Pervasive additive and non-additive effects within the HLA region contribute to disease risk in the UK Biobank

Guhan Ram Venkataraman[1+], Julia Eve Olivieri[2], Christopher DeBoever[1], Yosuke Tanigawa[1], Johanne Marie Justesen[1], Manuel A. Rivas[1+]

Affiliations:
1. Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA.
2. Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA.
+Corresponding author

Abstract

The human leukocyte antigen (HLA) region of the genome is one of the most disease-associated regions of the human genome, yet even well-studied alleles in the HLA region have unknown impact on disease. Here, we study the effect of 156 HLA alleles on 677 binary phenotypes for 337,138 individuals in the UK Biobank. We perform single-allele association analysis and subsequent Bayesian Model Averaging for conditional analysis, a) replicating 88 known associations between HLA alleles and cancer and other binary phenotypes, and b) discovering 90 novel associations, including to phenotypes not previously associated with the HLA region (e.g. anemias and acne), and associations to skin and reproductive tract cancers. Finally, we find many non-additive effects, suggesting a more complex landscape of disease modifying effects throughout the region. Our results demonstrate the HLA region's complexity and richness while underscoring its clinical relevance.

Introduction

The human leukocyte antigen (HLA) region of the genome is one of the most disease-associated but gene-dense and polymorphic regions of the human genome[1]. The downstream products of HLA genes generate and present peptides on the cell surface that can be recognized by T cell receptors, making these genes relevant to many disorders of the immune system[2]. Before the advent of high-throughput genome-wide association studies (GWAS), HLA polymorphisms were associated only with autoimmune disorders and cancers[3,4]. Since, GWAS has uncovered associations between HLA and common infections[5], shingles[6], chronic Hepatitis B[7], Epstein-Barr virus[8], and other diseases. The HLA region contains 1.5% of the genes currently in Online Mendelian Inheritance in Man (OMIM, a database of genetic disorders and traits focusing on gene-phenotype relationships) and accounts for roughly 1% (1,827/185,864) of the genome-wide significant SNP associations in the NHGRI/EBI GWAS catalog[9], underscoring its importance to disease. Despite being an attractive target for comprehensive GWAS, the HLA region presents many challenges that obscure its roles in disease pathogenesis; complex linkage disequilibrium (LD) structures, structural variation, closely-related genes, paralogs, segmental duplications, and violations of Hardy-Weinberg equilibrium make high-throughput genotyping methods such as fine-mapping and imputation challenging[10,11]. Despite custom-designed chips (e.g. the Illumina MHC SNP Panel[12]) and robust reference panels built to study the region[13,14], there exist well-studied alleles in the HLA region whose impact on disease are unknown. Thus, understanding the role of the HLA region is critical to assessing disease risk.

The UK Biobank dataset offers researchers genetic and phenotypic data for over 500,000 individuals, and additionally provides the opportunity to analyze HLA region data for 337,138 unrelated white British individuals, generated via imputation strategies that address the aforementioned challenges[13]. These data can be used to perform association analyses across the rich diversity of phenotypes in the UK Biobank, which are derived from cancer registry, hospital in-patient, primary care, and self-reported questionnaire data. Assessing relationships between HLA alleles and individual-level phenotype data can help pinpoint culprit alleles for disease associations via well-powered conditional analysis. We can additionally identify settings where HLA alleles exhibit non-additive effects on phenotypes, improving our ability to accurately integrate them in disease risk models.

Here, we performed a single-allele association analysis for 156 HLA alleles across 677 phenotypes in the UK Biobank and identified 1291 associations at a false discovery rate (FDR) of 5% after Benjamini-Yekutieli multiple testing correction. The associations spanned 113 HLA alleles (across all 11 Biobank-provided HLA loci) and 128 binary phenotypes. We performed Bayesian Model Averaging to narrow down to conditionally independent alleles and uncovered 178 high-confidence (posterior probability >= 80%) associations spanning 50 alleles (across all 11 HLA loci) and 78 binary phenotypes (88 [49%] of which were supported by the literature, **Supplementary Table 1**). We assessed these associations for non-additive effects, i.e. “disease contributions beyond the cumulative effect of individual alleles”[15], finding 25 associations with significant deviations from additivity (8 [32%] of which were supported by the

literature, **Supplementary Table 1**). 9 of these 25 non-additive effects drive intestinal malabsorption and/or celiac disease, but other associations affect a multitude of other autoimmune and endocrine disorders. For example, although HLA-B*27:05 is known to have strong effects on ankylosing spondylitis and iridicyclitis, and HLA-B*57:01 is likewise known to have strong effects on psoriasis, we find that these effects are also non-additive. This study associates HLA alleles with a variety of phenotypes in the UK Biobank through association analyses and model selection techniques, providing important insight into disease pathogenesis.

Materials and Methods

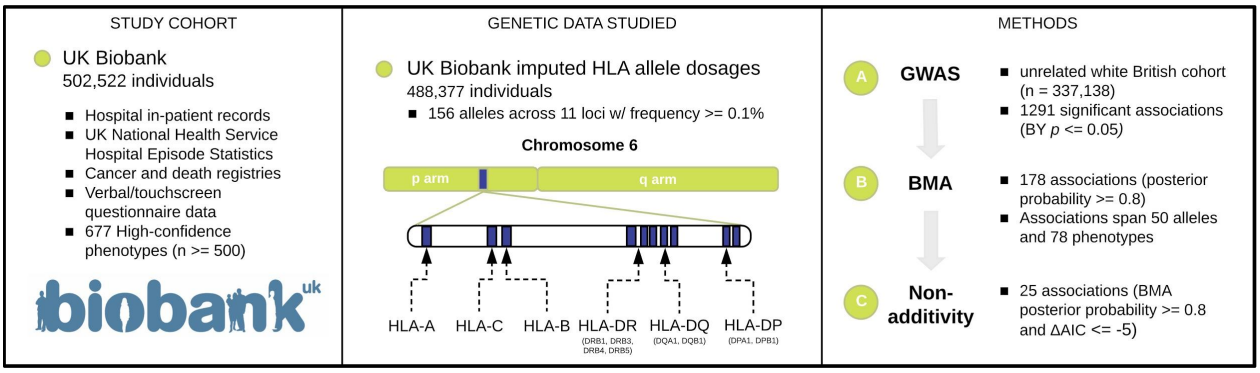

Figure 1. Overview of the study design.

Data

For unrelated white British individuals (n = 337,138) in the UK Biobank[16], three sets of phenotypes were defined: “high-confidence”, “time-to-event”, and “algorithmically-defined”. High-confidence phenotypes were defined by combining disease diagnoses (UK Biobank Field ID 41202) from the UK National Health Service Hospital Episode Statistics (ICD-10 codes H40.[0-6,8,9], H42.8, and Q15.0), the self-reported diagnosis questionnaire (UK Biobank Field ID 20002), and the UK Cancer Registry. Time-to-event phenotypes were derived from First Occurrence of Health Outcomes data as defined by 3-character ICD-10 codes in UK Biobank’s Category 1712. The First Occurrence data-fields were generated by mapping: read code information in the primary care data (Category 3000); ICD-9 and ICD-10 codes in hospital inpatient data (Category 2000); ICD-10 codes in death registry records (Field 40001, Field 40002); and self-reported medical condition codes (Field 20002), reported at baseline or subsequent UK Biobank assessment center visits as 3-character ICD-10 codes. Algorithmically-defined outcomes (based on data from Category 42) include phenotypes of select health-related events obtained through algorithmic combinations of coded information from the UK Biobank’s baseline assessment data collection. The data were derived from self-reported medical conditions, operations and medications together with linked data from hospital admissions and death registries. We included phenotypes with at least 500 cases among the white British cohort (**Supplementary Table 2**) and manually deduplicated several using the FuzzyWuzzy python package’s `partial_ratio` function, resulting in the 677 phenotypes used in the analysis.

HLA alleles HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4, and -DRB5 were imputed using the HLA*IMP:02 program[13]. HLA-IMP:02 is built on multiple reference panels and a graphical model of the HLA haplotype structure. The UK Biobank provides an imputed dosage file that contains the estimated dosage of 362 HLA alleles across 11 HLA loci from 488,377 individuals (while modeling imputation uncertainty). We included 156 alleles across all 11 loci that had a frequency of 0.1% or greater in our white British cohort (**Supplementary Table 2**). For consistency across models, and because the non-additivity analysis requires integral values for allele dosage, those dosages that were within 0.1 of 0, 1, or 2 for each allele were rounded, and the remaining nonzero entries were excluded. Erroneous total allele counts post-rounding were excluded.

Association analysis

We performed generalized linear model association analysis (with “Firth-fallback”) as implemented in PLINK v2.00aLM (March 14 2020)[17], for the above-described 677 phenotypes and 156 HLA alleles across UK Biobank white British individuals (n = 337,138). Firth-fallback is a hybrid algorithm which normally uses logistic regression but “falls back” to Firth's bias reduction method[18], equivalent to penalization of the log-likelihood by the Jeffreys prior, in two cases: (1), one of the cells in the 2x2 (allele count-by-case/control status) contingency table is empty; or (2), logistic regression fails to converge within the usual number of steps. We used age, sex, genotyping array, number and length of copy number variants, and the first ten genotype principal components as covariates in our analysis[19]. We additionally computed an additive model (with the same parameters as PLINK) in R in order to facilitate the subsequent non-additivity analysis. We used the Benjamini-Yekutieli (BY) multiple-testing FDR control method[20] to adjust the resultant p-values, selecting those associations with BY-corrected p-values less than 0.05 to control FDR at 5%.

Bayesian Model Averaging

Given the high LD between HLA alleles and the resultant high number of potentially spurious HLA allele-phenotype associations even despite FDR correction, we used a Bayesian Model Averaging (BMA) approach, implemented in the `bma` R package[21], to prioritize which HLA loci were most likely causal for each phenotype. We used BMA to train a model on each possible subset of significantly-associated alleles. The posterior probabilities of each model are reported, and the degree to which an allele is included across all models (the “allele posterior probability”) is then a measure of confidence in the allele-phenotype association.

We identified all of the allele-phenotype pairs that had BY-adjusted p-values less than or equal to 0.05 from the additive association analysis. Because BMA is exponential in complexity with respect to the number of alleles analyzed (requiring analysis of 2^n models for n alleles included in the analysis), we performed BMA only on the 20 alleles with the lowest BY-adjusted p-values from the additive association analysis for each phenotype. If there were less than two such alleles for a given phenotype, we did not perform BMA for that phenotype. Additionally, only the

models whose posterior probabilities were within a factor of 1/5 of that of the best model were kept for the final averaging. These requirements filtered our testing base down to 116 phenotypes, with 111 alleles included in at least one analysis. We used BMA with a binomial error distribution and link function and with age, sex, genotyping array, and the first ten genotype principal components as covariates (number and length of copy number variant covariates were excluded). We focused on allele-phenotype pairs with BMA posterior probabilities >= 0.8 in the subsequent non-additivity analysis.

Analysis of non-additive genetic effects

To assess whether certain allele-phenotype pairs exhibited non-additive effects on their phenotypes, we additionally performed logistic regressions in R using the glm function in R (family = "binomial") and providing dosages as factors (i.e., separate terms indicated whether a subject was heterozygous or homozygous for the HLA allele in question) and included age, sex, genotyping array, number and length of copy number variants, and the first ten genotype principal components as covariates (as in the single-allele association analysis). We computed an Akaike Information Criterion (AIC, a measure of goodness-of-fit[22]), comparing the non-additive and additive association models in R for model selection. To identify gene-phenotype associations with suspected departures from additivity, we identified allele-phenotype pairs where the BMA posterior probability was greater than 0.8 and the difference between the non-additive model AIC and additive model AIC (ΔAIC) was <= −5.

Literature review

For each allele-phenotype pair resulting from the BMA analysis, we conducted a comprehensive literature review to determine whether the association had been previously found. The HLA allele and the phenotype in question were entered into Google Scholar both with and without quotations, and aliases were checked (e.g. B*08:01, B*0801, B/08/01, and B/0801). Search results were manually inspected to determine associations (**Supplementary Table 1**).

Results

Association analysis

To assess the extent to which HLA alleles affect various phenotypes, we conducted an HLA-wide association study across 337,138 white British individuals in the UK Biobank (Methods). We used 156 HLA alleles (out of 175 total) that were present at greater than 0.1% minor allele frequency in the cohort and 677 de-duplicated binary phenotypes with more than 500 cases in the UK Biobank. These binary phenotypes included autoimmune, lymphatic, cardiovascular, dermal, skeletal, tissue, digestive, respiratory, renal, and endocrine disorders as well as many cancers. To control for false discovery, we corrected for multiple testing across the phenotypes

using the Benjamini-Yekutieli (BY) procedure at a FDR of 5%. We found 113 alleles that were associated with at least one of 128 binary phenotypes for a total of 1291 associations.

Bayesian Model Averaging

Given the high degree of LD between HLA alleles, we examined whether the identified associations were conditionally independent. We used Bayesian Model Averaging (BMA) on the top 20 BY-significant alleles for each phenotype that had at least 2 BY-significant allele associations. Across 116 test phenotypes and 111 alleles included in at least one analysis, we found 178 putatively associated allele-phenotype pairs among 50 distinct alleles and 78 distinct phenotypes. Of these 178 associations, 88 have been previously documented as associated allele-disease pairs, and 90 were novelly marked by our analysis as high probability of having causal signal (**Methods, Supplementary Tables 1-2, Figure 2A, Figure 2B**). Of note, 83 putatively causal allele-phenotype pairs with posterior probability 1 were not previously documented in the literature. These 83 allele-phenotype pairs mostly consist of associations that either feature alleles in LD with known associations or associations between alleles and closely-related phenotypes to known associations. For example, we find an association between **non-melanoma skin cancer and DQB1*03:02**, although the literature reports an association with an allele in LD with DQB1*03:02, DQB1*05:01[23]. Alleles associated to celiac disease, such as **B*08:01**[24], were found to be associated with **intestinal malabsorption and anemias**, which have close ties to celiac disease[25]. Additionally, alleles associated with Graves' disease and Hashimoto thyroiditis (such as **B*08:01**[26], **B*39:06**[27], **DQB1*06:04**[28], and **DRB1*03:01**[29]) were found to be associated with **hypo- and hyper-thyroidism** in the UK Biobank and account for 13 of the 83 high-probability associations. Finally, alleles associated with type 1 diabetes (**DQB1*02:01, DQB1*03:02, and DRB1*04:01**) were found to be associated with "**other disorders of pancreatic internal secretion,**" which is biologically related to type 1 diabetes.

Other associations among these 178 include those between HLA alleles and infectious diseases, asthma, systemic autoimmune disorders, and skin cancers (**Supplementary Table 1**). Among the strongest risk effects are associations between: **B*27:05 and ankylosing spondylitis** ($p_{BY} = 0$, $BMA_{postOR} = 9.02$, 95% CI = [8.19, 9.94])[30], **B*27:05 and iridocyclitis**, a type of uveitis ($p_{BY} = 2.94 \times 10^{-239}$, $BMA_{postOR} = 4.75$, 95% CI = [4.32, 5.23])[31], **B*39:01 and psoriatic arthropathy** ($p_{BY} = 1.06 \times 10^{-8}$, $BMA_{postOR} = 3.54$, 95% CI = [2.51, 5.0])[32] and **DQA1*03:01 and seropositive rheumatoid arthritis** ($p_{BY} = 6.46 \times 10^{-66}$, $BMA_{postOR} = 3.16$, 95% CI = [2.75, 3.63])[33]. We replicate previously-found strong protective effects of **DQB1*03:01** ($p_{BY} = 1.00 \times 10^{-70}$, $BMA_{postOR} = 0.42$, 95% CI = [0.34, 0.51]) and **DQA1*01:01** ($p_{BY} = 2.62 \times 10^{-51}$, $BMA_{postOR} = 0.71$, 95% CI = [0.61, 0.81])[15] on **celiac disease** and **malabsorption**. We also discover novel associations to these phenotypes: **DQB1*03:03** ($p_{BY} = 2.30 \times 10^{-22}$, $BMA_{postOR} = 0.47$, 95% CI = [0.36, 0.62]), DQB1*03:01 ($p_{BY} = 2.00 \times 10^{-53}$, $BMA_{postOR} = 0.42$, 95% CI = [0.34, 0.51]), **DRB4*99:01** ($p_{BY} = 4.47 \times 10^{-31}$, $BMA_{postOR} = 0.52$, 95% CI = [0.46, 0.58]), and **DRB4*01:03** ($p_{BY} = 1.40 \times 10^{-71}$, $BMA_{postOR} = 0.63$, 95% CI = [0.49, 0.81]) are all associated. Of note, we find new HLA associations to anemias, with which HLA alleles have not previously been associated with, except for the rare case of aplastic anemia (Latin America, France, Turkey, Brazil, and Barcelona have reported prevalences of only 1.6, 1.5, 0.64, 2.4, and 2.5 per

million for the disease)[34]. We find **B*08:01** (p_{BY} = 8.71 x 10^{-4}, BMA_{postOR} = 1.2, 95% CI = [1.11, 1.3]) and **B*39:06** (p_{BY} = 7.66 x 10^{-3}, BMA_{postOR} = 1.95, 95% CI = [1.49, 2.56]) to be associated with **vitamin B12 deficiency anemia** risk with moderate effects. In contrast, **DQA1*01:01** protects against **vitamin B12 deficiency anemia** (p_{BY} = 8.41 x 10^{-7}, BMA_{postOR} = 0.76, 95% CI = [0.68, 0.83]) and **pernicious anemia** as well (p_{BY} = 7.20 x 10^{-3}, BMA_{postOR} = 0.73, 95% CI = [0.63, 0.83]).

Several other associations are intriguing. The only previous association between the HLA region and acne is the discovery of the DPB1*04:02 allele itself while HLA-typing a patient with acne vulgaris[35]. Yet, we find a weak protective effect of **B*08:01 on acne** (p_{BY} = 4.21 x 10^{-3}, BMA_{postOR} = 0.85, 95% CI = [0.79, 0.91]). This allele is additionally found to be associated with multiple systemic autoimmune disorders, such as **connective tissue disorder** (p_{BY} = 3.45 x 10^{-10}, BMA_{postOR} = 1.32, 95% CI = [1.23, 1.42]), **sarcoidosis** (p_{BY} = 2.04 x 10^{-33}, BMA_{postOR} = 1.58, 95% CI = [1.36, 1.84]), **Sjogren's syndrome** (p_{BY} = 1.52 x 10^{-15}, BMA_{postOR} = 1.83, 95% CI = [1.61, 2.09]), and **systemic lupus erythematosus** (p_{BY} = 9.50 x 10^{-12}, BMA_{postOR} = 1.79, 95% CI = [1.55, 2.06]), suggesting diverse and widespread involvement in various immune and bodily functions (**Supplementary Table 1**). Additionally, we find that **B*57:01** (notably associated with a spike in Alanine aminotransferase levels, among other adverse drug reactions, to the anti-retroviral abacavir[36,37]) to be associated with **calculus of kidney and ureter**.

Associations between HLA and cancers are well-established, and here we recover several, including between **DQB1*05:01** and **DRB3*99:01** and **cervical cancer** and **DQA1*03:01** and **female genital tract cancer**. Furthermore, we find **DQB1*05:01** to be associated with **skin cancer**, **non-melanoma skin cancer**, and **non-hodgkin's lymphoma**. In addition, we identify several novel associations between skin cancer, non-melanoma skin cancer, and HLA allelotypes. Statistics on all cancer associations (as well as their novelty) are described in **Table 1**. Taken together, by applying BMA, we recover known associations as well as identify novel disease associations with HLA alleles. These additive associations could add to elucidation of disease risk across a multitude of complex traits.

HLA Allele	**Phenotype**	**BY-adjusted *p*-value**	**BMA OR [95% CI]**	**Novel?**
DQB1*05:01	cervical cancer	4.51 x 10^{-4}	0.76 [0.69, 0.83]	N[38]
DRB3*99:01	cervical cancer	2.21 x 10^{-3}	1.19 [1.13, 1.27]	N[39]
DQA1*03:01	female genital tract cancer	4.77 x 10^{-5}	1.16 [1.1, 1.22]	N[40]
DQB1*05:01	skin cancer	2.12 x 10^{-15}	1.11 [1.08, 1.15]	N[23]
DQB1*05:01	non-melanoma skin cancer	3.27 x 10^{-17}	1.13 [1.09, 1.17]	N[23]
DQB1*05:01	non--hodgkin's lymphoma	3.68 x 10^{-3}	1.31 [1.19, 1.45]	N[41]
B*44:03	skin cancer	3.65 x 10^{-5}	0.89 [0.85, 0.93]	Y

DQB1*03:02	skin cancer	2.69×10^{-8}	0.91 [0.87, 0.95]	Y
DQB1*03:02	non-melanoma skin cancer	3.46×10^{-8}	0.92 [0.88, 0.95]	Y
B*40:01	non-melanoma skin cancer	4.11×10^{-5}	0.89 [0.85, 0.94]	Y
B*44:03	non-melanoma skin cancer	5.32×10^{-5}	0.88 [0.84, 0.93]	Y

Table 1. Cancer associations found by BMA (posterior probability >= 0.8).

Figure 2A. Overview of BMA results. a) Direction and magnitude of HLA allele effects on phenotypes. x-axis indicates allelotype, y-axis indicates phenotype. Red shows risk effects; blue shows protective effects. A systematic literature search was conducted, yielding a set of previously reported and novel associations (**Supplementary Table 1**). Marker size indicates novelty of discovery - small (previously found) or large (novel). y-axis was clustered by mean effects of alleles on phenotypes.

Figure 2B. Spider plot showing BMA associations between HLA alleles and binary phenotypes. The phenotypes were grouped by category as in the legend. A line was drawn between an allele and a phenotype if the allele had a BMA posterior probability ≥ 0.8 for that phenotype. Novel associations (as determined in **Supplementary Table 1**) are shown as solid, opaque lines; previously-found associations are indicated as dashed, faded lines. Line width scales with BMA posterior mean effect size.

Figure 3. Non-additive dosage model associations. (A) Allele-phenotype pair. Light grey text represents a previously-found non-additive effect (**Supplementary Table 1**). **(B) Number of UK Biobank cases for the phenotype** x-axis is $\log_{10}$ scaled. **(C) Odds ratio and log odds ratio for additive (green) and genotypic model (blue, dosage 1 and yellow, dosage 2).** Graphical measure of model fit. **(D) ΔAIC (AIC additive model - AIC genotype model).** A more negative value represents a larger departure from additivity.

A
Global Biobank Engine Phenotype
HLA Allelotype
BMA_posterior_mean
0.4
0.0
-0.4
Novel Finding
N
Y

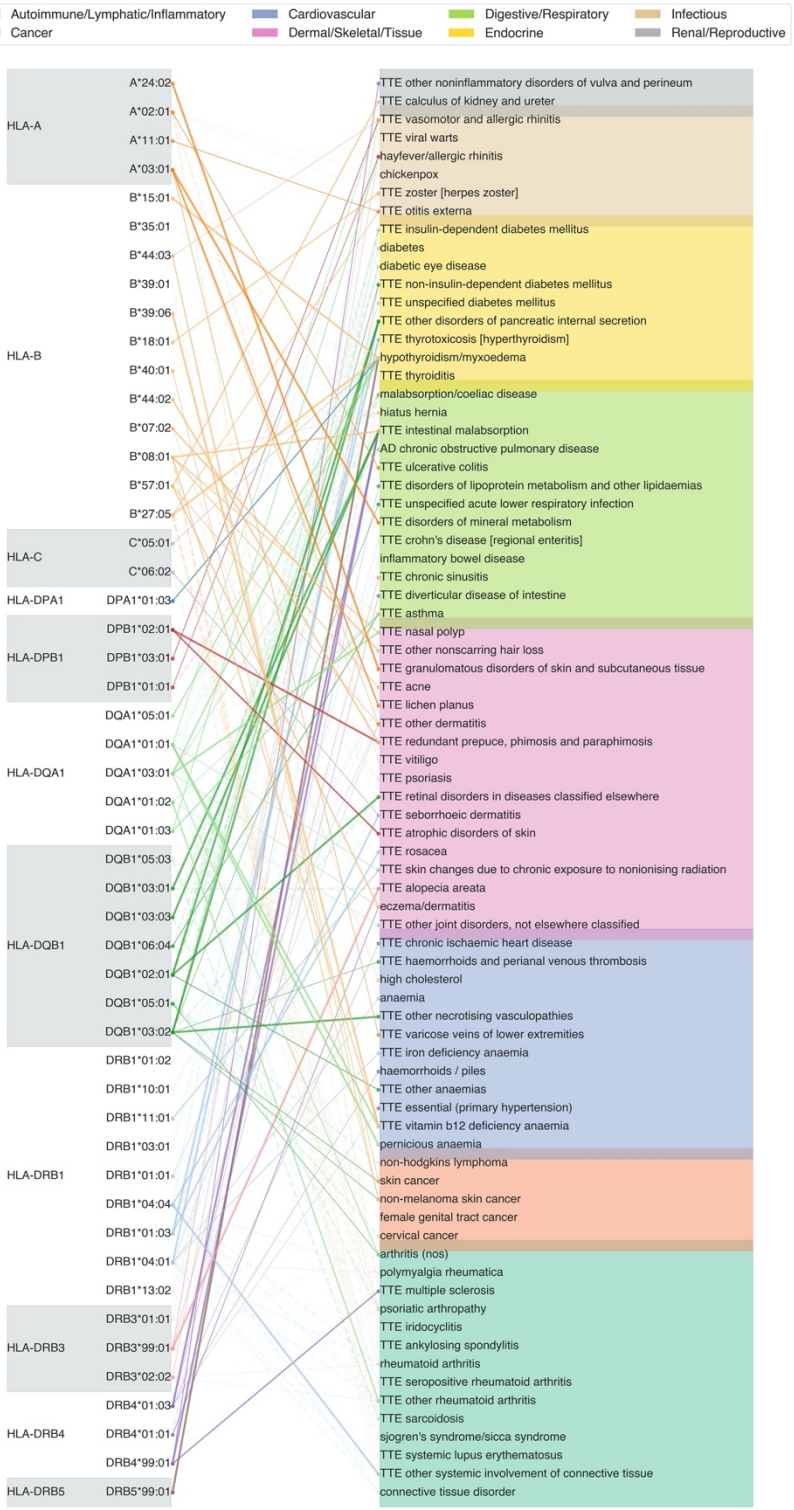
B
Autoimmune/Lymphatic/Inflammatory
Cancer
Cardiovascular
Dermal/Skeletal/Tissue
Digestive/Respiratory
Endocrine
Infectious
Renal/Reproductive
HLA-A
HLA-B
HLA-C
HLA-DPA1
HLA-DPB1
HLA-DQA1
HLA-DQB1
HLA-DRB1
HLA-DRB3
HLA-DRB4
HLA-DRB5

Figure 2

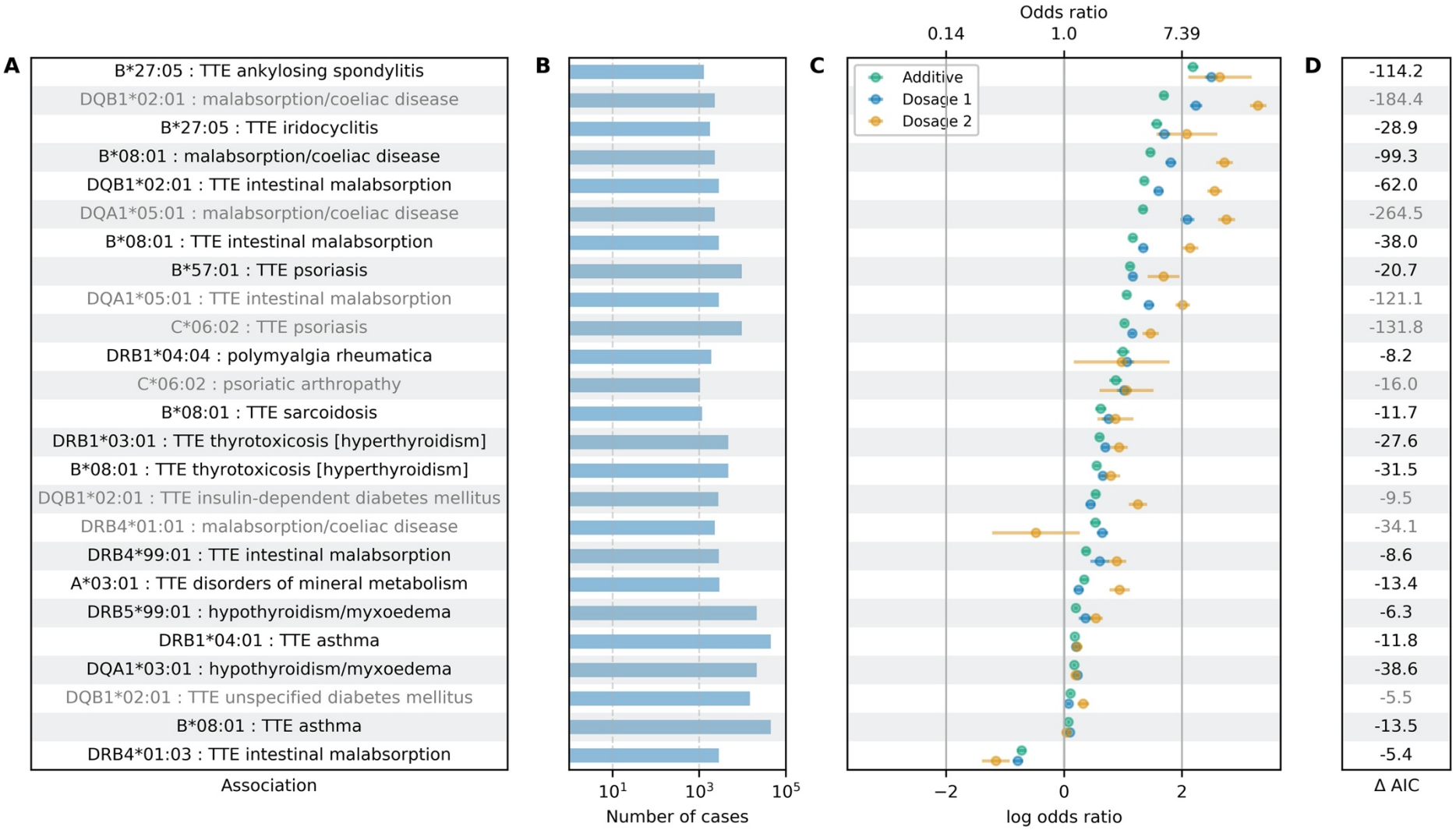
A
B
C
D
Odds ratio
0.14
1.0
7.39
Additive
Dosage 1
Dosage 2
B*27:05 : TTE ankylosing spondylitis
DQB1*02:01 : malabsorption/coeliac disease
B*27:05 : TTE iridocyclitis
B*08:01 : malabsorption/coeliac disease
DQB1*02:01 : TTE intestinal malabsorption
DQA1*05:01 : malabsorption/coeliac disease
B*08:01 : TTE intestinal malabsorption
B*57:01 : TTE psoriasis
DQA1*05:01 : TTE intestinal malabsorption
C*06:02 : TTE psoriasis
DRB1*04:04 : polymyalgia rheumatica
C*06:02 : psoriatic arthropathy
B*08:01 : TTE sarcoidosis
DRB1*03:01 : TTE thyrotoxicosis [hyperthyroidism]
B*08:01 : TTE thyrotoxicosis [hyperthyroidism]
DQB1*02:01 : TTE insulin-dependent diabetes mellitus
DRB4*01:01 : malabsorption/coeliac disease
DRB4*99:01 : TTE intestinal malabsorption
A*03:01 : TTE disorders of mineral metabolism
DRB5*99:01 : hypothyroidism/myxoedema
DRB1*04:01 : TTE asthma
DQA1*03:01 : hypothyroidism/myxoedema
DQB1*02:01 : TTE unspecified diabetes mellitus
B*08:01 : TTE asthma
DRB4*01:03 : TTE intestinal malabsorption
Association
10^1
10^3
10^5
Number of cases
−2
0
2
log odds ratio
-114.2
-184.4
-28.9
-99.3
-62.0
-264.5
-38.0
-20.7
-121.1
-131.8
-8.2
-16.0
-11.7
-27.6
-31.5
-9.5
-34.1
-8.6
-13.4
-6.3
-11.8
-38.6
-5.5
-13.5
-5.4
Δ AIC

Figure 3

Analysis of non-additive genetic effects

Several analyses have looked at the evidence of non-additive effects within the HLA region of the genome[15]; thus, we examined whether non-additive effects among the 178 associations identified in the BMA analysis contribute to disease risk. Where ΔAIC was defined as the Akaike Information Criterion goodness-of-fit[22] difference between the additive (standard GWAS using numerical values) model and the non-additive/"genotypic" model (in which rounded genotype dosage is considered as a factor), we selected those associations where the ΔAIC was <= -5, i.e. where the genotypic model fit was significantly better than that of the additive model.

We found 25 associations which met these criteria, of which 8 were previously described[15] (**Figure 3**). There is an enrichment for malabsorption (9) thyroid (4), and autoimmune (4) phenotypes, but psoriatic (3), asthma (2), and generic diabetes (2) phenotypes also feature multiple non-additive associations. We define heterozygote odds ratio (OR) to be the OR associated with Dosage 1 and homozygote OR to be the OR associated with Dosage 2, as in **Figure 3**. We replicate a known non-additive effect[15] of **C*06:02** on **psoriasis** (additive OR 2.79, 95% CI [2.69, 2.89]; heterozygote OR 3.19, 95% CI [3.05, 3.33]; homozygote OR 4.35, 95% CI [3.79, 4.99]). In addition, we identify multiple novel associations, such as **B*27:05** and **ankylosing spondylitis** (additive OR 8.93, 95% CI [8.13, 9.81]; heterozygote OR 12.24, 95% CI [10.94, 13.68]; homozygote OR 14.11, 95% CI [8.24, 24.18]); low homozygote counts result in a large CI for this homozygote OR. In relation to the association between **DRB5*99:01** and **hypothyroidism/myxedema** (additive OR 1.22, 95% CI [1.19, 1.26]; heterozygote OR 1.44, 95% CI [1.28, 1.62]; homozygote OR 1.72, 95% CI [1.53, 1.93]), the heterozygote and homozygote CIs overlap with each other but not with the additive CI. We see a significantly higher homozygote OR for **B*57:01** and **psoriasis** (additive OR 3.07, 95% CI [2.93, 3.23]; heterozygote OR 3.22, 95% CI [3.06, 3.39]; homozygote OR 5.42, 95% CI [4.14, 7.09]), representing another type of departure from additivity. Other non-additive associations have CIs that all overlap with each other, e.g. **B*27:05** and **iridocyclitis** (additive OR 4.83, 95% CI [4.4, 5.29]; heterozygote OR 5.5, 95% CI [4.96, 6.09]; homozygote OR 8.05, 95% CI [4.8, 13.51]).

Perhaps the most dramatic of these is the effect of **DRB4*01:01 on celiac disease** (additive OR 1.71, 95% CI [1.56, 1.87]; heterozygote OR 1.92, 95% CI [1.74, 2.12]; homozygote OR 0.62, 95% CI [0.3, 1.31]). The heterozygote and additive effect estimates are somewhat opposite the homozygous estimate, albeit the CI of the homozygote OR crossing 0 (**Figure 3**). However, 12 previously-known moderate non-additive associations[15] was not replicated in our analyses. These range in ΔAIC from -3.368 to +1.99 (**Supplementary Table 2**). Overall, the underlying biology of these non-additive associations could be important to disease pathogenesis.

Discussion

This study is a comprehensive overview of associations of 156 HLA alleles to 677 phenotypes in 337,138 individuals in the UK Biobank. Using single-allele association analysis and subsequent

BMA, we replicate 88 known associations between HLA alleles and cancer and other binary phenotypes and discover 90 novel associations; many of the novel associations feature phenotypes close to known associations (e.g. intestinal malabsorption and celiac disease, hypo- and hyperthyroidism), but some introduce associations that are not known (e.g. anemias and acne). For example, we find that DRB1*04:04 is associated with a family of autoimmune disorders such as rheumatoid arthritis, polymyalgia rheumatica, and connective tissue disorder, replicating signals found previously in the literature. We add to the body of known HLA associations with strong, novel effects such as that of B*39:06 on vitamin B12 deficiency anemia. Most importantly, we find several novel associations to skin and reproductive tract cancers that could prove impactful.

These genotype-phenotype associations associate the HLA region of the genome to many diverse phenotypes and suggest avenues for uncovering relevant, novel biology and developing therapeutics; however, there are many opportunities for future studies to extend this analysis, and avenues for improvement. Given that BMA is exponential in complexity with respect to the number of alleles analyzed, it is computationally infeasible to run the analysis on all BY-significant alleles for certain phenotypes. As such, it is likely that the 178 allele-phenotype associations we find are actually conservative, highlighting the impact of the HLA on disease risk. Additionally, the non-additivity analysis may be underpowered to determine homozygote ORs in some cases where those genotypes are low in number. It is surprising that we do not find any non-additive effects consistent with a recessive-only model, i.e. where the heterozygote OR overlaps 1. It is also unclear whether or not these non-additive effects can be classified as truly dominant, recessive, or co-dominant effects without additional formal model comparisons.

HLA alleles are oftentimes inherited as haplotypes that group several alleles in LD with each other. Better HLA typing could enable accurate haplotype frequency estimation and subsequent haplotype-based tests that enable testing of other targeted hypotheses of HLA diversity. Interactions between HLA allelotypes are also not studied here and would be a logical extension of this work; examples in which interaction effects are responsible for disease risk include the effect of an HLA-B/ERAP1 interaction on ankylosing spondylitis[42] and HLA-C/ERAP1 on psoriasis[43]. One of the functions of ERAP1 is to “trim” peptides before they are presented on HLA class I proteins. These studies therefore suggest that the risk effect mediated by certain HLA alleles also depends on peptide pre-processing. As better reference panels, superior imputation techniques, and larger population biobanks with more sequence data are developed, the HLA region will continue to grow in importance in human genetics, and down the line, as effective therapeutic hypotheses are generated, in clinical settings.

Author Contributions

M.A.R. conceived and designed the study. G.R.V., J.E.O., and C.D. analyzed the data. G.R.V., Y.T., and J.M.J. provided quality control analysis. The manuscript was written by G.R.V. and M.A.R. and revised by all the co-authors. All co-authors have approved of the final version of the manuscript.

Acknowledgments and Funding

This research has been conducted using the UK Biobank Resource under Application Number 24983, “Generating effective therapeutic hypotheses from genomic and hospital linkage data” (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf). We thank all of the participants in the UK Biobank study. This work was supported by the National Human Genome Research Institute (NHGRI) of the National Institutes of Health (NIH) under award R01HG010140. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. G.R.V. is supported by the National Library of Medicine (NLM) T15 Continuing Education Training Grant. J.M.J. is funded by grant NNF17OC0025806 from the Novo Nordisk Foundation and the Stanford Bio-X Program. M.A.R. is supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080).

Supplementary Materials

Supplementary Table 1. Results of systematic literature review. Allele-phenotype pairs were marked as “novel” if not found in the literature and “non-additive” if the allele was found to have a non-additive effect on the phenotype. PMIDs and URLs for all references were provided.

https://bit.ly/hla_sup_1

Supplementary Table 2. Significant BMA results table. Results from PLINK additive association analysis, BMA analysis, and non-additivity analysis for each allele-phenotype pair crossing BMA posterior probability 0.8. Number of cases per phenotype and allele frequencies per allele in the UK Biobank white British cohort also provided.

https://bit.ly/hla_sup_2

References

1. Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., et al. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat. Biotechnol. *33*, 1152–1158.

2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). T Cells and MHC Proteins (Garland Science).

3. Raychaudhuri, S., Sandor, C., Stahl, E.A., Freudenberg, J., Lee, H.-S., Jia, X., Alfredsson, L., Padyukov, L., Klareskog, L., Worthington, J., et al. (2012). Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. Nat. Genet. *44*, 291–296.

4. De Re, V., Caggiari, L., Simula, M.P., De Vita, S., Mazzaro, C., Lenzi, M., Massimo, G.M., Monti, G., Ferri, C., Zignego, A.L., et al. (2007). Role of the HLA class II: HCV-related disorders. Ann. N. Y. Acad. Sci. *1107*, 308–318.

5. Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y., and Hinds, D.A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat. Commun. *8*, 599.

6. Crosslin, D.R., Carrell, D.S., Burt, A., Kim, D.S., Underwood, J.G., Hanna, D.S., Comstock, B.A., Baldwin, E., de Andrade, M., Kullo, I.J., et al. (2015). Genetic variation in the HLA region is associated with susceptibility to herpes zoster. Genes Immun. *16*, 1–7.

7. Mbarek, H., Ochi, H., Urabe, Y., Kumar, V., Kubo, M., Hosono, N., Takahashi, A., Kamatani, Y., Miki, D., Abe, H., et al. (2011). A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population. Hum. Mol. Genet. *20*, 3884–3892.

8. Rubicz, R., Yolken, R., Drigalenko, E., Carless, M.A., Dyer, T.D., Bauman, L., Melton, P.E., Kent, J.W., Jr, Harley, J.B., Curran, J.E., et al. (2013). A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). PLoS Genet. *9*, e1003147.

9. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.

10. Pappas, D.J., Lizee, A., Paunic, V., Beutner, K.R., Motyer, A., Vukcevic, D., Leslie, S., Biesiada, J., Meller, J., Taylor, K.D., et al. (2018). Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J. *18*, 367–376.

11. Kennedy, A.E., Ozbek, U., and Dorak, M.T. (2017). What has GWAS done for HLA and disease associations? Int. J. Immunogenet. *44*, 195–211.

12. Jeanmougin, M., Noirel, J., Coulonges, C., and Zagury, J.-F. (2017). HLA-check: evaluating HLA data from SNP information. BMC Bioinformatics *18*, 334.

13. Dilthey, A., Leslie, S., Moutsianas, L., Shen, J., Cox, C., Nelson, M.R., and McVean, G.

(2013). Multi-population classical HLA type imputation. PLoS Comput. Biol. *9*, e1002877.

14. Degenhardt, F., Wendorff, M., Wittig, M., Ellinghaus, E., Datta, L.W., Schembri, J., Ng, S.C., Rosati, E., Hübenthal, M., Ellinghaus, D., et al. (2019). Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. Hum. Mol. Genet. *28*, 2078–2092.

15. Lenz, T.L., Deutsch, A.J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T.W.J., Abecasis, G., et al. (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat. Genet. *47*, 1085–1090.

16. DeBoever, C., Tanigawa, Y., Lindholm, M.E., McInnes, G., Lavertu, A., Ingelsson, E., Chang, C., Ashley, E.A., Bustamante, C.D., Daly, M.J., et al. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. Nat. Commun. *9*, 1612.

17. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

18. Firth, D. (1993). Bias Reduction of Maximum Likelihood Estimates. Biometrika *80*, 27–38.

19. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

20. Benjamini, Y., and Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann. Stat. *29*, 1165–1188.

21. Raftery, A.E., Madigan, D., and Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models. J. Am. Stat. Assoc. *92*, 179–191.

22. Akaike, H. (2011). Akaike's Information Criterion. International Encyclopedia of Statistical Science 25–25.

23. Wang, W., Ollila, H.M., Whittemore, A.S., Demehri, S., Ioannidis, N.M., Jorgenson, E., Mignot, E., and Asgari, M.M. (2018). Genetic variants in the HLA class II region associated with risk of cutaneous squamous cell carcinoma. Cancer Immunol. Immunother. *67*, 1123–1133.

24. Picascia, S., Sidney, J., Camarca, A., Mazzarella, G., Giardullo, N., Greco, L., Auricchio, R., Auricchio, S., Troncone, R., Sette, A., et al. (2017). Gliadin-Specific CD8 T Cell Responses Restricted by HLA Class I A*0101 and B*0801 Molecules in Celiac Disease Patients. The Journal of Immunology *198*, 1838–1845.

25. Martín-Masot, R., Nestares, M.T., Diaz-Castro, J., López-Aliaga, I., Alférez, M.J.M., Moreno-Fernandez, J., and Maldonado, J. (2019). Multifactorial Etiology of Anemia in Celiac Disease and Effect of Gluten-Free Diet: A Comprehensive Review. Nutrients *11*,.

26. Mehraji, Z., Farazmand, A., Esteghamati, A., Noshad, S., Sadr, M., Amirzargar, S., Yekaninejad, M.S., and Amirzargar, A. (2017). Association of Human Leukocyte Antigens Class I & II with Graves' Disease in Iranian Population. Iran. J. Immunol. *14*, 223–230.

27. Valdés-Corona, L.F., Hernández-Doño, S., Rodríguez-Reyna, T.S., García-Silva, R., Jakez,

J., Escamilla-Tilch, M., Lima, G., Llorente, L., Pineda, C., Yunis, E., et al. (2020). Aspartic acid70 in the HLA-DRB1 chain and Shared Epitope alleles partially explain the high prevalence of autoimmunity in Mexicans. Journal of Translational Autoimmunity 100057.

28. Ueda, S., Oryoji, D., Yamamoto, K., Noh, J.Y., Okamura, K., Noda, M., Kashiwase, K., Kosuga, Y., Sekiya, K., Inoue, K., et al. (2014). Identification of independent susceptible and protective HLA alleles in Japanese autoimmune thyroid disease and their epistasis. J. Clin. Endocrinol. Metab. *99*, E379–E383.

29. Macel, L.M.Z., Rodrigues, S.S., Dibbern, R.S., Navarro, P.A.A., and Donadi, E.A. (2001). Association of the HLA-DRB1*0301 and HLA-DQA1*0501 Alleles with Graves' Disease in a Population Representing the Gene Contribution from Several Ethnic Backgrounds. Thyroid *11*, 31–35.

30. Chen, B., Li, J., He, C., Li, D., Tong, W., Zou, Y., and Xu, W. (2017). Role of HLA-B27 in the pathogenesis of ankylosing spondylitis (Review). Mol. Med. Rep. *15*, 1943–1951.

31. Li, H., Li, Q., Ji, C., and Gu, J. (2018). Ankylosing Spondylitis Patients with HLA-B*2704 have More Uveitis than Patients with HLA-B*2705 in a North Chinese Population. Ocul. Immunol. Inflamm. *26*, 65–69.

32. FitzGerald, O., Haroon, M., Giles, J.T., and Winchester, R. (2015). Concepts of pathogenesis in psoriatic arthritis: genotype determines clinical phenotype. Arthritis Res. Ther. *17*, 115.

33. Yen, J.H., Chen, C.J., Tsai, W.C., Ou, T.T., Lin, C.H., Lin, S.C., and Liu, H.W. (2001). HLA-DQA1 genotyping in patients with rheumatoid arthritis in Taiwan. Kaohsiung J. Med. Sci. *17*, 183–189.

34. Dhaliwal, J.S., Wong, L., Kamaluddin, M.A., Yin, L.Y., and Murad, S. (2011). Susceptibility to aplastic anemia is associated with HLA-DRB1*1501 in an aboriginal population in Sabah, Malaysia. Hum. Immunol. *72*, 889–892.

35. Witter, K., Kirchner, E., Borelli, C., Messer, G., Albert, T., Zahn, R., and Kauke, T. (2009). In a study for acne vulgaris, sequence-based HLA typing showed a novel DPB1 allele, DPB1*2402. Tissue Antigens *74*, 354–356.

36. Mounzer, K., Hsu, R., Fusco, J.S., Brunet, L., Henegar, C.E., Vannappagari, V., Stainsby, C.M., Shaefer, M.S., Ragone, L., and Fusco, G.P. (2019). HLA-B*57:01 screening and hypersensitivity reaction to abacavir between 1999 and 2016 in the OPERA® observational database: a cohort study. AIDS Res. Ther. *16*, 1.

37. Cardone, M., Garcia, K., Tilahun, M.E., Boyd, L.F., Gebreyohannes, S., Yano, M., Roderiquez, G., Akue, A.D., Juengst, L., Mattson, E., et al. (2018). A transgenic mouse model for HLA-B*57:01–linked abacavir drug tolerance and reactivity. Journal of Clinical Investigation *128*, 2819–2832.

38. Gregoire, L., Lawrence, W.D., Kukuruga, D., Eisenbrey, A.B., and Lancaster, W.D. (1994). Association between HLA-DQB1 alleles and risk for cervical cancer in African-American women. Int. J. Cancer *57*, 504–507.

39. Bao, X., Hanson, A.L., Madeleine, M.M., Wang, S.S., Schwartz, S.M., Newell, F.,

Pettersson-Kymmer, U., Hemminki, K., Tiews, S., Steinberg, W., et al. (2018). HLA and KIR Associations of Cervical Neoplasia. J. Infect. Dis. *218*, 2006–2015.

40. Ferrera, A., Olivo, A., Alaez, C., Melchers, W.J., and Gorodezky, C. (1999). HLA DOA1 and DOB1 loci in Honduran women with cervical dysplasia and invasive cervical carcinoma and their relationship to human papillomavirus infection. Hum. Biol. *71*, 367–379.

41. Skibola, C.F., Akers, N.K., Conde, L., Ladner, M., Hawbecker, S.K., Cohen, F., Ribas, F., Erlich, H.A., Goodridge, D., Trachtenberg, E.A., et al. (2012). Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. Tissue Antigens *79*, 279–286.

42. Evans, D.M., Spencer, C.C.A., Pointon, J.J., Su, Z., Harvey, D., Kochan, G., Oppermann, U., Dilthey, A., Pirinen, M., Stone, M.A., et al. (2011). Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat. Genet. *43*, 761–767.

43. Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, Strange, A., Capon, F., Spencer, C.C.A., Knight, J., Weale, M.E., Allen, M.H., Barton, A., Band, G., Bellenguez, C., et al. (2010). A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat. Genet. *42*, 985–990.