The Role of the HLA Gene Region and Environmental Risk Factors in Follicular Non-Hodgkin Lymphoma

by

Nicholas Kipp Akers

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Health Sciences

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martyn T. Smith, Chair
Professor Lisa F. Barcellos
Professor Stephen M. Rappaport
Professor Christine F. Skibola

Spring 2014

Abstract

The Role of the HLA Gene Region and Environmental Risk Factors in Follicular Non-Hodgkin Lymphoma

by

Nicholas Kipp Akers

Doctor of Philosophy in Environmental Health Sciences

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Martyn T. Smith, Chair

The first genome-wide scans searching for follicular lymphoma (FL) risk factors revealed that a section of chromosome 6 powerfully impacts risk of this disease. Common genetic variants within the human leukocyte antigen (HLA) gene region were shown to be associated with an approximate doubling of individual disease odds. This dissertation aims to concurrently improve the resolution of, expand upon, clarify, and take the first steps in explaining these findings. Chapter 1 provides a review of the broadly relevant literature, including the epidemiology of FL and related lymphomas, the molecular immunology of FL, and the HLA gene region. Chapter 2 is a study making use of the highest possible resolution HLA genotyping methodology for its time, applied to an FL case-control study. This study not only increased our knowledge of known risk factors, it also was the first study to demonstrate an association of FL with variation at *HLA-DPB1*. Chapter 3 describes the method which will soon be used to localize to a single locus the associations which are ambiguously assigned to several genes. Using pilot data, this study demonstrate the feasibility of performing genetic ancestry matching and HLA imputations on historically stored samples. Chapter 4 uses data from several studies to identify two amino-acid positions, which may themselves explain a substantial portion of FL risk. The fact that these amino acid positions lie in the key peptide binding groove of *HLA-DRB1* gives some evidence that peptide binding is the mechanism by which these HLA associations are impacting FL development. Finally, in Chapter 5 the peptide binding properties of HLA class II alleles are computationally investigated, examining potential environmental and internal proteomes likely to be encountered by HLA proteins. This approach reveals that certain alleles which impact FL risk are predicted to be exceptionally strong or weak at binding peptides, and several candidate antigens are mined from the data. Concluding in Chapter 6, the state of HLA-FL research is summarized, and future research is recommended.

Dedication

I dedicate this dissertation to my loving grandparents, Carl and Nickie Akers. You're my greatest source of inspiration and you mean the world to me. I don't think I would have made it this far without you.

Table of Contents

Acknowledgements

This dissertation, and my identity as an academic researcher, is the product of collaboration with an amazing group of individuals. Let me thank some of the key figures that led to this work:

I am in debt to my dissertation committee for the feedback and encouragement which has drastically improved this work.

John Curry, you taught me what it means to work hard in the lab, how to make a hypothesis, how to perform an experiment. Much of what transpired here has its roots in conversations we had years ago. Lucia Conde, for teaching me the value of computers in biology, I suspect I won't forget any time soon. Fenna Sillé, thank you for being there to talk immunology whenever I needed, you're a fantastic person to work with and I hope you have your own lab someday soon, you deserve it!

Martyn Smith, you've been a fantastic calming force throughout the years we've worked together. You've taught me a million things, but I'm mostly thankful that you managed to guide me to where I am today.

Chris Skibola, in the massive world of cancer research it can be easy to feel insignificant, but you've always made time for me. For that, I thank you. You taught me how to succeed in a competitive world. The lessons you've given me have not always been easy, but I'm forever grateful for them.

My family; you are there when I need you, and you've given me the freedom to grow on my own. I'm so proud to be a part of our bizarre clan.

My friends; you kept me sane in the worst times, and without you I would not have had my best times. Thank you for your love and acceptance, it means everything to me.

Chapter 1

General Introduction

The Global Burden of Cancer and Non-Hodgkin Lymphoma

Carcinogenesis is a global killer, responsible for an estimated 8.2 million deaths in 2012. Individuals living today carry a greater than 10% chance of death from cancer before age 75 (1). Within the United States, malignant neoplasms represented 23% of all deaths in 2011 (2). Discovering the environmental causes of cancers remains a crucial public health goal. Tobacco smoke stands as a prime example of the ongoing benefits associated with the discovery of carcinogenic environmental exposures, with 157 million years of life saved in the half-century following the US surgeon general's first report on smoking and health in the United States (3).

Similar to lung cancer, non-Hodgkin lymphoma (NHL) incidence went up at an alarming rate from the 1960's until the early 1990's in the United States and elsewhere in the world (4,5). This trend would appear to indicate an environmental causal agent that grew more prevalent in this time-frame. Unfortunately, however, the cause of this period of increased incidence remains mostly unknown (6). Among United States cancer types, NHL was the 7th most frequent cancer by incidence (69,740 new cases) and total number of deaths caused (19,020 deaths) in 2013 (7).

NHL encompasses a heterogeneous group of cancers of the lymphatic system, excluding Hodgkin lymphoma. Follicular lymphoma (FL) is the second most common subtype of NHL, after diffuse large B-cell lymphoma (DLBCL). The exact proportion of NHL encompassed by FL varies, however, it appears that FL composes 15-32% of NHL cases, depending on the location (8–11). Very little research has been published regarding environmental risk factors specific to FL; as a result, the epidemiological literature for NHL as a whole will be examined here.

Classic epidemiology has revealed clues to the etiology of NHL. Men are at a slightly increased risk of disease, with 1.4 cases in men for every case in women (12). Age is a major risk factor, with incidence and mortality rates increasing to a peak after age 70 (7). There is a major trend towards higher NHL incidence in developed countries, with the U.S./Canada, Australia/New Zealand, and Europe having the highest incidences, and Asian and African countries containing the lowest incidences (13). NHL incidence in Caucasian-Americans is higher than in African-Americans, whose incidence, in turn, is much higher than that of most African countries (11). It is unclear if these effects are the result of differential reporting, environmental, or genetic effects; however, ample evidence exists indicating the latter two are key components of the disease.

Environmental Causes of NHL

The precipitous rise of NHL cases led to many studies investigating environmental causes of lymphoma. The strongest and most relevant findings will be presented here, however, excellent reviews exist which summarize the literature more completely (5,14).

Immunodeficiency

Immune deficiency is one of the best described and clearest risk factors for NHL. Transplant recipients have a relative risk of 20-120, depending on the site of transplant (12). HIV patients have a 14-350 fold increased risk of NHL, depending on the grade of NHL examined. Individuals born with genetic immune deficiencies are also at a great increased risk for NHL. For example, subjects with Wiscott-Aldrich Syndrome that live to age 30 approach 100% risk of NHL. It is clear that immune deficiency can be a causal factor for NHL and much can be learned from these examples. However, due to the rarity of the above exposures, the vast majority of NHL cases cannot be explained by these risk factors.

Infections

Infectious agents are another well-established risk factor which can currently explain only a small percentage of NHL cases. Some infectious agents are able to directly transform lymphocytes into lymphomas, including Epstein Barr Virus (EBV) (linked to Burkitt's lymphoma), human herpesvirus 8 (primary effusion lymphoma), and human T lymphotrophic virus type I (adult T-cell lymphoma) (15). In each of these cases, infection of lymphoma cells themselves has been demonstrated.

A second group of pathogens appear to cause lymphoma indirectly, by chronic infection. Although hepatitis C virus (HCV) is known to primarily infect liver hepatocytes, the virus appears to play a major role in some cases of lymphoma. Patients suffering from splenic lymphoma with villous lymphocytes who are positive for HCV undergo complete remission of the lymphoma after treatment with antiviral drugs. The same treatment has no effect on those suffering from the same disease but who are negative for HCV infection (16). This finding implies that HCV is necessary for some cases of lymphoma, despite no evidence that HCV infects lymphoma cells themselves. Similarly, *Helicobacter pylori* infection appears to have a strong relationship with gastric mucosa associated lymphoid tissue (MALT) NHL as well as non-MALT gastric NHL. It has been demonstrated that in gastric NHL, *H. pylori* infection precedes disease in 85% of cases (17). In most cases of gastric MALT NHL who test positive for *H. pylori*, antibiotic eradication of the bacterial infection leads to remission of the tumor (18). Similar associations of bacterial infections and rare NHL subtypes include *Campylobacter jejuni* (small intestine NHL), *Chlamydia psittaci* (ocular adnexa NHL), and *Borrelia afzelii* (cutaneous NHL) (15). It may be that other infections are contributing to more common subtypes of NHL, however, very little conclusive evidence for this exists.

Chemical Exposures

Chemical exposures have long been suspected as causal factors in NHL etiology, however, proving such a link has been difficult. Certain occupations, including farmers, forestry workers, and petroleum, plastics, rubber and synthetics industry workers appear to carry an increased risk of NHL. As a result, pesticides and organic solvents including benzene have been implicated (5), though reports have been inconsistent. For example, a 2007 review found that 93% of studies observe some elevation of NHL risk with benzene exposure yet only 53% of studies contain statistically significant associations. The authors suggest the healthy worker effect is responsible for driving a true association between benzene and NHL towards the null (19). Conversely, further meta-analyses attempting to answer the same question (20–22) found no significant increased risk of NHL in benzene exposed populations. Despite this, the benzene hypothesis remains attractive for several reasons. As a well characterized cause of acute myelogenous leukemia, benzene is a known carcinogen that targets cells in the bone marrow, likely causing chromosomal aberrations (23). Similar to leukemia, the cells of B-cell lymphomas differentiate in the bone marrow, and the FL subtype is characterized by a common chromosomal translocation (24). It may be the case that benzene exposure leads to specific NHL subtypes, but the association has been obscured by non-associated subtypes. Similar to benzene, exposure to pesticides, solvents, triclorethylene, styrene, nitrates, vinyl chloride and asbestos have been examined with relation to NHL in multiple studies, with heterogeneous findings. Recently, environmental tobacco smoke exposure was associated with FL, but not other subtypes of NHL (25).

Follicular Lymphoma Genetic Risk Factors

Much can be inferred from genetic studies of NHL, including common polymorphisms that put individuals at greater risk, and mutations that appear more commonly in tumor cells. Both serve as evidence of the molecular mechanisms at play in lymphomagenesis.

Perhaps the signature genetic event arising in FL tumor cells is a translocation of the short arms of chromosomes 14 and 18. This cytogenetic defect is present in about 90% of FL tumors(24), and appears to indicate a distinct subtype of diffuse large B cell lymphoma (DLBCL)(26). The primary effect of this translocation is to place the promoter for an immunoglobulin gene in front of the anti-apoptotic protein BCL2, leading to increased expression of the oncogene. What causes the t(14;18) translocation is widely debated, however, it is clear that this mutation is not sufficient to cause disease, as the translocation can be detected in the circulating blood of nearly 50% of the human population(24).

Based on familial aggregation and case-control studies, we can infer that NHL has a small, but important genetic basis for susceptibility. For example, individuals with a history of a first-degree relative with NHL are at a 1.7-fold increased risk of disease (10). Although this effect could reflect shared environmental factors, other research implicates at least a partial genetic effect. Case-control studies of NHL matched on ethnicity and other demographics have revealed key risk loci for the disease. DNA extracted from blood, saliva, or buccal cells is considered germ line, meaning it is the same sequence that an individual was born with. Because of this, researchers can be confident that any differences observed in DNA sequences between cases and controls predated disease onset.

Many genetic susceptibility factors for NHL have been published, however, genetic loci that have been validated in multiple studies are somewhat rare, and can be grouped into two major categories. Not surprisingly, polymorphisms affecting DNA repair genes make up one of these groups. Specifically, researchers have found that genetic changes which affect VDJ recombination breakpoint repair and recombination repair affect risk of NHL (27–29). DNA damage is a nearly ubiquitous aspect of cancer, and chromosomal translocations play a major role in lymphomas, so these findings do not come as a shock.

Polymorphisms in genes responsible for immune response, including the human leukocyte antigen (HLA) region, form a second group of susceptibility loci. Risk of DLBCL has been associated with genetic variants in the TNF and IL10 genes(30). The proteins these genes encode play key roles in immunoregulation and inflammation. Furthermore, polymorphisms in the CD40 and CD154 encoding genes have also been linked to NHL (31). These proteins play a key role in B cell growth and differentiation (Figure 1). In 2009, the first genome-wide association study (GWAS) of NHL was published (32), reporting a single nucleotide polymorphism (SNP) in the HLA class I region associated with FL (rs6457327, allelic OR = 0.59, *p* = 4.7×10^{-11}). A second GWAS revealed a second FL risk allele in the HLA class II region (33). This finding was validated in 8 study populations on 3 different continents, and has a population attributable risk (34) of ~9% using the risk and frequency data from that GWAS, indicating a substantial fraction of cases would not have occurred in the absence of this risk factor. Follow-up sequencing demonstrated that several HLA class II alleles may be associated with FL (see references (33,35) and chapter 2). The classic HLA genes are responsible for the presentation of proteins in the body as a mechanism of immune surveillance. At first daunting, these proteins, which appear to play a role in NHL risk, are all part of a group of key players in the germinal center reaction responsible for stimulating B cell growth and development.

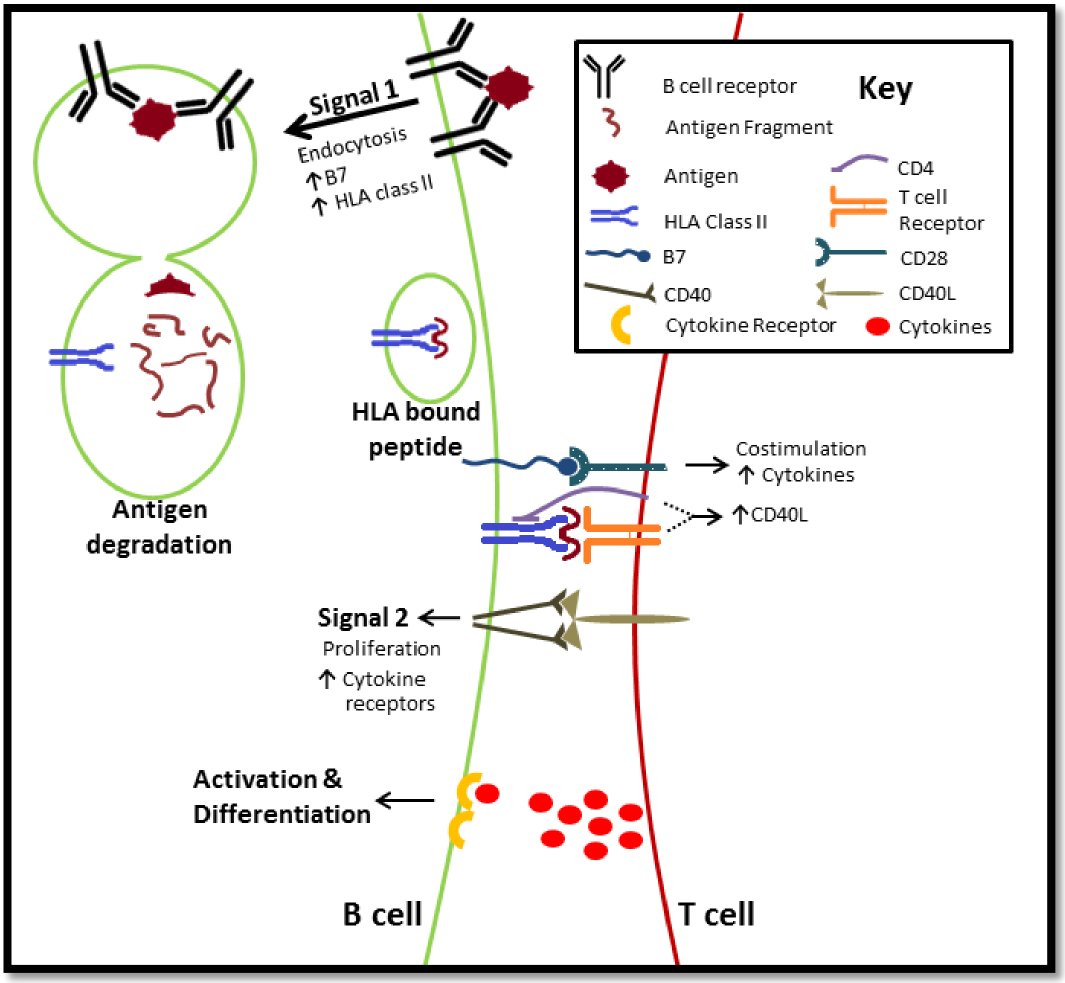


Figure 1: Proliferation and Differentiation Signals in B cells. Non-differentiated B cells require two signals to avoid apoptosis. The first signal comes from cross-linking of B-cell receptors when binding antigen. This leads to increased expression of B7 and HLA class II proteins, as well as endocytosis of the bound antigen. The antigen is degraded when the endosome fuses with a lysosome. A peptide fragment of that degraded antigen will then be bound to HLA class II proteins, which are brought to the surface. At the surface, HLA class II bound peptide interacts with the T cell receptor of CD4+ T cells. If the interaction is strong enough, CD4 will bind to the HLA class II protein, and B7 will act as a ligand for CD28. This causes increased expression of CD40L and cytokines (including IL2, IL4, and IL5) by the T cell. CD40L will then act as a ligand for CD40, which is the second signal needed by the B cell to avoid apoptosis. Cytokine receptors are expressed on the B cell, and it is the interaction of cytokines with their receptors which leads to activation and differentiation of B cells into proliferating effector B cells.

Molecular Biology of Follicular Lymphoma

Follicular lymphoma is a malignant growth of a B cells, which appear in a nodular, germinal center-like state (36). Much has been inferred about the development of FL cells based on the characteristics of these cells at the time of disease presentation. In order to appreciate the unique characteristics of a FL B cell, a background on the life cycle of a normal B cell is necessary.

B cell Maturation and the Germinal Center

B cells are the antibody producers of the immune system. Antibodies, also referred to as the B cell receptor when membrane-bound, are encoded by immunoglobulin genes and act to specifically target foreign bodies so they may be eliminated. B cells begin development in the bone marrow, where VDJ recombination, a process of genetic rearrangement, creates a unique DNA sequence for the B cell receptor. These cells are tested for auto-reactivity; those bearing antibodies which bind too strongly to self-proteins are edited or removed via apoptosis. Those B cells that are not self-reactive leave the bone marrow and enter the circulation, in order to encounter an antigen with their B cell receptor. An antigen is a protein which invokes an immune response, such as protein from virus, bacteria, fungi, or a different species. Often, antigens are brought to B cells by dendritic cells. When a B cell encounters an antigen, the antigen causes cross-linking of B cell receptors, the first of two important signals that prevent B cell apoptosis. With this first signal, the B cell migrates to a germinal center (Figure 2) (37).

During the germinal center reaction the B cell receptor's affinity for antigen is enhanced, and allows the B cell to differentiate into antibody producing plasma cells or long-lived memory B cells. The first step of this process is known as affinity maturation, where B cells rapidly divide while undergoing somatic hypermutation, a process where point mutations, insertions, and deletions accumulate in the immunoglobulin gene sequence. Many B cells are generated with slight changes to their antibody structure. Those B cells arising from this process with lower affinity for the antigen undergo apoptosis, while those with improved affinity will survive to present antigen fragments in their HLA class II receptor to helper T cells (Figure 2). This interaction with T cells is the second signal that causes a B cell to avoid apoptosis and proliferate. Specifically, the HLA class II-peptide complex is bound by the T cell receptor, which has similarly been trained to react with self-HLA proteins binding non-self peptides. Finally, B cells undergo class switching and mature into memory or plasma cells. Class switch recombination is a DNA excision event that allows different classes of antibody to be produced. There are five major classes of immunoglobulin, each with unique roles within the immune system (37,38).

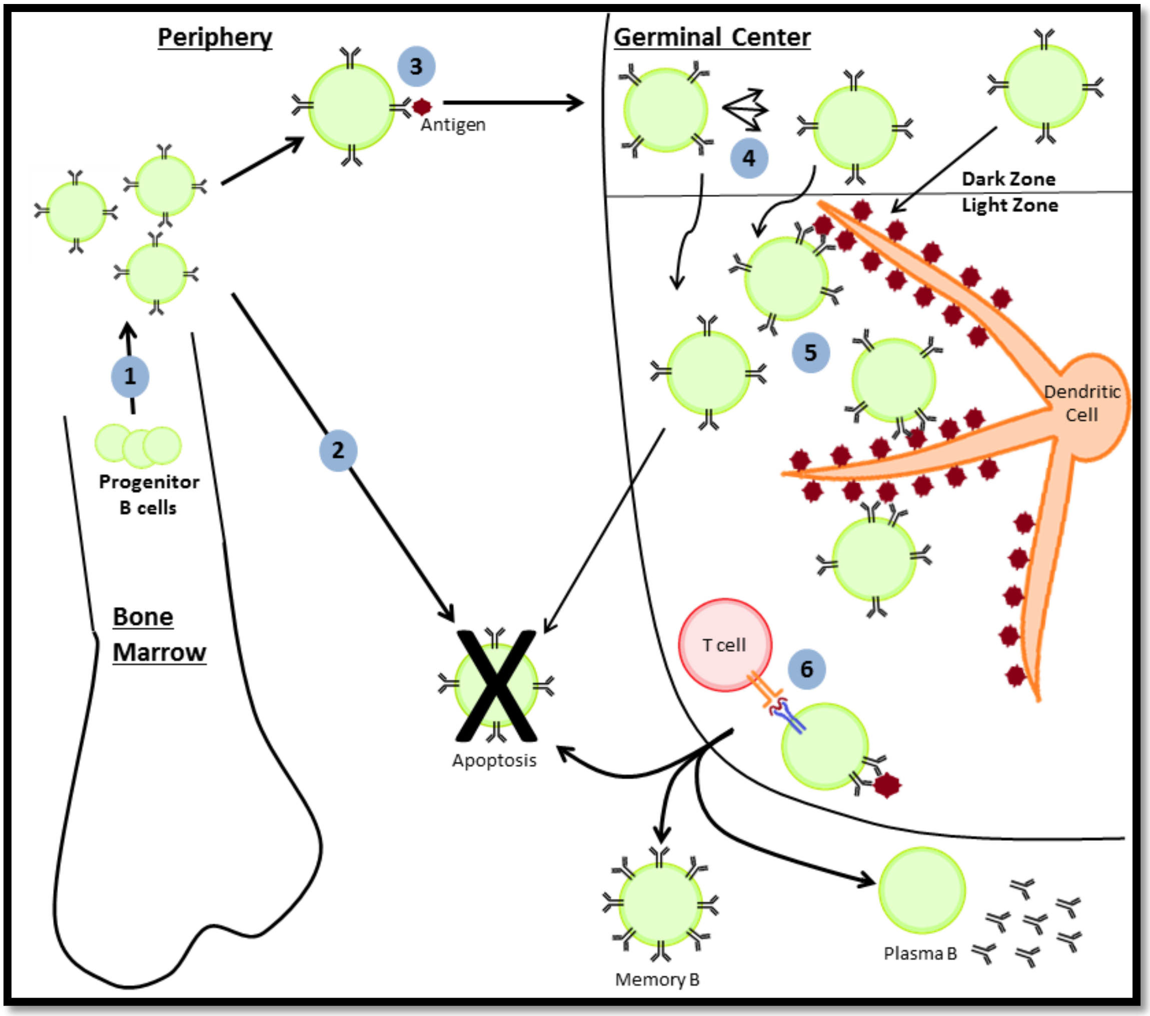


Figure 2: Life Cycle of a B cell. 1. B cells develop in the bone marrow, where they undergo V(D)J recombination to create unique B cell receptor sequences. The t(14:18) chromosomal translocation is thought to arise from errors in this recombination. Within the bone marrow, B cell receptors which react too strongly to self protein are selected against. 2. Mature, naive B cells exit the bone marrow and enter the periphery at a rate of ~5 x 10^6 per day. The vast majority (~90%) do not encounter antigen and undergo apoptosis. 3. Those that do encounter antigen migrate into a germinal center. 4. Within the dark zone of the germinal center, B cells rapidly proliferate while undergoing somatic hypermutation. This process creates point mutations, insertions, and deletions in the B cell receptor sequence in order to create clones with increased affinity for antigen. 5. As the B cells migrate from the dark zone to the light zone, their proliferation is now dependent on interactions with dendritic cells and T cells. Follicular dendritic cells carry many antigens to the germinal center on their cell surface, and B cells must compete to bind antigen. B cell receptors with high affinity for antigen will successfully bind antigen, while low affinity B cells will undergo apoptosis. 6. High-affinity B cells will then present peptides from the antigen on HLA class II proteins to a CD4+ T cell. This reaction can lead to class switch recombination of the immunoglobulin type, and differentiation into long lived memory B cells or plasma cells which secrete large amounts of antibody. A negative reaction with the T cell may also lead to apoptosis.

Antigens and Follicular Lymphoma

The aforementioned t(14:18) chromosomal translocation, present in nearly all FL tumors, as well as a large fraction of healthy individuals (24), is thought to occur very early in the life of a FL cell. The translocation involves the same immunoglobulin genes which are rearranged during VDJ recombination, a strong indication that the t(14:18) translocation is a malfunction of the early stage process. Roulland et al. demonstrated that the t(14:18) positive cells found in healthy individuals have unique characteristics resembling FL cells (39). FL B cells have undergone class switch recombination on both alleles, while normal cells undergoing the same process will only recombine the allele being expressed (40). When t(14:18) positive B cells in healthy individuals were examined, they, like FL B cells, had undergone class switch recombination on both alleles (39). These studies indicate that pre-FL cells may be arising at a fairly common rate in healthy individuals. It is therefore key to understand not only what influences the rate of t(14:18) positive B cells being produced, but also what impacts the transition from a non-malignant t(14:18) positive B cell to FL.

The fact that FL cells have undergone class switch recombination indicates that these cells have encountered antigen, and that they may also rely on antigen interaction for sustained growth and survival. Immunoglobulin gene sequencing of an FL patient indicated that mutations were accumulated sequentially, and preferentially towards non-synonymous changes (41). The finding of cells within an FL patient that share some, but not all, of the primary tumor cell mutations indicates that distinct populations of FL/pre-FL cells evolve within the body. Furthermore, it suggests these changes occur stepwise, rather than with a single round of somatic hypermutation. The mutations found in the immunoglobulin gene of FL cells also appeared to be more likely to cause amino-acid substitutions than would be expected by random chance. This statistical unlikelihood is explained by the authors as resulting from a selective force, likely clonal populations are being selected for the B cell receptor's affinity for antigen. Taken as a whole, this study indicates that FL cells encounter antigen at several points in their lifetime, undergoing somatic hypermutation to improve their affinity for the antigen.

A similar selective force appears to act on FL cells to ensure that they remain of the immunoglobulin M (IgM) class. Evidence exists that FL cells undergo the germinal center reactions of somatic hypermutation and class switch recombination several times. However, the majority of FL cells express surface IgM (39,40,42), indicating that this may also be an important factor to survival. IgM is one of 5 major classes of immunoglobulin, characterized as the first class produced in a primary response to antigen, and has the highest antigen binding power of all classes (37).

Researchers who have looked for the antigen specificity of FL cells have found some evidence of auto-reactivity. Dighiero et al. (43) tested the reactive potential of FL patient derived hybridomas to a panel of 5 common human proteins, finding that in 8 of 31 FL cases the immunoglobulin was specific enough to bind to at least one of the human proteins, and in 2 instances, bound to several of the human proteins. The weakness of this study is that reactivity was only tested against five human proteins; a much larger panel of proteins would be ideal. However, the large fraction of FL immunoglobulins reacting to the select panel would seem to indicate that the 26% reactive patients is actually an underestimate of the true number of FL patients with self-reactive immunoglobulins. Interestingly, immunoglobulin itself was the protein most commonly invoking a reaction. This may indicate that certain FL cells are self-reactive, and therefore self-reliant for antigenic stimulation. The aforementioned evidence that FL cells encounter antigen multiple times in their lifetime indicates the same antigen is present at multiple time points. A self-antigen satisfies this requirement, however, other routes may lead to multiple encounters with the same antigen, such as a chronic infection.

FL cells are dependent on germinal center interactions for survival. *In-vitro*, FL cells will die without co-culture of CD4+ T-cells (44). Perhaps not surprisingly, this proliferation appeared to be dependent on FL and T cells having similar antigen specificity. B cells were isolated from six FL patients and were tested for the ability to induce proliferation in two different CD4+ T cell clones. B cells from four patients induced proliferation in one of the T cell clones. When cultured with this T cell population, the four reactive B cell populations showed marked growth and proliferation, while the two non-reactive B cell populations did not. This effect could be interrupted by blocking the HLA class II receptor, which mediates T cell-B cell-antigen interactions. These findings indicate that FL proliferation is dependent on T cell interactions mediated by HLA class II. In a separate study, these T cell interactions were circumvented by treating FL cells with anti-CD40 antibody (to simulate CD154) and IL4 (45), two growth signals expected to be secreted in a germinal center reaction. The researchers found the FL cells to be resistant to death, relative to normal germinal center cells, but not proliferative until treated with molecules that simulate T cell feedback. Genetic changes such as the t(14:18) translocation will clearly cause cells to become resistant to death, but perhaps in FL, oncogenic proliferation is dependent on antigen-dependent reactions.

Recent studies have indicated that CD4+ T cell interactions may be more complicated than previously described. B-cell NHL tumors contain high levels of CD4+ regulatory T cells (T-regs) compared to control tissues, and these T-regs suppress proliferation of tumor fighting CD4+ and CD8+ T-cells (46,47). Under normal physiological conditions, T-regs function to limit immune response, as a means of preventing auto-immune reactions. In this case, however, they may be preventing the immune system from removing malignant cells. Evidence indicates that T-helper cells can be converted to T-regs by malignant FL B-cells in a process involving T-cell receptor stimulation (48). Because HLA class II molecules interact with the T-cell receptor, this suggests a second mechanism by which these proteins may couple with antigen to impact FL risk.

HLA Alleles and Immune Response

HLA proteins appear key to the proliferation of FL cells based on both molecular examinations of FL cells and genetic association studies. This gene family is well known for playing a prominent role in the immune system, for the population variability it contains, and for impacting the risk of numerous common diseases. These topics and their relevance to FL risk will be explored.

Molecular Role of HLA

The primary role of HLA molecules is to present antigen fragments to T cells as an immune surveillance mechanism. HLA class I proteins are expressed on most cell types and present primarily intracellular peptides. This serves to monitor for cellular infections with intracellular pathogens, as well as tumor antigens (49). Under normal conditions, HLA class II proteins are expressed only on professional antigen presenting cells, including B cells, dendritic cells, and macrophages. These cells are responsible for the uptake and presentation of extracellular antigens. Following endocytosis or phagocytosis, antigens are degraded within the endosome or lysosome into peptide fragments. These fragments are then loaded onto HLA class II molecules. Class I proteins are restricted to presentation of peptides 8-10 amino acids in length, however, class II proteins are able to present peptides of a wider range of 12-24 amino acids (50). These HLA bound peptides are brought to the cell surface to interact with T cells.

A positive HLA-T cell reaction is central to innate and humoral immune responses. In the case of B cell HLA class II interaction with CD4+ T cells, the result is a cascade of growth and proliferation signals, the details of which are outside the scope of this thesis. Briefly however, when the T cell receptor docks onto HLA class II, CD4 binds to the HLA class II protein. CD40 and the B7 family of proteins on the B cell interact with CD40L and CD28 on the T cell, respectively (Figure 1). These interactions are a form of bidirectional signaling, leading to proliferation of both interacting cells (37,51). Within the germinal center, this feedback system recognizes those B cells which have evolved (via somatic hypermutation) to most effectively target and present antigen. It is important then to understand what factors influence T cell response to HLA bound peptide.

HLA-T cell interaction is dependent on the T cell receptor affinity for the combination of HLA allele and bound peptide. What determines this affinity is still an important field of research. However, it is clear that how long a T cell receptor interfaces with HLA is key (52), with too short or too long interactions causing less proliferation. Other factors have also been demonstrated to be important to T cell activation, such as dynamic structural changes of the T cell receptor, weak and heterogeneous interactions between costimulatory molecules, and the distance between cell membranes (53–56). Together, these factors make prediction of T cell response to antigen extremely difficult.

Similar to the development of the B cell receptor, immature T cells rearrange their T cell receptor gene and undergo selection in the thymus based on T cell receptor affinity for HLA bound peptide. Specifically, T cells with too strong affinity for self-HLA protein are negatively selected (to prevent auto-immune responses), while T cells with too weak affinity for self-HLA protein do not mature and die off. This leaves only T cells with intermediate affinity which are driven to proliferate (50). The specificity that this process imparts on the circulating population of T cells has important consequences. T cell responses are not only antigen-specific, but also HLA protein specific. HLA proteins are encoded by numerous genes, each with numerous different alleles. Just as B cells are trained in the bone marrow to tolerate self-proteins, T cells are trained in the thymus to tolerate the HLA alleles carried by a given individual.

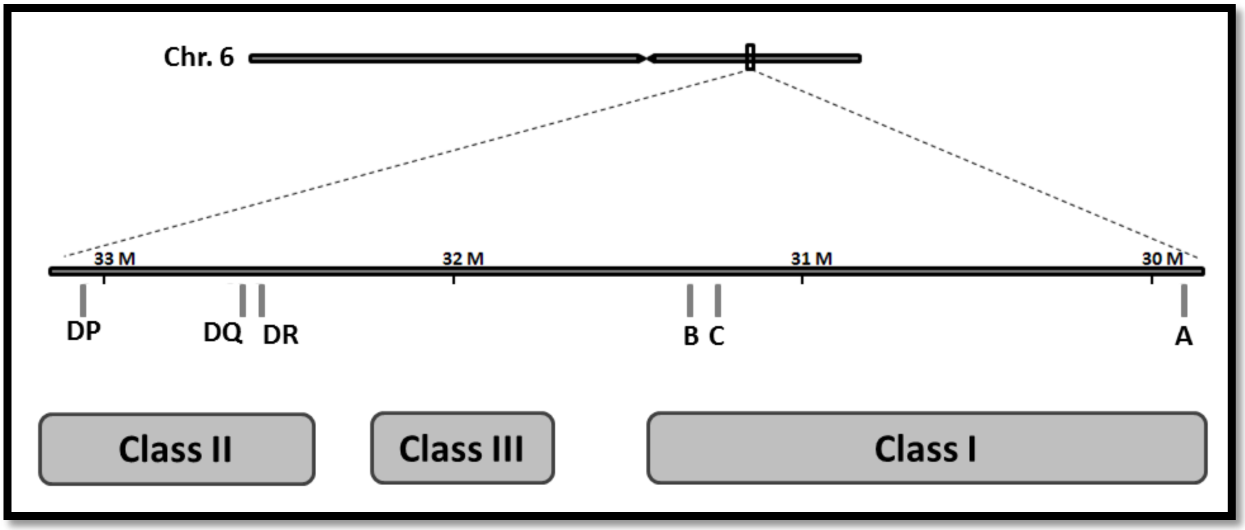


Figure 3. Map of the HLA region. The HLA region is located on the short arm of chromosome 6, spanning more than three million nucleotides. Class I genes, including *HLA-A, HLA-B, and HLA-C* are expressed on the surface of most cells, responsible for presenting peptides from intracellular antigens. Class II proteins are hetordimers, indicating that a functional *HLA-DR* protein requires amino acid chains from genes *HLA-DRA1* and *HLA-DRB1,* for example. Generally speaking, the most population sequence variation has been found on the beta genes, *HLA-DRB1, HLA-DQB1,* and *HLA-DPB1*. Class II proteins are responsible for presenting peptides from extracellular antigens. Class III includes the complement genes as well as TNF-α. The HLA region is very gene-rich, and many genes have been omitted from this diagram. Shown are the genes of focus in this dissertation.

Population Variability of the HLA Complex

Located on chromosome 6, the HLA complex spans over 3 million bases of the genome, and can be divided into 3 regions, named class I-class III (Figure 3). While the function of class I and II genes was covered in the previous section, class III genes are part of the complement system (50), and they will not be covered here. The genes of the class I and II regions display extensive sequence homology, indicative of a shared common ancestry. It has been postulated that the 6 genes and 10 pseudogenes of HLA class I, as well as the 11 genes and 8 pseudogenes of HLA class II resulted from extensive gene duplication events (57,58). Much of this duplication is likely to have occurred early in vertebrate evolution (where HLA genes are referred to by their more broad title, the major histocompatability complex [MHC]). This diversification of the proteins responsible for displaying antigen would have inferred a selective advantage by allowing individuals to present a broader spectrum of antigens, and thus improve immune response against infectious disease.

Similar logic can explain the value of populations carrying genetic diversity within a given HLA gene. Polymorphic alleles, i.e. different versions of a single gene, allow populations the benefit of greatly expanded variability in immune antigen presentation. As a result, an astounding number of alleles have been discovered in certain class I and class II genes (Table 1). The HLA gene region is the among the most polymorphic regions of the human genome (59,60), owing to the selective advantage for populations to diversify their alleles. Beyond pathogen-driven selection, selective mating may also play a role in generating the polymorphism of this region (61). Mice have been observed to mate non-randomly, with less observed homozygosity at MHC loci than would be expected by chance (62). Astoundingly, similar effects have been observed in humans, where studies indicate that women prefer the smell of HLA-mismatched men (63,64). A third source of HLA polymorphism likely comes from crosses of early *Homo sapiens* with closely related hominid species (65). The copious sequence variability observed in the HLA region does not however preclude extensive linkage among polymorphisms.

	Class I			Class II					
HLA Gene:	A	B	C	DRA	DRB1	DQA1	DQB1	DPA1	DPB1
Unique Alleles:	2,579	3,285	2,133	7	1,411	51	509	37	248
Unique Proteins:	1,833	2,459	1,507	2	1,047	32	337	19	205

Table 1: Polymorphism of the HLA region. The number of alleles (unique DNA sequence) and proteins (unique amino acid sequence) are shown for several HLA genes in class I and class II. These numbers represent the total number of unique sequences discovered, giving an indication of the extreme amount of population diversity within this gene group. This degree of polymorphism indicates that at this locus, population level diversity is positively selected for. This likely is due to the role of these proteins in the immune system recognizing infectious antigens. Data gathered from the IMGT/HLA database release 3.15 (66).

Linkage disequilibrium (LD) exists within the HLA region, often spanning multiple genes and obscuring causal loci in genetic association studies. The theoretical basis of LD is that mutations occur randomly, and are likely to arise just once. A new mutation will therefore be linked to all other polymorphisms on that chromosome, only separated by recombination occurring at a rate proportional to genetic distance (67). This simple model can be disrupted and LD generated by population effects including bottlenecks, genetic drift, and population mixture (68). When LD is high between multiple markers, a haplotype is said to exist within which genotypes of each polymorphism are highly informative of each other. Within the HLA region, extensive LD exists, particularly in Caucasian and Asian populations (69,70) . Common haplotypes link the genes HLA-DQB1, -DQA1, and -DRB1, as well as between HLA-B and –C. Lower LD exists between HLA-DRB1 and HLA-B. For example, within Caucasians, 14% of individuals carry the unique allele HLA-DRB1*15:01. Of these, 98% carry HLA-DQB1*05:01. In comparison, only 59% of DRB1*15:01 carriers also carry HLA-B*07:02 (69). If these alleles dissociated randomly, the overlap percentages would each be only 14%. This effect becomes problematic when a genetic association study discovers a locus within the HLA region impacting disease risk. The LD will obscure the exact locus causing the effect (assuming such an exact locus exists), and researchers are limited to knowing only that an association exists with some aspect of a given haplotype. This issue can sometimes be circumvented by looking in different populations, or by turning to studies of molecular effect.

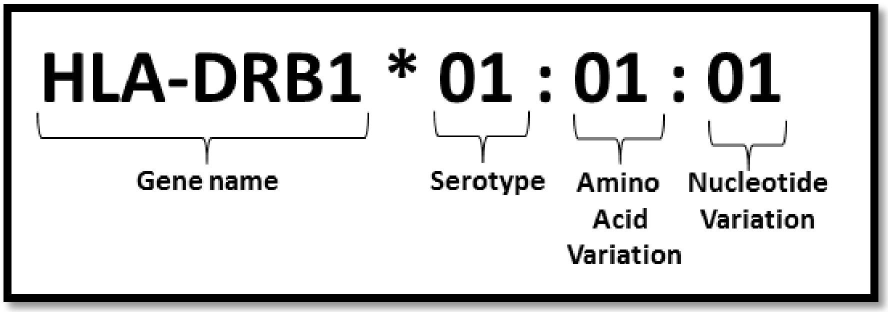


Figure 4. Understanding HLA Allele Nomenclature. With thousands of unique alleles, the HLA genes have specialized naming conventions. The gene name is listed first, followed by an asterisk (*). The first two digits after the asterisk indicate the serotype or allele group. This convention dates to genetic typing was not possible, and individuals were typed by immune response. The first four digits give all the information regarding amino acid sequence. Beyond 4 digits, the resolution is increased to the level of non-synonymous nucleotide changes. For example, HLA-DRB1*01:01:01 would have a different genetic sequence, but the same amino acid sequence, as HLA-DRB1*01:01:02.

HLA Disease Associations

Genetic association studies have linked variants in the HLA region with an overwhelming number of autoimmune and infectious diseases. Multiple sclerosis (71), narcolepsy (72), type 1 diabetes (73), malaria severity (74), progression to AIDS (75) and many others (76) have been associated with HLA genotypes. In many cases, these associations are believed to be attributable to an associated allele's ability to efficiently present key disease antigen, and thus illicit an immune response. In the examples of celiac disease the allele HLA-DQB1*02:01 is thought to uniquely present deamidated gliadin protein in a way that results in autoimmune attack in the small intestine (77). For the hepatitis B virus infection (HBV), HLA-DRB1*03:01 is linked with non-response to vaccination, possibly due to this allele’s low-affinity for HBV envelope ligands (78). In both cases, an allele puts carriers at increased risk of disease, likely because of inability to present key viral proteins to T cells.

Certain cancers have also been associated with HLA alleles, raising questions about the role of antigen presentation and immune response in these malignancies. Childhood B cell precursor acute lymphoblastic leukemia, a cancer with a similar cellular origin to FL, has been strongly associated with HLA-DPB1*06:01 (79). Although it has not been demonstrated, the authors surmise DPB1*06:01 may interact with a bioactive molecule to create an inflammatory, leukemia inductive state in the bone marrow. Cervical cancer, now known to be caused by infection with the herpes papilloma virus (HPV), also was associated with a DQB1-DRB1 haplotype (80). It is now understood that the HLA-cervical cancer association was actually due to these HLA alleles increasing likelihood of infection with the particularly carcinogenic HPV type 16. These varied and unique mechanisms of HLA-associated disease susceptibility indicate that it will not be a simple task to determine the molecular pathway on which HLA alleles impact FL risk.

As with cervical cancer, it may be that FL is caused directly by carcinogenic infection, and patients carrying certain HLA alleles are less able to clear the infection. A second, similar hypothesis is that infection is indirectly causing FL, by causing chronic inflammation, or by re-stimulating pre-FL B cells and inducing their continued growth and accumulation of mutations. It is important to realize that the antigen(s) indicated in FL may not be infectious in origin however. It may be that a self-antigen causes chronic stimulation under rare circumstances. Or that there is no single antigen to FL, that each pre-FL cell evolves within the body until it is responsive to any antigen. A simple explanation is that certain HLA alleles are in LD with promoter polymorphism which affects HLA gene expression and therefore FL risk. Finally, the complex interactions with CD4+ T cells, which are dependent on HLA class II protein, must not be ignored. If FL cells can reshape their tumor microenvironment, perhaps it is more the impact that an HLA allele has on the T cell which is crucial to FL development. These hypotheses are presented to demonstrate the breadth of questions that must now be asked with regards to the causes of FL.

It is clear that HLA class II alleles represent an ample opportunity for lymphoma researchers to learn about the origin of FL. However, before in-depth analyses of molecular dynamics, *in-vitro* examinations of HLA bound antigens with T cells, or molecular epidemiology to search for FL antigens can be performed, the fine details of the genetic association must be uncovered. Which allele or alleles affect FL risk, precisely? Within the associated haplotype, which gene contains the causal locus? Can the risk and protective associations be summarized by a single genetic site? By answering these questions, the monumental task of testing the aforementioned hypotheses can be reduced significantly. This thesis sets forth to answer these initial questions, as well as take the first steps towards understanding the molecular nature of HLA class II alleles impacting FL risk.

Bibliography

1. IARC. Fact Sheets by Population [Internet]. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012. [cited 2014 Jan 28]. Available from: http://globocan.iarc.fr/Pages/fact_sheets_population.aspx

2. National Vital Statistics Reports, Volume 61, Number 6, 10/10/2012 [Internet]. [cited 2014 Jan 28]. Available from: http://www.cdc.gov/nchs/data/nvsr/nvsr61/nvsr61_06.pdf

3. Holford TR, Meza R, Warner KE, Meernik C, Jeon J, Moolgavkar SH, et al. Tobacco control and the reduction in smoking-related premature deaths in the United States, 1964-2012. JAMA J Am Med Assoc. 2014 Jan 8;311(2):164–71.

4. Müller AMS, Ihorst G, Mertelsmann R, Engelhardt M. Epidemiology of non-Hodgkin's lymphoma (NHL): trends, geographic distribution, and etiology. Ann Hematol. 2005 Jan;84(1):1–12.

5. Ekström-Smedby K. Epidemiology and etiology of non-Hodgkin lymphoma--a review. Acta Oncol Stockh Swed. 2006;45(3):258–71.

6. Shiels MS, Engels EA, Linet MS, Clarke CA, Li J, Hall HI, et al. The epidemic of non-Hodgkin lymphoma in the United States: disentangling the effect of HIV, 1992-2009. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013 Jun;22(6):1069–78.

7. American Cancer Society. Cancer Facts & Figures 2013 [Internet]. Atlanta: American Cancer Society; 2013 [cited 2014 Feb 5]. Available from: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf

8. Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD, Linet MS. Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. Blood. 2006 Jan 1;107(1):265–76.

9. Anderson JR, Armitage JO, Weisenburger DD. Epidemiology of the non-Hodgkin's lymphomas: distributions of the major subtypes differ by geographic locations. Non-Hodgkin's Lymphoma Classification Project. Ann Oncol Off J Eur Soc Med Oncol ESMO. 1998 Jul;9(7):717–20.

10. Goldin LR, Landgren O, McMaster ML, Gridley G, Hemminki K, Li X, et al. Familial aggregation and heterogeneity of non-Hodgkin lymphoma in population-based samples. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2005 Oct;14(10):2402–6.

11. Groves FD, Linet MS, Travis LB, Devesa SS. Cancer surveillance series: non-Hodgkin's lymphoma incidence by histologic subtype in the United States from 1978 through 1995. J Natl Cancer Inst. 2000 Aug 2;92(15):1240–51.

12. Grulich AE, Vajdic CM. The epidemiology of non-Hodgkin lymphoma. Pathology (Phila). 2005 Dec;37(6):409–19.

13. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011 Apr;61(2):69–90.

14. Alexander DD, Mink PJ, Adami H-O, Chang ET, Cole P, Mandel JS, et al. The non-Hodgkin lymphomas: A review of the epidemiologic literature. Int J Cancer. 2007;120(S12):1–39.

15. Engels EA. Infectious Agents as Causes of Non-Hodgkin Lymphoma. Cancer Epidemiol Biomarkers Prev. 2007 Mar 1;16(3):401–4.

16. Hermine O, Lefrère F, Bronowicki J-P, Mariette X, Jondeau K, Eclache-Saudreau V, et al. Regression of splenic lymphoma with villous lymphocytes after treatment of hepatitis C virus infection. N Engl J Med. 2002 Jul 11;347(2):89–94.

17. Parsonnet J, Hansen S, Rodriguez L, Gelb AB, Warnke RA, Jellum E, et al. Helicobacter pylori Infection and Gastric Lymphoma. N Engl J Med. 1994 May 5;330(18):1267–71.

18. Wotherspoon AC, Doglioni C, Diss TC, Pan L, Moschini A, de Boni M, et al. Regression of primary low-grade B-cell gastric lymphoma of mucosa-associated lymphoid tissue type after eradication of Helicobacter pylori. Lancet. 1993 Sep 4;342(8871):575–7.

19. Smith MT, Jones RM, Smith AH. Benzene exposure and risk of non-Hodgkin lymphoma. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2007 Mar;16(3):385–91.

20. Alexander DD, Wagner ME. Benzene Exposure and Non-Hodgkin Lymphoma: A Meta-Analysis of Epidemiologic Studies: J Occup Environ Med. 2010 Feb;52(2):169–89.

21. Kane EV, Newton R. Benzene and the risk of non-Hodgkin lymphoma: A review and meta-analysis of the literature. Cancer Epidemiol. 2010 Feb;34(1):7–12.

22. Vlaanderen J, Lan Q, Kromhout H, Rothman N, Vermeulen R. Occupational Benzene Exposure and the Risk of Lymphoma Subtypes: A Meta-analysis of Cohort Studies Incorporating Three Study Quality Dimensions. Environ Health Perspect. 2010 Sep 29;119(2):159–67.

23. International Agency of Research on Cancer. IARC Monographs on the evaluation of carcinogenic risks to humans: Benzene. International Agency of Research on Cancer, Lyon; 2012. 249-294 p.

24. Staudt LM. A closer look at follicular lymphoma. N Engl J Med. 2007 Feb 15;356(7):741–2.

25. Diver WR, Teras LR, Gaudet MM, Gapstur SM. Exposure to Environmental Tobacco Smoke and Risk of Non-Hodgkin Lymphoma in Nonsmoking Men and Women. Am J Epidemiol. 2014 Feb 24;

26. Huang JZ, Sanger WG, Greiner TC, Staudt LM, Weisenburger DD, Pickering DL, et al. The t(14;18) defines a unique subset of diffuse large B-cell lymphoma with a germinal center B-cell gene expression profile. Blood. 2002 Apr 1;99(7):2285–90.

27. Hill DA, Wang SS, Cerhan JR, Davis S, Cozen W, Severson RK, et al. Risk of non-Hodgkin lymphoma (NHL) in relation to germline variation in DNA repair and related genes. Blood. 2006 Nov 1;108(9):3161–7.

28. Shen M, Purdue MP, Kricker A, Lan Q, Grulich AE, Vajdic CM, et al. Polymorphisms in DNA repair genes and risk of non-Hodgkin’s lymphoma in New South Wales, Australia. Haematologica. 2007 Sep 1;92(9):1180–5.

29. Skibola CF, Curry JD, Nieters A. Genetic susceptibility to lymphoma. Haematologica. 2007 Jul;92(7):960–9.

30. Rothman N, Skibola CF, Wang SS, Morgan G, Lan Q, Smith MT, et al. Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. Lancet Oncol. 2006 Jan;7(1):27–38.

31. Nieters A, Bracci PM, de Sanjosé S, Becker N, Maynadié M, Benavente Y, et al. A functional TNFRSF5 polymorphism and risk of non-Hodgkin lymphoma, a pooled analysis. Int J Cancer J Int Cancer. 2011 Mar 15;128(6):1481–5.

32. Skibola CF, Bracci PM, Halperin E, Conde L, Craig DW, Agana L, et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. Nat Genet. 2009 Aug;41(8):873–5.

33. Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet. 2010 Aug;42(8):661–4.

34. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998 Jan;88(1):15–9.

35. Akers NK, Curry JD, Conde L, Bracci PM, Smith MT, Skibola CF. Association of HLA-DQB1 alleles with risk of follicular lymphoma. Leuk Lymphoma. 2011 Jan;52(1):53–8.

36. Küppers R. Mechanisms of B-cell lymphoma pathogenesis. Nat Rev Cancer. 2005 Apr;5(4):251–62.

37. Thomas J Kindt, Richard A Godsby, Barbara A Osborne. Kuby Immunology. 6th ed. New York: W.H. Freeman and Company; 2007.

38. Klein U, Dalla-Favera R. Germinal centres: role in B-cell physiology and malignancy. Nat Rev Immunol. 2008 Jan;8(1):22–33.

39. Roulland S, Navarro J-M, Grenot P, Milili M, Agopian J, Montpellier B, et al. Follicular lymphoma-like B cells in healthy individuals: a novel intermediate step in early lymphomagenesis. J Exp Med. 2006 Oct 16;203(11):2425–31.

40. Vaandrager JW, Schuuring E, Kluin-Nelemans HC, Dyer MJ, Raap AK, Kluin PM. DNA fiber fluorescence in situ hybridization analysis of immunoglobulin class switching in B-cell neoplasia: aberrant CH gene rearrangements in follicle center-cell lymphoma. Blood. 1998 Oct 15;92(8):2871–8.

41. Bahler DW, Levy R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. Proc Natl Acad Sci U S A. 1992 Aug 1;89(15):6770–4.

42. Akasaka T, Akasaka H, Yonetani N, Ohno H, Yamabe H, Fukuhara S, et al. Refinement of theBCL2/immunoglobulin heavy chain fusion gene in t(14;18)(q32;q21) by polymerase chain reaction amplification for long targets. Genes Chromosomes Cancer. 1998 Jan;21(1):17–29.

43. Dighiero G, Hart S, Lim A, Borche L, Levy R, Miller RA. Autoantibody activity of immunoglobulins isolated from B-cell follicular lymphomas. Blood. 1991 Aug 1;78(3):581–5.

44. Umetsu DT, Esserman L, Donlon TA, DeKruyff RH, Levy R. Induction of proliferation of human follicular (B type) lymphoma cells by cognate interaction with CD4+ T cell clones. J Immunol Baltim Md 1950. 1990 Apr 1;144(7):2550–7.

45. Johnson PW, Watt SM, Betts DR, Davies D, Jordan S, Norton AJ, et al. Isolated follicular lymphoma cells are resistant to apoptosis and can be grown in vitro in the CD40/stromal cell system. Blood. 1993 Sep 15;82(6):1848–57.

46. Yang Z-Z, Novak AJ, Stenson MJ, Witzig TE, Ansell SM. Intratumoral CD4+CD25+ regulatory T-cell-mediated suppression of infiltrating CD4+ T cells in B-cell non-Hodgkin lymphoma. Blood. 2006 May 1;107(9):3639–46.

47. Hilchey SP, De A, Rimsza LM, Bankert RB, Bernstein SH. Follicular lymphoma intratumoral CD4+CD25+GITR+ regulatory T cells potently suppress CD3/CD28-costimulated autologous and allogeneic CD8+CD25- and CD4+CD25- T cells. J Immunol Baltim Md 1950. 2007 Apr 1;178(7):4051–61.

48. Ai WZ, Hou J-Z, Zeiser R, Czerwinski D, Negrin RS, Levy R. Follicular lymphoma B cells induce the conversion of conventional CD4+ T cells to T-regulatory cells. Int J Cancer J Int Cancer. 2009 Jan 1;124(1):239–44.

49. Hicklin DJ, Marincola FM, Ferrone S. HLA class I antigen downregulation in human cancers: T-cell immunotherapy revives an old story. Mol Med Today. 1999 Apr;5(4):178–86.

50. Steven G.E. Marsh, Peter Parham, Linda D. Barber. The HLA FactsBook. London, UK: Academic Press; 2000.

51. Khan N, Gowthaman U, Pahari S, Agrewala JN. Manipulation of Costimulatory Molecules by Intracellular Pathogens: Veni, Vidi, Vici!! Hobman TC, editor. PLoS Pathog. 2012 Jun 14;8(6):e1002676.

52. Kalergis AM, Boucheron N, Doucey MA, Palmieri E, Goyarts EC, Vegh Z, et al. Efficient T cell activation requires an optimal dwell-time of interaction between the TCR and the pMHC complex. Nat Immunol. 2001 Mar;2(3):229–34.

53. Choudhuri K, Wiseman D, Brown MH, Gould K, van der Merwe PA. T-cell receptor triggering is critically dependent on the dimensions of its peptide-MHC ligand. Nature. 2005 Jul 28;436(7050):578–82.

54. Krogsgaard M, Davis MM. How T cells “see” antigen. Nat Immunol. 2005 Mar;6(3):239–45.

55. Merwe PA van der, Davis SJ. MOLECULAR INTERACTIONS MEDIATING T CELL ANTIGEN RECOGNITION. Annu Rev Immunol. 2003 Apr;21(1):659–84.

56. Corr M, Slanetz A, Boyd L, Jelonek M, Khilko S, al-Ramadi B, et al. T cell receptor-MHC class I peptide interactions: affinity, kinetics, and specificity. Science. 1994 Aug 12;265(5174):946–9.

57. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, et al. Gene map of the extended human MHC. Nat Rev Genet. 2004 Dec;5(12):889–99.

58. Kulski JK, Gaudieri S, Bellgard M, Balmer L, Giles K, Inoko H, et al. The Evolution of MHC Diversity by Segmental Duplication and Transposition of Retroelements. J Mol Evol. 1997 Dec;45(6):599–609.

59. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001 Feb 15;409(6822):928–33.

60. Horton R, Niblett D, Milne S, Palmer S, Tubby B, Trowsdale J, et al. Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J Mol Biol. 1998 Sep;282(1):71–97.

61. Potts WK, Wakeland EK. Evolution of MHC genetic diversity: a tale of incest, pestilence and sexual preference. Trends Genet. 1993 Dec;9(12):408–12.

62. Potts WK, Manning CJ, Wakeland EK. Mating patterns in seminatural populations of mice influenced by MHC genotype. Nature. 1991 Aug 15;352(6336):619–21.

63. Wedekind C, Seebeck T, Bettens F, Paepke AJ. MHC-Dependent Mate Preferences in Humans. Proc R Soc B Biol Sci. 1995 Jun 22;260(1359):245–9.

64. Jacob S, McClintock MK, Zelano B, Ober C. Paternally inherited HLA alleles are associated with women’s choice of male odor. Nat Genet. 2002 Feb;30(2):175–9.

65. Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, et al. The Shaping of Modern Human Immune Systems by Multiregional Admixture with Archaic Humans. Science. 2011 Aug 25;334(6052):89–94.

66. Robinson J, Halliwell JA, McWilliam H, Lopez R, Parham P, Marsh SGE. The IMGT/HLA database. Nucleic Acids Res. 2013 Jan;41(Database issue):D1222–1227.

67. Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 2002 Jan;18(1):19–24.

68. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, et al. Linkage disequilibrium in the human genome. Nature. 2001 May 10;411(6834):199–204.

69. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol. 2007 Sep;68(9):779–88.

70. Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, et al. Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. J Immunol Baltim Md 1950. 1992 Jan 1;148(1):249–58.

71. Alcina A, Abad-Grau MDM, Fedetz M, Izquierdo G, Lucas M, Fernández O, et al. Multiple sclerosis risk variant HLA-DRB1*1501 associates with high expression of DRB1 gene in different human populations. PloS One. 2012;7(1):e29819.

72. Hor H, Kutalik Z, Dauvilliers Y, Valsesia A, Lammers GJ, Donjacour CEHM, et al. Genome-wide association study identifies new HLA class II haplotypes strongly protective against narcolepsy. Nat Genet. 2010 Sep;42(9):786–9.

73. Ettinger RA, Papadopoulos GK, Moustakas AK, Nepom GT, Kwok WW. Allelic variation in key peptide-binding pockets discriminates between closely related diabetes-protective and diabetes-susceptible HLA-DQB1*06 alleles. J Immunol Baltim Md 1950. 2006 Feb 1;176(3):1988–98.

74. Hill AV, Allsopp CE, Kwiatkowski D, Anstey NM, Twumasi P, Rowe PA, et al. Common west African HLA antigens are associated with protection from severe malaria. Nature. 1991 Aug 15;352(6336):595–600.

75. Carrington M, O'Brien SJ. The influence of HLA genotype on AIDS. Annu Rev Med. 2003;54:535–51.

76. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens. 2004 Dec;64(6):631–49.

77. Shan L, Molberg Ø, Parrot I, Hausch F, Filiz F, Gray GM, et al. Structural basis for gluten intolerance in celiac sprue. Science. 2002 Sep 27;297(5590):2275–9.

78. Godkin A, Davenport M, Hill AVS. Molecular analysis of HLA class II associations with hepatitis B virus clearance and vaccine nonresponsiveness. Hepatology. 2005 Jun;41(6):1383–90.

79. Taylor GM, Hussain A, Verhage V, Thompson PD, Fergusson WD, Watkins G, et al. Strong association of the HLA-DP6 supertype with childhood leukaemia is due to a single allele, DPB1*0601. Leukemia. 2009 May;23(5):863–9.

80. Apple RJ, Erlich HA, Klitz W, Manos MM, Becker TM, Wheeler CM. HLA DR–DQ associations with cervical carcinoma show papillomavirus–type specificity. Nat Genet. 1994 Feb;6(2):157–62.

Chapter 2

Multi-locus HLA Class I and II Allele and Haplotype Associations with Follicular Lymphoma[*]

C. F. Skibola[1], N. K. Akers[1], L. Conde[1], M. Ladner[2], S. K. Hawbecker[2], F. Cohen[2], F. Ribas[2], H. A. Erlich[2,3], D. Goodridge[4], E. A. Trachtenberg[2], M. T. Smith[1] & P. M. Bracci[5]

1 Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, USA
2 Center for Genetics, Children’s Hospital Oakland Research Institute, Oakland, CA, USA
3 Conexio Genomics, Perth, Australia
4 Roche Molecular Systems, Alameda, CA, USA
5 Department of Epidemiology and Biostatistics, School of Medicine, University of California San Francisco, San Francisco, CA, USA

[*] A similar version of this manuscript has been published: Skibola CF, Akers NK, Conde L, Ladner M, Hawbecker SK, Cohen F, et al. Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. Tissue Antigens. 2012 Apr;79(4):279–86.
This chapter is printed here with acknowledgement to all co-authors and Wiley Online Library.

Abstract

Follicular lymphoma (FL) is an indolent, sometimes fatal disease characterized by recurrence at progressively shorter intervals and is frequently refractive to therapy. Genome-wide association studies have identified single nucleotide polymorphisms (SNPs) in the human leukocyte antigen (HLA) region on chromosome 6p21.32-33 that are statistically significantly associated with FL risk. Low to medium resolution typing of single or multiple HLA genes has provided an incomplete picture of the total genetic risk imparted by this highly variable region. To gain further insight into the role of HLA alleles in lymphomagenesis and to investigate the independence of validated SNPs and HLA alleles with FL risk, high-resolution HLA typing was conducted using next-generation sequencing in 222 non-Hispanic White FL cases and 220 matched controls from a larger San Francisco Bay Area population-based case–control study of lymphoma. A novel protective association was found between the *DPB1**03:01 allele and FL risk [odds ratio (OR) = 0.39, 95% confidence interval (CI) = 0.21–0.68]. Extended haplotypes *DRB1**01:01-*DQA1**01:01-*DQB1**05:01 (OR = 2.01, 95% CI = 1.22–3.38) and *DRB1**15-*DQA1**01-*DQB1**06 (OR = 0.55, 95% CI = 0.36–0.82) also influenced FL risk. Moreover, *DRB1**15-*DQA1**01-*DQB1**06 was highly correlated with an established FL risk locus, rs2647012. These results provide further insight into the critical roles of HLA alleles and SNPs in FL pathogenesis that involve multi-locus effects across the HLA region.

Introduction

Follicular lymphoma (FL) is an indolent B-cell malignancy characterized by a highly variable clinical course and multiple relapses (1). Approximately one-third of FL cases transform to a more aggressive histology, usually diffuse large B-cell lymphoma (DLBCL), which is associated with a poor clinical outcome (2, 3). The molecular basis of FL has been fairly well characterized (4–6), although its root causes remain less clear. In recent genome-wide association studies (GWAS) of non-Hodgkin lymphoma (NHL) and validation within the large InterLymph consortium, we identified three independent susceptibility loci for FL on chromosome 6p21.3 in the human leukocyte antigen (HLA) class I and II regions (7–9). Located in the HLA class 1 region at 6p21.33 near psoriasis susceptibility region 1, rs6457327 was inversely associated with risk of FL (P-value = 4.7 × 10-11) (8). In the HLA class II region at 6p21.32, two single nucleotide polymorphisms (SNPs), rs10484561 and rs7755224, were associated with twofold increased risks of FL (P-values = 1.12 × 10-29 and 2.0 × 10-19, respectively) (7). rs10484561 and rs7755224 are in total linkage disequilibrium (LD) and are located 29 and 16 kb centromeric of *HLA-DQB1*, respectively. On the basis of a tag SNP analysis, we inferred that rs10484561 may be part of a high-risk extended haplotype, *DRB1**01:01-*DQA1**01:01-*DQB1**05:01 (7). Another class II locus in the *HLA-DQB1* region, rs2647012, was inversely associated with FL risk after adjusting for rs10484561 [Odds ratio (OR) = 0.70, P-value = 4 × 10-12] (9). In subsequent studies, we confirmed a positive association between FL risk and the *DQB1**05 allele group (P-value = 0.013) and identified the *DQB1**06 allele group as protective for FL (P-value = 4.5 × 10-5) (10). An independent study further supported *DRB1**01:01 as a risk locus for FL (11). Taken together, these studies suggest that genetic variation in the HLA region plays an important role in the etiology of FL.

HLA class I- and class II-restricted CD8+ and CD4+ T-cell responses are essential for the immune system to mount a successful antitumor immune defense or to remove infected cells. A defect in these important processes could allow pathogenic cells to escape host immune recognition that may increase the likelihood of lymphomagenesis. To further pinpoint risk-associated HLA alleles and haplotypes in the pathogenesis of FL, we investigated whether previously validated FL-associated GWAS SNPs (rs6457327, rs10484561, and rs2647012) and HLA alleles were independent risk factors for FL. To this end, we extended the analysis of our NHL case–control study to determine HLA class I (*HLA-A*, *-B*, and *-C*) and class II (*DRB1*/3/4/5, *DQA1*, *DQB1*, and *DPB1*) alleles using high-resolution HLA typing by next-generation Roche GS FLX 454 (Pleasonton, CA) sequencing in 222 non-Hispanic White FL cases and 220 controls frequency-matched by sex and age in 5-year groups.

Materials and methods

Study population

Samples sequenced at HLA included non-Hispanic White FL cases (n = 222) and frequency-matched controls (n = 200) who were part of a population-based case–control study of NHL conducted in the San Francisco Bay Area that included 2055 patients newly diagnosed with NHL from 2001 to 2006 frequency-matched to 2081 control participants. The majority of these FL cases were included in the previously described GWAS (92.8%) (7) and *DQB1* typing study (95.9%) (10). Eligible patients were identified by the Cancer Prevention Institute of California's rapid case ascertainment and by SEER abstract, were 20 to 85 years old at diagnosis, alive at first contact, residents of one of six Bay Area counties, able to complete an interview in English, and had no prior history of hematopoietic cancer or physician indicated contraindications to contact. Eligible controls were identified by random digit dial, and by random sampling of the Centers for Medicare and Medicaid Services lists for individuals aged 65 years or older and were frequencymatched to cases by 5-year age group, sex, and county of residence. Blood and/or buccal cells were collected from 85% of eligible study participants. Eligible NHL patients also provided consent (98%) to access their diagnostic materials to confirm diagnosis of NHL and for consistent classification of NHL subtype by the study pathologist using the WHO classification.

HLA genotyping by next-generation sequencing

High-resolution sequencing to obtain HLA genotypes (as in the IMGT/HLA database v3.6.0, http://www.ebi.ac.uk/imgt/ hla/) was carried out as previously described in detail (12, 13). Briefly, next-generation clonal sequencing of exonic amplicons was performed using the Roche 454 GS FLX massively parallel pyrosequencing system (14). Roche-developed polymerase chain reaction (PCR) primers to exons 2–4 for HLA class I (A, B, and C), exons 2–3 for class II *DQB1*, exon 2 for *DRB1*/3/4/5, *DQA1*, and *DPB1*; 11 multiplex identification tags were used in the 10-ng sample template amplifications. Primary HLA amplicons were purified to remove short artifacts, and then pooled in equimolar concentrations for emulsion PCR, bead recovery, and pyrosequencing. Sequence data analysis was accomplished using the ATF software (Conexio Genomics, Perth, Australia). In almost all HLA analyses to date, it has been cost-prohibitive to analyze all genomic regions for each gene to determine the unambiguous genotype of each sample, and until most of the genomic region of the genes is sequenced, there will always be a level of ambiguity due to the high degree of polymorphism of HLA genes. The Roche GS FLX 454 clonal sequencing of HLA described here consequently results in some residual ambiguity which, although limited compared to other sequencing methods, must still be reduced for analysis. To do this, the alleles analyzed here were called based on the most common 'lowest number' alleles from a list of possible genotypes derived by clonal sequence analysis of particular exons. The allelic genotype calls and the related total possible six digit alleles from resolved genotypes and unresolved ambiguity are listed for each locus in Tables S1–8, Supporting Information. The nature of the clonal sequencing dramatically reduces the level of possible ambiguity using traditional Sanger sequencing, and this is the first time that a complete ambiguity table has been reported for HLA genotypes in an association study. Genotypes derived from a total of 89 samples (27 for *HLA-A*, 39 for *HLA-B*, 18 for *HLA-C*, and 5 for *HLA-DRB1*) which failed at least one exon analysis by the clonal sequencing method were resequenced and retested using Luminex LABType SSO kits (One Lambda Inc., Canoga Park, CA). This method uses sequence-specific oligonucleotide probes bound to fluorescently coded microspheres to identify the HLA alleles in an amplified DNA sample, and alleles were identified using HLA FUSION software, v2.0.0 (One Lambda Inc.). Data from the LABType high-resolution bead kits were used in addition to the sequencing data to fill in exon gaps to resolve the genotypes at a level comparable to the 454 genotypes. The HLA nomenclature used for the current data in this article reflects the newest iteration of rules (2010) (http://hla.alleles.org/announcement.html) to describe HLA alleles, while all older data are presented as they were originally noted in the earlier publications.

Data Analysis

Haplotype frequencies for cases and controls were estimated using the iterative expectation-maximization algorithm implemented in the PYPOP software (15). LD between HLA alleles and rs6457327, rs10484561, and rs2647012 was measured in our control population using PYPOP that calculates Dand chi-squared values based on observed and expected frequencies of haplotypes. Deviations from Hardy–Weinberg equilibrium (HWE) in controls were tested with the ARLEQUIN software v3.5.1.2 (16) using a Markov chain method with exact P-value estimation (17). No significant departure from HWE was observed for any loci at a P < 0.001 level. For each individual allele or haplotype, the independence of the number of observed and unobserved counts in cases and controls was determined using the 'fisher.test' function from the 'STAT' package in R (http://stat.ethz.ch/R-manual/Rpatched/ library/stats/html/00Index.html). ORs and 95% confidence intervals (CIs) were estimated as further measures of the magnitude of the association between alleles or haplotypes and disease status. The 'p.adjust' function from the same package in R was used to adjust the P-values for the number of independent statistical tests at each locus using the Bonferroni correction.

Unconditional backward stepwise logistic regression methods in STATA version 11 (StataCorp, College Station, TX) were used to assess independence of individual risk loci. All established or suspected risk factors in the classic HLA regions (rs6457327, rs2647012, rs10484561-*DRB1**01:01-*DQA1**01:01-*DQB1**05:01, *DRB1**15-*DQA1**01-*DQB1**06, *DRB1**13-*DQA1**01-*DQB1**06, and *DPB1**03:01) were included and the final best fitting model was determined based on a likelihood ratio test. A P-value threshold of 0.10 was the criteria used for remaining in the model. For these analyses, one allele of each haplotype was used as a proxy for the haplotype as a whole (*DQB1**05:01, *DRB1**15, and *DRB1**13). Due to collinearity, *DRB1**15 and rs2647012 were assessed as a single variable where 0 indicated presence of neither allele, 1 indicated presence of rs2647012 alone, and 2 indicated the presence of both rs2647012 and *DRB1**15. All other alleles of interest were coded as present vs absent.

Results

The association results for all HLA class I and II alleles with P-values <0.05 are shown in Table 1. We identified a novel protective allele, *DPB1**03:01, associated with risk of FL (OR = 0.39, 95% CI = 0.21–0.68, adjusted P-value = 8.30 × 10-3; Table 2) that was not in significant LD with any HLA alleles previously shown to be associated with FL (D'= 0.30 with rs10484561, D' = 0.04 with rs6457327, and D' = 0.03 with rs2647012; Table 3). *DPB1* is located centromeric to *DRB1*, *DQA1*, and *DQB1* and is separated from these genes by a recombination hotspot (18). Using backward stepwise logistic regression methods to analyze HLA alleles and previously identified SNPs of interest, the final best fitting model showed that *DPB1**03:01 was independently associated with FL (Table 4).

Allele	Case alleles (freq)	Control alleles (freq)	[*]OR (95% CI)	*p*-value	[†]Bonf. *p*
*C*04:01*	70 (0.159)	44 (0.102)	1.67 (1.1-2.56)	1.56E-02	2.81E-001
*C*06:02*	48 (0.109)	24 (0.056)	2.08 (1.22-3.62)	4.50E-03	8.10E-002
*C*07:02*	38 (0.086)	61 (0.141)	0.58 (0.36-0.9)	1.38E-02	2.48E-001
*C*16:01*	7 (0.016)	18 (0.042)	0.37 (0.13-0.95)	2.56E-02	4.60E-001
*B*07:02*	34 (0.077)	58 (0.133)	0.54 (0.34-0.87)	7.95E-03	2.86E-001
*B*35:01*	43 (0.097)	26 (0.06)	1.7 (1-2.94)	4.45E-02	1.00E+000
*B*50:01*	8 (0.018)	1 (0.002)	8 (1.07-356.03)	3.81E-02	1.00E+000
*DRB1*01:01*	54 (0.122)	31 (0.07)	1.84 (1.13-3.02)	1.17E-02	3.16E-01
*DRB1*01:02*	17 (0.038)	5 (0.011)	3.48 (1.22-12.15)	1.56E-02	4.21E-01
*DRB1*15:01*	41 (0.093)	71 (0.161)	0.53 (0.34-0.81)	2.37E-03	6.40E-002
*DRB5*01:01*	42 (0.095)	71 (0.161)	0.55 (0.35-0.83)	3.43E-03	3.09E-02
*DRB3/4/5*ABSENT*	99 (0.224)	60 (0.136)	1.83 (1.27-2.65)	8.41E-04	7.57E-03
*DQA1*01:01*	97 (0.22)	55 (0.125)	1.98 (1.36-2.9)	2.39E-04	1.67E-03
*DQA1*01:02*	64 (0.145)	101 (0.23)	0.57 (0.4-0.82)	1.82E-03	1.27E-02
*DQB1*03:03*	29 (0.066)	15 (0.034)	1.98 (1.01-4.03)	4.33E-02	7.37E-001
*DQB1*05:01*	89 (0.201)	43 (0.098)	2.31 (1.54-3.51)	1.90E-05	3.23E-04
*DQB1*06:02*	41 (0.093)	69 (0.158)	0.55 (0.35-0.84)	4.21E-03	7.15E-002
*DPB1*03:01*	19 (0.043)	46 (0.105)	0.39 (0.21-0.68)	4.61E-04	8.30E-03
*DPB1*17:01*	2 (0.005)	9 (0.02)	0.22 (0.02-1.06)	3.71E-02	6.68E-001
*DPB1*20:01*	1 (0.002)	8 (0.018)	0.12 (0-0.92)	2.07E-02	3.72E-001

Table 1. HLA class I and II allele counts, odds ratios and 95% confidence intervals in follicular lymphoma cases and controls (2n =444 cases, 440 controls) from a population-based case-control study of non-Hodgkin lymphoma in the San Francisco Bay Area. Only alleles with p-values < 0.05 are shown. [*]Odds ratios (OR), 95% confidence interval (95% CI), and *p*-values were obtained using the 'fisher.test' function form he 'stat' package in R. [†]*p*-values were adjusted for the number of alleles tested using a Bonferroni correction.

Allele	**Case count (freq.)**	**Control Count (freq.)**	**[*]OR (95% CI)**	**[*]P-value**	**[†]Bonf. *p***
*DPB1*01:01*	22 (0.050)	25 (0.057)	0.87 (0.46-1.64)	6.56E-01	1.00E+000
*DPB1*02:01*	79 (0.179)	66 (0.150)	1.23 (0.85-1.79)	2.76E-01	1.00E+00
DPB1*03:01	**19 (0.043)**	**46 (0.105)**	**0.39 (0.21-0.68)**	**4.61E-04**	**8.30E-03**
*DPB1*04:01*	193 (0.437)	181 (0.411)	1.11 (0.84-1.46)	4.54E-01	1.00E+00
*DPB1*04:02*	51 (0.115)	46 (0.105)	1.12 (0.72-1.75)	6.67E-01	1.00E+00
*DPB1*05:01*	7 (0.016)	6 (0.014)	1.16 (0.33-4.23)	1.00E+00	1.00E+00
*DPB1*06:01*	10 (0.023)	3 (0.007)	3.37 (0.86-19.2)	8.99E-02	1.00E+00
*DPB1*09:01*	1 (0.002)	1 (0.002)	1 (0.01-78.27)	1.00E+00	1.00E+00
*DPB1*10:01*	9 (0.020)	7 (0.016)	1.29 (0.42-4.1)	8.02E-01	1.00E+00
*DPB1*11:01*	11 (0.025)	8 (0.018)	1.38 (0.5-3.99)	6.44E-01	1.00E+00
*DPB1*13:01*	17 (0.038)	7 (0.016)	2.47 (0.96-7.13)	6.04E-02	1.00E+00
*DPB1*14:01*	7 (0.016)	8 (0.018)	0.87 (0.27-2.77)	8.02E-01	1.00E+00
*DPB1*15:01*	3 (0.007)	4 (0.009)	0.75 (0.11-4.43)	7.25E-01	1.00E+00
*DPB1*16:01*	3 (0.007)	4 (0.009)	0.75 (0.11-4.43)	7.25E-01	1.00E+00
*DPB1*17:01*	2 (0.005)	9 (0.020)	0.22 (0.02-1.06)	3.71E-02	6.68E-01
*DPB1*19:01*	2 (0.005)	4 (0.009)	0.5 (0.05-3.48)	4.51E-01	1.00E+00
*DPB1*20:01*	1 (0.002)	8 (0.018)	0.12 (0-0.92)	2.07E-02	3.72E-01
*DPB1*23:01*	2 (0.005)	6 (0.014)	0.33 (0.03-1.85)	1.77E-01	1.00E+00

Table 2. *HLA-DPB1* allele counts, odds ratios and 95% confidence intervals in follicular lymphoma cases and controls (2n =444 cases, 440 controls) from a population-based case-control study of non-Hodgkin lymphoma in the San Francisco Bay Area. Only alleles with p-values < 0.05 are shown. [*]Odds ratios (OR), 95% confidence interval (95% CI), and *p*-values were obtained using the 'fisher.test' function form he 'stat' package in R. [†]*p*-values were adjusted for the number of alleles tested using a Bonferroni correction

As a follow-up to further explore independence between the GWAS SNP, rs10484561, and *DRB1*, *DQA1*, and *DQB1* alleles, we confirmed our previous tag SNP analysis (7) implicating the extended haplotype, *DRB1**01:01-*DQA1**01:01-*DQB1**05:01, as a risk factor for FL. Each allele of the haplotype was in strong LD with rs10484561 (D' = 0.93, 1.0, and 1.0, respectively; Table 3) and was associated with increased risk of FL (Table 5).

Allele 1	Allele 2	Haplotype Freq.	Observed	Expected	D'	ChiSq
rs10484561 ‘G’	_DQB1*05:01_	0.111	18.0	2.1	1.00	152
	DQA1*01:01	0.111	18.0	2.7	1.00	116
	DRB1*01:01	0.086	14.0	1.7	0.93	113
	DRB3/4/5*ABSENT	0.111	18.0	2.8	1.00	111
rs6457327 ‘A’	_C*07:02_	0.134	22.0	7.9	0.93	45
	_B*07:02_	0.122	20.0	6.8	1.00	44
	_C*03:04_	0.049	8.0	2.7	1.00	16
	_B*44:03_	0.049	8.0	2.7	1.00	16
	_C*02:02_	0.043	7.0	2.4	1.00	14
	_C*16:01_	0.043	7.0	2.4	1.00	14
rs2647012 ‘A’	_DQA1*01:02_	0.191	31.3	15.1	0.86	39
	DQB1*06:02	0.159	26	11.6	1.00	39
	DRB1*15:01	0.159	26	11.6	1.00	39
	DRB5*01:01	0.159	26	11.6	1.00	39
	DRB1*03:01	0.152	25	11.1	1.00	37
	DQB1*02:01	0.159	26	12.0	0.96	35
DPB1*03:01 (46)	rs10484561*T	0.108	17.5	16.9	0.30	0.23
	rs2647012*G	0.069	11.4	11.1	0.03	0.02
	rs6457327*C	0.082	13.5	13.2	0.04	0.02
DQB1*06	_DRB1*15_	0.174	76.0	22.0	0.95	224
DQB1*06:02	_DRB1*15:01_	0.158	69.0	11.2	1.00	423
DQB1*06:01	_DQB1*15:02_	0.018	8	0.1	1.00	438
DQB1*06	_DRB1*13_	0.098	43	13.9	0.8	95
DQB1*06:04	_DRB1*13:02_	0.032	14.0	0.7	0.93	281
DQB1*06:09	_DRB1*13:02_	0.014	6	0.3	1.0	127
DQB1*06:14	_DRB1*13:01_	0.005	2	0.1	1.0	38
DQB1*06:03	_DRB1*13:05_	0.002	1	0.04	1.0	21

Table 3. Linkage equilibrium (LD) values in the control population for selected allele combinations spanning the HLA region. LD valus were estimated with the Pypop software (http://www.pypop.org/).

	Stepwise logistic regression				
Allele*	Odds Ratio	Std. Error	95% CI	z-score	*p*-value
*DPB1*03:01*	0.29	0.11	0.14 - 0.6	-3.32	0.001
*DQB1*05:01*	1.75	0.56	0.94 - 3.29	1.75	0.08
rs6457327	0.66	0.14	0.44 - 1.01	-1.93	0.054
rs2647012 + *DRB1*15*	0.53	0.10	0.37 - 0.75	-3.54	<0.001

Table 4. Best fitting model from multivariable backward stepwise logistic regression of known HLA risk loci* associated with follicular lymphoma risk. *HLA alleles and SNPs coded as present/absent (a dominant effect model) included rs6457327, rs2647012, *DQB1**05:01 (as a proxy for the rs10484561-*DRB1**01:01-*DQA1**01:01-*DQB1**05:01 haplotype), *DRB1*15* (as a proxy for *DRB1**15-*DQA1**01-*DQB1**06), *DRB1*13* (as a proxi for *DRB1**13 - *DQA1**01 - *DQB1**06), and *DPB1**03:01. The rs2647012 and *DRB1**15 allele effects were assessed as 0 = carrier of neither allele, 1 = carrier of rs2647012, and 2 = carrier of both rs2647012 and *DRB1**15. Likelihood ratio test *p*-value cutoff of 0.10 was used as criteria for significant contribution to the model.

Haplotype and Individual Alleles	Case count	Control Count	*OR (95% CI)	*P-value	*Bonf. *p*
*DRB1*01 -DQA1*01-DQB1*05*	78 (0.178)	39 (0.089)	2.22 (1.45-3.43)	1.43E-04	2.23E-03
*DRB1*01:01*	54 (0.122)	31 (0.070)	1.84 (1.13-3.02)	1.17E-02	3.16E-001
*DRB1*01:02*	17 (0.038)	5 (0.011)	3.48 (1.22-12.2)	1.56E-02	4.21E-01
*DRB1*01:03*	10 (0.023)	5 (0.011)	1.99 (0.62-7.50)	2.98E-01	1.00E+00
*DQA1*01:01*	97 (0.22)	55 (0.125)	1.98 (1.36-2.90)	2.39E-04	1.67E-03
*DQB1*05:01*	89 (0.201)	43 (0.098)	2.31 (1.54-3.51)	1.90E-05	3.23E-04

Table 5. *HLA DRB1*01- DQA1*01-DQB1*05* haplotype and allele counts, odds ratios and 95% confidence intervals in follicular lymphoma cases and controls (2n=444 cases, 440 controls) from a population-based case-control study of non-Hodgkin lymphoma in the San Francisco Bay Area. *Odds ratios (OR), 95% confidence intervals (CI); p-values are based on a Bonferroni (Bonf) correction for the number of alleles tested at each locus (16 *DRB1-DQA1-DQB1* haplotypes [2 digit], 27 *DRB1* alleles [4 digit], 7 *DQA1* alleles [4 digit], and 17 *DQB1* alleles [4 digit]).

*DQB1**06 and *DRB1**13 have been reported as protective alleles for FL (10, 11). Because *DQB1**06 is known to exist in haplotypes with both *DRB1**13 and *DRB1**15 in Caucasian populations, it was unclear which haplotypes may be responsible for these associations. Here, we found that the *DQB1**06 and *DRB1**15 alleles were significantly associated with decreased FL risk (Table 6), and that although non-significant after correction, the frequency of the haplotypes *DRB1**15-*DQA1**01-*DQB1**06 and *DRB1**13-*DQA1**01-*DQB1**06 were similarly decreased in cases (Table 6). Logistic regression analysis showed that *DRB1**13-*DQA1**01-*DQB1**06 was no longer associated with FL risk after adjustment for other FL-associated HLA alleles (OR = 0.92, P = 0.83; Table 4). We also found that all carriers of the protective *DRB1**15:01-*DQA1**01:02-*DQB1**06:02 haplotype were carriers of the rs2647012 A allele, although the minor allele frequency of rs2647012 (0.40) was higher than the frequency of the linked *DRB1**15:01-*DQA1**01:02-*DQB1**06:02 haplotype (0.16). LD between rs2647012 and the individual alleles of the haplotype corroborated the high LD between rs2647012 and

DRB1*15:01-DQA1*01:02-DQB1*06:02 (D' = 1, 0.86, and 1, respectively; Table 3). Limiting the dataset to those individuals without DRB1*15:01-DQA1*01:02-DQB1*06:02 showed a modest effect of rs2647012 on FL risk (OR = 0.70, 95% CI = 0.45–1.1, P = 0.10).

Haplotype and Individual Alleles	Case Count (freq.)	Control Count (freq.)	*OR (95%CI)	*P-value	†Bonf. *p*
DRB1*15 - DQA1*01 - DQB1*06	45.0 (0.103)	76.0 (0.174)	0.55 (0.36-0.82)	3.18E-03	5.09E-02
DRB1*13 - DQA1*01 - DQB1*06	24.0 (0.055)	43.0 (0.098)	0.53 (0.30-0.92)	2.15E-02	3.44E-01
DRB1*15	45 (0.102)	79 (0.180)	0.52 (0.34-0.78)	9.60E-04	1.25E-002
DRB1*13	31 (0.07)	50 (0.114)	0.59 (0.36-0.96)	2.69E-02	3.50E-001
DQA1*01:02	64 (0.145)	101 (0.23)	0.57 (0.4-0.82)	1.82E-03	1.27E-02
DQA1*01:03	19 (0.043)	32 (0.073)	0.58 (0.3-1.07)	8.25E-02	5.78E-001
DQB1*06	73 (0.165)	122 (0.279)	0.51 (0.36-0.72)	6.52E-05	3.26E-04

Table 6. HLA-DRB1*15 and DRB1*13 haplotypes and constituent alleles and risk of follicular lymphoma (2n=444 cases, 440 controls). *Odd ratios (OR), 95% confidence intervals (CI) and p-values were obtained using the 'fisher.test' function from the 'stat' package in R. †*P*-values were adjusted for the number of alleles tested at each locus using a Bonferroni correction (16 *DRB1-DQA1-DQB1* haplotypes [2 digit], 13 *DRB1* alleles [2 digit], 7 *DQA1* alleles [4 digit], and 5 *DQB1* alleles [2 digit]).

For HLA class I loci, no significant associations with FL risk were found (Table 1). However, we found that the C*07:02 and B*07:02 alleles were linked to rs6457327 'A' carriers (D' = 0.93 and 1.0, respectively; Table 3). Restricting the dataset to those individuals without C*07:02 or B*07:02 made little change on the estimated risk statistic for rs6457327 (OR = 0.55, 95% CI = 0.30–1.00, P = 0.05).

Discussion

Previous GWAS and low to medium resolution HLA typing studies have identified major FL-susceptibility loci in the HLA class I and II regions. As a follow-up, we conducted next-generation, high-throughput HLA sequencing of class I (*HLA-A*, -B, and -C) and class II (*DRB1*/3/4/5, *DQA1*, *DQB1*, and *DPB1*) alleles to determine the independent role of HLA alleles and SNPs as susceptibility factors for FL. This study provides the first examination of *DPB1* alleles in FL cases, as well as the highest resolution and most complete characterization of HLA class I and II alleles to date. Here, we found that DPB1*03:01, DQB1*05:01, rs6457327, and DRB1*15 all independently influence FL risk. Specifically, we identified a novel, inverse association between the DPB1*03:01 allele and risk of FL that was independent of other HLA class II alleles based on LD and logistic regression analyses (Table 4). The low LD between *DPB1* and other class II loci is likely a result of the high level of recombination in the region (18). Interestingly, previous studies found that the DPB1*03:01 allele was positively associated with risk of nodular sclerosing HL (NSHL) (19, 20). Opposite effects with the same HLA alleles on the risk of FL and NSHL were also observed for the DRB1*15:01-DQA1*01:02-DQB1*06:02 haplotype (high risk for NSHL and low risk for FL) (21). These findings suggest that HLA class II alleles may modulate risk for NSHL and FL in a divergent manner.

Although non-significant, an inverse association with FL risk was found for *DPB1**20:01, an allele closely related to *DPB1**03:01, and positive associations were found with *DPB1**06:01 and *DPB1**13:01 (Table 2). Examining these alleles at the amino acid level shows that the *DPB1**03:01 and *DPB1**20:01 alleles that are overrepresented in controls possess a glutamic acid rather than a lysine residue at position 69. These amino acids are oppositely charged, and reside in binding pocket 4, suggesting this change may impact *DPB1* binding. Serological groupings may also be relevant at this locus (22). Characterizing each allele by *DPB1* serological group showed the DP3 group, containing the 56E and 85-87EAV sequences, represents only 11.8% of case alleles compared to 19.6% of control alleles. If validated, this may indicate a role for anti-DP serological activity in the etiology of FL.

This study also confirmed our previous report based on a tag SNP analysis (7) that the *DRB1**01:01-*DQA1**01:01-*DQB1**05:01 haplotype was associated with a twofold increased risk of FL, with *DQB1**05:01 being the most significantly associated allele in the risk haplotype. There is some indication that the risk haplotype includes *DRB1**01:02 and *01:03 (Table 5), although this finding will require replication in independent studies.

We further investigated the inverse associations between the *DQB1**06 and *DRB1**13 alleles and FL risk. As previously described in Caucasians (23), we found that *DRB1**13 was in strong LD with *DQB1**06:03, *06:04, and *06:09 alleles, whereas *DQB1**06:02 (the most common *DQB1**06 allele) was in high LD with *HLA-DRB1**15 (Table 3). Thus, we observed a decreased risk of FL with all alleles and haplotypes containing *DRB1**13 or *15 and *DQB1**06, with *DQA1**01:02 or *01:03 (Table 6). Due to the extensive LD across *DRB1*, *DQA1*, and *DQB1*, it is unclear which loci drive these haplotype–disease associations. However, *DRB1**13 did not affect FL risk after adjustment for other HLA alleles in logistic regression analyses suggesting that this association may be the result of confounding by other HLA alleles.

We also showed that carriers of the *DRB1**15:01-*DQA1**01:02-*DQB1**06:02 haplotype harbor the rs2647012 variant, which was previously reported as a protective allele for FL (9). This haplotype may be a causal variant driving the observed rs2647012 association with FL. Because there remained a modest reduction in FL risk for rs2647012 after adjusting for *DRB1**15:01-*DQA1**01:02-*DQB1**06:02, larger studies will be needed to determine the independent role of rs2647012 and the haplotype in disease risk. We further investigated LD between the HLA class I GWAS SNP, rs6457327 (8), and HLA class I alleles. Here, we found that the C*07:02 and B*07:02 alleles were in LD with the protective rs6457327 A allele. However, individuals with rs6457327 A had approximately the same risk regardless of C*07:02 and B*07:02 status, suggesting the role for a yet unidentified causal locus that is in LD with rs6457327. HLA class II alleles may influence FL risk through several modes of action including effects on of T-cell activation, antigen presentation of infectious or tumor-associated peptides, and HLA protein/gene expression. FL and Burkitt lymphoma disrupt normal HLA class II-mediated antigen presentation by B-cells and dendritic cells to CD4+ T-cells as a mechanism to hinder their recognition by the immune system (24). Underexpression of HLA class II on HL Reed–Sternberg cells is an independent adverse prognostic factor in classical HL (25), and loss of HLA class II expression on DLBCL tumor cells has been associated with poor survival (26). Further studies will be needed to clarify the functional role of HLA alleles in lymphomagenesis, which will likely expand our knowledge of the deregulated cellular processes that drive FL and its progression.

Use of cancer registry rapid case ascertainment and SEER abstracts to identify newly diagnosed NHL patients helped to diminish selection and participation bias in our study population, although patients with aggressive disease and poor prognosis are likely underrepresented. However, as FL in general is a more indolent lymphoma, effects of survival bias on case participation should not have affected these analyses. Further, bias effects were diminished by the high participation rate for biospecimen collection in participants (~87%). The small number of non-White participants precluded analyses by race and ethnicity. Despite evidence of internal consistency in the magnitude and direction of many of our results, we had low power to test associations for low frequency variants and results from analyses with few 'exposed' should be interpreted conservatively and require validation in further studies. In conclusion, these studies provide additional evidence that HLA alleles play essential roles in the pathogenesis of FL. As our findings show, this involves complex, multi-locus effects that span the HLA region. Because of the extensive and complex LD patterns within this region, studies in FL case–control populations from non-Caucasian ancestral pedigrees are underway that may help to distinguish between primary (causal) and secondary HLA signals. Because the causative alleles could be in non-coding (nc) regions that effect gene expression, studies are currently underway to test differential allelic gene expression of ncSNPs in high LD with HLA susceptibility alleles. Moreover, the contribution of HLA alleles in the pathogenesis of FL and other subtypes of NHL is a major focus of future studies within InterLymph where the HLA alleles identified here and in other independent case–control studies of NHL will be tested for further validation. Thus, we anticipate that substantial progress will be made in the near future that will help to elucidate the genetic basis of NHL. Such data will likely highlight pathways and components that may be amenable to therapeutic modulation.

Acknowledgments

This work was supported by National Institutes of Health grants: CA122663, CA154643-01A1, and CA104682 (CFS), and grants CA45614 and CA89745 from the National Cancer Institute, National Institutes of Health (PMB). CFS, MTS, and PMB designed the study. CFS drafted the manuscript with significant contributions from NKA and EAT. ML, SKH, FC, FR, DG, HE, and EAT performed high-resolution HLA sequencing analysis. NKA and LC analyzed the data. All authors reviewed and approved the final manuscript.

Conflict of interest

The authors have declared no conflicting interests.

Supplemental Tables and Figures

Supplemental Tables S1-S8 (145 pages) may be viewed online at : http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3293942/bin/NIHMS350823-supplement-Supp_Table_S1-_S8.doc

These tables give detailed and specific information about the ambiguity of the allele calls presented in this work.

References

1. Johnson PW, Rohatiner AZ, Whelan JS et al. Patterns of survival in patients with recurrent follicular lymphoma: a 20-year study from a single center. *J Clin Oncol* 1995: 13:140–7.
2. Bastion Y, Sebban C, Berger F et al. Incidence, predictive factors, and outcome of lymphoma transformation in follicular lymphoma patients. *J Clin Oncol* 1997: 15: 1587–94.
3. Horning SJ, Rosenberg SA. The natural history of initially untreated low-grade non-Hodgkin's lymphomas. *N Engl J Med* 1984: 311: 1471–5.
4. Fitzgibbon J, Iqbal S, Davies A et al. Genome-wide detection of recurring sites of uniparental disomy in follicular and transformed follicular lymphoma. *Leukemia* 2007: 21:1514–20.
5. Leich E, Salaverria I, Bea S et al. Follicular lymphomas with and without translocation t(14;18) differ in gene expression profiles and genetic alterations. *Blood* 2009: 114: 826–34.
6. O'Riain C, O'Shea DM, Yang Y et al. Array-based DNA methylation profiling in follicular lymphoma. *Leukemia* 2009:23: 1858–66.
7. Conde L, Halperin E, Akers NK et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* 2010: 42: 661–4.
8. Skibola CF, Bracci PM, Halperin E et al. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat Genet* 2009: 41: 873–5.
9. Smedby KE, Foo JN, Skibola CF et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS Genet* 2011: 7: e1001378.
10. Akers NK, Curry JD, Conde L, Bracci PM, Smith MT, Skibola CF. Association of *HLA-DQB1* alleles with risk of follicular lymphoma. *Leuk Lymphoma* 2011: 52:53–8.
11. Wang SS, Abdou AM, Morton LM et al. Human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology. *Blood* 2010: 115: 4820–3.
12. Bentley G, Higuchi R, Hoglund B et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens* 2009: 74: 393–403.
13. Holcomb CL, H¨oglund B, Anderson MW et al. A multi-site study employing high resolution HLA genotyping by next generation sequencing. *Tissue Antigens* 2011: 77: 206–17.
14. Margulies M, Egholm M, Altman WE et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005: 437: 376–80.
15. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update–a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* 2007:69 (Suppl 1): 192–7.
16. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 2005: 1: 47–50.
17. Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 1992: 48: 361–72.
18. Begovich AB, McClure GR, Suraj VC et al. Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. *J Immunol* 1992: 148: 249–58.
19. Oza AM, Tonks S, Lim J, Fleetwood MA, Lister TA, Bodmer JG. A clinical and epidemiological study of human leukocyte antigen-DPB alleles in Hodgkin's disease. *Cancer Res* 1994: 54: 5101–5.
20. Taylor GM, Gokhale DA, Crowther D et al. Further investigation of the role of *HLA-DPB1* in adult Hodgkin's disease (HD) suggests an influence on susceptibility to different HD subtypes. *Br J Cancer* 1999: 80: 1405–11.
21. Klitz W, Aldrich CL, Fildes N, Horning SJ, Begovich AB. Localization of predisposition to Hodgkin disease in the HLA class II region. *Am J Hum Genet* 1994: 54: 497–505.
22. Cano P, Fernandez-Vina M. Two sequence dimorphisms of *DPB1* define the immunodominant serologic epitopes of *HLA*-DP. *Hum Immunol* 2009: 70: 836–43.

23. Klitz W, Maiers M, Spellman S et al. New HLA haplotype frequency reference standards: high-resolution and large sample typing of HLA DR-DQ haplotypes in a sample of European Americans. *Tissue Antigens* 2003: 62: 296–307.
24. Amria S, Cameron C, Stuart R, Haque A. Defects in HLA class II antigen presentation in B-cell lymphomas*. Leuk Lymphoma* 2008: 49: 353–5.
25. Diepstra A, van Imhoff GW, Karim-Kos HE et al. HLA class II expression by Hodgkin Reed-Sternberg cells is an independent prognostic factor in classical Hodgkin's lymphoma. *J Clin Oncol* 2007: 25: 3101–8.
26. Rimsza LM, Farinha P, Fuchs DA, Masoudi H, Connors JM, Gascoyne RD. *HLA*-DR protein status predicts survival in patients with diffuse large B-cell lymphoma treated on the MACOP-B chemotherapy regimen. *Leuk Lymphoma* 2007: 48:542–6.

Chapter 3.

HLA Allele Typing African-American Follicular Lymphoma Cases With Formalin-Fixed, Paraffin-Embedded Tissues

Nicholas K. Akers[1], Jacques Riby[2], Martyn T. Smith[1], Christine F. Skibola[2].

[1]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, USA

[2]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Abstract

After extensive research, significant ambiguity remains over which genetic loci within the HLA class II region impact follicular lymphoma risk. In order to resolve this important question, genetic epidemiologists will need to move beyond Caucasian study populations, which have dominated this research field. Populations with African ancestry in particular are known to carry unique haplotypes that may be highly informative in the search for causal variants of follicular lymphoma. Obtaining sufficient subjects to make meaningful comparisons, however, may require supplementing existing lymphoma biobanks with historical samples stored as formalin-fixed, paraffin-embedded (FFPE) tissues. This sample type is abundant, yet unmatched to healthy controls. Furthermore, DNA from this source is technically difficult to work with, often fragmented to very short lengths. We present here a methodology to circumvent these issues; using array hybridization to determine genotypes and using publicly available data to create ancestry matched controls. With pilot data, we demonstrate that HLA class II allele calls from FFPE DNA can be obtained that are of comparable confidence to those derived from non-FFPE DNA. This study, even when reduced to just 16 quality control filtered cases and 64 ancestry matched controls, was able to detect previously described associations at *HLA-DQB1**06 and *HLA-DRB1**15. We present these results here and set forth guidelines for improving allele calls in larger, future studies using FFPE derived DNA.

Introduction

With the recent discovery of HLA class II alleles impacting follicular lymphoma (FL) risk (1,2), significant research has since been dedicated to better characterizing this important association. Studies that have been performed include genome-wide association (3,4), targeted sequencing (5), amino acid analysis (6), expression quantitative trait localization (7), and molecular characterization (8). Despite the successes of these studies, the HLA class II-FL association remains ambiguous. Of primary concern, there is not strong evidence to localize this association to a particular locus. Rather, researchers must rely on associated haplotypes, which span several polymorphic genes, including *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1*.

Resolution of this ambiguity will have multiple benefits to researchers of FL. As with all genetic associations, the motivation for discovery is translation of this newfound knowledge into viable preventative measures and improved medicines for treatment. These desirable outcomes require a fundamental understanding of the mechanism with which a genetic change is impacting disease risk. The complex role of the HLA class II protein group within the immune system obscures this mechanism. Likely a complete understanding the impact of these proteins on FL risk will be achieved only with the aid of *in-vitro* or animal model based studies. Until the precision of the HLA class II-FL associations is improved from haplotypes to individual genes, however, these mechanistic studies will be unable to know, with confidence, which genetic loci to test. As a result, each experiment is two-to-three times more labor intensive and costly.

The FL-associated haplotypes spanning *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* are virtually unbroken in Caucasian populations. Genetic association studies limited to this ethnicity (1,3,5) have little power to distinguish FL associated alleles of *HLA-DRB1* and *HLA-DQB1*. In contrast, African-American populations are predicted to harbor highly informative genotypes that are not typically found in other populations (9). Specifically, 9% of African-Americans have been recorded carrying *DQB1*05:01* without *DRB1*01:01*. The same percentage carry *DRB1*13* without *DQB1*06* (Table 1).

	FL Risk Allele Frequencies		**Haplotypes Frequencies**	
Population	DRB1*01	DQB1*05:01	DQB1*05:01 + NOT DRB1*01	DRB1*01 + NOT DQB1*05:01
Caucasian	0.12	0.12	0.01	0.00
AfrAm.	0.07	0.16	0.09	0.00
Asian-PI	0.03	0.08	0.06	0.00
Hispanic	0.08	0.11	0.04	0.00

	FL Protective Alleles Frequencies			**Haplotypes Frequencies**		
Population	DRB1*15	DRB1*13	DQB1*06	DQB1*06 + NOT DRB1*13/15	DRB1*13 + NOT DQB1*06	DRB1*15 + NOT DQB1*06
Caucasian	0.15	0.11	0.25	0.00	0.01	0.00
AfrAm.	0.14	0.18	0.27	0.05	0.09	0.00
Asian-PI	0.17	0.07	0.22	0.05	0.01	0.07
Hispanic	0.09	0.11	0.18	0.01	0.03	0.00

Table 1. FL associated allele frequencies in four U.S. populations. Data was taken from Maier et al.'s (9) survey of the United States National Bone Marrow Donor Program and includes >500 individuals for each ethnic group typed at *HLA-DRB1* and *HLA-DQB1*. Non-Caucasian populations need to be examined to distinguish FL associated *HLA-DRB1* alleles from their HLA-*DQB1* haplotype partners. PI: Pacific Islander.

An examination of HLA alleles in a population of African-American FL patients would likely provide key insights into the specific loci of association for this disease; however, such a population was not immediately available in either of the San Francisco-Bay Area based case-control studies of non-Hodgkin lymphoma (NHL) (10). African-Americans have a 2-3 fold lower incidence of FL than Caucasian-Americans (11), and are a minority population in the six counties within our study area. As a result, our study alone is underpowered to make comparisons of statistical significance.

One method of obtaining FL case subject DNA is to gather samples from formalin-fixed, paraffin-embedded (FFPE) tissue archives. These archives, originally designed to preserve morphological features of cells, go back as far as 100 years, and are vast in size, containing preserved tissues of virtually all known disease (12,13). In recent years, they have seen resurgence in use due to increased interest in the nucleic acids also preserved. An examination of PubMed publications shows an exponential increase in the use of the term "FFPE", beginning just after completion of the human genome project in 2001 (Figure 1).

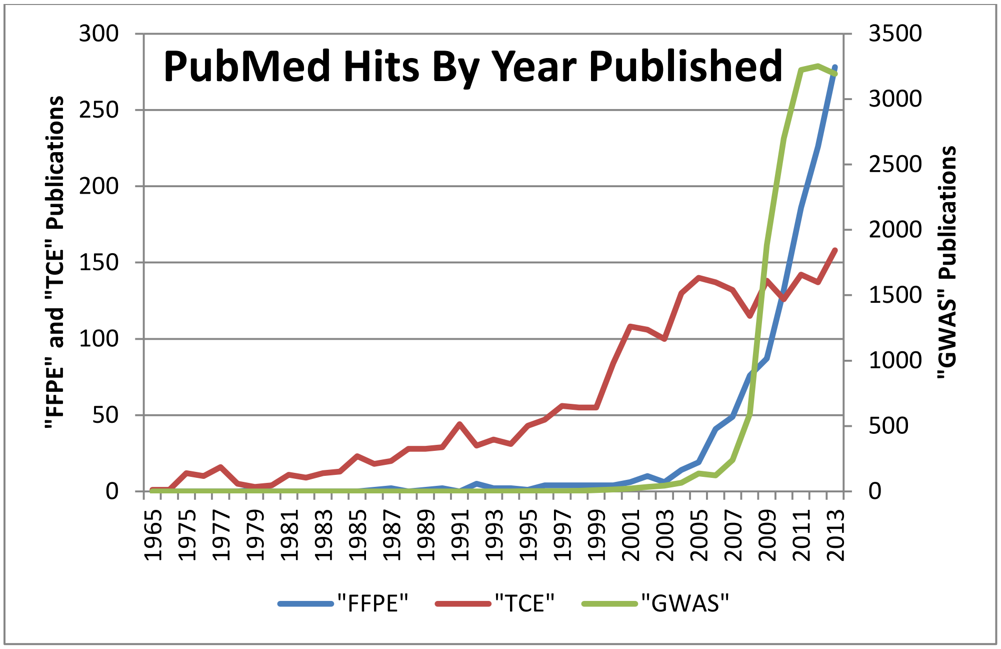


Figure 1. Exponential Growth of FFPE Research. Shown are the number of papers published each year containing the terms, "FFPE", "TCE" and "GWAS". "TCE" and "GWAS" were used as comparisons across the same research and publication environment. The left-hand y-axis scale refers to publications per year containing "FFPE" or "GWAS," while the right-hand y-axis scale refers to publications containing "GWAS." Data gathered from (14).

Although the FFPE tissue archive is vast, using these samples presents technical and methodological challenges. FFPE-extracted DNA and RNA are often of lower quality than nucleic acids preserved by other methods. The fixation process causes the formation of DNA-protein cross-links, and FFPE-extracted DNA is often very fragmented. These traits make amplification and hybridization less reliable in FFPE-extracted DNA (13,15). However, with the increased attention that this field has gained, there are many published guidelines and methodologies for optimal extraction (16–18), as well as numerous commercially available kits for sale.

Epidemiologically, there is a problem in obtaining cases that have not been matched to healthy controls when deriving DNA from FFPE tissues. Well-matched controls are crucial to draw meaningful conclusions from any case-control study. This issue is compounded by the increased importance of ethnicity matching when performing genetic analyses of an admixed population (19). To circumvent potential genetic biases arising in our data, a combination of careful control selection and population stratification adjustment must be employed. Population stratification may be measured by applying principle components analysis to ancestry informative genetic markers (20). The axes of variation calculated in this analysis can then be used to adjust measured genotypes by subtle ancestry differences between case and control populations (21). This general methodology has previously been successfully used to examine the HLA region in an admixed population, using less than 10,000 SNPs to infer population structures (22). An alternative to this 'virtual matching', axes of ancestral variation may be used to directly match specific controls to cases (23,24). This of course requires having controls with similar ancestry to each case. Making use of large, publicly available datasets should provide matches in most cases (23,25).

HLA typing of FFPE tissues is a technically difficult task due to the highly fragmented nature of FFPE-extracted DNA. Classic sequence-based typing techniques for HLA class II genes have relied on polymerase chain reaction (PCR) amplification of ~200 base pair products (26,27), a task that is not reliable with fragmented DNA. A viable alternative to amplification and sequencing techniques for FFPE DNA may be SNP genotyping. SNP arrays have shown utility as a method for genotyping FFPE DNA (28) as well as for inferring HLA allele types (29,30). 'Tag-SNP' HLA typing relies on population specific linkage disequilibrium (LD) and large training data sets to create HLA allele predictive algorithms that use on HLA region SNPs as input. Tag-SNP HLA typing has been developed into several software packages (31,32) specifically designed to predict HLA allele types from SNP-genotyping arrays. To the best of our knowledge, it has not been tested whether SNP array data from FFPE tissues is reliable for imputing HLA alleles.

We set forth to develop a protocol to robustly compare HLA types of African American FL case DNA extracted from FFPE to genetically matched controls. In completing this task, we present here a broadly applicable methodology that may be of interest to researchers using FFPE extracted DNA, performing HLA-tag SNP genotyping, or to those performing genetic analyses in an admixed population. The pilot data shown here indicates that this method is reliable and that HLA allele types imputed from FFPE extracted DNA are of comparable quality to other DNA sources.

Methods

FFPE Samples and DNA extraction

In total, 33 African-American FL-case FFPE tissue blocks from were gathered from the Los Angeles County Residual Tissue Repository. 10µM slices were taken from each sample and DNA was extracted using the High Pure FFPET DNA Isolation Kit (Roche Diagnostics—Indianapolis, IN, USA) according to the manufacturer's protocol. DNA was quantified using the Quant-It PicoGreen dsDNA assay kit (Thermo-Fisher Scientific, Waltham, MA, USA). DNA quality was assessed using the Infinium HD FFPE QC Assay kit (Illumina—San Diego, CA, USA), a real-time polymerase chain reaction (qPCR) assay that compares performance of FFPE extracted DNA to equivalent amounts of high-quality positive control DNA. Twelve FFPE DNA samples of varying quality were taken forward for genotyping, including one duplicate sample that was extracted twice.

Non-FFPE FL Samples

African-American FL cases were also drawn from a San Francisco Bay Area population-based study of Non-Hodgkin Lymphoma. Details of the study have been previously reported (10); however, cases for the research presented here were selected for FL diagnosis, with available DNA, and self-reported as African-American. Nine subjects were identified in this way, with

DNA extracted from whole blood, clotted blood, or buccal cells. DNA was previously extracted and stored at -80°C, but was quantified prior to genotyping for this study.

HLA-DQB1 Allele Typing

HLA-DQB1 alleles were typed in FL samples with sufficient DNA using two different assays. Multiplexed, ligation-dependent probe amplification (MLPA) is able to provide 2-4 digit typing resolution for this gene (33). An alternative assay using sequence-based typing (SBT) was also employed which generally is able to provide 4-digit resolution (27).

Genome-Wide Genotyping

Genotyping was performed on HumanOmniExpress 12.1 chips according to the Illumina Infinium HD Assay Ultra protocol (Illumina—San Diego, CA, USA). Chips were scanned using an Illumina HiScan, and quality control was assessed with Illumina Genome Studio software.

Non-FL Controls

Individual-level SNP data was obtained via the database of Genotypes And Phenotypes (www.dbgap.ncbi.nlm.nih.gov/) for consenting individuals participating in the Jackson Heart Study (data accession phs000499.v2.p1). The Jackson Heart Study is a cohort of non-institutionalized African-Americans, with genome-wide data available for some individuals (34). In total, data from 828 unrelated individuals was used in this study, after removal of related individuals and selection for consenting subjects with genome-wide SNP genotypes. The work presented here has been approved by the Committee for the Protections of Human Subjects at the University of California, Berkeley, and the University of Alabama, Birmingham, as well as the National Heart, Lung, and Blood Institute dbGAP data access committee.

HLA Region Imputation

SNP locations for all data sets used were normalized to the same genome-assembly, NCBI36, using the *liftover* tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Data were converted between formats using the April 15, 2014 release of PLINK 1.90 alpha (https://www.cog-genomics.org/plink2/) and gtool 0.7.5 (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html). To improve our ability to call HLA allele types, SNPs not genotyped in our samples were imputed using Impute2.3.0 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html)(35), using the 1000 Genomes Project pilot data as a reference (36). SNPs in the chromosome 6 interval from 25.5Mbp to 33.5Mbp were imputed, with only those imputed SNPs with greater than 90% probability used in the next phase of analysis.

HLA Alleles via Tag-SNPs

HLA-DRB1 and *HLA-DQB1* alleles were called for both FL cases and controls using the package HIBAG (32), on the statistical software R (http://www.r-project.org/). HIBAG infers HLA alleles from HLA region SNP genotypes using models trained on datasets containing both allele and SNP information. We used 4-digit *HLA-DRB1* and *HLA-DQB1* typing models trained on an African-American cohort with previously demonstrated high accuracy for this population (graciously provided by Dr. Albert Levin, Henry Ford Health System, Detroit, MI). HIBAG outputs the posterior probability of an allele call; using a posterior probability cutoff of >0.10, these models achieved 89.8 % accuracy at *HLA-DRB1* and 97.6% accuracy at *HLA-DQB1*, across a sample size of >150 individuals (37).

Association Testing with Adjustment for Population Stratification

Potential differences in population ancestry between our California-derived FL cases and the control population drawn from Mississippi were accounted for using Eigensoft v5.0.1 (http://www.hsph.harvard.edu/alkes-price/software/) (21). Genotypes for 11 diverse populations were downloaded from the International Hapmap Project Phase 3 (38), and offspring were removed, resulting in 970 individuals. This dataset was merged with our case and control data using Eigensoft *mergit*, resulting in 1,819 individuals with overlapping data at 153,315 SNP positions. Principle components analysis was performed using Eigensoft *smartpca*, with the special parameter, "lsqproject: YES" to use a least squares approach for missing data points. Principle components were generating using only HapMap data, and then projected onto our cases and controls. After 5 outlier removal iterations, 61/970 HapMap subjects were removed using *smartpca*'s default parameters. Regions of long linkage disequilibrium, including the HLA region, were excluded from this analysis as recommended by Price et al. (21). The principle components (or eigenvectors) were extracted for the case and control samples, and input into the Eigensoft *smarteigenstrat* program, with HLA allele call genotypes. The *smarteigenstrat* program then calculates a χ^2 value for each locus tested, using ancestry-adjusted genotypes and ancestry-adjusted case-status.

A second, perhaps more tangible, strategy for ancestry adjustment is to match controls to cases based on proximity in the multi-dimensional space of principle components (23,24). The benefit of this is the ability to perform a more classic comparison of alleles between cases and controls. To aid in this, a script was written in Perl (http://www.perl.org/) to match controls to each case, with variable inputs for the number of controls to match, the number of principle components to match on, and a maximum distance between matched samples (See Supplementary Materials). For each subject, values at each eigenvector were weighted by multiplying the square root of the eigenvalue by the unweighted eigenvector. The Euclidean distance in multidimensional eigenvector space was calculated for each case-control combination, and the controls closest to each case were selected as a comparison group (without allowing the same control to be used twice). The R command "fisher.test" in the "stat" package was then used to calculate odds ratios, confidence intervals, and p-values.

Results

FFPE DNA was initially assessed for performance using call rate on the Illumina OmniExpress chip. We observed a wide range in SNP call rates (0.53-0.92) that was strongly correlated with qPCR performance (R^2 = 0.78—Figure 2). In contrast, all SNP call rates in our African-American DNA extracted from non-FFPE tissues were >0.996.

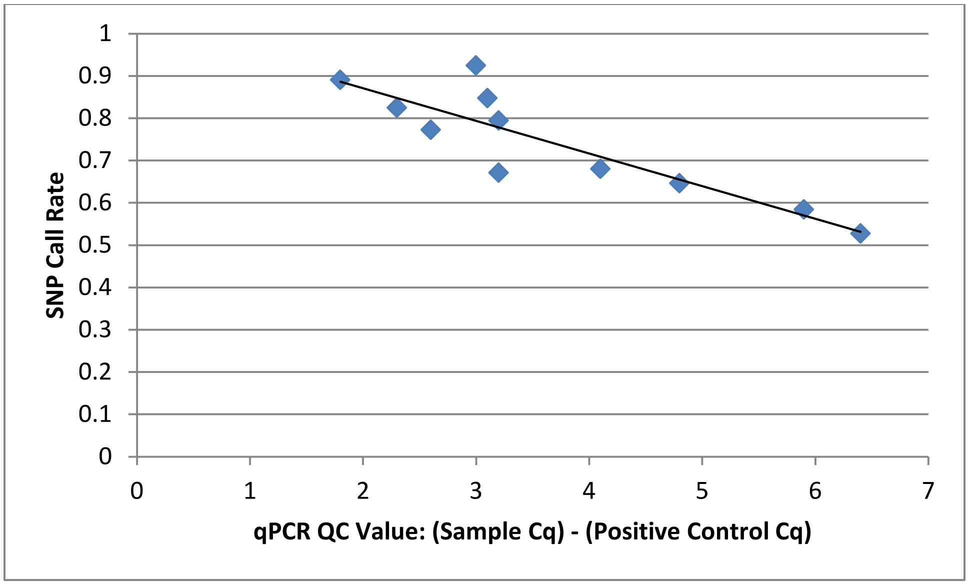


Figure 2. qPCR Performance Predicts SNP Call Success in FFPE DNA. The quality of FFPE DNA can be assessed by comparing the qPCR Cq values of FFPE DNA and high quality reference DNA. The x-axis here shows these differences with a lower value indicating higher quality DNA. A wide range of Cq values was input to assess the ability of the SNP array to genotype lower quality samples. There was a strong trend towards improved SNP call rate with better qPCR performance.

To improve our ability to call HLA alleles, ungenotyped SNPs in the HLA region were imputed using IMPUTE2. Using internal cross-validation built into IMPUTE2, we measured the concordance of known SNP genotypes with imputed calls on these genotypes when masked. Using a maximum probability cutoff of >0.9; control, non-FFPE case, and FFPE case samples had 98.2%, 97.4%, and 85.4% concordance, respectively.

HLA allele calls were imputed using HIBAG, which gives a posterior probability for the most-likely allele call. All samples examined were above a posterior probability cutoff of 0.10 for *HLA-DQB1*. All but 1 FFPE sample (9%) and 4 control samples (0.5%) were above this cutoff for *HLA-DRB1*. Interestingly, posterior probability of HLA allele calls was not strongly correlated with SNP call rate (Supplementary Figure 1). *HLA-DQB1* allele calls in FL cases were validated using laboratory data, with SBT and MLPA typing providing *HLA-DQB1* allele calls for 18/18 non-FFPE case alleles, and 16/24 FFPE case samples. FFPE DNA samples performed inconsistently in laboratory assays, exhibiting poor amplification during PCR steps. Of the 18 non-FFPE case alleles, HIBAG and laboratory assays agreed on 16 alleles (89%). Among FFPE derived samples, HIBAG called the same allele as laboratory assays in 13/16 alleles (81%) (Supplementary Table 1). Examining the larger population of control samples, imputed HLA allele frequencies aligned well with expected frequencies based on previously published frequencies (Table 2).

Allele	**N**	**Freq.**	**Exp.**
*DRB1**01	112	0.07	0.07
*DQB1**05:01	271	0.16	0.16
*DQB1**05:01 NOT *DRB1**01	142	0.09	0.09
*DRB1**13	277	0.17	0.18
*DRB1**15	217	0.13	0.14
*DQB1**06	436	0.26	0.27
*DRB1**13 NOT *DQB1**06	91	0.05	0.09
*DQB1**06 NOT *DRB1**13/15	63	0.04	0.05

Table 2. Imputed Allele Frequencies vs. Expected. Shown are allele counts and frequencies for 828 non FL African-American subjects, imputed based on SNP genotypes. The expected frequencies in the rightmost column are based on a self-reported African American group of bone marrow donors (9). Most alleles and haplotypes were very close in frequency to the previously published estimates for this ethnicity, giving confidence to this method of allele typing. N = Total number of alleles or haplotypes imputed, Freq.=Frequency, Exp.=Expected frequency.

Principle components analysis was used to infer ancestry among our cases and controls. Supplementary Table 2 shows the top 10 computed eigenvectors for the 11 worldwide populations in HapMap, and the proportion of variance explained by each eigenvector. This table demonstrates that after the top three eigenvectors, each additional component explains <1% of the total genetic variance. The eigenvectors calculated using HapMap3 samples were projected onto our FL cases and controls to assess ancestry of our subjects. Figure 3 shows the first four eigenvectors; however, similar charts for eigenvectors 5-10 can be seen in Supplementary Figures 2-4. There is a clear spread of ancestry among the African-American subjects, particularly along the first three eigenvectors. However, control subjects, case subjects, and African-American HapMap populations overlapped well, particularly in the first two eigenvectors. There was a group of four FL case subjects (including both versions of the one duplicate sample) that appeared to be outliers from the rest of the African-American samples. These cases were all derived from FFPE tissue, but were not associated with SNP call rate or any other notable trait (Supplementary Table 1). This could indicate issues with ancestry reports in the FFPE tissue archive.

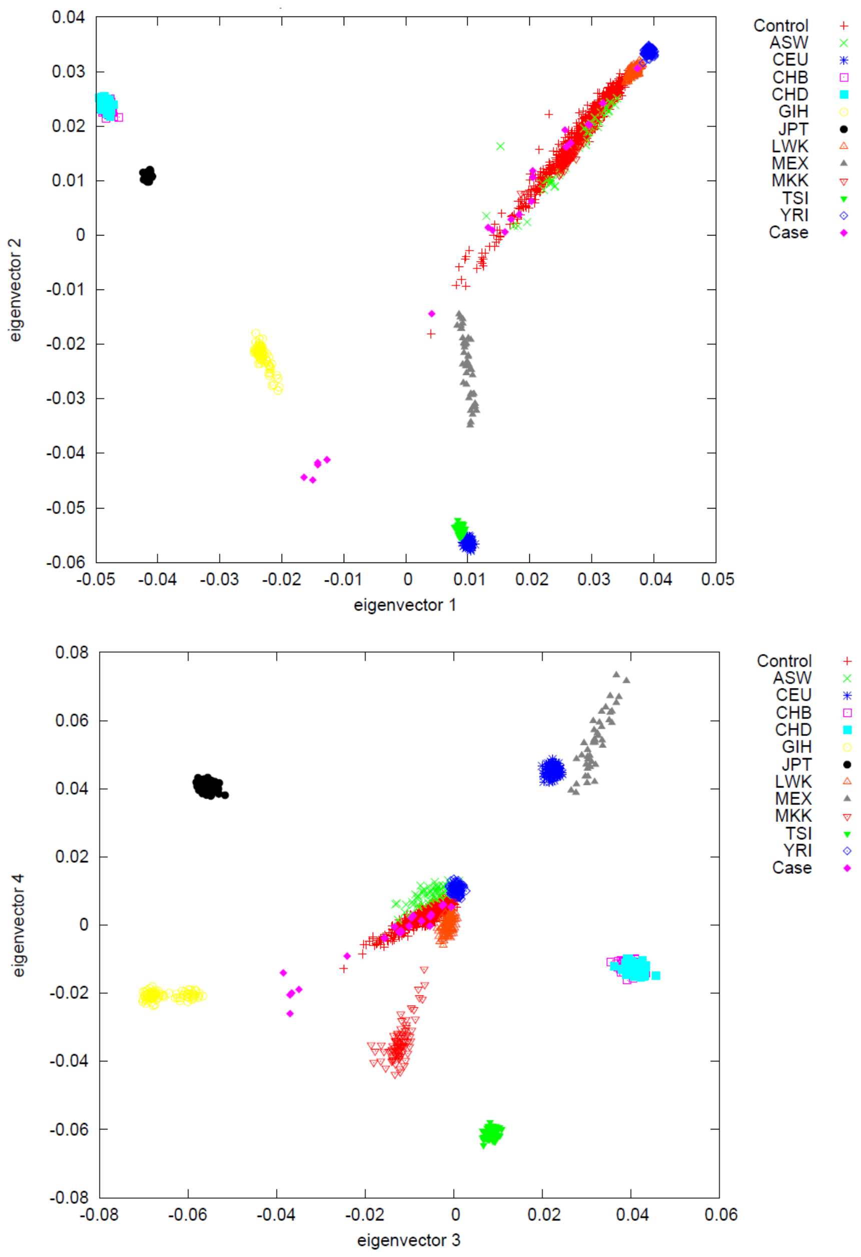


Figure 3. FL Cases and Controls Projected onto Top Axes of Variation. The upper panel plots the top two eigenvectors, while eigenvectors 3 and 4 are plotted in the bottom panel. Shown as pink diamonds are FL case subjects. These samples were recorded or self-reported as African-American, however a great deal of ancestral variation appears to exist in these samples, emphasizing the need to use genetic variation to assist in matching controls. Control subjects from the Jackson Heart Study had variation along the first 3 eigenvectors, but aligned well with the African-American HapMap population 'ASW'. These individual eigenvector values will be used to correct for population stratification or match controls on ancestry. Population codes: ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria.

Matching controls to cases on the top 3 principle components resulted in 5 FL samples (4 subjects and 1 duplicate) being removed as outliers for which no controls could be matched. The other 16 cases were matched to 4 controls each (Figure 4). Variation of these samples on eigenvectors 5-10 can be seen in Supplementary Figures 5-7.

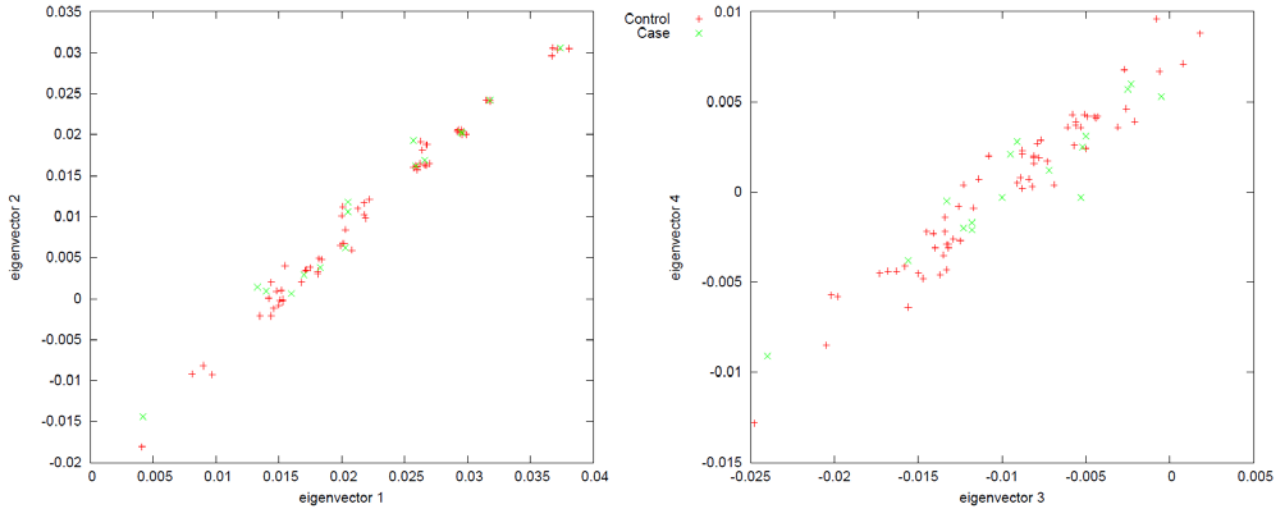


Figure 4. Ancestry Matched FL Cases and Controls. A reduction of Figure 3, this figure shows African-American FL cases derived from FFPE and non-FFPE sources (green) and the top four controls for each case (red), matched by weighted Euclidean distance to the top 3 eigenvectors. Each control was allowed to match to only one case.

Association tests were run on 8 HLA alleles and haplotypes using two methods, Fisher's exact test with 16 cases and 64 ancestry matched controls, and the *smarteigenstrat* χ^2 test making use of 21 cases and 828 controls. The results of these tests are shown in Table 4. Two alleles, *DRB1**15 and *DQB1**06 were significantly associated with FL (p<0.05) according to both methodologies. Using matched controls, the allele *DQB1**05:01 approached significance when considering all carriers (*p* = 0.08), or subjects that carry the allele without *DRB1**01 (p = 0.14).

Allele / Haplotype	Cases (%)	Controls (%)	OR	95% CI	*p*-value	Eig. χ^2	Eig. *p*
*DRB1**01	2 (0.13)	4 (0.06)	2.12	(0.26, Inf)	0.34	0.01	0.92
*DQB1**05:01	7 (0.44)	14 (0.22)	2.74	(0.88, Inf)	0.08	0.26	0.61
*DQB1**05:01 without *DRB1**01	5 (0.31)	10 (0.16)	2.42	(0.67, Inf)	0.14	0.88	0.34
*DRB1**13	4 (0.25)	20 (0.31)	0.74	(0, 2.37)	0.44	0.00	1.00
*DRB1**15	1 (0.06)	23 (0.36)	0.12	(0, 0.72)	0.02	4.83	0.03
*DQB1**06	3 (0.19)	35 (0.55)	0.19	(0, 0.67)	0.01	7.87	0.005
*DRB1**13 without *DQB1**06	1 (0.06)	7 (0.11)	0.55	(0, 3.79)	0.50	0.45	0.50
*DQB1**06 withot *DRB1**13/15	0 (0)	4 (0.06)	0.00	(0, 4.52)	0.40	1.40	0.23

Table 4. Association of HLA Class II Alleles with FL. The leftmost columns were calculated using ancestry-matched controls (N=64) to African-American FL cases (N=16). The R command 'fisher.test' was used to calculate odds ratios, 95% confidence intervals and 1-tailed p-values for each allele or haplotype. A subject was counted if they carried one or two copies of the allele. In the case of haplotypes, subjects were counted if they carried one or two copies of the first allele listed and zero copies of the second allele listed. The right-most columns were created using the *smarteigenstrat* ancestry correction to calculate χ^2 values and the corresponding p-values with 1 degree of freedom. Abbreviations: OR: odds ratio, CI: confidence interval, *p*: *p*-value, Inf: infinite. Eig: Eigenstrat.

Discussion

This paper presents a methodology that can be used in genetic epidemiological studies making use of FFPE tissues. We provide recommendations for optimizing quality of DNA extracted from FFPE tissues and show that even low-quality DNA can provide robust HLA allele calls. Using this methodology, we have created promising pilot data that aligns well with previously described associations, indicating that our strategy is sound.

Optimization of FFPE DNA extraction is crucial to success in downstream applications. We have found that tissue slices 10um or less more readily undergo protein digestion, and produce higher quality DNA. When it is impossible to obtain an entire tissue block, it is preferable to obtain a 10um tissue section, rather than a small 'core' of the entire tissue block. We found that working with tissue cores was prohibitively difficult, given the physical challenge of taking 10um sections from millimeter wide cores. It is advisable not to store tissues in thin sections for long, due to atmospheric oxidation; however, we have extracted robust, high-quality DNA from sections exposed to air for several days while in transit. All FFPE DNA will be fragmented as a result of the fixation process. Therefore, it is preferable to perform DNA extraction that removes small size DNA fragments. We've found that the Roche High Pure FFPET DNA Isolation Kit, which uses High Pure Filter Columns, effectively removes all DNA less than 100nt in length. This is a simple and elegant solution to removing DNA fragments which have potential to interfere with downstream applications. We show here that improved DNA quality is directly and strongly correlated with improved performance in genome-wide SNP genotyping assays (Figure 2).

Higher SNP call rates were associated with improved sample performance during imputation of ungenotyped SNPs, as measured by the concordance of known SNP genotypes with imputed genotypes. High quality imputations greatly aid in expanding the total number of SNPs in a region. This can become important when downstream applications rely on the overlap of specific SNPs, as is the case when calculating principle components across different populations, or performing HLA allele imputations using a SNP-based model.

Despite a number of our FFPE derived samples showing relatively poor SNP call rates, our HLA allele calling performed well overall, with a comparable successful call rate compared to non-FFPE DNA. The models we used to impute HLA alleles are reported to achieve 89.8 % accuracy at *HLA-DRB1* and 97.6% accuracy at *HLA-DQB1* when a posterior probability cutoff of 0.10 is used (37). At the *HLA-DRB1* locus, we observed ranges of posterior probability from 0.16-0.85 and 0.09-0.53 for non-FFPE and FFPE extracted samples, respectively. These ranges were 0.12-0.91 and 0.15-0.85 at *HLA-DQB1* (Supplementary Table 1). These values indicate that we could likely expect greater than 90% accuracy in *HLA-DRB1* call rates and greater than 98% accuracy in *HLA-DQB1* call rates. Testing *HLA-DQB1* allele imputations with other allele typing assays yielded lower than 98% agreement however. Alleles imputed from SNPs were in agreement with alleles typed using MLPA and/or SBT in 81% of FFPE sample alleles, and 89% of non-FFPE sample alleles. All discordant alleles had posterior probabilities <0.4, indicating a higher stringency cutoff may be appropriate. It should also be noted that no HLA typing method is 100% accurate, and there is no reported reliable method for HLA typing FFPE tissues. Given these issues, we consider our results a success.

The value of SNP-array based allele typing vs classic HLA typing in these FFPE samples is most clear when we examine the total call rate. A combination of MLPA and SBT, run twice for each sample, yielded 3/12 samples with no reliable *HLA-DQB1* allele calls and an additional 2/12 samples with just one allele called. HIBAG gave allele calls for each sample with posterior probabilities to indicate the quality of the call.

We observed four FFPE-derived FL cases with ancestry that was discordant from other observed African-American populations (Figure 3). These samples did not have exceptionally low SNP call rates (Supplementary Table 1), indicating that these samples may not be African-American, as reported. There are numerous reasons why historical samples may have incorrect ethnicity records: initial recording of ethnicity may have been performed improperly, discordant ethnicities may have been merged for simplicity, or patients themselves may have been motivated to report incorrect ethnicity information. This finding underscores the value of genome-wide ancestry information in a study such as this.

Association tests run using the samples we had available were remarkably consistent with previous findings. Table 4 shows that all previously described FL risk alleles had odds ratios

(OR) greater than 2.0, and all previously described FL protective alleles less than 0.75. Of course, there is limited power with such few cases to demonstrate that these ORs differ from 1.0 with any confidence. Despite this, our numbers do indicate that the alleles *HLA-DQB1**06 (OR=0.19, 95% confidence interval [CI] = (0, 0.67), *p* = 0.01) and *HLA-DRB1**15 (OR=0.12, 95% CI = (0, 0.72), *p* = 0.02) have ORs significantly less than 1.0 (α = 0.05). These results are consistent when eigenstrat genotype/phenotype ancestry corrections are made on our full dataset rather than using ancestry matched controls (Table 4). These allele associations are an encouraging sign that the methodology used in this study is working. Far greater sample numbers will be needed to resolve the ambiguity in FL associated HLA haplotypes; however, the early trends observed here indicate that this process can be successful when applied on a larger scale.

In summary, we show here that FFPE tissues represent a viable source of FL patient DNA, and this DNA produces high-quality HLA class II allele calls. The combination of genome-wide SNP arrays, SNP imputation, and tag-SNP HLA typing was robust even with low quality FFPE DNA. Concordance rates of HLA allele calls between this method and other HLA typing methods were similarly high in FFPE and non-FFPE extracted DNA. We show that genome-wide SNPs can be used to assess ancestry in FFPE extracted DNA, and that this process is crucial when selecting a non-diseased comparison group. Finally, we show that even with a very small number of cases, the effect of certain HLA alleles on FL risk becomes evident using this method. The use of FFPE tissues, publicly available sequence databases, and admixed populations in disease research will likely grow in the coming years. The techniques outlined here are broadly applicable to this related research, and can be considered guidelines for use of FFPE DNA in genetic epidemiology, particularly in an admixed population.

Acknowledgements

This work was supported by National Institutes of Health grants CA154643 and CA104682 (CFS).

Bibliography

1. Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet. 2010 Aug;42(8):661–4.
2. Wang SS, Abdou AM, Morton LM, Thomas R, Cerhan JR, Gao X, et al. Human leukocyte antigen class I and II alleles in non-Hodgkin lymphoma etiology. Blood. 2010 Jun 10;115(23):4820–3.
3. Smedby KE, Foo JN, Skibola CF, Darabi H, Conde L, Hjalgrim H, et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet. 2011 Apr;7(4):e1001378.
4. Christine F Skibola, Sonja I Berndt, Joseph Vijai, Lucia Conde, Zhaoming Wang, Meredith Yeager, et al. Genome-wide association study identifies new follicular lymphoma susceptibility loci. Nat Genet. Submitted;
5. Skibola CF, Akers NK, Conde L, Ladner M, Hawbecker SK, Cohen F, et al. Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. Tissue Antigens. 2012 Apr;79(4):279–86.
6. Foo JN, Smedby KE, Akers NK, Berglund M, Irwan ID, Jia X, et al. Coding variants at hexa-allelic amino acid 13 of HLA-DRB1 explain independent SNP associations with follicular lymphoma risk. Am J Hum Genet. 2013 Jul 11;93(1):167–72.
7. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013 Jan 10;92(1):126–30.
8. Sillé FCM, Conde L, Zhang J, Akers NK, Sanchez S, Maltbaek J, et al. Follicular lymphoma-protective HLA class II variants correlate with increased HLA-DQB1 protein expression. Genes Immun. 2013 Dec 5;
9. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol. 2007 Sep;68(9):779–88.
10. Skibola CF, Bracci PM, Halperin E, Nieters A, Hubbard A, Paynter RA, et al. Polymorphisms in the estrogen receptor 1 and vitamin C and matrix metalloproteinase gene families are associated with susceptibility to lymphoma. PloS One. 2008;3(7):e2816.
11. Groves FD, Linet MS, Travis LB, Devesa SS. Cancer surveillance series: non-Hodgkin's lymphoma incidence by histologic subtype in the United States from 1978 through 1995. J Natl Cancer Inst. 2000 Aug 2;92(15):1240–51.
12. Von Ahlfen S, Missel A, Bendrat K, Schlumpberger M. Determinants of RNA Quality from FFPE Samples. Fraser P, editor. PLoS ONE. 2007 Dec 5;2(12):e1261.
13. Lewis F, Maughan NJ, Smith V, Hillan K, Quirke P. Unlocking the archive - gene expression in paraffin-embedded tissue. J Pathol. 2001 Sep;195(1):66–71.
14. PubMed - NCBI [Internet]. [cited 2014 Mar 20]. Available from: http://www.ncbi.nlm.nih.gov/pubmed/?term=FFPE
15. Little SE, Vuononvirta R, Reis-Filho JS, Natrajan R, Iravani M, Fenwick K, et al. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. Genomics. 2006 Feb;87(2):298–306.
16. Kotorashvili A, Ramnauth A, Liu C, Lin J, Ye K, Kim R, et al. Effective DNA/RNA Co-Extraction for Analysis of MicroRNAs, mRNAs, and Genomic DNA from Formalin-Fixed Paraffin-Embedded Specimens. Wong C-M, editor. PLoS ONE. 2012 Apr 13;7(4):e34683.
17. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. 2010 Aug 1;38(14):e151–e151.

18. Tang W, David FB, Wilson MM, Barwick BG, Leyland-Jones BR, Bouzyk MM. DNA Extraction from Formalin-Fixed, Paraffin-Embedded Tissue. Cold Spring Harb Protoc. 2009 Feb 1;2009(2):pdb.prot5138–pdb.prot5138.
19. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, et al. Population stratification confounds genetic association studies among Latinos. Hum Genet. 2006 Jan;118(5):652–64.
20. Murray T, Beaty TH, Mathias RA, Rafaels N, Grant AV, Faruque MU, et al. African and non-African admixture components in African Americans and an African Caribbean population. Genet Epidemiol. 2010 Sep;34(6):561–8.
21. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9.
22. López Herráez D, Martínez-Bueno M, Riba L, García de la Torre I, Sacnún M, Goñi M, et al. Rheumatoid arthritis in Latin Americans enriched for Amerindian ancestry is associated with loci in chromosomes 1, 12, and 13, and the HLA class II region. Arthritis Rheum. 2013 Jun;65(6):1457–67.
23. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The Population Reference Sample, POPRES: A Resource for Population, Disease, and Pharmacological Genetics Research. Am J Hum Genet. 2008 Sep;83(3):347–58.
24. Imielinski M, Baldassano RN, Griffiths A, Russell RK, Annese V, Dubinsky M, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet. 2009 Dec;41(12):1335–40.
25. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007 Oct;39(10):1181–6.
26. Dinauer DM, Luhm RA, Uzgiris AJ, Eckels DD, Hessner MJ. Sequence-based typing of HLA class II DQB1. Tissue Antigens. 2000 Apr;55(4):364–8.
27. Van Dijk A, Melchers R, Tilanus M, Rozemuller E. HLA-DQB1 sequencing-based typing updated. Tissue Antigens. 2007 Apr;69 Suppl 1:64–5.
28. Lips EH, Dierssen JWF, van Eijk R, Oosting J, Eilers PHC, Tollenaar RAEM, et al. Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays. Cancer Res. 2005 Nov 15;65(22):10188–91.
29. De Bakker PIW, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, et al. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet. 2006 Oct;38(10):1166–72.
30. Leslie S, Donnelly P, McVean G. A statistical method for predicting classical HLA alleles from SNP data. Am J Hum Genet. 2008 Jan;82(1):48–56.
31. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP--an integrated framework for imputing classical HLA alleles from SNP genotypes. Bioinforma Oxf Engl. 2011 Apr 1;27(7):968–72.
32. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG-HLA genotype imputation with attribute bagging. Pharmacogenomics J. 2014 Apr;14(2):192–200.
33. Akers NK, Curry JD, Smith MT, Bracci PM, Skibola CF. Multiplexed, ligation-dependent probe amplification for rapid and inexpensive HLA-DQB1 allelotyping. Tissue Antigens. 2011 Oct;78(4):275–80.
34. Musunuru K, Lettre G, Young T, Farlow DN, Pirruccello JP, Ejebe KG, et al. Candidate gene association resource (CARe): design, methods, and proof of concept. Circ Cardiovasc Genet. 2010 Jun;3(3):267–75.
35. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012 Aug;44(8):955–9.
36. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.
37. Levin A et al. Submitted;

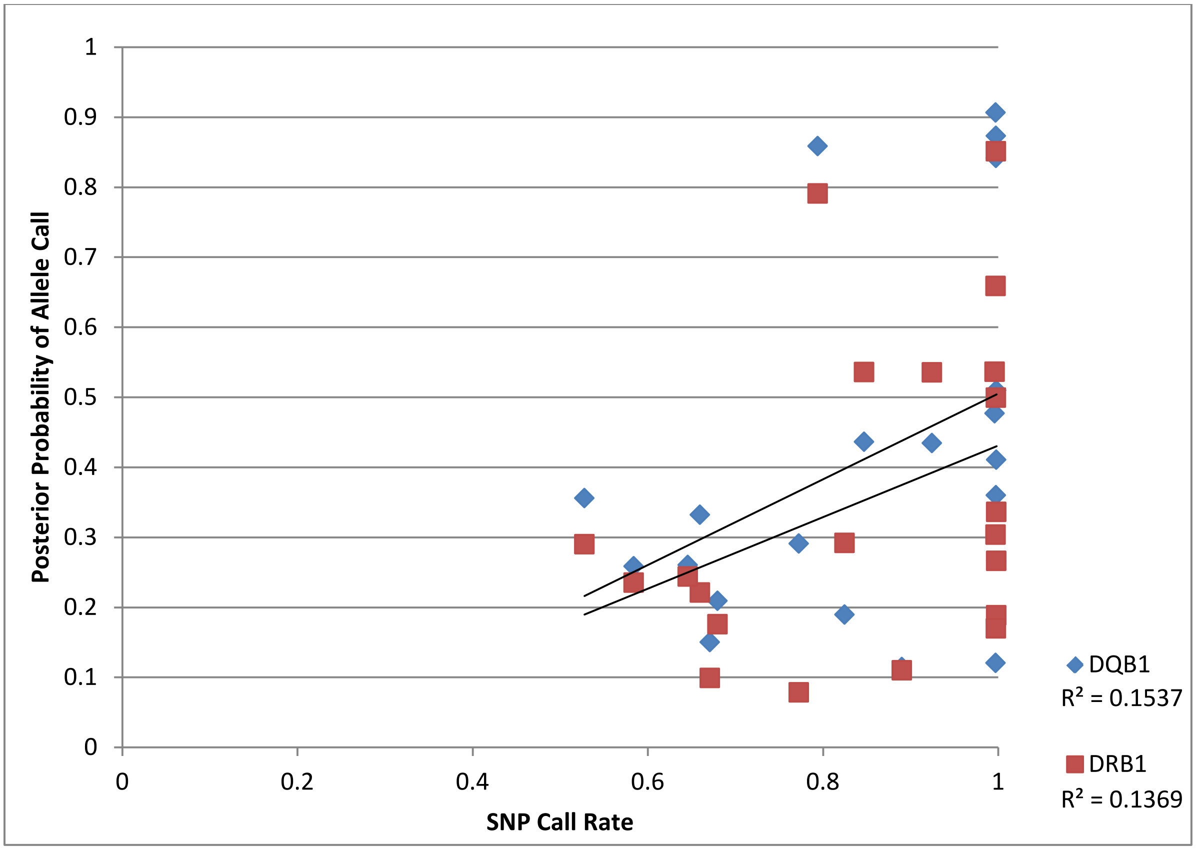
38. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2;467(7311):52–8.

	SNP	SNP Imputed Alleles						PCR-Based Allele Typing		
	Call	*HLA-DRB1*			*HLA-DQB1*			*HLA-DQB1* Alleles		Discordant
ID	Rate	A1	A2	Prob.	A1	A2	Prob.	SBT	MLPA	Alleles
FL-1	0.996	08:04	11:01	0.54	03:01	03:01	0.48	03:01	03 03	0/2
FL-2	0.998	11:01	12:01	0.50	03:01	05:01	0.84	03:01 05:01	03 05	0/2
FL-3	0.998	04:01	09:01	0.27	02:02	02:02	0.27	02:01 03:01	02 03	1/2
FL-4	0.998	07:01	12:01	0.19	02:02	05:01	0.51	02:01 05:01	02 05	0/2
FL-5	0.998	03:02	09:01	0.34	02:02	04:02	0.41	02:01 04:02	02 04	0/2
FL-6	0.997	01:01	08:04	0.66	03:01	05:01	0.91	03:01 05:01	03 05	0/2
FL-7	0.997	03:02	12:01	0.85	04:02	05:01	0.87	04:02 05:01	04 05	0/2
FL-8	0.998	07:01	11:01	0.17	02:02	03:01	0.36	02:01 03:01	02 03	0/2
FL-9	0.997	08:04	15:03	0.30	03:01	04:02	0.12	02:01 04:02	02 04	1/2
FFPE-1	0.925	13:02	13:04	0.54	03:01	06:09	0.43	03:01 06:03 06:09	03 06:05/06:09	0/2
FFPE-2	0.528	11:02	13:01	0.29	03:01	06:03	0.36	02:01 03:01	NA	1/2
FFPE-3	0.646	01:01	11:01	0.24	03:01	05:01	0.26	05:01 06:04	NA	1/2
FFPE-4	0.660	03:01	13:04	0.22	02:01	03:01	0.33	NA	NA	0/0
FFPE-5	0.584	03:01	12:01	0.24	03:01	05:01	0.26	06:04 05:01	02 05	0/1
FFPE-6	0.680	03:01	12:01	0.18	02:01	05:01	0.21	02:01		0/1
FFPE-7	0.847	13:02	13:04	0.54	03:01	06:09	0.44	NA	03 06:05/06:09	0/2
FFPE-8	0.825	11:01	11:01	0.29	03:01	03:01	0.19	NA	03 03	0/2
FFPE-9	0.890	13:02	14:01	0.11	05:03	06:03	0.11	02:01 05:03 06:03	05 06:03	0/2
FFPE-10	0.794	01:01	07:01	0.79	02:02	05:01	0.86	NA	NA	0/0
FFPE-11	0.772	07:01	07:01	0.08	02:02	02:02	0.29	NA	NA	0/0
FFPE-11 (dup.)	0.671	07:01	13:01	0.10	02:02	03:01	0.15	NA	03 05	1/2

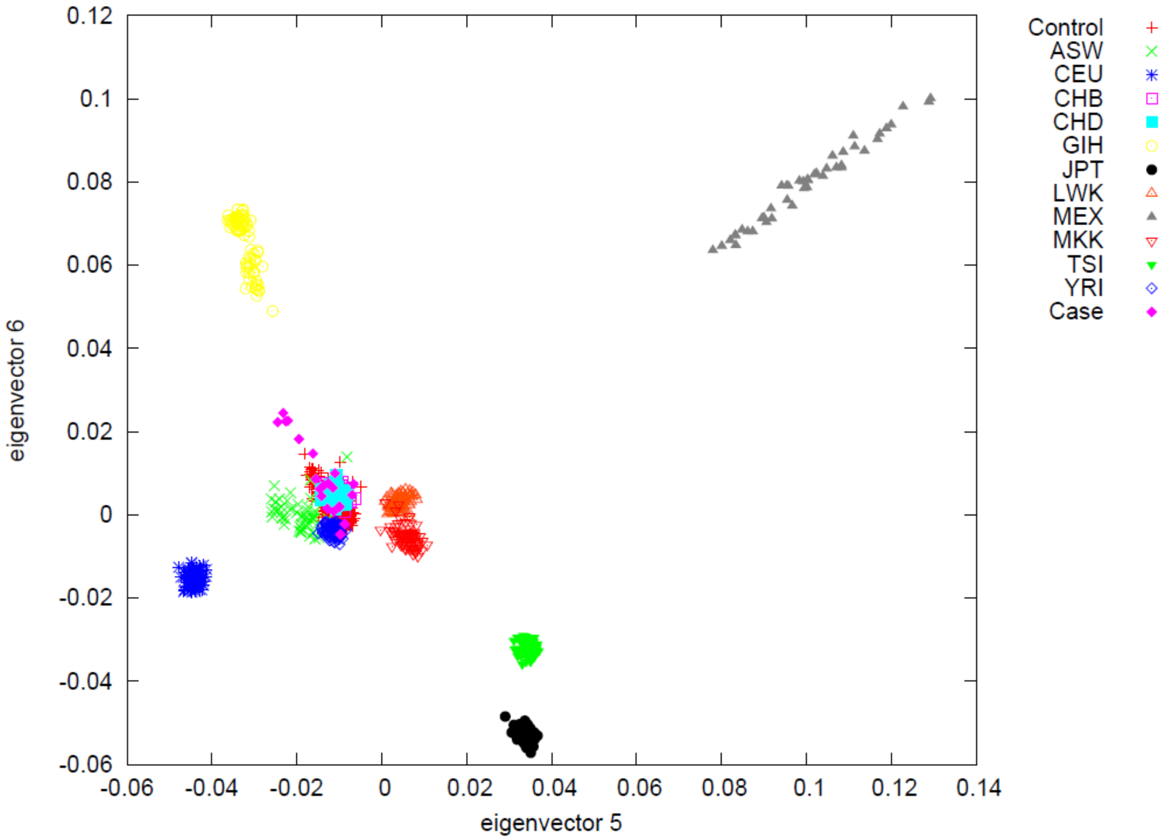
Supplemental Table 1. FL sample information. For each FL subject examined in this study, SNP call rates and imputed alleles at *HLA-DRB1* and *HLA-DQB1* based on those SNPs are given. Also shown are the laboratory determined, PCR-based *HLA-DQB1* alleles determined using sequence-based typing (SBT) and multiplexed-ligation dependent probe amplification (MLPA). When alleles could not be determined using an assay, 'NA' is shown. Discordant alleles were counted when both SBT and MLPA agreed on an allele, but SNP imputed alleles disagreed with that allele. When MLPA and SBT disagreed, that sample was not counted (e.g. FFPE-5). The final five samples (FFPE-8-11) were observed to be outliers in PCA analysis (Figure 3). Prob.=Posterior probability of allele call determined by HIBAG.

Eigenvector	Eigenvalue	% Variance	Cumulative % Variance
1	76.4	8.4	8.4
2	30.2	3.3	11.7
3	12.3	1.4	13.1
4	7.5	0.8	13.9
5	6.9	0.8	14.7
6	6.0	0.7	15.4
7	3.3	0.4	15.7
8	1.7	0.2	15.9
9	1.4	0.2	16.1
10	1.4	0.2	16.2

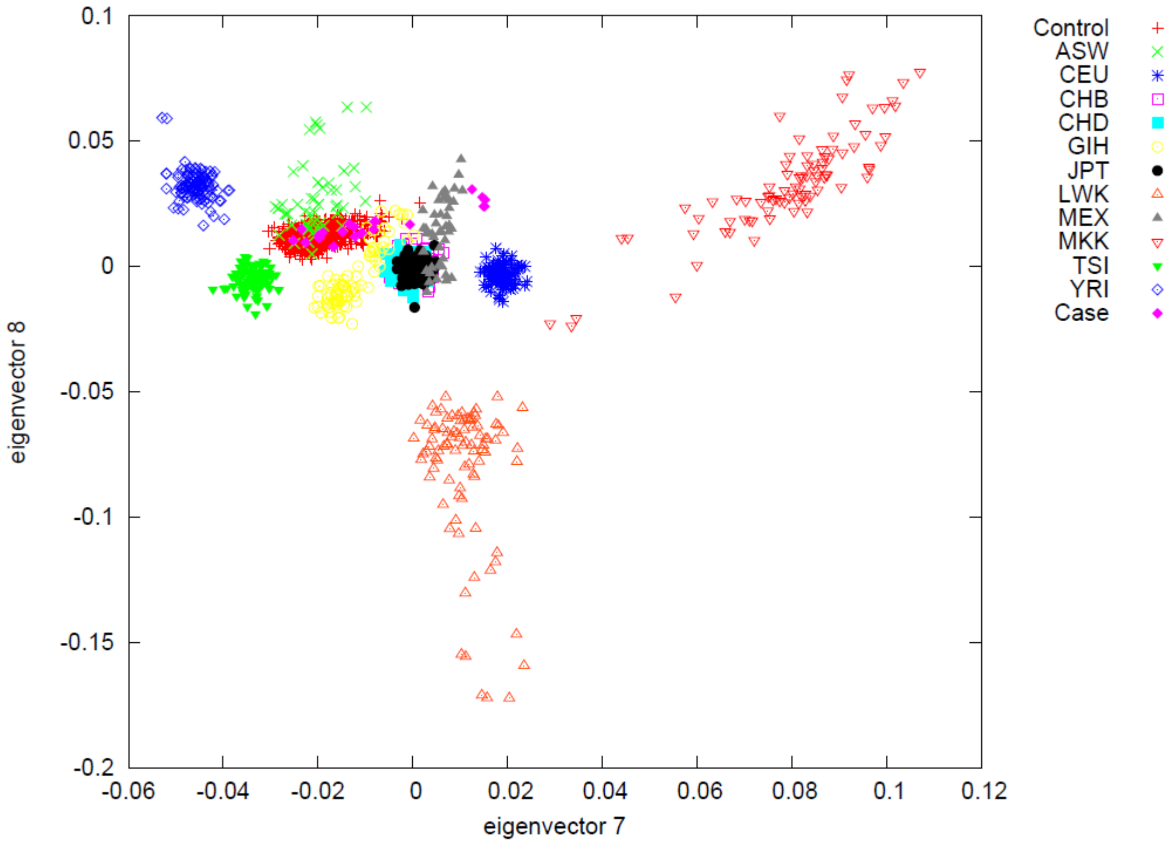
Supplementary Table 2. Principle Components Variance Explained. The Eigensoft script *smartpca* calculated eigenvalues and eigenvectors for the HapMap3 dataset. Proportion of variance explained by eigenvector K was calculated by dividing eigenvalue K by the sum of all eigenvalues. With this methodology, we can see that after 3 principle components there is diminishing value to considering more eigenvectors.



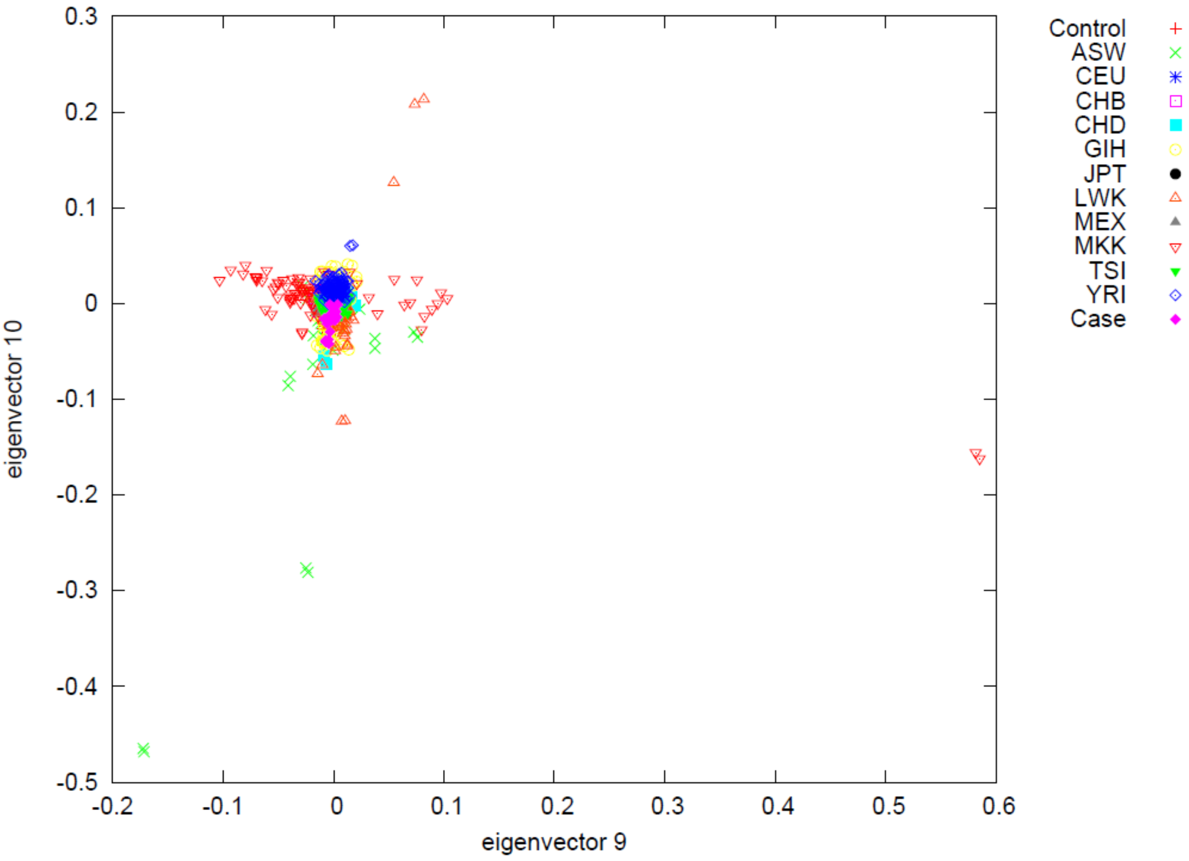
Supplementary Figure 1. SNP Call Rate and HLA Allele Call Confidence. Initial SNP call rates from the Illumina Omni-Express Chip are plotted against the posterior probability of the most likely HLA allele calls from HIBAG. The samples plotted here are the 21 African-American FL cases from FFPE and non FFPE tissues. There is a trend towards more data providing higher confidence allele calls; however, the correlation is low, indicating call rate is not the major source of variation in this data. Certain alleles are likely more difficult to call with confidence.



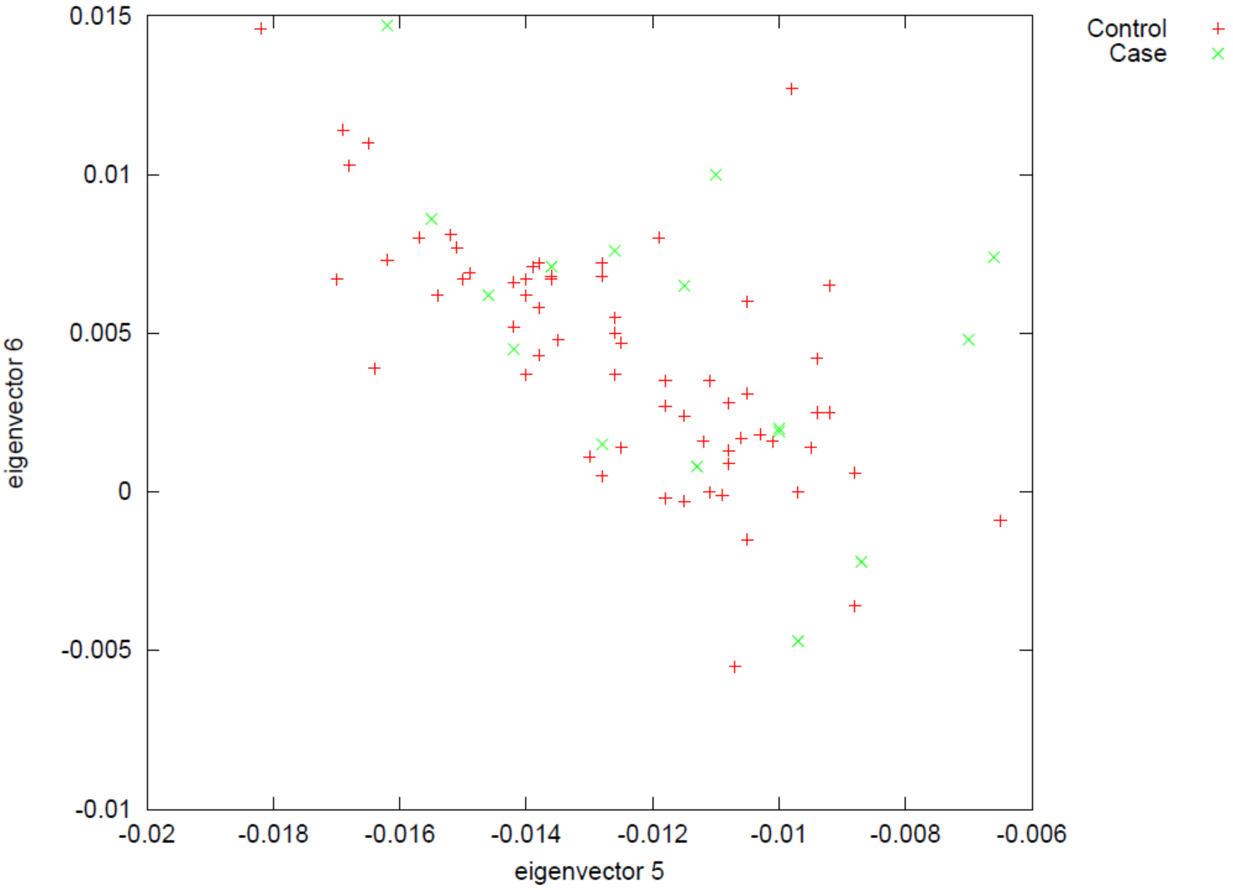
Supplementary Figure 2. FL Cases and Controls with HapMap Samples. Eigenvectors were calculated using HapMap3 samples, and then projected onto African American FL cases and controls. Eigenvectors 5 and 6 are shown here. Population codes: ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria.



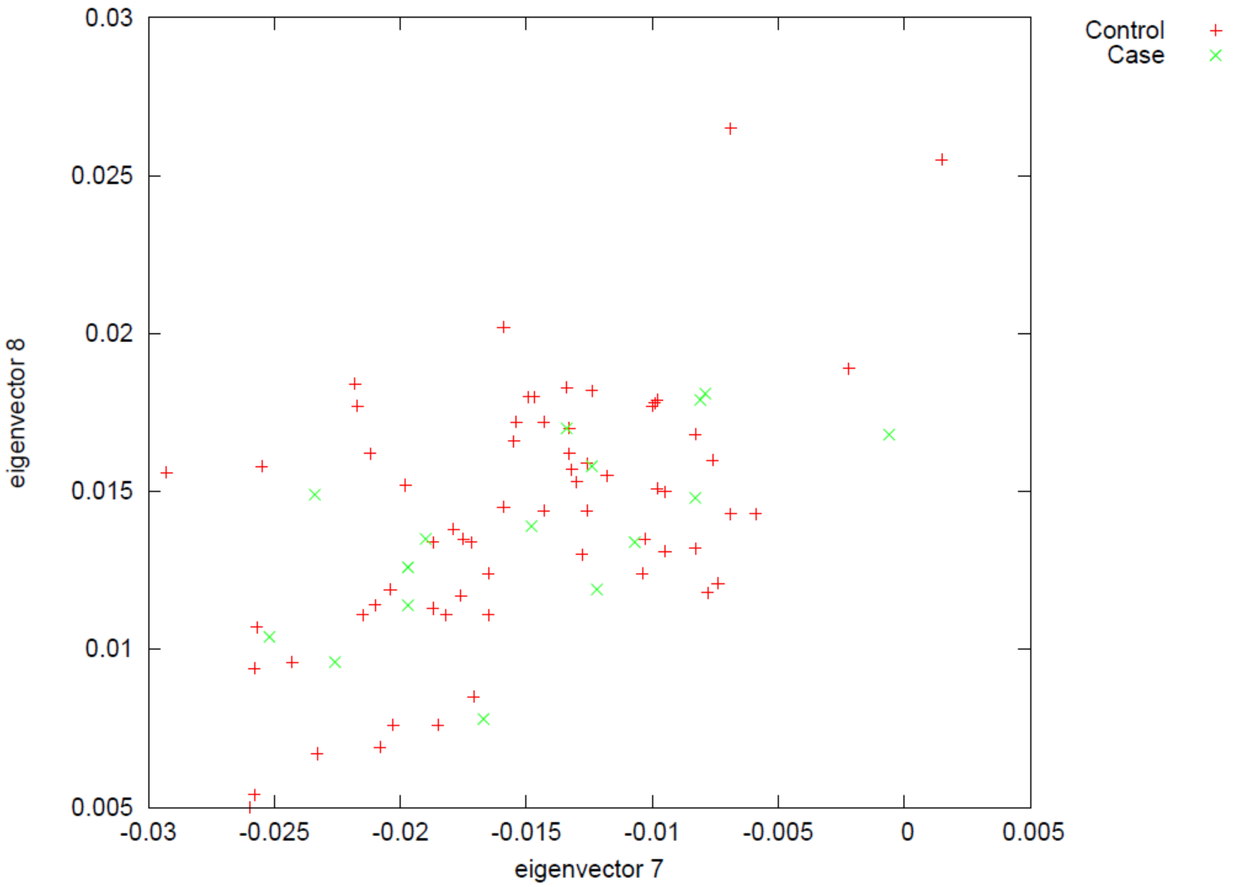
Supplementary Figure 3. FL Cases and Controls with HapMap Samples. Eigenvectors were calculated using HapMap3 samples, and then projected onto African American FL cases and controls. Eigenvectors 7 and 8 are shown here. Population codes: ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria.



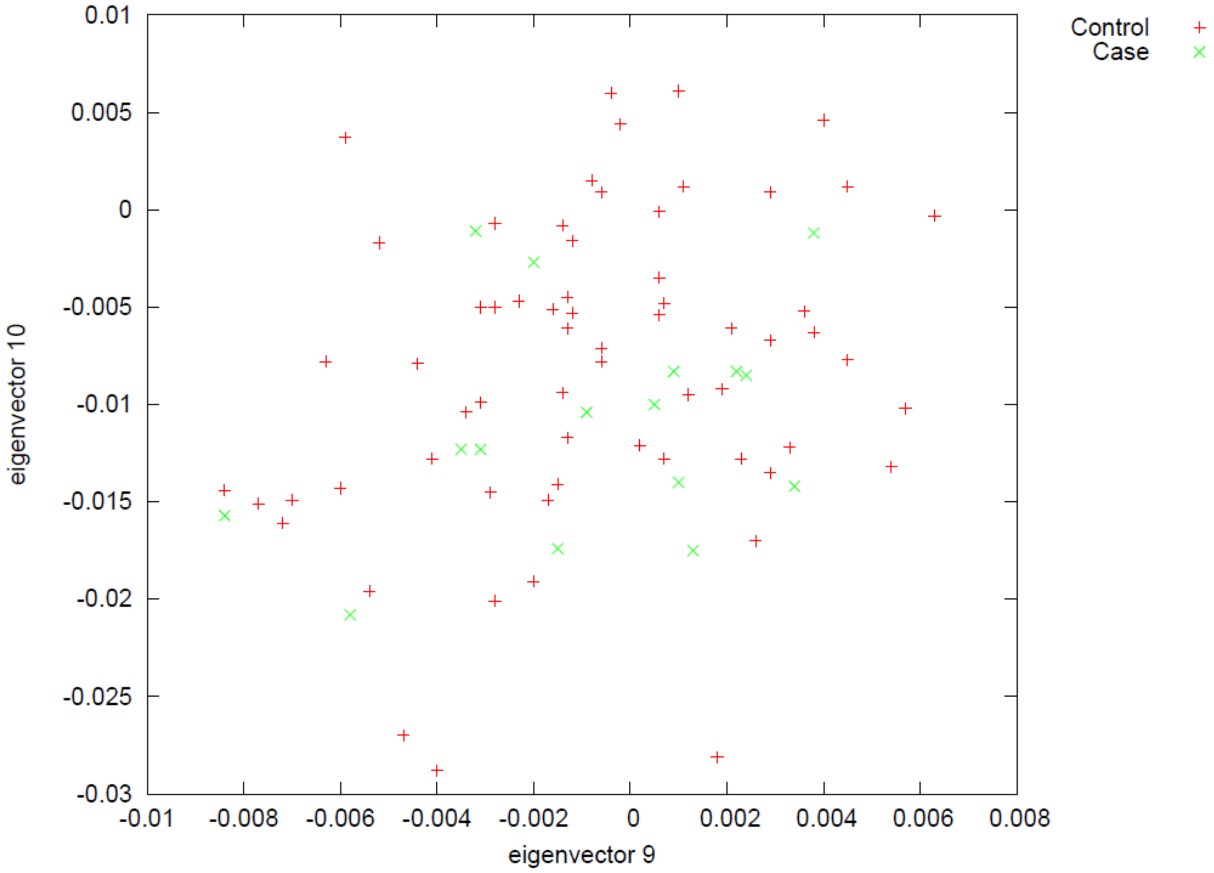
Supplementary Figure 4. FL Cases and Controls with HapMap Samples. Eigenvectors were calculated using HapMap3 samples, and then projected onto African American FL cases and controls. Eigenvectors 9 and 10 are shown here. Population codes: ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MEX: Mexican ancestry in Los Angeles, California; MKK: Maasai in Kinyawa, Kenya; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria.



Supplementary Figure 5. FL Ancestry Matched FL Cases and Controls. A reduction of Supplementary Figure 2, this shows African-American FL cases derived from FFPE and non-FFPE sources (green) and the top four controls for each case (red), matched by weighted Euclidean distance to the top 3 eigenvectors. Each control was allowed to match to only one case. Despite these eigenvectors not being used in the control matching process, the two populations do not appear different.



Supplementary Figure 6. FL Ancestry Matched FL Cases and Controls. A reduction of Supplementary Figure 3, this shows African-American FL cases derived from FFPE and non-FFPE sources (green) and the top four controls for each case (red), matched by weighted Euclidean distance to the top 3 eigenvectors. Each control was allowed to match to only one case. Despite these eigenvectors not being used in the control matching process, the two populations do not appear different.



Supplementary Figure 6. FL Ancestry Matched FL Cases and Controls. A reduction of Supplementary Figure 4, this shows African-American FL cases derived from FFPE and non-FFPE sources (green) and the top four controls for each case (red), matched by weighted Euclidean distance to the top 3 eigenvectors. Each control was allowed to match to only one case. Despite these eigenvectors not being used in the control matching process, the two populations do not appear different.

Supplementary Material: Code Used to Match Cases and Controls. The code below, named pcamatcher.pl, can be used in a linux computing environment to create controls matched to cases on the top eigenvectors from Eigensoft output.

```
#! /usr/bin/perl
use warnings;

#pcamatcher.pl by Kipp Akers
# script to match cases to controls on principle components, using
eigensoft.evec file
# usage: pcamatcher.pl input.evec numberofpcs youroutputname

open (INPUT, $ARGV[0]) or die $!;
$outfile = $ARGV[2];
$totalPCs = $ARGV[1];
print "reading input file $ARGV[0]\n";
open (OUT, ">$outfile") or die $!;
$counter = 0;
$numcontrols =0;
$numcases = 0;
$controlstomatch = 4; #adjustable: number of controls to match
@comparison = ();
$distcutoff = .1;   #adjustable: max distance between a case and control.
0.1 seems reasonable.
%usedcontrols = ();
print OUT "CaseID\tMatchedControlID\tEuclidDist\tMatchRank\tOther Parameters:
Matching $controlstomatch controls on $totalPCs PCs with distance cutoff
$distcutoff\n";

while (<INPUT>) {
    $counter++;
    chomp;
    $_ =~ s/^\s+//;
    if ($counter ==1) {
        @eigenvalues = split(/\s+/, $_);#we'll use these as weights: 0=trash
    }
    else {
        my @wholeline= split(/\s+/, $_); #0-ID #2-(x-1): eigenvals,#[-
1]:phenotype
        if ($wholeline[-1] =~ m/Control/) {
            $numcontrols++;
            $controls[$numcontrols][0] = $wholeline[0]; #ID name
            foreach my $x (1..$totalPCs) {
                $controls[$numcontrols][$x] =
$wholeline[$x]*sqrt($eigenvalues[$x]);
                 #this builds an array of arrays: $controls[indiv
number][PCnum] = weighted eigenscore
            }
        }
                if ($wholeline[-1] =~ m/Case/) {
                        $numcases++ ;
                        $cases[$numcases][0] = $wholeline[0]; #ID name
                        foreach my $x (1..$totalPCs) {
                                $cases[$numcases][$x] =
$wholeline[$x]*sqrt($eigenvalues[$x]);
```

```
                        }         #this builds an array of arrays:
$cases[indiv number][PCnum] = weighted eigenscore
                }
    }
}
close (INPUT);
open (OUT2, ">allcombinations.txt") or die $!;
foreach my $caseindex (1..$numcases) { #cycle through all the cases
    %control_dist = ();
    @casescores = ();
    foreach my $PCs (1..$totalPCs) {
        push (@casescores, $cases[$caseindex][$PCs]); #create an array,
@casescores, with all the weighted vals.
    }
    foreach my $controlindex (1..$numcontrols) { #cycle through all controls
        @comparison = @casescores;
        foreach my $PCs (1 ..$totalPCs) { #this loop should create a
2xtotalPCs array ie (case1, case2, case3, control1, control2, control3)
            push (@comparison, $controls[$controlindex][$PCs]);
        }
        $compdistance = &distance(@comparison); #calculate euclidian distance
        $control_dist{$controls[$controlindex][0]} = $compdistance ;#hash of
distances from case.  key is controlID
    }
    #print OUT2 "$cases[$caseindex][0]\n";  #this block can be used to print
out all the combinations, if needed
    #while ( my ($key, $value) = each%control_dist) {
        #print OUT2 "$key\t$value\n";
    #}
    my @sortedmatches = sort { $control_dist{$a} <=> $control_dist{$b} } keys
(%control_dist) ; #sort the controls by distance
    $tempcounter = 0;
    $tempcounter2 =0;
    while ($tempcounter2 < $controlstomatch) {
        #$tempcounter++;
        if ($control_dist{$sortedmatches[$tempcounter]} > $distcutoff) { #if
its a bad enough match
            print "No more controls found for $cases[$caseindex][0] after
$tempcounter2 successes. Lowest distance:
$control_dist{$sortedmatches[$tempcounter]}\n";
            $tempcounter2 = $controlstomatch;
        }
        else {
            if (exists $usedcontrols{$sortedmatches[$tempcounter]} ) {#if the
control has already been matched
                #print "subtracing";
            }
            else {
                print OUT
"$cases[$caseindex][0]\t$sortedmatches[$tempcounter]\t$control_dist{$sortedma
tches[$tempcounter]}\tnthbest:$tempcounter\n"; #print the matches
                $usedcontrols{$sortedmatches[$tempcounter]} = 1;
                $tempcounter2++;
            }
        }
        $tempcounter++;
    }
```

```
}



#this subroutine copied directly from Orwant et al "Mastering Algorithms with
Perl"
# distance( @p ) computes Euclidean distance between 2 d-dimensional points
# example: 3-D points : ( $x0, $y0, $z0, $x1, $y1, $z1)

sub distance {      #this subroutine copied directly from Orwant et al
"Mastering Algorithms with Perl"
    my @p = @_;     #the coordinates of the points
    my $d = @p / 2; #the number of dimensions
    # the case of 2 dimensions is optimized
    return sqrt( ($_[0] - $_[2])**2 + ($_[1] - $_[3])**2 )
        if $d == 2;

    my $S = 0;  #the sum of the squares
    my @p0 = splice @p, 0, $d;  #the starting point

    for ( my $i = 0; $i < $d; $i++ ) {
        my $di = $p0[ $i ] - $p[ $i ]; #Difference....
        $S += $di * $di;    #squared and summed
    }
    return sqrt( $S );
}
```

Chapter 4.

Coding Variants at Hexa-allelic Amino Acid 13 of HLA-DRB1 Explain Independent SNP Associations with Follicular Lymphoma Risk*

Jia Nee Foo,[1,17] Karin E. Smedby,[2,17] Nicholas K. Akers,[3] Mattias Berglund,[4] Ishak D. Irwan,[1] Xiaoming Jia,[5,6] Yi Li,[1] Lucia Conde,[7] Hatef Darabi,[8] Paige M. Bracci,[9] Mads Melbye,[10] Hans-Olov Adami,[8,11] Bengt Glimelius,[4] Chiea Chuen Khor,[1,12,13,14] Henrik Hjalgrim,[10] Leonid Padyukov,[15] Keith Humphreys,[8] Gunilla Enblad,[4] Christine F. Skibola,[7,18] Paul I.W. de Bakker,[5,6,16,18] and Jianjun Liu[1,14,18]

1 Human Genetics, Genome Institute of Singapore, Agency for Science, Technology, and Research, Singapore 138672, Singapore;
2 Clinical Epidemiology Unit, Department of Medicine, Karolinska Institutet, Stockholm SE-171 77, Sweden;
3 Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, USA;
4 Department of Radiology, Oncology, and Radiation Science, Uppsala University, Uppsala SE-751 85, Sweden;
5 Division of Genetics, Brigham andWomen's Hospital, Harvard Medical School, Boston, MA 02115, USA;
6 Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA;
7 Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA;
8 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-171 77, Sweden;
9 Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94107, USA;
10 Department of Epidemiology Research, Statens Serum Institut, Copenhagen DK-2300, Denmark;
11 Department of Epidemiology, Harvard University, Boston, MA 02115, USA;
12 Singapore Eye Research Institute, Singapore 168751, Singapore; 13Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119074, Singapore;
14 Saw Swee Hock School of Public Health, National University Health Systems, National University of Singapore, Singapore 117597, Singapore;
15 Rheumatology Unit, Department of Medicine, Karolinska Institutet and Karolinska University Hospital Solna, Stockholm SE-171 76, Sweden;
16 Department of Medical Genetics and of Epidemiology, University Medical Center Utrecht, Utrecht 3584 CX, the Netherlands
17 These authors contributed equally to this work
18 These authors contributed equally to this work and are co-senior authors

*A similar version of this manuscript has been published: Foo JN, Smedby KE, Akers NK, Berglund M, Irwan ID, Jia X, et al. Coding variants at hexa-allelic amino acid 13 of HLA-DRB1 explain independent SNP associations with follicular lymphoma risk. Am J Hum Genet. 2013 Jul 11;93(1):167–72.
This chapter is printed here with acknowledgement to all co-authors and Elsevier.

Abstract

Non-Hodgkin lymphoma represents a diverse group of blood malignancies, of which follicular lymphoma (FL) is a common subtype. Previous genome-wide association studies (GWASs) have identified in the human leukocyte antigen (HLA) class II region multiple independent SNPs that are significantly associated with FL risk. To dissect these signals and determine whether coding variants in HLA genes are responsible for the associations, we conducted imputation, HLA typing, and sequencing in three independent populations for a total of 689 cases and 2,446 controls. We identified a hexa-allelic amino acid polymorphism at position 13 of the HLA-DR beta chain that showed the strongest association with FL within the major histocompatibility complex (MHC) region (multiallelic $p = 2.3 \times 10^{-15}$). Out of six possible amino acids that occurred at that position within the population, we classified two as high risk (Tyr and Phe), two as low risk (Ser and Arg), and two as moderate risk (His and Gly). There was a 4.2-fold difference in risk (95% confidence interval = 2.9–6.1) between subjects carrying two alleles encoding high-risk amino acids and those carrying two alleles encoding low-risk amino acids ($p =1.01 \times 10^{-14}$). This coding variant might explain the complex SNP associations identified by GWASs and suggests a common HLA-DR antigen-driven mechanism for the pathogenesis of FL and rheumatoid arthritis.

Four genome-wide association studies (GWASs) have recently revealed complex associations between genetic variants in the human leukocyte antigen (HLA) region and follicular lymphoma (FL) risk, (1–4) particularly two independent associations tagged by rs10484561 (1) and rs2647012 (2) within the HLA class II region. Further imputation with tag SNPs (1,5) and HLA typing (6) revealed that coding-sequence variation in the molecules encoded by the extended *HLA-DRB1*0101-HLA-DQA1*0101-HLA-DQB1*0501* haplotype might be responsible for the association at rs10484561 and that the *DRB1*15-DQA1*01-DQB1*06* haplotype might partly explain the association at rs2647012 (6). A recent analysis also observed the association between gene-expression changes and rs2647012, but not rs10484561 (7). These previous findings indicate an important role of genetic variation in the HLA class II region in FL pathogenesis, but the underlying causal variants that drive the association are still unknown. Each extended haplotype and classical HLA allele is defined by a precise combination of coding differences at various amino acid (AA) positions in the encoded HLA molecules, and it is possible that changes at the AA level might impact antigen binding and therefore influence disease pathogenesis through altered immune response.

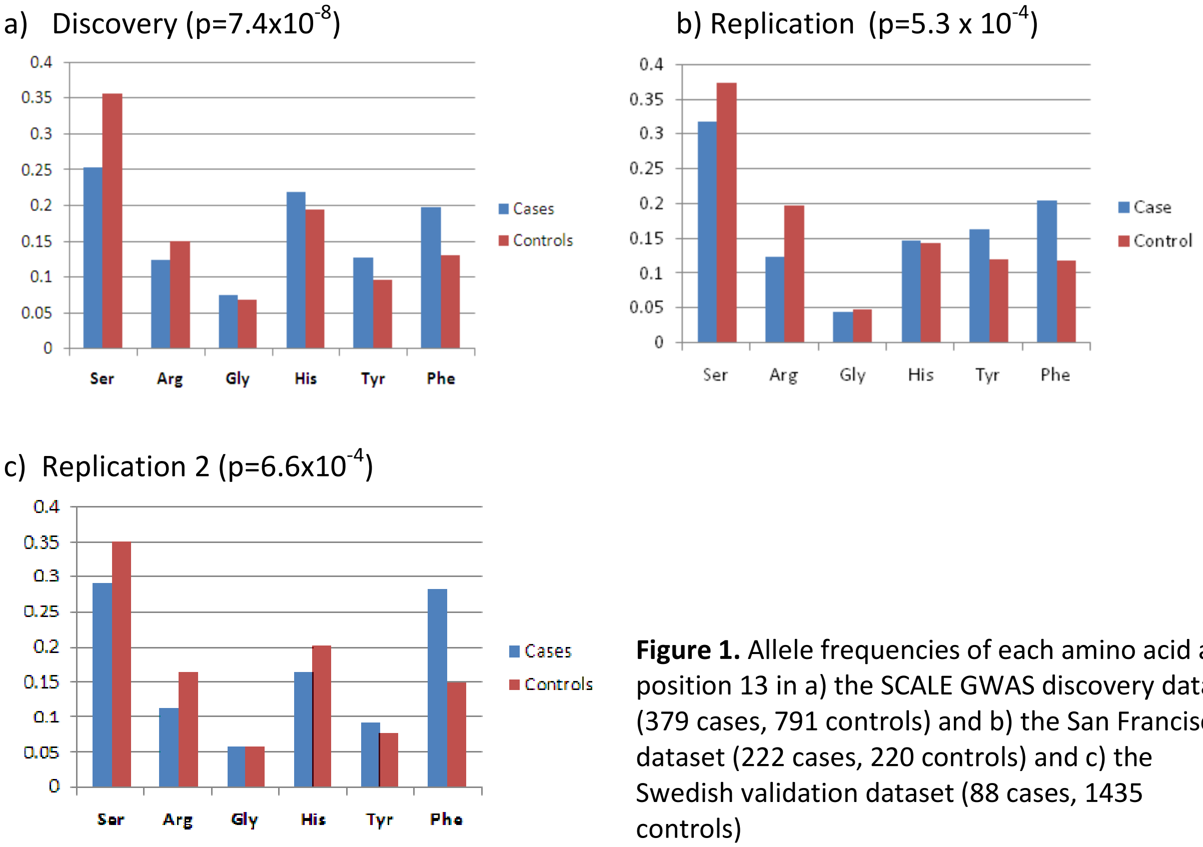
To determine whether specific coding variants within HLA genes contribute to the diverse association signals, we imputed dense SNPs (by using 1000 Genomes Project data (8)) and classical HLA alleles and coding variants across the HLA region (b36, chr6: 20–40 Mb) in our GWAS discovery sample of 379 cases and 791 controls (2) from Sweden and Denmark (the Scandinavian Lymphoma Etiology [SCALE] study). Imputation of classical HLA alleles and their constituent single-nucleotide variants and corresponding AAs was performed with BEAGLE as previously described (9,10). The reference panel was constructed with the use of genotype data from the Major Histocompatibility Complex (MHC) Working Group of the Type 1 Diabetes Genetics Consortium; these data consist of 2,537 SNPs genotyped in the MHC region and classical types at 4-digit resolution from 11,173 individuals. On the basis of the EMBL-EBI Immunogenetics HLA Database, AA variants (based on codons) were coded as binary markers (present or absent) in the reference panel. The final reference panel for imputation contained 2,767 unrelated founder individuals of European descent from collections across Europe, the United Kingdom, and North America (9,10) and data for 263 classical HLA alleles and 372 AA positions. We used default parameters for BEAGLE (ten iterations of phasing and imputation and testing four pairs of haplotype pairs for each individual at each iteration) (9,10) (average BEAGLE r^2 = 0.96). SNPs were imputed with IMPUTE2 (v.2.2.2) (11) on the basis of phased genotype data from the 1000 Genomes Project Phase I integrated variant set v.3 (8) (average info score for SNPs with MAF > 0.01 = 0.90). We set all SNP genotypes imputed with genotype probability lower than 0.9 to missing and analyzed all imputed HLA genotypes. SNPs imputed with an info score < 0.8 in IMPUTE2, >10% missing data, or MAF < 0.01 were excluded from further analyses. A total of 71,954 SNPs (average info score = 0.97), 263 classical HLA alleles (two- and four-digit resolution), and 372 AA positions were imputed. Of these, 86 AA positions were “multiallelic,” i.e., had three to six different residues encoded at each position across all our samples.

We conducted trend tests on all the imputed biallelic variants while adjusting for PC1-3 as covariates (2) (genomic inflation in the GWAS = 1.028). For multiallelic AA sites (triallelic, quadrallelic, penta-allelic, or hexa-allelic), we performed the global multiallelic test at each site and further tested every possible combination of one, two, or three encoded AAs against the rest. Association tests were performed on both the best-guess genotypes and the allelic dosages as determined from the imputed genotype probabilities, which accounted for any uncertainties in the imputed genotypes. The results were checked for consistency between the two methods, and the results from best-guess genotypes were presented. We also performed global multiallelic (genotypic) tests while taking into account all alleles at each multiallelic position. The multiallelic test was performed as follows: convert k-alleles to k-1 bialleles, invoke the glm function in R to estimate the multivariate model, and use the likelihood ratio test to compute the global multiallelic test p values. We performed all conditional logistic regression analyses with PLINK (12) by entering any five out of six alleles encoding AA 13 as covariates. Pairwise linkage disequilibrium (LD; r^2) between SNPs and alleles were measured in PLINK.

	No. of Cases (Freq.)	No. of Controls (Freq.)	OR (95% CI)	P
Ser + Arg at AA 13				
Discovery	379 (0.379)	791 (0.509)	0.591 (0.495-0.707)	7.83 x 10^{-9} Dosage: 8.26 x 10^{-9}
Replication 1	222 (0.439)	220 (0.573)	0.584 (0.446-0.766)	1.01 x 10^{-4}
Replication 2	88 (0.403)	1,435 (0.515)	0.640 (0.469-0.872)	4.69 x 10^{-3}
Meta-analysis	**P**	**OR**	**P_{het}**	**I^2**
	6.51x10^{-14}	0.598	0.893	0
Tyr + Phe at AA 13				
Discovery	379 (0.326)	791 (0.228)	1.660 (1.363-2.020)	4.87 x10^{-7} Dosage: 4.69x10^{-7}
Replication 1	222 (0.365)	220 (0.239)	1.822 (1.355-2.450)	7.15x10^{-5}
Replication 2	88 (0.375)	1,.435 (0.226)	1.954 (1.437-2.660)	1.90 x10^{-5}
Meta-analysis	**P**	**OR**	**P_{het}**	**I^2**
	2.00 x10^{-14}	1.760	0.659	0

Table 1. Association Statistics from Testing Protective Alleles versus All Others and Risk Alleles versus All Others at AA 13 of HLA-DRB1. Abbreviations: OR, odds ratio; CI, confidence interval; AA, amino acid; P_{het}, Cochran's Q test p value; and I^2, inconsistency measure.

Among all the variants tested, the top signal of association came from a combination of two possible genotypes, coding for either Ser or Arg at the hexa-allelic AA 13 encoded by exon 2 of *HLA-DRB1*. This variant showed stronger association (Table 1; $p_{Ser+Arg13} = 7.8 \times 10^{-9}$, odds ratio [OR] = 0.59; 95% confidence interval [CI] = 0.50–0.70) than any other HLA variant or SNP (genotyped or imputed) tested in our study and the strongest association in global multiallelic tests across all AA positions (multiallelic $p = 7.4 \times 10^{-8}$). AA 13 genotypes accounted for slightly more variance in FL risk (3.1%) than both rs2647012 and rs10484561 combined (3.0%). The strongest-associated imputed classical HLA allele across all two- and four-digit alleles was *HLA-DQB1∗06* (OR = 0.63; 95% CI = 0.51–0.78; $p = 2.05 \times 10^{-5}$), which could be accounted for by rs2647012 alone ($OR_{adj\ rs2647012}$ = 0.86; 95% CI = 0.65–1.15; $p_{adj\ rs2647012}$ = 0.32). The top imputed SNP was rs9268839, in complete LD (r^2 = 1; crude OR = 1.64; 95% CI = 1.37–1.95; p = 4.57 × 10^{-8}) with rs9378212 (which we previously reported and validated technically by Taqman genotyping (2)) and partially correlated with rs2647012 (r^2 = 0.56) and rs10484561 (r^2 = 0.15) ($OR_{adj\ 2snps}$ = 1.25; 95% CI = 0.94–1.64; $p_{adj\ 2snps}$ = 0.12). The FL-associated HLA SNPs reported in a recent GWAS (4) were also not independent of these previously reported SNPs in our data set (Table S1).


a) Discovery ($p=7.4x10^{-8}$)

b) Replication ($p=5.3 \times 10^{-4}$)

c) Replication 2 ($p=6.6x10^{-4}$)

Figure 1. Allele frequencies of each amino acid at position 13 in a) the SCALE GWAS discovery dataset (379 cases, 791 controls) and b) the San Francisco dataset (222 cases, 220 controls) and c) the Swedish validation dataset (88 cases, 1435 controls)

Six alleles were present at AA position 13 within our samples and in the reference panel (9,10); those encoding Ser and Arg showed low risk of FL, those encoding Tyr and Phe showed high risk, and those encoding His and Gly showed moderate risk (Figure 1, Table S2, and Figure S2). Conditioning upon the alleles at position 13 was sufficient to eliminate all the association signals within the vicinity in our study (chr6: 32–33 Mb; Table S3 and Figure S1). Dosage analyses (Table 1), HLA typing, (10) and sequencing analyses (13,14) confirmed high accuracy of the imputation (BEAGLE r^2 = 0.99) at this position and high genotype concordance across these platforms (~98%; see below).

To validate the associations, we evaluated genotypes at AA 13 in two independent data sets of 222 FL cases and 220 controls from San Francisco (6) (replication 1) and 88 FL cases and 1,435 controls from Sweden (replication 2) by using a combination of Sanger sequencing (replication 2 cases), (13,14) canonical HLA typing (replication 2 controls), (10) and GS-FLX sequencing (replication 1) (6). Replication 1 included non-Hispanic white FL cases and controls who were part of a population-based case-control non-Hodgkin lymphoma (NHL) study conducted in the San Francisco Bay Area from 2001 to 2006 (6). Controls were matched according to the number or frequency of cases in 5-year age groups, sex, and county of residence. We carried out HLA typing in these samples by using the Roche high-resolution HLA primer set and subsequently performed GS-FLX sequencing as previously described (6). Replication 2 included 19- to 83-year old (median age = 58 years) FL cases of Swedish descent and with available fresh frozen tumor material assembled in the Uppsala-Örebro region from 1970 to 2006 (15); these cases were typed by PCR amplification and Sanger sequencing (replication 2). Primer sequences for the amplification of HLA-DRB1exon 2 were obtained from previously published work (13,14). PCR products were purified with the use of AMPure XP (Agencourt) solid-phase-reversible-immobilization beads and run on a Bioanalyzer (Agilent) or 15% polyacrylamide gel for ensuring that excess primers were removed prior to sequencing. Sequence reads were visually inspected at the codon position encoding AAs 11 and 13, and genotypes were manually called (Figure S3) on the basis of expected codons at both positions and flanking ones according to sequences from the EMBL-EBI Immunogenetics HLA Database. For controls in replication 2, we used a second independent set of 1,435 Swedish control subject samples collected within the Epidemiological Investigation of Rheumatoid Arthritis (EIRA) study from 1995 to 2006. HLA typing was carried out in the control samples from the EIRA study with the use of sequence-specific primer PCR (DR low-resolution kit; Olerup SSP) as previously described (16). All discovery and replication studies were conducted in accordance with the ethical standards of the respective institutional review board of each institution, and informed consent was obtained from study participants.

We first tested the accuracy of the imputed genotypes by amplification and Sanger sequencing (13,14) of DRB1 exon 2 in a subset of our discovery samples. We observed high concordance between imputed genotypes and those inferred from the sequence chromatograms (Figure S3) (n = 92, 98.4% allelic, 97% genotypic concordance) and those inferred from HLA types (8) (Table S6; n = 596, 98.6% allelic, 97.3% genotypic concordance), confirming the high accuracy of the imputed genotypes and the reliability of the association results in our samples. To assess the accuracy of the AA genotypes inferred from canonical HLA typing data, we also Sanger sequenced a subset of 189 EIRA controls and found the alleles to be 97.6% concordant; hence, there are not likely to be any major biases because of the use of different genotyping platforms in replication 2.

We observed consistent direction of associations at all of the high- and low-risk residues at AA 13 across the three study populations (Figure 1, Table S2). Although the association did not reach statistical significance for some of the individual alleles (Table S2), it remained significant for the combined protective (encoding Ser or Arg) and risk (encoding Tyr or Phe) alleles in all the studies (Table 1). In the meta-analysis of all three populations, there was strong evidence of association between the AA 13 polymorphism and FL risk and no evidence of heterogeneity (multiallelic $p = 2.3 \times 10^{-15}$; $p_{Ser+Arg13} = 6.5 \times 10^{-14}$; $OR_{Ser+Arg13} = 0.60$; heterogeneity $I^2 = 0$) (Table

1). Across the three data sets, there was a 4.2-fold difference in risk (95% CI = 2.9–6.1) between subjects carrying two high-risk alleles and those carrying two low-risk alleles (Cochran-Mantel-Haenszel $p = 1.01 \times 10^{-14}$).

Conditional analyses followed by meta-analyses across all three data sets suggested that the AA 13 polymorphism might fully explain associations observed at SNPs rs2647012 ($p = 4.72 \times 10^{-11}$ before adjustment; $p = 0.804$ after adjustment) and rs10484561 ($p = 2.61 \times 10^{-11}$ before adjustment; $p = 0.356$ after adjustment) (Table S3). However, there was evidence of heterogeneity in the results of the conditional analysis across the three data sets ($Q < 0.05$; $I^2 > 70\%$). In particular, there was residual association at rs10484561 and rs2647012 after adjustment for the genetic effects at AA 13 in the replication 1 data set (Table S3). Conversely, conditioning on genotypes at rs2647012 (multiallelic $p = 1.68 \times 10^{-7}$), rs10484561 (multiallelic $p = 7.58 \times 10^{-7}$), or both SNPs (multiallelic $p = 0.005$) did not fully eliminate the association observed at AA 13, suggesting that the genotypes at AA 13 are well tagged, but not fully tagged, by these two SNPs.

Further haplotype analysis indicated that the minor C allele at rs10484561 partially tags the haplotypes carrying the high-risk allele encoding Phe (OR = 1.81; 95% CI = 1.51–2.18), whereas the minor T allele at rs2647012 partially tags the haplotypes carrying the low-risk alleles encoding Ser (OR = 0.69; 95% CI = 0.59–0.81) and Arg (OR = 0.67; 95% CI = 0.55–0.82) (Table S4). For this analysis, genotypes (six imputed alleles at AA 13 and two SNPs) were phased with the PHASE program (17) with default parameters. Phased haplotypes were then tested for association with FL with the use of logistic regression analysis in PLINK. Only haplotypes with minor allele frequencies > 1% in controls were analyzed. It is interesting to note that some rare (allele frequencies ~2%) haplotypes, such as Phe-C-A (Table S4), are not tagged by rs10484561 and rs2647012 but showed diverse associations across the three study populations, and this contributed to the large heterogeneity of the results of conditional association analysis across the three data sets above. Further studies with much larger sample sizes will be needed for evaluating the effects of these rare haplotypes. Taken together, our data suggest that the hexa-allelic AA 13 polymorphism might be the primary driver of the association within the region, whereas the diverse associations at multiple SNPs might be due to their differential tagging effects of various AA risk variants.

AA 13 is located in the middle of the peptide binding groove of the HLA-DR heterodimer molecule (Figure S2) and is thus well positioned to directly interact with bound peptides. This position, together with positions 70, 71, and 74, has been shown to play important roles in the binding-specificity profile of pocket 4, which is one of the most important pockets for antigen interaction and presentation by the HLA-DR molecule (18). Chemical properties of the side chain at this position might have a direct effect on antigen binding and recognition within the binding groove. The high-risk alleles at AA 13 both encode AAs with bulky hydrophobic side chains (Phe and Tyr), whereas the low-risk alleles encode AAs with polar or charged side chains (Figure S2). The roles of the moderate-risk alleles (encoding His and Gly) will need to be further explored in larger samples and/or functional assays. Although our results seem to suggest that variation at position 13 plays an important role in influencing FL risk, other nearby residues within the extended HLA haplotype might also influence peptide binding, and detailed functional work will be needed for proving the importance of this single position.

Residues at AA 13 are in high LD with nearby residues at AA 11, the second-best-associated coding variant (multiallelic $p = 3.6 \times 10^{-7}$) in discovery GWASs (Table S5). Given the close proximity and tight LD between the two variants, we were unable to distinguish the effects without additional functional investigation. Variants at HLA-DRB1 AAs 11 and 13 have previously been shown to associate with risk of rheumatoid arthritis (RA) (10). RA is an autoimmune disorder with a well-established correlation with NHL risk (mostly with the diffuse large B cell NHL subtype, but also with FL) (19). Although the patterns of association at AAs 11 and 13 differ between RA and FL, (10) the shared associations between the variants at HLA-DRB1 AAs 11–13 and FL and RA clearly suggest the involvement of common HLA-DR antigen-driven pathogenesis in the two different diseases. Further studies on large cohorts with detailed clinical information on both diseases are likely to reveal information on the shared or distinct etiological mechanisms of RA and lymphoma pathogenesis. In vitro biochemical studies will be needed for testing the effects of mutants with different combinations of AAs at these two positions on binding of a relevant autoantigen or tumor antigen to the beta chain of HLA-DR.

We have performed comprehensive imputation of SNPs, classical HLA alleles, and HLA coding variants by using a large reference panel of European subjects. Recent evaluation studies have demonstrated high accuracy of the imputations of these coding variants (20). Although our imputation confidence was high (BEAGLE $r^2 = 0.96$), there might have been minor inaccuracies owing to the nature of the imputation. It is possible that the GWAS samples might have been differentiated (in terms of population structure) relative to the reference samples used for the HLA imputations. We expect that this would have led to a consistent loss of imputation accuracy and therefore of power in both cases and controls and hence do not expect a differential bias in terms of imputation accuracy between cases and controls (a differential bias could inflate type 1 error). Nonetheless, we have confirmed the accuracy of genotypes at the most significant imputed variants by direct Sanger resequencing and additional HLA typing using sequence-specific primer PCR and have demonstrated high concordance. We have also replicated the results in two independent sample collections that were directly genotyped with experimental methods. We anticipate that there might be haplotypes or variants, especially rare ones with allele frequencies < 1%, that are poorly tagged and hence might have been missed in this analysis. In particular, there might be additional rare causal variants that could account for the residual association observed in replication 1. Our current sample size is, however, not large enough for a conclusive assessment of these possibilities, and the analysis of much larger reference panels and GWAS sample sizes will be needed for accurately evaluating the associations of these rare haplotypes.

In summary, through a comprehensive imputation and further experimental validation analysis of the HLA region (Figure S4), we have shown that the variants at a single hexa-allelic AA position (13) of HLA-DRB1 influence FL risk. This AA might account for most of the currently observed independent SNP signals previously identified through GWASs in populations of European descent. There was, however, an indication of residual associations in the replication

1 data set, and confirming these results therefore warrants further study. Nevertheless, here we show strong evidence that coding variants at a single AA position of HLA-DRB1 contribute to multiple association signals observed for FL. Our findings further suggest that this multiallelic AA polymorphism might explain a significant portion of the genetic associations observed for FL within the HLA class II region.

Acknowledgments

This work was funded by the Agency for Science and Technology and Research of Singapore, the Swedish Cancer Society (090659), the Swedish Research Council (K2008-64X-20737-01-2, 523-2006-972), and the Danish Medical Research Council (FSS 09-63424). The collection of blood samples in the founding case-control study was supported by the National Cancer Institute (NCI) (CA069269-01). Sample collection in the Epidemiological Investigation of Rheumatoid Arthritis study was supported by the Swedish Council for Working life (2008-0567), AFA Insurance, and the COMBINE project, funded by Vinnova. The San Francisco study was supported by the NCI, National Institutes of Health (CA122663, CA104682, CA45614, and CA89745). The Uppsala study was supported by the Lions Cancer Research Foundation, the Selanders Research Foundation, and the Research Foundation of the Department of Oncology of Uppsala University Hospital. P.I.W.d.B. is supported, in part, by a VIDI Award from the Netherlands Organization for Scientific Research (NWO-016.126.354).

References

1. Conde L., Halperin E., Akers N.K., Brown K.M., Smedby K.E., Rothman N., Nieters A., Slager S.L., Brooks-Wilson A., Agana L. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat. Genet*.2010;42:661–664.

2. Smedby K.E., Foo J.N., Skibola C.F., Darabi H., Conde L., Hjalgrim H., Kumar V., Chang E.T., Rothman N., Cerhan J.R. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. *PLoS Genet*. 2011;7:e1001378.

3. Skibola C.F., Bracci P.M., Halperin E., Conde L., Craig D.W., Agana L., Iyadurai K., Becker N., Brooks-Wilson A., Curry J.D. Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat. Genet*.2009;41:873–875.

4. Vijai J., Kirchhoff T., Schrader K.A., Brown J., Dutra-Clarke A.V., Manschreck C., Hansen N., Rau-Murthy R., Sarrel K., Przybylo J. Susceptibility loci associated with specific and shared subtypes of lymphoid malignancies. *PLoS Genet*. 2013;9:e1003220.

5. de Bakker P.I., McVean G., Sabeti P.C., Miretti M.M., Green T., Marchini J., Ke X., Monsuur A.J., Whittaker P., Delgado M. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet*. 2006;38:1166–1172.

6. 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. Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. *Tissue Antigens*. 2012;79:279–286.

7. Conde L., Bracci P.M., Richardson R., Montgomery S.B., Skibola C.F. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. *Am. J. Hum. Genet*.2013;92:126–130.

8. Abecasis G.R., Altshuler D., Auton A., Brooks L.D., Durbin R.M., Gibbs R.A., Hurles M.E., McVean G.A., 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature*.2010;467:1061–1073.

9. Pereyra F., Jia X., McLaren P.J., Telenti A., de Bakker P.I., Walker B.D., Ripke S., Brumme C.J., Pulit S.L., Carrington M., International HIV Controllers Study The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*. 2010;330:1551–1557.

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

11. Howie B., Fuchsberger C., Stephens M., Marchini J., Abecasis G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet*. 2012;44:955–959.

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

13. Sayer D.C., Whidborne R., De Santis D., Rozemuller E.H., Christiansen F.T., Tilanus M.G. A multicenter international evaluation of single-tube amplification protocols for sequencing-based typing of HLA-DRB1 and HLA-DRB3,4,5. *Tissue Antigens*. 2004;63:412–423.

14. Sayer D., Whidborne R., Brestovac B., Trimboli F., Witt C., Christiansen F. HLA-DRB1 DNA sequencing based typing: an approach suitable for high throughput typing including unrelated bone marrow registry donors. *Tissue Antigens*. 2001;57:46–54.

15. Thunberg U., Enblad G., Turesson I., Berglund M. Genetic variation in tumor necrosis factor and risk of diffuse large B-cell lymphoma and follicular lymphoma: differences between subgroups in Swedish patients. *Leuk. Lymphoma*.2010;51:1563–1566.

16. Lundström E., Källberg H., Smolnikova M., Ding B., Rönnelid J., Alfredsson L., Klareskog L., Padyukov L. Opposing effects of HLA-DRB1∗13 alleles on the risk of developing anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum*. 2009;60:924–930.

17. Stephens M., Smith N.J., Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum*. Genet. 2001;68:978–989.

18. Sturniolo T., Bono E., Ding J., Raddrizzani L., Tuereci O., Sahin U., Braxenthaler M., Gallazzi F., Protti M.P., Sinigaglia F., Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol*.1999;17:555–561.

19. Baecklund E., Backlin C., Iliadou A., Granath F., Ekbom A., Amini R.M., Feltelius N., Enblad G., Sundström C., Klareskog L. Characteristics of diffuse large B cell lymphomas in rheumatoid arthritis. *Arthritis Rheum*.2006;54:3774–3781.

20. Jia X., Han B., Onengut-Gumuscu S., Chen W.M., Concannon P.J., Rich S.S., Raychaudhuri S., de Bakker P.I. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS ONE*. 2013;8:e64683.

Supplemental Tables and Figures

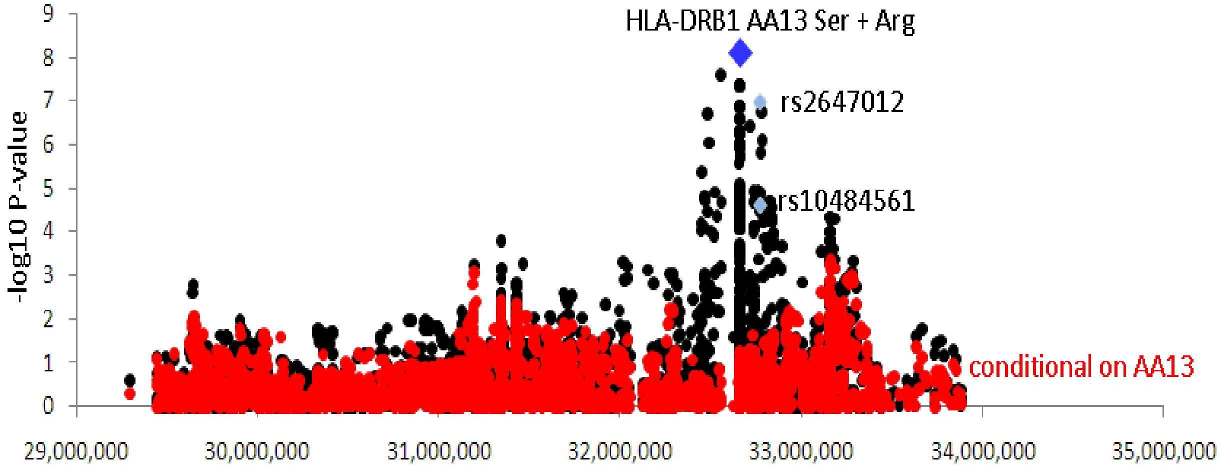


Figure S1. Region Association Plot before and after Conditioning on Alleles at Position 13. Regional association plot of bi-allelic logistic regression test P-values done on imputed amino acid alleles and SNPs surrounding rs2647012 and rs10484561 before (black) and after (red) conditioning on alleles at position 13 in SCALE GWAS samples

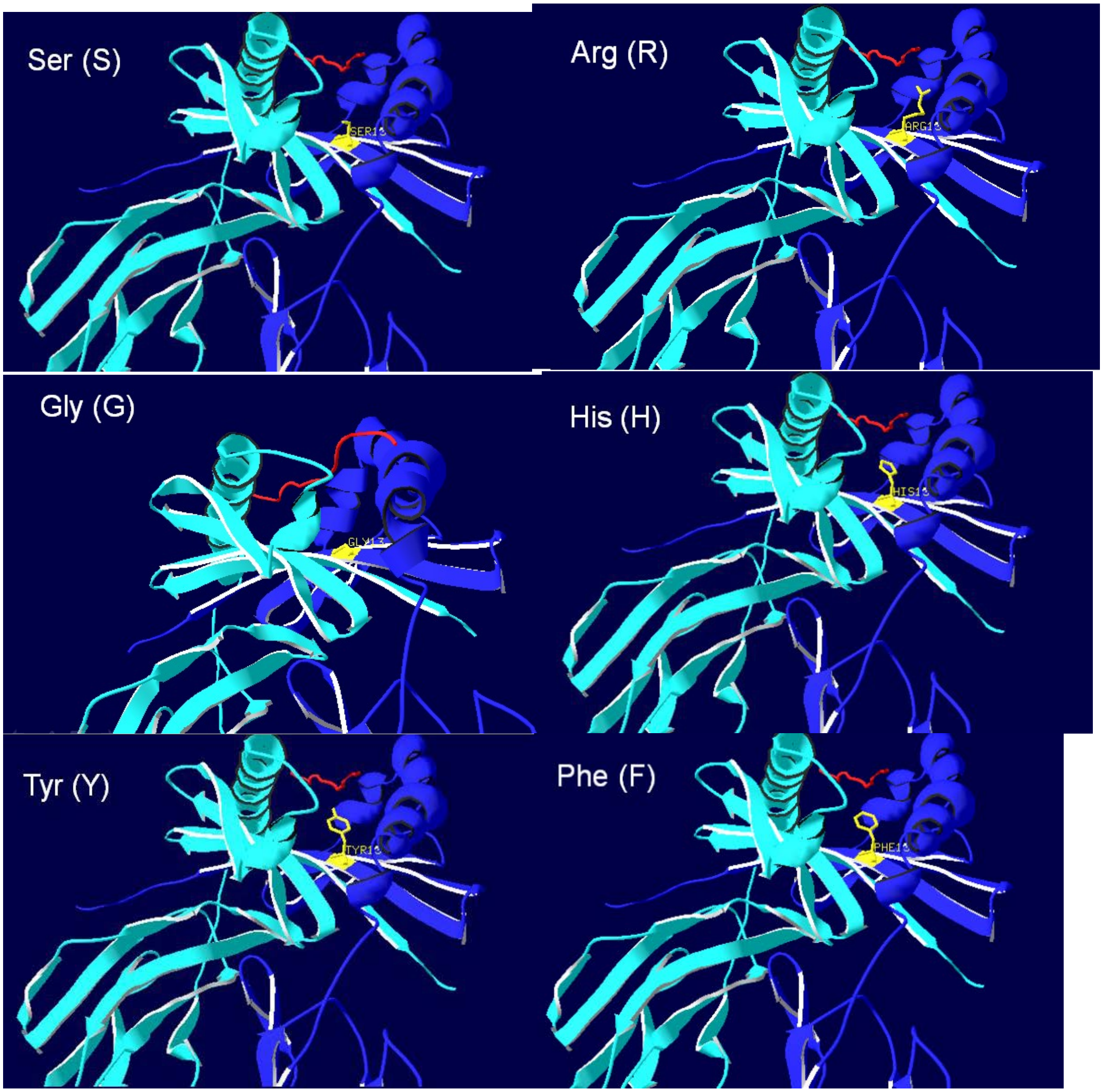


Figure S2. Six Possible Encoded Amino Acids at Position 13 in the HLA-DR Binding Groove. Amino acid 13 (yellow) in the binding groove of structure (PDB ID: 1AQD) of HLA-DR1 (DRA, DRB1 0101, Phe at position 13) protein (extracellular domain) complexed with endogenous peptide (red). Other amino acid alleles were mutated *in silico* using the DeepView/Swiss-PdbViewer program (http://spdbv.vital-it.ch/).

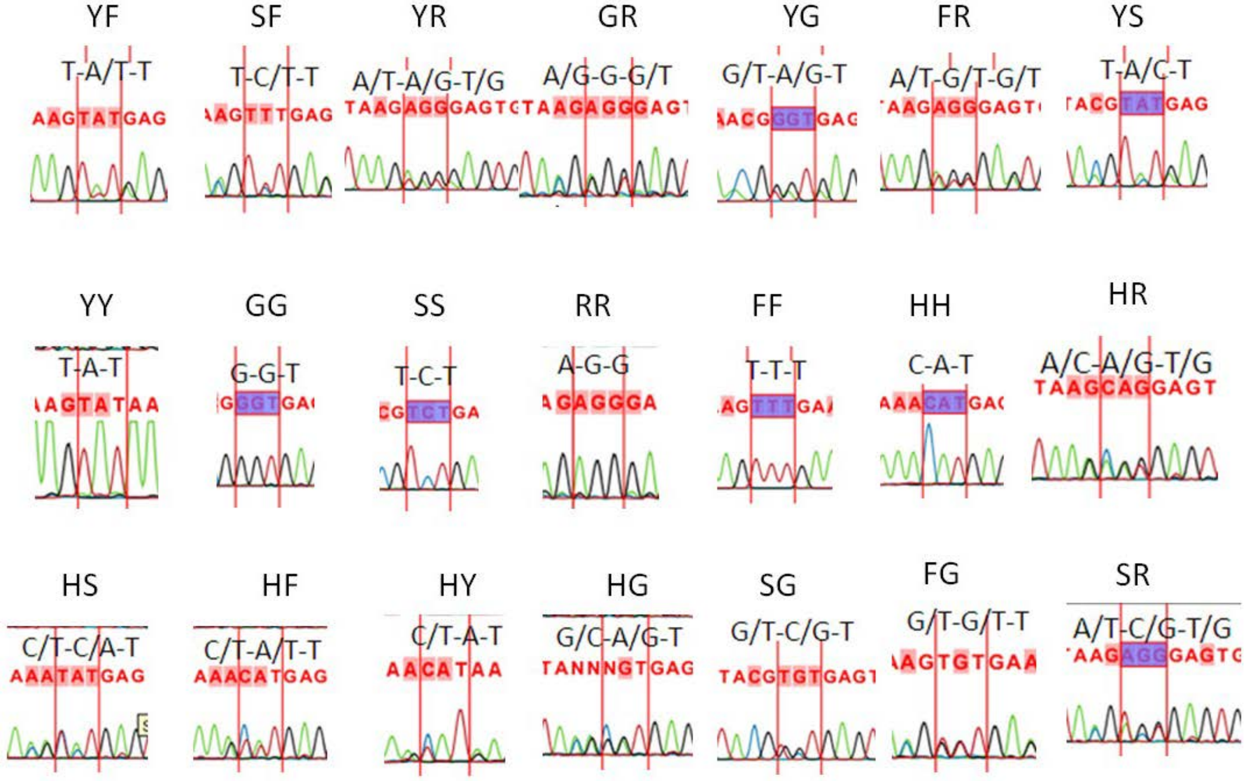


Figure S3. Genotyping of Amino Acid 13 Based on Sequence Chromatograms. All 21 genotypes (the two residues present in each individual, each coded as Y=Tyr, F=Phe, G=Gly, S=Ser, R=Arg, H=His) could be called and distinguished by visual inspection of the chromatograms.

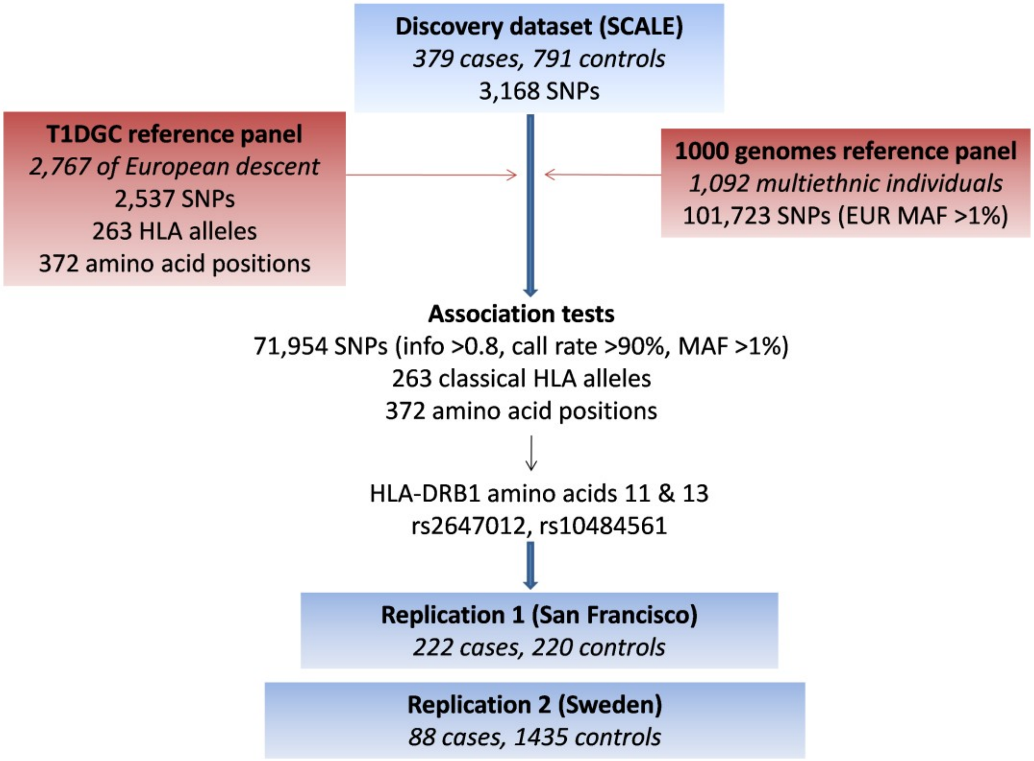

Discovery dataset (SCALE)
379 cases, 791 controls
3,168 SNPs
T1DGC reference panel
2,767 of European descent
2,537 SNPs
263 HLA alleles
372 amino acid positions
1000 genomes reference panel
1,092 multiethnic individuals
101,723 SNPs (EUR MAF >1%)
Association tests
71,954 SNPs (info >0.8, call rate >90%, MAF >1%)
263 classical HLA alleles
372 amino acid positions
HLA-DRB1 amino acids 11 & 13
rs2647012, rs10484561
Replication 1 (San Francisco)
222 cases, 220 controls
Replication 2 (Sweden)
88 cases, 1435 controls

Figure S4. Flowchart on Study Design and Data Acquisition

Flowchart on the imputation, association testing (showing filters for SNPs imputed by IMPUTEv2) and selection of variants for validation in additional samples. The datasets (and sample sizes) used as reference panels for imputation are shown in red boxes and the discovery or replication datasets used for association testing are shown in blue boxes.

Table S1. Conditional Analysis on FL-Associated SNPs in the HLA Class II Region

	Position				OR condition	P condition	r^2 / D'	r^2 / D'	r^2 / D'
SNP	(b37)	A1	OR (95% CI)	P	rs2647012 +	rs2647012 +	rs2647012	rs10484561	rs9378212
	chr6				rs10484561	rs10484561			
rs4530903	32581889	T	1.545 (1.183-2.018)	0.00139	0.509 (0.186-1.392)	0.188	0.085/0.979	0.92/0.991	0.154/0.989
rs9268853	32429643	C	1.302 (1.087-1.559)	0.00409	1.311 (0.995-1.727)	0.054	0.324/0.991	0.07/0.962	0.577/1
rs2621416	32741868	C	1.419 (1.163-1.732)	0.00058	1.056 (0.840-1.328)	0.642	0.207/0.843	0.147/0.641	0.187/0.611
rs2647046	32668336	A	0.617(0.510-0.747)	6.84E-07	NA	NA	0.991/0.996	0.09/0.934	0.558/0.986
rs9276490	32718681	A	0.695 (0.576-0.840)	0.000161	0.871 (0.703-1.080)	0.208	0.263/0.603	0.077/0.735	0.082/0.322
rs7453920	32730012	A	0.689 (0.570-0.833)	0.000116	0.854 (0.687-1.060)	0.152	0.27/0.605	0.075/0.738	0.079/0.319

Linkage disequilibrium and conditional analyses on rs2647012, rs10484561 and rs9378212 and the HLA SNPs reported in a recent FL GWAS by Vijai et al.[4] that were either genotyped (only rs7453920) or imputed in the SCALE discovery GWAS dataset.

Table S2. Association Results at Amino Acids 11 and 13 in the Discovery and Replication Data Sets

	Discovery GWAS (379 cases/791 controls)			Replication 1 (222 cases/220 controls)			Replication 2 (88 cases/1,435 controls)			Meta-analysis		
	Freq Cases/ Controls	OR (95% CI)	P	Freq Cases/ Controls	OR (95% CI)	P	Freq Cases/ Controls	OR (95% CI)	P	OR	P_{het}	I^2
Asp11	0.022/ 0.017	1.311 (0.716-2.402)	0.380	0.009/ 0.016	0.558 (0.161-1.935)	0.358	0.040/ 0.021	1.928 (0.871-4.270)	0.105	1.326	0.218	26.3
Gly11	0.128/ 0.097	1.358 (1.039-1.773)	0.025	0.162/ 0.121	1.388 (0.955-2.017)	0.085	0.091/ 0.077	1.180 (0.705-1.974)	0.528	1.338	$4.36x10^{-3}$	0
Leu11	0.168/ 0.106	1.724 (1.338-2.222)	$2.59x10^{-5}$	0.182/ 0.093	2.211 (1.465-3.337)	1.58×10^{-4}	0.227/ 0.115	2.149 (1.498-3.082)	$3.22x10^{-5}$	1.922	$4.82x10^{-12}$	0
Pro11	0.124/ 0.151	0.803 (0.624-1.032)	0.087	0.124/ 0.198	0.568 (0.391-0.826)	3.03×10^{-3}	0.114/ 0.165	0.654 (0.408-1.050)	0.079	0.710	$4.26x10^{-4}$	16.4
Ser11	0.330/ 0.426	0.666 (0.555-0.799)	$1.17x10^{-5}$	0.360/ 0.421	0.775 (0.590-1.018)	0.067	0.358/ 0.411	0.805 (0.588-1.100)	0.173	0.717	$1.67x10^{-6}$	0
Val11	0.228/ 0.204	1.150 (0.937-1.411)	0.181	0.158/ 0.152	1.042 (0.723-1.504)	0.824	0.165/ 0.211	0.740 (0.493-1.112)	0.147	1.050	0.558	44.3
Ser13	0.255/ 0.358	0.615 (0.506-0.747)	$9.30x10^{-7}$	0.315/ 0.375	0.763 (0.576-1.012)	0.061	0.290/ 0.350	0.761 (0.546-1.062)	0.108	0.677	$1.18x10^{-7}$	5.9
Arg13	0.124/ 0.151	0.803 (0.624-1.032)	0.087	0.124/ 0.198	0.568 (0.391-0.826)	3.03×10^{-3}	0.114/ 0.165	0.654 (0.408-1.050)	0.079	0.710	$4.26x10^{-4}$	16.4
Gly13	0.075/ 0.068	1.111 (0.795-1.553)	0.538	0.045/ 0.045	0.991 (0.526-1.865)	0.977	0.057/ 0.057	1.000 (0.521-1.922)	0.999	1.069	0.629	0
His13	0.220/ 0.195	1.161 (0.943-1.428)	0.159	0.146/ 0.143	1.026 (0.705-1.494)	0.892	0.165/ 0.203	0.774 (0.513-1.166)	0.221	1.060	0.489	34.0
Tyr13	0.128/ 0.097	1.358 (1.039-1.773)	0.025	0.162/ 0.121	1.388 (0.955-2.017)	0.085	0.091/ 0.077	1.180 (0.705-1.974)	0.528	1.338	$4.36x10^{-3}$	0
Phe13	0.198/ 0.132	1.649 (1.303-2.087)	$3.09x10^{-5}$	0.203/ 0.118	1.881 (1.296-2.731)	8.90×10^{-4}	0.284/ 0.148	2.160 (1.547-3.016)	$6.11x10^{-6}$	1.820	$6.71x10^{-12}$	0

P_{het}: Cochrane's Q test P-value; I^2: inconsistency

Table S3. Association Statistics at rs2647012, rs10484561, and rs9268839/rs9378212 before and after Adjustment for Alleles at Amino Acid Position 13

Sample collection Cases/Controls	**Before adjustment**				**After adjustment**			
rs2647012	**OR (95% CI)**			**P**	**OR (95% CI)**			**P**
Discovery 379 / 791	0.603 (0.500-0.727)			$1.10\text{x}10^{-7}$	0.851 (0.582-1.244)			0.405
Replication 1 222 / 220	0.551 (0.418-0.726)			$2.33\text{x}10^{-5}$	0.636 (0.432-0.936)			0.0216
Replication 2 88 /1435	0.831 (0.609-1.135)			0.245	1.771 (0.965-3.250)			0.065
Meta-analysis	**P**	**OR**	$\mathbf{P_{het}}$	$\mathbf{I^2}$	**P**	**OR**	$\mathbf{P_{het}}$	$\mathbf{I^2}$
Fixed effects	$4.72\text{x}10^{-11}$	0.628	0.125	51.99	0.207	0.853	0.021	74.29
Random effects	$3.33\text{x}10^{-5}$	0.639			0.804	0.938		

rs10484561	**OR (95% CI)**			**P**	**OR (95% CI)**			**P**
Discovery 379 / 791	1.686 (1.324-2.148)			$2.33\text{x}10^{-5}$	1.558 (0.864-2.808)			0.140
Replication 1 222 / 220	2.570 (1.701-3.883)			$7.41\text{x}10^{-6}$	6.006 (2.226-16.2)			$4.00\text{x}10^{-4}$
Replication 2 88 /1435	1.791 (1.237-2.595)			$2.05 \text{ x } 10^{-3}$	0.614 (0.311-1.212)			0.160
Meta-analysis	**P**	**OR**	$\mathbf{P_{het}}$	$\mathbf{I^2}$	**P**	**OR**	$\mathbf{P_{het}}$	$\mathbf{I^2}$
Fixed effects	$2.61\text{x}10^{-11}$	1.857	0.220	33.97	0.104	1.402	0.0009	85.75
Random effects	$9.14\text{x}10^{-8}$	1.900			0.356	1.698		

rs9268839/rs9378212	**OR (95% CI)**			**P**	**OR (95% CI)**			**P**
Discovery 379 / 791	1.637 (1.372-1.953)			$4.57\text{x}10^{-8}$	NA			NA

Given the high heterogeneity observed across the three datasets, both the fixed effects and random effects models were considered in the meta-analysis. P_{het}: Cochrane's Q test P-value; I^2: inconsistency.

Table S4. Haplotype Analysis of rs10484561 and rs2647012 with the Six Amino Acid Residues at Position 13 in a Meta-analysis across All Three Data Sets

Phased haplotype: AA13 – rs2647012 –	Discovery		Replication 1		Replication 2		Meta-analysis		
	MAF Cases/ Controls	OR (95% CI) P	MAF Cases/ Controls	OR (95% CI) P	MAF Cases/ Controls	OR (95% CI) P	OR (95% CI)	P	P_{het}/ I^2
Ser – C – A	0.050/ 0.068	(0.49-1.06) 0.098	0.119/ 0.111	(0.72-1.60 0.722	0.028/ 0.077	(0.14-0.86) 0.022	0.81 (0.62-1.05)	0.117	0.059/ 64.7
Ser – T – A	0.201/ 0.290	0.62 (0.50-0.76) 6.62E-06	0.194/ 0.264	0.69 (0.51-0.94) 0.018	0.253/ 0.266	0.94 (0.66-1.32) 0.706	0.69 (0.59-0.81)	2.69E-06	0.132/ 50.6

P_{het}: Cochrane's Q test P-value; I^2: inconsistency

Table S5. Haplotype Analysis of Alleles at Amino Acids 11 and 13 in the Meta-analysis across All Three Data Sets

Haplotype	Average Frequency in controls	P	OR (95% CI)	P_{het}	I^2	LD (r^2) in SCALE
Leu11-Phe13	0.105	$4.82x10^{-12}$	1.922 (1.597-2.313)	0.468	0	0.79
Asp11-Phe13	0.018	0.218	1.326 (0.847-2.078)	0.257	26.3	0.095
Gly11-Tyr13	0.098	0.004	1.338 (1.095-1.634)	0.870	0	1
Val11-His13	0.180	0.489	1.060 (0.898-1.252)	0.220	34.0	0.95
Ser11-Gly13	0.057	0.629	1.069 (0.816-1.400)	0.929	0	0.12
Pro11-Arg13	0.171	$4.26x10^{-4}$	0.710 (0.586-0.859)	0.302	16.4	1
Ser11-Ser13	0.361	$1.18x10^{-7}$	0.677 (0.586-0.782)	0.345	5.9	0.74

P_{het}: Cochrane's Q test P-value; I^2: inconsistency; LD: linkage disequilibrium between the encoded residues at positions 11 and 13.

Table S6. Amino Acid Residue Found at Positions 11 and 13 of Each Classical *HLA-DRB1* Allele

HLA type	11	13
DRB1*01:01	Leu	Phe
DRB1*01:02	Leu	Phe
DRB1*01:03	Leu	Phe
DRB1*01:04	Leu	Phe
DRB1*03:01	Ser	Ser
DRB1*04:01	Val	His
DRB1*04:02	Val	His
DRB1*04:03	Val	His
DRB1*04:04	Val	His
DRB1*04:05	Val	His
DRB1*04:06	Val	His
DRB1*04:07	Val	His
DRB1*04:08	Val	His
DRB1*07	Gly	Tyr
DRB1*08	Ser	Gly
DRB1*09	Asp	Phe
DRB1*10:01	Val	Phe
DRB1*11:01	Ser	Ser
DRB1*11:02	Ser	Ser
DRB1*11:03	Ser	Ser
DRB1*11:04	Ser	Ser
DRB1*12:01	Ser	Gly
DRB1*13:01	Ser	Ser
DRB1*13:02	Ser	Ser
DRB1*13:03	Ser	Ser
DRB1*13:04	Ser	Ser
DRB1*14:01	Ser	Ser
DRB1*14:02	Ser	Ser
DRB1*14:03	Ser	Ser
DRB1*14:04	Ser	Gly
DRB1*15:01	Pro	Arg
DRB1*16:01	Pro	Arg

Data from the EMBL-EBI Immunogenetics HLA Database (http://www.ebi.ac.uk/imgt/hla/)

Chapter 5.

Computational Prediction of Candidate Follicular Lymphoma Antigens

Nicholas K. Akers[1], Martyn T. Smith[1], Christine F. Skibola[2].

[1]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, CA 94720, USA

[2]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA

Abstract

The recent association of HLA class II alleles with altered risk of follicular lymphoma gives strength to the hypothesis that this is an antigen-driven disease. Although much research has been dedicated to characterizing the antigen-peptide binding preferences of HLA class II alleles, the ability of each allele to bind likely environmental antigens has never been characterized. Using the software NetMHCII, we predicted the binding of 23 HLA class II alleles to 11,545 protein sequences, selected as likely to be presented via the class II pathway. We found that significant variability in capacity to bind antigen peptides exists from allele-to-allele. The follicular lymphoma risk allele, HLA-DRB1*01:01, is consistently predicted to bind more peptides with higher affinity than any other class II allele. DQB1*05:01, in high linkage disequilibrium with DRB1*01:01 particularly in European populations, was routinely predicted to bind among the least number of peptides per protein. These findings may indicate that a more general mechanism of antigen-induced lymphomagenesis may influence follicular lymphoma risk, rather than specific antigen binding. Given the important role of HLA alleles in influencing susceptibility to a number of adverse health outcomes, these findings may be broadly applicable for researchers in the fields of infectious diseases, vaccine design, and immunology.

Introduction

Follicular lymphoma (FL) is a cancer of immune system B cells. Although the disease is usually indolent, approximately 30% of cases will transform to a more aggressive subtype of lymphoma with poor survival (1). It is estimated that 10,000-22,000 new cases of FL are diagnosed in the United States each year (2,3). Although treatment options have improved, FL for the most part remains an uncurable disease. Uncovering the causal factors for this disease would benefit not only the patients of FL, but possibly other subtypes of lymphoma as well.

The recent discovery that carrying certain HLA alleles has a profound impact on an individual's risk of developing FL indicates an antigen-based etiology for this disease. The haplotype HLA-DQB1*05:01-DRB1*01 has been demonstrated to nearly double FL risk (4–6) while HLA-DQB1*06-DRB1*15 and DPB1*03:01 are associated with decreased risk of FL (6,7). Recently, these findings have been confirmed in the largest study of FL to date, with indications that amino acid changes in the HLA peptide binding groove are key to FL risk (8). In previous studies, we have shown that carriers of the protective haplotype HLA-DQB1*06:02-DRB1*15:01 express higher levels of HLA-DQB1, suggesting that perhaps greater HLA class II expression was protective for FL (9,10). Other FL-associated alleles, however, were not associated with gene expression changes.

The HLA class II-FL association is not altogether surprising in light of the molecular features of this disease. FL cells exist in a germinal center state (11), dependent on interactions with germinal center T-cells for survival (12,13). These interactions are mediated by HLA class II proteins on the FL cell surface. Furthermore, FL B-cells have undergone somatic hypermutation and class switch recombination to enhance their B-cell receptor affinity for antigen (14,15). This implies that FL cells have successfully bound antigen peptide in the HLA class II receptor for T-cells in the past. These results prompted researchers to hypothesize that FL was an antigen-driven disease (11), though the more general hypothesis of antigen-driven lymphoproliferative disease is much older (16).

If FL is being driven by a particular antigen, discovering the nature of this antigen could have a major impact on our understanding of this disease. Unfortunately, classic epidemiology does not present a strong case for any one antigen. Studies of common infectious agents have not provided strong evidence of association with FL (17). Dietary antigens also have not been associated with FL (18,19). These failings may have to do with the long latency period of FL, the difficulty in characterizing exposure to dietary and infectious antigens, or it may be that FL is driven by self-antigen. Previous evidence does indicate that FL B-cells are prone to self-reactivity (20). The discovery of HLA alleles impacting FL risk gives us an alternative method for discovering antigens that may be driving FL.

Each HLA allele has a unique amino acid sequence and therefore a unique, three-dimensional peptide binding pocket. Each allele's peptide binding pocket has been shown to possess different 'preferences' for the biochemical aspects of the peptides it binds. Stronger affinity of allele for peptide is associated with increased likelihood of immune response (21). This feature is likely the reason for many of the HLA allele associations with infectious and autoimmune disease (22). One allele may be binding a disease-relevant antigen with particularly high or low affinity, changing the risk of disease for those who harbor the allele. By measuring the affinity of alleles for specific peptides *in-vitro*, a wealth of data was produced characterizing the specifics of allele peptide binding preferences (23,24).

Machine learning methods have generated formulas for predicting the strength with which an HLA allele will bind a peptide sequence. At least 34 unique software packages exist to predict peptide binding epitopes of alleles (http://mba.biocuckoo.org/links.php). Restricting to software capable of providing quantitative binding predictions for HLA-DR, -DQ, and -DP reduces this list significantly. Of the remaining, *NetMHCII* stands out for its ability to predict many alleles and for outperforming competitors in benchmark studies (25,26). Using four unique datasets, *NetMHCII* was consistently better able to predict peptide binding across HLA-DR alleles. *NetMHCII* uses *NN-align*, a neural network based, machine learning methodology, to predict class II binding affinities. This program will move progressively through a protein sequence and calculate predicted 50% inhibitory concentration (IC_{50}) for each peptide subsequence of the protein. A lower IC_{50} value is indicative of a more strongly bound peptide.

Despite the availability of HLA peptide prediction algorithms and of proteome sequences, there has been no prior study of proteome-wide HLA class II binding affinities that we are aware of. This more basic application of the concepts presented here may have implications for researchers of other HLA associated diseases, vaccine design, and population genetics. Similar research has been performed examining HLA class I alleles (27), where researchers found that large variation peptide binding capacity exists for class I alleles, and that using IC_{50} cutoffs is more appropriate than allele rank cutoffs.

Using state-of-the-art HLA-peptide binding prediction software, we sought to determine the ability of FL-associated HLA alleles to bind potentially antigenic proteins from human and pathogenic datasets. Our hypothesis was that FL-associated alleles present a fraction of antigens uniquely (either stronger or weaker) and that these antigens will represent strong candidates for further study.

Methods

All computing was performed in Red Hat Enterprise Linux Server release 6.5 (http://www.redhat.com/), on a Rocks version 6.1 cluster (http://www.rocksclusters.org/wordpress/?page_id=449). This cluster has 12 nodes, each with 8 Intel Xeon E5520 2.27GHz CPUs.

Antigenic Proteomes

Protein sequences were downloaded from the Uniprot database (http://www.uniprot.org/), selecting for sequences likely to be presented via the class II pathway. These queries resulted in 2,339 human, 6,287 bacterial, and 2,919 viral sequences (Table 1) which were downloaded directly in FASTA format. Restriction to only hand-reviewed sequences significantly reduced our dataset, however, non-reviewed sequences cause erratic patterns in our data. This is likely the result of sequence overlap and over-represented sequences (Supplementary Figure 1).

Human Antigens	**Results**
(taxonomy: 9606 AND reviewed:yes AND fragment:no**)** AND **(**go:"extracellular region [0005576]" OR go:"cell outer membrane [0009279]" OR location:"Secreted [SL-0243]"**)**	2,339
Bacterial Antigens	**Results**
(taxonomy:"Bacteria [2]" AND reviewed:yes AND fragment:no**)** AND **(**go:"extracellular region [0005576]" OR go:"cell outer membrane [0009279]" OR go:"cell wall [0005618]" OR location:"Secreted [SL-0243]"**)**	6,287
Viral Antigens	**Results**
(taxonomy:"Viruses [10239]" AND host:9606 AND fragment:no AND reviewed:yes**)** AND **(**keyword:"Virion [KW-0946]" OR keyword:capsid**)**	2,919
Total:	11,545

Table 1. Search terms used for gathering antigenic proteomes. The Uniprot database (http://www.uniprot.org/) was queried to reduce proteomic datasets to phyla-specific proteins likely to be encountered by antigen-presenting cells in humans. The search term "9606" is the organism code for humans. The right column lists the number of antigenic sequences found.

NetMHCII

NetMHCII version 2.2 was downloaded from (http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?netMHCII). Scripts were written in Bash (https://www.gnu.org/software/bash/bash.html) to test each downloaded protein sequence for each available HLA class II allele (Table 2). Bound cores below a 50% inhibitory concentration (IC_{50}) threshold of 2,000nM were kept for further analysis. Variable peptide length was tested by repeating this procedure with –l (peptide length) option set to 12, 15, 18, and 21 amino acids. These lengths were selected as a reasonable range of previously described, naturally bound peptides (28–30), though ~15 amino acids is generally agreed as the most common length of bound peptides.

Allele	NetMHCII Key	Allele	NetMHCII Key
DPB1*01:01	HLA-DPA10201-DPB10101	DRB1*01:01	HLA-DRB10101
DPB1*02:02	HLA-DPA10103-DPB10201	DRB1*03:01	HLA-DRB10301
DPB1*03:01	HLA-DPB10301-DPB10401	DRB1*04:01	HLA-DRB10401
DPB1*04:01	HLA-DPA101-DPB10401	DRB1*04:04	HLA-DRB10404
DPB1*04:02	HLA-DPA10301-DPB10402	DRB1*04:05	HLA-DRB10405
DPB1*05:01	HLA-DPA10201-DPB10501	DRB1*07:01	HLA-DRB10701
DQB1*02:01	HLA-DQA10501-DQB10201	DRB1*08:02	HLA-DRB10802
DQB1*03:01	HLA-DQA10501-DQB10301	DRB1*09:01	HLA-DRB10901
DQB1*03:02	HLA-DQA10301-DQB10302	DRB1*11:01	HLA-DRB11101
DQB1*04:02	HLA-DQA10401-DQB10402	DRB1*13:02	HLA-DRB11302
DQB1*05:01	HLA-DQA10101-DQB10501	DRB1*15:01	HLA-DRB11501
DQB1*06:02	HLA-DQA10102-DQB10602		

Table 2. HLA alleles tested. In this table are the 23 HLA class II proteins tested in this study, with the NetMHCII key used to call for this allele, and the simplified allele names used throughout this text.

Data Processing

Scripts were written in perl (www.perl.org) to aggregate and simplify *NetMHCII* output data. Output lines with redundant inner 'core' sequence but slightly different peptide sequence were reduced to just one peptide. The number of peptides bound under affinity IC_{50} = 50nM (strong binders), IC_{50} = 500nM (weak binders) and IC_{50} < 1100nM (predicted to bind) were counted, with affinity cutoffs as described in (26,31). The broad IC_{50} cutoff of 1100nM was used in an attempt to capture a high percentage of true binders, at the risk of also capturing some false-positives. Previous, similar analyses have found that appropriate cutoffs vary by allele, ranging as high as 944nM for class I alleles (27). These values were then adjusted by the length of the protein sequence, in order to normalize the data (Supplementary Figure 1). Finally, data were visualized and statistical comparisons were made in R. All scripts, as well as the analysis pipeline used in this project are available in the Supplementary Materials.

Results

In order to make meaningful comparisons between alleles, a biologically relevant affinity cutoff was needed. Comparing the profiles of all alleles tested as strong (IC_{50} <50nM), weak (IC_{50} < 500nM) and 'predicted to bind' (IC_{50} <1100nM), affinities demonstrated a major shift in the data as the affinity cutoff was increased (Figure 1). Restriction to just strongly bound peptides resulted in zero data points for many allele-protein combinations. Increasing our data to include all peptides predicted to bind to some extent (IC_{50} <1100nM) resulted in much more normally distributed data for each allele. The length of bound peptide used for predictions did not influence this. Figure 1 shows predictions for the bacterial protein dataset, however, this trend also was similar for human (Supplementary Figure 2) and viral (Supplementary Figure 3) proteins.

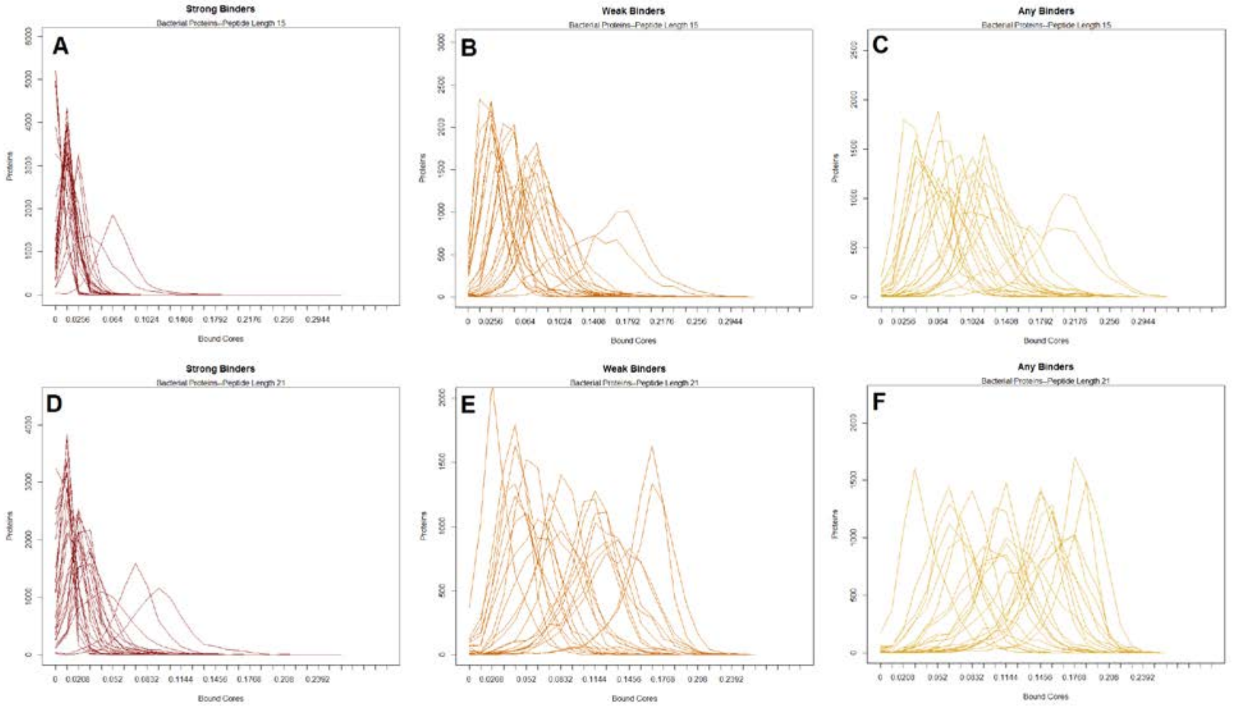


Figure 1. Strength of Peptide Binding. The number of peptides bound per protein is graphed in the histograms. The scale of the x-axis is the number of bound cores, adjusted for length of the protein. For example, 0.1 would indicate 10 peptides predicted to bind for a 100 amino-acid length protein. For a very low affinity (IC_{50} <50nM) cutoff value, the data are bound by zero on the left (A,D) indicating many alleles are not predicted to bind any peptides from a given protein IC_{50} =50. This problem was partly resolved by examining weakly (IC_{50} <500nM) bound peptide predictions (B,D), however for some alleles more than others. By raising the affinity threshold to IC_{50} < 1100nM, the histograms appeared normally distributed without being bound by zero, and remained within a biologically relevant range (C,F). These trends held when we examined peptide lengths = 15 (A,B,C) and 21 (D,E,F). The data shown here are for bacterial proteins only (n=6,287); however, human and viral proteomes gave similar results (Supplementary Figures 2,3).

The length of peptide used to make binding predictions was next examined. Box-plots of binding predictions for all alleles tested were created at 4 different peptide lengths (12, 15, 18, and 21 amino acids). These plots indicate that the length of peptide used to predict binding affinity does impact the predicted binding profile quite strongly for some alleles (Figure 2). In particular, several alleles that are predicted to bind very few 12- and 15-mer peptides were predicted to bind a substantially greater number of 18- and 21-mer peptides. Conversely, those alleles predicted to bind the greatest number of 12- and 15-mer peptides appeared to have a slightly decreased affinity for 18- and 21-mer peptides. Based on these observations, it is clear that binding affinity prediction can be strongly influenced by the length of peptide used. Figure

2 shows this trend in the bacterial protein dataset, however the trend was similar for human and viral proteins (Supplementary Figures 4,5).

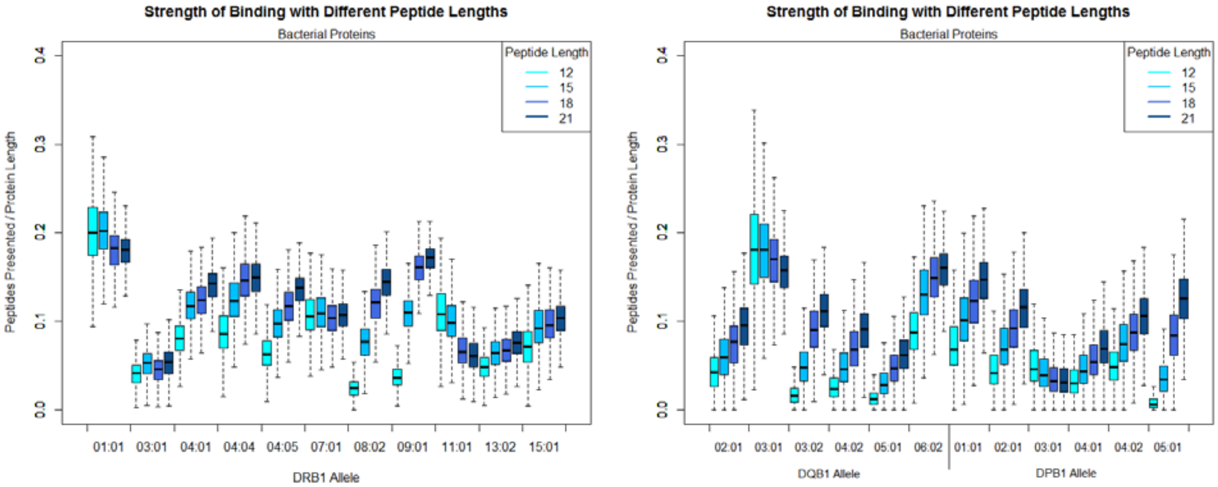


Figure 2. Peptide length influence on HLA class II binding. Each box represents peptide prediction data for one allele, at one peptide length, and 6,287 bacterial proteins. For each protein, the number of unique peptides which are predicted to be bound at IC_{50} <1100nM was divided by the length of the protein. Each box contains a line for the median, and is bound at the 1st and 3rd quartile (i.e., each box contains 50% of all data). The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range (the length of the box).

The source of proteins was examined to determine if phylum-specific trends in peptide binding prediction exist for HLA class II alleles. Human, bacterial, and viral proteomes as described in the methods are plotted in Figure 3 for each allele, at predicted peptide lengths of 15 and 21 amino acids. Slight differences indicate that most DRB1 and DPB1 alleles will bind viral protein peptides with more affinity than bacterial or human protein peptides. DQB1 alleles, conversely, bind bacterial proteins peptide with higher affinity than human and viral protein peptides. However, this trend is dwarfed by the effect that the allele itself has in prediction of binding affinity, regardless of the protein source.

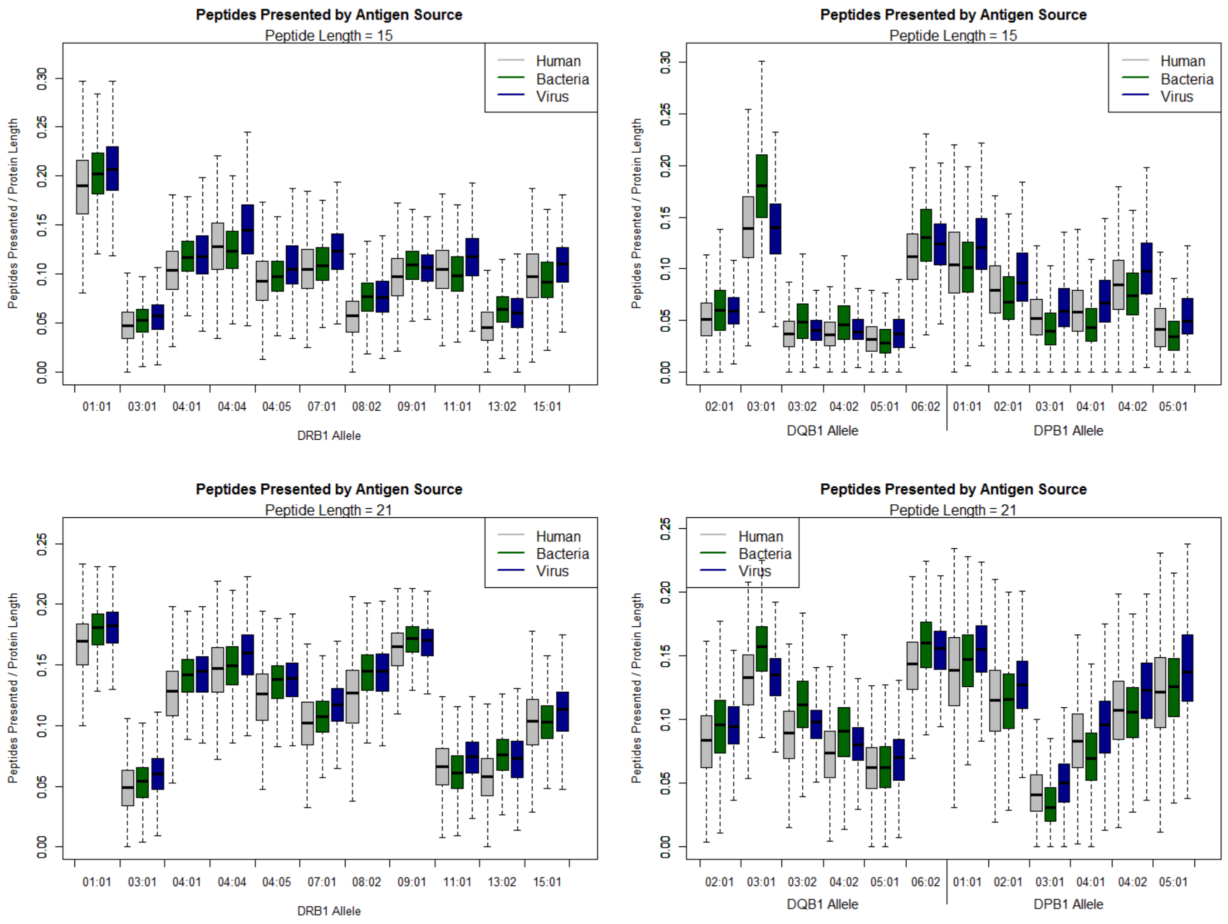


Figure 3. Influence of protein source on predicted binding affinity. Three protein datasets sourced from humans (n = 2,339), bacteria (n = 6,287), and viruses (n = 2,919) are shown in allele-specific binding prediction boxplots. For each protein, the adjusted number of peptides predicted to bind at IC_{50} <1100nM are shown.

Figures 2 and 3 indicated that there was substantial allele-to-allele variation in the number of peptides predicted to be bound for a given protein. Table 3 illustrated the potential for proteome-wide differences in peptide-binding capacity influencing FL risk. Alleles were ranked by their median predicted binding capacity for human, bacterial and viral proteomes at two different peptide lengths. The *DRB1*01:01* allele, associated with increased FL risk, was consistently predicted to bind the most peptides per protein, across datasets. Interestingly, the haplotype partner of this allele, *DQB1*05:01*, was consistently predicted to bind among the least peptides per protein.

	Peptide Length 15			Peptide Length 21			FL Risk Affect
Allele	Human	Bacteria	Virus	Human	Bacteria	Virus	(reference)
DRB1*01:01	1	1	1	1	1	1	Increased Risk (4,6,8)
DQB1*05:01	23	23	23	20	20	21	
DRB1*15:01	9	11	9	13	15	14	Decreased Risk (5–8)
DQB1*06:02	4	3	4	4	3	4	
DPB1*03:01	16	21	17	23	23	23	Decreased Risk (6,8)

Table 3. Allele ranks for peptide binding. For all alleles tested (n = 23) the median of predicted peptides bound was calculated for three phyla and 2 peptide lengths. The values shown here are the allele ranks by median value. DQB1-DRB1 haplotypes are boxed, with their reported effect on FL risk shown in the far-right column.

The degree to which each allele is predicted to present _unique_ peptides was also investigated. For each protein sequence, the number of peptide sequences presented by a given allele, but not presented by any other allele (below affinity IC$_{50}$ = 1100nM) was counted. Figure 4 summarizes this data, showing that most alleles will present 0-2 unique peptides per protein. However, there are a few notable outliers, with _DQB1*03:01_ presenting the greatest number of unique peptides per protein, by median.

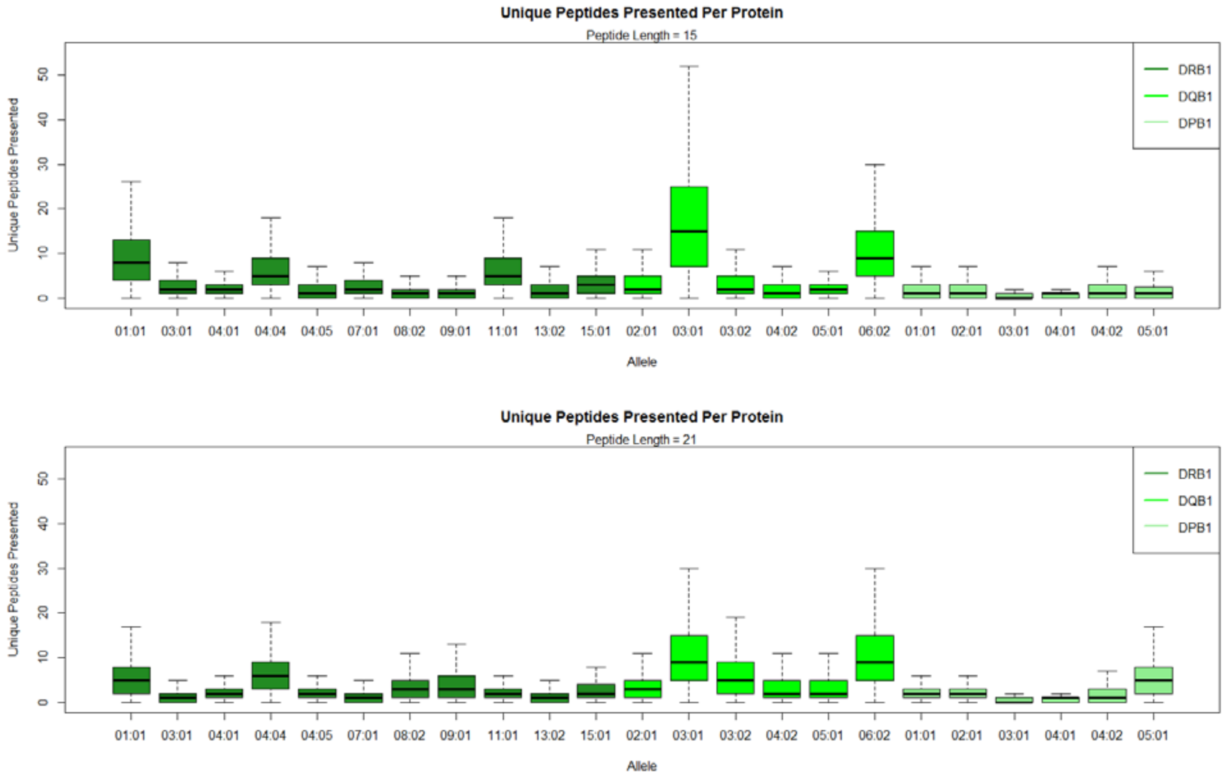


Figure 4. Unique peptides presented per protein. For all alleles and all proteins tested (n=11,545), the number of unique peptides presented is shown. A peptide presented by one allele, but not any other has the potential to invoke a unique immune response that may influence disease outcome. Although the median number of unique alleles presented per protein was less than 3, there are some outliers. These values coincide strongly with the binding affinity values in Figure 3 and Table 3.

Supplementary Tables 1 and 2 contain lists of proteins which may serve as candidate FL antigens. These tables were compiled by searching for proteins with exceptional peptide binding profiles for FL-associated alleles. This includes proteins with significantly greater or fewer peptides presented by FL-associated alleles (Supplementary Table 1), and proteins with greater or fewer unique peptides presented by FL-associated alleles (Supplementary Table 2). These differences were assessed using a Z-score, calculated across all HLA Class II alleles tested.

Discussion

We examined 23 HLA class II alleles for meaningful differences in the affinity for which they are predicted to bind human, viral, and bacterial proteins. The computational pipeline for this project created peptide binding affinity predictions for numerous sequences within each of 11,545 proteins, at 4 different peptide lengths, and for 23 different alleles. This resulted in ~2x10^9 unique peptide binding predictions. The data presented here indicate that there is a large amount of allele-to-allele variation in the predicted affinity of HLA class II alleles for natural antigen peptides. This finding was resilient even after correction for source of protein, length of predicted peptide, and the affinity cutoff used to count bound peptides. Given the impact that genetic variation at the HLA locus has on risk of FL, other diseases, and vaccine efficacy, these differences in peptide binding may be widely informative toward our understanding of the etiology of HLA-allele associated health outcomes.

A peptide binding affinity cutoff was necessary to enumerate total bound peptides per protein/allele combination. Figure 1 shows that once adjusted for protein length, "Strong Binders", described as having IC_{50} < 50nM, was too stringent of a cutoff. For many allele/protein combinations, no peptides are predicted to be bound below this threshold, resulting in histograms strongly bound on the left by zero (Figure 1A,D). The IC_{50} cutoff of 1100nM was selected for further analyses in an attempt to capture a high percentage of true binders for all alleles tested, even those with relatively weak binding profiles. This cutoff also created normally distributed data for all alleles.

HLA class II proteins are known to bind a wide range of peptide lengths (28,30), and this parameter appears to effect the predicted peptide binding capacity of alleles. Figure 2 indicates that the effect of increasing the length of peptide when making binding predictions has a dramatic effect for several alleles. This result, while intriguing, does not fully align with previously observed data. Chicz and colleagues (29) observed the naturally bound repertoire of peptides for several of these alleles, including *DRB1*03:01, DRB1*04:01, DRB1*07:01*, and *DRB1*15:01*. Each of these alleles was observed binding the most peptides from 15-18 amino acids in length. If *DRB1*04:01* bound longer peptides with a higher affinity, as predicted in Figure 2, this could be expected to have higher affinity for naturally bound peptides. This is not, however, the case in the Chicz et al. data, which shows *DRB1*04:01* binds relatively few peptides greater than 20 amino acids in length. There are two reasonable explanations for this disagreement. The relatively small number of long peptides in the data used to train NetMHCII may be causing a bias, and these predictions should be considered lower confidence. A second explanation is that peptide binding assays are not fully reflective of the peptide selection process within the cell, and shorter, lower affinity peptides are more available *in-vivo*. While we have more confidence in shorter, 15-mer peptide predictions, two peptide lengths, 15 and 21, have subsequently been shown in each analysis.

Researchers have hypothesized that certain HLA genes have co-evolved with infectious antigens, and that, for example, *HLA-B* has evolved specifically to bind viral antigens (32,33). Our data provide some evidence of gene-wide trends toward phylum-specificity. Five of six tested DQB1 alleles bind more peptides from bacterial proteins, by median, than human or viral proteins. Conversely, 8/11 DRB1 alleles and all DPB1 alleles bind more peptides from viral proteins than human or bacterial proteins (Figure 3). This trend holds when alleles are averaged across each gene (Supplementary Figure 6). Using the Students 2-tailed T-test, DQB1 alleles on average bound significantly more peptides from bacterial proteins than human or viral sources, while DRB1 and DPB1 bound significantly more peptides from viral proteins than human or bacterial sources (all p-values $<3.5*10^{-26}$—Supplementary Table 3). However, it should be noted that this phylum effect appears to be dwarfed by the allele-specific variability in overall peptide binding capacity.

The most striking trend observed in the data was the wide range across alleles in the number of peptides which are predicted to be bound. The potential impacts of this are numerous. Presenting more peptides per protein could have several consequences. Antigens will be more likely to be HLA-bound and presented to the immune system. However, an immunogenic peptide may go unrecognized when diluted among several other strongly bound peptides from the same protein. Autoimmunity may be more likely, as a diverse array of self-peptides may be more likely to escape negative selection. Additionally, a stronger, more diverse presentation of peptides is likely to impact T-cell selection in the thymus.

The FL risk allele *DRB1**01:01 is predicted to bind the most peptides per protein, regardless of source or peptide length, while its haplotypic partner, *DQB1**05:01, is predicted to bind among the least (Table 3). This may indicate that *DRB1**01:01 increases FL risk via enhanced presentation, or that *DQB1**05:01 increases FL risk through inability to present peptides. Genetic epidemiological studies will be needed to determine the potential causal FL risk allele; however, antigen binding capacity is an appealing hypothesis to explain the association of either allele. These results suggest that FL-associated alleles impact FL risk through general antigen interactions, rather than impacting the presentation of one or a few antigens. Similar to the risk alleles, the FL protective *DQB1*06:02* has high overall affinity for peptides, whereas its haplotypic partner, *DRB1*15:01,* has average affinity for peptides. It may be that FL is affected by which gene presents certain peptides, DQB1 or DRB1.

If over- or under-presentation of a FL-specific antigen impacts disease risk, we may see this antigen stand out in our data. By comparing the number of peptides bound to FL-associated alleles with summary data of all alleles, a z-score may be calculated. Similarly, we can find in our data those proteins for which FL associated alleles are predicted to present unique peptides, not presented by any other alleles (Figure 4). If FL progression is impacted by a unique allele/peptide interaction with the immune system, it is likely that this peptide is presented by only an FL associated allele. As an example, deamidated gliadin peptides which are uniquely immunogenic in celiac disease (34), are predicted to be bound by only celiac disease associated allele *DQB1*02:01* (data not shown). Candidate antigen proteins were compiled for FL, calculated based on the total number of peptides presented (Supplementary

Table 1) or the number of unique peptides presented (Supplementary Table 2), compared to all other alleles for each tested protein. These proteins may serve as interesting candidates for further functional testing. For example, *DRB1*01:01* is predicted to bind more peptides than average for the *IL1B*, *IL36A*, and *IL8* proteins (Uniprot IDs P01584, Q9UHA7, P10145: Supplementary Table 1). These interleukin proteins are active in regulating immune response, and are likely to be encountered by pre-FL B-cells. It may be that *DRB1*01:01* can elicit immunogenicity and growth signals by strongly presenting these proteins.

*DQB1*05:01* is predicted to only bind one peptide from the three B melanoma antigen proteins (Uniprot IDs Q13072, Q86Y28, and Q86Y27: Supplementary Table 1). This protein family is only expressed in the testis and certain melanoma, bladder and lung cancer tissues (35) where it is targeted by T-cells as a cancer antigen. Although expression of this protein family has not been detected in lymphoma tissues, *DQB1*05:01* may be similarly unable to present peptides from lymphoma cancer antigens.

In conclusion, we demonstrate here that there exists a wide range in the predicted binding capacity of HLA class II alleles. Interestingly, the antigen binding capacities of the FL-risk alleles DRB1*01:01 and DQB1*05:01 appear to fall at the opposite ends of the spectrum, indicating that there may a relationship between FL risk and the capacity of HLA class II proteins to bind antigen peptides. To the best of our knowledge, this is the first large-scale study comparing predicted binding capacity of HLA class II alleles. The results shown here have a broad application to numerous fields that rely on HLA class II allele specificity, including research in infectious diseases, vaccine design, and immunology.

Acknowledgements

This work was supported by National Institutes of Health grants CA154643 and CA104682 (CFS).

Bibliography

1. Bastion Y, Sebban C, Berger F, Felman P, Salles G, Dumontet C, et al. Incidence, predictive factors, and outcome of lymphoma transformation in follicular lymphoma patients. J Clin Oncol Off J Am Soc Clin Oncol. 1997 Apr;15(4):1587–94.

2. American Cancer Society. Cancer Facts & Figures 2013 [Internet]. Atlanta: American Cancer Society; 2013 [cited 2014 Feb 5]. Available from: http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-036845.pdf

3. Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD, Linet MS. Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. Blood. 2006 Jan 1;107(1):265–76.

4. Conde L, Halperin E, Akers NK, Brown KM, Smedby KE, Rothman N, et al. Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. Nat Genet. 2010 Aug;42(8):661–4.

5. Akers NK, Curry JD, Conde L, Bracci PM, Smith MT, Skibola CF. Association of HLA-DQB1 alleles with risk of follicular lymphoma. Leuk Lymphoma. 2011 Jan;52(1):53–8.

6. Skibola CF, Akers NK, Conde L, Ladner M, Hawbecker SK, Cohen F, et al. Multi-locus HLA class I and II allele and haplotype associations with follicular lymphoma. Tissue Antigens. 2012 Apr;79(4):279–86.

7. Smedby KE, Foo JN, Skibola CF, Darabi H, Conde L, Hjalgrim H, et al. GWAS of follicular lymphoma reveals allelic heterogeneity at 6p21.32 and suggests shared genetic susceptibility with diffuse large B-cell lymphoma. PLoS Genet. 2011 Apr;7(4):e1001378.

8. Christine F Skibola, Sonja I Berndt, Joseph Vijai, Lucia Conde, Zhaoming Wang, Meredith Yeager, et al. Genome-wide association study identifies new follicular lymphoma susceptibility loci. Nat Genet. Submitted;

9. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013 Jan 10;92(1):126–30.

10. Sillé FCM, Conde L, Zhang J, Akers NK, Sanchez S, Maltbaek J, et al. Follicular lymphoma-protective HLA class II variants correlate with increased HLA-DQB1 protein expression. Genes Immun. 2013 Dec 5;

11. Küppers R. Mechanisms of B-cell lymphoma pathogenesis. Nat Rev Cancer. 2005 Apr;5(4):251–62.

12. Umetsu DT, Esserman L, Donlon TA, DeKruyff RH, Levy R. Induction of proliferation of human follicular (B type) lymphoma cells by cognate interaction with CD4+ T cell clones. J Immunol Baltim Md 1950. 1990 Apr 1;144(7):2550–7.

13. Johnson PW, Watt SM, Betts DR, Davies D, Jordan S, Norton AJ, et al. Isolated follicular lymphoma cells are resistant to apoptosis and can be grown in vitro in the CD40/stromal cell system. Blood. 1993 Sep 15;82(6):1848–57.

14. Roulland S, Navarro J-M, Grenot P, Milili M, Agopian J, Montpellier B, et al. Follicular lymphoma-like B cells in healthy individuals: a novel intermediate step in early lymphomagenesis. J Exp Med. 2006 Oct 16;203(11):2425–31.

15. Bahler DW, Levy R. Clonal evolution of a follicular lymphoma: evidence for antigen selection. Proc Natl Acad Sci U S A. 1992 Aug 1;89(15):6770–4.

16. DAMESHEK W, SCHWARTZ RS. Leukemia and auto-immunization- some possible relationships. Blood. 1959 Oct;14:1151–8.

17. Engels EA. Infectious Agents as Causes of Non-Hodgkin Lymphoma. Cancer Epidemiol Biomarkers Prev. 2007 Mar 1;16(3):401–4.

18. Alexander DD, Mink PJ, Adami H-O, Chang ET, Cole P, Mandel JS, et al. The non-Hodgkin lymphomas: A review of the epidemiologic literature. Int J Cancer. 2007;120(S12):1–39.

19. Ekström-Smedby K. Epidemiology and etiology of non-Hodgkin lymphoma--a review. Acta Oncol Stockh Swed. 2006;45(3):258–71.

20. Dighiero G, Hart S, Lim A, Borche L, Levy R, Miller RA. Autoantibody activity of immunoglobulins isolated from B-cell follicular lymphomas. Blood. 1991 Aug 1;78(3):581–5.

21. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, et al. The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J Immunol Baltim Md 1950. 1994 Dec 15;153(12):5586–92.

22. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens. 2004 Dec;64(6):631–49.

23. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics. 2008;9 Suppl 12:S22.

24. El-Manzalawy Y, Dobbs D, Honavar V. On evaluating MHC-II binding peptide prediction methods. PloS One. 2008;3(9):e3268.

25. Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinformatics. 2009;10:296.

26. Nielsen M, Lundegaard C, Lund O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics. 2007;8:238.

27. Paul S, Weiskopf D, Angelo MA, Sidney J, Peters B, Sette A. HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity. J Immunol. 2013 Dec 15;191(12):5831–9.

28. Chicz RM, Urban RG, Lane WS, Gorga JC, Stern LJ, Vignali DA, et al. Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. Nature. 1992 Aug 27;358(6389):764–8.

29. Chicz RM. Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles. J Exp Med. 1993 Jul 1;178(1):27–47.

30. Steven G.E. Marsh, Peter Parham, Linda D. Barber. The HLA FactsBook. London, UK: Academic Press; 2000.

31. Wang P, Sidney J, Kim Y, Sette A, Lund O, Nielsen M, et al. Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics. 2010;11(1):568.

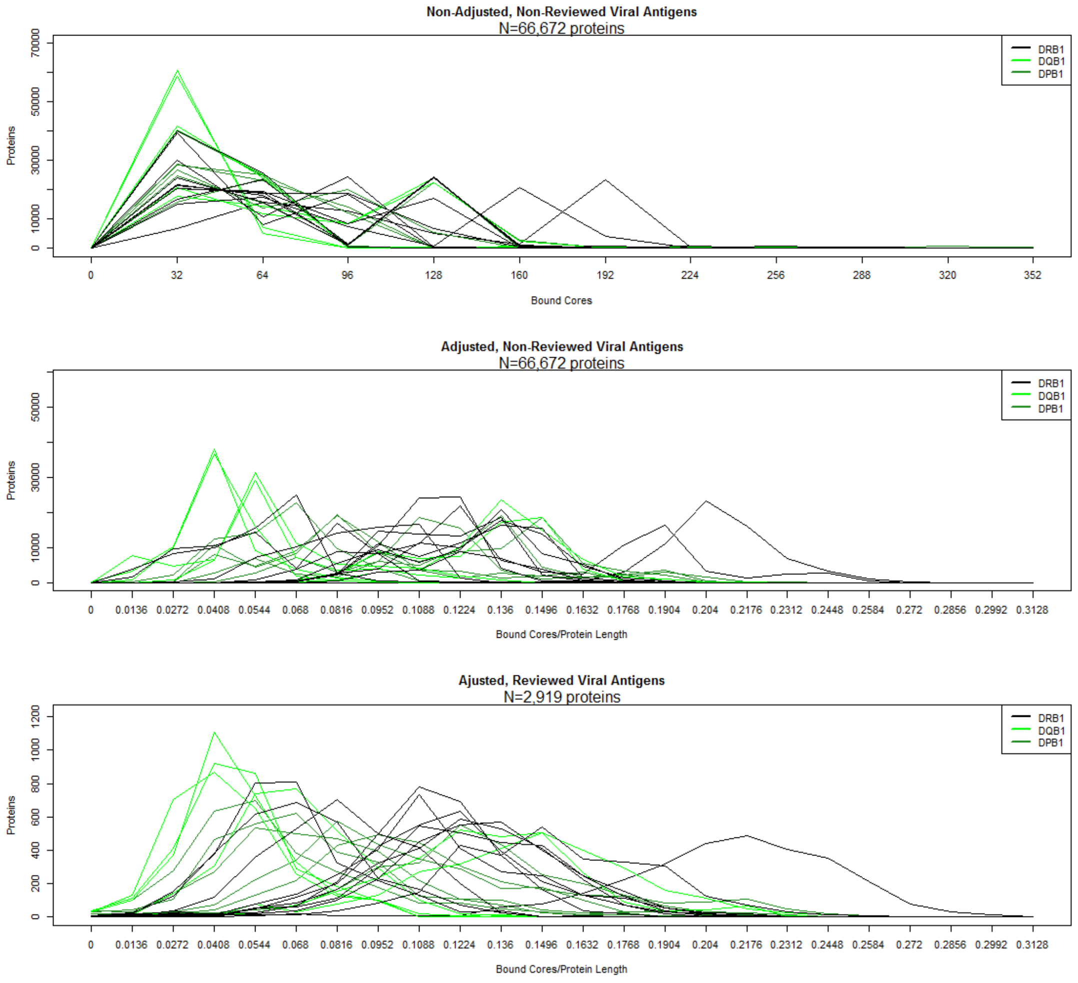
32. Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, et al. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature. 2004 Dec 9;432(7018):769–75.

33. Goulder PJR, Walker BD. HIV and HLA Class I: An Evolving Relationship. Immunity. 2012 Sep;37(3):426–40.

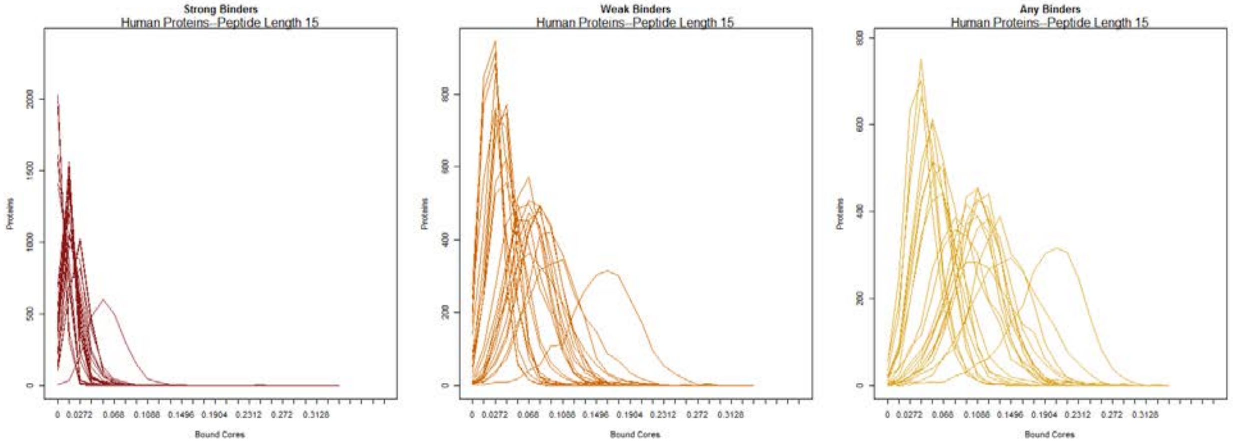
34. Shan L, Molberg Ø, Parrot I, Hausch F, Filiz F, Gray GM, et al. Structural basis for gluten intolerance in celiac sprue. Science. 2002 Sep 27;297(5590):2275–9.

35. Ruault M, van der Bruggen P, Brun M-E, Boyle S, Roizès G, De Sario A. New BAGE (B melanoma antigen) genes mapping to the juxtacentromeric regions of human chromosomes 13 and 21 have a cancer/testis expression profile. Eur J Hum Genet EJHG. 2002 Dec;10(12):833–40.

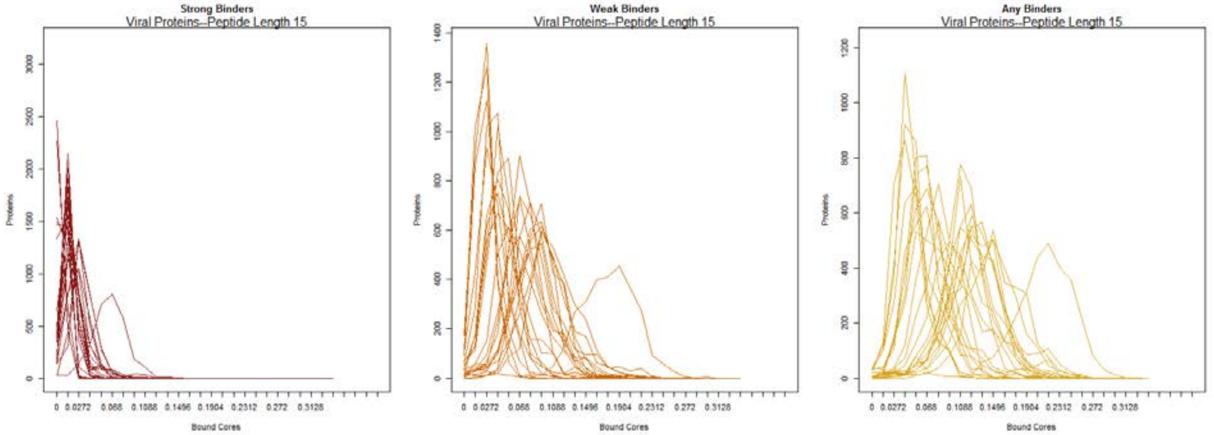
Supplementary Tables and Figures



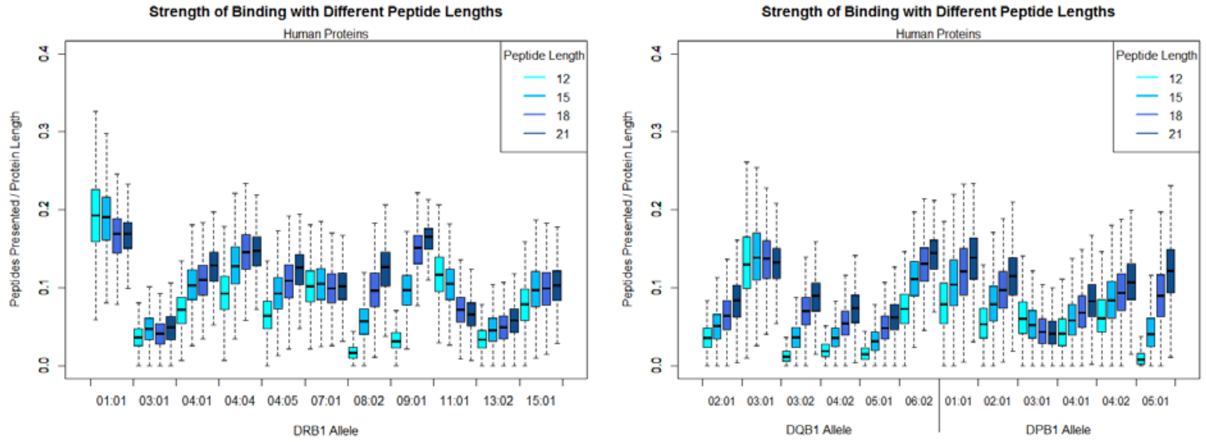
Supplementary Figure 1. **Normalization of Peptide Binding Data.** Observing total peptide binding predictions from non-reviewed Uniprot viral sequences (top panel), an erratic distribution pattern emerged (top panel). This was greatly aided by normalization of bound cores to the total length of each protein (middle panel). The total number of cores able to be bound by a single allele is greatly influenced by protein length, and adjusting for this variable makes allele-specific differences in binding capacity far more clear. Limiting our dataset to only Uniprot hand-reviewed sequences also aided in normalization of this data (bottom panel). This may be due to highly similar sequences uploaded to the Uniprot database multiple times, causing local spikes in the data. As a result of this observation, we decided to focus our study on only hand-reviewed sequences in the database.



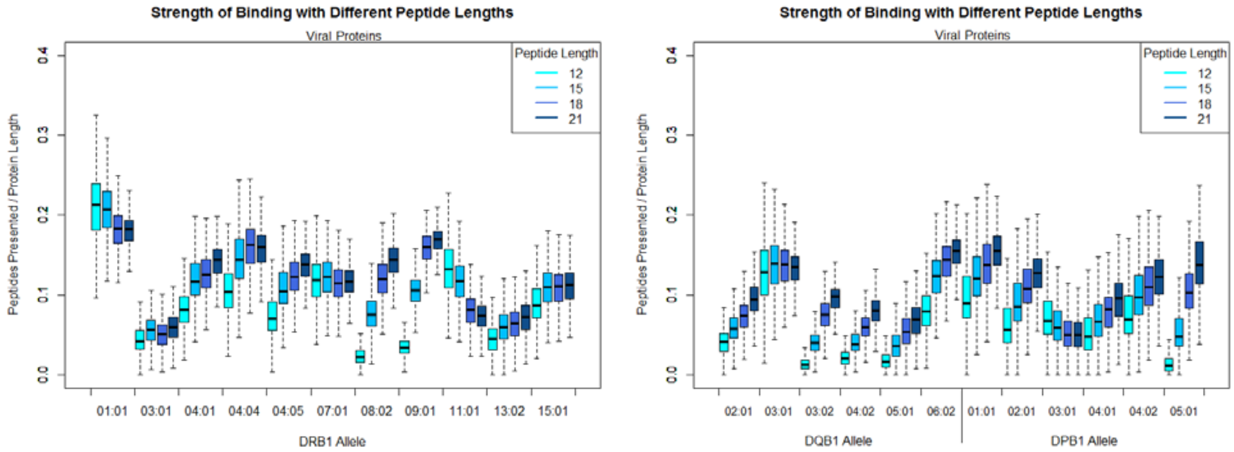
Supplementary Figure 2. The number of peptides bound per protein is graphed in histograms. The scale of the x-axis is number of bound cores, adjusted for length of the protein. For example 0.1 would indicate 10 peptides predicted to bind for a 100 amino-acid length protein. For too low of an affinity (IC_{50} <50nM) cutoff value, the data is bound by zero on the left (A,D) indicating many alleles are not predicted to bind any peptides from a give protein below IC_{50} =50. This problem is mostly fixed by examining weakly (IC_{50} <500nM) bound peptide predictions (B,D), however for some alleles more than others. Raising the affinity threshold to IC_{50} < 1100nM left us with histograms that appeared normally distributed without being bound by zero, while remaining within a biologically relevant range (C,F).



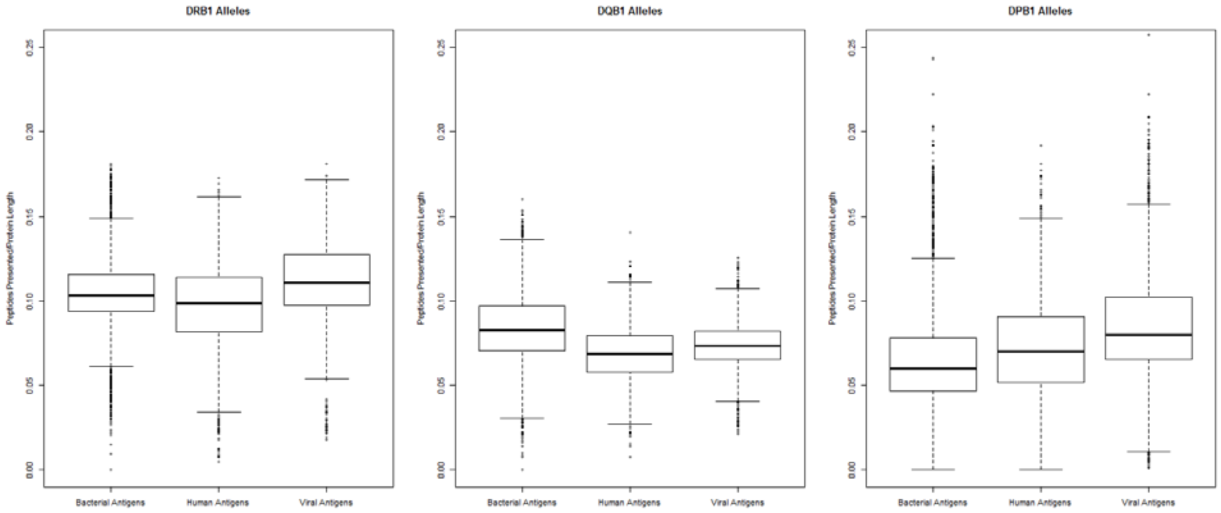
Supplementary Figure 3. Similar to Supplementary Figure 2, this figure shows decreasingly zero-bound, normally distributed data as the affinity cutoff for categorizing bound cores is increased, for viral protein datasets.



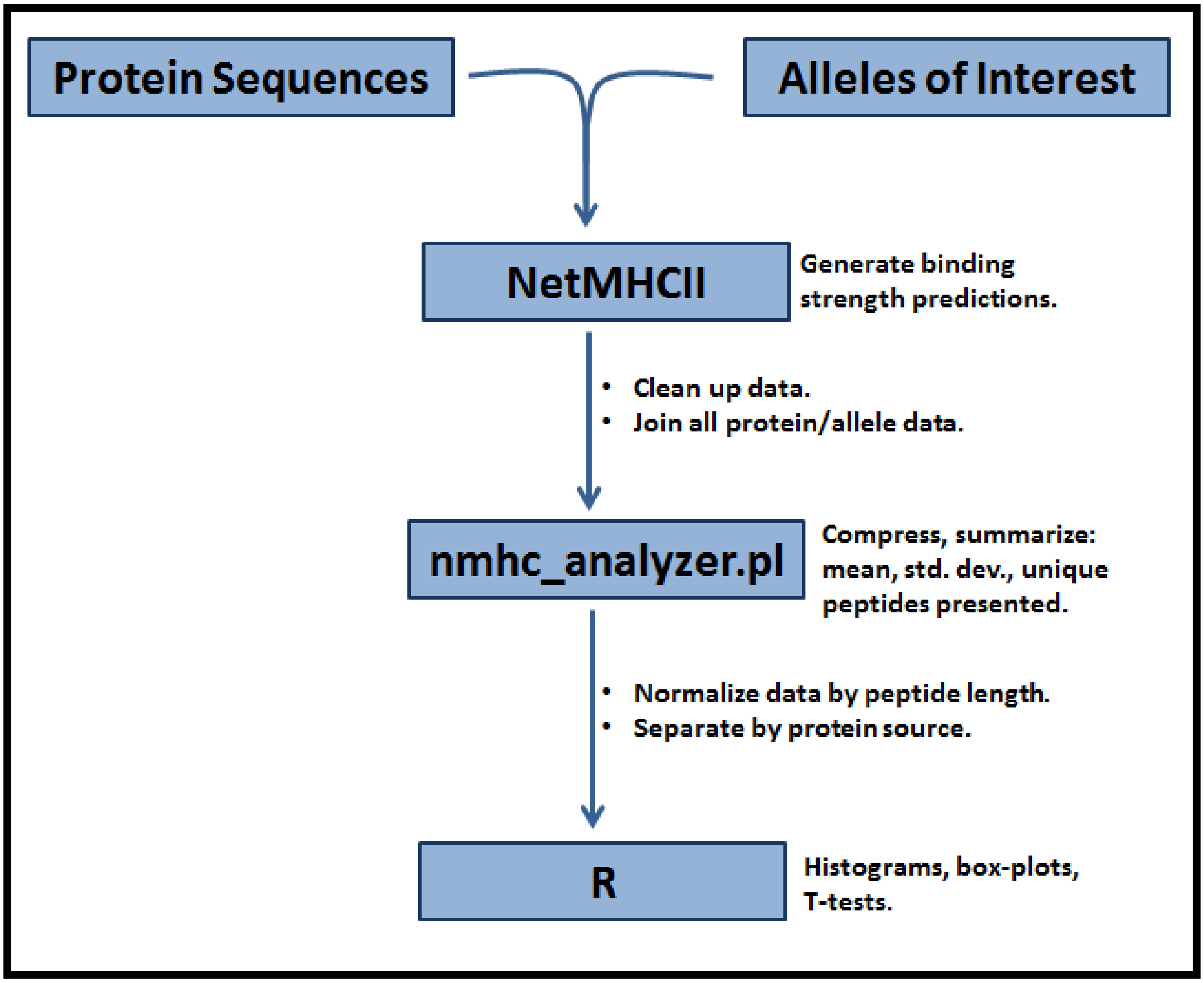
Supplementary Figure 4. Peptide length influence on HLA class II binding human proteins. Each box represents peptide prediction data for one allele, at one peptide length, and 2,339 human proteins. For each protein, the number of unique peptides which are predicted to be bound at IC_{50} <1100nM was divided by the length of the protein. Each box contains a line for the median, and is bound at the 1^{st} and 3^{rd} quartile (ie. each box contains 50% of all data). The whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range (the length of the box).



Supplementary Figure 5. Peptide length influence on HLA class II binding proteins. Similar to Supplementary Figure 4, this figure shows that similar to human and bacterial protein datasets, peptide length influences predicted binding affinity in viral proteins.



Supplementary Figure 6. Gene-wide differences in peptide presentation. Shown are box-plots of the average number of peptides presented for each of 2,339 human, 6,287 bacterial and 2,919 viral proteins across three HLA class II genes. The number of peptides presented per protein has been adjusted by the total length of each protein. This figure includes data from 11 HLA-DRB1, 6 HLA-DQB1, and 6 HLA-DPB1 alleles. The peptide length used in these predictions was 15 amino acids.


Protein Sequences
Alleles of Interest
NetMHCII
Generate binding strength predictions.
• Clean up data.
• Join all protein/allele data.
nmhc_analyzer.pl
Compress, summarize: mean, std. dev., unique peptides presented.
• Normalize data by peptide length.
• Separate by protein source.
R
Histograms, box-plots, T-tests.

Supplementary Figure 7. Analysis Pipeline. This figure demonstrates the data pipeline used to process 11,545 protein sequences and 23 HLA Class II alleles binding predictions from NetMHCII. A combination of allele and protein sequence is fed into NetMHCII, generating binding predictions. This output is cleaned up, and joined with the data from all other allele/protein combinations. The nmhc_analyzer.pl script is used to compress this data and summarize each protein using a number of metrics. This script also searches for candidate antigen proteins that were presented exceptionally by FL associated alleles. Data was then output, where it was normalized by peptide length, separated by protein source, and fed into R. R was used to create histograms, box-plots, and T-tests presented in the text.

DQB1* 05:01

Rank	Bound	Avg.	S.D	Z	ID	Protein	Species
1	1	3.91	1.31	-2.22	Q13072	B melanoma antigen 1	Homo sapiens
2	0	9.65	4.37	-2.21	P62158	Calmodulin	Homo sapiens
3	3	33.04	13.68	-2.20	P08758	Annexin A5	Homo sapiens
4	0	3.48	1.59	-2.19	P86395	Bacteriocin SRCAM 37	Paenibacillus polymyxa
5	1	3.65	1.23	-2.16	Q86Y28	B melanoma antigen 4	Homo sapiens
6	2	7.04	2.36	-2.13	P0C046	Virulence factor EsxA	Staphylococcus aureus
7	2	7.04	2.36	-2.13	Q5HJ91	Virulence factor EsxA	Staphylococcus aureus
8	2	7.04	2.36	-2.13	Q6GCJ0	Virulence factor EsxA	Staphylococcus aureus
9	2	7.04	2.36	-2.13	Q6GK29	Virulence factor EsxA	Staphylococcus aureus
10	2	7.04	2.36	-2.13	Q7A1V4	Virulence factor EsxA	Staphylococcus aureus
11	2	7.04	2.36	-2.13	Q7A7S4	Virulence factor EsxA	Staphylococcus aureus
12	2	7.04	2.36	-2.13	Q99WU4	Virulence factor EsxA	Staphylococcus aureus
13	2	13.2	5.31	-2.10	Q9K1F0	Outer membrane protein assembly factor BamE	Neisseria meningitidis serogroup B
14	1	9.83	4.25	-2.08	P09312	Envelope protein US9	Varicella-zoster virus
15	1	9.83	4.25	-2.08	Q77NN6	Envelope protein US9	Varicella-zoster virus
16	4	21.9	8.68	-2.06	P20338	Ras-related protein Rab-4A	Homo sapiens
17	8	43.7	17.44	-2.05	P50453	Serpin B9	Homo sapiens
18	1	3.78	1.38	-2.02	Q86Y27	B melanoma antigen 5	Homo sapiens
19	0	3.09	1.53	-2.01	O33690	Competence-stimulating peptide	Streptococcus oralis
20	3	9.13	3.05	-2.01	P10599	Thioredoxin	Homo sapiens
21	1	22.09	10.51	-2.01	P52823	Stanniocalcin-1	Homo sapiens

DRB1* 01:01

Rank	Bound	Avg.	S.D	Z	ID	Protein	Species
1	46	14.83	8.11	3.85	P05856	Protein Nef	HIV type 1 group M subtype B
2	40	13.09	7.24	3.72	P69440	Adenylate kinase	Mycobacterium tuberculosis
3	23	6.00	4.82	3.52	Q30KP8	Beta-defensin 136	Homo sapiens
4	221	85.26	38.70	3.51	Q9ULD4	Bromodomain and PHD finger-containing protein 3	Homo sapiens
5	40	13.61	7.58	3.48	P20886	Protein Nef	HIV type 1 group M subtype B
6	79	34.35	12.90	3.46	O34656	Spore coat protein I	Bacillus subtilis
7	42	13.26	8.42	3.41	Q9UHA7	Interleukin-36 alpha	Homo sapiens
8	53	21.52	9.24	3.41	O53894	Response regulator MprA	Mycobacterium tuberculosis
9	47	15.30	9.35	3.39	B1JEP9	Outer-membrane lipoprotein LolB	Pseudomonas putida
10	39	15.52	6.94	3.38	Q9IDV1	Protein Nef	HIV type 1 group N
11	39	13.09	7.67	3.38	O89293	Protein Nef	HIV type 1 group M subtype F1

12	22	8.30	4.08	3.35	P10145	Interleukin-8	Homo sapiens
13	41	14.52	7.90	3.35	P19546	Protein Nef	HIV type 1 group M subtype B
14	38	14.87	6.92	3.34	Q9QBY9	Protein Nef	HIV type 1 group M subtype F2
15	47	15.30	9.48	3.34	B0KND8	Outer-membrane lipoprotein LolB	Pseudomonas putida
16	40	14.91	7.51	3.34	P19545	Protein Nef	HIV type 1 group M subtype B
17	153	58.43	28.35	3.34	Q13421	Mesothelin	Homo sapiens
18	36	13.09	6.88	3.33	P36921	Cell wall enzyme EbsB	Enterococcus faecalis
19	48	16.48	9.47	3.33	Q1IEY4	Outer-membrane lipoprotein LolB	Pseudomonas entomophila
20	36	13.48	6.79	3.31	Q9Q713	Protein Nef	HIV type 1 group M subtype H
21	81	33.30	14.41	3.31	Q14515	SPARC-like protein 1	Homo sapiens
22	37	13.83	7.03	3.30	P04602	Protein Nef	HIV type 1 group M subtype D
23	43	15.74	8.28	3.29	P03407	Protein Nef	HIV type 1 group M subtype B
24	49	16.00	10.03	3.29	A5VYG2	Outer-membrane lipoprotein LolB	Pseudomonas putida
25	44	15.61	8.63	3.29	Q70627	Protein Nef	HIV type 1 group M subtype B
26	48	15.91	9.78	3.28	Q88PX4	Outer-membrane lipoprotein LolB	Pseudomonas putida
27	39	16.52	6.85	3.28	D0ZWR8	Salmonella pathogenicity island 2 protein C	Salmonella typhimurium
28	39	16.52	6.85	3.28	P0CZ04	Salmonella pathogenicity island 2 protein C	Salmonella typhimurium
29	36	13.17	6.97	3.27	P05858	Protein Nef	HIV type 1 group M subtype B
30	37	13.17	7.28	3.27	Q9WC61	Protein Nef	HIV type 1 group M subtype J
31	38	13.22	7.58	3.27	Q9QSQ6	Protein Nef	HIV type 1 group M subtype F1
32	46	16.09	9.15	3.27	Q888C4	Outer-membrane lipoprotein LolB	Pseudomonas syringae pv. tomato
33	54	23.22	9.42	3.27	P01584	Interleukin-1 beta	Homo sapiens
34	112	43.74	20.91	3.27	A5VZC8	Membrane-bound lytic murein transglycosylase F	Pseudomonas putida
35	28	8.74	5.91	3.26	P01599	Ig kappa chain V-I region Gal	Homo sapiens
36	85	35.43	15.21	3.26	P16779	Protein UL38	Human cytomegalovirus
37	37	14.17	7.01	3.26	P04604	Protein Nef	HIV type 1 group M subtype D
38	51	17.78	10.22	3.25	P0A672	Iron-dependent repressor IdeR	Mycobacterium tuberculosis
39	117	47.65	21.34	3.25	B1KLC4	Membrane-bound lytic murein transglycosylase F	Shewanella woodyi

40	112	43.17	21.19	3.25	Q88P17	Membrane-bound lytic murein transglycosylase F	Pseudomonas putida
41	78	29.04	15.09	3.25	P00813	Adenosine deaminase	Homo sapiens
42	110	43.04	20.65	3.24	B1JDH3	Membrane-bound lytic murein transglycosylase F	Pseudomonas putida
43	45	15.48	9.11	3.24	P04324	Protein Nef	HIV type 1 group M subtype B
44	28	11.74	5.03	3.23	Q5EBL8	PDZ domain-containing protein 11	Homo sapiens
45	43	15.30	8.57	3.23	C3KDC3	Outer-membrane lipoprotein LolB	Pseudomonas fluorescens
46	62	28.04	10.52	3.23	Q9PCR3	Monofunctional biosynthetic peptidoglycan transglycosylase	Xylella fastidiosa
47	48	17.35	9.52	3.22	A4XR58	Outer-membrane lipoprotein LolB	Pseudomonas mendocina
48	77	31.52	14.12	3.22	P44567	Lipid A biosynthesis	Haemophilus influenzae
49	46	16.39	9.21	3.22	Q48MV7	Outer-membrane lipoprotein LolB	Pseudomonas syringae pv. phaseolicola
50	34	10.61	7.29	3.21	Q8WXF3	Relaxin-3	Homo sapiens

DRB1* 15:01 Rank	**Bound**	**Avg.**	**S.D**	**Z**	**ID**	**Protein**	**Species**
1	33	10.96	7.27	3.03	Q17RF5	Uncharacterized protein C4orf26	Homo sapiens
2	19	5.78	4.70	2.81	Q14508	WAP four-disulfide core domain protein 2	Homo sapiens
3	17	6.57	4.07	2.57	Q8N690	Beta-defensin 119	Homo sapiens
4	26	9.78	6.48	2.50	P48061	Stromal cell-derived factor 1	Homo sapiens
5	46	21.70	9.81	2.48	Q8IZ96	CKLF-like MARVEL transmembrane domain-containing protein 1	Homo sapiens
6	16	6.22	3.95	2.47	O95925	Eppin	Homo sapiens
7	58	25.74	13.40	2.41	P10966	T-cell surface glycoprotein CD8 beta chain	Homo sapiens
8	16	4.87	4.66	2.39	Q8IUB3	Protein WFDC10B	Homo sapiens
9	2	0.43	0.66	2.36	Q7M0J9	Anantin	Streptomyces coerulescens
10	23	8.43	6.33	2.30	Q08648	Sperm-associated antigen 11B	Homo sapiens
11	20	8.35	5.08	2.29	P08493	Matrix Gla protein	Homo sapiens
12	43	20.74	9.97	2.23	P23560	Brain-derived neurotrophic factor	Homo sapiens
13	89	43.26	20.61	2.22	Q11203	CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltransferase	Homo sapiens
14	11	4.39	3.01	2.19	Q30KQ9	Beta-defensin 110	Homo sapiens

15	31	14.57	7.49	2.19	Q9NY56	Odorant-binding protein 2a	Homo sapiens
16	41	22.48	8.64	2.14	Q9BT67	NEDD4 family-interacting protein 1	Homo sapiens
17	21	8.35	6.15	2.06	Q9UNK4	Group IID secretory phospholipase A2	Homo sapiens
18	9	4.30	2.29	2.05	P15515	Histatin-1	Homo sapiens
19	118	60.96	27.90	2.04	P48740	Mannan-binding lectin serine protease 1	Homo sapiens
20	41	18.57	11.01	2.04	Q99674	Cell growth regulator with EF hand domain protein 1	Homo sapiens
DQB1* 06:02							
Rank	**Bound**	**Avg.**	**S.D**	**Z**	**ID**	**Protein**	**Species**
1	10	1.48	2.31	3.68	P85148	Bacteriocin E50-52	Enterococcus faecium
2	3	0.39	0.78	3.33	P01560	Heat-stable enterotoxin ST-2	Escherichia coli
3	24	7.17	6.11	2.75	P96363	ESAT-6-like protein EsxJ	Mycobacterium tuberculosis
4	24	7.22	6.13	2.74	O07932	Putative ESAT-6-like protein 10	Mycobacterium tuberculosis
5	10	3.91	2.25	2.70	Q03709	Lysis protein for colicin E7	Escherichia coli
6	22	7.52	5.37	2.70	Q9HD89	Resistin	Homo sapiens
7	24	7.35	6.31	2.64	O05299	ESAT-6-like protein EsxK	Mycobacterium tuberculosis
8	16	4.52	4.36	2.63	Q66A25	Major outer membrane lipoprotein	Yersinia pseudotuberculosis serotype I
9	16	4.52	4.36	2.63	Q8ZDZ6	Major outer membrane lipoprotein	Yersinia pestis
10	24	7.39	6.32	2.63	P95243	Putative ESAT-6-like protein 7	Mycobacterium tuberculosis
11	18	6.30	4.47	2.62	Q01523	Defensin-5	Homo sapiens
12	44	17.70	10.32	2.55	Q0T677	Cell division inhibitor SulA	Shigella flexneri serotype 5b
13	44	17.70	10.32	2.55	Q83RX1	Cell division inhibitor SulA	Shigella flexneri
14	37	16.22	8.18	2.54	A7MFW4	Cell division inhibitor SulA	Cronobacter sakazakii
15	11	3.00	3.16	2.53	B0B816	Small cysteine-rich outer membrane protein OmcA	Chlamydia trachomatis serovar L2
16	11	3.00	3.16	2.53	P0DJI1	Small cysteine-rich outer membrane protein OmcA	Chlamydia trachomatis
17	37	12.74	9.61	2.53	Q672H9	Protein VP2	Sapporo virus
18	44	17.83	10.36	2.53	Q31YM1	Cell division inhibitor SulA	Shigella boydii serotype 4
19	5	2.13	1.14	2.52	P0CJ68	Humanin-like protein 1	Homo sapiens
20	5	2.13	1.14	2.52	P0CJ75	Humanin-like protein 8	Homo sapiens
21	10	3.91	2.43	2.51	P15176	Lysis protein for colicin E9	Escherichia coli
22	5	1.83	1.27	2.51	P0CJ76	Humanin-like protein 9	Homo sapiens
23	44	17.96	10.42	2.50	A7ZK62	Cell division inhibitor SulA	Escherichia coli O139:H28

24	44	17.96	10.42	2.50	A7ZYR0	Cell division inhibitor SulA	Escherichia coli O9:H4
25	44	17.96	10.42	2.50	B1IVX4	Cell division inhibitor SulA	Escherichia coli
26	44	17.96	10.42	2.50	B1LJ40	Cell division inhibitor SulA	Escherichia coli
27	44	17.96	10.42	2.50	B1X8R2	Cell division inhibitor SulA	Escherichia coli
28	44	17.96	10.42	2.50	B2TTT7	Cell division inhibitor SulA	Shigella boydii serotype 18
29	44	17.96	10.42	2.50	B5YT88	Cell division inhibitor SulA	Escherichia coli O157:H7
30	44	17.96	10.42	2.50	B6I933	Cell division inhibitor SulA	Escherichia coli
31	44	17.96	10.42	2.50	B7LE58	Cell division inhibitor SulA	Escherichia coli
32	44	17.96	10.42	2.50	B7M887	Cell division inhibitor SulA	Escherichia coli O8
33	44	17.96	10.42	2.50	B7N3C1	Cell division inhibitor SulA	Escherichia coli O17:K52:H18
34	44	17.96	10.42	2.50	C4ZQ84	Cell division inhibitor SulA	Escherichia coli
35	44	17.96	10.42	2.50	P0AFZ5	Cell division inhibitor SulA	Escherichia coli
36	44	17.96	10.42	2.50	P0AFZ6	Cell division inhibitor SulA	Escherichia coli O157:H7
37	44	17.96	10.42	2.50	Q1JQN1	Cell division inhibitor SulA	Escherichia coli
38	44	17.96	10.42	2.50	Q3Z3G3	Cell division inhibitor SulA	Shigella sonnei
39	44	18.04	10.46	2.48	B7UN37	Cell division inhibitor SulA	Escherichia coli O127:H6
40	43	17.70	10.25	2.47	A1A9M7	Cell division inhibitor SulA	Escherichia coli O1:K1 / APEC
41	43	17.70	10.25	2.47	B7MIB2	Cell division inhibitor SulA	Escherichia coli O45:K1
42	43	17.70	10.25	2.47	Q1RDQ5	Cell division inhibitor SulA	Escherichia coli
43	44	18.30	10.50	2.45	B7MS69	Cell division inhibitor SulA	Escherichia coli O81
44	44	18.30	10.50	2.45	Q0TJA2	Cell division inhibitor SulA	Escherichia coli O6:K15:H31
45	44	18.30	10.50	2.45	Q8FJ79	Cell division inhibitor SulA	Escherichia coli O6:H1
46	11	3.04	3.30	2.41	C4PRC2	Small cysteine-rich outer membrane protein OmcA	Chlamydia trachomatis serovar B
47	11	3.04	3.30	2.41	P0CC05	Small cysteine-rich outer membrane protein OmcA	Chlamydia trachomatis
48	9	3.39	2.33	2.41	P10099	Lysis protein for colicin E8	Escherichia coli
49	42	18.13	9.96	2.40	B7NM15	Cell division inhibitor SulA	Escherichia coli O7:K1
50	43	17.70	10.59	2.39	B7LNW8	Cell division inhibitor SulA	Escherichia fergusonii

Supplementary Table 1. Candidate FL Antigens—Number of Peptide Presented. This table gives the top 50 candidate antigens for each of 4 FL associated alleles. For each protein, the total number of 15mer peptides predicted to be presented with affinity IC_{50} < 1,100nM was calculated for each allele. Protein wide averages and standard deviations were then calculated based on all 23 HLA class II alleles tested. Based on the number of peptides bound by the FL associated allele, the average, and standard deviation for a given protein, z-scores were calculated. Shown are the top 50 z-scores with absolute value >2.0 for 4 FL associated alleles. There was no overlap between the four lists.

DRB1* 01:01 rank	Unique	Avg.	S.D.	Z	ID	Protein Name	Species
1	7	0.96	1.55	3.90	P0A5X0	30S ribosomal protein S10	Mycobacterium tuberculosis
2	11	1.48	2.47	3.86	O11459	Membrane-associated protein VP24	Zaire ebolavirus
3	4	0.39	0.94	3.84	P17797	Outer membrane lipoprotein virB7	Agrobacterium tumefaciens
4	10	1.22	2.30	3.83	P35964	Virion infectivity factor	HIV type 1 group M subtype B
5	17	3.04	3.65	3.82	P16442	Histo-blood group ABO system transferase	Homo sapiens
6	19	3.26	4.16	3.78	P00813	Adenosine deaminase	Homo sapiens
7	3	0.35	0.71	3.71	P0A3W4	Outer membrane lipoprotein virB7	Rhizobium radiobacter
8	3	0.35	0.71	3.71	P0A3W5	Outer membrane lipoprotein virB7	Agrobacterium tumefaciens
9	19	2.70	4.40	3.70	P50469	M protein, serotype 2.2	Streptococcus pyogenes
10	10	1.57	2.29	3.68	Q6V1Q3	Membrane-associated protein VP24	Zaire ebolavirus
11	15	2.83	3.31	3.68	Q02104	Lipase 1	Psychrobacter immobilis
12	7	1.17	1.59	3.67	A8MT79	Putative zinc-alpha-2-glycoprotein-like 1	Homo sapiens
13	8	1.26	1.84	3.66	P55145	Mesencephalic astrocyte-derived neurotrophic factor	Homo sapiens
14	8	1.35	1.82	3.65	Q9UHA7	Interleukin-36 alpha	Homo sapiens
15	7	0.96	1.66	3.63	P19875	C-X-C motif chemokine 2	Homo sapiens
16	17	2.65	3.97	3.61	P13050	IgA receptor	Streptococcus pyogenes
17	18	2.87	4.19	3.61	O75493	Carbonic anhydrase-related protein 11	Homo sapiens
18	9	1.35	2.12	3.60	Q05322	Membrane-associated protein VP24	Zaire ebolavirus
19	11	1.70	2.62	3.55	P05898	Virion infectivity factor	HIV type 1 group M subtype B
20	18	2.57	4.38	3.53	P16946	Virulence factor-related M protein	Streptococcus pyogenes serotype M49
21	3	0.43	0.73	3.52	Q8N104	Beta-defensin 106	Homo sapiens
22	12	1.87	2.88	3.52	Q6UX52	Uncharacterized protein C17orf99	Homo sapiens
23	12	2.04	2.84	3.51	P46233	Sodium-type flagellar protein MotY	Vibrio parahaemolyticus serotype O3:K6
24	10	1.74	2.36	3.50	P22107	TraT complement resistance protein	Salmonella typhimurium
25	6	1.26	1.36	3.50	P22301	Interleukin-10	Homo sapiens
26	4	0.39	1.03	3.49	P0DMC3	Apelin receptor early endogenous ligand	Homo sapiens

27	14	3.04	3.14	3.49	P15088	Mast cell carboxypeptidase A	Homo sapiens
28	8	1.26	1.94	3.48	P68608	DNA-directed RNA polymerase 22 kDa subunit	Vaccinia virus
29	9	1.87	2.05	3.48	Q06092	Protein UL24 homolog	Human herpesvirus 6A
30	8	1.57	1.85	3.47	Q9ZCD8	Phospholipase D	Rickettsia prowazekii
31	4	0.57	0.99	3.46	Q8NES8	Beta-defensin 124	Homo sapiens
32	8	1.61	1.85	3.45	Q14116	Interleukin-18	Homo sapiens
33	2	0.22	0.52	3.44	P81052	Bacteriocin leucocin-B	Leuconostoc mesenteroides
34	12	2.48	2.78	3.43	Q6GE14	Gamma-hemolysin component A	Staphylococcus aureus
35	9	1.65	2.14	3.43	Q8Z6A0	Outer-membrane lipoprotein LolB	Salmonella typhi
36	5	1.00	1.17	3.43	Q2FWV6	Staphylococcal complement inhibitor	Staphylococcus aureus
37	5	1.00	1.17	3.43	Q99SU9	Staphylococcal complement inhibitor	Staphylococcus aureus
38	13	2.30	3.13	3.42	Q1PDC8	Matrix protein VP40	Lake Victoria marburgvirus
39	3	0.35	0.78	3.42	A9Q0M7	Bacteriocin ubericin-A	Streptococcus uberis
40	3	0.35	0.78	3.42	P06962	Lysis protein for colicin A	Citrobacter freundii
41	3	0.35	0.78	3.42	P36500	Lantibiotic salivaricin-A	Streptococcus salivarius
42	9	1.61	2.17	3.41	P03253	Protease	Human adenovirus C serotype 5
43	7	1.26	1.68	3.41	P04597	Virion infectivity factor	HIV type 1 group M subtype D
44	5	0.78	1.24	3.40	Q77375	Protein Vpr	HIV type 1 group O
45	13	2.04	3.23	3.40	Q9UBU2	Dickkopf-related protein 2	Homo sapiens
46	9	1.57	2.19	3.39	P03252	Protease	Human adenovirus C serotype 2
47	3	0.52	0.73	3.39	P45453	Competence pheromone	Bacillus subtilis
48	11	2.65	2.46	3.39	Q79FX8	Hydroxymycolate synthase MmaA4	Mycobacterium tuberculosis
49	9	1.61	2.19	3.38	A9MW01	Outer-membrane lipoprotein LolB	Salmonella paratyphi B
50	9	1.61	2.19	3.38	B4SUG5	Outer-membrane lipoprotein LolB	Salmonella newport

DQB1* 06:02 rank	**Unique**	**Avg.**	**S.D.**	**Z**	**ID**	**Protein Name**	**Species**
1	7	0.57	1.47	4.37	P85148	Bacteriocin E50-52	Enterococcus faecium
2	5	0.48	1.08	4.18	P06963	Lysis protein for colicins E2 and E3	Escherichia coli
3	12	1.96	2.42	4.15	Q04470	Type-2Aa cytolytic delta-endotoxin	Bacillus thuringiensis subsp. kyushuensis
4	10	1.09	2.15	4.14	P04118	Colipase	Homo sapiens
5	7	0.74	1.54	4.06	P01215	Glycoprotein hormones alpha chain	Homo sapiens

6	4	0.39	0.89	4.05	Q03709	Lysis protein for colicin E7	Escherichia coli
7	9	1.87	1.77	4.04	Q9JFA5	Core protein D3	Vaccinia virus
8	10	1.91	2.02	4.00	O57210	Core protein D3	Vaccinia virus
9	10	1.91	2.02	4.00	P21009	Core protein D3	Vaccinia virus
10	7	0.74	1.57	3.98	P0A564	6 kDa early secretory antigenic target	Mycobacterium tuberculosis
11	7	0.74	1.57	3.98	P0A565	6 kDa early secretory antigenic target	Mycobacterium bovis
12	11	1.78	2.33	3.95	Q3IK97	Outer-membrane lipoprotein LolB	Pseudoalteromonas haloplanktis
13	4	0.48	0.90	3.92	P33835	Virion membrane protein A9	Variola virus
14	12	1.65	2.66	3.90	Q82TQ2	Outer-membrane lipoprotein LolB	Nitrosomonas europaea
15	7	0.83	1.59	3.89	P10513	Pilin	Escherichia coli
16	9	1.09	2.04	3.87	Q92M53	Outer membrane lipoprotein omp10 homolog	Rhizobium meliloti
17	33	5.09	7.22	3.86	P18047	Fiber protein 1	Human adenovirus F serotype 40
18	5	0.39	1.20	3.85	P0A310	Bacteriocin sakacin-A	Lactobacillus sakei
19	5	0.39	1.20	3.85	P0A311	Bacteriocin curvacin-A	Lactobacillus curvatus
20	11	1.43	2.48	3.85	P16795	Glycoprotein N	Human cytomegalovirus
21	10	1.57	2.19	3.85	A7MFW4	Cell division inhibitor SulA	Cronobacter sakazakii
22	33	5.13	7.24	3.85	P16071	Hemagglutinin-neuraminidase	Human parainfluenza 1 virus
23	10	1.30	2.27	3.84	A8AIC1	Cell division inhibitor SulA	Citrobacter koseri
24	9	1.13	2.05	3.84	P06311	Ig kappa chain V-III region IARC/BL41	Homo sapiens
25	4	0.39	0.94	3.84	P21185	Lysis protein for colicin E1*	Shigella sonnei
26	7	0.83	1.61	3.83	P04737	Pilin	Escherichia coli
27	27	4.39	5.91	3.82	P13119	Flagellin B	Rhizobium meliloti
28	18	3.22	3.87	3.82	Q8YB48	Glucose/galactose transporter	Brucella melitensis biotype 1
29	12	2.26	2.58	3.77	P18773	Esterase	Acinetobacter lwoffii
30	4	0.43	0.95	3.77	P13345	Lysis protein for colicin E6	Escherichia coli
31	9	1.35	2.04	3.76	B7MS69	Cell division inhibitor SulA	Escherichia coli O81
32	9	1.35	2.04	3.76	Q0TJA2	Cell division inhibitor SulA	Escherichia coli O6:K15:H31
33	9	1.35	2.04	3.76	Q8FJ79	Cell division inhibitor SulA	Escherichia coli O6:H1
34	21	3.65	4.62	3.76	P12446	Polyprotein p42 [Cleaved into: Protein M1'	Influenza C virus
35	13	2.39	2.82	3.76	Q9PK23	Putative outer membrane protein TC_0650	Chlamydia muridarum
36	7	0.87	1.63	3.76	P14496	Pilin	Escherichia coli
37	8	1.22	1.81	3.75	P60672	Envelope protein A28 homolog	Yaba monkey tumor virus

38	13	2.26	2.86	3.75	P05777	Matrix protein 1	Influenza A virus
39	9	1.43	2.02	3.75	P19249	Thermostable direct hemolysin 1	Vibrio parahaemolyticus serotype O3:K6
40	22	4.09	4.79	3.74	Q65RJ5	Membrane-bound lytic murein transglycosylase F	Mannheimia succiniciproducens
41	9	1.48	2.02	3.72	A7ZK62	Cell division inhibitor SulA	Escherichia coli O139:H28
42	9	1.48	2.02	3.72	A7ZYR0	Cell division inhibitor SulA	Escherichia coli O9:H4
43	9	1.48	2.02	3.72	B1IVX4	Cell division inhibitor SulA	Escherichia coli
44	9	1.48	2.02	3.72	B1LJ40	Cell division inhibitor SulA	Escherichia coli
45	9	1.48	2.02	3.72	B1X8R2	Cell division inhibitor SulA	Escherichia coli
46	9	1.48	2.02	3.72	B2TTT7	Cell division inhibitor SulA	Shigella boydii serotype 18
47	9	1.48	2.02	3.72	B5YT88	Cell division inhibitor SulA	Escherichia coli O157:H7
48	9	1.48	2.02	3.72	B6I933	Cell division inhibitor SulA	Escherichia coli
49	9	1.48	2.02	3.72	B7LE58	Cell division inhibitor SulA	Escherichia coli
50	9	1.48	2.02	3.72	B7M887	Cell division inhibitor SulA	Escherichia coli O8

Supplementary Table 2. Candidate FL Antigens—Uniquely Presented Peptides. This table gives the top 50 candidate antigens for FL associated alleles DRB1*01:01 and DQB1*06:02. For each protein, the total number of 15mer peptides predicted to be presented with affinity IC_{50} < 1,100nM was calculated for each allele. Within those, the subset of 'unique peptides' presented by one allele, but no others, was tabulated. Averages and standard deviations were calculated based for all 23 HLA class II alleles tested. Based on the number of unique peptides bound by the FL associated allele, the average, and standard deviation for a given protein, z-scores were calculated. Shown are the top 50 z-scores for DRB1*01:01 and DQB1*01:01. There was no overlap between the two lists.

Comparison	T-statistic	Degrees of Freedom	2-Tailed p-value
DRB1: Human vs. Virus	-21.8	4890.4	2.9E-100
DRB1: Bacteria vs. Virus	-13.8	5126.4	7.8E-43
DRB1: Human vs. Bacteria	-13.2	3642.9	1.1E-38
DQB1: Human vs. Virus	-11.9	4760.0	4.5E-32
DQB1: Bacteria vs. Virus	25.1	7667.2	2.1E-133
DQB1: Human vs. Bacteria	-34.3	5232.3	1.8E-232
DPB1: Human vs. Virus	-15.1	5217.1	9.4E-51
DPB1: Bacteria vs. Virus	-28.7	5013.9	1.3E-167
DPB1: Human vs. Bacteria	10.7	4150.6	3.5E-26

Supplementary Table 3. T-tests comparing means of peptides presented. T-tests were performed to compare human, bacterial, and viral protein peptide presentation within the HLA-DRB1, DQB1, and DPB1 genes. Each test was performed using the R command 't.test', with the alternative hypothesis being that the true difference in means between each group was not equal to zero. The command default of using Welch modification to calculate the degrees of freedom was also used. A negative T-statistic indicates the mean of the group listed first is lower than the group listed second.

Supplementary Materials: netmhc_submit.sh. This short script is an overall wrapper, designed to feed fasta sequences into *netMHCII*, take the output, and send it to *netmhcreader.pl*

```
#!/bin/sh
#$ -cwd
#$ -V
#$ -j y
#$ -S /bin/sh
#netmhc_submit.sh

FILENAME=$BLASTDB/uniprot-all-antigens.fasta #our FASTA sequences to be
tested
ALLELE=HLA-DRB10101,HLA-DRB10301  #the HLA alleles being ran
PREFIX=dr1
FILE1=$PREFIX-tempfile
FILE2=$PREFIX-tempfile2
LIMIT=0.295  #cutoff of IC50 = 2000

for LENGTH in 12 15 18 21 #cycle through different peptide lengths
do
    OUTFILE=$PREFIX-$LENGTH
    count0=
    exec 3<&0
    exec 0< $FILENAME
    echo "START $(date)" >>log-$OUTFILE  #create a log file.
    cnt=0  #seed a counting variable
    while read LINE     #go through the FASTA file
    do
    if [[ "$LINE" == \>* ]]; then   #when you see a sequence header
            cnt=$((cnt + 1))    #skip the first header (it would send nothing
to netmhcii)
        if [[ "$cnt" -gt 1 ]]; then
            netMHCII -a $ALLELE -l $LENGTH -t $LIMIT -s $FILE1 > $FILE2
#call netMHCII
            perl netmhcreader.pl $FILE2 $COLLECT $OUTFILE #send the output to
a perl script for reading.
        fi
            echo "$LINE" > $FILE1
    else
            echo "$LINE" >> $FILE1
    fi
    done
    netMHCII -a $ALLELE -t $LIMIT -s $FILE1 > $FILE2    #for the last
sequence
    perl netmhcreader.pl $FILE2 $OUTFILE
    exec 0<&3
    echo "END netmhcii & netmhcreader $(date)">>log-$OUTFILE
done
```

Supplementary Materials: netmhcreader.pl. This script takes the output from *NetMHCII*, which is designed to be visually examined, and reformats it for ease of use downstream.

```
#! /usr/bin/perl
# netmhcreader.pl by kipp akers
#use this script to scan through netMHCii output files, removing all the extra stuff.
#usage: perl netmhcreader.pl netmhcii.output output.filename

#Collect all the gathered data
$filename = $ARGV[1]; #argument 2 is the filename prefix
$filename = $filename . "tophits.txt";
open(TEMPOUT, $ARGV[0]) or die "can't open your file: $!"; #create temporary file
open (OUTPUT, ">>$filename") or die "write-to file error: $! \n"; #create the output file
        while ($_ = <TEMPOUT>) { #read the file
                $_ =~ s/^\s+//; #remove leading whitespace
                if ($_ =~ /^HLA-/) { #if the line starts with HLA- (ie. if it has peptide binding data)
                                print OUTPUT "$_"; #collect it.
                }
        }
        ##print "there are $counter lines in $file\n";
close (OUTPUT);
close(TEMPOUT);
```

Supplementary Materials: nmhc_analyzer.pl. This script in perl will take in large sets of NetMHCII output data (adjusted by netmhcreader.pl) and perform a number of tasks. The first phase greatly compresses the data into a much smaller format. The second phase sorts the data and adds in missing lines when a protein/allele combination is not predicted to bind any proteins. Finally the third phase calculates a number of statistics for each protein sequence, asking if our alleles of interest bind this protein uniquely/more strongly etc.

```
#! /usr/bin/perl
use warnings;
use Scalar::Util qw(looks_like_number);
#nmhc_analyzer.pl by kipp akers
#take an netmhcii (sorted- -s flag used) output and extract meaningful stats.
#USAGE r2analyzer.pl <netMHCiisortedresultsfile> optional:SKIP

####PHASE 1===OPEN INDIVIDUAL NETMHCII FILES AND PROCESS EACH ALLELE/PROTEIN
COMBO INTO 1 LINE.
#ADJUSTABLE VARIABLES#
$cutoff1 = 50;
$cutoff2 = 500;
$cutoff3 = 1100;
$topnumb = 5;

#SET SOME VARIABLES
@split = NULL;
@nsplit = NULL;
$currentallele=NULL;
$count = 0;
$count1 = 0;
$count2 = 0;
$count3 = 0;
@allelelist = qw(
    HLA-DPA101-DPB10401
    HLA-DPA10103-DPB10201
    HLA-DPA10201-DPB10101
    HLA-DPA10201-DPB10501
    HLA-DPA10301-DPB10402
    HLA-DPB10301-DPB10401
    HLA-DQA10101-DQB10501
    HLA-DQA10102-DQB10602
    HLA-DQA10301-DQB10302
    HLA-DQA10401-DQB10402
    HLA-DQA10501-DQB10201
    HLA-DQA10501-DQB10301
    HLA-DRB10101
    HLA-DRB10301
    HLA-DRB10401
    HLA-DRB10404
    HLA-DRB10405
    HLA-DRB10701
    HLA-DRB10802
    HLA-DRB10901
    HLA-DRB11101
    HLA-DRB11302
    HLA-DRB11501
);
```

```
if ($ARGV[1] ne "SKIP"){ ####Skip trick.  Use argument 1 SKIP to activate

#READ AND PROCESS NETMHCII DATA
open (READFILE, "<", $ARGV[0]) || die "cannot open the read file: $!\n";
open (OUTFILE, ">", "step1out.txt") || die "cannot open writefile1: $!\n";
    while (<READFILE>) {
        chomp;
        $count++;
        $_ =~ s/WB\s+|SB\s+//; #remove the binder designations from netMHCII
        @split = split (/\s+/, $_); # 0:allele 1:amino# 2:peptide 3:core
4:affinity 5:aff.in.nM 6:%random 7:ID
        if (($currentallele ne NULL) && (($currentallele ne $split[0]) ||
($split[7] !~ $proteinid))) {
                    #if it's a new allele/protein combination, do math,
output it all to one line, then clear the variables.
            #CALCULATE STATS AND OUTPUT
            $avg1 = &average(\@cutoff1_affs); #average of binding affinity
under cutoff1.
            $std1 = &stdev(\@cutoff1_affs);
            $med1 = &median(\@cutoff1_affs);
            $avg2 = &average(\@cutoff2_affs);
            $std2 = &stdev(\@cutoff2_affs);
            $med2 = &median(\@cutoff2_affs);
            $avg3 = &average(\@cutoff3_affs);
            $std3 = &stdev(\@cutoff3_affs);
            $med3 = &median(\@cutoff3_affs);
            $avgt = &average(\@topcount_affs);
            $stdt = &stdev(\@topcount_affs);
            $medt = &median(\@topcount_affs);
            print OUTFILE
"$proteinid\t$currentallele\t$count1\t$avg1\t$med1\t$std1\t$count2\t$avg2\t$m
ed2\t$std2\t$count3\t$avg3\t$med3\t$std3\t$avgt\t$medt\t$stdt\t@coresarray\n"
;
            #CLEAR VARIABLES
            %cores = ();
            $currentallele = $split[0];
            @nsplit = split (/\|/, $_); #pull the name from between brackets
            $proteinid = $nsplit[1];
            $count1 = 0;
            $count2 = 0;
            $count3 = 0;
            $topcount = 0;
            @cutoff1_affs = ();
            @cutoff2_affs = ();
            @cutoff3_affs = ();
            @topcount_affs = ();
            @coresarray = ();
        }
        #PROPERLY INITIATE
        if ($currentallele eq NULL) {#for the first line.
            $currentallele = $split[0];
            @nsplit = split (/\|/, $_); #pull the name from between brackets
            $proteinid = $nsplit[1];
        }

        ####MEAT N POTATOES--Where most stuff happens.
        next if (exists $cores{$split[3]});  #if the core has already been
counted, skip it.
        $cores{$split[3]} = 1;#if the core is new, create a new hash key-
pair. split[3] is the core sequence
        $topcount++; #to index the the properties of the top X bound cores, I
use this.
        #Begin cutoff specific accounting
        if ($split[5] < $cutoff1) { #if the affinity is less than my set
cutoff
            $count1++;  #count it
            push (@cutoff1_affs, $split[5]); #add it to the array
        }
        if ($split[5] < $cutoff2) {
            $count2++;
            push (@cutoff2_affs, $split[5]);
        }
        if ($split[5] < $cutoff3) {
            $count3++;
            push (@cutoff3_affs, $split[5]);
            push (@coresarray, $split[3]);
        }
        if ($topcount <= $topnumb) { #if we haven't reached the X in our top
X bound peptides
            push (@topcount_affs, $split[5]);
        }
    }
#CALC STATS AND OUTPUT (FOR LAST PROTEIN/ALLELE COMBO)
$avg1 = &average(\@cutoff1_affs);
$std1 = &stdev(\@cutoff1_affs);
$med1 = &median(\@cutoff1_affs);
$avg2 = &average(\@cutoff2_affs);
$std2 = &stdev(\@cutoff2_affs);
$med2 = &median(\@cutoff2_affs);
$avg3 = &average(\@cutoff3_affs);
$std3 = &stdev(\@cutoff3_affs);
$med3 = &median(\@cutoff3_affs);
$avgt = &average(\@topcount_affs);
$stdt = &stdev(\@topcount_affs);
$medt = &median(\@topcount_affs);
print OUTFILE
"$proteinid\t$currentallele\t$count1\t$avg1\t$med1\t$std1\t$count2\t$avg2\t$m
ed2\t$std2\t$count3\t$avg3\t$med3\t$std3\t$avgt\t$medt\t$stdt\t@coresarray\n"
;
close OUTFILE;
close READFILE;

###PHASE 2 SORT THE OUTFILE/ADD MISSING LINES.
### make sure you have the env var LANG set to C before this.  I have
modified my .bashrc file.
#SORT
system("sort -b -k 1 -k 2 step1out.txt >sortedtemp.txt"); #sort on protein id
then on allele
#ADD IN MISSING LINES##
    #if no cores bind below the set threshold nmhc outputs nothing.  I want
zeros in my file.
open (TEMP, "<", "sortedtemp.txt") || die "cannot open sortedtemp.txt for
reading: $!\n";
```

```
open (OUTFILE2, ">", "step2out.txt") || die "cannot create step2out.txt:
$!\n";
$proteinid = '';
$tempcounter = 0;
@missing = ();
while (<TEMP>) { #read through the file
        chomp;
        while ($_ !~ /$allelelist[$tempcounter]/) {#if the line doesn't match
the predicted allele (ie if there is a missing line)
                if ($proteinid ne '') {
                        print OUTFILE2
"$proteinid\t$allelelist[$tempcounter]\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0
\t0\n"; # filler line of zeros
                        $tempcounter++; #tempcounter keeps track of what
allele we should be at.
                }
                else {
                        push(@missing, $tempcounter); #@missing is the array
of alleles with missing data.
                        $tempcounter++;
                }
                if ($tempcounter > 22) { #if we're on the last allele-#23
                        $tempcounter = 0; #clear everything.
                        $proteinid = '';
                }
        }
        @split = split (/\s+/, $_); # 0:proteinid 1:alleleid
        $proteinid = $split[0];
        if (@missing) { #if the missing array has values
                foreach (@missing) {
                        print OUTFILE2
"$proteinid\t$allelelist[$_]\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\t0\n";
                }
                @missing = ();
        }
        $tempcounter++;
        if ($tempcounter >22) {
                $tempcounter = 0;
                $proteinid = '';
        }
        print OUTFILE2 "$_\n";
}
close TEMP;
close OUTFILE2;
###End SKIP trick.  this can be moved around for convenience when
troubleshooting.
}
###PHASE 3 LOOK FOR INTERESTING PROTEINS###
#ADJUSTABLE VARIABLES
#only change if you run more alleles than 6dp/6dq/11dr
@dp = (1..6);
@dq = (7..12);
@dr = (13..23);
$dr01 = 13;
$dq05 = 7;
$dq06 = 8;
$dr15 = 23;
```

```
@allnr = (1,2,3,4,5,6,8,9,10,11,12,14,15,16,17,18,19,20,21,22,23); #all non-
risk alleles
@allnp = (1,2,3,4,5,6,7,9,10,11,12,13,14,15,16,17,18,19,20,21,22);  #all non-
protective alleles.
#SET VARIABLES
$cycler = 0;
@AoH = ();
@dataline = ();
$count = 0;
$drtopcount = 0;
$dqtopcount = 0;

##OPEN UP OUTPUT FILES/READFILES###
open (OUTFILE3, ">", "unique_cores.txt") || die "cannot open step 3 output
file: $!\n";
    print OUTFILE3 "Comparison\tProteinID\tCore_seq\tRank\n";
open (OUTFILE4, ">", "numberof_cores.txt") || die "cannot open step 3 output
file: $!\n";
    print OUTFILE4
"Comparison\tProteinID\tRisk_Allele_Cores\tNR_Average\tNR_stdev\n";
open (OUTFILE5, ">", "uniques_per_protein.txt") || die "cannot open step 3
output file:$!\n";
open (READFILE2, "<", "step2out.txt") || die "cannot open the read file2:
$!\n"; #open the sorted file with my peptide binding summaries
while (<READFILE2>) {
    chomp;
    $count++;
    $cycler++;
    if ($cycler > 23) { #after I've indexed all the alleles for a given
protein.
       ###is there a difference between risk/non-risk and average for this
protein?
        $avgnumcores = &average(\@allnumofcores);
        $stdnumcores = &stdev(\@allnumofcores);
        #DQ05 vs all alleles
        if (abs($avgnumcores-$allnumofcores[($dq05-1)]) > (2*$stdnumcores)){
#sign. test: if DQ05 has abs(z-score) >2
            print OUTFILE4 "DQ05vsALL\t$dataline[0]\t$allnumofcores[($dq05-
1)]\t$avgnumcores\t$stdnumcores\n";
        }
        #DR01 vs all alleles
        if (abs($avgnumcores-$allnumofcores[($dr01-1)]) > (2*$stdnumcores)){
            print OUTFILE4 "DR01vsALL\t$dataline[0]\t$allnumofcores[($dr01-
1)]\t$avgnumcores\t$stdnumcores\n";
        }
        #DQ06 vs. all alleles
        if (abs($avgnumcores-$allnumofcores[($dq06-1)]) > (2*$stdnumcores)){
                        print OUTFILE4
"DQ06vsALL\t$dataline[0]\t$allnumofcores[($dq06-
1)]\t$avgnumcores\t$stdnumcores\n";
        }
        #DR15 vs. all alleles
        if (abs($avgnumcores-$allnumofcores[($dr15-1)]) > (2*$stdnumcores)){
#sign test
                            print OUTFILE4
"DR15vsALL\t$dataline[0]\t$allnumofcores[($dr15-
1)]\t$avgnumcores\t$stdnumcores\n";
```

```
        }
        ###CHECK THE CORES FOR UNIQUENESS
        #tally the number of unique cores presented
        foreach $allele (1..23) {
            $uniques_for_this_allele= 0;
            while (($key, $value) = each %{$AoH[$allele]}) {#this cycles 
through the cores of our $AoH[risk allele]
                                $testcount =0;
                                foreach $x (1..23) { #cycle through all the 
comparison alleles
                                        if (exists $AoH[$x]{$key}) { #if 
another allele presents the core
                                                $testcount++; #tally
                                        }
                                }
                             if ($testcount == 1) { #ie if the above core is 
unique
                                        $uniques_for_this_allele++;
                                 }
                         }
            print OUTFILE5 
"Allele:$allele\t$dataline[0]\t$uniques_for_this_allele\n";
        }

        ##DQ05 vs ALLNR
        while (($key, $value) = each %{$AoH[$dq05]}) {#this cycles through 
the cores of our $AoH[risk allele]
            $testcount =0;
            foreach $x (@allnr) { #cycle through all the comparison alleles
                if (exists $AoH[$x]{$key}) { #if another allele presents the 
core
                    $testcount++; #tally
                }
            }
            if ($testcount == 0) { #ie if the above core is unique
                my $adjvalue = ($value-16);
                if ($adjvalue < $topnumb) {##topcores vs. all cores
                    print OUTFILE3 
"DQ05vsAll\t$dataline[0]\t$key\t$adjvalue\n";
                }
            }
        }
        ##DR01 vs ALLNR
        while (($key, $value) = each %{$AoH[$dr01]}) {#this cycles through 
the cores of our $AoH[risk allele]
            $testcount =0;
            foreach $x (@allnr) { #cycle through all the comparison alleles
                if (exists $AoH[$x]{$key}) { #if another allele presents the 
core
                    $testcount++; #tally
                }
            }
            if ($testcount == 0) { #ie if the above core is unique
                my $adjvalue = ($value-16);
                if ($adjvalue < $topnumb) {##topcores vs. all cores
                    print OUTFILE3 
"DR01vsALL\t$dataline[0]\t$key\t$adjvalue\n";
```

```
                }
            }
        }
        ##DR15 vs ALLNP
        while (($key, $value) = each %{$AoH[$dr15]}) {#this cycles through
the cores of our $AoH[risk allele]
                $testcount =0;
                foreach $x (@allnp) { #cycle through all the comparison
alleles
                        if (exists $AoH[$x]{$key}) { #if another allele
presents the core
                                $testcount++; #tally
                        }
                }
                if ($testcount == 0) { #ie if the above core is unique
                        my $adjvalue = ($value-16);
                        if ($adjvalue < $topnumb) {##topcores vs. all cores
                                print OUTFILE3
"DR15vsALLnp\t$dataline[0]\t$key\t$adjvalue\n";
                        }
                }
        }
        ##DQ06 vs ALLNP
        while (($key, $value) = each %{$AoH[$dq06]}) {#this cycles through
the cores of our $AoH[risk allele]
                $testcount =0;
                foreach $x (@allnp) { #cycle through all the comparison
alleles
                        if (exists $AoH[$x]{$key}) { #if another allele
presents the core
                                $testcount++; #tally
                        }
                }
                if ($testcount == 0) { #ie if the above core is unique
                        my $adjvalue = ($value-16);
                        if ($adjvalue < $topnumb) {##topcores vs. all cores
                                print OUTFILE3
"DQ06vsALLnp\t$dataline[0]\t$key\t$adjvalue\n";
                        }
                }
        }
        #RESET VARIABLES
        @AoH = ();
        $drtopcount = 0;
        $dqtopcount = 0;
        $cycler = 1;
        @allnumofcores =();
        @drnumofcores = ();
        @dqnumofcores = ();
        #sleep(1);

    }
    ##TEMPORARY STORAGE
    @dataline = split (/\s+/, $_); #take the line and split it
    ##########CORES###############
    for my $i (17..$#dataline) {#17+ on are the core seqs
```

```
        my $core = $dataline[$i];#this for loop should put all core data for
1 protein into an AoHs (array of hashes)
        $AoH[$cycler]{$core} = $i; #dataformat:  $AoH[allele#]{core
sequence}=rank+16 if it exists.
    }
    #PRESENTABLES
    push (@allnumofcores, $dataline[10]); #note the allele number =/= array
index here, b/c index starts at zero
        #dataline[10] is the number of (cutoff3) bound peptides.
}

sub average{ ###compliments to kate in edwards lab at SDSU.  Adapted for
missing data.
        my($data) = @_;
        if (not @$data) {
                return(0);
        }
        my $total = 0;
        my $number = 0;
        foreach (@$data) {
                if (looks_like_number($_)) {
                        $total += $_;
                        $number++;
                }
        }
        my $average = $total / $number;
        return $average;
}

sub median { ##adapted.  NOTE that this relies on the ordered data I know I
have.
    my($data) = @_;
    if (not @$data) {
            return(0);
    }
    my $midpoint = @$data/2;
    if ($midpoint =~ m/\D/ ) {#if it has a decimal
        $upper = $midpoint - 0.5;#minus because arrays start at zero
        $median = @$data[$upper];
    }
    else {
        $lower = @$data[$midpoint];
        $upper = $midpoint - 1; #minus because arrays start at zero
        $upper = @$data[$upper];
        $median = ($upper + $lower)/2;
    }
    return $median;
}
sub stdev{ ###compliments to kate in edwards lab at SDSU
        my($data) = @_;
        if(@$data == 1){
                return 0;
        }
        my $average = &average($data);
        my $sqtotal = 0;
        my $number = 0;
        foreach(@$data) {
                if (looks_like_number($_)) {
                        $number++;
                        $sqtotal += ($average-$_) ** 2;
                }
        }
        my $std = ($sqtotal / ($number-1)) ** 0.5;
        return $std;
}
```

Chapter 6

Summary and Conclusions

This dissertation set forth to bring research of HLA alleles and their effect on FL risk one step closer to positively impacting the public. Preventing lymphoma and other HLA associated diseases remains a public health priority. With a better understanding of how HLA genetic changes are impacting individual risk of FL, we position ourselves to make great improvements in the lives of many individuals. Advancements in our knowledge of how this gene region impacts this one lymphoma subtype have the potential to be widely applicable to a number of common cancers and autoimmune diseases.

The preceding chapters have demonstrated that there are several independent risk factors for FL within the HLA class II gene region. A locus increasing FL risk and a locus decreasing FL risk both appear to exist in the region spanning *HLA-DQB1* and *HLA-DRB1*. These associations may in fact both be localized to a single amino acid position of *HLA-DRB1*. A third, independent, FL associated allele appears to exist at *HLA-DPB1*.

A growing body of evidence appears to indicate that peptide presentation plays a major role in the etiology of FL. This hypothesis is supported in this dissertation by indications that the amino-acid positions most associated with FL reside within the peptide binding groove of *HLA-DRB1*. Variation at these residues has previously been shown to influence peptide binding, and is shown here to effect individual FL risk. Further evidence of the importance of peptide binding to FL is given using a computational approach. When tested for their predicted ability to bind environmental proteomes, FL-associated alleles *HLA-DRB1**01:01 and *HLA-DQB1**05:01 were shown to be consistent outliers. These alleles are predicted to bind environmental peptides stronger and weaker, respectively, than their non-FL associated counterparts. This indicates that FL risk may be attenuated by the strength with which an individual binds an antigen, or group of antigens.

For a number of reasons, these are very exciting conclusions. The discovery of multiple, independent, risk loci for this disease within the HLA region underscores two major points. First, given the population frequency and the degree of increased risk associated with these alleles, it appears that a large percentage of population risk for FL is attributable to this gene region. When coupled with the recent discovery that t(14:18) positive cell counts predict FL risk up to 15 years in advance of disease presentation (1), we likely have the ability to characterize individual risk for FL using just a small amount blood and simple genotyping. Secondly, the breadth of these associations indicates that these proteins may be central to development of FL. Molecular characterization of this disease association will be crucial in determining not only how to treat FL, but how to prevent it as well.

A number of studies were performed concurrently with this dissertation, and their findings are a great supplement to the research presented here. Sillé et al. demonstrated that *in-vitro*, the FL protective haplotype spanning *HLA-DRB1* and *HLA-DQB1* is associated with increased expression of *HLA-DQB1* (2). This finding expanded upon and confirmed a previous report making use of publicly available RNA-Seq data (3). The implication of this research is that the causal locus for this protective association is a regulatory feature which increases *HLA-DQB1* expression. One can further hypothesize that increased *HLA-DQB1* protein levels are FL-protective. These findings do not at all contradict the research presented in this dissertation.

The data presented in Chapter 5 indicates that the FL risk haplotype spanning *HLA-DRB1**01 and *HLA-DQB1**05:01 may be functioning via particularly strong or weak peptide binding. In that chapter, FL protective alleles *HLA-DRB1**15:01 and *HLA-DQB1**06:02 were average compared to other HLA Class II alleles. Perhaps these two HLA associations are functioning in different ways to affect the same molecular pathway. *HLA-DQB1**06:02 may be FL-protective by being highly expressed and presenting peptides with average strength, while *HLA-DQB1**05:01 induces FL risk by being expressed at average levels but presenting peptides with below average strength. This is just one hypothesis to explain the observed data. A second hypothesis is that increased *HLA-DQB1* expression is protective by recognizing and targeting pre-FL B cells, while *HLA-DRB1**01's strong binding of peptides creates enhanced antigenic stimulation, leading to a pro-inflammatory state and increasing FL risk. There appears to be much left to discover within this field.

Fascinating research has come from an extremely high-powered genome-wide association study of FL (4). By making use of exceptionally large numbers of FL cases and controls, these investigators were able to detect more subtle FL-associated loci outside of the HLA region. In doing so, FL risk alleles were discovered in genes impacting B-cell differentiation and signaling, as well as known B cell lymphoma oncogenes. This research is still under review, but the findings will no doubt play a large role in directing the next wave of lymphoma research. As more and more genomic associations with FL are discovered, the pathways on the causal map of this disease will be become clear. That information will be an invaluable resource to molecular biologists and medical researchers studying the events that lead to formation of FL.

The role of regulatory T cells in the development of FL remains a fascinating research area which may be dependent on HLA class II alleles. Research shows that FL cells enable the conversion of helper T cells to regulatory T cells (5), presumably a feature that improves survival by suppressing anti-tumor immune response. More recent research has focused on reversing the FL tolerant state within FL patients by triggering certain toll-like receptors (6). With this treatment, regulatory T cells were inhibited and effector T cell function was restored. This strategy has had success *in-vitro*; however, it remains to be seen if these results will hold when treating patients. With a diverse treatment population, the effect of HLA variability may need to be accounted for. It is clear that interactions between FL cells and T cells are HLA class II dependent. This was initially shown many years ago (7), but confirmed once again more recently (6). The extent to which genetic variability at the HLA class II locus affects T cell interactions in FL patients remains understudied.

Similarly, other cellular residents of the FL microenvironment may be attenuating FL risk via HLA class proteins. The survival time for FL patients is highly dependent on the population of non-malignant cells within the FL tumor (8). While the presence of T cells generally indicates favorable outcome, macrophages and dendritic cells are associated with shorter survival. Of course, the latter two are both professional antigen-presenting cells which express HLA class II proteins. Perhaps relatedly, it was recently reported that *CD14*+ dendritic cells localizing to the follicle were associated with shorter times to transformation among FL patients (9). Transformation is generally associated with a worse prognosis and shorter survival times. Previous research has shown that more aggressive B cell lymphoma is associated with increased cells in the blood positive for *CD14* and with low *HLA-DRB1* (10). It can be hypothesized that HLA class II alleles are attenuating FL risk via non-B cell antigen presenting cells. This hypothesis coincides with what we know about the cellular function of HLA class II proteins on these cells and the research showing the impact these cells have on FL progression and survival. The difficulty of working with these cell types likely explains why this topic remains under-utilized in research.

Future studies aiming to build off the research presented here should begin with an expansion of the work presented in Chapter 3. Discovering the causal locus of association within HLA haplotypes is an important task for all HLA-FL research. The exact genetic change that impacts FL risk is unlikely to be revealed with genetic epidemiology alone; however, reducing the associated haplotype to a single associated gene will simplify this task greatly. A well-designed study of FL in a non-Caucasian population could have strong power to detect the difference between FL associated alleles at *HLA-DRB1* and *HLA-DQB1*. This research is currently underway, with African-American and Asian-American FL case DNA being gathered for genome-wide genotyping.

Further work is also needed to better characterize the bound peptide profile of cells in FL and non-FL states. Chapter 5 demonstrates that certain HLA alleles are likely to present antigen peptides with higher affinity, and that this may be influencing FL risk. A follow-up study to characterize the HLA Class II bound peptide repertoire, both *in-vitro* and *in-vivo,* is needed. An *in-vitro* system would allow the testing of antigen presenting cells for their ability to process and present candidate FL antigens, such as those highlighted in Chapter 5. The greatest impact research, however, is likely to come from an examination of those peptides being presented by HLA class II proteins in FL patient tumor material. This experimental design would allow an unprecedented look at the antigens which are present and impacting the growth and proliferation of FL cells and the FL microenvironment. It is possible that a few specific antigens dominate HLA Class II proteins in fully developed FL cells. It is also possible that FL cells have mutated to the point where any non-specific antigen will lead to FL cell proliferation. Either finding would be an important discovery. The methodology of such an experiment is complex, requiring collaboration between cancer biologists, immunologists and mass spectrometry experts. Furthermore, acquiring fresh or frozen FL tumor material, in sufficient quantities to strip and quantify bound peptides, is a difficult task. However, the potential for discovery of such an experiment far outweighs the difficulty of performing it, and I hope to see such research soon.

1. Roulland S, Kelly RS, Morgado E, Sungalee S, Solal-Celigny P, Colombat P, et al. t(14;18) Translocation: A Predictive Blood Biomarker for Follicular Lymphoma. J Clin Oncol Off J Am Soc Clin Oncol. 2014 May 1;32(13):1347–55.

2. Sillé FCM, Conde L, Zhang J, Akers NK, Sanchez S, Maltbaek J, et al. Follicular lymphoma-protective HLA class II variants correlate with increased HLA-DQB1 protein expression. Genes Immun. 2013 Dec 5;

3. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013 Jan 10;92(1):126–30.

4. Christine F Skibola, Sonja I Berndt, Joseph Vijai, Lucia Conde, Zhaoming Wang, Meredith Yeager, et al. Genome-wide association study identifies new follicular lymphoma susceptibility loci. Nat Genet. Submitted;

5. Ai WZ, Hou J-Z, Zeiser R, Czerwinski D, Negrin RS, Levy R. Follicular lymphoma B cells induce the conversion of conventional CD4+ T cells to T-regulatory cells. Int J Cancer J Int Cancer. 2009 Jan 1;124(1):239–44.

6. Voo KS, Foglietta M, Percivalle E, Chu F, Nattamai D, Harline M, et al. Selective targeting of Toll-like receptors and OX40 inhibit regulatory T cell function in follicular lymphoma: Inhibition of regulatory T cells in follicular lymphoma. Int J Cancer. 2014 Apr;n/a–n/a.

7. Umetsu DT, Esserman L, Donlon TA, DeKruyff RH, Levy R. Induction of proliferation of human follicular (B type) lymphoma cells by cognate interaction with CD4+ T cell clones. J Immunol Baltim Md 1950. 1990 Apr 1;144(7):2550–7.

8. Dave SS, Wright G, Tan B, Rosenwald A, Gascoyne RD, Chan WC, et al. Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. N Engl J Med. 2004 Nov 18;351(21):2159–69.

9. Smeltzer J, Jones J, Ziesmer SC, Grote DM, Xiu B, Ristow K, et al. Pattern of CD14+ follicular dendritic cells and PD1+ T cells independently predicts time to transformation in follicular lymphoma. Clin Cancer Res [Internet]. 2014 Apr 11 [cited 2014 May 1]; Available from: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-13-2367

10. Lin Y, Gustafson MP, Bulur PA, Gastineau DA, Witzig TE, Dietz AB. Immunosuppressive CD14+HLA-DRlow/- monocytes in B-cell non-Hodgkin lymphoma. Blood. 2011 Jan 20;117(3):872–81.