Fine-mapping of HLA associations with chronic lymphocytic leukemia in US populations

Loren Gragert,1 Stephanie Fingerson,1 Mark Albrecht,1 Martin Maiers,1 Matt Kalaycio,2 and Brian T. Hill2

1Bioinformatics Research, National Marrow Donor Program/Be The Match, Minneapolis, MN; and 2Hematologic Oncology and Blood Disorders, Taussig Cancer Institute, Cleveland Clinic, Cleveland, OH

Chronic lymphocytic leukemia (CLL) displays remarkable ethnic predisposition for whites, with relative sparing of African-American and Asian populations. In addition, CLL displays among the highest familial predispositions of all hematologic malignancies, yet the genetic basis for these differences is not clearly defined. The highly polymorphic HLA genes of the major histocompatibility complex play a central role in immune surveillance and confer risk for autoimmune and infectious diseases and several different cancers, the role for which in the development of CLL has not been extensively investigated. The National Marrow Donor Program/Be The Match has collected HLA typing from CLL patients in need of allogeneic hematopoietic stem cell transplant and has recruited millions of volunteers to potentially donate hematopoietic stem cells. HLA genotypes for 3491 US white, 397 African-American, and 90 Hispanic CLL patients were compared with 50,000 controls per population from the donor registry. We identified several HLA alleles associated with CLL susceptibility in each population, reconfirming predisposing roles of HLA-A*02:01 and HLA-DRB4*01:01 in whites. Associations for haplotype DRB4*01:01−DRB1*07:01−DQB1*03:03 were replicated across all 3 populations. These findings provide a comprehensive assessment of the role of HLA in the development of severe CLL. (Blood. 2014;124(17):2657-2665)

Introduction

Chronic lymphocytic leukemia (CLL) is the most common form of leukemia in the United States and Europe1 and is characterized by clonal proliferation of malignant B lymphocytes.2 There is significant heterogeneity in disease characteristics and prognosis for CLL, with many patients never requiring treatment. The median age of diagnosis is 72 years, and the disease occurs more commonly in men.3 CLL occurs more commonly in Western populations, is less common in African Americans, and is relatively rare in Asians.4,6 Differences in incidence persist despite immigration,7-11 and first-degree relatives of CLL patients have a two- to sevenfold elevated risk of developing the disease,12,13 implicating genetic predisposition rather than environmental factors in the pathogenesis of CLL.

HLA plays a central role in immune surveillance, and HLA polymorphisms may impact the ability of the immune system to identify malignant cells and target them for T cell-mediated elimination.14 HLA class I proteins (HLA-A, -C, and -B) present peptides from endogenous proteins to cytotoxic T lymphocytes. HLA class II proteins (HLA-DRB3/4/5, -DRB1, and -DQB1) present peptides derived from exogenous proteins to CD4+ helper T cells. Downregulation of HLA is a well-defined mechanism of immune surveillance by viral infections and in neoplastic cells.15,16 Specifically, downregulation of HLA class I proteins has been shown in CLL cells.17 HLA gene polymorphism has been found to be associated with cancer, auto-immune, and infectious diseases.18

Key Points

- Polymorphisms in HLA genes may impact the ability of the immune system to detect malignant cells and direct T cells to eliminate them.
- Several HLA alleles and haplotypes are associated with development of chronic lymphocytic leukemia across different US populations.

Consortia have identified several single nucleotide polymorphisms (SNPs) in the major histocompatibility complex (MHC) near the HLA loci associated with CLL in genome-wide association studies (GWASs). Although the specific HLA alleles found have not yet been identified,20-22 1 recent CLL GWAS used HLA assignments imputed from SNPs.23

Classical HLA typing is best suited for HLA association studies because information from coding sequences is ascertained directly rather than imputed from linkage with nearby SNPs. However, most CLL association studies using classical HLA typing have had an inadequate number of samples to detect associations for less common alleles. Most of these studies have relied on a German cohort of <200 patients.24 Because HLA genes are highly polymorphic, even classical HLA typing cannot easily resolve all ambiguities within...
a locus; therefore, most HLA association studies for CLL have thus only reported HLA associations at the allele family level. In addition, alleles also cannot readily be phased across loci into haplotypes experimentally.25

A large number of HLA-typed CLL cases have been accumulated by the National Marrow Donor Program (NMDP)/Be The Match from patients in need of allogeneic hematopoietic stem cell transplant (allo-SCT) since 1987. Although a minority of CLL patients ultimately undergo HLA typing in anticipation of potential allo-SCT, sampling the frequency of HLA alleles in this selected patient population can provide significant insight into severe disease biology. To address the issue of unresolved allelic ambiguity and haplotype phasing in registry HLA typing, we developed methods to incorporate this typing ambiguity into a statistical model, allowing us to report high-resolution allele, haplotype, and genotype associations. We also apply factor analysis to group highly correlated associations in the context of the high linkage disequilibrium of the HLA system.

We find several HLA alleles that are associated with an increased susceptibility to the development of severe CLL, as well as several alleles that are protective across populations.

Methods

Study population

The Institutional Review Board of the National Marrow Donor Program approved this study. Informed consent was obtained in accordance with the Declaration of Helsinki. Cases consisted of CLL patients >18 years of age at diagnosis for whom a preliminary search had been conducted with the NMDP to identify a matched unrelated donor for possible allo-SCT. A total of 50 000 controls from each self-identified race/ethnic background were selected randomly from the NMDP volunteer adult donor registry recruited since 2005. Controls were matched for patient age quartiles and gender proportions, although the maximum control age (60 years) was younger than the oldest cases due to registry recruitment policy. Both cases and controls were HLA typed using DNA-based methods (either Sanger sequence-based typing or sequence-specific oligonucleotide).25

HLA genotype information

This study used multivariate logistic regression on independent sets of cases and controls for each population to identify HLA associations that are either protective or predisposing with respect to the development of CLL.

To resolve case/control HLA typing ambiguity, we first enumerated a list of all high resolution 6-locus (A−C−B−DRB3/4/5−DRB1−DQB1) phased genotypes consistent with the individual’s DNA-based typing.26 Probability mass was then assigned to each genotype in proportion to population haplotype frequencies calculated from the NMDP registry.15 Under the assumption of Hardy-Weinberg equilibrium, genotype probabilities were derived as the product of haplotype frequencies for each possible genotype. At this point, the HLA information for each individual is contained in the distribution of possible genotypes. In the context of this imputation using population haplotype frequencies, the probability distribution at each locus is highly skewed toward a single high resolution allele pair, given the HLA typing resolution of NMDP samples.27,28

To obtain a fixed genotype for analysis, we took a realization of this probability mass function to assign a single genotype for each individual. We repeated this realization to generate 5 imputed HLA datasets, each containing a single unambiguous phased HLA genotype per individual. We then applied traditional logistic regression methods to each of these five imputed HLA datasets and combined the results using multiple imputation methods designed for statistical inference in the context of incomplete data.29

HLA variant assessments

Predisposing or protective effects with respect to CLL were identified at (1) the individual allele level (A, C, B, DRB3/4/5, DRB1, and DQB1 loci); (2) haplotype combinations of these loci; and (3) genotypes (both specific allele combinations and overall heterozygosity and homozygosity). Associations for these groupings were tested at 2 levels of HLA typing resolution: allele family and high resolution. All assessments were conditional on self-reported race (ie, run separately by race/ethnic group). DRBX*NNNN designates absence of any DRB3/4/5 gene on the chromosome.

We tested killer cell immunoglobulin-like receptor (KIR) ligand categories for HLA-B and HLA-C. HLA-B alleles were categorized as having either Bw4 or Bw6 epitopes. Bw4 epitopes were further subdivided by the amino acid at position 80 (eg, B80 for isoleucine at position 80). KIR3DL1 binds HLA-Bw4 group alleles but not HLA-Bw6 group alleles.28 For HLA-C alleles, we tested HLA-C1 or HLA-C2 KIR-binding epitopes based on the amino acid at position 80. KIR2DL2 binds HLA-C1 group alleles, whereas KIR2DL1 binds HLA-C2 group alleles.29

Statistical analysis

The odds ratio (OR) of each HLA variant (listed in assessments) was estimated using logistic regression for the 5 imputed HLA datasets; reported ORs were adjusted for the covariates of age (continuous), gender, and geographical location (4 levels: west, east, midwest, and south) using an additive model specification. The model was further augmented in a secondary analysis to include an interaction term between the covariate of age and the predisposing/protective effect of interest.

Statistical inference for model parameters was assessed using an F test on the test statistic derived from multiple imputation that accounts for the within-sample and between-sample variance of the 5 imputed HLA datasets.32 P values for protective and predisposing ORs reported throughout the results section were adjusted for multiple testing using a false discovery rate (FDR) method with a FDR threshold of 5%.33 P values for assessing the significance of the age and gender interaction terms were not adjusted for multiple testing; 2-tailed P < .05 was considered significant.

As a last step toward identifying the correlation structure of protective or predisposing HLA variants, factor analysis34 was applied to a reduced data set containing only the significant associations detected during analysis. Each significant association was coded with an indicator variable (present/absent) for every individual in the patient dataset (controls were excluded). Given that many HLA variants are highly correlated and, therefore, representative of a single underlying process, factor analysis was used to determine both the true number of underlying processes and group HLA variants by these underlying processes.

For all significant groupings, factor analysis loadings were used to determine the number of groups representing underlying processes and assign HLA variants to these groups. A scree plot was used to identify the number of active factors (ie, underlying processes), whereby a flattening or “knee” in the percent variation explained by sequential loadings was used to determine a cutoff for the number of active factors.35 Factors were grouped by assignment to their maximum loading, permitting the loading was >0.2; factors having weaker loadings less than this or negative loadings were assigned to an independent group.

Results

Potential allo-SCT CLL cases are enriched for men of younger age than the general CLL population

Cases consisted of 3616 US white, 413 African-American, and 97 Hispanic CLL patients. For each population, 50 000 controls were randomly selected from the NMDP registry. Baseline demographics for cases and controls are shown in Table 1. Asian/Pacific Islander patients were excluded from this study due to inadequate sample size (28) and high population substructure. Among whites, the median age of cases was 51 years (range, 18-76 years), which is significantly lower than the median age of CLL, which is 72 years. The median age
of controls was 51 years (range, 18-60 years). The US geographic distributions for cases and controls are also shown and were similar between the groups. The gender distribution of cases was 74% men, whereas the controls were 75% men. This ratio is slightly higher than the rate of 65% men for CLL in the general population.26

**US white population contains HLA alleles that are protective against development of CLL and others that confer increased risk**

In whites, the population with the largest number of cases, we found many significant HLA associations. Our association study found 28 protective or predisposing high-resolution alleles for CLL in whites (Figure 1A). Twelve alleles were protective (A*01:01, C*05:01, C*07:01, C*16:01, B*27:05, DRB1*01:01, DRB1*04:03, DRB1*13:01, DQB1*03:01, DQB1*05:01, DQB1*06:03, and DQB1*06:04) and 16 predisposing (A*02:01, C*05:01, C*07:01, C*16:02, B*14:01, B*15:01, DRB4*01:01, DRB1*04:01, DRB1*04:02, DRB1*07:01, DRB1*08:01, DQB1*03:02, DQB1*03:03, DQB1*04:02, and DQB1*05:04). The frequency of each individual HLA allele in controls and cases is indicated, as are ORs and P values.

A full listing of significantly associated HLA alleles and haplotypes, grouped using factor analysis, is in supplemental Table 1 available on the Blood Web site. The most predisposing HLA allele was DQB1*05:04 (OR = 5.62, P = 5.69 x 10^{-4}), whereas the allele with the greatest protective effect was DRB1*04:03 (OR = 0.67, P = .043).

We also identified many HLA alleles that had suggestive associations for CLL but did not reach statistical significance when the FDR was applied to consider multiple testing (supplemental Table 2).

**Antigen level analysis reveals additional associations**

Because many HLA alleles are infrequent, some associations can only observed when alleles are grouped into allele families. These allele families were originally derived from serologic antigen groups, with new alleles named based on sequence similarity with already reported alleles.20 Although the vast majority of allele family level associations were highly correlated with other high-resolution associations (supplemental Table 1), we were able to detect some haplotype associations where no high-resolution variant within the allele family reached significance; for example, A*02–C*03–B*15–DRB4*01–DRB1*04–DQB1*03 in whites (OR = 1.35, P = .025). Examination of allele family level associations also allows for more direct comparison with previous reports.

Increased statistical power may be achieved when grouping several low-frequency alleles that have similar impact on disease. However, some strong high-resolution allelic associations are not detected at the antigen level because of opposing effects of alleles within the same antigen group; for example, DRB1*04 in whites contains the predisposing alleles DRB1*04:01 (OR = 1.15) and DRB1*04:02 (OR = 1.63) but also a protective allele DRB1*04:03 (OR = 0.67).

**Novel factor analysis groups correlated alleles and associated haplotypes**

We applied factor analysis in an attempt to automatically group the multitude of significantly associated HLA alleles and haplotypes that are highly correlated by linkage disequilibrium. Considering these groupings, we can identify as many as 47 potentially independent associations for CLL in whites and 6 in African Americans. The longest extended haplotype associations within each factor analysis group are listed in Tables 2-4, with a full listing of allele and haplotype associations provided in supplemental Table 1. For example, the first factor analysis grouping for whites included 35 different allele and haplotype associations, all contained in the extended haplotype A*26:01–C*12:03–B*38:01–DRB4*01:01–DRB1*04:02–DQB1*03:02. Each factor analysis group tends to represent a single haplotype, and the groups often include some of the constituent associated alleles if those alleles are not frequently found on other haplotypes.

Linkage disequilibrium between specific HLA alleles is often very high, making it extremely challenging to distinguish causative HLA alleles from other alleles that are simply linked on the sample haplotype with the true causative allele. Often there are not sufficient cases in which one allele is present, and another allele is absent to separate the individual contribution of alleles. Haplotype-level associations were often more extreme than allele level; for example, A*02:01–B*15:01–DRB1*04:01 had an OR of 1.41, but A*02:01, B*15:01, and DRB1*04:01 had lower ORs of 1.20, 1.26, and 1.15, respectively (supplemental Table 1).

**DRB4*01:01–DRB1*07:01–DQB1*03:03 is a universally predisposing haplotype for CLL in whites, African Americans, and Hispanics**

We identified an HLA class II haplotype, DRB4*01:01–DRB1*07:01–DQB1*03:03, with a strongly predisposing impact in whites (OR = 1.49, P = 1.79 x 10^{-7}), African Americans (OR = 28.03, P < 2 x 10^{-16} at 2-digit resolution), and Hispanics (OR = 13.86, P = 9.59 x 10^{-8}) (supplemental Table 1). The frequency of this haplotype is 3.4% in whites, 0.3% in African Americans, and 1.2% in Hispanics.19 The DRB4*01:01 and DQB1*03:03 alleles on this same haplotype were also found to be predisposing in both whites (OR = 1.17, P = 3.94 x 10^{-5} and OR = 1.40, P = 1.85 x 10^{-7},
Minority populations harbor specific HLA alleles that alter CLL risk

We also found that DRB1*09:01 (OR = 2.00, \( P = 2.31 \times 10^{-4} \)) predisposes African Americans to the development of CLL (Figure 1B). No DRB1*09:01 association was identified in whites, but the allele is in near complete linkage disequilibrium (\( D' > 0.999 \)) with DRB4*01:01,\(^{19} \) an allele found to be predisposing in both white and African-American populations. DRB1*09:01 is at 3% frequency in African Americans, but only 0.1% frequency in whites, limiting the statistical power to detect DRB1*09:01 associations across populations.

C*04:01 was found to be uniquely protective in African Americans (OR = 0.72, \( P = .025 \); Figure 1B), where the allele is at 21% frequency in controls but only 17.1% of cases.\(^{19} \) Interestingly, B*45:01 (OR = 2.98, \( P = .015 \)) and C*07:04 (OR = 5.60, \( P = 5.71 \times 10^{-4} \)) were uniquely predisposing in Hispanics (supplemental Table 1).

Haplotype common in Ashkenazi Jews is associated with an increased risk of CLL

CLL incidence has been reported to be much higher in Jewish populations than in neighboring Arab populations.\(^{37} \) We identified as predisposing the haplotype A*26:01–C*12:03–B*38:01–DRB4*01:01–DRB1*04:02–DQB1*03:02 (OR = 2.05,
<table>
<thead>
<tr>
<th>Variant</th>
<th>OR</th>
<th>Control frequency</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>FDR P</th>
<th>Factor analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A<em>26:01–C</em>12:03–B<em>38:01–DRB4</em>01:01–DRB1<em>04:02–DQB1</em>03:02</td>
<td>2.05</td>
<td>0.0035</td>
<td>1.23</td>
<td>3.41</td>
<td>1.94E-04</td>
<td>Group 1</td>
</tr>
<tr>
<td>C<em>16:01–B</em>44:03–DRB4*01:01</td>
<td>0.79</td>
<td>0.0290</td>
<td>0.63</td>
<td>0.99</td>
<td>3.05E-02</td>
<td>Group 2</td>
</tr>
<tr>
<td>A<em>01:01–C</em>07:01–B<em>08:01–DRB1</em>03:01–DQB1*02:01</td>
<td>0.83</td>
<td>0.0636</td>
<td>0.70</td>
<td>0.99</td>
<td>2.07E-02</td>
<td>Group 3</td>
</tr>
<tr>
<td>DRB3<em>02:02–DRB1</em>13:01–DQB1*06:03</td>
<td>0.79</td>
<td>0.0401</td>
<td>0.65</td>
<td>0.96</td>
<td>6.79E-03</td>
<td>Group 4</td>
</tr>
<tr>
<td>B<em>52:01–DQB5</em>01:02–DRB1<em>15:02–DQB1</em>06:01</td>
<td>1.54</td>
<td>0.0064</td>
<td>1.05</td>
<td>2.26</td>
<td>1.45E-02</td>
<td>Group 5</td>
</tr>
<tr>
<td>A<em>02–C</em>03–B*15</td>
<td>1.27</td>
<td>0.0280</td>
<td>1.01</td>
<td>1.58</td>
<td>2.13E-02</td>
<td>Group 6</td>
</tr>
<tr>
<td>DRB<em>X</em>NNN–DRB1<em>01:01–DQB1</em>05:01</td>
<td>0.84</td>
<td>0.0808</td>
<td>0.73</td>
<td>0.97</td>
<td>5.36E-03</td>
<td>Group 7</td>
</tr>
<tr>
<td>A<em>02–C</em>03–B<em>15–DRB4</em>01–DRB1<em>04–DQB1</em>03</td>
<td>1.35</td>
<td>0.0142</td>
<td>1.02</td>
<td>1.78</td>
<td>2.55E-02</td>
<td>Group 8</td>
</tr>
<tr>
<td>C<em>02:02–B</em>27:05–DRB3*02:02</td>
<td>0.43</td>
<td>0.0047</td>
<td>0.20</td>
<td>0.92</td>
<td>1.59E-02</td>
<td>Group 9</td>
</tr>
<tr>
<td>C<em>02–B</em>27–DRB4<em>01–DRB1</em>04–DQB1*03</td>
<td>0.57</td>
<td>0.0071</td>
<td>0.34</td>
<td>0.95</td>
<td>2.25E-02</td>
<td>Group 10</td>
</tr>
<tr>
<td>C<em>03:04–DRB4</em>01:01</td>
<td>1.18</td>
<td>0.0416</td>
<td>1.00</td>
<td>1.40</td>
<td>4.23E-02</td>
<td>Group 11</td>
</tr>
<tr>
<td>C<em>05:01–B</em>44:02–DRB1<em>11:01–DQB1</em>03:01</td>
<td>0.48</td>
<td>0.0042</td>
<td>0.23</td>
<td>0.99</td>
<td>4.45E-02</td>
<td>Group 12</td>
</tr>
<tr>
<td>C<em>05:01–B</em>44:02–DRB1<em>13:01–DQB1</em>06:03</td>
<td>0.62</td>
<td>0.0083</td>
<td>0.39</td>
<td>0.99</td>
<td>4.04E-02</td>
<td>Group 13</td>
</tr>
<tr>
<td>DRB3<em>03:01–DRB1</em>13:02–DQB1*06:04</td>
<td>0.82</td>
<td>0.0363</td>
<td>0.67</td>
<td>0.99</td>
<td>3.94E-02</td>
<td>Group 14</td>
</tr>
<tr>
<td>DRB5<em>01–DRB1</em>15–DQB1*05:01</td>
<td>4.80</td>
<td>0.0007</td>
<td>2.09</td>
<td>11.03</td>
<td>1.86E-06</td>
<td>Group 15</td>
</tr>
<tr>
<td>A<em>02:01–DRB1</em>07:01</td>
<td>1.39</td>
<td>0.0314</td>
<td>1.13</td>
<td>1.70</td>
<td>9.57E-06</td>
<td>Group 16</td>
</tr>
<tr>
<td>C<em>01–B</em>27–DRB<em>X</em>NN–DRB1*01</td>
<td>0.62</td>
<td>0.0110</td>
<td>0.41</td>
<td>0.94</td>
<td>1.16E-02</td>
<td>Group 17</td>
</tr>
<tr>
<td>C<em>03–B</em>15–DRB4<em>01–DRB1</em>04–DQB1*03</td>
<td>1.37</td>
<td>0.0231</td>
<td>1.10</td>
<td>1.71</td>
<td>3.09E-04</td>
<td>Group 18</td>
</tr>
<tr>
<td>DRB3<em>02–DRB1</em>11–DQB1*03</td>
<td>0.89</td>
<td>0.0107</td>
<td>0.80</td>
<td>1.00</td>
<td>4.00E-02</td>
<td>Group 19</td>
</tr>
<tr>
<td>DRB4<em>01–DRB1</em>04–DQB1*02</td>
<td>2.02</td>
<td>0.0012</td>
<td>1.04</td>
<td>3.92</td>
<td>3.17E-02</td>
<td>Group 20</td>
</tr>
<tr>
<td>A*02:01</td>
<td>1.20</td>
<td>0.2670</td>
<td>1.08</td>
<td>1.34</td>
<td>3.29E-02</td>
<td>Group 21</td>
</tr>
<tr>
<td>C<em>03:03–B</em>15:01–DRB4*01:01</td>
<td>1.68</td>
<td>0.0057</td>
<td>1.11</td>
<td>2.52</td>
<td>2.39E-03</td>
<td>Group 22</td>
</tr>
<tr>
<td>C<em>12:03–B</em>35:03</td>
<td>1.75</td>
<td>0.0027</td>
<td>1.06</td>
<td>2.86</td>
<td>1.78E-02</td>
<td>Group 23</td>
</tr>
<tr>
<td>C<em>07:01–B</em>18:01</td>
<td>0.73</td>
<td>0.0216</td>
<td>0.56</td>
<td>0.95</td>
<td>8.97E-03</td>
<td>Group 24</td>
</tr>
<tr>
<td>C<em>01:02–B</em>27:05–DRB<em>X</em>NNN–DRB1*01</td>
<td>0.60</td>
<td>0.0090</td>
<td>0.36</td>
<td>0.99</td>
<td>4.23E-02</td>
<td>Group 25</td>
</tr>
</tbody>
</table>

The longest extended haplotype association for each factor analysis group is listed. When multiple haplotypes had the same number of loci within the same factor analysis group, the most significant association is provided. When no haplotypes were contained in a factor analysis group, the most significant allele association is provided. Lower 95% CI and upper 95% CI indicate the ORs for the 95% CI. FDR P value is the significance after FDR correction for multiple testing. A complete listing of allele and haplotype associations for each factor analysis group with additional information on uncorrected P values, case frequencies, and interactions for age, geography, and gender is provided in supplemental Table 1. CI, confidence interval.

The frequency of cases is provided after each haplotype variant in parentheses. For personal use only.

From www.bloodjournal.org by guest on October 28, 2017. For personal use only.
individual HLA alleles. Several alleles were found to have differing associations by age (supplemental Table 1). For example, A*23:01 has a protective association in only the whites <50 years of age (OR = 0.616, \(P = .014\)). Stratified ORs for variants with age interactions are reported in supplemental Table 3.

In addition, because of the well-documented male predominance of CLL, we sought to determine if gender was associated with any of the specific HLA alleles. Several alleles were found to have differing associations by gender (supplemental Table 1). B*27:05 may lead to lower CLL risk in women (OR = 0.55, \(P = .00056\)) than in men. Stratified ORs for variants with age interactions are reported in supplemental Table 4.

### Homozygosity at HLA class I is associated with development of CLL

We found overall homozygosity at all 3 HLA class I loci was associated with increased risk of CLL in whites: HLA-A (OR = 1.19, \(P = 1.67 \times 10^{-3}\)), HLA-B (OR = 1.16, \(P = .024\)), and HLA-C (OR = 1.16, \(P = .0076\); Table 5). One specific homozygote genotype, A*02:01\(^a\)+A*02:01\(^b\), was found to have increased risk in both whites (OR = 1.26, \(P = .0059\)) and Hispanics (OR = 1.84, \(P = .041\) at 2-digit resolution). In addition, we found several other specific HLA genotype associations (supplemental Table 5).

### Discussion

In this study, we reconfirmed several HLA associations seen in previous CLL genetic association studies and identified many others that were previously unobserved due to low statistical power. The large number of disease cases and donor controls across multiple populations coupled with high-quality classical HLA typing make the NMDP dataset a powerful source for inference of lower-frequency haplotype and genotype-level HLA associations with severe CLL for the first time.

GWASs for whites with CLL have previously identified associations with several SNPs in the HLA region. Because these SNPs are often in high linkage disequilibrium with specific HLA alleles, using software such as HLA-IMPAIR, \(^{59}\) it is possible to impute HLA alleles from GWAS SNP data. We reconfirm with very high confidence a A*02:01 association from a GWAS study of 517 CLL cases and additional replication study specifically typing the associated rs6904029 SNP, \(^{23}\) finding close agreement in ORs, with \(OR = 1.32\) in the study by Di Bernardo et al \(vs\) 1.20 in our study. A*02:01 is the most common HLA-A allele in whites, with a frequency of 26.7% in the general population \(^{19}\) and 30.4% in whites with severe CLL.

GWAS using GEC/Interlymph Consortium samples have yielded 6 SNP associations in the HLA region, \(^{20-22}\) but we are unable to compare results because these studies did not impute HLA alleles from the SNP data. HLA imputation, or more ideally classical HLA typing, on the GEC/Interlymph Consortium samples would enable adequate replication of our study. The GWAS cases included older patients and those with less severe disease than those included in this study based on data collected from the NMDP for a cohort of CLL patients who have likely failed prior therapies and were in need of allo-SCT. Candidate genes involved in apoptosis and hematopoetic cell homeostasis were identified as conferring increased risk in these GWAS.

Findings from previous CLL association studies using classical HLA typing are often consistent with our results. The first

### Table 3. Extended HLA haplotype associations from factor analysis for African Americans

<table>
<thead>
<tr>
<th>Variant</th>
<th>OR</th>
<th>Control frequency</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>FDR P</th>
<th>Factor analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB4<em>01:01–DRB1</em>09:01–DQB1*02:01</td>
<td>2.24</td>
<td>0.084</td>
<td>1.00</td>
<td>4.99</td>
<td>7.1E-02</td>
<td>Group 4</td>
</tr>
<tr>
<td>DRB5<em>01:01–DRB1</em>13:02–DQB1*06:04</td>
<td>1.07</td>
<td>0.211</td>
<td>0.54</td>
<td>0.98</td>
<td>2.46E-02</td>
<td>Group 3</td>
</tr>
<tr>
<td>DRB5<em>01–DRB1</em>07–DQB1*03</td>
<td>2.83</td>
<td>0.039</td>
<td>9.64</td>
<td>81.50</td>
<td>0.00E+00</td>
<td>Group 1</td>
</tr>
<tr>
<td>C*04:01</td>
<td>0.72</td>
<td>0.211</td>
<td>0.54</td>
<td>0.98</td>
<td>2.46E-02</td>
<td>Group 4</td>
</tr>
<tr>
<td>DRB4*01</td>
<td>1.37</td>
<td>0.007</td>
<td>1.04</td>
<td>1.80</td>
<td>1.39E-02</td>
<td>Group 5</td>
</tr>
<tr>
<td>A<em>03–DRB1</em>07</td>
<td>2.24</td>
<td>0.084</td>
<td>1.00</td>
<td>4.99</td>
<td>7.1E-02</td>
<td>Group 6</td>
</tr>
</tbody>
</table>

The longest extended haplotype association for each factor analysis group is listed. When multiple haplotypes had the same number of loci within the same factor analysis group, the most significant association is provided. When no haplotypes were contained in a factor analysis group, the most significant allelic association is provided. Lower 95% CI and upper 95% CI indicate the ORs for the 95% CI. FDR \(P\) value is the significance after FDR correction for multiple testing. A complete listing of allele and haplotype associations for each factor analysis group with additional information on uncorrected \(P\) values, case frequencies, and interactions for age, geography, and gender is provided in supplemental Table 1.

### Table 4. Extended HLA haplotype associations from factor analysis for Hispanics

<table>
<thead>
<tr>
<th>Variant</th>
<th>OR</th>
<th>Control frequency</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>FDR P</th>
<th>Factor analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>C<em>04:01–B</em>04:03–DRB4<em>01:01–DRB1</em>07:01–DQB1*02:01</td>
<td>4.18</td>
<td>0.007</td>
<td>1.48</td>
<td>11.78</td>
<td>3.93E-03</td>
<td>Group 1</td>
</tr>
<tr>
<td>DRB4<em>01:01–DRB1</em>07:01–DQB1*03:03</td>
<td>13.86</td>
<td>0.003</td>
<td>4.15</td>
<td>46.30</td>
<td>9.59E-09</td>
<td>Group 2</td>
</tr>
<tr>
<td>C<em>04:01–B</em>04:03</td>
<td>3.47</td>
<td>0.009</td>
<td>1.16</td>
<td>10.37</td>
<td>1.70E-02</td>
<td>Group 5</td>
</tr>
<tr>
<td>B<em>03–DRB4</em>01–DRB1*07</td>
<td>5.76</td>
<td>0.006</td>
<td>2.07</td>
<td>16.03</td>
<td>1.27E-04</td>
<td>Group 4</td>
</tr>
<tr>
<td>A<em>24:02–DRB1</em>07:01</td>
<td>4.27</td>
<td>0.008</td>
<td>1.65</td>
<td>11.08</td>
<td>1.29E-03</td>
<td>Group 1</td>
</tr>
<tr>
<td>C<em>07–DRB5</em>01</td>
<td>2.25</td>
<td>0.040</td>
<td>1.04</td>
<td>4.86</td>
<td>2.96E-02</td>
<td>Group 6</td>
</tr>
<tr>
<td>C*07:04</td>
<td>5.60</td>
<td>0.006</td>
<td>1.67</td>
<td>18.74</td>
<td>5.71E-04</td>
<td>Group 7</td>
</tr>
</tbody>
</table>

The longest extended haplotype association for each factor analysis group is listed. When multiple haplotypes had the same number of loci within the same factor analysis group, the most significant association is provided. When no haplotypes were contained in a factor analysis group, the most significant allelic association is provided. Lower 95% CI and upper 95% CI indicate the ORs for the 95% CI. FDR \(P\) value is the significance after FDR correction for multiple testing. A complete listing of allele and haplotype associations for each factor analysis group with additional information on uncorrected \(P\) values, case frequencies, and interactions for age, geography, and gender is provided in supplemental Table 1.
predisposing HLA association found with CLL using DNA-based typing was reported by British investigators. We reconfirm the A2→B62→DR4 haplotype reported (OR = 4.1) in our data at high resolution as A*02:01→B*15:01→DRB1*04:01 (OR = 1.41, \(P = 0.0049\)). A predisposing C*16 association (OR = 2.69) in a study of Southeastern Spanish patients was also seen in US whites for C*16:02 (OR = 1.83, \(P = 0.0062\)); however, C*16:01 was protective (OR = 0.79, \(P = 0.0098\)).

Recent studies have suggested that high levels of HLA homozygosity is associated with the development of CLL and DCLCL, which we also report for all HLA class I loci. Although consanguinity among cases cannot be formally assessed, it would not account for the magnitude of this observed risk (OR = 1.16-1.19).

The predisposing association for DR53 (DRB4) identified in whites by Dorak et al was also replicated as DRB4*01:01 (OR = 1.17, \(P = 3.95 \times 10^{-5}\)), whereas the protective DR52 (DRB3) association was observed in several HLA class II haplotypes: DRB3*03:01→DBR1*13:02→DQB1*06:04 (OR = 0.82, \(P = 0.039\)), DRB3*02:02→DRB1*13:01→DQB1*06:03 (OR = 0.79, \(P = 0.0068\), and DRB3*02→DRB1*11→DQB1*03 (OR = 0.89, \(P = 0.04\). However, a predisposing DRB3-containing haplotype, DRB3*02→DRB1*11→DQB1*06 (OR = 8.14, \(P = 2.68 \times 10^{-8}\)), was also observed. The DRB4 locus was also found on multiple predisposing haplotypes containing alleles in the DRB1*04, DRB1*07, and DRB1*09 allele families. However, among studied populations, DRB1*09 was only frequent enough in African Americans to detect a predisposing effect for that of the DRB4*01:01→DRB1*09:01 haplotype. Looking across populations, we observed a universal association for the DRB4*01:01→DRB1*07:01→DQB1*03:03 haplotype in whites, Hispanics, and African Americans. Concordance in associations across these populations suggests similar disease etiology in all ethnic groups studied.

The novel application of factor analysis to this HLA association study gives an automated determination of the number of unique HLA haplotypes associated with CLL. Allele-level ORs were often less extreme than haplotype level effects, which may either result from additive effects of ORs from individual constituent alleles, or it may be that certain alleles are also in linkage disequilibrium with alleles on other loci that have opposing impact on disease risk. The factor analysis groupings suggest that many HLA allele associations are best described in the context of their membership within a constellation of specific haplotypes rather than as independent findings. Our hybrid approach combining the advantages of both high-resolution allele and antigen level associations with factor analysis allowed us to avoid any disadvantages that could result from choosing to analyze at any particular level of resolution over another.

Alleles that more avidly bind a wider range of peptides may protect from infectious diseases and cancers while predisposing for development of autoimmune diseases. We identified the B*27:05 allele as the second most protective allele for CLL (OR = 0.77). B*27:05 has been found to be predisposing for ankylosing spondylitis, whereas it is protective for HIV progression to AIDS, and now for CLL.

Dorak et al reported that the A1→B8→DR3 haplotype was identified as trending toward significantly protective, which was confirmed in our study as A*01:01→C*07:01→B*08:01→DRB1*03:01→DQB1*02:01 (OR = 0.83, \(P = 0.02\)). However, a recent CLL association result from MD Anderson indicated that the A1→C7→B8 haplotype was predisposing for CLL. We note the MD Anderson dataset included only patients that proceeded to allogeneic transplant, which likely biases cases toward more common HLA alleles, whereas the NMDP dataset included patients that did not identify a matched donor and/or did not proceed to transplant.

Differences in CLL incidence may also be caused by allelic differences in downregulation of HLA expression. Maintenance of HLA-Bw4 allelic expression is beneficial for leukemic cells because these HLA alleles provide an inhibitory signal to natural killer cells via the KIR3DL1 ligand. In CLL, HLA alleles in the Bw4 group are selectively downregulated less often than Bw6 alleles, protecting leukemic cells from natural killer-mediated lysis and increasing presentation of CLL-specific peptides to cytotoxic T lymphocytes. We found Bw4 to be protective with threonine at position 80 (OR = 0.88, \(P = 0.0013\)). In ovarian cancer, downregulation of A*02 has been associated with poor prognosis. Together, these observations suggest that immune evasion by HLA downregulation may be a mechanism by which specific HLA alleles confer an increased risk.

Incidence of CLL varies widely among world populations, occurring at higher rates in whites and found least often in populations of the Far East. For Hispanic and Asian populations, far fewer cases were available in the NMDP dataset than would be expected by overall patient population proportions, consistent with observations that CLL is seen at much lower incidence compared with whites. Due to low sample sizes, there has been no assessment of the role of HLA in the development of CLL among world populations; therefore, global CLL sample collection efforts and studies are necessary. Combining whole exome sequencing and HLA peptide binding analysis is a promising approach to identify interactions between potentially differing population frequencies of CLL-specific somatic mutations and presentation of resulting tumor neoantigens by HLA alleles, which are already known to differ in frequency across populations.

With an incidence rate ratio of 0.73, CLL is less common in African Americans compared with the US white population, which has an incidence ratio of 1.89. In a recent review of the clinical course of African Americans with CLL, investigators from the MD Anderson Cancer Center and Duke University Medical Center reported that African Americans tend to have more unfavorable clinical and biologic characteristics, an earlier need of treatment, fewer remissions, and inferior overall survival compared with nonblack patients.

Despite a lower overall incidence of CLL in African Americans, the somewhat more aggressive clinical course suggests that these differences may not be accountable strictly by differences in socioeconomic status. We report for the first time that the HLA alleles predisposing to the development of CLL in African Americans are different than those for whites, and this may have implications for mechanism of evasion of immune surveillance that normally functions to suppress burgeoning malignancies. The predisposing DRB4*01:01 is less common in African Americans

### Table 5. Locus-level HLA associations for overall heterozygosity and homozygosity in whites

<table>
<thead>
<tr>
<th>Variant</th>
<th>OR</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*HOMO</td>
<td>1.19</td>
<td>1.07</td>
<td>1.33</td>
<td>0.0016</td>
</tr>
<tr>
<td>A*HETERO</td>
<td>0.84</td>
<td>0.76</td>
<td>0.92</td>
<td>0.0016</td>
</tr>
<tr>
<td>C*HOMO</td>
<td>1.16</td>
<td>1.02</td>
<td>1.31</td>
<td>0.0077</td>
</tr>
<tr>
<td>C*HETERO</td>
<td>0.86</td>
<td>0.77</td>
<td>0.96</td>
<td>0.0077</td>
</tr>
<tr>
<td>B*HOMO</td>
<td>1.16</td>
<td>1.02</td>
<td>1.32</td>
<td>0.024</td>
</tr>
<tr>
<td>B*HETERO</td>
<td>0.86</td>
<td>0.74</td>
<td>1.00</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Lower 95% CI and upper 95% CI indicate the ORs for the 95% CI. HOMO, associations for all homozygous genotypes at each locus; HETERO, all heterozygous genotypes.
(18.2% allele frequency) than whites (29.7%), which may explain some of the difference in incidence. In addition, C*04:01 is uniquely protective in African Americans and found at the relatively high frequency (21%) in this population.

In conclusion, we report the largest HLA association study of CLL to date. We find several HLA alleles that predispose to the development of severe disease in whites and African Americans, as well as others that are protective. Future studies are needed to confirm these HLA associations among unselected cases that would include those with less severe CLL and patients unfit for allo-SCT due to advanced age. Additional controls over age 60 could improve adjustment for age. Correlation of HLA alleles with cancer cytogenetics that were unavailable in this dataset could reveal prognostic indicators for progression of CLL to more severe disease.53,55

Although knowledge of HLA associations in cancer does not immediately suggest clinical intervention, these findings may direct further study of the role of the adaptive immune response in against CLL including both allogeneic stem cell transplantation and adoptive immunotherapy through use of chimeric antigen receptor T cells.56 Identifying which proteins are mutated or differentially expressed in CLL cells and which endogenous peptides derived from these cancer-related proteins are presented to T cells by different HLA alleles50,57 is an important step in identification of specific peptide-HLA complexes that could be targeted for personalized immunotherapy.58

References

3. Shanafelt TD, Geyer SM, Kay NE. Prognostic indicators for progression of CLL to more severe disease.54,55
Fine-mapping of HLA associations with chronic lymphocytic leukemia in US populations

Loren Gragert, Stephanie Fingerson, Mark Albrecht, Martin Maiers, Matt Kalaycio and Brian T. Hill