

## Toward understanding MHC disease associations: Partial resequencing of 46 distinct HLA haplotypes

Wade P. Smith<sup>a</sup>, Quyen Vu<sup>a</sup>, Shuying Sue Li<sup>b</sup>, John A. Hansen<sup>a</sup>, Lue Ping Zhao<sup>b</sup>,  
Daniel E. Geraghty<sup>a,\*</sup>

<sup>a</sup> *Clinical Research Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA*

<sup>b</sup> *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109-1024, USA*

Received 31 August 2005; accepted 29 November 2005

Available online 23 January 2006

### Abstract

We carried out a resequencing project that examined 552 kb of sequence from each of 46 individual HLA haplotypes representing a diversity of HLA allele types, generating nearly 27 Mb of fully phased genomic sequence. Haplotype blocks were defined extending from telomeric of HLA-F to centromeric of HLA-DP including in total 5186 MHC SNPs. To investigate basic questions about the evolutionary origin of common HLA haplotypes, and to obtain an estimate of rare variation in the MHC, we similarly examined two additional sets of samples. In 19 independent HLA-A1, B8, DR3 chromosomes, the most common HLA haplotype in Northern European Caucasians, variation was found at 11 SNP positions in the 3600-kb region from HLA-A to DR. Partial resequencing of 282 individuals in the gene-dense class III region identified significant variability beyond what could have been detected by linkage to common SNPs.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Genes; MHC classes I and II; Haplotypes; Major histocompatibility complex; Linkage disequilibrium; Polymorphism, single nucleotide; Genomics; Sequence analysis, DNA

Genetic and evolutionary studies of the immune response have often centered on the major histocompatibility complex (MHC), comprising the class I, II, and III regions, and including the class I and II genes that encode the human leukocyte antigen (HLA) molecules involved in immune recognition [1]. In addition to these fundamental immune response genes, the MHC region contains over 140 genes, about 40% of which function in a variety of ways contributing to both innate and adaptive immunity [2]. For the past several decades, the unusual levels of polymorphism at the HLA class I and II loci have facilitated a large number of studies of the relationships between genetic variability among specific MHC loci and interesting immunological phenotypes related to autoimmunity, transplantation, and infectious disease [3–7]. Despite many significant efforts in search of the causative relationship class I or II polymorphism might have with disease, none of the highly

associated HLA alleles have been demonstrated directly as causative of disease. This is most likely due to the fact that the HLA class I and class II genes are not always necessary and are never sufficient to cause disease (e.g., most A1, B8, DR3-carrying individuals will not develop type 1 diabetes), and therefore variability at other loci, within the MHC or elsewhere in the genome, must contribute to disease phenotype [8].

With the allelic polymorphism among the MHC class I and II genes now being very well defined [9], attention has turned to the other MHC genes and also to whole MHC sequences for further information on new variability relevant to immunity and disease [10,11]. Several haplotype maps of the human MHC have been described recently [12,13], characterizing the linkage disequilibrium relationships among several thousand single-nucleotide polymorphisms (SNPs) and adding cohesiveness to our knowledge of the underlying linkage disequilibrium (LD) block structure long recognized within the human MHC [14,15]. Using haplotype-tagged SNPs [16,17], it is possible to sample a subset of SNPs within a haplotype block and derive

\* Corresponding author.

E-mail address: [geraghty@fhcrc.org](mailto:geraghty@fhcrc.org) (D.E. Geraghty).

information about all the SNPs in that block [18]. If haplotype-tagged SNPs across the MHC can be identified at sufficient density it should be possible to perform association and linkage analysis in comparisons with disease and controls to detect common variants outside of HLA class I and II loci that are associated with disease. Exactly such studies are now being carried out in several labs assuming the common disease/common variant hypothesis will hold true in these searches [19]. These searches are usually limited to SNP typing technology, which cannot detect previously unknown variation. An equally plausible hypothesis recognizes that there are a much larger number of rare SNPs in the genome and although it is usually not possible to associate them individually with disease, it may be possible to associate the aggregate of rare SNPs. Indeed, exactly such a case has been found in the analysis of obesity in which direct sequencing of well-chosen candidate genes led to the identification of numerous rare alleles causing both syndromic and common obesity [20]. Similarly, rare DNA sequence variants were shown to contribute collectively to variation in plasma levels of high-density lipoprotein cholesterol [21]. Although the possibility of detecting rare variants using common variant tagged SNPs and SNP typing technology has been raised [22], such methods may currently apply only in the hypothetical due to practical constraints in predicting haplotypes with certainty.

These considerations provided further motivation to understand better the genetic diversity of the MHC and to develop tools that could be routinely used to identify functional variants in the MHC other than the class I and II loci. Derivative data and tools could then be used both to aid clinical studies and to advance our understanding of the evolutionary history of humans through comparative genomic analysis. Toward these ends, we established a resequencing project that examined 552 kb of MHC sequence from each of 48 individual chromosomes representing a diversity of HLA allele types, generating nearly 27 Mb of genomic sequence. From these data we were able to build haplotype blocks extending telomeric of HLA-F to the centromeric sequences near HLA-DP defining linkage relationships among several thousand new common variants in the MHC. To answer basic questions about the evolutionary origin of disease-related MHC haplotypes and to demonstrate the effectiveness of the unique approach used for SNP genotyping, we similarly examined 19 examples of the most common MHC haplotype HLA-A1, B8, DR3, a haplotype that is known to be associated with autoimmune disease, including type 1 diabetes (T1D). The results of these studies further motivated an extended look at the variability that can be found beyond that detected using common SNPs as markers. These data and methods together are discussed in terms of their suitability for detecting polymorphism in the MHC relevant to immunological phenotypes and therefore fundamental to studies of infectious disease, autoimmunity, and transplantation.

## Results

The term haplotype has long been used to refer to collections of HLA alleles, spanning the entire MHC [23].

This term refers to the HLA genes, which are routinely typed in transplantation, and while two individuals with the same HLA alleles are said to have the same MHC haplotypes, they do not necessarily contain identical sequences throughout the MHC. The question as to the degree of similarity between chromosomes with the same MHC haplotype is an open area of research and is the same as asking how often particular HLA alleles come together through recombination and whether one contiguous MHC has been maintained through time [24]. The term haplotype can also refer to sequence identity between two individuals over a short distance. We use the term in both ways throughout and clarify the usage when necessary.

### MHC SNPs

PCR resequencing of 557 segments (amplicons) averaging 1 kb in length spread throughout the MHC yielded 552 kb of DNA sequence for each of 48 chromosomes (Fig. 1) for a total of approximately 27 million bp of MHC sequence. Nearly all of this sequence was derived from unique sequences, including gene, pseudogene, promoter, and other unique sequences of unknown function. From these data it was possible to assemble completely phased SNP data for almost all of the regions spanned by the 48 chromosomes, excepting 2% of the data from families as described under Materials and methods. In that regard, almost all of the haplotype information for the 5186 SNPs identified is deduced and not inferred. Three of the 48 chromosomes had the same HLA-A, B, C, DRB1, DQB1 genotype, and thus the total dataset represented 46 distinct HLA haplotypes, providing what is in essence a partial genomic sequence (~15% of the complete sequences, 30% of the unique sequences) of 48 different 4.0-Mb MHC haplotypes. Each of these 48 haplotypes was derived from 557 amplicons, and each amplicon in turn can be described in terms of its allelic variation of its constituent SNPs. Twenty-eight of the amplicons contained no variation at all, while the remaining amplicons contained an average of 9.7 SNPs. A simple breakdown of the distribution of these SNPs over genes, pseudogenes, and other sequences within the MHC is summarized in Table 1. With a sample size of 48 chromosomes, the 5186 SNPs represent 87% of the SNPs with a minor allele frequency at or above 1% and 99% of the SNPs present at or above 5% [25] in the human population in the regions sequenced.

All of the SNPs included in the analysis lie at least 10 kb outside of the sequences flanking the classical class I and II HLA genes. As is indicated in Fig. 1 and in agreement with previous studies the SNP density near the class I and II loci is very high [26,27]. In addition, the HLA genes themselves being the most polymorphic loci in the genome are highly adequate markers for these regions. Therefore, considering this and the technical complexities of analyzing the high-density SNPs near the class I and II genes, these adjacent SNP data were not included in this analysis (although data were acquired from some of these regions). Overall there were 9.4 SNPs per sequenced kilobase

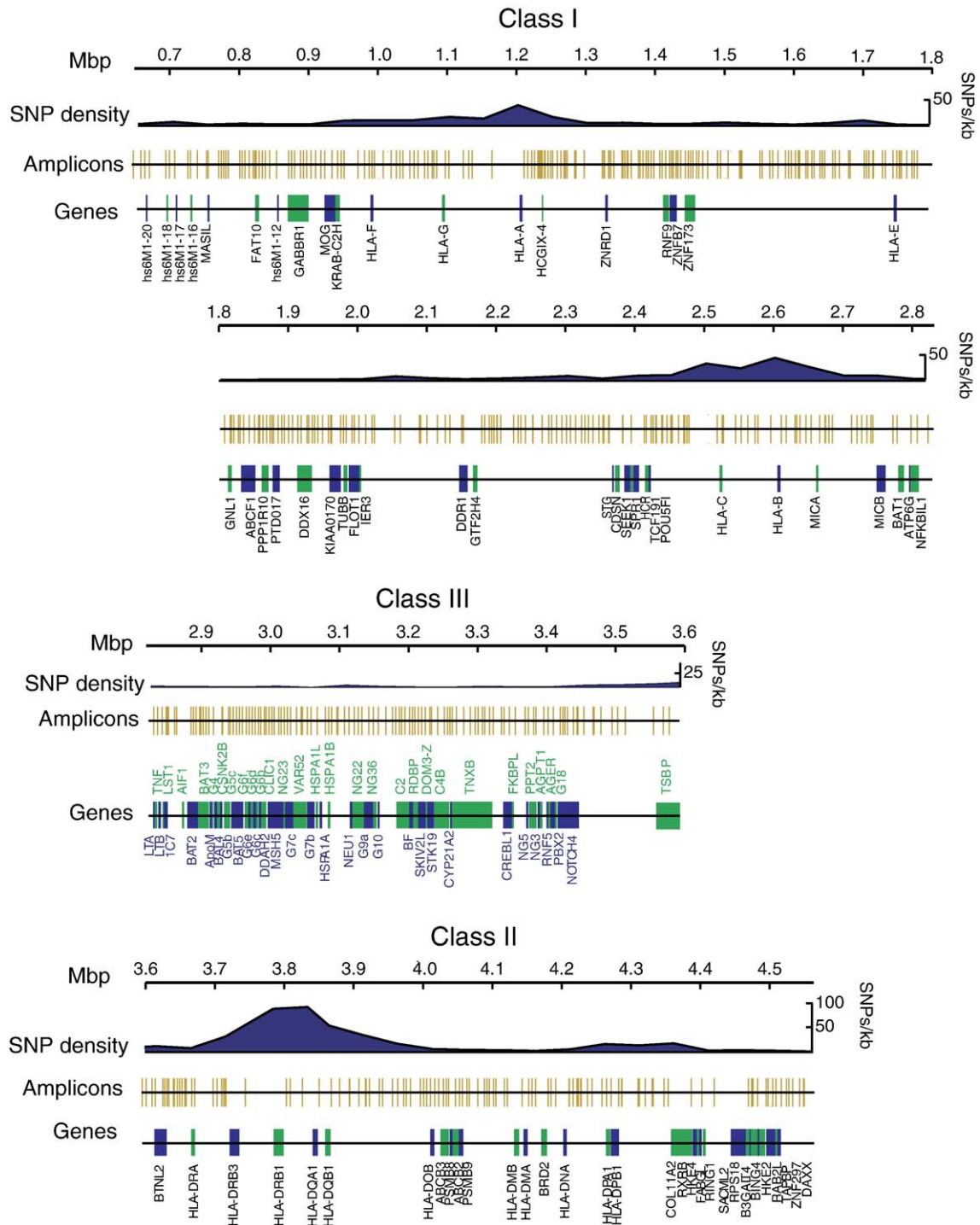


Fig. 1. An overview of the density of variant sites in MHC assembled from 48 chromosomes. Gene locations are indicated beneath each segment and MHC numbering is according to the convention established by the Sanger Center MHC consensus located at <http://www.vega.sanger.ac.uk> [10]. Vertical orange bars on the line labeled Amplicons denote location of the 557 amplicons analyzed. Each amplicon spans on average 1 kb representing about 15% of the total MHC sequence or 1 kb of sequencing per 7.1 kb. The graph on the SNP density bar represents the measured number of variant sites among 46 MHC haplotypes, with the vertical axis representing the number of SNPs/kb.

found on average. To estimate the total number of SNPs between these 48 chromosomes we need to correct for the fact that the amplicons are more closely spaced in regions of low SNP density. A simple estimate allows each amplicon to represent the number of SNPs between half the distance to the next most telomeric amplicon and half the distance to the next most

centromeric amplicon. Using this estimation we find that the average number of SNPs per kilobase throughout the MHC is 13.3 and there are a total of 51,990 SNPs between these 48 chromosomes. There was an average of 2.4 SNPs/kb identified between any two individual chromosomes among the 48 analyzed.

Table 1  
MHC SNP and amplicon distribution from 48 phased chromosomes

	SNPs	Amplicons
Genes	1188	188
Pseudogenes	921	52
Unknown function <sup>a</sup>	3077	317
Total	5186	557

<sup>a</sup> Including multispecies conserved sequences.

### Linkage disequilibrium, haplotype blocks, and tagged amplicons

To define the local MHC structure and identify the minimal number of informative amplicons capable of defining regional LD, we computed the linkage disequilibrium index (LDI) between adjacent amplicons, inferred their haplotype-block structures, and identified haplotype-tagging amplicons (Fig. 2). Because these data were fully phased, multiple SNP alleles within an amplicon were treated as a single locus with multiple alleles, allowing us to evaluate LD among amplicons

rather than individual SNPs. While  $D'$  is often used, it cannot be easily adapted to work with multiple SNP loci or multiple amplicons. Instead, we used a modified measure of linkage disequilibrium, known as the linkage disequilibrium index based upon the concept of entropy (see Materials and methods).

Upon calculating LDI we identified a set of haplotype-tagging amplicons, providing an economy that will allow the simultaneous detection of other SNPs in most segments of the MHC that are in linkage disequilibrium. The variance in block size observed was very large, extending from a few kilobases to over 150 kb, as illustrated in Fig. 2, in which it is evident that the large blocks include regions with previously well established linkage disequilibrium (e.g., HLA-DR-DQ) [28]. One interesting aspect of the block distribution is contained in the class III region where amplicons were less likely to fit into a haplotype block, resulting in shorter, sparsely distributed blocks, consistent with significant levels of recombination. A number of secondary associations with MHC-related diseases have been detected in this region [8].

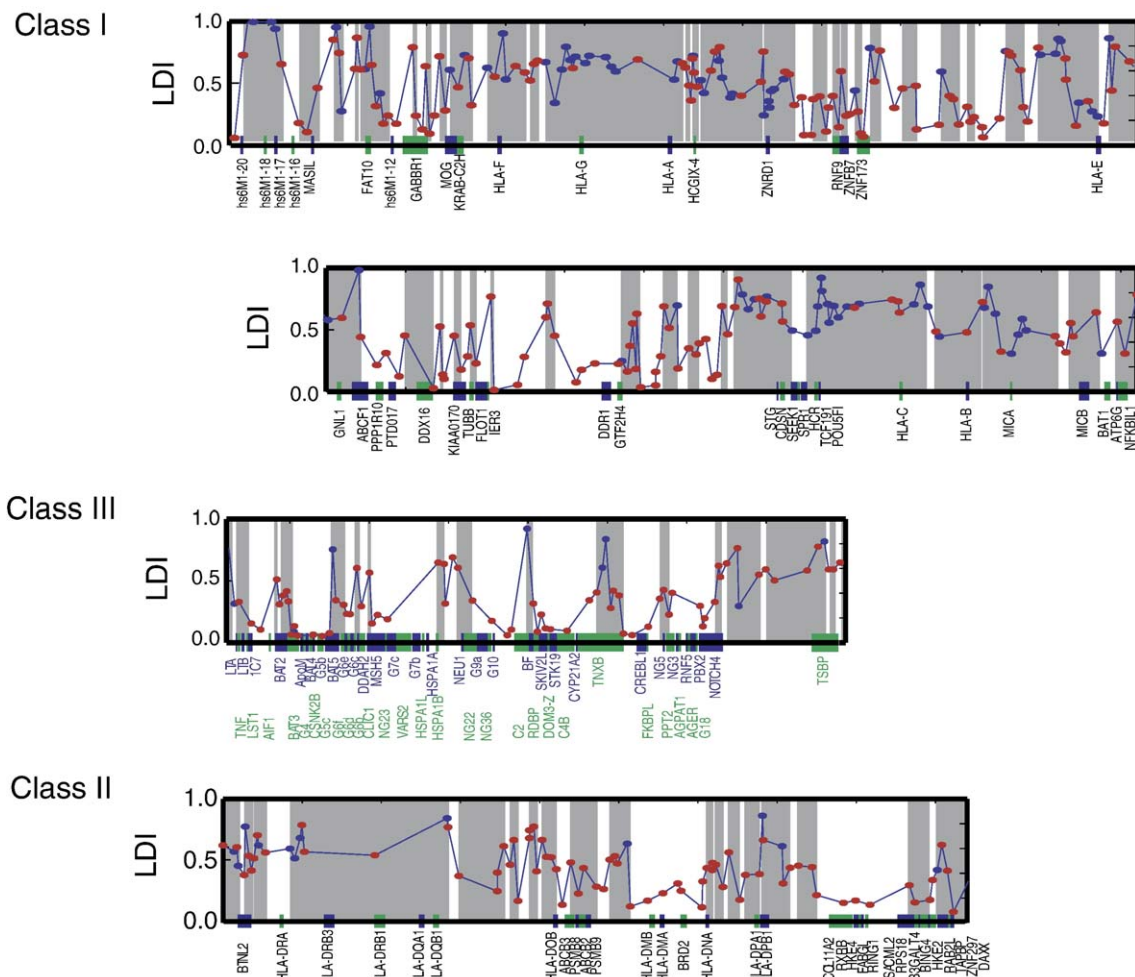


Fig. 2. Observed linkage disequilibrium index (LDI) and haplotype blocks in the MHC. An LDI value of 1 corresponds to perfect linkage disequilibrium, while a value of 0 corresponds to linkage equilibrium. One value for the LDI is plotted for each of the amplicons, and each plotted value corresponds to the LDI of the amplicon at that point with the following amplicon. Amplicons are then grouped into haplotype blocks, shown in gray. Blocks that contain only one amplicon are indicated simply as the amplicon in the areas with white background. Tagged amplicons within blocks are red points. Note that haplotypes within each amplicon are included in the analysis only if there are more than five copies, to ensure the robustness of haplotype block detection.



Table 2  
Tabulation of blocks and block sizes in the MHC derived from 46 HLA haplotypes

Block size (No. of amplicons)	Count	Percentage	Block size in kilobases			
			Mean	Min	Max	Mean No. of tagged amplicons
1	74	48.37	–	–	–	1
2	26	16.99	10.3	2.7	48.5	2
3	16	10.46	13.4	7.1	25.9	2.9
4	15	9.80	24.0	6.2	58.4	3.1
5	8	5.23	30.1	14.0	74.6	3.6
6	2	1.31	46.6	45.2	48.1	3.5
7	2	1.31	41.5	33.9	49.1	3.5
8	4	2.61	87.9	35.3	198.9	4.8
10	2	1.31	67.4	40.0	94.8	3.5
11	1	0.65	77.4	77.4	77.4	5
12	1	0.65	76.2	76.2	76.2	5
16	1	0.65	175.5	175.5	175.5	3
20	1	0.65	150.5	150.5	150.5	4

Note that haplotypes within each amplicon are included in the analysis only if there are more than five copies, to ensure the robustness of haplotype block detection.

One hundred fifty-three amplicon-defined SNP blocks were identified among the 48 phased chromosomes studied based on 410 amplicons, each selected because it contained more than five haplotypes (Table 2). Seventy-four (48%) of the 153 blocks consisted of a single amplicon, and 26 (17%) consisted of 2 amplicons. The average size of all 2-amplicon blocks was 13.4 kb with a range of 2.7–48.5 kb. The number of tagging amplicons for the largest blocks plateaus near 4, as shown in Table 2. For example, the largest block is composed of 20 amplicons, and sequencing of only 4 of the 20 amplicons yields 95% of the SNP information of the entire block. In contrast, the 8 5-amplicon blocks still require sequencing of an average of 3.6 amplicons to yield maximum information. All but 1 of the 26 2-amplicon blocks required sequencing of both amplicons for full information reporting. The total number of tagged amplicons for this dataset is 303 of 410 (74%) (Fig. 2) with 55 of the 107 unnecessary amplicons residing within the 6 largest blocks. The average size of the blocks containing 2 or more amplicons was 28 kb, which is similar to other regions in the human genome [29] and to two recent MHC SNP-typing studies [12,30].

#### MHC SNP analysis of the HLA-A1, B8, DR3 common extended haplotype

The relatively high prevalence of a few common extended conserved (CEH) haplotypes in human populations has long been a subject of interest with regard to their relationships to immune selection and immune-mediated disease [15]. The most frequent of these is the HLA-A1, B8, DR3, found at a frequency of about 7% in European Caucasians. Three copies of the 8.1 haplotype were present within our original set of 48 chromosomes, and analysis of these revealed virtually identical sequences throughout most of the MHC. To examine further this haplotype toward understanding its origin and potentially its age in the human population, we added 16 8.1 chromosomes from unrelated individuals to this analysis. A second goal was to

identify the boundaries of conservation within each of these CEH chromosomes, should they be contained within the regions that were sampled, to provide an initial focus for MHC SNP analysis of potential relevance to disease.

When the 19 8.1 chromosomes were analyzed at all 557 amplicon positions, it was immediately apparent that the region from HLA-A to HLA-DQ was highly conserved, while the variability outside of these loci was indistinguishable from any random set of 19 unrelated chromosomes, indicating that recombination has frequently taken place beyond these boundaries (Fig. 3). Virtually overlapping patterns of SNP variability were observed centromeric of HLA-DQ including DP. Conversely, the sequences beyond HLA-A were highly conserved until beyond the HLA-F locus, where variability overall was low. Of 317 SNPs identified among the 19 8.1 chromosomes, only 11 SNPs were found in the conserved region between HLA-A and DQ defined by 393 kb of measured sequence. Thus the diversity among the 19 8.1 chromosomes, a total of 11 SNPs, represents an average of 3.8 SNPs per pair of chromosomes over the 2.6-Mb region between HLA-A and HLA-DQ. These data are consistent with previous studies of LD between the HLA class I, class II, and class III (C4, Bf, TNF) genes, demonstrating that the Caucasian 8.1 haplotype is highly conserved and thus its preponderance (about 7% of the Caucasian haplotypes) is likely due to a profound expansion that must have occurred in relatively recent times. Assuming that SNPs accrue at a rate of 3/genome/generation [31], one can calculate how many generations are required for there to be an 11/19 chance that one SNP appears in a given number of bases. Every base has a  $1 \times 10^{-9}$  chance of being polymorphic in a given generation; there were  $393 \times 10^3$  bases sequenced between HLA-A and HLA-DR; after  $t$  generations the likelihood that we will see an SNP in that stretch is  $(1 \times 10^{-9}) (393 \times 10^3) t$ , and the number of generations required to accumulate an 11/19 chance is  $t = (11/19) / [(1 \times 10^{-9}) (393 \times 10^3)]$ . With 1470 generations at 16 years/generation we estimate that these haplotypes could have

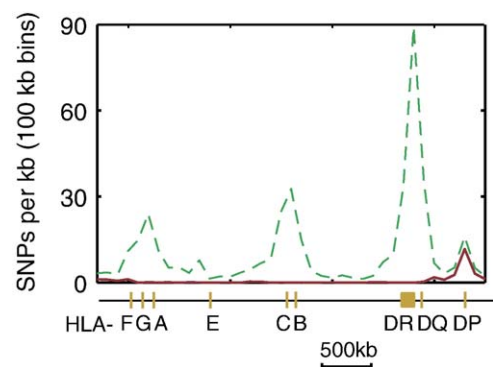


Fig. 3. The A1, B8, DR3 haplotypes are highly conserved between HLA-Q and HLA-DR. Variability between 19 8.1 chromosomes is presented with similar analysis of 19 HLA chromosomes analyzed in this study. The numbers of SNPs at continuous positions extending over the 4.0-Mb MHC region are plotted on the vertical against positions within the MHC on the horizontal. The solid red line represents analysis of the 19 8.1 chromosomes and the dashed green line is derived from 19 randomly chosen chromosomes from the original set of 48.

diverged from a single, common ancestor  $23,500 \pm 10,000$  years ago, where the uncertainty is a 95% confidence value and is dominated by the sampling error from the small number of chromosomes involved.

In Table 3 the numbers of SNPs among the 8.1 chromosomes and among all haplotypes examined are broken down over the entire region surveyed and within the conserved region between HLA-A and DQ. This breakdown resulted in three groupings, SNP alleles that were identified among the 8.1 chromosomes and distinguished from one another, SNP alleles that were unique to a subset of the 8.1 chromosomes, and SNP alleles that were unique to all of the 8.1 chromosomes. The alleles of SNPs unique to all 8.1 haplotypes analyzed include 25 positions that do not vary among the 19 chromosomes but of which the genotype is different from that found in any of the other 45 chromosomes analyzed (e.g., at one position a G is found in all 8.1 chromosomes and a T in all other chromosomes examined). Thus, any one of these 25 SNP positions could be used to identify the 8.1 chromosomes among the sample size of 46 haplotypes reported here. Of the 11 additional SNPs differing among the 8.1 chromosomes, 8 of these are unique among the 46 HLA haplotypes to a subset of the 8.1 chromosomes. These 8.1-specific SNPs were distributed across the MHC with no evidence of clustering that might suggest mutational hot spots (Fig. 4).

#### Rare SNPs in the MHC

Because the approach used in this study for genotyping was direct sequencing it was possible to detect rare SNPs, including those identified in the 8.1 analysis as discussed above. One of the reasons for continuing to use this method beyond an initial SNP discovery phase was evident from the data acquired during the sequencing of 8.1 chromosomes. In this case, if only common SNPs identified in the preliminary analysis of 48 chromosomes had been used to analyze the set of 19 8.1 chromosomes, only 3 of the 11 SNPs (Table 3) would have been identified. When extrapolating from the 15% of the region sampled, we can predict that there are about 73 SNPs that would be missed in the 8.1 haplotypes using a common variant association approach to detect SNPs.

The degree to which rare SNPs would not be detected using methods for SNP typing other than direct sequencing motivated a further study to estimate better the number of rare SNPs found per kilobase per individual. For this study, 15 amplicon

Table 3

	Complete MHC	HLA-A to DRB1
<i>A. SNP frequencies in the HLA-A1, B8, DR3 (8.1) haplotype<sup>a</sup></i>		
Number of SNPs	317	11
Number of variant amplicons	108	11
<i>B. Genotypes found only in 8.1 haplotypes among the 48 haplotypes examined</i>		
Unique genotypes possessed by all 8.1 chromosomes	25	23
Unique genotypes possessed by a subset of the 8.1 chromosomes	21	8

<sup>a</sup> Analysis included 21 8.1 haplotypes.

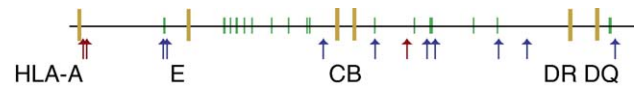


Fig. 4. Identification of SNPs within the 8.1.3 haplotypes. The positions of the HLA class I and II genes within the MHC are indicated on the line by larger vertical bars and each is identified beneath with the corresponding letter designating the HLA locus. Blue arrows below the line mark the positions of amplicons that contained SNPs unique to the 8.1 chromosomes, red arrows mark additional positions that differed among some of the 8.1 chromosomes in polymorphic positions known from the original 45 non-8.1 chromosomes, and green vertical lines on the line indicating the MHC mark positions where the genotype was the same for all of the 8.1 chromosomes examined but differed from all of the other 45 HLA haplotypes analyzed in this study.

sequences were sampled in 282 unrelated individuals mostly of Caucasian origin. From the 4.2 Mb of sequence derived, 45 SNPs were identified that were not observed among the previously established set of 64 SNPs identified in the same group of amplicons (Table 4). Of these, 36 were present at less than 1.1% of the population sampled. These SNPs could be further broken down into substitutions affecting coding sequences and noncoding sequences. The group of coding SNPs was equally weighted toward synonymous versus nonsynonymous substitutions, with 25 synonymous and 22 nonsynonymous. Only 7 of the 22 coding SNPs were present in more than 3% of the chromosomes. The total number of nonsynonymous substitutions in this group of chromosomes was 3.9%, or 7.5% of the individuals examined (all except one were heterozygous). This rate of nonsynonymous coding

Table 4  
Rare SNP analysis in 18 amplicons in the MHC class III region

Gene	Sequencing		SNPs				NS minor allele prevalence <sup>g</sup>	
	CDS bp <sup>a</sup>	bp seq'd <sup>b</sup>	Non-CDS <sup>c</sup>	CDS <sup>d</sup>	SS <sup>e</sup>	NS <sup>f</sup>	<1.1%	>1.1%
BAT2	1753 (27%)	4122	16	11	4	7	2 (4)	5 (399)
C6orf29	875 (41%)	1816	4	6	2	4	3 (3)	1 (230)
MICA	200 (17%)	811	13	4	3	1	0	1 (147)
AGER	466 (38%)	998	3	3	0	3	3 (7)	0
NCR3	453 (75%)	895	2	4	2	2	1 (6)	1 (9)
C6orf25	419 (63%)	1163	2	4	2	2	2 (2)	0
C2	399 (18%)	1030	3	1	0	1	1 (1)	0
DOM3Z	462 (39%)	808	2	3	2	1	1 (4)	0
C4B	859 (16%)	1848	4	5	4	1	1 (5)	0
BF	308 (13%)	1152	6	2	2	0	0	0
BAT3	374 (30%)	724	5	2	2	0	0	0
VARS2	430 (13%)	1216	1	1	1	0	0	0
NEU1	565 (15%)	1121	1	1	1	0	0	0
Total			62	47	25	22	14 (32)	8

282 cell lines in 18 amplicons were analyzed.

<sup>a</sup> Coding sequence (CDS) base pairs (bp) sequenced; the percentage of the total CDS from each gene is in parentheses.

<sup>b</sup> Total base pairs sequenced in the region of the gene.

<sup>c</sup> SNPs that are not located in CDS.

<sup>d</sup> Total SNPs in CDS.

<sup>e</sup> Synonymous CDS SNPs (SS) do not affect amino acid sequence.

<sup>f</sup> Nonsynonymous CDS SNPs (NS) alter amino acid sequence.

<sup>g</sup> Breakdown of rare and common NS and total number of chromosomes that have these SNPs.

substitutions suggests a potential source of functional variation that, when considered cumulatively, could significantly contribute to phenotype.

## Discussion

These studies were aimed at establishing and validating tools for examining genomic diversity across the human MHC, ultimately with the aim of using them in investigations of disease associations with the MHC. The tools in this case constituted the primer pairs that defined amplicons of unique sequence that were useful to derive data from any number of individual MHCs. The data generated represent the SNPs identified from the analysis of 48 distinct MHC haplotypes. Our goal was to define amplicons spread throughout the MHC at a relatively even spacing that would together comprise a significant portion of the MHC within the limits of practical constraints. Since the MHCs examined were either derived from homozygous HLA cells or based on the study of pedigree-defined HLA haplotypes, it was possible to deduce haplotypes for nearly all of the SNP data and phase this with the known HLA allelic information. The defined amplicons and data can be used to focus on the region as a whole or on any subset toward clinical and disease mapping applications in the HLA region.

Establishing linkage disequilibrium statistics among the 5186 SNPs spanning the MHC identified in this study can give an indication of recombination and evolutionary history of the MHC [12,30] but, importantly for practical studies, will allow for the selection of a subset of tagged SNPs that can be used to impart efficiency in any genotyping strategy [18,32]. To calculate LD we used the LDI as a potentially more powerful statistical method since it can accommodate more than two adjacent SNPs when testing for LD. In addition, instead of tagged SNPs we calculated LDI based on amplicons whereby the alleles at a given amplicon were tested for LD with adjacent amplicons downstream and upstream. This analysis was more appropriate for our study, since the chromosomal haplotype for each amplicon was fully recovered when we sequenced the amplicon, in contrast to SNP typing assays, which yield information only about the targeted base. Therefore, similar to the use of tagged SNPs, using sequence-based methods for querying tagged amplicons will yield information about other amplicons that are in LD, since the LDI gives information on how well the haplotypes or blocks present in one amplicon predict those in another. Although the multilocus ability of the LDI is most appropriate for our investigations involving resequencing, our SNP data are present in dbSNP and the entire set of over 250,000 genotypes is available in dbMHC (under the handle “FHCRC”), allowing any desired measure of LD to be applied to this dataset.

Cullen et al. have previously characterized linkage and haplotype blocks within the MHC class II region by genotyping sperm with microsatellites to map recombination locations [33]. Spots of high recombination should correspond to regions of low LDI, and all three of the spots described by Cullen et al. are consistent with the LDI map generated by the current dataset. A 200-kb segment of class II was mapped in more detail using

SNPs [34] with similar results, indicating that SNP searches for recombination hot spots will yield results similar to those of sperm typing with microsatellites. A larger study of recombination in the MHC mapped the entire region [35] and found one additional hot spot, located between LTA and BAT2, which was also found to be a region of low LDI in our analysis. Contrary to our results, however, that study found that the class III region is a relatively cool recombination area, while we find a low LDI throughout this region. Interestingly, while Cullen et al. did not observe any double recombination in the MHC, our analysis showed that cell line PLH (IHW090407) underwent a double recombination event between HLA-A and HLA-B extending from HCGIX-4 to HLA-E. It is not known if this was the result of double recombination in one generation or two separate recombinations over more than one generation.

Linkage disequilibrium among HLA loci has been a hallmark of the MHC, and the idea of block structures for the MHC was extensively analyzed and discussed well before a similar focus was brought to bear on the human genome [23,36–38]. One of the most interesting MHC haplotypes described is the HLA-A1, B8, DR3 haplotype, which is common among European Caucasians and is associated with several immune-mediated diseases including type 1 diabetes [3,6,39]. By sequencing 19 8.1 chromosomes from unrelated individuals we were able to obtain evidence of a rapid expansion of an ancestral chromosome in relatively recent times. This means that the HLA-A1, B8, and DR3 alleles are in strong linkage disequilibrium due to a recent expansion rather than having been derived from distinct haplotypes through multiple independent events in which selection drove the linkage of these alleles. This finding could be interpreted to mean that there is not necessarily any functional significance to the linkage of the three loci, but rather that possibly only one of the A1, B8, DR3 loci was under strong positive selection, which drove an expansion in recent times. Consistent with this possibility is the observation of a number of other conserved haplotypes that include DR3 but have lost HLA-A1 or HLA-B8.

Using simple assumptions, we estimated the age of the 8.1 haplotype at about 23,500 years, a time that is well within estimates of the human migration and expansion into Europe. These considerations might suggest then that around this time, strong selection for part or all of the 8.1 haplotype began. In that regard, the recent hypothesis that type 1 diabetes may have been selected for as an adaptive measure in cold climates raises an intriguing speculation [40]. In that study, cryoprotective adaptation based on higher glucose levels is proposed to have protected ancestral northern Europeans from the effects of the ice age around 14,000 years ago. Therefore, when life expectancy was short, factors predisposing to type 1 diabetes are proposed to have provided a survival advantage. This time frame is within the limits of error of our estimate of the beginning of the expansion of the 8.1 chromosome. It may be possible to define more precisely the period and the length of time such selection was operating by comparative study of 8.1 chromosomes with other HLA-DR3 and HLA-B8 positive haplotypes and building phylogenetic trees based on the sharing of SNP-defined construction of MHC blocks.



While the anthropological origins of the CEHs are interesting in their own right, the extreme similarities among the 8.1 chromosomes identify an important consideration in the design of strategies to identify genes residing in the MHC, other than the class I or II loci, that are associated with disease, of which a number of efforts are ongoing [41]. As mentioned, the 8.1 chromosome is not only highly represented in the European Caucasians, but it is also significantly associated with type 1 diabetes. Therefore it is of interest to compare type 1 diabetic 8.1 chromosomes with 8.1 chromosomes derived from normal controls. However, our findings indicate that the most commonly proposed approach, using haplotype-tagged common SNPs to map new loci, may not be satisfactory to identify what would presumably be SNPs that are unique to the T1D 8.1 chromosomes. Of 11 SNPs we identified that were 8.1 specific, only 3 of these would be eligible for inclusion in a common SNP panel and none of the 11 were in the typing panel used in the most recent MHC SNP study [12]. Based on the number of SNPs found unique to 8.1 and considering that we sampled 15% of the conserved region, we can calculate that there are as many as 53 8.1-specific SNPs among the 19 chromosomes tested here. Therefore, direct sequence analysis may be the only method that will be adequate for detecting differences between disease-associated and normal 8.1 chromosomes or to rule out any differences and thus conclude that there are no MHC resident loci other than the class II genes associated with T1D, at least between HLA-A and DQ.

Our analysis of the 8.1 CEH also suggests the importance of defining the extent of similarity across the MHC and the value of defining boundaries similarly for other known disease-associated CEHs. Definition of boundaries can provide a focus on the appropriate regions that can be applied toward answering similar questions about those disease associations. For example, the sequences immediately centromeric of HLA-DQ appear to have undergone significant recombination among all the initial 48 chromosomes and to a very similar extent among the group of 19 8.1 chromosomes (Fig. 3), indicating that loci within this region may not be candidates for association with T1D. Alternatively, this could be interpreted to focus studies more aggressively toward proving or disproving that there is no significant T1D association for any of the SNP blocks centromeric to DQ with T1D. Thus further examination of both normal and disease-associated 8.1 chromosomes might focus a search on specific regions, in addition to the need to rely on sequence-based methods for genotyping. From our preliminary analysis of other CEHs, the 8.1 haplotype provides an example of the type of information we can expect from a detailed study of several other CEHs (unpublished data).

Our preliminary work defining MHC variation was carried out using sequencing for SNP discovery, as are most current state-of-the-art approaches for SNP discovery. We extended the use of this approach to genotyping in the MHC rather than relying on other fixed-platform SNP typing methods for the reasons alluded to above to test the hypothesis that differences that do exist between CEHs may not be identifiable by SNP typing platforms that detect only common tagged SNPs. In addition, it is apparent from several studies that a number of

different rare variants from multiple genes can collectively contribute toward disease [20,21]. This includes the variants that are individually rare in the population and have very low heterozygosity but whose total number is large enough to provide many polymorphisms in a pedigree. A consideration of the potential number of distinct phenotypes associated with the MHC, which contains over 50 genes with immune-related functions, makes the possibility of a similar rare-SNP hypothesis for MHC disease associations quite plausible. In this study 105 common SNPs of varying frequencies in the gene-dense class III region were examined, and 45 new SNPs were identified through the use of sequence-based analysis. The majority of these were rare (<1%), but their combined cumulative frequency added significantly to the overall variability within the population. In addition, a significant number of these were nonsynonymous changes in coding sequences and thereby could have a direct functional consequence. Thus, resequencing can provide access to this abundant pool of rare variants—something that is impossible to achieve with conventional SNP maps using fixed-platform technologies.

The flexibility of the resequencing approach means that new information can be a routine by-product of sequence-based genotyping without any redesign in the underlying technology for acquiring genetic data. In fact, the ability to develop new assays is routine, with minimal costs, and can be added at any stage in a study. Instead of looking at over a thousand SNPs simultaneously on a single individual, as is the case for some fixed-platform methods, the amplicon-based approach can provide the flexibility and convenience of targeting selected MHC regions or genes. This approach is also adaptable to a change in the SNP composition to be queried in response to new information acquired as a study progresses. Such changes might require an entirely new application of some fixed-based SNP platforms to the study population or possibly the application of a different technology altogether to deal with the smaller number of SNPs. Sequenced-based genotyping requires nothing new as the application of new amplicons would seamlessly integrate into the ongoing process. In that regard, studies of disease association in the MHC may benefit from direct knowledge of the DNA sequence to make the most effective use of the extensive diversity that is the MHC.

## Materials and methods

### *Cell lines used*

A total of 47 Epstein–Barr virus-transformed B cell lines available through the International Histocompatibility Working Group (IHWG.org) were used in this study for primary SNP discovery (see online Supplementary Material). Twenty-seven of these were selected from a set originally part of the 10th International HLA Workshop (IHWS) panel identified as individually containing MHCs that were identical by descent (IBD) and were chosen as a group that represented diversity in their HLA types. Of these, 23 proved to contain homozygous sequences over the entire MHC region studied and 3 were homozygous from our most telomeric amplicon up to and including the HLA-DR region. Two of the 27 were found to have undergone intra-HLA recombination. BRIP (IHW09044) was homozygous only in a 1.5-Mb region extending from our most telomeric amplicon to the gene GTF2H4, and another,



PLH (IHW09047), apparently is the result of a double recombination event between HLA-A and HLA-B as heterozygous sequences were found from HCGIX-4 to HLA-E and homozygous throughout the rest of the region. An additional 20 cell lines were chosen from members of five selected pedigrees, which, by combining family relationships and HLA typing data, allowed for the resolution of individual haplotypes for the four parental chromosomes and in one case the additional grandparental chromosome. All of the sequence data and derivative SNP typings were resolved into phases either as homozygous sequences from the IBD lines or from segregation analysis in informative families. In the latter, in which all members of a family were heterozygous, which accounted for less than 2% of the SNP positions identified, it was not possible to assign SNPs to haplotypes and the haplotype was inferred to be consistent with the most common of the other unambiguous haplotype data available in the typing set. From this group, three haplotypes representing the HLA-A1, B8, DR3 (referred to here as the 8.1 haplotype) were available. To focus specifically on 8.1, this number was supplemented for analysis by eight additional homozygous 8.1 individuals (not derived from consanguineous matings). To quantify the number of rare SNPs in the class III region, 282 anonymous DNAs from healthy individuals were used.

#### Primer design and amplicons sampled for MHC variability

Amplicons were generated by defining PCR primers derived from the genomic template that yielded robust products of the predicted size first from 27 (MHC) identical by descent cell lines. Our goal then was to define amplicons spread throughout the MHC at a relatively even spacing that would together comprise a significant portion of the MHC but still stay within the limits of practical cost constraints. Primer pairs were designed to amplify intervals from 800 bp to 1.2 kb, using the completed MHC consensus as template [2] using software developed in-house that incorporated the fundamentals of Primer3 [42]. Repetitive regions were avoided by the use of RepeatMasker [43] and polynucleotide and multisatellite sequences were avoided or directed toward the center of amplicons. No preference was given to genes or noncoding regions, although all of the sequences examined were unique or low repetitive and included promoter, enhancer, multispecies conserved sequences [44,45], and other unique sequences of unknown function. Primer pairs were tested first on DNAs from 4 individual cell lines to determine whether they were capable of generating reliable and robust products (amplicons) and derivative sequence data. About 80% of all primer pairs passed this phase and were then used to generate PCR products and sequences from the remaining 44 cell lines. Of the 20% that failed one or more DNAs in the initial test, a single redesign was able to recover 75% of these, resulting in an overall success rate of 95%. These percentages did not apply to the highly polymorphic regions of the MHC such as that near HLA-B and -C [26], where many of the primers worked well on IBD DNAs but suffered from allele dropout due to having been derived from sequences that were polymorphic, resulting in preferential amplification of certain haplotypes. Over the MHC region examined, this occurred in 58 of the 576 amplicons designed (those that passed the initial test set of 4). Data from families derived from these amplicons were not included in the analysis reported here. This group of highly diverged amplicons included up to 167 SNPs/kb, which derived unambiguously from the group of 27 IBD DNAs. The primer sequences are available in the online supplementary material.

#### Sequence analysis for SNP discovery and genotyping

Resequencing for SNP discovery and genotyping was done with high-throughput methods that used Beckman Fx robots for PCR assembly and for sequence reaction assembly and ABI 3730xl sequencers for analysis, essentially according to the manufacturer's instructions and to well-established basic procedures [46]. Briefly, PCRs were assembled in a PCR clean room using a Beckman Fx robot according to strategies facilitated and automated by software built in the lab referred to as a Genetics Management System (GeMS) and modified substantially over previously developed software and concepts [47]. This system was used for experimental design, laboratory procedure tracking, and data storage in a relational database and as a framework for other exiting data analysis tools used throughout all phases of a resequencing project. This

system allowed precise control over the laboratory flow and was essential in establishing an overall low frequency of genotyping failures when using sequencing. In these studies, the failure rate was 1%, while the genotyping error rate (as judged by independent genotyping of the same samples) was less than 0.2%. The sample failure rate is considerably better than many fixed-platform technologies and the error rate compares very favorably. PCR products were treated with Exo-SAP and cycle sequencing reactions assembled from PCRs using a Beckman Fx robot were performed in 96-well thermocyclers (Perkin-Elmer 9700 or 2700) using cycle sequencing kits and protocols developed by PE-Applied Biosystems (BigDye V3.1) at between 1/8 and 1/24 reaction per sample. Reaction products were filtered through Sephadex columns and trace data were collected on an ABI 3700 or ABI 3730xl and chromatograms transferred automatically to an established database (GeMS).

#### SNP base-calling from sequence trace data

Both homozygous and heterozygous traces from each amplicon were first combined and analyzed for sequence content using Phred and Phrap [48] and data viewed using Consed [49]. A significant step toward making a sequencing project of this size manageable in a small lab was the in-lab development of a program for heterozygous sequence analysis (heterozygous trace resolution or HTR), which was able to perform most of the analysis of sequence-based SNP discovery and typing for heterozygous DNAs. This software is similar in intent to Polyphred [50] and Mutation Surveyor by Soft Genetics, LLC, and is specifically adapted for high-throughput analysis. We first tested the program on a subset of 26 amplicons, encompassing 389 polymorphic positions. These tests demonstrated that HTR was able to reduce the number of genotype errors made by other solutions by well over an order of magnitude, making only 5% as many false-positive position calls and improving on the rate of false-negative calls by a factor of 2 (false-negative calls were near zero in this comparison). Given the extent to which these error types are time consuming to correct manually, these improvements provide substantial productivity gains.

To provide a more comprehensive examination of the HTR interpretation of heterozygous data, all of the ambiguities found were resolved by manual interpretation. Of the 21 chromosomes contained within the five families, 18 were present in at least two individuals, providing an internal check of consistency of data calls. The known HLA type of each cell line was used to infer the phase for each family chromosome using locally written software. We identified 152 of 92,700 genotypes—or 0.2%—in which the duplicate-sequenced chromosomes were given different base calls by either the HTR software or manual interpretation. This value was essentially the same as we have previously experienced when independently analyzing the same sample by manual trace interpretations. All genotype positions conformed to expected Mendelian inheritance patterns within families.

#### LDI, haplotype blocks, and identification of tagged amplicons

The method of computing the LDI was developed to quantify genetic diversity at multiple loci and may be thought of as an extension of the traditional measurement of  $D'$ . In this analysis we used LDI to measure LD between amplicons, each of which includes multiple SNPs. Suppose an array of multiple SNPs ( $A_1-A_2-\dots-A_p$ ) from an amplicon  $A$  and multiple SNPs ( $B_1-B_2-\dots-B_q$ ) from an amplicon  $B$ , given the primary objective to measure LDI between these two amplicons, we computed their entropies, denoted as

$$E_A = - \sum_i p_{Ai} \log(p_{Ai}) \text{ and } E_B = - \sum_i p_{Bi} \log(p_{Bi}),$$

where  $p$  is the haplotype frequency for associated haplotypes. Now one can also compute their joint entropy  $E_{AB}$ , for joint haplotypes of  $A$  and  $B$ . Under the null hypothesis that amplicons  $A$  and  $B$  are at linkage equilibrium (LE), the joint diversity is maximized and equals the summation of individual entropies, i.e.,  $E_{\max} = E_A + E_B$ . If both amplicons are at perfect LD, their joint entropy would be minimized, and its value is denoted as  $E_{\min}$ . The observed entropy,  $E_{AB}$ , falls between  $E_{\min}$  and  $E_{\max}$ . Therefore, if we define LDI as

$$\text{LDI}(A-B) = \frac{E_{\max} - E_{AB}}{E_{\max} - E_{\min}}, \quad (1)$$

LDI is within the range (0,1) with value 0 under LE and value 1 under perfect LD.

To detect haplotype blocks, we used a  $\chi^2$  test to guide the detection, rather than using pair-wise measurement  $D'$  as did Gabriel et al. [51]. A major difference specific to this analysis is the building of haplotype blocks directly from amplicons rather than SNPs. In that case each amplicon consisting of multiple alleles is measured for LD with neighboring amplicons and grouped into a single block if LD as computed by the  $\chi^2$  test is calculated as high. Following this definition, the basic procedure systematically tests if an adjacent amplicon should be merged into the existing haplotype block. The rejection of the null hypothesis of LE implies that the amplicon should be concatenated into the block currently being tested due to LD with the block. Otherwise, a new block is formed starting at that amplicon.

Due to high LD among amplicons within the block, fewer amplicons are actually needed to re-create the entire haplotype block such that there is virtually no or nonsignificant loss of genetic diversity within the blocks. To identify such a set of haplotype-tagging amplicons within each block, we systematically compute entropy values with one amplicon being deleted at a time and leave that amplicon only if the deletion of that amplicon does not reduce the original entropy with all amplicons within the block by, e.g., 5%. Applying this procedure to all of the amplicon-based haplotype blocks, we can identify such a set of “tagged amplicons.” It is important to note that the set of tagged amplicons is not unique. Among those equivalent sets, one can choose the “preferred set” per selection criteria, such as experimental robustness. This procedure, together with the one for detecting haplotype blocks, has been applied to analyze HapMap project data and will be reported separately (personal communication, L. P. Zhao).

## Acknowledgments

We thank Skylar Nelson for excellent technical assistance and Ruihan Wang for outstanding contributions to data analysis. D.E.G. was supported by AI 49245 and AI 33484, W.P.S. was supported by T32 HG00035. L.P.Z. and S.L. were supported by CA106320 and A1049213 and J.A.H. was supported by AI33484, AI49213, and CA15704.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.ygeno.2005.11.020](https://doi.org/10.1016/j.ygeno.2005.11.020).

## References

- [1] P. Parham, Virtual reality in the MHC, *Immunol. Rev.* 167 (1999) 5–15.
- [2] MHC Consortium, Complete sequence and gene map of a human major histocompatibility complex. The MHC Sequencing Consortium, *Nature* 401 (1999) 921–923.
- [3] B. Dupont, A. Svejgaard, HLA and disease, *Transplant. Proc.* 9 (1977) 1271–1274.
- [4] E.J. Yunis, B. Dupont, J. Hansen, Immunogenetic aspects of allotransplantation, *Adv. Exp. Med. Biol.* 73 Pt B (1976) 231–251.
- [5] J.A. Hansen, K. Yamamoto, E. Petersdorf, T. Sasazuki, The role of HLA matching in hematopoietic cell transplantation, *Rev. Immunogenet.* 1 (1999) 359–373.
- [6] C.E. Larsen, C.A. Alper, The genetics of HLA-associated disease, *Curr. Opin. Immunol.* 16 (2004) 660–667.
- [7] S. Segal, A.V. Hill, Genetic susceptibility to infectious disease, *Trends Microbiol.* 11 (2003) 445–448.
- [8] T. Shiina, H. Inoko, J.K. Kulski, An update of the HLA genomic region, locus information and disease associations: 2004, *Tissue Antigens* 64 (2004) 631–649.
- [9] S.G. Marsh, Nomenclature for factors of the HLA system, update May 2004, *Hum. Immunol.* 65 (2004) 1546–1547.
- [10] C.A. Stewart, et al., Complete MHC haplotype sequencing for common disease gene mapping, *Genome Res.* 14 (2004) 1176–1187.
- [11] R. Daza-Vamenta, G. Glusman, L. Rowen, B. Guthrie, D.E. Geraghty, Genetic divergence of the rhesus macaque major histocompatibility complex, *Genome Res.* 14 (2004) 1501–1515.
- [12] M.M. Miretti, et al., A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms, *Am. J. Hum. Genet.* 76 (2005) 634–646.
- [13] E.C. Walsh, et al., An integrated haplotype map of the human major histocompatibility complex, *Am. J. Hum. Genet.* 73 (2003) 580–590.
- [14] R. Dawkins, et al., Genomics of the major histocompatibility complex: haplotypes, duplication, retroviruses and disease, *Immunol. Rev.* 167 (1999) 275–304.
- [15] E.J. Yunis, et al., Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks, *Tissue Antigens* 62 (2003) 1–20.
- [16] J.M. Chapman, J.D. Cooper, J.A. Todd, D.G. Clayton, Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power, *Hum. Hered.* 56 (2003) 18–31.
- [17] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, E.S. Lander, High-resolution haplotype structure in the human genome, *Nat. Genet.* 29 (2001) 229–232.
- [18] F. Takeuchi, et al., Linkage disequilibrium grouping of SNPs reflecting haplotype phylogeny for efficient selection of tag SNPs, *Genetics* 16 (2005) 16.
- [19] N. Risch, K. Merikangas, The future of genetic studies of complex human diseases, *Science* 273 (1996) 1516–1517.
- [20] M.M. Swarbrick, C. Vaisse, Emerging trends in the search for genetic variants predisposing to human obesity, *Curr. Opin. Clin. Nutr. Metab. Care* 6 (2003) 369–375.
- [21] J.C. Cohen, et al., Multiple rare alleles contribute to low plasma levels of HDL cholesterol, *Science* 305 (2004) 869–872.
- [22] S. Lin, A. Chakravarti, D.J. Cutler, Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies, *Nat. Genet.* 36 (2004) 1181–1188.
- [23] R.L. Dawkins, et al., Disease associations with complotypes, supratypes and haplotypes, *Immunol. Rev.* 70 (1983) 1–22.
- [24] P. Price, et al., The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases, *Immunol. Rev.* 167 (1999) 257–274.
- [25] L. Subrahmanyam, M.A. Eberle, A.G. Clark, L. Kruglyak, D.A. Nickerson, Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus, *Am. J. Hum. Genet.* 69 (2001) 381–395.
- [26] T. Guillaudeux, M. Janer, G.K. Wong, T. Spies, D.E. Geraghty, The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism, *Proc. Natl. Acad. Sci. U. S. A.* 95 (1998) 9494–9499.
- [27] R. Horton, 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.* 282 (1998) 71–97.
- [28] B. Carlsson, J. Wallin, J. Bohme, E. Moller, HLA-DR-DQ haplotypes defined by restriction fragment analysis: correlation to serology, *Hum. Immunol.* 20 (1987) 95–113.
- [29] G.R. Abecasis, et al., Extent and distribution of linkage disequilibrium in three genomic regions, *Am. J. Hum. Genet.* 68 (2001) 191–197.
- [30] A. Stenzel, et al., Patterns of linkage disequilibrium in the MHC region on human chromosome 6p, *Hum. Genet.* 114 (2004) 377–385.
- [31] A.L. Hughes, M.K. Hughes, Natural selection on the peptide-binding regions of major histocompatibility complex molecules, *Immunogenetics* 42 (1995) 233–243.
- [32] K. Lindblad-Toh, et al., Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse, *Nat. Genet.* 24 (2000) 381–386.
- [33] M. Cullen, et al., Characterization of recombination in the HLA class II region, *Am. J. Hum. Genet.* 60 (1997) 397–407.
- [34] A.J. Jeffreys, L. Kauppi, R. Neumann, Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex, *Nat. Genet.* 29 (2001) 217–222.
- [35] M. Cullen, S.P. Peretto, W. Klitz, G. Nelson, M. Carrington, High-resolution patterns of meiotic recombination across the human major histocompatibility complex, *Am. J. Hum. Genet.* 71 (2002) 759–776.

- [36] C.A. Alper, Z.L. Awdeh, E.J. Yunis, Complotypes, extended haplotypes, male segregation distortion, and disease markers, *Hum. Immunol.* 15 (1986) 366–373.
- [37] C.A. Alper, E. Fleischnick, Z. Awdeh, A.J. Katz, E.J. Yunis, Extended major histocompatibility complex haplotypes in patients with gluten-sensitive enteropathy, *J. Clin. Invest.* 79 (1987) 251–256.
- [38] R.L. Dawkins, et al., Some disease-associated ancestral haplotypes carry a polymorphism of TNF, *Hum. Immunol.* 26 (1989) 91–97.
- [39] R.L. Dawkins, et al., Supratypes and ancestral haplotypes in IDDM: potential importance of central non-HLA MHC genes, *J. Autoimmun.* 3 (Suppl. 1) (1990) 63–68.
- [40] S. Moalem, K.B. Storey, M.E. Percy, M.C. Peros, D.P. Perl, The sweet thing about Type 1 diabetes: a cryoprotective evolutionary adaptation, *Med. Hypotheses* 65 (2005) 8–16.
- [41] S. Onengut-Gumuscu, P. Concannon, Mapping genes for autoimmunity in humans: type 1 diabetes as a model, *Immunol. Rev.* 190 (2002) 182–194.
- [42] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, *Methods Mol. Biol.* 132 (2000) 365–386.
- [43] A.F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, *Curr. Opin. Genet. Dev.* 9 (1999) 657–663.
- [44] W.J. Murphy, G. Bourque, G. Tesler, P. Pevzner, S.J. O'Brien, Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps, *Hum. Genom.* 1 (2003) 30–40.
- [45] E.H. Margulies, E.D. Green, Detecting highly conserved regions of the human genome by multispecies sequence comparisons, *Cold Spring Harbor Symp. Quant. Biol.* 68 (2003) 255–263.
- [46] D.E. Geraghty, R. Daza, L.M. Williams, Q. Vu, A. Ishitani, Genetics of the immune response: identifying immune variation within the MHC and throughout the genome, *Immunol. Rev.* 190 (2002) 69–85.
- [47] D.E. Geraghty, et al., Data acquisition, data storage, and data presentation in a modern genetics laboratory, *Rev. Immunogenet.* 2 (2000) 532–540.
- [48] B. Ewing, L. Hillier, M.C. Wendl, P. Green, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res.* 8 (1998) 175–185.
- [49] D. Gordon, C. Abajian, P. Green, Consed: a graphical tool for sequence finishing, *Genome Res.* 8 (1998) 195–202.
- [50] D.A. Nickerson, V.O. Tobe, S.L. Taylor, PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing, *Nucleic Acids Res.* 25 (1997) 2745–2751.
- [51] S.B. Gabriel, et al., The structure of haplotype blocks in the human genome, *Science* 296 (2002) 2225–2229.