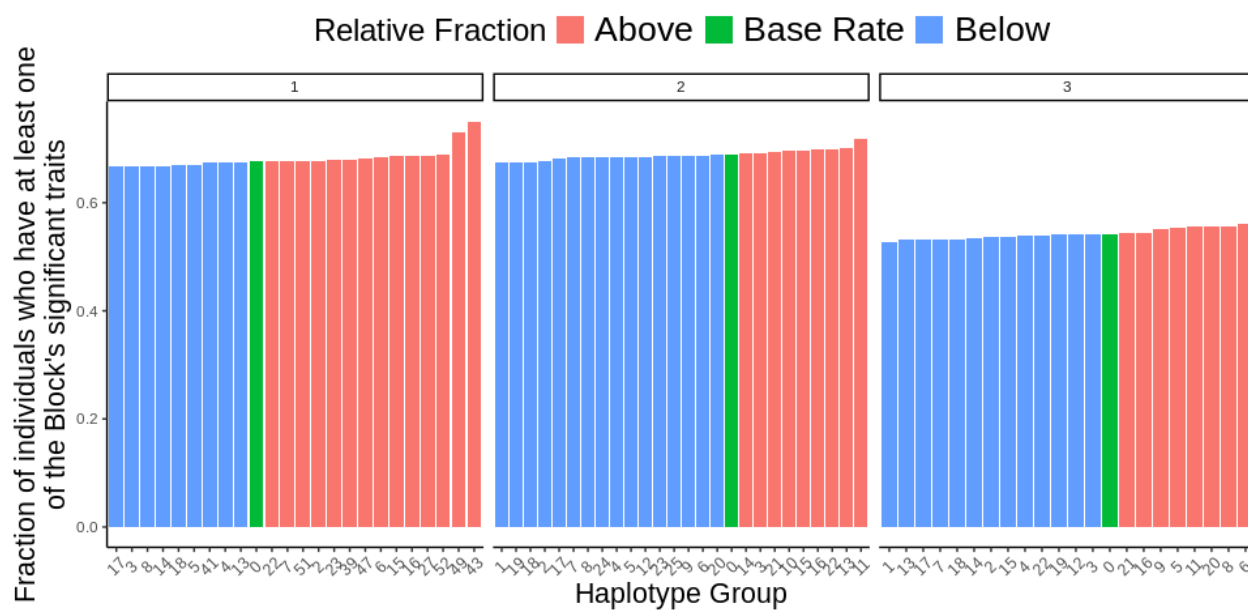


Figure S6: Haplotype group disease burden. Fraction of individuals in each haplotype group who had a diagnosis of at least one of the block's significant traits, for each block. The overall disease proportion, or base rate, was defined as the fraction of individuals in all of FinnGen who had a diagnosis of at least one of the block's significant traits and is shown in green. Haplotype groups with burden below the block's base rate are in blue and those above are in red.

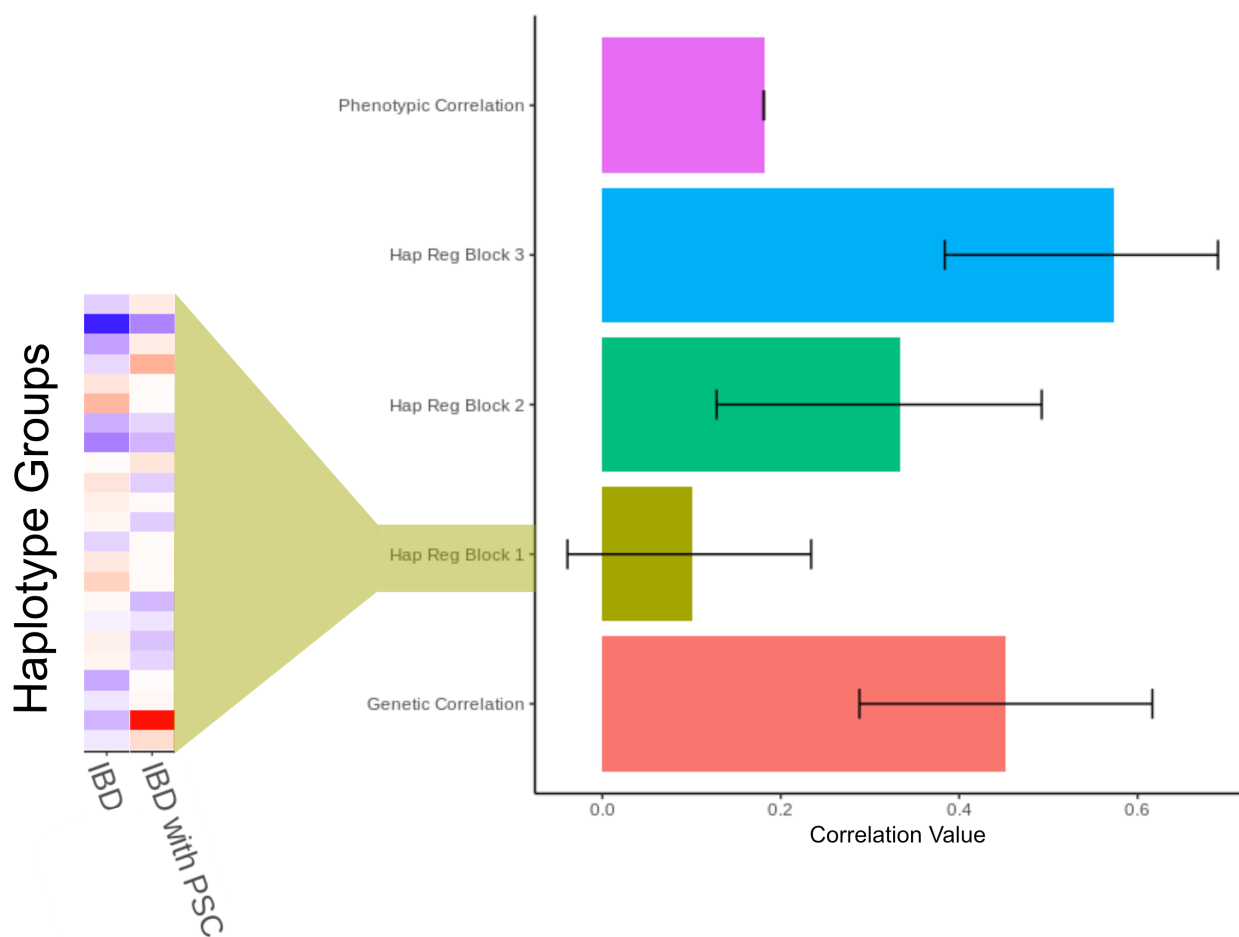


Figure S7: Comparison of IBD and IBD with PSC. Correlation measures between IBD and IBD with PSC. The inset for the haplotype group regression correlation for Block 1 corresponds to the Z-scores for individual haplotype groups in Block 1.

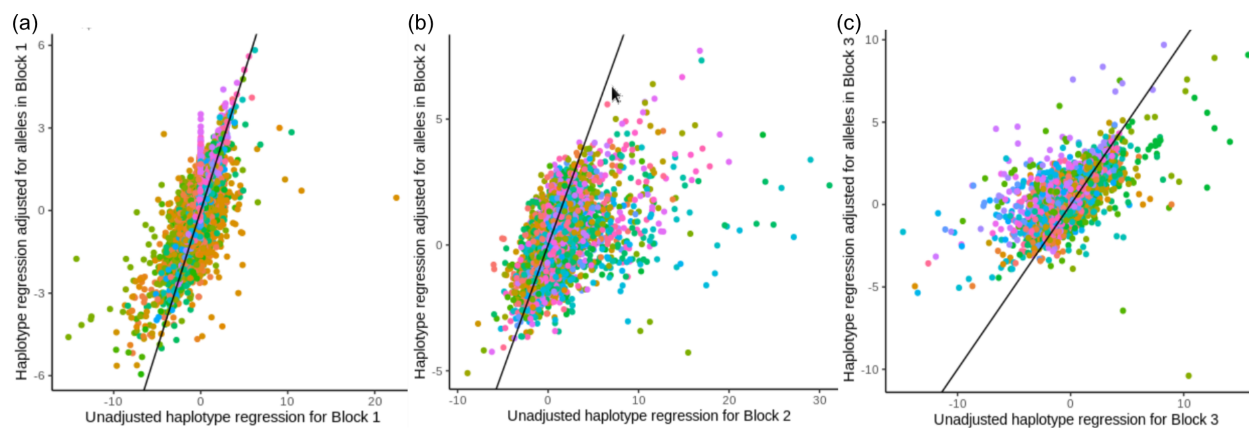


Figure S8: Allele adjusted haplotype group regressions. Comparison of the effects of the haplotype group regressions before and after adjusting for the classical HLA alleles across traits for each block. The points on the scatter plots correspond to Z-scores for different combinations of haplotype group trait associations.

583 **Supplementary Tables**

584 Supplementary Table 1: Enrichment of GWAS hits in the HLA region for each trait group for all
585 traits in FinnGen with at least one GWAS hit ($MAF > 1\%$) anywhere in the genome.

586 Supplementary Table 2: Manual trait group classification for the 269 non-redundant diseases
587 with at least one significantly associated SNP ($P < 10^{-6}$; $MAF > 1\%$) in the HLA region, by
588 pathophysiology first (Category) and then by affected organ system (Subcategory).

589 Supplementary Table 3: Haplotype and haplotype group statistics and assignments. Haplotype
590 statistics are for all haplotypes with > 10 total copies for privacy policy reasons.

591 Supplementary Table 4: Regression results for haplotype groups across all 3 blocks. The first
592 tab has the data plotted in the heatmap of Figure 5, which is the values of the regression Z-scores
593 rescaled to add back in the dropped haplotype group for each block. The next two tabs have the
594 (non-rescaled) regression results for all traits, with and without jointly modeling with the relevant
595 classical HLA alleles in the block.

596 Supplementary Table 5: Regression results for the SNP-trait associations for significant SNP
597 associations remaining after step-wise conditional analysis in the HLA region.

598 Supplementary Table 6: Regression results for all allele associations for all traits, for both the
599 approach jointly modeling alleles within a given block together (tab 1) and for the approach with
600 one allele per regression (tab 2).

601 Supplementary Table 7: List of FinnGen contributors.

602 Supplementary Information

603 Supplementary Information 1: Related haplotypes were clustered into haplotype groups for each
604 block (Supplementary Table 3). This resulted in 53 haplotype groups for Block 1, with a mean
605 of 15,554 total copies and a maximum of 81,997 copies (in haplotype group 15). Block 2 had 25
606 haplotype groups, with a mean of 32,974 copies and a maximum of 77,943 (in haplotype group
607 2). There were 22 haplotype groups for Block 3, with a mean of 37,471 copies and a maximum of
608 76,329 (for haplotype group 21). In Block 1, the mean number of trait associations per haplotype
609 group was 6.25 traits, and haplotype group 14 had the maximum number of significant ($|Z| > 4$)
610 associations with 25 trait associations. In Block 2, the mean trait associations per haplotype group
611 was 8.4, and haplotype group 11 had the most with 30 trait associations. Block 3 had a mean of
612 6.8 trait associations per haplotype group, with haplotype group 8 having the maximum number of
613 associations at 19. Across blocks, the mean number of significant trait associations for each of the
614 block's relevant haplotype groups was 7.2.

615 Supplementary Information 2: Multiple haplotype groups were positively associated with some
616 traits and negatively associated with others. Haplotype group 22 in Block 1 is another example of
617 a group with both positive and negative associations, including 8 traits with association $Z > 2$, and
618 13 with $Z < -2$. This haplotype group was associated with increased risk of Celiac disease, Graves
619 disease and thyrotoxicosis. However, it was also associated with increased risk of hypothyroidism,
620 Sjogren's, and lichen planus. It was again negatively associated with traits like spondylopathies,
621 iridocyclitis, rheumatoid arthritis, but also Type 1 diabetes, chronic tonsil/adenitis, and retinal
622 disorders. In Block 2, haplotype group 10 is positively associated ($Z > 2$) with 13 traits, including
623 sexually transmitted diseases, chronic hepatitis, and Immune disorders. It is negatively associated
624 with 14 traits, including many rheumatic disorders, as well as papulosquamous disorders, and psoriatic
625 arthropathies. Similarly, haplotype group 7 of Block 3, is positively associated with 10 traits,
626 such as multiple sclerosis, degenerative CNS disorders, and demyelinating diseases, and negatively
627 associated with 16 traits, such as type 1 diabetes, retinal disorders, Lichen sclerosus, and juvenile
628 arthritis.

629 References

- 630 1. Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., Lush,
631 M. J., Povey, S., Talbot, C. C., Wright, M. W., et al. (2004). Gene map of the extended human
632 MHC. *Nature Reviews Genetics* 5, 889–899. [10.1038/nrg1489](https://doi.org/10.1038/nrg1489).
- 633 2. Neefjes, J., Jongasma, M. L. M., Paul, P., and Bakke, O. (2011). Towards a systems under-
634 standing of MHC class I and MHC class II antigen presentation. *Nature Reviews Immunology*
635 11, 823–836. [10.1038/nri3084](https://doi.org/10.1038/nri3084).
- 636 3. Ishigaki, K., Lagattuta, K. A., Luo, Y., James, E. A., Buckner, J. H., and Raychaudhuri,
637 S. (2022). HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors.
638 *Nature Genetics* 54, 393–402. [10.1038/s41588-022-01032-z](https://doi.org/10.1038/s41588-022-01032-z).
- 639 4. Fan, W.-L., Shiao, M.-S., Hui, R. C.-Y., Su, S.-C., Wang, C.-W., Chang, Y.-C., and Chung,
640 W.-H. (2017). HLA Association with Drug-Induced Adverse Reactions. *Journal of Immunology*
641 Research 2017. [10.1155/2017/3186328](https://doi.org/10.1155/2017/3186328).
- 642 5. Parham, P. and Guethlein, L. A. (2018). Genetics of Natural Killer Cells in Human Health,
643 Disease, and Survival. *Annual Review of Immunology* 36, 519–548. [10.1146/annurev-immunol-](https://doi.org/10.1146/annurev-immunol-042617-053149)
644 [042617-053149](https://doi.org/10.1146/annurev-immunol-042617-053149).
- 645 6. Butler-Laporte, G., Farjoun, J., Nakanishi, T., Lu, T., Abner, E., Chen, Y., Hultström, M.,
646 Metspalu, A., Milani, L., Mägi, R., et al. (2023). HLA allele-calling using multi-ancestry whole-
647 exome sequencing from the UK Biobank identifies 129 novel associations in 11 autoimmune
648 diseases. *Communications Biology* 6, 1–17. [10.1038/s42003-023-05496-5](https://doi.org/10.1038/s42003-023-05496-5).
- 649 7. Karnes, J. H., Bastarache, L., Shaffer, C. M., Gaudieri, S., Xu, Y., Glazer, A. M., Mosley, J. D.,
650 Zhao, S., Raychaudhuri, S., Mallal, S., et al. (2017). Phenome-wide scanning identifies multiple
651 diseases and disease severity phenotypes associated with HLA variants. *Science Translational*
652 *Medicine* 9. [10.1126/scitranslmed.aai8708](https://doi.org/10.1126/scitranslmed.aai8708).
- 653 8. Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshiba, S., Narita, A.,
654 Konuma, T., Yamamoto, K., Akiyama, M., et al. (2021). A cross-population atlas of genetic
655 associations for 220 human phenotypes. *Nature Genetics* 53, 1415–1424. [10.1038/s41588-021-](https://doi.org/10.1038/s41588-021-00931-x)
656 [00931-x](https://doi.org/10.1038/s41588-021-00931-x).
- 657 9. Kennedy, A. E., Ozbek, U., and Dorak, M. T. (2017). What has GWAS done for HLA and
658 disease associations? *International Journal of Immunogenetics* 44, 195–211. [10.1111/iji.12332](https://doi.org/10.1111/iji.12332).
- 659 10. Hurley, C. K. (2021). Naming HLA diversity: A review of HLA nomenclature. *Human Immunol-*
660 *ogy. Defining and Characterizing HLA Diversity* 82, 457–465. [10.1016/j.humimm.2020.03.005](https://doi.org/10.1016/j.humimm.2020.03.005).
- 661 11. Pierini, F. and Lenz, T. L. (2018). Divergent Allele Advantage at Human MHC Genes: Sig-
662 natures of Past and Ongoing Selection. *Molecular Biology and Evolution* 35, 2145–2158. [10.](https://doi.org/10.1093/molbev/msy116)
663 [1093/molbev/msy116](https://doi.org/10.1093/molbev/msy116).

- 664 12. Manczinger, M., Boross, G., Kemény, L., Müller, V., Lenz, T. L., Papp, B., and Pál, C. (2019).
665 Pathogen diversity drives the evolution of generalist MHC-II alleles in human populations.
666 PLOS Biology 17. 10.1371/journal.pbio.3000131.
- 667 13. Özer, O. and Lenz, T. L. (2021). Unique Pathogen Peptidomes Facilitate Pathogen-Specific
668 Selection and Specialization of MHC Alleles. *Molecular Biology and Evolution* 38, 4376–4387.
669 10.1093/molbev/msab176.
- 670 14. Miyadera, H. and Tokunaga, K. (2015). Associations of human leukocyte antigens with autoim-
671 mune diseases: challenges in identifying the mechanism. *Journal of Human Genetics* 60, 697–
672 702. 10.1038/jhg.2015.100.
- 673 15. Radwan, J., Babik, W., Kaufman, J., Lenz, T. L., and Winternitz, J. (2020). Advances in the
674 Evolutionary Understanding of MHC Polymorphism. *Trends in Genetics* 36, 298–311. 10.1016/
675 j.tig.2020.01.008.
- 676 16. Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent
677 selection and polymorphism of major histocompatibility complex loci. *Genetics* 124, 967–978.
678 10.1093/genetics/124.4.967.
- 679 17. Fortier, A. L. and Pritchard, J. K. (2022). Ancient Trans-Species Polymorphism at the Major
680 Histocompatibility Complex in Primates. Preprint at bioRxiv. 10.1101/2022.06.28.497781.
- 681 18. Arden, B. and Klein, J. (1982). Biochemical comparison of major histocompatibility complex
682 molecules from different subspecies of *Mus musculus*: evidence for trans-specific evolution of
683 alleles. *Proceedings of the National Academy of Sciences* 79, 2342–2346. 10.1073/pnas.79.7.
684 2342.
- 685 19. Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., Seventer, G. van, and Klein, J. (1988). Nu-
686 cleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evo-
687 lution. *The EMBO Journal* 7, 2765–2774. 10.1002/j.1460-2075.1988.tb03131.x.
- 688 20. Ohashi, A., Murayama, M. A., Miyabe, Y., Yudoh, K., and Miyabe, C. (2024). Streptococcal
689 infection and autoimmune diseases. *Frontiers in Immunology* 15. 10.3389/fimmu.2024.1361123.
- 690 21. Hillary, R. P., Ollila, H. M., Lin, L., Desestret, V., Rogemond, V., Picard, G., Small, M., Arnulf,
691 I., Dauvilliers, Y., Honnorat, J., et al. (2018). Complex HLA association in paraneoplastic
692 cerebellar ataxia with anti-Yo antibodies. *Journal of Neuroimmunology* 315, 28–32. 10.1016/
693 j.jneuroim.2017.12.012.
- 694 22. Santambrogio, L. and Marrack, P. (2023). The broad spectrum of pathogenic autoreactivity.
695 *Nature Reviews Immunology* 23, 69–70. 10.1038/s41577-022-00812-2.
- 696 23. Bakkalci, D., Jia, Y., Winter, J. R., Lewis, J. E., Taylor, G. S., and Stagg, H. R. (2020). Risk
697 factors for Epstein Barr virus-associated cancers: a systematic review, critical appraisal, and
698 mapping of the epidemiological evidence. *Journal of Global Health* 10. 10.7189/jogh.10.010405.

- 699 24. Khan, G. and Hashim, M. J. (2014). Global burden of deaths from Epstein-Barr virus at-
700 tributable malignancies 1990-2010. *Infectious Agents and Cancer* 9, 38. 10.1186/1750-9378-9-
701 38.
- 702 25. Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year
703 2002. *International Journal of Cancer* 118, 3030–3044. 10.1002/ijc.21731.
- 704 26. Bjornevik, K., Cortese, M., Healy, B. C., Kuhle, J., Mina, M. J., Leng, Y., Elledge, S. J.,
705 Niebuhr, D. W., Scher, A. I., Munger, K. L., et al. (2022). Longitudinal analysis reveals high
706 prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* 375, 296–301. 10.
707 1126/science.abj8222.
- 708 27. Bjornevik, K., Münz, C., Cohen, J. I., and Ascherio, A. (2023). Epstein-Barr virus as a leading
709 cause of multiple sclerosis: mechanisms and implications. *Nature Reviews Neurology* 19, 160–
710 171. 10.1038/s41582-023-00775-5.
- 711 28. Brown, M. A., Kenna, T., and Wordsworth, B. P. (2016). Genetics of ankylosing spondylitis–
712 insights into pathogenesis. *Nature Reviews Rheumatology* 12, 81–91. 10.1038/nrrheum.2015.
713 133.
- 714 29. Noble, J. A. and Valdes, A. M. (2011). Genetics of the HLA Region in the Prediction of Type
715 1 Diabetes. *Current Diabetes Reports* 11, 533–542. 10.1007/s11892-011-0223-x.
- 716 30. Ziade, N. (2023). Human leucocyte antigen-B27 testing in clinical practice: a global perspective.
717 *Current Opinion in Rheumatology* 35, 235–242. 10.1097/BOR.0000000000000946.
- 718 31. Raiteri, A., Granito, A., Giamperoli, A., Catenaro, T., Negrini, G., and Tovoli, F. (2022).
719 Current guidelines for the management of celiac disease: A systematic review with comparative
720 analysis. *World Journal of Gastroenterology* 28, 154–175. 10.3748/wjg.v28.i1.154.
- 721 32. Amstutz, U., Shear, N. H., Rieder, M. J., Hwang, S., Fung, V., Nakamura, H., Connolly, M. B.,
722 Ito, S., Carleton, B. C., and CPNDS clinical recommendation group (2014). Recommendations
723 for HLA-B*15:02 and HLA-A*31:01 genetic testing to reduce the risk of carbamazepine-induced
724 hypersensitivity reactions. *Epilepsia* 55, 496–506. 10.1111/epi.12564.
- 725 33. D’Antonio, M., Reyna, J., Jakubosky, D., Donovan, M. K., Bonder, M.-J., Matsui, H., Ste-
726 gle, O., Nariai, N., D’Antonio-Chronowska, A., and Frazer, K. A. (2019). Systematic genetic
727 analysis of the MHC region reveals mechanistic underpinnings of HLA type associations with
728 disease. *eLife* 8. 10.7554/eLife.48476.
- 729 34. Bettens, F., Ongen, H., Rey, G., Buhler, S., Sollet, Z. C., Dermitzakis, E., and Villard, J.
730 (2022). Regulation of HLA class I expression by non-coding gene variations. *PLOS Genetics*
731 18. 10.1371/journal.pgen.1010212.
- 732 35. Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. (2018). HLA variation and disease.
733 *Nature Reviews Immunology* 18, 325–339. 10.1038/nri.2017.143.

- 734 36. Jin, Y., Roberts, G. H. L., Ferrara, T. M., Ben, S., Geel, N. van, Wolkerstorfer, A., Ezzedine,
735 K., Siebert, J., Neff, C. P., Palmer, B. E., et al. (2019). Early-onset autoimmune vitiligo asso-
736 ciated with an enhancer variant haplotype that upregulates class II HLA expression. *Nature*
737 *Communications* 10, 391. [10.1038/s41467-019-08337-4](https://doi.org/10.1038/s41467-019-08337-4).
- 738 37. Sekar, A., Bialas, A. R., De Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley,
739 K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex
740 variation of complement component 4. *Nature* 530, 177–183. [10.1038/nature16549](https://doi.org/10.1038/nature16549).
- 741 38. Gupta, A. and Thelma, B. K. (2016). Identification of critical variants within SLC44A4, an
742 ulcerative colitis susceptibility gene identified in a GWAS in north Indians. *Genes & Immunity*
743 17, 105–109. [10.1038/gene.2015.53](https://doi.org/10.1038/gene.2015.53).
- 744 39. Zhang, X., Lucas, A. M., Veturi, Y., Drivas, T. G., Bone, W. P., Verma, A., Chung, W. K.,
745 Crosslin, D., Denny, J. C., Hebring, S., et al. (2022). Large-scale genomic analyses reveal
746 insights into pleiotropy across circulatory system diseases and nervous system disorders. *Nature*
747 *Communications* 13, 3428. [10.1038/s41467-022-30678-w](https://doi.org/10.1038/s41467-022-30678-w).
- 748 40. Ritari, J., Koskela, S., Hyvärinen, K., FinnGen, n., and Partanen, J. (2022). HLA-disease
749 association and pleiotropy landscape in over 235,000 Finns. *Human Immunology* 83, 391–398.
750 [10.1016/j.humimm.2022.02.003](https://doi.org/10.1016/j.humimm.2022.02.003).
- 751 41. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic,
752 D., Delaneau, O., O’Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping
753 and genomic data. *Nature* 562, 203–209. [10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z).
- 754 42. Hirata, J., Hosomichi, K., Sakaue, S., Kanai, M., Nakaoka, H., Ishigaki, K., Suzuki, K.,
755 Akiyama, M., Kishikawa, T., Ogawa, K., et al. (2019). Genetic and phenotypic landscape
756 of the major histocompatibility complex region in the Japanese population. *Nature Genetics*
757 51, 470–480. [10.1038/s41588-018-0336-0](https://doi.org/10.1038/s41588-018-0336-0).
- 758 43. Mozzi, A., Pontremoli, C., and Sironi, M. (2018). Genetic susceptibility to infectious diseases:
759 Current status and future perspectives from genome-wide approaches. *Infection, Genetics and*
760 *Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*
761 66, 286–307. [10.1016/j.meegid.2017.09.028](https://doi.org/10.1016/j.meegid.2017.09.028).
- 762 44. Apps, R., Qi, Y., Carlson, J. M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G. Q.,
763 Goulder, P., Brumme, Z. L., et al. (2013). Influence of HLA-C Expression Level on HIV Control.
764 *Science* 340, 87–91. [10.1126/science.1232685](https://doi.org/10.1126/science.1232685).
- 765 45. Tian, C., Hromatka, B. S., Kiefer, A. K., Eriksson, N., Noble, S. M., Tung, J. Y., and Hinds,
766 D. A. (2017). Genome-wide association and HLA region fine-mapping studies identify suscep-
767 tibility loci for multiple common infections. *Nature Communications* 8, 599. [10.1038/s41467-](https://doi.org/10.1038/s41467-017-00257-5)
768 [017-00257-5](https://doi.org/10.1038/s41467-017-00257-5).
- 769 46. Binder, M. D., Fox, A. D., Merlo, D., Johnson, L. J., Giuffrida, L., Calvert, S. E., Akker-
770 mann, R., Ma, G. Z. M., ANZgene, Perera, A. A., et al. (2016). Common and Low Fre-
771 quency Variants in MERTK Are Independently Associated with Multiple Sclerosis Suscepti-

- 772 bility with Discordant Association Dependent upon HLA-DRB1*15:01 Status. *PLOS Genetics*
773 12. 10.1371/journal.pgen.1005853.
- 774 47. Bosca-Watts, M. M., Minguez, M., Planelles, D., Navarro, S., Rodriguez, A., Santiago, J.,
775 Tosca, J., and Mora, F. (2018). HLA-DQ: Celiac disease vs inflammatory bowel disease. *World*
776 *Journal of Gastroenterology* 24, 96–103. 10.3748/wjg.v24.i1.96.
- 777 48. Lundström, E., Gustafsson, J. T., Jönsen, A., Leonard, D., Zickert, A., Elvin, K., Sturfelt, G.,
778 Nordmark, G., Bengtsson, A. A., Sundin, U., et al. (2013). HLA-DRB1*04/*13 alleles are asso-
779 ciated with vascular disease and antiphospholipid antibodies in systemic lupus erythematosus.
780 *Annals of the Rheumatic Diseases* 72, 1018–1025. 10.1136/annrheumdis-2012-201760.
- 781 49. Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK
782 Biobank. *Nature Genetics* 50, 1593–1599. 10.1038/s41588-018-0248-z.
- 783 50. Rioux, J. D., Goyette, P., Vyse, T. J., Hammarström, L., Fernando, M. M. A., Green, T.,
784 De Jager, P. L., Foisy, S., Wang, J., Bakker, P. I. W. de, et al. (2009). Mapping of multiple
785 susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of*
786 *the National Academy of Sciences* 106, 18680–18685. 10.1073/pnas.0909307106.
- 787 51. Debebe, B. J., Boelen, L., Lee, J. C., IAVI Protocol C Investigators, Thio, C. L., Astemborski,
788 J., Kirk, G., Khakoo, S. I., Donfield, S. M., Goedert, J. J., et al. (2020). Identifying the immune
789 interactions underlying HLA class I disease associations. *eLife* 9. 10.7554/eLife.54558.
- 790 52. Busch, R., Kollnberger, S., and Mellins, E. D. (2019). HLA associations in inflammatory arthri-
791 tis: emerging mechanisms and clinical implications. *Nature Reviews Rheumatology* 15, 364–
792 381. 10.1038/s41584-019-0219-5.
- 793 53. Queiro, R., Morante, I., Cabezas, I., and Acasuso, B. (2016). HLA-B27 and psoriatic disease: a
794 modern view of an old relationship. *Rheumatology* 55, 221–229. 10.1093/rheumatology/kev296.
- 795 54. Kim, Y. S., Hurley, E. H., Park, Y., and Ko, S. (2023). Primary sclerosing cholangitis (PSC)
796 and inflammatory bowel disease (IBD): a condition exemplifying the crosstalk of the gut–liver
797 axis. *Experimental & Molecular Medicine* 55, 1380–1387. 10.1038/s12276-023-01042-9.
- 798 55. Conigliaro, P., D’Antonio, A., Pinto, S., Chimenti, M. S., Triggianese, P., Rotondi, M., and
799 Perricone, R. (2020). Autoimmune thyroid disorders and rheumatoid arthritis: A bidirectional
800 interplay. *Autoimmunity Reviews* 19. 10.1016/j.autrev.2020.102529.
- 801 56. The China Consortium for the Genetics of Autoimmune Thyroid Disease (2011). A genome-
802 wide association study identifies two new risk loci for Graves’ disease. *Nature Genetics* 43, 897–
803 901. 10.1038/ng.898.
- 804 57. Smith, C. J., Sinnott-Armstrong, N., Cichońska, A., Julkunen, H., Fauman, E. B., Würtz, P.,
805 and Pritchard, J. K. (2022). Integrative analysis of metabolite GWAS illuminates the molecular
806 basis of pleiotropy and genetic correlation. *eLife* 11. 10.7554/eLife.79348.

- 807 58. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Ben-
808 ner, C., O’Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient
809 whole-genome regression for quantitative and binary traits. *Nature Genetics* 53, 1097–1103.
810 10.1038/s41588-021-00870-7.
- 811 59. Kurki, M. I., Karjalainen, J., Palta, P., Sipilä, T. P., Kristiansson, K., Donner, K. M., Reeve,
812 M. P., Laivuori, H., Aavikko, M., Kaunisto, M. A., et al. (2023). FinnGen provides genetic
813 insights from a well-phenotyped isolated population. *Nature* 613, 508–518. 10.1038/s41586-
814 022-05473-8.
- 815 60. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A Simple New Approach to
816 Variable Selection in Regression, with Application to Genetic Fine Mapping. *Journal of the*
817 *Royal Statistical Society* 82, 1273–1300. 10.1111/rssb.12388.
- 818 61. Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H., Ripke, S., Yang, J., Patterson, N., Daly, M. J.,
819 Price, A. L., and Neale, B. M. (2015). LD Score Regression Distinguishes Confounding from
820 Polygenicity in Genome-Wide Association Studies. *Nature Genetics* 47, 291–295. 10.1038/ng.
821 3211.