UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Technology

Arne Kukkonen

Genome-Wide Association Study for Detecting Autoimmune-Disease-Associated Genetic Pattern Differences in Specific HLA Type Carriers

Bachelor's thesis (12 ECTS)

Curriculum Science and Technology

Supervisor:
Erik Abner, PhD

Tartu 2023

Genome-wide Association Study for Detecting Autoimmune-Disease-Associated Genetic Pattern Differences in Specific HLA Type Carriers

Abstract:

The HLA locus variants are one of the strongest genetic predictors for most, if not all, human autoimmune diseases. The HLA locus genes include the antigen-presenting cell surface peptide encoding genes, which form an essential component in the maturation of the T-cell population in the thymus, and their subsequent activation in the periphery.

Leveraging the modern population-wide genotype information that capture even the most polymorphic loci, this work sets the aim to design a case-control genome-wide association study (GWAS), that would result in the detection of non-HLA genetic variants that have a statistically different effect on an autoimmune disease in the carriers of certain HLA types, in comparison to the non-carriers. For the purpose of this aim, study groups are assembled based on specific HLA allele doses, so that for 42 HLA allele types selected for this study there are 42 HLA-specific groups where every individual is a carrier of at least one copy of the HLA allele type. The effect sizes from the summary statistics of the HLA-specific GWASs are compared to a general population GWAS (which is done on all the participants of the Estonian Biobank in this case). The variants are considered relevant to this aim if their effect size is statisticallt different in the HLA-specific groups than they are in the general population GWAS.

Keywords:

Genome-wide association study, human leukocyte antigens, epistasis, autoimmune diseases

Geneetiline assotsiatsiooniuuring tuvastamaks HLA-dega kooslevaid geene autoimmuunsetes haigustes

Lühikokkuvõte:

Inimese leukotsüüdi antigeeni (HLA) lookuse geenivariatsioon moodustab ühe tugevaima geneetilise teguri enamikes, kui mitte kõigis, teadaolevates autoimmuunsetes haigustes. HLA lookus hõlmab endas geene, mis toodavad antigeene-esitlevaid pinnavalke, mis moodustavad olulise komponendi T-raku populatsiooni kujunemisel tüümuses ja/või nende aktiveerimisel perifeerias.

Kasutades moodsat ja täpset genotüübi-informatsiooni, millel on suutlikus kaardistada isegi kõige muutlikemaid lookuseid, on antud uurimistöö eesmärgiks disainida geneetiline assotsiatsiooniuuring (GWAS), millel oleks võime tuvastada geenivariante, mille mõju autoimmuunhaiguse eelsoodumusele oleks oluliselt teistsugune kindla HLA-tüübi kandjates, võrreldes nende HLAde mittekandjatega. Ned tulemused oleksid viitavad võimalikule epistaasile HLA ja mitte-HLA geenide vahel, ja käesoleva uuringuga tuvastatud variante saab käsitleda prioriteetsetena edasistes uuringutes, millel on võime täpsemalt käsitleda geen-geen interaktsioonide olemasolu.

Selle töö põhimõte seisneb selekteeritud HLA-tüüpide alleelidooside alusel koostatud rühmades, selliselt, et 42-le selle uuringu jaoks valitud HLA alleeli tüübile vastab 42 HLA-spetsiifilist rühma, kus iga inimene on vähemalt ühe alleelikoopia kandja. HLA-spetsiifiliste rühmadega tehtud GWASidest saadud variantide mõjude väärtuseid võrreldakse üldpopulatsiooni esindavast GWASist (tehtud kõigi Eesti geenivaramus osalejatega) pärinevate mõjude väärtustega. Potensiaalse epistaasi vaatenurgast on antud projekti raames huvipakkuvad need lookused, millel on statistiliselt oluliselt erinev seos haigusega HLA-spetsiifilises rühmas.

Võtmesõnad:
Geneetiline assotsiatsiooniuuring, inimese leukotsüüdi antigeenid (HLA), epistaas, autoimmuunsed haigused

TABLE OF CONTENTS

INTRODUCTION

The human leukocyte antigen (HLA) system contains a series of polymorphic loci on the short arm of chromosome 6 that encode for surface proteins that have the function in the human adaptive immune system of presenting antigenic peptide fragments to the T-cell receptors (TCRs) [1]. The HLA correlation with autoimmune diseases has been published in the scientific literature since the 1970' [2] [3], yet, to this day, the exact molecular mechanisms that would provide a causal explanation remain elusive for many of the HLA-associated diseases. In the past, the efforts to investigate the HLA role in disease pathologies have been challenged by the dense clustering of the genes in the MHC locus, strong linkage disequilibrium and high polymorphism within the alleles [4]. The common feature across all HLA-associated autoimmune diseases is tissue damage by autoreactive T-cells, and with the advances in immunology and genomics, several hypotheses have emerged that could explain the generation of effector T-cells against self-motifs. Based on the review of the relevant literature, it is plausible to speculate that mutations in HLA or in other protein-encoding genes may lead to epistatic interactions in the form of TCR-self-peptide-HLA binding that results in the generation of autoreactive T-cells. The present research leverages the high-quality genomic data for the Estonian population in the Estonian Biobank (EstBB) to search for variants in protein-coding genomic regions that have significantly different association with autoimmune diseases in specific HLA type carriers compared to non-carriers, which would be suggestive of the presence of potential gene-gene interactions between HLA types and non-HLA peptides.

1 LITERATURE REVIEW

1.1 The Human Leukocyte Antigens (HLAs) & The Adaptive Immune System

The function of the HLA molecules in the adaptive immune system is to present peptide fragments, obtained from the degradation of proteins in the cytoplasm, on the surface of all cells in the body (with the exception of central nervous system cells) to the TCR of T-cells. The HLA molecules can be viewed as the means by which the adaptive immune system scans the proteome of the organism to maintain a library of epitopes that are either labeled as self – and thus must be subject to protection; or pathogenic – which need to be eliminated (along with the cells that contain them) [5]. All HLA molecules share the essential functional features of having a peptide-presenting cleft structure, into which a peptide cargo is non-covalently loaded in the cytoplasm, before the HLA-peptide complex is transported to the cell surface and presented to the T-cell receptor (TCR). The HLA system has two classes of HLA peptides: the HLA class 1 peptides are expressed on the surface of all nucleated cells and platelets (except central nervous system cells) and present peptides derived from the proteins in the cytosol to the $CD8^{+}$T cell TCR. The class 2 HLAs are expressed on antigen presenting cells (APCs), which include B lymphocytes, dendritic cells, macrophages, monocytes, Langerhans cells, endothelial cells, and thymic epithelial cells [6]; the class 2 HLAs present peptides derived from the proteins in the intracellular vesicles, such as internalized by phagocytosis or pathogens living in macrophage vesicles, to the $CD4^{+}$T cell TCR.

The 3.6 megabase HLA locus on the short arm of chromosme 6 contains about 200 genes [7], of which 40 encode the leukocyte antigens. The class 1 genes encode for the α chain of the class 1 HLA molecule; the class 1 β chain is encoded by the beta-2-microglobulin (*B2M*) gene on chromosome 15 [8]. The class 2 HLA genes encode for both the α and β chains of the HLA molecule, as designated by the third letter in the class 2 HLA name (A and B, for α and β, respectively; see section 1.1.1 for HLA nomenclature). The α chain of the class I molecule has two peptide-binding domains (α1 and α2), an immunoglobulin-like domain (α3), the transmembrane region (TM), and the cytoplasmic tail. Each of the class II α and β chains has four domains: the peptide-binding domain (α1 or β1), the immunoglobulin-like domain (α2 or β2), the transmembrane region (TM), and the cytoplasmic tail (Figure 1).

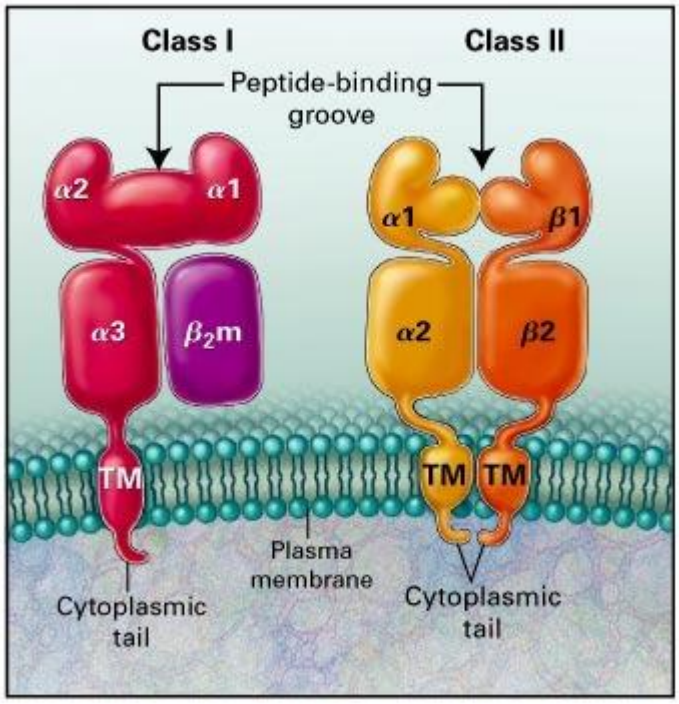

Figure 1 *Structure of class 1 and class 2 HLA molecules (figure obtained from Klein, J. and Sato, A., 2000 [8])*

The HLA molecule peptide-binding groove comprises an anti-parallel β strand floor and two α helices which presents a single protein fragment to the TCR (Figure 2). Most of the polymorphism in the HLA genes happen in defined locations within the HLA peptide-binding groove, and a single amino acid substitution has been shown via *in silico* experiments to change 20-30% of the peptide repertoire that the HLA can bind to [9]. This is in concordance with genomic association studies that have identified HLA risk alleles that differ by a single amino acid substitution. Both the presentation of alternative antigens as well as unconventional antigen orientations resulting from polymorphisms in the peptide binding groove have potential implications in the development of

autoimmune diseases (see section 1.2)

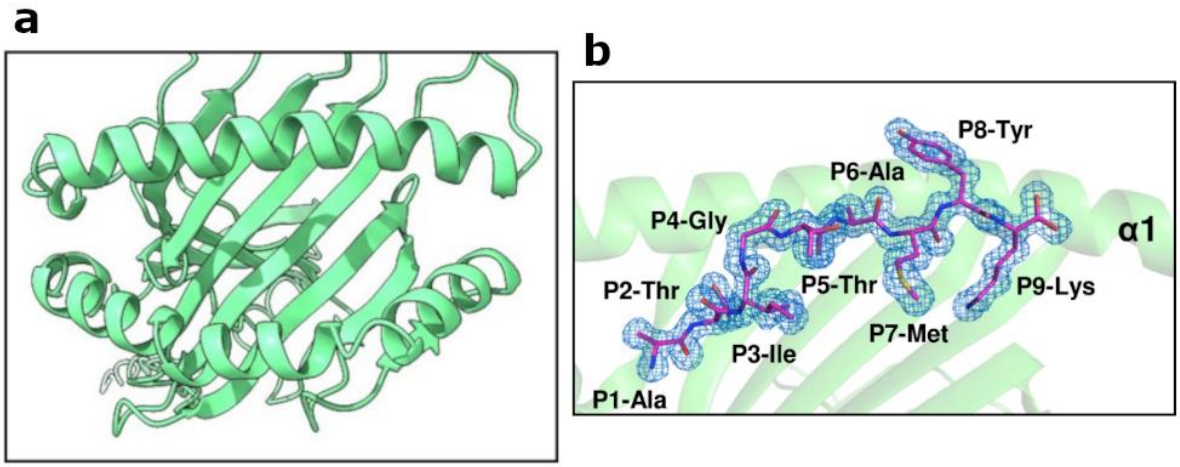

Figure 2 ***a)*** *The antigen-presenting cleft of HLA-A in top view (obtained from AlphaFold protein structure database; UniProt nr: A0A6N0A1S6).* ***b)*** *shows the atomic structure model of the Epstein Barr virus fragment ATIGTAMYK presented by HLA-A*11:01 (from Huan, et al., 2019 [10]; PDB ID: 6JOZ)*

In general terms, the two types of receptors in the adaptive immune system that can recognize antigens are the surface immunoglobulin of B cells and the TCR of T cells. The B cells recognize antigens that are outside the endogenous cells, such as extracellular bacteria, and B cells are primarily involved in the humoral immunity of the adaptive immune system and the release of antibodies in response to an antigen; the B cell receptors do not directly interact with the pHLA complexes. The TCR recognizes antigens generated inside the endogenous cells, which are presented to the TCR by HLA, such as virus peptides and peptides obtained by phagocytosis; T-cells are involved in the cell-mediated immunity – activating their function in response to the interaction with pHLA complexes [11]. The T-cell types are defined by their characteristic pattern of cytokine production and function. The three main types of T-cells are the cytotoxic T-cells ($CD8^+$T cells), the T-helper (Th) cells and the regulatory T-cells (T_{reg}). Upon $CD8^+$ TCR binding to a complementary pHLA complex, the $CD8^+$T cells activate to kill the target cell by the release of perforin and granzyme, that cause cell lysis; and granulysin, which signals the targeted cell to enter apoptosis [12]. The Th cells have no cytotoxic activity, but instead regulate the activity of other cells: macrophages, $CD8^+$T cells and the antibody-producting B cells. Th cells have an important role in the humoral response, as there are only few antigens that can directly stimulate

B-cells, and in most cases the B-cells rely on the signals from the Th cells to start proliferating and producing antibodies. T_{reg} cells are essential for the maintenance of peripheral tolerance and the prevention of autoimmune reactions. T_{reg} cells suppress immune responses by four modes of action: producing inhibitory cytokines; cytolysis of effector T-cells by granzyme A/B and perforin secretion; metabolic disruption of effector T-cells; and by modulating dendritic cell maturation and function [13].

The TCR comprises of two chains, TCRα and TCRβ, which are co-expressed with CD3 chains γ, δ, ε and ζ. TCRα and TCRβ each have a variable region and a constant region; the majority of the TCR interactions with the pHLA complex happens within the three complementarity-determining regions of the variable region of TCR. Upon TCR binding, the immunoreceptor tyrosine-based activation motifs (ITAMs) in CD3 chains are phosphorylated, triggering signaling cascades that result in T cell activation. The total number of possible unique TCRs is likely in the range of 10^{11}-10^{12}, and their affinities to target motifs are not very specific - each TCR has the potential to recognize close to a million different peptides [14]. This inherent cross-reactivity of T-cells allows a limited number of T-cells to provide comprehensive immune cover, but in case of malfunction (likely due to genetically predisposed complications in the negative thymic selection), this heterologous immunity is also observed to initiate autoimmunity in response to viral infection through a process called molecular mimicry, where the viral proteins share target motifs with self-peptides [15]. Epidemiological studies have identified viral infections as a significant risk factor in autoimmune diseases, and the molecular mimicry hypothesis has also been supported by evidence from animal studies (see section 1.2).

The T-cell repertoire is audited in the thymus' medulla and cortex, where the developing T-cells (thymocytes) undergo negative and positive selection, respectively. The negative selection removes unwanted T-cells, such as those with a potential to become autoreactive; and positive selection coordinates the generation of functional and self-tolerant T-cells. A crucial determinant in the thymocyte selection is the strength of the affinity between TCR and the self-pHLA complex on the APC. The thymocytes that do not interact with the APCs at all undergo death by neglect and a weak interaction between self-pHLA complex is required to promote the positive selection. APCs that are strong agonists to maturing T-cells, via self-pHLA-TCR interactions, either remove the autoreactive T-cells by apoptosis or divert the T-cell fate towards a regulatory T cell (T_{reg}). There seemingly exists a relatively wide window of TCR-pHLA affinity stregths that lead to the

generation of T_{reg} lineage, between the low and high affinity strengths that lead to the positive and negative selections, respectively [16].

1.1.1 HLA nomenclature

The current HLA nomenclature includes five standard fields (Figure 3). The first field represents the locus name; for class 2 HLAs the first field consists of three letters, beginning with letter D, and ending with a number; and for class 1 the first field consists of only one letter. In the first field of class 2 HLA name, after the letter D, the following two letters represent: the locus family (letters M, O, P, Q, or R), and whether the gene encodes the α or β chain (A or B, respectively). The four numeric fields, after the first field, following the asterisk delimiter are: 2) the allele or antigen group; 3) the specific amino acid sequence of the allele; 4) presence of synonymous polymorphisms; 5) and differences in non-coding regions. The numeric fields are delimited by a colon (":"). For example, HLA-DRB1*13:01 represents a class 2 (letter D), family R, β-chain (letter B) encoding gene, together forming the DRB1 locus (the DRB loci were named DRB1, DRB2, DRB3, DRB4 and DRB5 [17]), allele group 13 and a specific allele 01. HLA-DRB1*13:01:02 differs from HLA-DRB1*13:01 by a synonymous mutation. As a class 1 HLA name example, HLA-B*27 represents a class 1, B locus, α chain encoding gene (as all class 1 HLA genes only encode the α chain; the β chain of class 1 molecules comes from the *B2M* gene on chromosome 15), allele group 27. The nomenclature describes the typing of the HLA alleles in different levels of resolution. "Two-digit" typing provides the first field in the molecular-based nomenclature (example: HLA-A*01). "Four-digit" typing distinguishes alleles based on the sequence of the peptide-binding region of the HLA molecule, corresponding to the first two fields of the molecular nomenclature (example: HLA-B*57:01), and so on [18].

In the current project, the alleles were considered in two-digit resolution.

Figure 3 *The different fields used in standard HLA nomenclature. The yellow field indicates the gene name or locus, A in this example. The asterisk separator indicates that this is molecular typing. Field one (blue) gives the allele group, field two (pink) indicates the specific HLA protein, field three (red) synonymous polymorphisms, and field four (green) differences in non-coding regions. [18]*

Historically, HLA typing was done by serologic assays, where serum from sensitized patients, typically women who were sensitized during pregnancy to paternal HLA antigens expressed by their fetuses, was incubated with lymphocytes from various donors in the presence of a complement [18]. Other established typing methods before NGS technology were oligonucleotide probe hybridization, PCR amplification with sequence-specific primers. HLA genotyping is made challenging by the high polymorphism and the high gene density in the locus [19] [20], and these methods were only able to produce ambiguous results. With the advent of NGS technology, customizable high-density genotyping chips, and the advancement of bioinformatic tools, the resolution and efficiency of HLA typing has got to a level that enables robust genome-wide population studies using accurate population-specific reference panels [21].

1.2 HLA-dependent Autoimmune Disease Pathogenesis

Current evidence from the literature allows to consider the HLA-related autoimmune disease mechanisms in 6 broad categories (which are not mutually exclusive), as follows:

1) TCR-pHLA interactions that mediate the escape of T-cells from the negative thymic selection.

The failure of the negative thymic selection to eliminate autoreactive T-cells is a key aspect in the development of autoimmune diseases. As mentioned in section 1.1, weak interaction between TCR and self-pHLA is necessary to prevent thymocytes to die from neglect and subject them to positive selection. There is evidence that variants in HLA peptide-binding domain, in conjunction with specific TCR types, could alter the affinity of TCR to self-pHLA complexes, which could help T-

cells evade the thymic surveillance. Multiple sclerosis is a demyleating central nervous system disease where the immune system attacks the myelin basic protein (MBP). The HLA-DR15 haplotype (alleles HLA-DRB1*15:01 and HLA-DRB5*01:01) is found to be a significant risk factor in multiple sclerosis [22]. A transgenic mouse study expressing HLA-DRB15 together with TCR type Ob.1A12 concluded evidence for the pathological role of self-MBP presentation in multiple scerosis [23]. Molecular studies have shown low affinity and off-centre binding topology between TCR Ob.1A12 and the pHLA complex with reduced surface area, which is could allow the effector cells to escape negative thymic selection [24] [25]. In another discovered multiple sclerosis pathway, the TCR MS2-3C8 has an affinity towards MBP epitopes presented by HLA-DRB1*0401. In this case, the HLA-peptide complex itself is very unstable, but the strong binding of the TCR MS2-3C8 is thought to stabilize the complex, while still producing weak enough signal in the thymus to not evoke negative selection [26] [27].

In type 1 diabetes (T1D) it is found that the TCR type 1E6 mediates pancreatic beta cell killing via recognition of amino acids 15–24 of preproinsulin (PPI), presented by HLA-A*0201 [28]. Although it's been found that TCR 1E6 has a standard docking orientation with the HLA-A*0201-restricted PPI epitope, the affinity is extremely low, which could also lead to the evasion of the autoreactive T-cells from negative thymic selection [29].

2) TCR – pHLA interactions that result in cross-reactive T-cells.

In addition to the significant heritability in multiple sclerosis, epidemiological studies have identified a strong risk factor being the infection with Epstein-Barr virus [30] [31]. Investigations into the interplay between HLA alleles and TCR types provide insights to how genetic predisposition can lead to low-specificity effector T-cells that can become cross-reactive in response to the so-called molecular mimics – proteins from pathogens that share limited structural homology between self-peptides. The Hy.1B11 is a T-cell clone from multiple sclerosis patients that binds to the HLA-DQ1 restricted MBP on amino acid positions 85-99. However, a crystallography study has shown that the binding of Hy.1B11 to HLA-DQ1 is tilted and the interaction is limited to only three peptides on MBP, which renders the T-cells less specific to MBP and gives them the potential to react with a much wider repertoire of peptides [32]. Hy.1B11 has been shown to also react with HLA-DQ1-restricted herpesvirus, adenovirus and Pseudomonas peptides [33]. An in vitro binding assay involving the TCR clones Ob.1A12 and Ob.2F3 detected cross-reactivity between several microbial proteins and the HLA-DR2 restricted MBP [34]. An in

vivo study with humanized Ob TCR-DR2b transgenic mice found that multiple sclerosis-like symptoms were evoked with the M. avium peptide almost as severely as with the MBP peptide [35].

3) Mutations in HLA peptide-binding groove that lead to increased self-antigen presentation

Celiac disease is a chronic immune response to ingested gluten that results in intestinal tissue damage by autoreactive T-cells, with the primary antigen being the transglutaminase 2 enzyme that deamidates ingested gluten. The HLA-DQ2.5 haplotype (which comprises any alleles of HLA-DQA1*05 and HLA-DQB1*02) and the HLA-DQ8 haplotype (any allele of HLA-DQA1*03; and the HLA-DQB1*03:02 allele) are identified as the risk HLA genotypes in celiac disease [36]. Furthermore, all the T-cells derived from the intestine of celiac disease patients recognize the gluten restricted by DQ8 and DQ2 [37] [38] [39]. Meta-analyses have determined that the HLA risk variants together account for 25% of the celiac disease heritability in European populations, and the HLA-DQ2.5 heterodimer receptor is present in 90-95% of the patients. The DQ2 and DQ8 genotypes are strong negative predictors of celiac disease (their absence predicts the absence of the disease with nearly a 100% accuracy), but, due to the high prevalence of the haplotypes in the general population, the risk haplotypes are not strong positive predictors [40]. One particular HLA variant associated with celiac disease is a DQ8 with an absent canonical aspartic acid residue at position 57 in the β chain. A study demonstrated that this DQ8 variant has a preference to recruit TCRs which have a negative charge at position 3 of CDR3β loop, which could contribute to the pathology by stabilizing of the TCR-deamidated gluten-HLA complex [41].

Rheumatoid arthritis is a chronic joint inflammation that causes the progressive destruction and deformation of joints. Most of the rheumatoid arthritis patients have antibodies against citrullinated peptides, termed seropositive for the anti-cyclic citrullinated peptide (anti-CCP); the patients lacking anti-CCP are termed seronegative. The genetic heritability of seropositive rheumatoid arthritis is close to 60%, most of which is accounted for by the HLA locus [42] [43]. Conditional and haplotype analyses identified that three amino acid positions (11, 71 and 74) in HLA-DRβ1 and single–amino-acid polymorphisms in HLA-B (at position 9) and HLA-DPβ1 (at position 9), together explain 12.7% of the phenotypic variance of rheumatoid arthritis. All those variants reside in the peptide-binding groove, and involve class 1 and class 2 HLA genes, which suggests the role of both CD8+ and CD4+ TCR-HLA interactions in the disease pathology [44].

4) Differential HLA density leads to autoreactivity.

It's been found that transcriptional regulation of HLA has significance in autoimmune diseases, notably the XL9 *cis*-regulatory region between HLA-DRB1 and HLA-DQB loci that modulates the transcription of HLA-DQ and HLA-DR genes. A GWAS combined with high-coverage sequencing with 773 cases and 576 controls reported peak signal in specific XL9 haplotypes in association with systemic lupus erythematosus (SLE). The study's RNA-seq expression analysis agreed with the quantitative flow cytometry results, which detected 2.5-fold increase in the HLA-DR surface density on monocyte-derived dendritic cell cultures from the disease-associated XL9 genotype individuals [45].

5) HLA stability

Unstable forms of HLAs have been associated with autoimmune diseases, while stable forms seem to confer protective effects. The 47α residue has been identified as a key regulator of HLA-DQ molecule stability, and variations in it have been found strongly associated with T1D [46]. HLA-DM is an important chaperone that functions to stabilize the antigen-presenting HLAs and load the peptide cargo into the HLA groove in the lysosome [47]. The mutations on HLA-DQ could exert their effects by hindering the HLA-DM action of protein exchange and stabilization. The structural characteristics of the disease-associated HLA-DQ variants imply that the change in specific hydrophobic residues, capable of forming H-bonds, may render the DM activity inefficient. The resulting increased presentation of unstable self-peptide complexes may lead to compromised negative selection, or alternatively, unstable HLA may be more prone to DM-independent protein exchange, thus leading to increased self-antigen presentation [48].

6) HLA interaction with microbiome

The microbiome has important implications in immune function [49] and the proper colonization of the infant colon at an early age is a requirement for the development of a healthy adult immune system [50]. The HLA molecules prime the immune response against pathogenic microbes, as well as regulate the tolerance to symbiotic bacteria, that in turn influences a diverse set of processes. In a rat model study of gastrointestinal inflammation, the HLA-B27 transgenic rats with the colon bacteria *Bacteroides vulgatus* showed spontaneous occurrence of inflammation, whereas the HLA-B27 population lacking the $CD4^{+}$T cells did not show disease response to *B. vulgatus* in their colon. No development of disease was found in germ-free environments or other tested common colon

bacteria [51] [52], which indicates that the HLA-B27 presenting peptides from otherwise non-pathogenic bacteria to $CD4^{+}$T cells could play a major role in the pathology. In celiac disease, multiple studies have shown that the infants with higher genetic risk to celiac disease (the DQ2 and DQ8 HLA haplotypes) had an altered microbiome composition regardless of the feeding type [53]. A microbiome analysis using next generation sequencing concluded that the specific HLA-DQ genotype was likely causal in shaping the microbiome [54] [40].

1.3 Genome-wide Association Testing

Genome-wide association study (GWAS) is a study design that is aimed at discovering associations between genotypes and phenotypes by comparing the allele frequencies of genetic variants between cohorts. The statistical power to detect associations between DNA variants and a trait determined by the experimental sample size, the distribution of effect sizes of (unknown) causal genetic variants, the frequency of those variants, and the linkage disequilibrium (LD) between observed genotyped DNA variants and the unknown causal variants [55] [56]

A GWAS consists of a series of association tests for each variant studied, where the variant is considered the explanatory variable (as the number of copies of the allele), and the phenotype status is considered as the dependent variable. Depending on whether the phenotype is continuous or binary, linear mixed models (LMM) or mixed logistic regression (MLR) models are used in GWAS respectively [56] [57]. In general terms, the "mixed" effects model refers to the fact that confounding effects on the explanatory variables are accounted for in the regression model (such as population stratification in GWAS). Chen, et al. showed in 2016 that fitting LMM to binary traits fails to account for population stratification and inflates type 1 error rate [58]. The linear mixed model BOLT-LMM [59] has still demonstrated decent performance with binary cohorts but only when cases and controls are balanced, while loss of power occurs with imbalanced cohorts. Mixed logistic regression models, such as SAIGE [60] and REGENIE [61] are able to perform very well with imbalanced case-control ratios.

The controlling for population stratification in GWAS cohorts can be done with either individual-level data, where the principal components of the genomic relatedness matrix are taken as covariables in the mixed effects model; or the stratification control can be done on the association test results (summary statistics) to correct for the "p-value inflation". A popular summary-statistics approach has been Genomic Control [62], where the p-value at each variant is adjusted by a

common inflation factor. More recently it's been shown that in majority of the GWASs, the inflation is not due to population stratification, but instead caused by the legitimate polygenic nature of most traits [63]. Genomic control fails to differentiate between the inflation caused by bias and polygenicity, and it's been suggested that LD-score regression should be preferred over genomic control to correct for p-value inflation in GWAS summary statistics.

Testing millions of variants in a GWAS creates a multiple testing challenge, which calls for a method to control for the large number of false positives with minimal trade-off in statistical power. If the tests were independent, the Bonferroni correction would control at level α for n tests with a significance level of α/n. Since in GWAS the linkage disequilibrium renders the tests on genomic variants non-independent, a Bonferroni correction would result in a great loss of power. In 2005 the HapMap Consortium used permutation testing of high-density genotypes in 10 genomic regions to obtain the so-called "effective number of independent tests" [64], which allows for a Bonferroni corrected significance level that better accounts for linkage disequilibrium. This effective number of independent tests was found to be 150 per 500 kb, scaling it up to 3 Gb gets a significance level of 5.5×10-8, which has since become a standard for common-variant (MAF≥5%) GWAS. This number, or close to this number, of independent tests has been separately obtained in significance level calculations for European ancestry common variant GWASs [65].

One factor to take into account with GWASs done on any biobank data is that the biobanks have shown difficulty in having a sample that is representative of the true population. This bias is thoroughly assessed in the case of UK biobank, that invited about 9 million people to participate, of which only about 500,000 (5.5%) participated. The resulting sample was found to be biased to be generally more healthy and having a higher socioeconomic status, which, in turn, has been shown to distort GWAS results, especially with lifestyle- and behaviour-related traits [66]. This voluntary participation bias is likely to extend to all the biobanks in the world, at least to some degree.

A standard procedure for any GWAS to increase the reliability of the associations, as well as to increase the power to discover new associations, is to replicate the study on mutliple cohorts, and to conduct meta-analyses accross multiple GWAS summary statistics. For example, the generalized-method-of-moments-estimator-based meta-analysis tool MTAG has been demonstrated to increase the discovery of trait-associated loci several-fold, compared to single

GWASs, and it performs especially well when the polygenic traits share some, or all, of the causal loci [67].

Since the completion of the HapMap project and the advancement of sequencing technologies, GWASs have identified many novel genetic variants that provide preliminary information about the underlying mechanisms in autoimmune diseases such as celiac disease, T1D, multiple sclerosis, rheumatoid arthritis, Crohn's disease and systemic lupus erythematosus [68]. The novel loci discovered by GWASs can guide the more expensive molecular studies needed to eventually map the genetic architecture and detect the functional consequences in the cells/molecules involved.

One of the characteristics of GWAS findings is that, due to LD, GWASs usually give many significant hits in one locus and it's not clear from just the summary statistics which ones of them are relevant to the disease pathology, especially since most (>90%) of the variants are in the non-coding DNA regions. Several strategies are being actively developed for the efficient and rapid fine-mapping of the causal variants in the associated loci. For example, the massively parallel reporter assay (MPRAs) has been successfully used test thousands of loci and tens of thousands of variants to pin-point causal variants, in both protein-coding- and non-protein-coding DNA regions, whose precise function can be subsequently determined by genetically editing cell lines and model organisms [69] [70]. Genome-wide association studies have been also successfully utilized as a part of a strategy to re-purpose drugs. [71] [72] and determining candidate genes that enable better informed treatment strategies and guide drug development processes [73].

1.4 Imputation

The power of a GWAS is also affected by the reliability of the genotype data which depends the method by which the genotype was obtained. Whole-genome sequencing (WGS) provides the most reliable genotype information, but considering the need to scale this to hundreds of thousands of individuals, a more cost-effective solution is to leverage the linkage-disequilibrium (LD) pattern in a population to predict the genetic variants based on marker loci. In genotype imputation, a population-specific reference panel is created from the whole-genome sequences of a population sample, and the complete set of variants can be predicted by genotyping only for a relatively small subset of markers [74].

It is also important to consider the information about which set of the inherited chromosomes the haplotype originates from, because the nucleotide content and the gene copy number differ between the two sets. This is referred to as the phase information between the maternally- and paternally-derived sequences, and specialized phasing software, using algorithms such as the hidden Markov model, have been developed and widely-utilized as a standard procedure in imputation [75] [76].

WGS is still exclusively used for rare-variant association studies, but for variants even as infrequent as 1/1000, it's been demonstrated that there is practically no gain in statistical power when WGS is used over imputation [55].

1.5 Principal Component Analysis (PCA)

Population stratification is the systemic difference in allele frequencies in population (dividing the population into strata), which can generate spurious associations in regression models. The idea to use PCA on genomic data was first published by Menozzi, et al. in 1978 [77] and it's been since shown that PCA can be effectively used to account for population stratification in GWASs, by including the principal components of a population genomic covariance matrix as covariates in mixed effects models [78].

In individual-level genomics data, the principal component analysis is based on the eigendecomposition of the genomic covariance matrix, where the rows represent individuals and the columns represent the marker allele doses (0, 1, 2 for biallelic variants) [79]. Technically, PCA itself is not dimensionality reduction, but instead it's a data transformation technique that makes the data amendable to dimesnionality reduction. Principle component analysis finds new axis (dimensions) to the data space, that are ordered hierarchically according to the amount of variance they describe in the data. The first axis (the first principal component) describes the most amount of variance, meaning, the mean squared error from all the data points to that axis is minimized; the second axis describes less amount of variance than the first one, and the third one less still, etc. The PCA allows to perform dimensionality reduction by choosing the number of principal components that describe the amount of variance in the data to a degree that we are satisfied with [80] [81].

For individual-level genomic data, it's recommended that the genomic markers subject to PCA are not in strong LD with one another – on one hand so that the principal components would not capture too much correlation between markers in individual loci, and it is also computationally efficient to include only the SNPs most relevant to the stratification modeling. Pairwise LD pruning of SNPs

above some LD threshold is recommended before conducting PCA on genomic data [82].

1.6 Variant Effect Prediction

Variant effect prediction aims to prioritize the genomic variants according to the severity of the effect they have on the phenotype of interest. In the early days, the identification of influential variants was done by literature search of the most statistically robust findings, and this laborious work was usually limited to only a handful of variants [83]. The 1000 Genomes Consortium published findings in 2015 that a typical human genome differs at 4.1-5.0 million variant sites from the global variant reference of over 88 million variants [84]. Many of those variants have benign effects that are not likely relevant to the phenotype studied, and rapid methods for prioritizing and providing biological context to variants are essential tools in areas such as modern sequence-based clinical diagnostics and genome-wide association studies.

Variant effect prediction can be performed ad-hoc or post-hoc according to when it is performed relative to the genomic analysis. The ad-hoc approach is commonly used in association tests for rare variants to increase the statistical power by „collapsing“ a set of rare variants into groups based on their functional annotation [85]. In genome-wide association studies that perform agnostic tests on common variants, post-hoc variant annotation is done on the significant associations in the summary statistics of the test. Several bioinformatics tools have been developed for variant annotation; their general principle is to provide references to public databases, that contain aggregated information on known variants, and provide computationally-derived impact scores that reflect how deleterious the mutation is. Most of the variant tools focus on the annotation of SNPs; INDELs are also covered by some tools, however, structural variant (SV) calling is currently limited to copy number variants, and this is performed only by some of the latest tools [86] [87]. Among the widely-used annotation tools are ANNOVAR [88], SNPeffect [89] and the Ensembl Variant Effect Predictor (VEP) [90].

SNPeffect predicts the effects of variants only in protein-coding genes. It uses computational algortihms that rely on the data from the UniProt human variation database to predict protein aggregation and amyloid formation (TANGO and WALTZ, respectively), also chaperone binding (LIMBO) and structural stability (FoldX). The annotation tools VEP and ANNOVAR predict variant effects in protein-coding as well as non-coding regions. They rely on several gene annotation databases, such as GENCODE [91] and Ref-Seq [92]. GENCODE combines

experimental molecular studies and transcriptome analysis with the Ensembl automatic annotation pipeline to annotate coding- and non-coding regions. The experimental information for the variants in GENCODE has been collected from different studies that have collectively mapped the variant characteristics to the human genome; these studies have included DNase hypersensitivity assays for chromatin openness, transcription factor binding assays to identify promoter sites, and the identification of polymerase-bound genomic regions to locate enhancer elements. Transcriptome analysis is used to identify protein-coding regions, and regulatory regions that, for example, encode miRNA. There is some inconsistency between the major databases (GENCODE, Ref-Seq, UCSC, CCDS and AceView) in terms of the number of loci identified as protein-coding. This is largely due to different methods of how the mRNA transcript data is interpreted [91]. In addition, all of the databases are subject to constant updates; mostly, the number of known transcript isoforms per loci increases as experimental data continues to accumulate.

In addition to database references, variant annotation tools can give several computationally-derived scores that predict the severity of the variant. Annotation tools, such as ANNOVAR and VEP, harbor pre-calculated scores for variants in the updated databases, and these can be included in the annotation output by the user. Some of the computational scores are highlighted here. PolyPhen2 (Polymorphism Phenotyping v2, [93]) predicts how an amino acid substitution (a missense mutation) affects the function of the gene product. It's machine learning model is trained on a set of annotated protein sequences of healthy and disease-associated sequences from UniProt, incorporating also information about the protein structure, and whether the mutation has occurred in a polymorphic or evolutionarily conserved site. ClinVar [94] is a variant severity score that specifically considers disease-causing mutations, and it gives a prediction of disease significance based on a public archive of annotated alleles, which is maintained at the National Center for Biotechnology Information (NCBI). The Combined Annotation Dependent Depletion (CADD) [95] framework combines the diverse set of developed annotations into one single score; it contains 60 different annotations in categories of epigenetic modifications, functional prediction (which includes PolyPhen2), evolutionary conservation, and the gene sequence context (indel length, CpG content, etc.). CADD's machine-learning model is trained on a set of alleles which have not been annotated based on experimental data, but instead the healthy reference variants are taken to be the variants in the human genome that have persisted since the split with the human-ape common ancestor (that is, their current observed allele frequency is 95-100%), as millions of years of purifying selection has presumably kept only the non-deleterious ones in the population.

3 AIMS OF THE THESIS

It is well dodumented that certain HLA alleles are correlated with autoimmune diseases and numerous studies have provided evidence that mutations in non-HLA loci can be significant co-factors to autoimmune disease susceptibility. The advancement of sequencing, genotyping technology, and bioinformatics tools have made it only recently possible to robustly conduct genetic association analysis accross many alleles and phenotypes with large cohorts, that have the capacity to provide novel insights into the genetic architecture of diseases which have a heritable component. The current study aims to agnostically search for signals of autoimmune disease associations, using the cohort of the EsBB, which would be suggestive of the presence of possible gene-gene interactions between HLA and non-HLA protein-encoding genes.

To satisfy this aim, a GWAS (using mixed logistic regression) was designed in this project which comprises of two types of case-control study groups: HLA-specific study groups, where individuals are the carriers of a specific HLA type (at least one allele copy), and a study group comprising of the whole EstBB population. The effect sizes of the autoimmune-disease-associated variants from the HLA-specific GWASs were compared against the effect sizes of these variants from the whole-population GWAS; the variants were considered relevant to the aim if the 95% confidence intervals did not overlap in the comparison between the groups.

The present research does not aim to make the distinction whether the genetic patterns are due to multiplicative (epistatic) or additive effects, but a locus that has a significantly stronger association with a disease in a HLA-specific cohort would be deemed of interest in the context of further analysis, and the results of this work can prioritise a set of variants to be subject to studies that involve more complex modeling of gene-gene interactions between HLA and non-HLA proteins.

4 EXPERIMENTAL PART

4.1 MATERIALS AND METHODS

4.1.1 Association Testing Software

The association test software used for this study was Regenie [61] with saddle point approximation (SPA) correction.

The cohort covariates included in the mixed logistic regression were sex, age, age-squared + 10 principal components.

The association test software was run in the University of Tartu High-Performace Computing server.

4.1.2 EstBB hg19 Cohort Data

The Estonian Biobank is a population-based biobank with 212,955 participants (data freeze 2022v2). All biobank participants have signed a broad informed consent form and information on ICD-10 codes is obtained via regular linking with the national Health Insurance Fund and other relevant databases, with majority of the electronic health records having been collected since 2004

[96]. The analyses in this work were only carried out on participants with reliably imputed HLA alleles (n=196,293).

All EstBB participants have been genotyped at the Core Genotyping Lab of the Institute of Genomics, University of Tartu, using Illumina Global Screening Array v3.0_EST. Samples were genotyped and the PLINK format files were created using Illumina GenomeStudio v2.0.4. Individual sample exclusion criteria was: call-rate < 95% (2.52% of the samples); sex based on heterozygosity of X chromosome not matching sex in phenotype data; HWE p-value < 1e-4 (autosomal variants only), and minor allele count = 0. Genotyped variant positions were in build hg19. Phasing was performed using the Eagle v2.4.1 software. Imputation was performed with Beagle v5.4 software and default settings, in batches of 10,000. A population specific reference panel consisting of 2,056 WGS samples was utilized for imputation and standard Beagle hg37 recombination maps were used. Altogether, 30,367,921 SNPs are imputed per individual from 296,903 SNP markers that overlap between the chip array and the WGS reference panel. Based on principal component analysis, samples who were not of European ancestry were removed. Duplicate and monozygous twin detection was performed with KING 2.2.7 [PMID:20926424], and one sample was removed out of the pair of duplicates. Imputation INFO score was calculated using bcftools +impute-info command. The HLA alleles are imputed using the SNP2HLA software [97]. PCA was performed in PLINK on LD-pruned variants (LD-pruning performed using PLINK function --indep-pairwise with 100kb window size; variant count step-size 10; r2 threshold 0.1): plink2 --extract plink2.prune.in --pca 50 approx.

4.1.3 Selecting the SNPs for the study

The analyses in this project concentrated on protein structure-altering genetic variants, as these SNPs generally provide the most straightforward biological interpretation options. From the total number of 39,274,267 SNPs in the Estonian Biobank for the hg19 genome assembly, this study included only exonic variants that had MAF>0.05 and the imputation INFO score ≥ 0.9. After applying MAF and INFO score filters, 5,886,414 entries remained. To select variants residing in exome, overlapping chromosome and position numbers (hg19 genome assembly contains only one variant per genomic position) with the exome sites from GnomAD (gnomad.exomes.r2.1.1.sites.vcf.bgz; contains 17,209,972 sites) were obtained. There were 655,342 positions that matched the selection criteria for this study.

4.1.4 Variant Effect Prediction

The variant effect prediction was done with Ensembl Variant Effect Predictor (VEP) web interface; the input to VEP was a VCF file in standard format for all the 655,342 exonic variants that were selected from the Estonian Biobank (section 4.1.3).

This study considered the following variant types: missense, frameshift, stop-gained, stop-lost, start-lost, inframe deletion, inframe insertion. The most_severe VEP output filter was applied, which gives only the most severe consequence per input variant as predicted by VEP. From the 655,342 exonic variants, 224,799 were missense; 6,957 were frameshift; 5,740 were stop-gained; 3,184 were inframe deletions; 1,157 were inframe insertions; 892 were start-lost; 452 were stop-lost.

4.1.5 Defining the Relevant Findings

The GWAS summary statistics were filtered with a significance level 5×10^{-8}.

To satisfy the aim of finding effects that are tied to specific HLA loci (indicative of epistasis), a comparison of the effect sizes was done between the GWAS with the whole EstBB population and the GWASs of HLA-specific study groups. The variant associations were considered relevant to the study aim if the 95% confidence intervals of the effect sizes did not overlap between the HLA and population groups.

Variant associations in the HLA locus on chromosome 6 were excluded form the summary statistics as the study focuses only on epistatic interactions that happen between HLA and non-HLA genes.

4.1.6 Ethics

The activities of the EstBB are regulated by the Human Genes Research Act, which was adopted in 2000 specifically for the operations of EstBB. Individual level data analysis in EstBB was carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics and Human

Research (Estonian Ministry of Social Affairs), using data according to release application 6-7/GI/8592 from the Estonian Biobank.

4.1.6 Data Handling Tools

The filtering and handling of the variant and cohort data was done using R v. 3.6.0; R data.table v. 1.14.2; R dplyr v. 1.0.8. The bar plots and the heatmap were done using R ggplot2 v. 3.4.2. The UpSet plot was done using R ComplexHeatmap from BioConductor.

4.2 RESULTS

Previously, EstBB has identified HLA associations to different diseases by carrying out PheWAS association analyses between 507 imputed HLA alleles against 193 different ICD-10 based diagnoses using logistic-regression-based analyses [98].

To compile study groups with HLA types that are most relevant to autoimmune disease susceptibility, the HLAs were selected based on the PheWAS (Figure 4). The criterium of an allele being included in the current project was that the number of associated diseases according to the PheWAS ≥ 1, and the imputation score $DR^2 \geq 0.8$ (see section 4.1.2 for imputation methods in EstBB). 42 HLA types were included in the project; the current project considered the HLA alleles with 2-digit resolution (see section 1.1.1 for HLA nomenclature).

The study groups were selected from the EstBB cohort based on the specific HLA allele dose: 42 groups were formed for every HLA allele type included in the study, with every individual in a group having an imputed/genotyped HLA allele dose of >0.9; (Figure 5). The GWASs done with the HLA study groups included only the autoimmune diseases which were associated with the HLA type in the PheWAS (Figure 4). In total, 245 GWASs were conducted across the 42 HLA-type groups and 52 diseases (as defined by ICD10 annotation) included in the study.

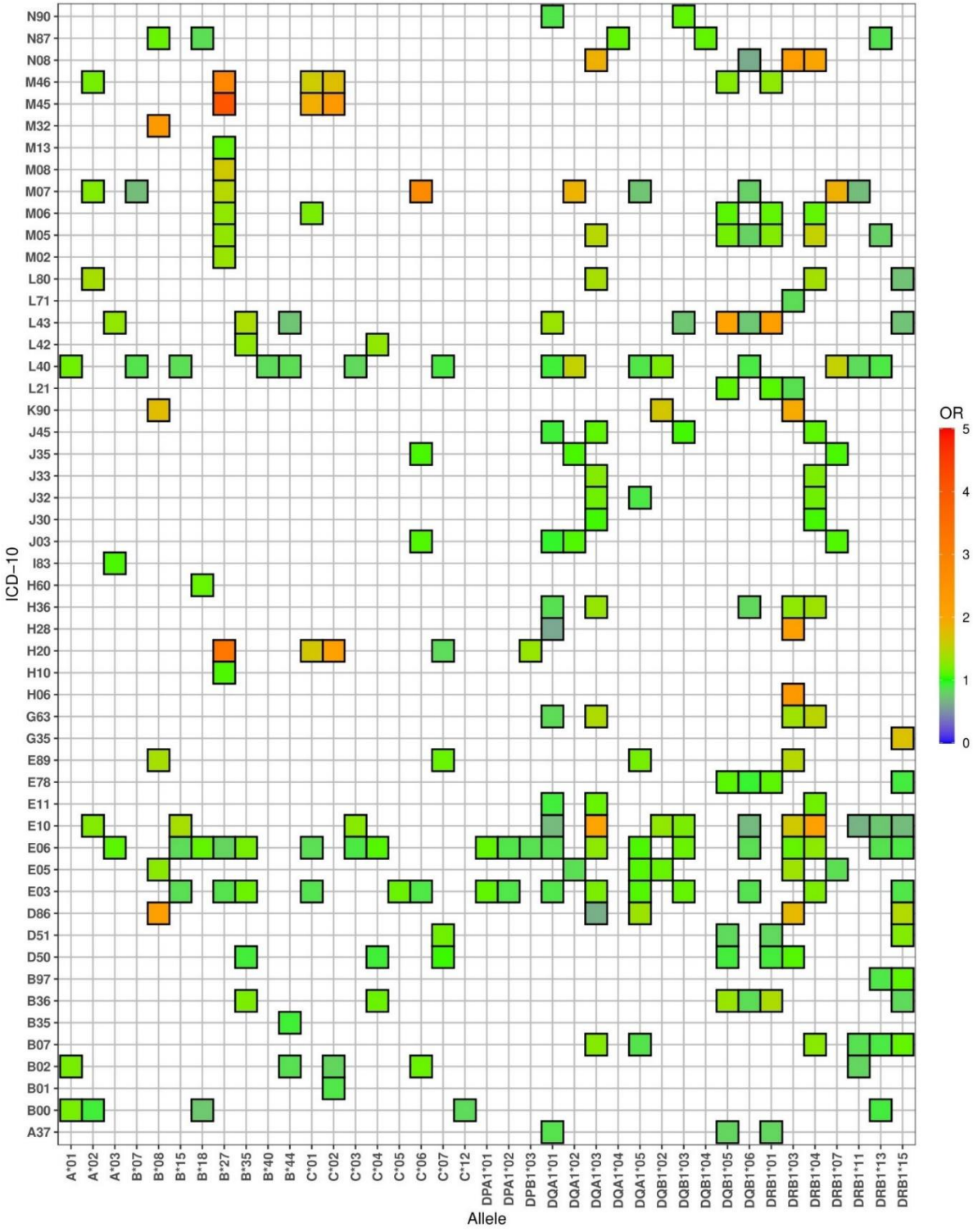

Figure 4 *Heatmap showing the PheWAS results previously done on EstBB population. This heatmap includes all the alleles, and their respective disease-associations (as odds-ratios), according to which the HLA-specific groups were made.*

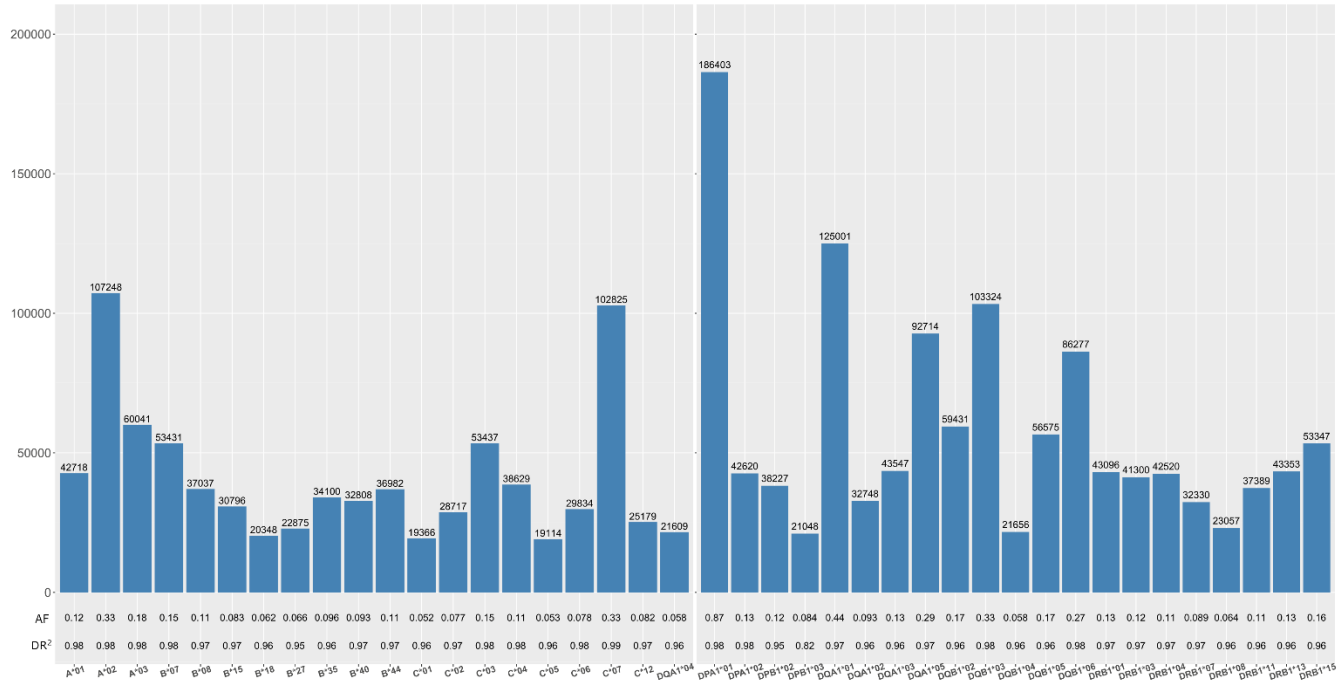

Figure 5 *The HLA-specific groups that were included in the study, along with their allele frequency (AF) and imputation score (DR^2; imputed with SNP2HLA)*

The GWASs gave 221 hits that reached the $p=5\times10^{-8}$ ($-\log_{10}(p) = 7.3$) significance level. 220 of them were in the variant class „missense“ (as annotated in VEP; see METHODS 4.1.4) and 1 hit was in the variant class „inframe deletion“. For other variant classes included in the study, none of the GWASs produced hits above the significance level.

After the comparison of the effect sizes between the HLA-specific and whole-population groups (see METHODS 4.1.5), 4 out of the 221 results remained as having a significantly different effect size in between the HLA and whole-EstBB-population study groups.

All of the significant comparison results were for the SNP rs2476601 in the gene *PTPN22*. The associated HLA types were DRB1*03 (diseases T1D and retinopathy), DQA1*05 (thyroiditis) and DQB1*02 (T1D) (Figure 6)

The PolyPhen2 score predicts rs2476601 as BENIGN, score = 0.029; CADD score = 18.22; ClinVar identifies the variant as „risk factor“ for T1D and thyroiditis (no rs2476601 ClinVar data found for retinopathy).

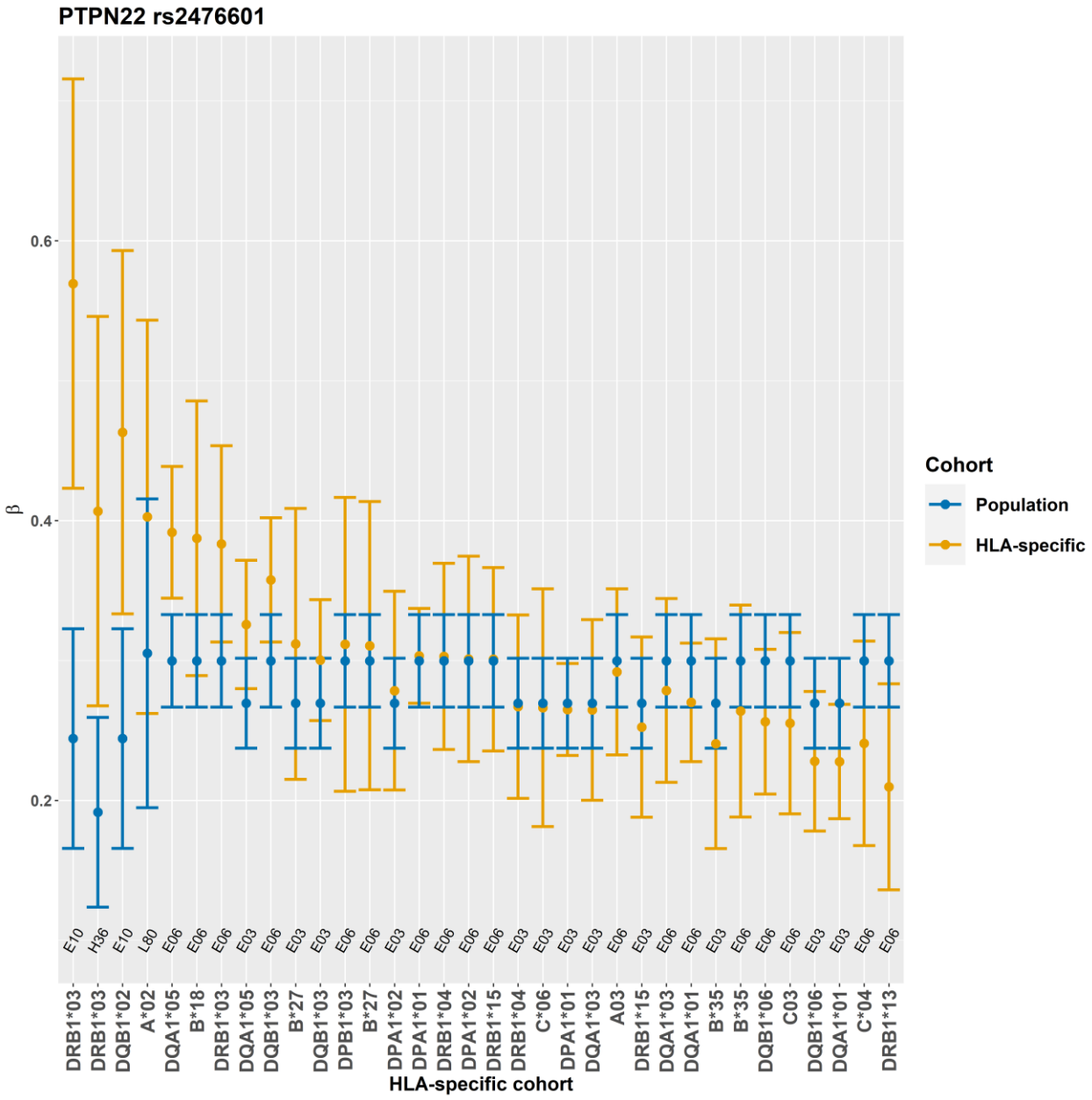

Figure 6 *The effect sizes (y axis) and their respective 95% confidence intervals showing the effect size comparison between HLA-specific and the whole EstBB population sample group GWASs. The table depicts all (34) of the R620W hit comparisons from the total of 221 that reached GWAS significance. See Table 1 for ICD10 code definitions. All of the 4 results with significant effect size differences were with the PTPN22 variant R620W.*

There was also a large number of loci, which had a relatively large effect size difference in the HLA-specific groups, but which did not reach statistical difference in the current framework due

to overlapping 95% confidence intervals in the comparison with the whole-population GWAS (Figure 7)

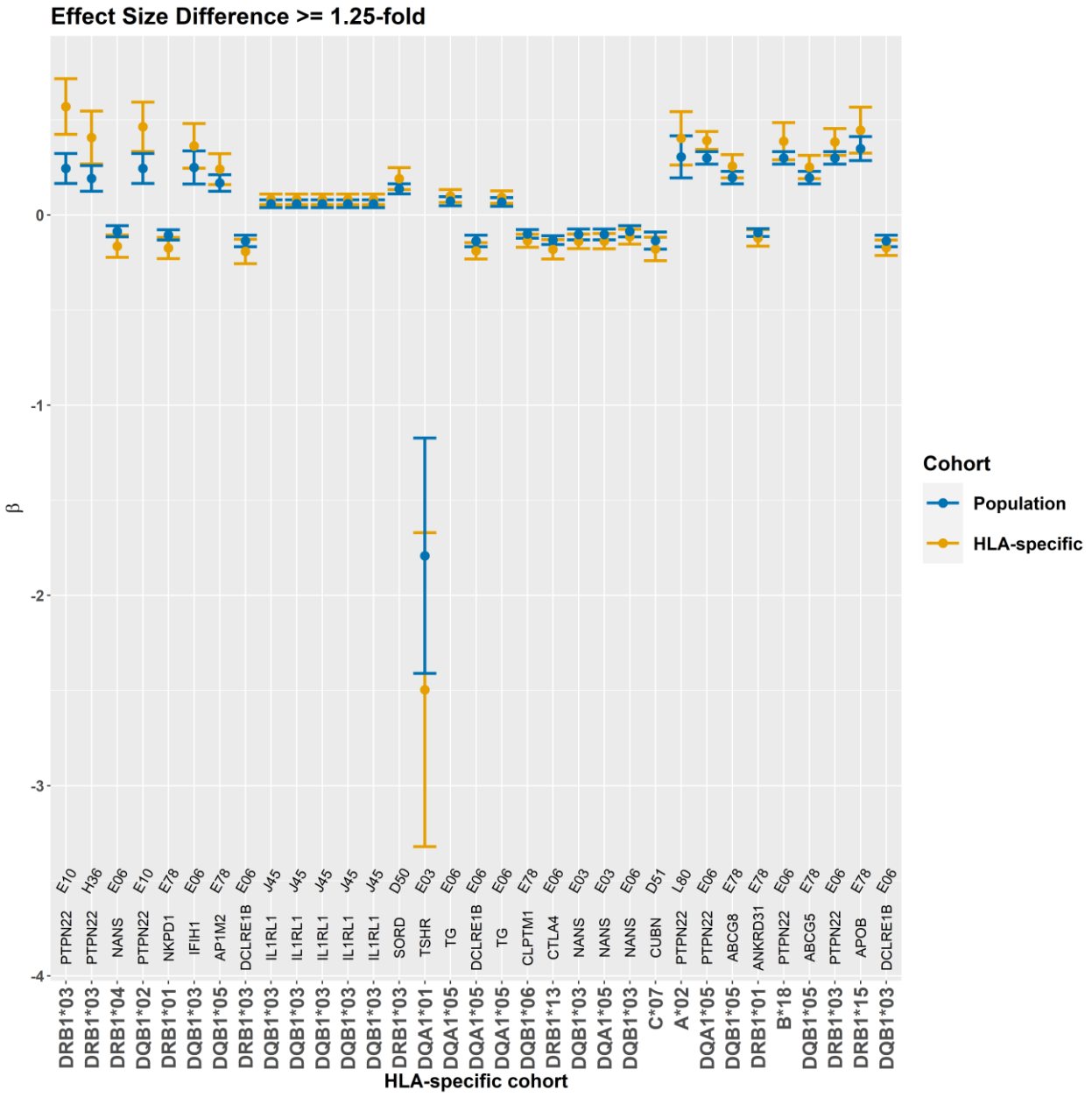

Figure 7 *The effect size comparison data for all the gene loci which had the mean effect size at least 1.25-times different from the effect sizes in the whole-population GWAS. See Table 1 for ICD10 code definitions.*

ICD10	DISEASE
E10	Type 1 diabetes mellitus
H36	Retinopathy, unspecified
E06	Thyroiditis
E78	Disorders of lipoprotein metabolism and other lipidemias
J45	Asthma
D50	Iron deficiency anemia
E03	Hypothyroidism
D51	B12 deficiency anemia
L80	Vitiligo

Table 1 *The ICD10 code definitions for all the diseases that are included in figures 6 and 7*

The overlap of disease cases between the EstBB cohort and the HLA-specific groups was also examined. The total number of T1D cases in EstBB is 2,733, of which ~1/3 are present in the HLA DRB1*03 and DQB1*02 groups (Figure 8). The amount of thyroiditis cases in the DQA1*05 group

was 8,604, which is 49% from the 17,469 cases in EstBB (data not shown); the amount of retinopathy cases in the DRB1*03 group was 925, which is 24% of the 3,849 (data not shown).

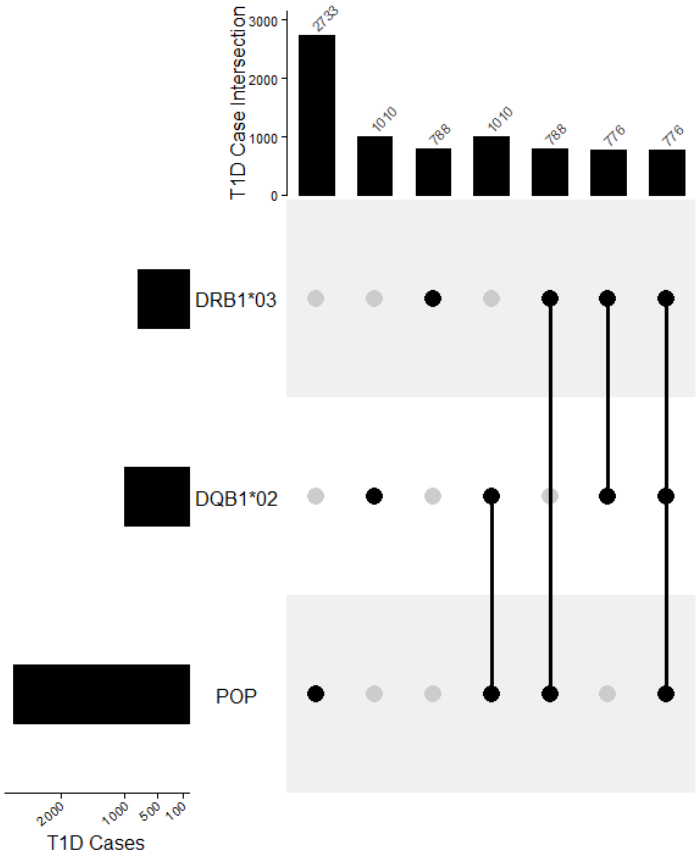

Figure 8 *an UpSet plot showing overlap between T1D cases between the EstBB population reference group (POP) and the two HLA-specific group where the R620W signal reached significance. The sample overlaps are in „intersect", meaning that the overlaps are not mutually exclusive; see the intersection modes in the ComplexHeatmap R package documentation [99]*

5 DISCUSSION

These findings agree with the literature, which concludes that autoimmune disease heritability is explained 2-30% by HLA alleles, strongest which is known to be for T1D by the HLA DR and DQ haplotypes (~30%) [6]. The *PTPN22* variant R620W has been identified as one of the strongest non-HLA genetic risk factors in several autoimmune diseases [113].

The *PTPN22* variant R620W was first published as a T1D associated variant in 2004 [100] and it has been since determined to have a strong association with several autoimmune diseases [101] [102] [103] [104] [105]. The *PTPN22* gene on chromosome 1 encodes the lymphoid-specific tyrosine phosphatase (Lyp) which has an important role in the activation pathway of T-cells. Lyp negatively regulates (dephosphorylates) the Syk and Src family kinases, which are the main mediators of T-cell activation triggered by TCR engagement [106] [107]. The variant rs2476601 (also named C1858T; R620W; 1858C>T) is a missense C → T mutation at position 1858 of the *PTPN22* gene that results in arginine → tryptophan amino acid change at position 620 of Lyp. Studies with cells from patients with T1D, rheumatoid arthritis and thyroiditis are reporting reduced IL-2 production and decreased TCR signal response (blunted calcium flux and transcription factor activity), but an expanded pool of memory T-cells [108]. Mice expressing $Ptpn22^{R619W}$ (the mouse orthologue of R620W) have been shown to have an expanded population of $CD4^+$ and $CD8^+$ effector T-cells, and a downregulated population of the immunosuppressing T_{reg} cells [109].

The published data on gene-gene interactions suggests that the joint-effect between HLA locus and *PTPN22* in autoimmune diseases may be additive (not epistatic) [110] [111] [112] [113] [114], however, the precise role of *PTPN22* variants in the development of autoimmune diseases is generally not very well defined and still under active research, thus the locus should be kept in relevance for further studies. Furthermore, the current study resulted in several loci within the 221 statistically significant hits, which, although, did not pass the criterium of synergy under current framework (due to overlapping 95% confidence intervals in the whole-population comparison), still should be considered as potential candidates for further gene-gene interaction analysis. Such as the *NANS* gene locus (lead SNP rs1058446), which showed a 2-fold mean effect size increase in HLA-DRB1*04 group for thyroiditis, and the *NKPD1* locus (lead SNP rs28469095) showing 1.65-fold mean effect increase in the DRB1*01 group in association with lipoprotein metabolism disorders (Figure 7). In the result interpretation, it should be taken into account that the EstBB

population reference did not exclude the disease cases that were also the study group HLA carriers, and considering that the disease case overalp is substantial (Figure 8), there is likely some inflation in type 2 error rate. Moreover, since it's been determined that single amino acid substitutions in the HLA peptide binding groove can have dramatic effects on the TCR/antigen interactions, a future study should consider the HLA alleles with higher resolution to increase power. Also, the present study included only non-synonymous variants in protein-coding regions, and future analysis could benefit from also investigating the pathogenic associations involving regulatory DNA regions (with the inclusion of appropriate fine-mapping measures [69] [70])

Inherently, the agnostic GWAS results in the framework of this analysis are only suggestive of gene interactions, and the distinction between epistatic and additive genetic interactions cannot be inferred from these results directly, but instead these results serve to guide further studies with loci/variants where gene-gene interaction may be present. For future studies, the use of specialized statistical frameworks should be examined, such as the Partial Least Squares Path Modeling statistic mPLSM [115] or PCA-based approaches [116]. Recently these statistical methods have also been pointed out to have inferior power to detect gene-gene interactions compared to deep learning methods, which may be more suitable for this task instead [117].

6 SUMMARY

The project's HLA-specific GWAS design and the subsequent variant effect size comparison with the EstBB population reference was aimed at detecting variants that would differ in their association strength to autoimmune diseases in individuals that carry specific HLA types. The analysis resulted in the detection of significant synergy between the *PTPN22* gene variant R620W with T1D (HLA study groups DRB1*03 and DQB1*02), retinopathy (HLA study group DRB1*03) and thyroiditis (HLA study group DQA1*05). Based on the results, as well as the relevant published literature, it's feasible to suggest that the variant R620W may involve synergy with certain HLA types.

The present study also detected several loci with substantial mean effect difference in the HLA-specifc groups, despite having overlapping 95% confidence intervals with the EstBB population GWAS reference, which should also be prioritized in epistatic effects modeling.

In summary, the present work concluded that several non-HLA variants show substantially stronger GWAS effect sizes with respect to autoimmune diseases among the carriers of certain HLA types. These variants are suggested to be subjected to further gene-gene interaction analysis which can lead to more precise mapping of the genetic architecture behind these diseases.

REFERENCES

[1] P. J. Delves, "Human Leukocyte Antigen System," September 2021. [Online]. Available: https://www.merckmanuals.com/professional/immunology-allergic-disorders/biology-of-the-immune-system/human-leukocyte-antigen-hla-system.

[2] R. L. P and S. A., "GENETICS OF HLA DISEASE ASSOCIATION," *Annual Review of Genetics,* vol. 15, pp. 169-187, 1981.

[3] F. J. Forbes and J. P. Morris, "Leukocyte Antigens In Hodgkins Disease," *The Lancet,* vol. 296, no. 7678, pp. 849-851, 1970.

[4] J. Trowsdale and J. C. Knight, "Major Histocompatibility Complex Genomics and Human Disease," *Annual Review of Genomics and Human Genetics,* vol. 15, pp. 301-323, 2013.

[5] Pishesha, Novalia et al., "A guide to antigen processing and presentation," *Nature Reviews Immunology,* vol. 22, p. 751–764, 2022.

[6] Anaya, Juan-Manuel et al, Autoimmunity: From Bench to Bedside, Bogota: El Rosario University Press, 2013.

[7] The MHC sequencing consortium, "Complete sequence and gene map of a human major histocompatibility complex," *Nature,* vol. 401, p. 921–923, 1999.

[8] J. Klein and A. Sato, "The HLA System," *The New England Journal of Medicine,* vol. 343, pp. 702-709, 2000.

[9] H. W. M. van Deutekom and C. Kesmir, "Zooming into the binding groove of HLA molecules: which positions and which substitutions change peptide binding most?," *Immunogenetics,* vol. 67, p. 425–436, 2015.

[10] Huan, Xuelu et al., "Crystal structure of suboptimal viral fragments of Epstein Barr Virus Rta peptide-HLA complex that stimulate CD8 T cell response," *Nature Scientific Reports*, vol. 9, 2019.

[11] Janeway, Charles et al., Immunobiology: The Immune System in Health and Disease. 5th edition, New York: Garland Science, 2001.

[12] Warrington, Richard et al., "An introduction to immunology and immunopathology," *Allergy, Asthma & Clinical Immunology,* vol. 7, 2011.

[13] Vignali, Dario A. A. et al., "How regulatory T cells work," *Nature Reviews Immunology,* vol. 8, p. 523–532, 2008.

[14] A. V. Joglekar and G. Li, "T cell antigen discovery," *Nature Methods,* vol. 18, p. 873–880, 2021.

[15] A. K. Sewell, "Why must T cells be cross-reactive?," *Nature Reviews Immunology,* vol. 12, p. 669–677, 2012.

[16] Klein, Ludger et al., "Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see)," *Nature Reviews Immunology,* vol. 14, p. 377–391, 2014.

[17] K. C. Hurley, "Naming HLA diversity: A review of HLA nomenclature," *Human Immunology,* vol. 82, no. 7, pp. 457-465, 2021.

[18] K. Madden and D. Chabot-Richards, "HLA testing in the molecular diagnostic laboratory," *Virchows Archiv, Springer,* vol. 474, pp. 139-147, 2019.

[19] Warren, Rene L. et al., "Derivation of HLA types from shotgun sequence," *Genome Medicine,* vol. 4, 20212.

[20] K. K. Mittal and K. Daljit, "Serological Testing of the HLA Antigens," *Current Trends in Histocompability. Springer,* pp. 209-229, 1981.

[21] Szolek, András et al., "OptiType: precision HLA typing from next-generation sequencing data," *Bioinformatics*, vol. 30, no. 23, pp. 3310-3316, 2014.

[22] Stürner, Klarissa, Anja et al., "Is multiple sclerosis progression associated with the HLA-DR15 haplotype?," *Multiple Sclerosis Journal*, 2019.

[23] Madsen, Lars, S et al., "A humanized model for multiple sclerosis using HLA-DR2 and a human T-cell receptor," *Nature Genetics*, vol. 23, p. 343–347, 1999.

[24] M. Hahn, "Unconventional topology of self peptide–major histocompatibility complex binding by a human autoimmune T cell receptor," *Nature Immunology*, vol. 6, p. 490–496, 2005.

[25] Appel, Heiner et al., "Kinetics of T-cell Receptor Binding by Bivalent HLA-DR·Peptide Complexes That Activate Antigen-specific Human T-cells," *Journal of Biological Chemistry*, vol. 275, no. 1, pp. 312-321, 2000.

[26] Quandt, Jaqueline, A. et al., "Unique Clinical and Pathological Features in HLA-DRB1*0401–restricted MBP 111–129–specific Humanized TCR Transgenic Mice," *Journal of Experimental Medicine*, p. 223–234, 2004.

[27] Yin, Y et al., "Structure of a TCR with high affinity for self-antigen reveals basis for escape from negative selection," *The EMBO Journal*, vol. 30, pp. 1137-1148, 2011.

[28] Skowera, Ania et al., "CTLs are targeted to kill β cells in patients with type 1 diabetes through recognition of a glucose-regulated preproinsulin epitope," *The Journal of Clinical Investigation*, 2008.

[29] Bulek, Anna M et al., "Structural basis for the killing of human beta cells by CD8+ T cells in type 1 diabetes," *Nature Immunology*, vol. 13, p. 283–289, 2012.

[30] Bjornevik, Kjetil et al., "Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis," *Science*, vol. 375, no. 6578, pp. 296-301, 2022.

[31] H. B. Warner and R. I. Carp, "MULTIPLE SCLEROSIS AND EPSTEIN-BARR VIRUS," *The Lancet,* vol. 318, no. 8258, p. 1290, 1981.

[32] Sethi, Dhruv K. et al., "A highly tilted binding mode by a self-reactive T cell receptor results in altered engagement of peptide and MHC," *Journal of Experimental Medicine,* vol. 208, p. 91–102, 2011.

[33] K. W. Wucherpfennig and J. L. Strominger, "Molecular Mimicry in T Cell-Mediated Autoimmunity: Viral Peptides Activate Human T Cell Clones Specific for Myelin Basic Protein," *Cell,* vol. 80, pp. 695-705, 1995.

[34] Hausmann, Stefan et al., "Structural Features of Autoreactive TCR That Determine the Degree of Degeneracy in Peptide Recognition," *The Journal of Immunology,* vol. 162, no. 1, p. 338–344, 1999.

[35] Harkiolaki, Maria et al., "T Cell-Mediated Autoimmune Disease Due to Low-Affinity Crossreactivity to Common Microbial Peptides," *Immunity,* vol. 30, no. 3, pp. 348-357, 2009.

[36] Cheong, Rok Seon et al., "Celiac disease risk stratification based on HLA-DQ heterodimer (HLA-DQA1 ~ DQB1) typing in a large cohort of adults with suspected celiac disease," *Human Immunology,* pp. 59-64, 2020.

[37] Lundin, K E et al., "Gliadin-specific, HLA-DQ(alpha 1*0501,beta 1*0201) restricted T cells isolated from the small intestinal mucosa of celiac disease patients.," *Journal of Experimental Medicine,* vol. 178, p. 187–196, 1993.

[38] van de Wal, Yvonne et al., "Small intestinal T cells of celiac disease patients recognize a natural pepsin fragment of gliadin," *PNAS,* vol. 95, pp. 10050-10054, 1998.

[39] K. E. A. e. a. Lundin, "T cells from the small intestinal Mucosa of a DR4, DQ7/DR4. DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8," *Human Immunology,* vol. 41, pp. 285-291, 1994.

[40] L. Espino and C. Núñez, "The HLA complex and coeliac disease," *International Review of Cell and Molecular Biology,* 2020.

[41] Hovhannisyan, Zaruhi et al., "The role of HLA-DQ8 β57 polymorphism in the anti-gluten T-cell response in coeliac disease," *Nature,* vol. 456, pp. 534-538, 2008.

[42] Bax, Marieke et al., "enetics of rheumatoid arthritis: what have we learned?," *Immunogenetics,* vol. 63, p. 459–466, 2011.

[43] Smolen, Josef S. et al., "Rheumatoid arthritis," *Nature Reviews Disease Primers,* vol. 4, 2018.

[44] Raychauduri, Ray et al., "Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis," *Nature Genetics,* vol. 44, pp. 291-296, 2012.

[45] Raj, P. et al., "Regulatory polymorphisms modulate the expression of HLA class II molecules and promote autoimmunity," *eLife Genetics and Genomics,* 2016.

[46] Miyadera, Hiroko et al., "Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA," *The Journal of Clinical Investigation,* 2014.

[47] A. B. Vogt and H. Kropshofer, "HLA-DM – an endosomal and lysosomal chaperone for the immune system," *Trends in Biochemical Sciences,* vol. 24, no. 4, pp. 150-154, 1999.

[48] Z. Zhou and P. E. Jensen, "Structural characteristics of HLA-DQ that may impact DM editing and susceptibility to type-1 diabetes," *Frontiers Immunology,* vol. 4, 2013.

[49] Zheng, Danping et al., "Interaction between microbiota and immunity in health and disease," *Nature Cell Research,* vol. 30, p. 492–506, 2020.

[50] A. W. Walker, "The importance of appropriate initial bacterial colonization of the intestine in newborn, child, and adult health," *Nature Pediatric Research,* vol. 82, p. 387–395, 2017.

[51] Hoentjen, Frank et al., "CD4+ T lymphocytes mediate colitis in HLA-B27 transgenic rats monoassociated with nonpathogenic Bacteroides vulgatus," *Inflammatory Bowel Diseases,* vol. 13, p. 317–324, 2007.

[52] Rath, Heiko C. et al., "Differential Induction of Colitis and Gastritis in HLA-B27 Transgenic Rats Selectively Colonized with Bacteroides vulgatus or Escherichia coli," *ASM Journals, Infection and Immunity,* vol. 67, no. 6, 1999.

[53] Sánchez, Ester et al., "Influence of environmental and genetic factors linked to celiac disease risk on infant gut colonization by Bacteroides species," *Applied and Environmental Microbiology,* vol. 77, no. 15, pp. 5316 - 5323, 2011.

[54] Olivares, M et al., "The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease," *BMJ Journals,* vol. 64, no. 3, pp. 406-417, 2015.

[55] Visscher, Peter M et al., "10 Years of GWAS Discovery: Biology, Function, and Translation," *The American Journal of Human Genetics,* vol. 101, no. 1, 2017.

[56] Uffelmann, Emil et al., "Genome-wide association studies," *Nature Reviews Methods Primers,* vol. 1, 2021.

[57] Milet, Jacqueline et al., "Mixed logistic regression in genome-wide association studies," *BMC Bioinformatics,* vol. 21, 2020.

[58] Chen, Han et al., "Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Regression," *American Journal of Human Genetics,* vol. 98, no. 4, pp. 653-666, 2016.

[59] Loh, Po-Ru et al., "Efficient Bayesian mixed-model analysis increases association power in large cohorts," *Nature Genetics,* vol. 47, p. 284–290, 2015.

[60] Zhou, Wei et al., "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies," *Nature Genetics,* vol. 50, p. 1335–1341, 2018.

[61] Mbatchou, Joelle et al., "Computationally efficient whole-genome regression for quantitative and binary traits," *Nature Genetics,* vol. 53, p. 1097–1103, 2021.

[62] B. Devlin and K. Roeder, "Genomic Control for Association Studies," *Biometrics,* pp. 997-1004, 2004.

[63] Bulik-Sullivan, Brendan K et al., "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies," *Nature Genetics,* vol. 47, p. 291–295, 2015.

[64] The International HapMap Consortium, "A haplotype map for the human genome," *Nature,* vol. 437, p. 1299–1320, 2005.

[65] Fadista, João et al., "The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants," *European Journal of Human Genetics,* vol. 24, p. 1202–1205, 2016.

[66] T. e. a. Schoeler, "Participation bias in the UK Biobank distorts genetic associations and downstream analyses," *Nature Human Behaviour,* 2023.

[67] P. e. a. Turley, "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nature Genetics,* vol. 50, p. 229–237, 2018.

[68] G. Lettre and J. D. Rioux, "Autoimmune diseases: insights from genome-wide association studies," *Oxford Journals Human Molecular Genetics,* pp. 116-121, 2008.

[69] Tewhey, Ryan et al., "Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay," *Cell,* vol. 165, no. 6, pp. 1519-1529, 2016.

[70] Mouri, Kousuke et al., "Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells," *Nature Genetics,* vol. 54, p. 603–612, 2022.

[71] Okada, Yukinori et al., "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature,* vol. 506, p. 376–381, 2013.

[72] Imamura, Minako et al., "Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes," *Nature Communications,* vol. 7, 2016.

[73] K. Sonehara and Y. Okada, "Genomics-driven drug discovery based on disease-susceptibility genes," *Inlammation and Regeneration,* vol. 41, 2021.

[74] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies," *Nature Reviews Genetics,* vol. 11, pp. 499-511, 2010.

[75] Tewhey, Ryan et al., "The importance of phase information for human genomics," *Nature Reviews Genetics,* vol. 12, p. 215–223, 2011.

[76] S. R. Browning and B. L. Browning, "Haplotype phasing: existing methods and new developments," *Nature Review Genetics,* vol. 12, p. 703–714 , 2011.

[77] Menozzi, P et al., "Synthetic Maps of Human Gene Frequencies in Europeans," *Science,* vol. 201, no. 4358, pp. 786-792, 1978.

[78] Price, Alkes L et al., "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics,* vol. 38, p. 904–909, 2006.

[79] Patterson, Nick et al., "Population Structure and Eigenanalysis," *PLOS Genetics,* 2006.

[80] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine,* vol. 2, pp. 559-572, 1901.

[81] H. Hotelling, "Analysis of a Complex of Statistical Variables Into Principal Components," *Journal of Educational Psychology,* vol. 24, pp. 417-441; 498-520, 1933.

[82] Reed, Eric et al., "A guide to genome-wide association analysis and post-analytic interrogation," *Statistics in Medicine,* vol. 34, no. 28, pp. 3769-3792, 2015.

[83] M. Butkiewicz and S. W. Bush, "In Silico Functional Annotation of Genomic Variation," *Current Protocols in Human Genetics,* vol. 88, 2018.

[84] T. 1. G. P. Consortium, "A global reference for human genetic variation," *Nature,* vol. 526, p. 68–74, 2015.

[85] Bansal, Vikas et al., "Statistical analysis strategies for association studies involving rare variants," *Nature Reviews Genetics,* vol. 11, p. 773–785, 2010.

[86] Pabringer, Stephan et al., "A survey of tools for variant analysis of next generation genome sequencing data," *Briefings in Bioinformatics,* vol. 15, pp. 256-278, 2012.

[87] Mahmoud, Medhat et al., "Structural variant calling: the long and the short of it," *Genome Biology,* vol. 20, 2019.

[88] H. Yang and K. Wang, "Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR," *Nature Protocols,* vol. 10, p. 1556–1566, 2015.

[89] De Baets, Greet et al., "SNPeffect 4.0: on-line prediction of molecular and," *Nucleic Acid Research,* vol. 40, 2012.

[90] Hunt, Sarah E. et al., "Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor—A tutorial," *Human Mutation,* vol. 43, no. 8, pp. 986-997, 2021.

[91] Harrow, Jennifer et al., "GENCODE: The reference human genome annotation for The ENCODE Project," *Cold Spring Harbor Laboratory Press,* vol. 22, pp. 1760-1774, 2012.

[92] Pruitt, Kim D et al., "RefSeq: an update on mammalian reference sequences," *Nucleic Acid Research,* vol. 42, 2014.

[93] Adzhubei, Ivan et al, "A method and server for predicting damaging missense mutations," *Nature Methods,* vol. 7, pp. 248-249, 2010.

[94] Landrum, Melissa J et al., "ClinVar: improvements to accessing data," *Nucleic Acids Research,* vol. 48, 2019.

[95] Rentzsch, Philipp et al., "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Research,* vol. 47, 2019.

[96] Leitsalu, Liis et al., "Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu," *Interntional Journal of Epidemiology,* vol. 44, no. 4, p. 1137–1147, 2014.

[97] jia, Xiaoming et al., "Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens," *PLOS ONE,* 2013.

[98] Butler-Laporte, Guillaume et al., "HLA allele-calling using whole-exome sequencing identifies 129 novel associations in 11 autoimmune diseases: a multi-ancestry analysis in the UK Biobank (preprint)," *medRxiv,* 2023.

[99] Z. e. a. Gu, "ComplexHeatmap Complete Reference," [Online]. Available: https://jokergoo.github.io/ComplexHeatmap-reference/book/upset-plot.html.

[100] Bottini, Nunzio et al., "A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes," *Nature Genetics,* vol. 36, p. 337–338, 2004.

[101] Smyth, Deboral et al., "Replication of an Association Between the Lymphoid Tyrosine Phosphatase Locus (LYP/PTPN22) With Type 1 Diabetes, and Evidence for Its Role as a General Autoimmunity Locus," *Diabetes,* vol. 53, no. 11, p. 3020–3023, 2004.

[102] Ladner, Martha B. et al., "Association of the single nucleotide polymorphism C1858T of the PTPN22 gene with type 1 diabetes," *Human Immunology,* vol. 66, no. 1, pp. 60-64, 2005.

[103] Qu, H et al., "Confirmation of the association of the R620W polymorphism in the protein tyrosine phosphatase PTPN22 with type 1 diabetes in a family based study," *Journal of Medical Genetics,* vol. 42, pp. 266-270, 2005.

[104] Zheng, Weipeng et al., "Genetic Association Between a Lymphoid Tyrosine Phosphatase (PTPN22) and Type 1 Diabetes," *Diabetes,* vol. 54, no. 3, p. 906–908, 2005.

[105] Bottini, Nunzio et al., "Role of PTPN22 in type 1 diabetes and other autoimmune diseases," *Seminars in Immunology,* vol. 18, no. 4, pp. 207-213, 2006.

[106] J.-F. Cloutier and A. Veillette, "Cooperative inhibition of T-cell antigen receptor signaling by a complex between a kinase and a phosphatase," *Journal of Experimental Medicine,* vol. 189, no. 1, pp. 111 - 121, 1999.

[107] J.-R. Hwang, "Recent insights of T cell receptor-mediated signaling pathways for T-cell activation and development," *Experimental & Molecular Medicine,* vol. 52, p. 750–761, 2020.

[108] Burn, Garth L et al., "Why is PTPN22 a good candidate susceptibility gene for autoimmune disease?," *FEBS Letters,* vol. 585, no. 23, pp. 3689-3698, 2011.

[109] Sanchez-Blanco, Cristina et al., "Protein tyrosine phosphatase PTPN22 regulates LFA-1 dependent Th1 responses," *Journal of Autoimmunity,* vol. 94, pp. 45-55, 2018.

[110] Bjørnvold, M et al., "Joint effects of HLA, INS, PTPN22 and CTLA4 genes on the risk of type 1 diabetes," *Diabetologia,* vol. 51, no. 4, pp. 589-596, 2008.

[111] Portuesi, Rosalba et al., "Assessment of Type 1 Diabetes Risk Conferred by HLA-DRB1, INS-VNTR and PTPN22 Genes Using the Bayesian Network Approach," *PLOS ONE,* 2013.

[112] Morgan, Ann W. et al., "Reevaluation of the interaction between HLA–DRB1 shared epitope alleles, PTPN22, and smoking in determining susceptibility to autoantibody-positive and autoantibody-negative rheumatoid arthritis in a large UK Caucasian population," *Arthritis & Rheumatology,* vol. 60, no. 9, pp. 2565-2576, 2009.

[113] Chinoy, H. et al., "The protein tyrosine phosphatase N22 gene is associated with juvenile and adult idiopathic inflammatory myopathy independent of the HLA 8.1 haplotype in

British Caucasian patients," *Arthritis & Rheumatology*, vol. 58, no. 10, pp. 3247-3254, 2008.

[114] A. T. Lee, "The PTPN22 R620W polymorphism associates with RF positive rheumatoid arthritis in a dose-dependent manner but not with HLA-SE status," *Genes & Immunity*, vol. 6, p. 129–133, 2005.

[115] Li, F. et al., "A powerful latent variable method for detecting and characterizing gene-based gene-gene interaction on multiple quantitative traits," *BMC Genetics*, vol. 14, 2013.

[116] J. e. a. Li, "Identification of gene-gene interaction using principal components," *BMC Proceedings,* vol. 3, 2009.

[117] T. e. a. Cui, "Gene–gene interaction detection with deep learning," *Nature Communications Biology,* vol. 5, 2022.

[119] J. Lempainen and R. Veijola, "The heterogeneous pathogenesis of type 1 diabetes mellitus," *Nature Reviews Endocrinology* , vol. 15, pp. 635-650, 2019.

[120] J. F. Cloutier and A. Veillette, "Association of inhibitory tyrosine protein kinase p50csk with protein tyrosine phosphatase PEP in T cells and other hemopoietic cells.," *The EMBO Journal,* vol. 15, pp. 4909-4918, 1996.

[121] Kirino, Yohei et al., "Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1," *Nature Genetics,* vol. 45, pp. 202-207, 2013.

[122] Hughes, Travis et al., "Evidence for gene–gene epistatic interactions among susceptibility loci for systemic lupus erythematosus," *Arthritis & Rheumatology,* vol. 64, no. 2, pp. 485-492, 2011.

[123] Matzaraki, Vasiliki et al., "The MHC locus and genetic susceptibility to autoimmune and infectious diseases," *Genome Biology,* vol. 18, 2017.

[124] Miles, John J et al., "Understanding the complexity and malleabilityof T-cell recognition," *Immunology and Cell Biology,* vol. 93, pp. 433-441, 2015.

[125] Tuteja, Sachleen et al., "A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequence (NGS) based genetic testing," *Journal of Pathology Informatics,* vol. 13, 2022.

[126] Zhu, Jinfang et al., "Differentiation of Effector CD4 T Cell Populations," *Annual Reviews Immunology,* vol. 28, pp. 445-449, 2010.

[127] Tizaoui, Kalthoum et al., "Genetic Polymorphism of PTPN22 in Autoimmune Diseases: A Comprehensive Review," *Medicina,* vol. 58, no. 8, p. 1034, 2022.

NON-EXCLUSIVE LICENCE TO REPRODUCE THESIS AND MAKE THESIS PUBLIC

I, Arne Kukkonen,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

„Genome-Wide Association Study for Detecting Autoimmune-Disease-Associated Genetic Pattern Differences in Specific HLA Type Carriers“,

supervised by Erik Abner, PhD.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Arne Kukkonen

26/05/2023