Single Nucleotide Polymorphism
Blocks and Haplotypes: Human
MHC Block Diversity

E.J. Yunis\textsuperscript{1,2}, J. Zuñiga\textsuperscript{1,2,3}, C.E. Larsen\textsuperscript{2,4}, M. Fernández-Viña\textsuperscript{5}, J. Granados\textsuperscript{6}, Z.L. Awdeh\textsuperscript{4}, and C.A. Alper\textsuperscript{2,4}

\textsuperscript{1}Dana-Farber Cancer Institute, Boston, MA, USA
\textsuperscript{2}Harvard Medical School, Boston, MA, USA
\textsuperscript{3}Instituto Nacional de Enfermedades Respiratorias, Mexico City, Mexico
\textsuperscript{4}The CBR Institute for Biomedical Research, Boston, MA, USA
\textsuperscript{5}University of Texas, Houston, TX, USA
\textsuperscript{6}Instituto Nacional de Ciencias Médicas y de la Nutrición Salvador Zubirán, Mexico City, Mexico

1 Introduction 194

2 The Major Histocompatibility Complex (MHC) 195
2.1 MHC Blocks 198
2.2 HLA Blocks 198
2.2.1 HLA-Cw/B Block 198
2.2.2 HLA-DR/DQ Block 199
2.3 Common HLA Haplotypes Defined Serologically and Deduced HLA-Cw/B, HLA-DRB1/DQB1 Common Haplotypes and their Frequencies 201
2.4 The Aggregate Block Frequency (ABF) Measures Genetic Diversity 201
2.5 Class III Region Blocks 202
2.5.1 TNF Block 202
2.5.2 Complotype Block 202

3 Conserved Extended Haplotypes (CEHs) 202
3.1 Extension of the HLA-Cw/B Block to HLA-A and of the HLA-DR/DQ Block to HLA-DPB1 203
3.1.1 Frequency Distribution of MHC Blocks, CEH Fragments, and CEHs 203
3.1.2 The Global LD Measurement Related to the Size of the Database 203
3.1.3 Association of CEHs with Disease 203
Keywords

Alleles
Variants of genes, alternative forms of a gene that occur at a given location of a chromosome.

Conserved Extended Haplotypes
Relatively large stretches (1.5 Mb or more) of conserved DNA sequence, in which essentially all allelic markers within the genetic region are identical and shared among a relatively large number of unrelated individuals (typically, although not definitively, greater than or equal to 0.5% of a given population). Operationally, human MHC CEHs are identified by identical markers on a single chromosome at HLA-B, complotype and HLA-DRB1.

Disease Association
Genetic markers significantly associated with diseases.

Fragments of Conserved Extended Haplotypes
Genetic regions with different sizes of DNA blocks.

genes
Self-reproducing hereditary characters involved in production of proteins.

Genetic Recombination and Hotspots
Points of crossing-over between chromosomes during meiosis determined by family studies. Meiosis is the creation of sex cells by replication of chromosomes followed by
cell division and production of gametes; sperms and eggs. Hotspots refer to hypothetical chromosomal regions with higher rates of recombination.

**Genetic Stratification and Population Admixture**
Studies of DNA, RNA, or proteins that characterize different nationalities or races.

**HLA Polymorphism**
Allelic or specificity variation within specific genes or loci of the human MHC.

**Human Genetic Diversity**
Genetic markers that differ among individuals.

**Linkage**
Alleles on the same chromosome close enough to be inherited together.

**Linkage Disequilibrium**
Nonrandom association of inherited alleles in a population.

**Major Histocompatibility Complex (MHC)**
Polymorphic genes located in the human chromosome 6 involved in immune responses and in transplantation.

**Population Genetics**
Genetic differences among individuals of different nationalities or racial groups.

We describe genetically fixed segments of DNA within the major histocompatibility complex (MHC) extending to 3.2 Mb of DNA from HLA-A to HLA-DPB1. These are variable sized DNA fragments that vary in frequency in different ethnicities or races. Within the region, one 1.5-Mb block is relatively frequent, the DNA conserved extended haplotypes (CEHs). These span from HLA-Cw, B to HLA-DRB1, DQB1 and include the polymorphisms of complement genes (complotypes) and TNF (tumor necrosis factor genes). This segment is informative for mapping disease susceptibility, immune responses, and allotransplantation matching. Other chromosomal regions (paralogous) have conserved genes, including those genetically related to those of the MHC, which also form DNA blocks that could be involved in immune functions and show disease association. DNA blocks can be used to measure human diversity. For example, the aggregate frequency of DNA blocks (ABF), determine the degree of genetic diversity in different populations; Africans, Asians, and Hispanics have higher genetic diversity than Caucasian Americans. The use of single nucleotide polymorphisms (SNPs) to determine DNA blocks (by measurement of linkage disequilibrium (LD)) is limited when ignoring the well-documented variability of frequency and size of DNA blocks among different populations.
Single Nucleotide Polymorphism Blocks and Haplotypes: Human MHC Block Diversity

1 Introduction

Recent interest in mapping human disease genes using single nucleotide polymorphisms (SNPs) has prompted us to review the structure of the human major histocompatibility complex (MHC) in order to discuss the limitations of such methods. SNPs occur every 200 to 2000 bp of human genomic DNA, and groups of relatively nearby SNPs have been analyzed as haplotypes on the basis of the measurement of linkage disequilibrium (LD). Genetic variation is nonrandomly distributed in the genome, and SNP density and distribution varies. The haplotypes deduced are said to be islands of nucleotide segments that are stable and form blocks. Claims are also made that meiotic crossing-over occurred between these blocks at what are called “hotspots” with the assumption that historic recombinations between the blocks explain and define operationally such blocks. Thus defined, the sizes of such blocks, using LD measurements, are between 5 to approximately 200 kb in size or more than 800 kb. A major drawback of all these reports of SNP blocks is that they disregard the well-documented existence of DNA blocks in the MHC, which vary in size among individuals and vary in frequency in different ethnicities.

We summarized in two previous reports some of the structural elements and polymorphic character of the human MHC regions between HLA-A and HLA-B, and between HLA-B and HLA-DRB1. These two regions have comparable sizes: 1460 kb from HLA-A to HLA-B and 1375 kb from HLA-B to HLA-DRB1. We also described the existence of small blocks and other relatively fixed genetic fragments in the human MHC: A/Cw = 1365 kb, Cw/B = 95 kb, TNF region ≥ 7 kb, complotype = 120 kb, and DR/DQ = 150 kb. We used the known frequencies of the latter four different small blocks to deduce the frequencies of common MHC haplotypes with sizes up to 3.4 Mb. We described a model using those four blocks for studying the population-based structure of the MHC with 16 potential kinds of block combinations that are either fragments of or individual blocks of conserved extended haplotypes (CEHs) with a total length of 1470 kb. This segment of the MHC (from HLA-Cw to HLA-DQB1) has been studied more extensively than the segment between HLA-A and HLA-B despite the fact that the latter was discovered much earlier.

MHC haplotype blocks and the larger CEHs are usually inherited intact as a unit, and the allele frequency distribution of particular MHC locus combinations in individuals is nonrandom. Specific alleles at various MHC loci tend to occur together on the same chromosome much more or much less frequently than the product of their individual frequencies. The LD between HLA-A and HLA-B and between HLA-B and HLA-DRB1 has been well determined in some populations. By contrast, recent studies describing “blocks” of conserved DNA sequence (15–150 kb) within the human genome separated by sites of recombination (deduced) based on LD analysis of SNPs, suggest the existence of uniform lengths of conserved DNA sequences. Some investigators believe that each of these short blocks can be studied with high-throughput screening without the need for doing pedigree analysis. This, however, has questionable utility in the MHC. Of interest is a related SNP approach, analyzing a larger stretch of DNA (3.5 Mb) within the MHC, which was
reported recently, with the finding of what were called surrogate haplotypes arbitrarily anchored at HLA-B. Unfortunately, the haplotypes found did not always correlate with well-defined common haplotypes and CEHs found by pedigree analysis.

The small blocks of the MHC have different sizes and their variants differ in frequency in populations where they behave as fixed genetic units. These genetic units have been used as markers of human diversity, of ethnicity and/or nationality, of immune response, and of disease risk. They can also be used for genetic stratification studies to confirm the similarity or degree of admixture of populations. The HLA-Cw/B, complement, and DR/DQ small blocks and the larger CEHs or common MHC haplotypes are important, because some of their variants are ethnic specific, while others are shared between different ethnicities. By contrast, SNP analysis in its infancy has rarely been used to distinguish individual variations in and to identify the population-specific distribution of the sizes of conserved DNA sequence longer than that of the small blocks. Furthermore, little effort has been made to understand the discrepancies between measurements of SNP LD as compared with methods of direct counting of different sizes of blocks in the MHC. Our view is that a thorough understanding of the relatively well-characterized MHC blocks in different human populations should serve investigators to better understand the sizes of DNA blocks throughout the genome.

2 The Major Histocompatibility Complex (MHC)

The human MHC is located within chromosomal region 6p21.3, spans at least 3.4 Mb of DNA and contains many genes and pseudogenes. In addition to the classical HLA genes, the MHC contains more than 140 other functional genes, many of which have immune function. The MHC genes are divided into three regions (class I, II, and III). The class I region is located at the telomeric end of the MHC, while the class II is at the centromeric end, with the class III region located between the class I and class II regions (Fig. 1). HLA allelic variants were first studied using serological reagents detecting human HLA

---

**Fig. 1** Map of the human major histocompatibility complex (MHC) in 6p21.3. The physical distances between specific loci are shown. Distances are approximate and vary at many locations in different haplotypes as a result of various DNA insertions, deletions, and gene duplications. Modified from Yunis, E.J. et al. (2003) *Tissue Antigens.*
specificities, but HLA alleles can only be determined with precision using DNA methods. Genetic variants (alleles) at more than one locus inherited together on a parental chromosome are referred to as haplotypes.

MHC haplotypes (a term derived from haploid cells) found in an individual can only be defined by pedigree analysis in family studies, but several studies of large populations of unrelated subjects have been published that predicted frequencies of haplotypes calculated by statistical analysis. The original description of an HLA haplotype in 1967 using genotyped data that was obtained by pedigree analysis explained the coinheritance of alleles of two closely linked loci in several unrelated individuals. The concept of LD refers to the nonrandom association between alleles (usually at relatively nearby genetic loci). Therefore, this analytical method was quickly exploited to analyze preferential association between alleles at two or more loci based solely on phenotypic data (i.e. HLA typing of unrelated individuals, in which haplotype phase is not accurately known) has been described for the HLA region by several authors. LD analysis has played an important role in both gene mapping and as one approach to identifying susceptibility markers in complex diseases. The extent of LD is given by delta \( D = f(AB) - (f(A) \times f(B)) \), where larger delta values indicate greater LD. Thus, the LD of a two-locus haplotype, \( AiBj \) will be:

\[
LD(AiBj) = HF(AiBj) - (f(Ai) \times f(Bj)),
\]

where \( HF \) is the haplotype frequency and \( f(Ai) \) and \( f(Bj) \) are the frequencies of the \( Ai \) and \( Bj \) alleles. The standard delta value \( (D) \) is related to the frequency of alleles \( f(A) \) and \( f(B) \). The allele frequency effect can be partially compensated by calculating a normalized LD value \( (D') \) that reflects the relative LD irrespective of allele frequencies. \( D' \) is calculated as:

\[
D' = D/D_{max},
\]

where \( D_{max} \) is the maximum \( D \) value possible. The definition of “strong” LD varies by the specific type of locus analysis under study. The highly polymorphic classical MHC haplotypes (and even their small blocks) containing alleles in strong LD generally have a \( D' > 0.10 \). For statistical tests of the significance, chi-square values are calculated using \( 2 \times 2 \) contingency tables converted to \( p \) values.

LD determinations between HLA-A and HLA-B and between HLA-B and HLA-DRB1 have been well described. Despite the opportunity for genetic recombination within the MHC seen in human sperm, the effect of meiotic recombination on the population distribution of MHC haplotypes is still uncertain. For example, an apparent “hotspot” of genetic recombination was found in human sperm within the class III region of the MHC near LTA, whereas a lower than expected recombination rate
was seen between HLA-A and HLA-B. This would suggest less relative fixity between HLA-Cw and HLA-DRB1 as compared with the class I region. Nevertheless, this recombination site has not been confirmed in family studies. In the absence of both consistent family study confirmation and studies of the effect of sperm recombination on fertilization success and on pregnancy outcome, sperm recombination studies should not be generalized to describe the effect of recombination on haplotype frequencies in the population. Furthermore, the existence of CEHs, found in family studies, within the region between HLA-Cw and HLA-DRB1 (often without a frequent HLA-A allele) and the comparable LD between HLA-A and HLA-B alleles as compared with that between HLA-B and HLA-DRB1 alleles suggests that the genetic fixity of the two regions is similar.

We believe that it is necessary to study more intensively the genetic region between HLA-Cw and HLA-A to define new class I blocks and to analyze the possibility that the hotspots or regions of historical recombination represent small block boundaries. Recently, other recombination hotspots have been described. However, some of the described hotspots need to be confirmed with finer methods to establish their precise position. The physical distances for the two regions, HLA-A to HLA-Cw/B and HLA-Cw/B to HLA-DR/DQ are nearly identical. Pedigree analyses by family study have shown that both regions have approximately 1% recombination frequencies. If hotspots of genetic recombination between HLA-B and HLA-DRB1 had a significant effect on the population frequencies of specific haplotypes, one would not have expected that the degree of nonrandom association would be higher in the latter. The full extent of LD within the MHC was further demonstrated with the elucidation of genetic polymorphism in the class III region intermediate between HLA-B and HLA-DRB1. Class III region genes include the complement genes C2, C4A, C4B, and BF as well as the cytokine genes TNF, LTA, and LTB. The haplotypes formed by complement genes in MHC class III region are called complements [K]. There are 14 relatively high-frequency complements in Caucasians that occur with a combined frequency of 0.96. MHC haplotypes defined by pedigree analysis containing essentially fixed DNA sequence across class I, II, and III loci found at relatively high frequency in specific populations are referred to as CEHs or ancestral haplotypes (AHs).

The presence in the population of CEHs, each defined by pedigree analysis, provides evidence for the DNA fixity in specific haplotypes of unrelated individuals. Two recent studies using LD analysis failed to find a significant amount of nonrandom association within the HLA-Cw to HLA-DRB1 region other than in small blocks, although a third study found such associations. The former results are in apparent conflict with those reported by us and by others. We are not surprised that statistical analyses of a relatively few haplotypes (or unrelated individuals) fail to detect more than a few nonrandom associations of alleles of the MHC.

Reports describing the existence of “blocks” of conserved DNA sequence in the range of 5 to 150 kb within the human genome, separated by sites of high recombination activity, are based solely on LD analysis applied to SNP data. Most of these reports have concluded that such blocks represent relatively uniform lengths of conserved DNA sequences maintained throughout the human population as haplotypes. Many investigators now believe that such short blocks can be studied
with high-throughput screening without pedigree analysis. Although these techniques have been used recently to detect some of the previously demonstrated associations in the already well-characterized MHC genes and variants, their current utility, at least within the MHC, is uncertain. For example, LD analysis of SNP typing results, from unrelated Caucasian donors, was used recently across a much longer stretch of DNA (3.5 Mb) within the MHC to study “surrogate” HLA-B haplotypes. Genetic fixity within the MHC is so great that even LD measurements, taken over long distances of computer-constructed surrogate haplotypes (i.e. not defined by pedigree analysis) predicted to varying degrees the existence of specific common MHC haplotypes, although the haplotypes predicted often did not conform to high-frequency haplotypes seen in larger population studies based on direct haplotype counting through pedigree analysis. Another report, using pedigree-defined SNP-typed MHC haplotypes to construct an “integrated haplotype map” of the MHC, claimed that genetic fixity is higher within the MHC than throughout the rest of the human genome. However, those results did not achieve statistical significance, and the report only identified one common haplotype (although shorter) found in larger haplotype studies. SNP haplotype analysis, in its infancy, clearly still has much to prove before it might be considered a useful method for providing new insights into structure–function relationships, at least within the already well-characterized MHC.

2.1 MHC Blocks

A block size range of 5 to 150 kb in the human MHC have been well defined. These blocks are different from those recently reported, are never characterized solely by SNP typing, are highly polymorphic, and are diverse in different ethnic groups. Specific MHC blocks with specific alleles of one locus are often “haplospecific” for particular CEHs. Frequently, one allele at one MHC locus is highly correlated with an entire CEH. In other cases, a specific group of alleles within the same MHC block or in separate blocks is required to have a high correlation with a specific CEH. However, definitive characterization of haplotypes requires the analysis of families. Some haplotypes are not CEHs but could contain fragments or small blocks from specific CEHs. The best determination of haplotypes would require family studies.

2.2 HLA Blocks

In previous reports, we have summarized the two best characterized blocks of HLA loci, but we will emphasize in the following that the frequency distribution of the inheritance of the two blocks together has not been done in a critical manner. Instead, it has been possible to deduce the resulting haplotypes for the purpose of comparing frequencies of blocks and haplotypes in several ethnicities.

2.2.1 HLA-Cw/B Block

The HLA-Cw/B block is a relatively small segment of DNA (95 kb in length). Figure 2 shows the most frequent HLA-Cw/B block associations in four different ethnic groups of Americans, which, as reported before, are found in strong ($D' > 0.10$) and significant ($p > 0.0005$) LD. African-Americans, Asian-Americans, and Caucasian Americans share several high-frequency small blocks. Also shown
are the frequent HLA-Cw/B blocks of Hispanic Americans to show haplotypes that indicate ethnic admixture. However, some blocks are much more frequently found in only one or two of the groups. Certain blocks are of higher frequency: (HLA-Cw*0702, B*0702) \((f = 0.11)\), (HLA-Cw*0701, B*0801) \((f = 0.11)\), and (HLA-Cw*0502, B*4402) \((f = 0.09)\) in Caucasian Americans; (HLA-Cw*0401, B*5301) \((f = 0.09)\), (HLA-Cw*0202, B*1503) \((f = 0.07)\), and (HLA-Cw*1701, B*4201) \((f = 0.05)\) in African-Americans; and (HLA-Cw*0302, B*5801) \((f = 0.07)\), (HLA-Cw*0702, B*3802) \((f = 0.07)\), and (HLA-Cw*0102, B*4601) \((f = 0.06)\) in Asian-Americans.

### 2.2.2 HLA-DR/DQ Block

The HLA-DR/DQ region, spanning approximately 150 kb, is the longest of the four MHC blocks. The definition of this block is based on published frequencies of HLA-DRB1, -DQB1 haplotypes in three different American populations. HLA-DRB3/4/5 and -DQA1 are other well-characterized polymorphic loci located within this DNA segment. Figure 3 summarizes the frequency data for the most frequent DR/DQ haplotypes in various ethnic groups. Some of these DR/DQ blocks are found much more frequently in certain ethnic groups. For example, (HLA-DRB1*0401, DQB1*0301) is primarily found in Caucasian Americans \((f = 0.03)\); (HLA-DRB1*1503, DQB1*0602) is primarily found in African-Americans \((f = 0.07)\); and (HLA-DRB1*0901, DQB1*0303) is primarily found in Asian-Americans \((f = 0.10)\). Other DR/DQ haplotypes are found frequently in more than one ethnic group. There are also a few haplotypes shared between Americans of all major ethnicities. For North American Hispanics, some of the haplotypes are ethnically specific and probably resulted from admixture with Amerindians. This is easily demonstrated when studying Mexicans, Mexican Americans, or other Hispanic Americans. In such cases the degree of admixture with African, Asian, and Amerindian haplotypes is variable (Fig. 3).
Shared and ethnic-specific DRB1*, DQB1* haplotypes based on Yunis EJ et al. (B, C), Zuñiga J, Azocar J, Yunis EJ (unpublished results P.R.A.), Vargas-Alarcon G et al., Arnaiz-Villena A et al., Hollenbach JA., et al.


<table>
<thead>
<tr>
<th>DRB1*</th>
<th>DQB1*</th>
<th>CA</th>
<th>AA</th>
<th>AsA</th>
<th>HA</th>
<th>P.R.A</th>
<th>MM</th>
<th>Mazatecan</th>
<th>Zapotec</th>
<th>Mixe</th>
<th>Mixtec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1501, 0602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0101, 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0301, 0201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1101, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1301, 0603</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0701, 0201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0401, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0401, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0404, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1302, 0604</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0701, 0202</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0701, 0303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1001, 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1101, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1102, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0302, 0402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1102, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0804, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1503, 0602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1301, 0303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1101, 0602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0102, 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0901, 0303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0405, 0401</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0803, 0601</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1202, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1401, 0503</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0802, 0402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0407, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1602, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1406, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0411, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1402, 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1502, 0601</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0403, 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1602, 0502</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frequency ≥ 2.0%

Fig. 3  Common HLA-DRB1*, DQB1* haplotypes in different populations.
2.3 Common HLA Haplotypes Defined Serologically and Deduced HLA-Cw/B, HLA-DRB1/DQB1 Common Haplotypes and their Frequencies

We previously summarized the frequency distribution of common HLA haplotypes defined by serological specificities for HLA-A, B, and DR. The allele-level typing of these haplotypes has not been determined systematically in a large database. Figure 4 shows the probable allele-level typing of the most common (serologically defined) HLA haplotypes in four American ethnic groups. The alleles assigned were determined by a combination of deduction and specific studies of individual haplotypes, often found in homozygous cell lines. Not shown in the figure is the fact that there is a large number of Hispanic common haplotypes with a frequency of less than 0.005, which is an indication of the higher genetic diversity of Hispanic Americans.

2.4 The Aggregate Block Frequency (ABF) Measures Genetic Diversity

The sum of frequencies of small blocks, larger blocks, or entire haplotypes within

<table>
<thead>
<tr>
<th>HLA-A</th>
<th>HLA-B*, HLA-Cw*</th>
<th>HLA-DRB1*</th>
<th>HLA-DQB1*</th>
<th>CA</th>
<th>AA</th>
<th>AsA</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0201</td>
<td>0702 0702</td>
<td>1501 0602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0301</td>
<td>0702 0702</td>
<td>1501 0602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0101</td>
<td>0801 0701</td>
<td>0301 0201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3301</td>
<td>1402 0802</td>
<td>0102 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0201</td>
<td>1501 0303</td>
<td>0401 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0301</td>
<td>3501 0401</td>
<td>0101 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0201</td>
<td>4402 0501</td>
<td>0401 0301</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2902</td>
<td>4403 1601</td>
<td>0701 0202</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0101</td>
<td>5701 0602</td>
<td>0701 0303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3001</td>
<td>4201 1701</td>
<td>0302 0402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3001</td>
<td>1302 0602</td>
<td>0701 0202</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2402</td>
<td>0702 0702</td>
<td>0101 0501</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3303</td>
<td>4403 1403</td>
<td>1302 0604</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3303</td>
<td>4403 1403</td>
<td>0701 0202</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0207</td>
<td>4601 0102</td>
<td>0901 0303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2402</td>
<td>5201 1202</td>
<td>1502 0601</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3303</td>
<td>5801 0302</td>
<td>0301 0201</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0201</td>
<td>3501 0401</td>
<td>0407 0302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0201</td>
<td>3512 0401</td>
<td>0802 0402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ Frequency \geq 1.0\% \], [*Haplotypes with frequency \geq 2.0\%.


Frequencies based on Yunis EJ et al., and Cao K et al.

Fig. 4 Common HLA-A*, B*, Cw*, DRB1*, DQB1* haplotypes in different populations.
the MHC, defined by two or more loci sharing significant LD, is named ABF (aggregate block frequency). The frequencies higher than 1% (arbitrarily) of variants of particular blocks are added to compare the degree of diversity of populations studied. It is of interest that the ABFs of individual HLA blocks usually do not show much difference in all ethnicities studied. However, the ABFs of two or more combined blocks show that the Caucasian Americans have less diversity than African- or Asian-Americans. Of interest, the Mexican and Puerto Rican Americans appear to have a higher level of diversity by the use of available information of the DR/DQ block. In preliminary studies, we have determined that the Puerto Rican population of Massachusetts shows a high genetic diversity due to the admixture of Caucasian, African, and Asian haplotypes that are present in frequencies of less than 1%. The ABF for the common haplotypes from HLA-A to HLA-DQB1 is 0.17 in Caucasian Americans, whereas it is 0.03 in African-Americans and 0.10 in Asian-Americans. The ABFs for common haplotypes of Hispanic Americans is approximately 0.06.

### 2.5 Class III Region Blocks

#### 2.5.1 TNF Block

The shortest conserved MHC block studied is of 7 kb size and contains the genes for TNF-α and TNF-β (lymphotoxin-α). The TNF (tumor necrosis factor) microsatellites a, b, and c are upstream (a, b) or within (c) the LTA gene. Also, the TNF block includes the SNPs of the TNF promoter region. Specific variants of this block are present in all populations while others may be population-specific. Specific variants of this block are also haplospecific. For example, the nucleotide A (substituting for the common G nucleotide) at the −243 position is a marker for the African CEH [HLA-A*3001, Cw*1701, B*4201, FC(1,90)0, DRB1*0302, DQB1*0402]. Also, the TNF*237(G-A) SNP is haplospecific for the HLA-B*4001 CEH found often in Caucasians and for at least some of the HLA-B48 haplotypes of East Asians found also in Colombian Amerindians. Likewise, a TNF*-307(G-A), *862(C-A), and *856(C-T) combination is found in both Caucasian and Asian haplotypes. It is not surprising that variants of this block are in LD with both complotypes and the HLA-Cw/B block variants.

#### 2.5.2 Complotype Block

The four MHC-encoded complement genes (C4B, C4A, BF, and C2) are located between HLA-DRB1 (525-kb centromeric to the complement genes) and TNF (350-kb telomeric to the complement genes). This block is variable in length from 75 to 120 kb. The gene order is Telomere-C2-BF-C4A-CYP21A-C4B-CYP21B-Centromere. Complotypes are inherited together as a unit, essentially without recombination. The most frequent haplotypes are SC31 and FC31 detected in all ethnic groups and SC01 frequent among the British. Two complotypes are ethnic specific, the SC21 of the Ashkenazi Jews and the F1C30 of Basques and Sardinians.

### 3 Conserved Extended Haplotypes (CEHs)

Family studies have identified MHC haplotypes that show a significant degree of sequence conservation across a large stretch of the MHC in the region between HLA-Cw, B, TNF, HSP-70, complotype, and DR,DQ. Twelve of the CEHs found in
Caucasians are listed in Fig. 5. For detailed information, consult our earlier reviews.

3.1 Extension of the HLA-Cw/B Block to HLA-A and of the HLA-DR/DQ Block to HLA-DPB1

Figure 4 lists the high-frequency CEHs and common HLA haplotypes that extend to HLA-A. Inclusion of HLA-A alleles diminishes the ABF. At least four CEHs found in Caucasians extend to the HLA-DP region. Strong LD between the HLA-DR/DQ region and HLA-DPB1 was shown for several ethnic groups during the 11th international histo compatibility workshop (IHWS), and more recently in Japanese. LD predictions, however, have reported a lack of nonrandom association between HLA-DR/DQ and DP, perhaps reflecting high levels of recombination within this region and/or the inability of statistical analysis to find genetic fixity or nonrandom associations in relatively small databases of haplotypes or unrelated individuals. In Caucasians, the CEH [HLA-A*0101, Cw*0701, B*0801, SC01, DRB1*0301, DQB1*0201] frequent in both the British and in Ashkenazi Jews, often extends to HLA-DPB1*0401. Also, the CEH [HLA-A*3002, Cw*0501, B*1801, F1C30, DRB1*0301, DQB1*0201] of the Basques and Sardinians often extends to HLA-DPB1*0202.

3.1.1 Frequency Distribution of MHC Blocks, CEH Fragments, and CEHs

Table 1 shows the frequency distribution of DNA blocks within a region of MHC of approximately 1.5 MB, between HLA-Cw and HLA-DQB1, in a study of 372 haplotypes determined in 92 pedigree-analyzed Caucasian families. Figure 5, showing the distribution of these haplotypes as analyzed by our previously published model, demonstrates that, of the 12 most frequent Caucasian CEHs, only 9 CEHs show nonrandom association with significant LD (p < 0.0005). Interestingly, approximately one-third of the haplotypes are composed of combinations of blocks in random association. Therefore, some CEHs cannot be found with statistical significance in this limited database. However, the small HLA-Cw/B and DR/DQ blocks were found to have significant LD corroborating those described previously. These data are the first demonstration of the frequency distribution of blocks in genotyped families (n = 92), although the data for complotypes were deduced from previous studies. The ABF measurement is comparable to those that had been deduced by us from separate reports of HLA-Cw/B and DR/DQ blocks. The high frequency of the complotype SC31 explains why CEHs carrying it require a larger database to produce LD to reach statistical significance.

3.1.2 The Global LD Measurement Related to the Size of the Database

As had been reported before, global LD measurements of large size DNA blocks or haplotypes are not informative when the size of the data studied is small. However, there were several small blocks and fragments of CEHs that can be found even with a limited number of haplotypes as demonstrated in the analyses of the data shown in Fig. 5 and Table 1.

3.1.3 Association of CEHs with Disease

Disease susceptibility or protective alleles of particular loci are included in several CEHs or their associated fragments. Important studies of pedigree-analyzed CEHs have been conducted to localize disease susceptibility genes, including type 1 diabetes, IgA deficiency,
### Block combinations

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
<th>XI</th>
<th>XII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cw</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TNF microsatellites

| 2, 3, 1 | 11, 4, 1 | 6, 5, 1 | 2, 5, 2 | 4, 5, 1 | 7, 4.1 | 7, 4.1 | 2, 1, 2 | 2.1, 2 | 5, 5, 2 | 10, 4, 1 | 5, 5, 2 |

### Complotype

| SC01 | SC31 | SC30 | SC61 | SC02 | FC31 | SC31 | SC2(1,2) | SC33 | SC31 | SC31 | FC(3,2)0 |

### HLA alleles

- **Cw**: 0701, 0702, 0501, 0602, 0304, 1601, 0602, 0802, 0304, 0401, 0401, 0401
- **B**: 0801, 0702, 4402, 5701, 4001, 4403, 1302, 1402, 1501, 3501, 3501, 3501
- **DRB1**: 0301, 1501, 0401, 0701, 1302, 0701, 0701, 0102, 1401, 1101, 0401, 0101
- **DQB1**: 0201, 0602, 0301, 0303, 0604, 0202, 0202, 0501, 0503, 0301, 0301, 0501
- **DQA1**: 0501, 0102, 0302, 0201, 0102, 0201, 0201, 0101, 0104, 5, 302, 0101

#### Cw, B TNF C' DR/DQ 109 Conserved

- Total: 109
- Extended haplotypes: 130
- Shared haplotypes: 61

#### Fig. 5

Model of four blocks or fragments of DNA within the MHC in CEH. Using 16 kinds of haplotypes, the solid lines represent stretches of conserved DNA.
Tab. 1  Frequencies of MHC blocks (Cw/B, TNF, complotype, and class II genes) in a Caucasian population of 92 families.

<table>
<thead>
<tr>
<th>Blocks</th>
<th>Published</th>
<th>Observed</th>
<th>Blocks</th>
<th>Observed</th>
<th>Blocks</th>
<th>Observed</th>
<th>Blocks</th>
<th>Published</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF</td>
<td></td>
<td></td>
<td>HF</td>
<td></td>
<td>HF</td>
<td></td>
<td>HF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0701,0801</td>
<td>0.111</td>
<td>0.108</td>
<td>2,3,1</td>
<td>0.118</td>
<td>SC01</td>
<td>0.120</td>
<td>0301,0201,0501</td>
<td>0.056</td>
<td>0.112</td>
</tr>
<tr>
<td>0702,0702</td>
<td>0.107</td>
<td>0.140</td>
<td>11,4,1</td>
<td>0.188</td>
<td>SC31</td>
<td>0.340</td>
<td>1501,0602,0102</td>
<td>0.085</td>
<td>0.139</td>
</tr>
<tr>
<td>0501,4402</td>
<td>0.091</td>
<td>0.070</td>
<td>6,5,1</td>
<td>0.075</td>
<td>SC30</td>
<td>0.067</td>
<td>0401,0301,0302</td>
<td>0.030</td>
<td>0.086</td>
</tr>
<tr>
<td>0602,5701</td>
<td>0.039</td>
<td>0.032</td>
<td>2,5,2</td>
<td>0.026</td>
<td>SC61</td>
<td>0.030</td>
<td>0701,0303,0201</td>
<td>0.024</td>
<td>0.032</td>
</tr>
<tr>
<td>0304,4001</td>
<td>0.064</td>
<td>0.026</td>
<td>4,5,1</td>
<td>0.038</td>
<td>SC02</td>
<td>0.030</td>
<td>1302,0604,0102</td>
<td>0.023</td>
<td>0.024</td>
</tr>
<tr>
<td>1601,4403</td>
<td>0.017</td>
<td>0.026</td>
<td>7,4,1</td>
<td>0.048</td>
<td>FC31</td>
<td>0.035</td>
<td>0701,0201,0201</td>
<td>0.031</td>
<td>0.043</td>
</tr>
<tr>
<td>0602,1302</td>
<td>0.013</td>
<td>0.019</td>
<td>2,1,2</td>
<td>0.035</td>
<td>SC2(1,2)</td>
<td>0.016</td>
<td>0102,0501,0101</td>
<td>0.010</td>
<td>0.021</td>
</tr>
<tr>
<td>0802,1402</td>
<td>0.032</td>
<td>0.022</td>
<td>5,5,2</td>
<td>0.040</td>
<td>SC33</td>
<td>0.016</td>
<td>1401,0302,0501</td>
<td>–</td>
<td>0.021</td>
</tr>
<tr>
<td>0304,1501</td>
<td>0.011</td>
<td>0.019</td>
<td>10,4,1</td>
<td>0.019</td>
<td>FC(3,2)0</td>
<td>0.010</td>
<td>1101,0301,05</td>
<td>0.037</td>
<td>0.032</td>
</tr>
<tr>
<td>0401,3501</td>
<td>0.064</td>
<td>0.070</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0101,0501,0101</td>
<td>0.060</td>
<td>0.027</td>
</tr>
</tbody>
</table>


*cYunis, E.J. et al. (2005).

*dComplotypes were deduced on the basis of the association between Cw/B, TNF, and complotypes. The term “blocks” was defined as published. LDs were significant between Cw/B and TNF, and complotypes. The term “blocks” was defined as published.
dermatitis herpetiformis, and gluten-sensitive enteropathy. Likewise, fragments of the CEH [HLA-Cw*0303, B*5502, SB45, DRB1*1401, DQB1*0503] served to map pemphigus vulgaris (PV) genetic susceptibility in non-Jewish patients to the DR, DQ region. But, PV in Jewish patients is associated with the entire Ashkenazi CEH [HLA-A*2601, Cu*1203, B*3801, SC21, DRB1*0402, DQB1*0302]. One of the most frequent CEH of Caucasians, [HLA-A*0301, Cu*0702, B*0702, SC31, DRB1*0501, DQB1*0602], has been found to be associated with multiple sclerosis (MS) and with ragweed allergy. In another study, we had prior evidence that the CEH [HLA-B8, SC01, DR3] was increased among nonresponders to hepatitis B vaccine and that such subjects were more homozygous than expected. We prospectively immunized homozygotes and heterozygotes and none of the homozygotes responded, confirming the dominant role of the MHC in the immune response.

5 Paralogous Chromosomal Regions

Paralogous chromosomal regions are those genomic regions found in a single species, which are thought to have a common genetic origin, but which have been separated by mechanisms of duplication from an ancestral genetic unit. Other hypotheses suggest that perhaps independently duplicated genes were grouped during evolution by selective forces. Paralogous genes have been found in different chromosomal segments in humans. The early findings demonstrate the existence in chromosomes 11 and 12 of duplicated chromosomal segments in which two pairs of duplicate genes were contained. In general, the block duplication mechanism can better explain why genes paralogous to some found in the MHC are clustered in three other specific regions of the human genome.

Several published studies have described the existence of three chromosomal regions paralogous to the MHC: 1q21-25, 9q33-34, and 19p13.4–13.1. The possible origin of MHC paralogous regions on chromosomes 1, 9, and 19 could be from two rounds of duplication of the whole genome (2R hypothesis). These findings are supported by the identification of paralogous genes in Drosophila melanogaster and in other species in human chromosomes 1, 9, and 19. Phylogenetic analyses have revealed that the MHC and its paralogous regions have a common (ancestral) genomic region that was duplicated as
a block after the divergence of cephalochordates and vertebrates. Duplications (polyploidizations) could have occurred between 766 and 528 Myr ago. Interestingly, the MHC paralogous region in chromosome 9q32-34 retains an ancestral state in the context of organization and gene substitution patterns compared with both the human 1p31-p11, 19p13, and 6p21.3 regions and with the amphibious, chordate, and gnathostomata proto-MHC regions. The MHC paralogous genes are located in the classical MHC class I and class II regions, but also can be found in neighboring regions of the MHC. Gene families such as NOTCH, BRD, and PBX have copies in all of the MHC paralogous regions, but some genes that have paralogous copies in chromosomes 1, 9, and 19 are not found in the MHC.

Among the other previously mentioned paralogous regions, the 1q21-25 is unique because it contains histocompatibility-like loci in the CD1 and MR1 regions. CD1 molecules show structural and functional similarities in relation to HLA class I molecules. The CD1 molecules can present certain glycolipids to cytotoxic T lymphocytes. Genomic sequence of the human chromosome 9 reveals a highly polymorphic structure with a total of 31 genes paralogous to genes found in the human MHC. Interestingly, in chromosome 9, more than 90 genes associated with different human diseases have been mapped. These include AIF1-L, BAT2-L, C5, PBX3, BRD3, COL1A5, and RXRA. This fact is important in attempting to decipher the possible influence of several loci in polygenic diseases. Also of interest is that, in the extended MHC class I region of chromosome 6, families of olfactory receptor genes have been mapped also in the chromosome 9 paralogous region. This fact could be a determinant in the species preservation and reproduction, as recent reports suggested that mate selection could be influenced by HLA segregation of specific parental haplotypes that could be in LD with specific olfactory related genes in the proximity of MHC.

6 Discussion

High-resolution definition of the most frequent MHC haplotypes extending from at least HLA-Cw to HLA-DQB1 is incomplete at the phenotype level, much less at the genotype level using pedigree analysis. We summarized the current knowledge of allelic variants of well-studied loci in common HLA haplotypes and, to a limited extent only, in MHC CEHs. Using the frequencies of nonrandomly associated alleles of two close loci, HLA-Cw, -B in one example and HLA-DRB1, -DQB1 in another, we described blocks of DNA measuring approximately 100 to 150 kb each. We deduced that the existence of common HLA haplotypes, defined as the nonrandom association of these two blocks, is related to genetic fixity within the MHC spanning at least 1.4 Mb as exemplified by the existence of CEHs. When typing of the intermediate class III region of the MHC (complotypes, TNF, and HSP-70) and haplotype identification by pedigree analysis are conducted, approximately 85% of common HLA haplotypes are the same as genetically fixed CEHs. We showed in Fig. 5 that approximately 22% of common HLA haplotypes are not CEHs because the intermediate regions have different degrees of LDs (often related to separate CEH variants).

Although several of the MHC loci are among the most polymorphic known in the human genome, they represent only a small percentage of the total MHC DNA
sequence. Formal proof of the essential fixity of the complete DNA sequence for specific MHC CEHs shall require DNA sequencing of the entire region in several independent examples. This work has begun on consanguineous cell lines, with the MHC Haplotype Project. If one expands the CEH definition to include the loci from HLA-A to HLA-DPB1, the fixed DNA length would be more than 3 Mb of DNA.

The frequency of block combinations varies between major ethnic groups and/or in different nationalities. The preponderance of common haplotypes in Caucasians may be due to the fact that more Caucasians than Africans or Asians have been studied or may simply reflect a lower genetic diversity among Caucasians. The preponderance of common haplotypes in Caucasians may be due to the fact that more Caucasians than Africans or Asians have been studied or may simply reflect a lower genetic diversity among Caucasians.

Identifying CEHs in a population depends on the size of the haplotype database. Lower frequency CEHs can only be identified in larger databases. It is important to keep in mind that the number of and the particular types of CEHs (or common haplotypes) detected will also depend on the type (e.g. ethnicity, disease status) of the population studied. Some CEHs, such as the first six listed in Table 1 and Fig. 5, are relatively frequent in healthy Caucasians. In a database of 500 Caucasian haplotypes, we would predict the aggregate frequency of all of the CEHs listed in Fig. 5 (not including HLA-A) to be at least 30%. In this review, we reported an analysis of 372 haplotypes derived from pedigree analysis of 92 parents, although the complotypes were deduced from probable nonrandom associations using our previous studies. Only nine CEHs showed significant nonrandom associations. The number of individuals studied would also affect the distribution of CEHs, their fragments, and small blocks. The arrows in Fig. 5 suggest sites of possible historical recombination that in many recent articles using SNPs are referred to as hotspots.

Structural analysis of MHC haplotypes is incomplete, but we believe that the available evidence demonstrates that there are fixed genetical segments of DNA within the MHC, the frequency of which vary in different human populations. Some CEHs have shown significant genetic fixity across blocks of DNA as large as 3.2 Mb from HLA-A to -DPB1. However, the most completely described segment spans HLA-Cw and -DQB1, and includes polymorphisms of complement genes (complotypes) and TNF. The boundaries of this 1.5-Mb block mark CEHs that are relatively frequent, which have proven informative for mapping disease susceptibility and immune response genes and they could be used for allotransplantation matching.

The concept of the aggregate frequency of individual common block variants spanning the MHC segment between HLA-Cw/B and HLA-DR/DQ, or ABF, may be used to determine the degree of relative diversity of Americans of Asian, Caucasian, African, or Hispanic ancestry. Likewise, we suggest that determining the ABF of a combination of small blocks can be used to estimate the degree of genetic fixity of even larger segments of DNA called common HLA haplotypes (which are usually representative of CEHs). Thus, ABF is a measurement of the frequency of DNA blocks of any particular size range, based on the aggregate frequency of nonrandomly associated alleles at two or more loci, which estimates the frequency of various sizes of genetically fixed DNA within the MHC of well-defined populations. The
result can be used to compare human diversity among well-defined populations.

Other DNA methods, such as RFLP (restriction fragment length polymorphism), SNP analysis, and microsatellite polymorphisms, have also been useful in defining extended genomic variation throughout the MHC at sites other than the classical HLA loci (e.g. complement and TNF locus typing).Allele-level typing of HLA-A, -Cw, -B, -DRB1, and/or -DQB1 has been reported for a limited number of Americans of different ethnicities. At least three characteristics of these MHC haplotype blocks are quite different from those recently reported. MHC blocks described here are never defined solely by SNP typing and are highly polymorphic within well-characterized populations. In contrast to two to five common SNP-defined haplotype variants present in diverse ethnic groups, MHC blocks have many more common variants even within a single ethnic group. Finally, MHC haplotype variants exhibit a high degree of isolated population and/or ethnic specificity. Furthermore, despite both the likelihood of many recombination “hotspots” within the MHC and this great diversity within specific MHC blocks, many CEHs exist at relatively high frequency. Specific alleles or specific MHC blocks are often “haplospecific” for particular CEHs. In many cases, one particular allele at one locus is highly correlated with an entire MHC CEH. In other cases, a specific set of alleles within a single MHC block or in more than one block is required to achieve a high correlation with a particular CEH. However, haplotype determination by family study is required to rigorously define a complete haplotype.

Many MHC haplotypes are not CEHs but contain only fragments or even only a single small block from any specific CEH. Non-CEH haplotypes containing CEH fragments are represented in Fig. 5 by any haplotype except #1 (the full CEH) or #16 (containing no part of any CEH). For example, not all haplotypes containing the (HLA-Cw*0701, B*0801) block exist solely as the CEH [HLA-Cw*0701, B*0801, TNF 2,3,1, SC01, DRB1*0301, DQB1*0201] (although, we argue, the majority are contributed by this CEH). The distribution in frequencies between any given CEH and its specific blocks depends on several factors, including at least the degree to which a particular block exists in more than one extended haplotype (i.e. is not haplospecific) and the degree of fragmentation of the extended haplotype in the study population. If a disease gene is mapped to a CEH, the number of haplotypes bearing fragments of that CEH that is required to localize the gene will be greater than the number of “random” haplotypes required. If the gene is already mapped to a small block, variants of that block that are not associated with any CEH required to localize the gene should be fewer, because the complete set of these “random” variants should share little that is unique except for the susceptibility locus. For these reasons, it is probably more difficult to map susceptibility genes in American Caucasians than in other American ethnicities simply because the level of genetic diversity is lower in Caucasian Americans. We also believe that most of the diversity is possibly related to the degree of genetic admixture of individuals recruited, and this admixture could be within nationalities or between ethnicities. An example of this complexity can be the use of the term Hispanic to describe individuals that originated from Spanish-speaking countries. Of course, such individuals probably resulted from different degrees of population admixture.
producing stratification, and they represent many different ethnicities. The use of SNPs or short tandem repeats or variants of genes that could have been conserved in all humans or within a particular ethnicity will need to be used to rule out population genetic stratification in every study population. In this case, DNA blocks that are conserved in the genome need to be identified in order to distinguish them from DNA blocks that show variability caused by population admixture or by natural selection.

CEHs might result from many possible mechanisms, including recent population bottlenecks, recombination suppression, preferential transmission, migration and admixture, and/or genetic drift or natural selection. Duplication of genes involved in immune functions argues in favor of selection for blocks of DNA, which vary in different populations. Paralogous genes to some found in the human MHC with equal or related functions are located in other chromosomes (Chromosomes 1, 9, and 19). Such duplicated genes add another dimension of difficulty to disease gene localization because there may be several genes involved in a disease, which are molecularly related. For example, the NOTCH genes are found in the four paralogous chromosomal regions. One caveat, however, is that the MHC paralogous region of chromosome 9 contains many other genes that may be in LD with NOTCH1, but the equivalent genes to those are not present in chromosome 6. In the case of schizophrenia, NOTCH4 could be in LD with another marker different from that of NOTCH1 and it could suggest a difference in the populations studied. However, the BRD2 and BRD3 genes of chromosomes 6 and 9, respectively, are at similar distances to the C4 genes of chromosome 6 and the C5 of chromosome 9. The latter region is expected to have genes in LD that may be acting synergistically.

7  
Future Studies

Complete human population studies based on the CEH and DNA blocks concept are needed. Requirements will include collecting data on the basis of geography (nationality), language, and ethnicity, as well as identifying other (non-MHC) genetic markers. The latter include the Y chromosome, mitochondrial DNA, blood groups, and short tandem repeats outside the MHC region. Such analysis could derive the age of haplotypes and population diversity relationships. Genetic admixtures need to be analyzed in order to know the markers that indicate genetic drifts or old admixtures. We believe that any unique marker of the polymorphic loci of a haplotype would be a marker of the age of the MHC haplotype. Two or more ancient mutations or gene conversions inherited together will have random frequencies, but during evolution they are fixed at the population level and may prove to have selective advantage. Natural selection is often mentioned as a major influence in MHC evolution and in the formation of CEH. However, nonrandom association between distant loci may be the result of directional selection (molecular cooperation during the immune response). In this context, the possible role of paralogous genes may be operating. Alternatively, balancing selection by a higher replacement rate of different alleles at certain loci could have been an influence, particularly a departure from neutral expectation of HLA frequencies.
8 Summary

We described an updated interpretation or definition of MHC or HLA haplotypes. Although the information is incomplete, we believe that there are genetically fixed MHC haplotypes of variable sizes in the human population and that the stretch of fixed DNA could be up to 3.2 Mb of DNA in several haplotypes that extend from HLA-A to HLA-DPB1. However, the fixity of the DNA between HLA-Cw and HLA-DQB1 has been studied more extensively, using polymorphisms of the complement system (complotype) and the TNF region to define that stretch of fixed DNA of approximately 1.5 Mb. The boundaries of this large block, between HLA-Cw and HLA-DQ, mark MHC CEHs with fixed (or nearly so) DNA that are relatively frequent and have been used for mapping studies of disease and immune responses, and they could be used for matching for allotransplantation. The aggregate block frequencies (ABFs) of both the HLA-Cw/B and HLA-DRB1/DQB1 small blocks were investigated in three different ethnic groups living in the United States and in five African populations. The results suggest that the aggregate frequency of the most common small blocks in strong and significant LD may be used to determine the degree of relative diversity of Asian, Caucasian, or African-Americans and Hispanic populations. Likewise, we suggest that determining the ABF of a combination of specific small blocks can be used to determine the degree of genetic fixity of even larger segments of DNA. Thus, ABF is a measurement of the total frequency of fixed genetic (high population frequency) variants of a particular DNA block, to estimate the frequency of various sizes of DNA within the MHC of well-defined populations. This measurement adds a new dimension in the studies of human diversity. Essentially, we have developed a model of combinations of blocks of 16 kinds of MHC haplotypes, with one of them being the CEH, seven of them being CEH fragments containing combinations of the CEH blocks, seven more having only one CEH block or two separated blocks, and one of them containing only rare blocks that do not show any LD. We used as an example the higher degree of genetic diversity of the Mexican Mestizos as compared with Caucasian Americans or specific Amerindian subgroups to demonstrate that diversity as measured by ABF can vary related to the contribution of the genetic diversity of the populations that contributed to a particular population admixture. We anticipate that genetic diversity will also differ between supposedly ethnically similar individuals when they are recruited from different geographic regions. Although not defined in the same manner, as recently reported blocks of SNPs, previously identified MHC blocks are essentially genetic units. We discussed the problem of attempting to use SNP blocks to localize disease genes without previous assignment of MHC blocks determined by high-resolution typing and preferably by pedigree analysis. We suggest that after this is done, it is possible to map disease or protection genes using positional cloning and high-throughput technologies. Disease genes could be located in either large or small blocks of DNA, but we believe that it will be easier to look for them in random blocks not associated with high-frequency CEHs. We also discussed the possibility of applying the techniques and analytical methodologies developed for the MHC to other chromosomes of the human genome, with the thought that other regions paralogous to the MHC might be
best suited for initial work. Studies will be required to compare genes that are involved in the same pathways or networks, especially those that have evolved with high levels of polymorphism. Individual locus and haplotype variation in these paralogous regions characteristic of both specific ethnicities and specific diseases need to be studied to determine whether paralogous genes with similar or related functions might explain ethnic variation in the pathogenesis of certain diseases such as autoimmune diseases.

**Acknowledgment**

Yunis and Zuñiga were supported by Public Health Service grants HL29583 and HL59838 from the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) and a gift from Granados. Larsen, Awdeh, and Alper were supported by Public Health Service grant HL29583 from the National Heart, Lung, and Blood Institute of the NIH.

See also Genetic Analysis of Populations; Genetic Variation and Molecular Evolution; Genetics, Molecular Basis of; Human Genetic Variation and Disease; Targeted Therapy: Genomic Approaches.

**Bibliography**

**Book and Reviews**


Primary Literature


