Am. J. Hum. Genet. 73:580–590, 2003

An Integrated Haplotype Map of the Human Major Histocompatibility Complex

Emily C. Walsh,[1] Kristie A. Mather,[2] Stephen F. Schaffner,[1] Lisa Farwell,[1] Mark J. Daly,[1] Nick Patterson,[1] Michael Cullen,[3] Mary Carrington,[3] Teodorica L. Bugawan,[4] Henry Erlich,[4] Jay Campbell,[1] Jeffrey Barrett,[1] Katie Miller,[1] Glenys Thomson,[2] Eric S. Lander,[1] and John D. Rioux[1]

[1]Center for Genome Research, Whitehead Institute for Biomedical Research, Cambridge, MA; [2]Department of Integrative Biology, University of California, Berkeley; [3]Basic Research Program, Science Applications International Corporation–Frederick, Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD; and [4]Roche Molecular System, Department of Human Genetics, Alameda, CA

Numerous studies have clearly indicated a role for the major histocompatibility complex (MHC) in susceptibility to autoimmune diseases. Such studies have focused on the genetic variation of a small number of classical human-leukocyte-antigen (*HLA*) genes in the region. Although these genes represent good candidates, given their immunological roles, linkage disequilibrium (LD) surrounding these genes has made it difficult to rule out neighboring genes, many with immune function, as influencing disease susceptibility. It is likely that a comprehensive analysis of the patterns of LD and variation, by using a high-density map of single-nucleotide polymorphisms (SNPs), would enable a greater understanding of the nature of the observed associations, as well as lead to the identification of causal variation. We present herein an initial analysis of this region, using 201 SNPs, nine classical *HLA* loci, two *TAP* genes, and 18 microsatellites. This analysis suggests that LD and variation in the MHC, aside from the classical *HLA* loci, are essentially no different from those in the rest of the genome. Furthermore, these data show that multi-SNP haplotypes will likely be a valuable means for refining association signals in this region.

Introduction

The major histocompatibility complex (MHC) represents the most intensively studied 4 Mb in the human genome. Associations between autoimmune disease and alleles of genes in the region are among the most consistent findings in human genetics (Price et al. 1999; Beck and Trowsdale 2000). Historically, attempts to characterize the region have focused on a handful of highly variable, classical human-leukocyte-antigen (*HLA*) genes (class-I genes: *HLA-A, HLA-B,* and *HLA-C*; and class-II genes: *HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1,* and *HLA-DPB1*). These genes encode cell-surface molecules that present antigenic peptides to T cells, thereby initiating acquired immune response to invading pathogens and other foreign antigens. However, the classical *HLA* loci represent a minority of the genes found in the MHC region, since at least another 120 genes are present (Beck and Trowsdale 2000). By focusing on just the classical *HLA* genes, one may overlook other disease-influencing variation in the region. A more uniform, comprehensive map of the commonly linked variation—that is, a haplotype map—will help to discriminate between causal alleles and variation that is merely in linkage disequilibrium (LD) with them. Such a resource will also allow a more complete description of the haplotype structure and, potentially, insight into the evolutionary and recombinational history of the region.

With these goals in mind, we set out to build a SNP haplotype map of the region. To be able to integrate this map with the wealth of findings from association studies, we genotyped 201 reliable, polymorphic, evenly spaced SNPs (target density: one SNP every 20 kb) in 136 independent chromosomes also genotyped for nine *HLA* genes, two *TAP* genes (involved in antigen processing), and 18 microsatellites. Markers were genotyped in families (18 multigenerational European pedigrees) to allow direct assessment of chromosomal phase and, thus, simple reconstruction of haplotypes. Using these SNP data, we examined the haplotype patterns of the region and mapped these patterns, relative to both genetic and physical distance, as assayed by an exceedingly high-resolution recombination map (fig. 1). This recombination map is the result of the analysis of 20,000 sperm meioses from 12 men (Cullen et al. 2002). Although this SNP density does not represent a "com-

Received May 27, 2003; accepted for publication June 30, 2003; electronically published August 14, 2003.

Address for correspondence and reprints: Dr. John D. Rioux, Whitehead Institute/MIT, Center for Genome Research, One Kendall Square, Building 300, Cambridge, MA 02139. E-mail: rioux@genome.wi.mit.edu

0002-9297/2003/7303-0011$15.00

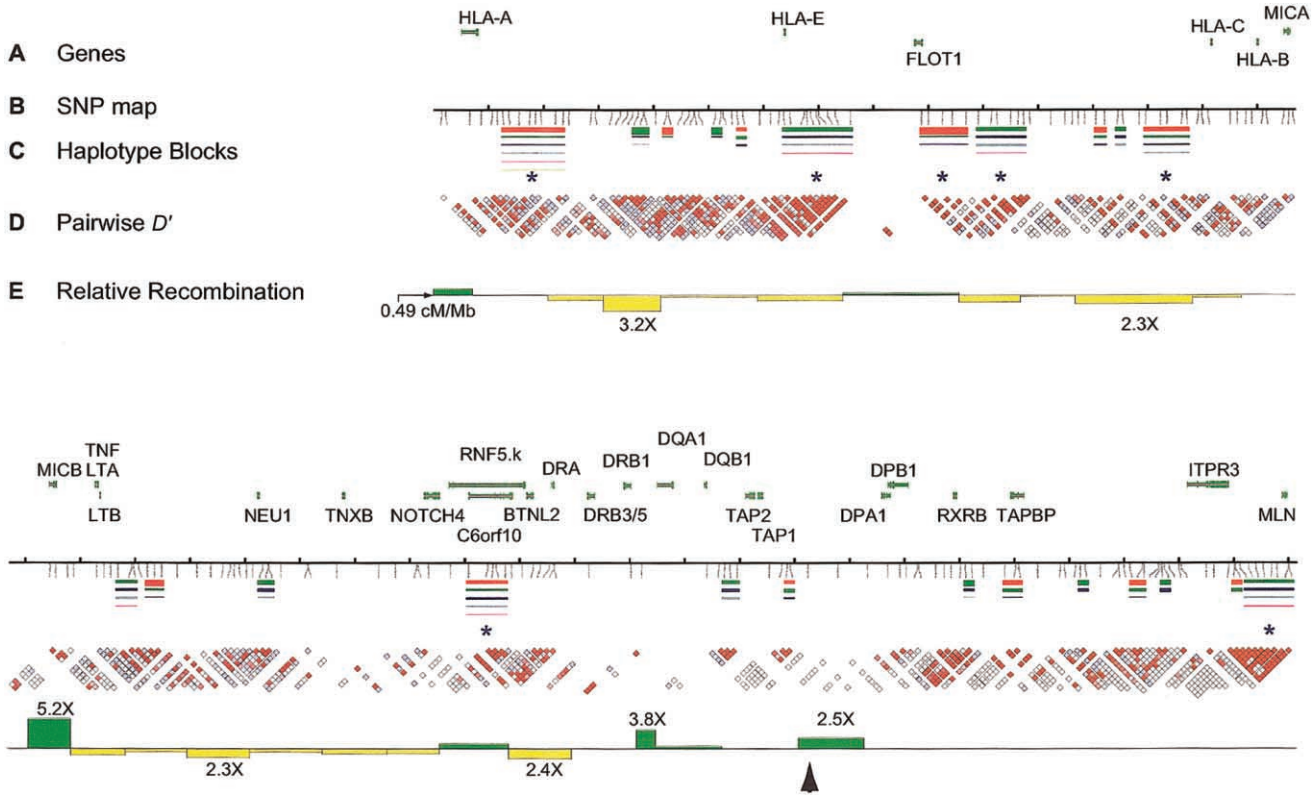


Figure 1 Integrated SNP map of the 4-Mb MHC in CEPH Europeans. *A,* Location and exon-intron structure provided for a subset of genes above the map, for positional reference. *B,* The 201 reliable, polymorphic SNPs are indicated on the map with ticks below the line. Ticks above the line are placed with 100-kb spacing. *C,* Haplotype blocks indicated below and common haplotype variants (>3% frequency) shown as colored lines (thickness indicates relative population frequency). Colors serve only to distinguish haplotypes and do not indicate block-to-block connections. Asterisks are found below the seven largest haplotype blocks. *D,* Pairwise D' values (Lewontin 1964) for SNPs indicated below the haplotype blocks. Note that each block represents a single D' calculation and is placed in the middle between the two SNPs analyzed. This is the same information as contained in the traditional D' plot in the IDRG supplementary figure 1; only the data have been plotted in the *X*-axis, for ease of viewing with respect to the physical map. Red indicates strong LD and high confidence of the D' estimate ($D' > 0.95$; LOD ≥ 3.0). Blue indicates strong LD with low confidence of the estimate of D' ($D' = 1$; LOD < 3.0). White indicates weak LD. *E,* Relative recombination rate, which is based on the sperm meiotic map, indicated in bar-graph form, where the value on the line is the regional average, 0.49 cM/Mb. Green bars indicate recombination rates >0.49 cM/Mb, and yellow bars indicate rates <0.49 cM/Mb. The black arrowhead denotes approximate location of well-mapped recombination rate from Jeffreys et al. (2001). SNP marker density in that region is too low to comment on any similarities between our studies. Note that five of seven long haplotype blocks map to regions where the recombination rate is ≤0.49 cM/Mb. The remaining two long blocks are found in domains where recombination rates are 0.64 cM/Mb and 0.83 cM/Mb (rates below or near, respectively, the genomewide average).

plete" haplotype map, it is a large first step toward a comprehensive characterization of the patterns of common variation in the MHC. Here, we use this map to first explore the structure of LD in the region, with respect to both haplotype blocks and extended haplotypes. Next, we examined SNP-haplotype variation in the MHC, first considering regions between the classical *HLA* loci and then examining SNP-haplotype variation across these genes. We also examined whether the SNP-haplotype diversity near classical *HLA* loci contained enough information to predict the *HLA* allele carried on the chromosome.

Materials and Methods

DNA Samples

Samples were obtained from the Coriell Cell Repository and drawn from the collection of Utah CEPH pedigrees of European descent. One hundred thirty-six independent, grandparental chromosomes were used for haplotype construction. Of these chromosomes, 96 were in common with Gabriel et al. (2002) and, therefore, were used for comparison with the genomewide LD structure.

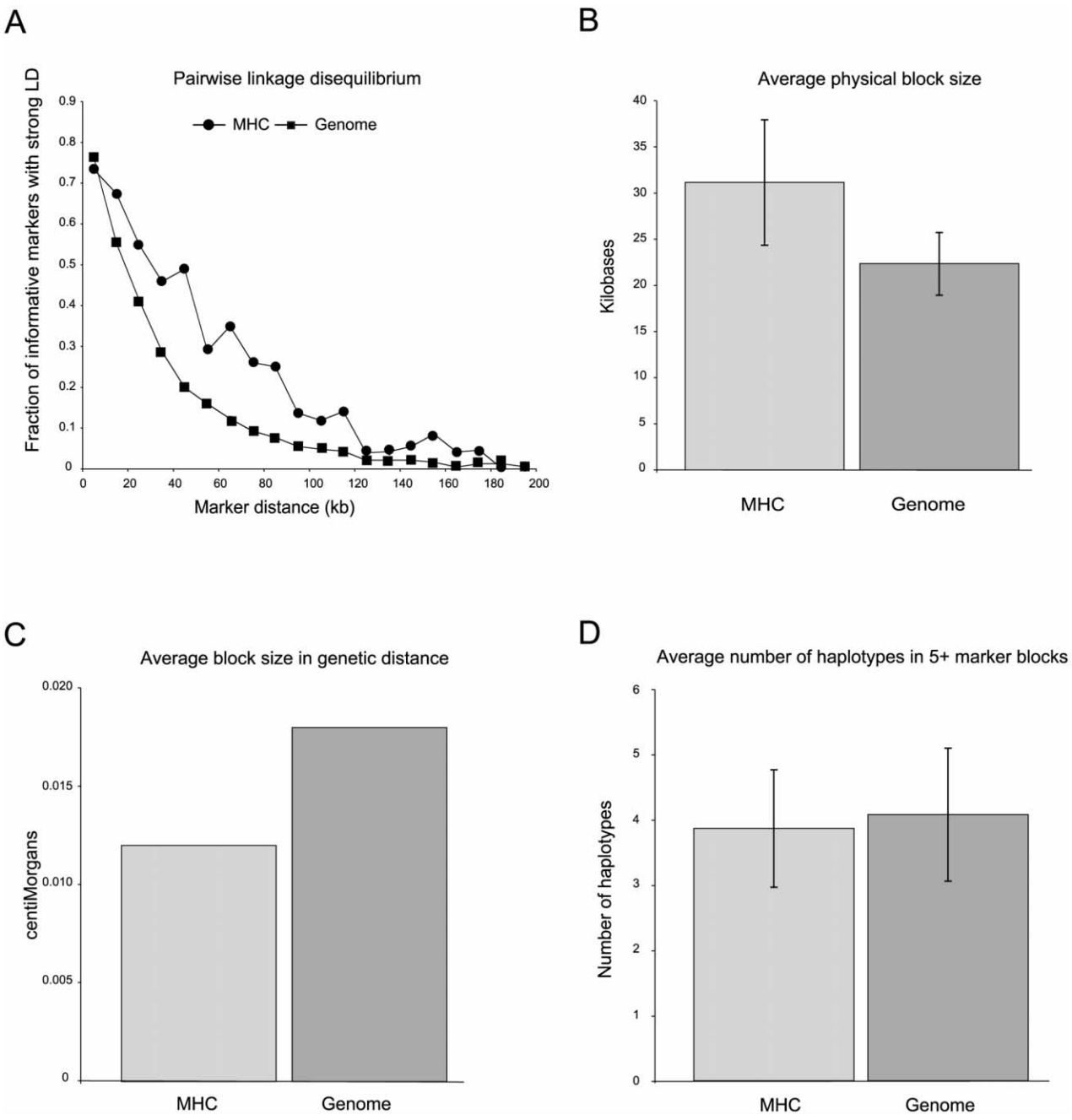


Figure 2 Block comparison between the MHC and other autosomal regions from a genomewide survey. *A*, Plot of LD by physical distance revealing that LD is extended in the MHC. *B*, Accordingly, the average physical length of blocks in the MHC is longer than in the rest of the genome. *C*, However, measured by genetic distance, we observe that block size in the MHC is somewhat less than in the rest of the genome. *D*, The number of haplotype variants in blocks not spanning classical *HLA* genes is the same as elsewhere in the genome.

Identifiers for all individuals can be found at the Inflammatory Disease Research Group (IDRG) Web site.

Genotyping and Data Checking

All SNPs for which genotyping was attempted were publicly available at the dbSNP Web site. SNPs were selected mainly to achieve a desired spacing (1/20 kb); however, SNPs with more than one submitter were preferentially chosen. SNP primers and probes were designed in multiplex format (average fivefold multiplexing) with SpectroDESIGNER software (Sequenom). A total of 435 assays were designed. Assays were considered successful

and genotype data were included in our analyses if they passed all of the following criteria: (1) a minimum of 75% of all genotyping calls were obtained, (2) markers did not deviate from Hardy-Weinberg equilibrium, and (3) markers had no more than one Mendelian error. These criteria defined 201 successful assays. Genotype calls for successful markers were then set to zero for any single Mendelian error. All of these working assays had minor allele frequencies >5%, and 89% of these assays had minor allele frequencies >10%. Overall, for successful markers, 97.6% of all attempted genotypes were obtained. The entire list of SNP assays, as well as detailed genotyping information, can be found at the IDRG Web site.

Four-digit *HLA* types were determined for *HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DMB1, HLA-DQA1, HLA-DQB1, HLA-DPA1,* and *HLA–DPB1,* as described elsewhere (Begovich et al. 1992; Carrington et al. 1994, 1999; Moonsamy et al. 1997; Bugawan et al. 2000). Typing was performed twice independently, and conflicting types were resolved, in most cases, by two independent retyping experiments. *TAP1* and *TAP2* were genotyped as described elsewhere (Carrington et al. 1993). D6S2971, D6S2749, D6S2874, D6S273, D6S2876, D6S2751, D6S2741, and D6S2739 were typed as described elsewhere (Martin et al. 1998). Genotyping details for the 11 remaining microsatellites can be found as supplemental information on the IDRG Web site. D6S2972 and D6S265 genotypes were typed twice (IDRG; Martin et al. 1998), and conflicts were resolved by retyping. Alias details for all microsatellites are provided elsewhere (Cullen et al. 2003).

D′ *Confidence Limits, Definition of Haplotype Blocks, and Structure Comparison*

Pairwise *D′* values—estimates of the strength of LD (Lewontin 1964)—for SNP markers were assessed and haplotype blocks were defined as per Gabriel et al. (2002). In brief, *D′* confidence limits were determined by calculating the probability of the observed data for all possible values of *D′,* from which an overall probability distribution was determined. For all blocks identified, the outermost marker pair was required to be in strong LD, with an upper confidence limit (CU) > 0.98 and a lower confidence limit (CL) > 0.7. Blocks defined by only two markers required confidence bounds of CL > 0.8 and CU > 0.98 and an intervening distance of ⩽20 kb; for three consecutive markers, all pairs had to have confidence bounds of CL > 0.5 and CU > 0.98 and an intervening distance of <30 kb; and for four markers, the fraction of informative pairs in strong LD (CL > 0.7 and CU > 0.98) was required to be >95%, with an intervening distance of <30 kb. For runs of five or more markers, the fraction of informative pairs in strong LD was required to be >95%, and markers were allowed to span any distance.

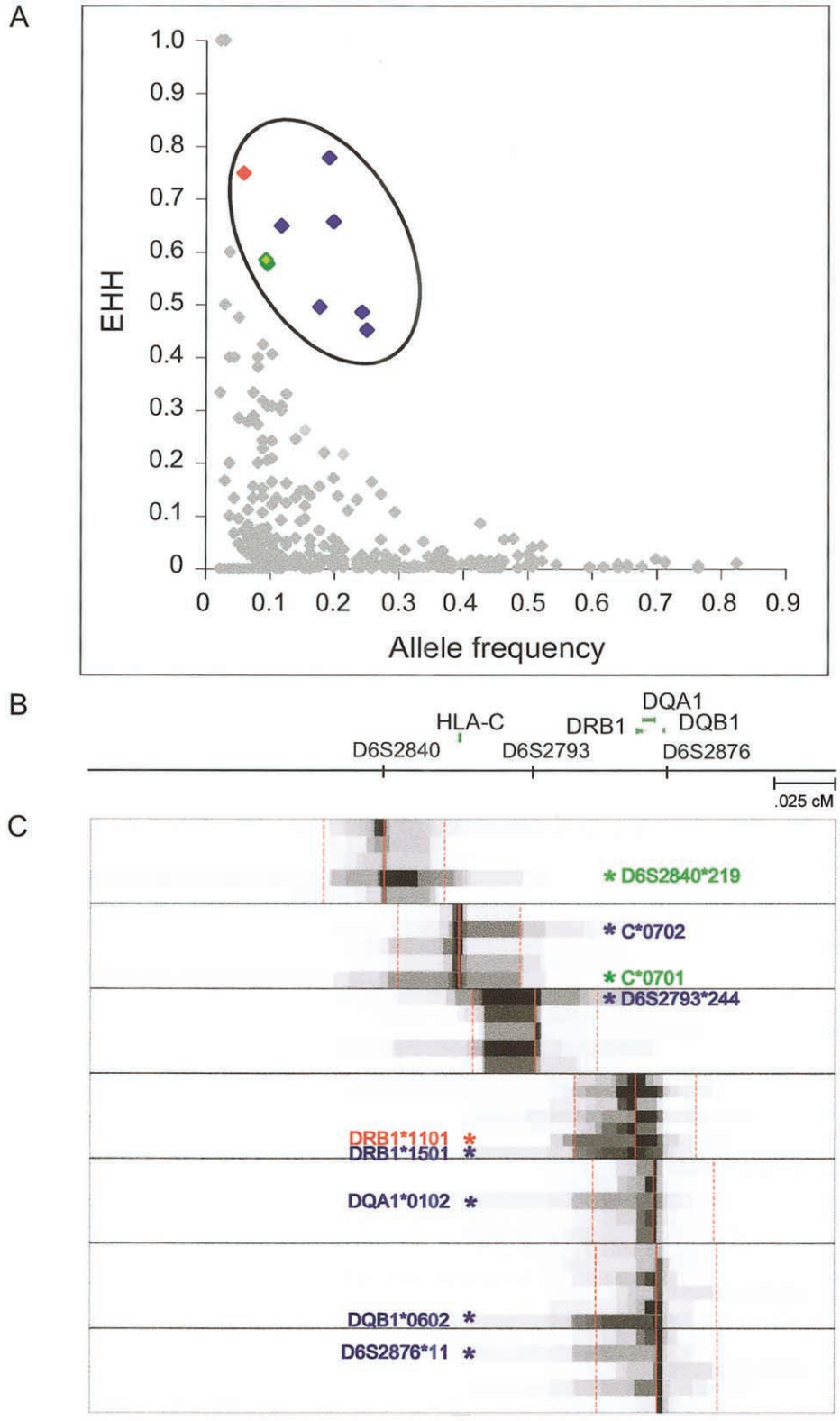
SNP genotypes from Gabriel et al. (2002) were used for comparison of haplotype block structure. As the density of coverage was different between these two studies, 20 data sets were derived from the Gabriel et al. (2002) data by randomly removing markers to achieve the same average spacing and spacing distribution. Since there were two existing 100-kb gaps in our SNP coverage, owing to a lack of available SNPs to type near *FLOT1* and *DQB1,* comparison was done by segmenting the MHC into three parts at these gaps.

Phase Inference for Extended-Haplotype-Homozygosity Analysis

Initial SNP, *HLA, TAP,* and microsatellite chromosomal phasing was done, on the basis of segregation analysis, using the Genehunter program (Kruglyak et al. 1996). The bulk of genotypes—91.6% of SNP genotypes and 95% of *HLA*, *TAP,* and microsatellite genotypes—were phased with family information. Apart from initial phasing with family information, *HLA, TAP,* and microsatellite genotypes were not phased further, and the 5% of genotypes that were indeterminate were considered "ambiguous" in further analyses. Further haplotype inference of SNP genotyping data was performed with a procedure that is based on a probability model for haplotypes proposed elsewhere (Fearnhead and Donnelly 2001). This model can be regarded as a refinement that allows for recombination of the model used in the well-known program, PHASE (Stephens et al. 2001). Both unphased and missing SNP data were inferred in this manner. Since we have a dense set of markers and most markers are in strong LD with several other markers, we do not believe that the phasing has introduced serious bias into our results.

Extended-Haplotype-Homozygosity Analysis

Extended-haplotype-homozygosity (EHH) analysis was performed, as described elsewhere (Sabeti et al. 2002), for each haplotype block, microsatellite, *HLA,* and *TAP* allele, with cM estimations used as distance. Grandparental chromosomes from all families were analyzed. However, some microsatellite types (D6S258, D6S2840, D6S2814, D6S2793, D6S1666, D6S1701, D6S1560, and D6S1542) were not determined for five of these families (1346, 1345, 1420, 1350, and 13292). Rather than infer genotypes, we left these genotypes as "null calls." As mentioned above, 5% of microsatellite, *HLA,* and *TAP* genotypes could not be phased with family information. Since EHH is a cumulative statistic, these heterozygotes and missing data are predicted to result in a conservative estimate of EHH values.


A
EHH
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Allele frequency
B
HLA-C
DRB1
DQA1
DQB1
D6S2840
D6S2793
D6S2876
.025 cM
C
* D6S2840*219
* C*0702
* C*0701
* D6S2793*244
DRB1*1101 *
DRB1*1501 *
DQA1*0102 *
DQB1*0602 *
D6S2876*11 *

Outlying variants were chosen on the basis of two criteria designed to pick alleles with high EHH values for their frequency class. First, as a simple approximation of the distribution, we ranked scores by EHH value times allele frequency. Outliers had values >4.5 SDs above the mean. Second, all variants were sorted by frequency into 5% bins. Outliers had EHH values ≥4.79 SDs above the mean for the remaining values in that bin.

Analysis of SNP Haplotypes around HLA-A, HLA-B, HLA-C, *and* HLA-DRB1

Subsequent to the initial SNP genotyping and analysis of the entire region, additional SNP genotyping was performed near *HLA-A, HLA-C, HLA-B,* and *HLA-DRB1* to assess the correlation between the *HLA* genotype and local SNP haplotype. Multiblock SNP haplotypes include information from the blocks indicated, as described, as well as that from any intervening SNPs not in those blocks. "Leave-one-out" cross-validation was performed using the LeaveOneOut program (E. C. Walsh, IDRG Web site). (In brief, a single chromosome is selected from the data set. The remaining samples are used to build a predictor. This predictor is then used to predict the *HLA* genotype of the sample that has been removed. If the SNP haplotype occurred once, it is not considered in the test.) For each locus, prediction was performed with 10^6 iterations. (See the IDRG Web site for the LeaveOneOut program and genotyping details.)

Results

Structure of LD in the MHC, Compared with the Genome at Large

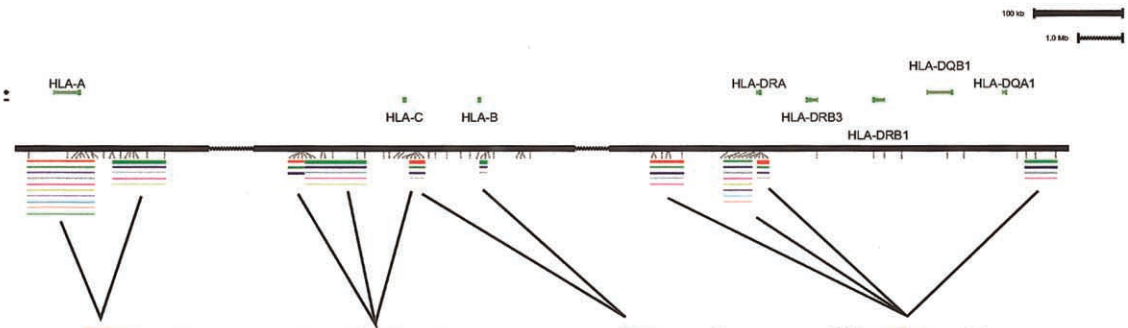
Recent studies have shown that LD extends across long segments of the genome (Daly et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Phillips et al. 2003). Within such segments, a small number of distinct, common patterns of sequence variation (haplotype alleles) are observed in the general population. Between these segments are short intervals where recombination is apparently most active in creating assortments of these patterns (Daly et al. 2001; Jeffreys et al. 2001; Gabriel et al. 2002). Operationally, it is not necessary to test each variant within an LD segment for association with disease phenotype. Rather, a small subset of variants that identifies all common haplotype alleles within a segment can be used.

We wanted to compare the LD structure in the MHC with that of the genome as a whole. To this end, we compared our SNP haplotype data from the MHC (fig. 1) with the data set from Gabriel et al. (2002), as this data set offers a genomewide comparison in which the same CEPH samples were genotyped. We also used the empirical definition of an LD segment or "haplotype block" described in Gabriel et al. (2002), as it provides a common measure for comparison of genomic regions (see the "Materials and Methods" section). Because the SNP coverage in our present study is less dense than that of Gabriel et al. (2002), we randomly selected subsets of markers from the Gabriel et al. (2002) study to create a data set with a spacing similar to that of the present study and thus appropriate for comparison (see the "Materials and Methods" section). Given the SNP coverage used, we do not detect all haplotype blocks. At this density, only 25% of the MHC and 14.5% of the Gabriel et al. (2002) data set is found to lie in blocks, compared with 85% when using the full density in the Gabriel et al. (2002) data set.

Our analysis shows that LD extends over greater physical distances in the MHC than elsewhere in the genome (fig. 2*A*; IDRG supplementary fig. 1). We identify 17 LD segments in the region that meet the criteria of haplotype blocks (Gabriel et al. 2002) (fig. 1). These MHC blocks are longer, on average, in physical distance than those found in the rest of the genome, although this finding does not reach significance, likely because of our small sample size (average length of 31.1 kb vs. 22.3 kb) (fig. 2*B*).

Despite being longer in physical distance, haplotype blocks in the MHC are actually shorter, in terms of genetic

Figure 3 EHH analysis of haplotype blocks, microsatellites, *HLA* genes, and *TAP* genes in the region. EHH is computed as the percentage of instances in which two randomly selected chromosomes with the same variant locus have identical alleles at all SNPs assayed to a particular distance from that locus (e.g., an EHH of 0.5 at marker X means that 50% of possible pairings of a particular variant exhibit sequence identity from the locus to marker X.) *A,* Points representing the EHH at a distance of 0.25 cM from an allele at a particular locus. Outlying variants are indicated in color. The nine outlying variants define three extended haplotypes. Blue points indicate variants that map on the *DRB1*1501* haplotype (associated with SLE and MS). Overlapping green points indicate variants *C*0701* and *D6S2840*219,* which are both found on a haplotype associated with autoimmune diabetes, SLE, and hepatitis (DR3). The red point indicates *DRB1*1101* (associated with pemphigoid disease). *B,* Recombination-distance-based map of the region. Microsatellites and genes for outlying EHH variants are indicated by ticks and above-line graphics, respectively. *C,* EHH values for loci that have at least one outlying variant. Outlying variants were seen at 7 of the 48 independent loci tested. The *X*-axis denotes distance in cM. EHH values are converted to grayscale values: EHH of 1 = black, EHH of 0.5 = 50% grayscale. Solid red lines indicate the locus about which values were derived. The red dotted lines indicate 0.25-cM distance at which outliers were assessed. Two *HLA-C* alleles, *C*0702* and *C*0701,* are extended, as are two *DRB1* alleles, *DRB1*1501* and *DRB1*1101.* The other *HLA* gene alleles with extended haplotypes are *DQA1*0102* and *DQB1*0602.* The microsatellite alleles with extended haplotypes are *D6S2793*244, D6S2876*11,* and *D6S2840*219.* Asterisks highlighting alleles are color coded by haplotype, as in *A.*



A

HLA-A		%
0101	G	92%
0101	O	2.6%
0101	D	2.6%
0101	C	2.6%
0199	M	100%
0201	D	93%
0201	G	2.3%
0201	P	2.3%
0201	I	2.3%
0205	E	100%
0206	D	100%
0299	D	100%
0301	C	93%
0301	J	7%
1101	J	100%
2301	F	66%
2301	L	33%
2402	A	100%
2403	A	100%
2501	L	100%
2601	L	100%
2902	B	100%
3001	N	100%
3101	I	100%
3201	K	100%
6899	H	100%
68011	H	100%
68012	H	100%

HLA-C		%
0102	K	50%
0102	R	50%
0303	A	100%
0401	L	80%
0401	M	20%
0501	C	88%
0501	N	6%
0501	L	6%
0602	D	94%
0602	C	6%
0701	E	14%
0701	F	71%
0701	O	4.7%
0701	G	4.7%
0701	K	4.7%
0702	G	85%
0702	E	3.8%
0702	J	3.8%
0702	P	3.8%
0702	D	3.8%
0704	H	100%
0802	S	100%
1202	M	100%
1203	J	100%
1402	C	100%
1502	K	75%
1502	E	25%
1601	I	100%
1701	Q	100%
2022	M	100%
0304/03041	B	100%

HLA-B		%
0702	A	89%
0702	L	3.6%
0702	M	3.6%
0702	N	3.6%
0801	B	94%
0801	A	6%
1302	A	100%
1401	O	100%
1402	P	100%
1501/1510	Q	6%
1501/1510	E	94%
1801	I	40%
1801	J	40%
1801	R	20%
2705	J	100%
3501	H	100%
3503	H	100%
3701	C	100%
3906	D	100%
4001	E	100%
4002	S	100%
4101	I	100%
4402	F	100%
4403	T	100%
4404	J	100%
4901	L	100%
5001	G	100%
5701	I	100%
5801	H	100%
5101/03	J	100%
5101/03	K	100%
5201/52011	J	100%
5501/02	J	100%

DRB1		%
0101	A	91%
0101	T	9%
0103	M	100%
0301	K	75%
0301	L	17%
0301	N	8%
0401	C	86%
0401	N	7%
0401	U	7%
0402	C	100%
0403	C	100%
0404	B	36%
0404	C	64%
0407	C	100%
0701	C	42%
0701	D	42%
0701	V	8%
0701	W	8%
0799	Q	100%
0801	M	100%
1001	R	100%
1101	O	100%
1103	O	100%
1104	S	100%
1201	P	100%
1301	E	33%
1301	F	33%
1301	G	33%
1302	H	50%
1302	M	25%
1302	X	25%
1303	J	100%
1401	L	100%
1501	N	93%
1501	V	3.5%
1501	Y	3.5%
1601	O	100%

B

HLA-A			%
A	2402	7	87.5%
A	2403	1	12.5%
B	2902	3	100%
C	0301	14	93.3%
C	0101	1	6.7%
D	0201	41	89%
D	0299	3	6.5%
D	0206	1	2.2%
D	0101	1	2.2%
E	0205	2	100%
F	2301	2	100%
G	0101	22	95.6%
G	0201	1	4.4%
H	68012	3	60%
H	6899	1	20%
H	68011	1	20%
I	3101	3	75%
I	0201	1	25%
J	1101	5	83.3%
J	0301	1	6.7%
K	3201	3	100%
L	2501	4	44.4%
L	2601	4	44.4%
L	2301	1	11.1%
M	0199	1	100%
N	3001	1	100%
O	0101	1	100%
P	0201	1	100%

HLA-C			%
A	0303	6	100%
B	0304/03041	6	83.3%
B	0399	1	16.7%
C	0501	14	87.4%
C	0602	1	6.3%
C	1402	1	6.3%
D	0602	15	93.8%
D	0702	1	6.2%
E	0701	3	60%
E	0702	1	20%
E	1502	1	20%
F	0701	15	100%
G	0702	23	95.8%
G	0701	1	4.2%
H	0704	3	100%
I	1601	2	100%
J	1203	2	66.7%
J	0702	1	33.3%
K	1502	3	60%
K	0701	1	20%
K	0102	1	20%
L	0401	8	88.9%
L	0501	1	11.1%
M	0401	2	33.3%
M	1202	2	33.3%
M	02022	2	33.3%
N	0501	1	100%
O	0701	1	100%
P	0702	1	100%
Q	1701	1	100%
R	0102	1	100%
S	0802	1	100%

HLA-B			%
A	0702	23	92%
A	0801	1	4%
A	1302	1	4%
B	0801	15	100%
C	3701	2	100%
D	3906	2	100%
E	1501	7	54.8%
E	4001	5	38.5%
E	1510	1	7.7%
F	4402	19	100%
G	5001	3	100%
H	3501	7	70%
H	3503	2	20%
H	5801	1	10%
I	5701	9	75%
I	1801	2	16.6%
I	4101	1	8.3%
J	1801	2	16.6%
J	5201/52011	2	16.6%
J	5101/03	5	41.7%
J	2705	1	8.3%
J	4404	1	8.3%
J	5501/02	1	8.3%
K	5101/03	1	100%
L	0702	1	50%
L	4901	1	50%
M	0702	1	100%
N	0702	1	100%
O	1401	1	100%
P	1402	1	100%
Q	1501	1	100%
R	1801	1	100%
S	4002	1	100%
T	4403	1	100%

DRB1			%
A	0101	10	100%
B	0404	4	100%
C	0404	7	23.3%
C	0401	12	40%
C	0403	3	10%
C	0407	2	6.6%
C	0701	5	16.6%
C	0402	1	3.3%
D	0701	5	100%
E	1301	2	100%
F	1301	2	100%
G	1301	2	100%
H	1302	2	100%
J	1303	2	100%
K	0301	9	100%
L	0301	2	28.6%
L	1401	5	71.4%
M	0801	4	66.6%
M	0103	1	16.6%
M	1302	1	16.6%
N	1501	25	92.6%
N	0401	1	3.7%
N	0301	1	3.7%
O	1103	2	15.4%
O	1101	8	61.5%
O	1601	3	23%
P	1201	1	100%
Q	0799	1	100%
R	1001	1	100%
S	1104	1	100%
T	0101	1	100%
U	0401	1	100%
V	0701	1	50%
V	1501	1	50%
W	0701	1	100%
X	1302	1	100%

distance. The average recombination rate in the MHC is 0.49 cM/Mb, versus 0.81 cM/Mb in the genome as a whole (Cullen et al. 2002; Kong et al. 2002). Given this difference in recombination rate, we found that blocks in the MHC have an average length of 0.012 cM, whereas the average is 0.017 cM for the genomewide control data set (significance not tested) (fig. 2*C*). Furthermore, the distribution of recombination across the region correlates well with most of the long blocks (fig. 1, *asterisks*) in the region. Six of the seven largest blocks (≥75 kb) lie in areas where recombination rate is well below the genome average of 0.81 cM/Mb. Moreover, five of these blocks lie in regions where the recombination rate is below the MHC regional average of 0.49 cM/Mb. The remaining large block falls into a region where the rate is 0.83 cM/Mb. We conclude that extent of LD in the MHC is longer in physical distance but not in genetic distance than elsewhere in the genome.

Extended-Haplotype Analysis

We next looked for alleles of haplotype blocks, microsatellites, or classical *HLA* genes that occur on haplotypes that extend across multiple blocks. Such so-called "extended haplotypes" are believed to represent a common feature of the MHC (Alper et al. 1992). To analyze the long-range structure of the region, we used EHH analysis, which determines the length of the chromosomal haplotypes that extend from a specific allele at a particular locus (Sabeti et al. 2002). High-frequency, extended haplotypes may result from positive selection or haplotype-specific recombination suppression. Positive selection brings rare alleles to higher frequency in relatively few generations, thus affording fewer opportunities for recombination events to separate an allele from its original chromosomal context. Alternatively, haplotype-specific recombinational suppression may result in high-frequency, extended haplotypes by reducing the number of recombination events a given haplotype will undergo. Since we have a detailed sperm-typing recombination map of the region, we used this to control for positional variation in average recombination rates that would artificially affect the length of haplotypes. Utilizing the integrated haplotype map, we scanned across the entire MHC, using each *HLA* gene, *TAP* gene, microsatellite, and haplotype block as an independent locus from which to determine EHH values, assessing every allele from a total of 46 loci.

The 50 regions in the Gabriel et al. (2002) data set each span only 250 kb and are, therefore, not long enough to serve as a suitable control data set for this analysis. Thus, we compared the EHH values of haplotype, microsatellite, and gene alleles within the MHC data set with each other and identified allelic variants that are outliers, on the basis of statistical rank of the EHH value at 0.25 cM, relative to allele frequency (see the "Materials and Methods" section) (fig. 3*A*). We identified nine alleles that map onto three different extended haplotypes (fig. 3*B*). It is striking that six of these nine variants map to a single multigene haplotype (*HLA-C*0702–D6S2793*244–DRB1*1501–DQA1*0102–DQB1*0602–D6S2876*11* [hereafter referred to as "DR2"]). Every element in the DR2 haplotype has an EHH value at least 4.8 SDs above the mean EHH for other variants with the same allele frequency. Two of the remaining outlying alleles map to a single haplotype (*D6S2840*219–C*0701*), and the last outlying allele is *DRB1*1101*.

As noted above, there are at least two possible underlying causes for these extended haplotypes. One possibility is that a variant on the haplotype has experienced recent positive selection. It is interesting that each of the three extended haplotypes has been implicated elsewhere in autoimmune disease (Thorsby 1997; Klein and Sato 2000). The DR2 haplotype is associated with systemic lupus erythematosus (SLE [MIM 152700]) and multiple sclerosis (MS [MIM 126200]) susceptibility, and it is protective for type I diabetes (IDDM [MIM 222100]) (Thorsby 1997; Chataway et al. 1998; Haines et al. 1998; Barcellos et al. 2002). *DRB1*1101* is associated with pemphigoid vulgaris, and *D6S2840*219–C*0701* is associated with autoimmune diabetes (MIM 275000) and thyroid disease (MIM 140300) (Drouet et al. 1998; Price et al. 1999; Okazaki et al. 2000). Thus, these three haplotypes appear to have functional consequences for the human immune system. Although these haplotypes are associated with autoimmune diseases at present, it is possible that, under certain conditions, these functional differences were (and perhaps still are) beneficial for disease resistance and, therefore, may have undergone positive selection in the past.

The other possibility is that these extended haplotypes are subject to allele-specific recombination suppression. By examining the individual recombination rates used to construct the recombination map, we observe that, of the 12 individuals examined, the single individual bear-

Figure 4 Correlation of *HLA* alleles to SNP haplotype background. Map of region showing placement of SNPs and haplotypes assayed is shown for reference. Multi-SNP haplotypes are coded by single capital letters. *A,* SNP-*HLA* haplotypes sorted by *HLA* allele. Percents indicate the percentage of a particular *HLA* allele that falls on the indicated SNP haplotype. *B,* SNP-*HLA* haplotypes sorted by SNP haplotype allele. Percents indicate the percentage of a particular SNP haplotype allele that bears the indicated *HLA* allele. Counts are overall number of chromosomes bearing the SNP-*HLA* haplotype indicated.

ing *DRB1*1501* showed many fewer recombination events across the MHC than did the others, although this difference did not significantly deviate from the mean. This suggests that allele-specific recombination suppression could be a possibility in this case (M. Cullen, unpublished material). Further sperm typing of additional individuals bearing each of these extended haplotypes should resolve whether the underlying cause of this extended haplotype is haplotype-specific recombinational suppression or whether recent positive selection is more likely.

Common Patterns of Sequence Variation in the MHC in Regions between the Classical HLA *Loci*

We next compared the haplotype block variation in the MHC with the rest of the genome. With our initial coverage, we did not identify blocks that spanned classical loci. We observe that these blocks have the same number of common patterns of sequence variation (haplotype alleles) as found in other regions of the genome (3.9 vs. 4.1 for blocks with five or more markers) (fig. 1*D*). Furthermore, we see the same percentage of rare haplotype alleles in both data sets (3%), which indicates that the MHC, aside from the classical loci, does not appear to have an excess of rare haplotype variants detectable at our current marker density. Our observation that the diversity of haplotypes outside the classical loci is typical of the rest of the genome is perhaps surprising, given the high level of variation at the classical *HLA* genes.

Common Variation in Regions Spanning the Classical HLA *Loci*

We separately analyzed the SNP haplotype diversity in regions spanning the classical *HLA* genes (but outside the highly variable exons) to understand how this variation is structured. For this purpose, it was necessary to increase our density of SNP coverage by three- to fivefold around the four *HLA* genes chosen for analysis, *HLA-A, HLA-C, HLA-B,* and *HLA-DRB1*. One motivating question in this analysis was whether SNP haplotypes spanning classical *HLA* loci contained enough information to predict *HLA* alleles. If so, it might be possible to use high-throughput SNP genotyping as a first-pass surrogate for traditional *HLA* gene molecular typing (e.g., probe-based typing or direct sequencing) in disease-association studies. For one of these classical genes, *HLA-A,* we found a single 7-SNP haplotype block spanning the locus. This 7-SNP *HLA-A* block has only six common variants, and those are predictive of the correct *HLA-A* allele 66.2% of the time, as shown by cross-validation analysis (LeaveOneOut [see the "Materials and Methods" section]). To capture more of the variation at this locus, we included the genotype information for a neighboring block and examined the SNP haplotypes that comprised the combinations of alleles of these two blocks. We greatly improved our success of prediction from 66.2% to 82.6% of all *HLA-A* alleles present.

Using such multiblock haplotypes for all four classical *HLA* loci studied, we find that multiblock SNP haplotypes can, in many cases, act as surrogate markers for *HLA* alleles. For example, the *HLA-A*0101* allele occurs on the "G" SNP haplotype (comprising the haplotype alleles of two blocks) 92% of the time (fig. 4*A*), and the "G" SNP haplotype correlates to *HLA-A*0101* 95.6% of the time (fig. 4*B*). We used the cross-validation analysis to estimate our success rate. Even with our current coverage, we can accurately predict *HLA* alleles by SNP haplotype 75%–84% of the time (*HLA-A*: 82.6%; *HLA-B*: 79.8%; *HLA-C*: 84.3%; and HLA-*DRB1*: 75.0%). If we consider only those haplotypes bearing common *HLA* alleles (allele frequency >5%), we are accurate at a much higher rate (*HLA-A*: 96.2%, *HLA-B*: 98.8%; *HLA-C*: 96.0%; and *HLA-DRB1*: 82.2%), which suggests that the bulk of our prediction failures reflect an inability to predict low-frequency variants. These data suggest that two elements are needed to improve our predictive power: (1) a larger data set, which would increase the number of observations of rare *HLA* variants, and (2) increased marker density that would provide additional SNP haplotype information, as evidenced by the case of *HLA-A* above.

Discussion

Here, we present a first-pass integrated map of the SNP, microsatellite, and *HLA* variation in the MHC. Using this map, we show that, aside from the classical *HLA* loci, the variation and LD structure of the MHC are not different from a genomewide control data set. Specifically, whereas LD appears to extend over longer physical distances in the MHC, this seems to be accounted for by the reduced recombination rate in the region. Furthermore, we show that, in the regions that do not span classical *HLA* loci, the number of common haplotype alleles in the MHC are not different from the rest of the genome.

It is important to note that we also demonstrate that multiblock SNP haplotypes contain considerable predictive information for common *HLA* alleles at *HLA-A, HLA-B, HLA-C,* and *HLA-DRB1*. Although direct molecular typing of classical *HLA* loci will likely remain the method of choice for clinical practice, multiblock SNP haplotypes should enable cost-efficient, large-scale exploration of the variation at the classical *HLA* loci and beyond. An additional implication of these results is that multiblock SNP haplotypes may be sufficient to identify low-frequency variants throughout the genome. Such low-frequency variants would likely be missed in single, block-based, common variant analysis; however,

their contribution to disease could be assayed by use of multiblock haplotypes in analysis.

What is the future utility of this integrated variation map of the MHC? In the 50 years of study since its first discovery, the MHC has been implicated in almost every human inflammatory and autoimmune disease. Historically, the MHC has been studied by typing of the classical *HLA* genes and microsatellites. However, only rarely has this analysis definitively identified causal variation. Often, association studies using these methods implicate more than one allele at a single locus as influencing disease susceptibility. Although this may represent allelic heterogeneity underlying disease, reinterpreting such results with attention to shared SNP-haplotype variation might point to additional hypotheses regarding the causal variant. For instance, one may find that two different disease-predisposing *HLA* alleles share a common SNP haplotype, which suggests that a variant carried on that haplotype may, in fact, be the underlying cause of disease.

Another common finding in MHC studies is that an extended haplotype, rather than a single variant, is associated with disease. A uniform map of the variation in the region would allow fine mapping of association signals on the basis of rare recombinant chromosomes. Because SNPs are more abundantly present, reliably typed, and cost efficient than microsatellites, they are an excellent choice for this sort of large-scale, high-density genotyping. A denser sampling of all the haplotype variation in the region will allow researchers to fully consider all of the 120 genes that lie in the MHC, rather than to focus solely on the classical *HLA* loci.

What are the next steps in the development of this map? The current map identifies haplotype blocks covering only 24.5% of the MHC, because the marker density is not yet sufficient to have fully captured all the common variation in the region. On the basis of the estimated average size of blocks in this region, SNP coverage must be increased fourfold to reach saturation. Attaining such density will certainly require additional SNP discovery efforts, such as those of the MHC Haplotype Sequencing Project (Allcock et al. 2002). Furthermore, this map is based on the genotypes only of individuals of European ancestry. Variation in other populations must also be examined to unify MHC association results between populations.

Ultimately, a full understanding of the patterns of LD and haplotype diversity of this region should allow the identification of a subset of SNPs required for future disease studies. This will allow MHC-association studies to be completed cost effectively by using a combination of haplotype-tagging and *HLA* allele-tagging SNPs. Although we used a large number of SNPs to construct this current map, and more SNPs will be needed to fully describe the haplotype structure of the region, we estimate that 10–15 SNPs per locus will be sufficient for common, classical HLA alleles. Moreover, in cases where there is already significant association to a particular locus, these informative SNPs may be used to map outward from the original signal and delimit the region of association. We estimate that a few dozen SNPs might be needed in such endeavors. In conclusion, SNP-based haplotype approaches will allow the examination of larger disease cohorts and enable the identification of rare recombinant haplotypes that would refine association signals and potentially identify the causal alleles for MHC-associated diseases.

Acknowledgments

We thank the members of the Inflammatory Disease Research Group, of the Whitehead Center for Genome Research, and Didier Stainier, for critical reading of this work and this manuscript. Many thanks to Andrew Kirby and Leslie Gaffney, for invaluable help with figures, and to Sheila Guschwan McMahon, for help with the Web site. E.W. is supported by a postdoctoral training fellowship from the Cancer Research Institute. K.M. is supported by a Howard Hughes Medical Institute predoctoral fellowship and National Science Foundation–Doctoral Dissertation Improvement grant. K.M. and G.T. are supported by National Institutes of Health (NIH) grant GM35326. This project has been funded, in part, with federal funds from the National Cancer Institute, NIH, under contract N01-CO-12400 (article H.36 of the prime contract). The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Electronic-Database Information

URLs for data presented herein are as follows:

Coriell Institute, http://locus.umdnj.edu/ccr/
dbSNP, http://www.ncbi.nlm.nih.gov/SNP/
IDRG, http://www-genome.wi.mit.edu/mpg/idrg/projects/hla.html (for supplementary "Materials and Methods" information and pairwise D' analysis for 201 reliable, polymorphic SNP assays in 18 multigenerational European CEPH families)
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for SLE, MS, IDDM, thyroid disease, and autoimmune diabetes)

References

Allcock RJ, Atrazhev AM, Beck S, de Jong PJ, Elliott JF, Forbes S, Halls K, Horton R, Osoegawa K, Rogers J, Sawcer S, Todd JA, Trowsdale J, Wang Y, Williams S (2002) The MHC haplotype project: a resource for HLA-linked association studies. Tissue Antigens 59:520–521

Alper CA, Awdeh Z, Yunis EJ (1992) Conserved, extended MHC haplotypes. Exp Clin Immunogenet 9:58–71

Barcellos LF, Oksenberg JR, Green AJ, Bucher P, Rimmler JB, Schmidt S, Garcia ME Lincoln RR, Pericak-Vance MA,

Haines JL, Hauser SL, Multiple Sclerosis Genetic Group (2002) Genetic basis for clinical expression in multiple sclerosis. Brain 125:150–158

Beck S, Trowsdale J (2000) The human major histocompatability complex: lessons from the DNA sequence. Annu Rev Genomics Hum Genet 1:117–137

Begovich AB, McClure GR, Suraj VC, Helmuth RC, Fildes N, Bugawan TL, Erlich HA, Klitz W (1992) Polymorphism, recombination, and linkage disequilibrium within the HLA class II region. J Immunol 148:249–258

Bugawan TL, Klitz W, Blair A, Erlich HA (2000) High-resolution HLA class I typing in the CEPH families: analysis of linkage disequilibrium among HLA loci. Tissue Antigens 56: 392–404

Carrington M, Colonna M, Spies T, Stephens JC, Mann DL (1993) Haplotypic variation of the transporter associated with antigen processing (TAP) genes and their extension of HLA class II region haplotypes. Immunogenetics 37:266–273

Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ (1999) HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science 283:1748–1752

Carrington M, Stephens JC, Klitz W, Begovich AB, Erlich HA, Mann D (1994) Major histocompatibility complex class II haplotypes and linkage disequilibrium values observed in the CEPH families. Hum Immunol 41:234–240

Chataway J, Feakes R, Coraddu F, Gray J, Deans J, Fraser M, Robertson N, Broadley S, Jones H, Clayton D, Goodfellow P, Sawcer S, Compston A (1998) The genetics of multiple sclerosis: principles, background and updated results of the United Kingdom systematic genome screen. Brain 121:1869–1887

Cullen M, Malasky M, Harding A, Carrington M (2003) High-density map of short tandem repeats across the human major histocompatibility complex. Immunogenetics 54:900–910

Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. Am J Hum Genet 71:759–776

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544–548

Drouet M, Delpuget-Bertin N, Vaillant L, Chauchaix S, Boulanger MD, Bonnetblanc JM, Bernard P (1998) HLA-DRB1 and HLA-DQB1 genes in susceptibility and resistance to cicatricial pemphigoid in French Caucasians. Eur J Dermatol 8: 330–333

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299–1318

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Haines JL, Terwedow HA, Burgess K, Pericak-Vance MA, Rimmler JB, Martin ER, Oksenberg JR, Lincoln R, Zhang DY, Banatao DR, Gatto N, Goodkin DE, SL H (1998) Linkage of the MHC to familial multiple sclerosis suggests genetic heterogeneity. The Multiple Sclerosis Genetics Group. Human Molecular Genetics 7:1229–1234

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222

Klein J, Sato A (2000) The HLA system: second of two parts. N Engl J Med 343:782–786

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363

Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49:49–67

Martin MP, Harding A, Chadwick R, Kronick M, Cullen M, Lin L, Mignot E, Carrington M (1998) Characterization of 12 microsatellite loci of the human MHC in a panel of reference cell lines. Immunogenetics 47:131–138

Moonsamy PV, Klitz W, Tilanus MG, Begovich AB (1997) Genetic variability and linkage disequilibrium within the DP region in the CEPH families. Hum Immunol 58:112–121

Okazaki A, Miyagawa S, Yamashina Y, Kitamura W, Shirai T (2000) Polymorphisms of HLA-DR and -DQ genes in Japanese patients with bullous pemphigoid. J Dermatol 27:149–156

Phillips MS, Lawrence R, Sachidanandam R, Morris AP, Balding DJ, Donaldson MA, Studebaker JF, et al (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. Nat Genet 33:382–387

Price P, Witt C, Allcock R, Sayer D, Garlepp M, Kok CC, French M, Mallal S, Christiansen F (1999) The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol Rev 167:257–274

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Thorsby E (1997) Invited anniversary review: HLA associated diseases. Hum Immunol 53:1–11