Immunogenetics (2006) 58: 241–251
DOI 10.1007/s00251-005-0076-z

ORIGINAL PAPER

**Chul-Woo Pyo · Luke M. Williams · Yuki Moore · Hironobu Hyodo · Shuying Sue Li · Lue Ping Zhao · Noriko Sageshima · Akiko Ishitani · Daniel E. Geraghty**

# HLA-E, HLA-F, and HLA-G polymorphism: genomic sequence defines haplotype structure and variation spanning the nonclassical class I genes

Received: 30 August 2005 / Accepted: 1 December 2005 / Published online: 29 March 2006


**Abstract** Despite several studies that defined the polymorphism of the nonclassical human leukocyte antigen-E (HLA-E), HLA-F, and HLA-G genes, most polymorphisms thus far examined in correlative studies were derived from the coding sequences of these genes. In addition, some discrepancies and ambiguities in the available data have persisted in current databases. To expand the data available and to resolve some of the discrepant data, we have defined protocols that allow for the amplification of 6 to 7 kb of contiguous genomic sequence for each gene, including all of the coding and intron sequences, approximately 2 kb of 5' flanking promoter sequence, and 1 kb of 3' flanking sequence. Using long-range polymerase chain reaction (PCR) protocols, generating either one or two PCR products depending on the locus, amplified genomic DNA was directly sequenced to completion using a set of about 30 primers over each locus to yield contiguous sequence data from both strands. Using this approach, we sequenced 33 genomic DNAs, from Asian, African American, and Caucasian samples. The results of this analysis confirmed several previously reported coding sequence variants, identified several new allelic variants, and also defined extensive variation in intron and flanking sequences. It was possible to construct haplotype maps and to identify tagging single nucleotide polymorphisms that can be used to detect the composite variation spanning all three genes.

**Keywords** MHC · Nonclassical class I · Genomics · SNPs · Alleles

C.-W. Pyo · L. M. Williams · Y. Moore · H. Hyodo · D. E. Geraghty (✉)
The Clinical Research Division,
Fred Hutchinson Cancer Research Center,
1100 Fairview Ave.,
N. Seattle, WA 98109-1024, USA
e-mail: geraghty@fhcrc.org
Tel.: +81-744-298843
Fax: +81-744-291116

S. S. Li · L. P. Zhao
The Public Health Sciences Division,
Fred Hutchinson Cancer Research Center,
1100 Fairview Ave.,
N. Seattle, WA 98109-1024, USA

N. Sageshima · A. Ishitani
Nara Medical University,
Kashihara, Nara 634, Japan

## Introduction

The major histocompatibility complex (MHC) in humans contains three nonclassical class I genes, human leukocyte antigen (HLA-E), HLA-F, and HLA-G, all of which are located within the class I region, and together with the classical class I antigens constitute the complete list of active class I genes in the human (Geraghty 1993; Geraghty et al. 1987, 1990; Koller et al. 1988). Each of the nonclassical class I genes can be distinguished from classical class I by their expression patterns, and for HLA-E and HLA-G, with respect to peptide binding properties and function. HLA-E is ubiquitously expressed (Koller et al. 1988; Lee et al. 1998a), and HLA-G is expressed specifically in placental tissue (Ishitani and Geraghty 1992; Kovats et al. 1990; Le Bouteiller et al. 1996). HLA-G binds a relatively narrow range of peptides probably serving as structural components rather than for antigen presentation (Ishitani et al. 2003; Lee et al. 1995), while HLA-E complexes with nonamer peptide derived from the signal sequence of other MHC class I genes, including HLA-A, HLA-B, HLA-C, and HLA-G, but excluding HLA-F (Lee et al. 1998a). The function of HLA-G is not yet clear but it may act as an inhibitory ligand by interacting with the Ig-like transcript 2 (ILT2) and ILT4 receptors (Shiroishi et al. 2003). HLA-E is a ligand for the lectin receptor CD94 combined with different NKG2 subunits acting to inhibit and activate primarily natural killer cells (Kaiser et al. 2005; Lee et al. 1998b).

In addition to functional differences from the classical class I antigens, and indeed, reflecting those differences, the HLA-E, HLA-F, and HLA-G loci have relatively little

allelic polymorphism in their coding sequences (Marsh et al. 2005). Despite additional allelic variants having been described over time, the original description of only two nonsynonymous HLA-E variants (Geraghty et al. 1992b) was apparently correct, as concluded in recent studies (Grimsley et al. 2002). Variation at HLA-G has also been extensively analyzed, and limited polymorphism has been observed in a number of different populations (Ishitani et al. 1999). HLA-F also shows little variation, and is highly conserved in distantly related nonhuman primates (Daza-Vamenta et al. 2004). Derivative in part of their proposed involvement in the immunology of pregnancy, the genetics of HLA-G, and to a lesser extent HLA-E, have been examined for association with a number of interesting clinical outcomes. Of several outcomes clinically relevant to pregnancy, recurrent spontaneous abortion (RSA) and preeclampsia have been studied with mixed results. In the former, heterozygote advantage is proposed in one study (Tripathi et al. 2004), while coding region variants were significantly associated in individuals with five or more RSAs (Pfeiffer et al. 2001). In the latter, HLA-G variation in the 3' untranslated region was predictive of preeclampsia (Hylenius et al. 2004). These polymorphisms may be associated with differential function or may reflect linkage disequilibria with other HLA variants, either or both resulting in consequences to the pregnancy (Agrawal and Pandey 2003). Among several immunological phenotypes associated with the MHC, variation within HLA-G has been implicated for involvement in asthma and as a bronchial hyperresponsiveness susceptibility gene (Nicolae et al. 2005), and a deletion in the 3' untranslated region of HLA-G has been associated with Pemphigus (Gazit et al. 2004). Coding variation in HLA-E is limited to two alleles; however, these variants have been shown to yield differential surface protein levels depending also on the combined expression with certain HLA alleles (Strong et al. 2003), and a single study has shown association of these variants with nasopharyngeal carcinoma (Hirankarn et al. 2004).

Most of the existing polymorphism data are limited, and while very few specific coding variants may have functional consequences affecting protein expression, promoter sequence variants and other sequence changes that might quantitatively or qualitatively affect gene expression have been little examined. In addition, resultant from several studies that defined the polymorphism of the nonclassical HLA-E, HLA-F, and HLA-G genes, some discrepancies and ambiguities in the available data still persist. To expand the data available and to resolve some of the discrepant data, we have defined protocols that allowed for the amplification of 6 to 8 kb of contiguous genomic sequence for each gene, including all of the coding and intron sequences and extensive flanking sequences. By examining several DNAs from individuals with both identical by descent (IBD) MHCs and from families, we defined completely phased variation spanning each of these loci for 33 chromosomes. We further identified new allelic coding variants, were unable to confirm a subset of previously identified variants, and established a resource of data and cloned allele types.

## Materials and methods

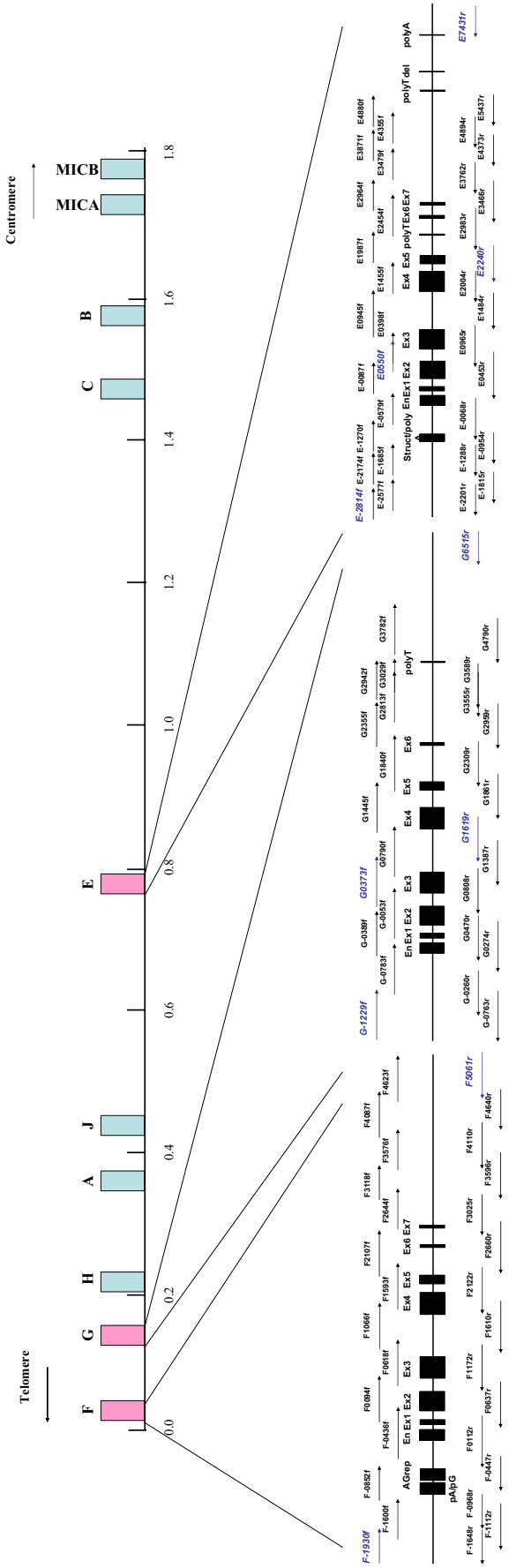
### Cell lines used

A subset of the Epstein–Barr virus transformed B cell lines available through the International Histocompatibility Working Group (IHWG) (http://ihwg.org/) were used in this study for primary single nucleotide polymorphisms (SNP) discovery (Table 1). Nine of these were selected from a set originally part of a tenth International HLA Workshop panel identified as individually containing MHCs that were IBD, and were chosen as a group that represented diversity in their HLA types. An additional 19 cell lines were chosen from members of five selected pedigrees, which, by combining family relationships and HLA typing data, allowed for the resolution of individual haplotypes for the four parental chromosomes, and in one case the additional grandparental chromosome (Table 1). Finally, five cell lines with MHCs IBD of Japanese origin were included. All of the sequence data and derivative SNP

**Table 1** DNA resources used in this study

| MHC Homozygous cell lines | | | | Family cell lines | | |
|---|---|---|---|---|---|---|
| Japanese | Caucasian | Afr-Am 022 | Afr-Am 024 | CEPH 1416 | CEPH 1408 | Seattle-Caucasian |
| 1471 | ARBO | JHUAA0393 (Pt) | JHAA0527 (Pt) | IHW1182 (Pt) | IHW1141 (Pt) | HAN1901 (Pt) |
| 1199 | MT14 | JHUAA394 (Pt) | JHUA0528 (Pt) | IHW1175 (Pt) | IHW1143 (Pt) | HAN1902 (Pt) |
| 595 | LBF | JHUAA396 (Ch) | JHUAA0529 (Ch) | IHW1181 (Ch) | IHW1152 (Ch) | HAN1904 (Ch) |
| 1350 | HOM2 | | JHUAA0530 (Ch) | IHW1184 (Ch) | | HAN1905 (Ch) |
| 435 | BM15 | | | IHW1173 (G.Pt) | | |
| | STE | | | | | |
| | WDV | | | | | |
| | SCHU | | | | | |
| | WT24 | | | | | |

*Afr-Am* African-American, *Pt* parent, *Ch* child, *G.Pt* grandparent




◀ **Fig. 1** Sequence analysis of the HLA-E, HLA-F, and HLA-G loci from genomic DNAs to identify new polymorphisms. Above are indicated the positions of the HLA loci within the class I portion of the MHC. Each of HLA-E, HLA-F, and HLA-G genes are expanded beneath with the exon–intron structures indicated by *black boxes* with an annotation above each. *Arrows* above and below each gene indicate the positions of primers used in long range amplification of each locus (*blue arrows*) and those used for sequencing of the amplified product (*black arrows*). The DNA sequences of all primers are provided as an appendix to this paper. Conditions for amplification and sequencing are provided in "Materials and methods"

typings were resolved into phases either as homozygous sequences from the IBD lines or from segregation analysis in the informative families.

## Long-range PCR amplification

Five sets of primers were developed that can be used to completely amplify coding and surrounding regions of HLA-E, HLA-F, and HLA-G. The amplifications of HLA-E and HLA-G each require two overlapping PCR reactions for completion, whereas HLA-F requires one reaction (Fig. 1). Long-range PCR was done using the TaKaRa LA PCR kit ver 2.1 (TaKaRa Bio, Otsu, Shiga, Japan) using the following protocol in the listed order, mixing thoroughly before the addition of Taq Polymerase: combine 10 mM deoxyribonucleotide triphosphates, 2× GC buffer I, 300 nM each of forward and reverse primers, and 2.5 U Taq Polymerase (TaKaRa LA Taq). PCR was carried out in a total volume of 50 µl with 120 ng of genomic DNA. Using an ABI 9600, the long-range PCR thermal cycler program was: 94°C for 4 min, 94°C for 10 s, 65°C for 30 s, 68°C for 18 min, the previous three steps were repeated 10 times, 94°C for 10 s, 65°C for 30 s, 68°C for 18 min + 20 s/cycle, the previous three steps were repeated 20 times, and 68°C for 7 min.

## Sequence analysis for SNP discovery and genotyping

Primers used for internal sequencing of HLA 5' product E, 3' product E, F, 5' product G, and 3' product G can be found in the attached appendix. Primer nomenclature (Fig. 1) was derived from the position of the 5' base of each primer within the chromosome 6 consensus sequence (NT_007592) with the numbering increasing from +1 for the A in the antithymocyte globulin start codon for each locus, and decreasing from −1 for the base immediately preceding the start codon counting in the reverse direction. Note that primer names and positions in future releases of this protocol will be relative to a consensus sequence that will be specific for that protocol.

Resequencing was done with high throughput methods according to the manufacturer's instructions and to well-established basic procedures (Geraghty et al. 2002). Briefly, PCR reactions were assembled in a PCR clean room using a Beckman Fx robot according to strategies facilitated and automated by software built in the lab, referred to as a genetics management system, and modified

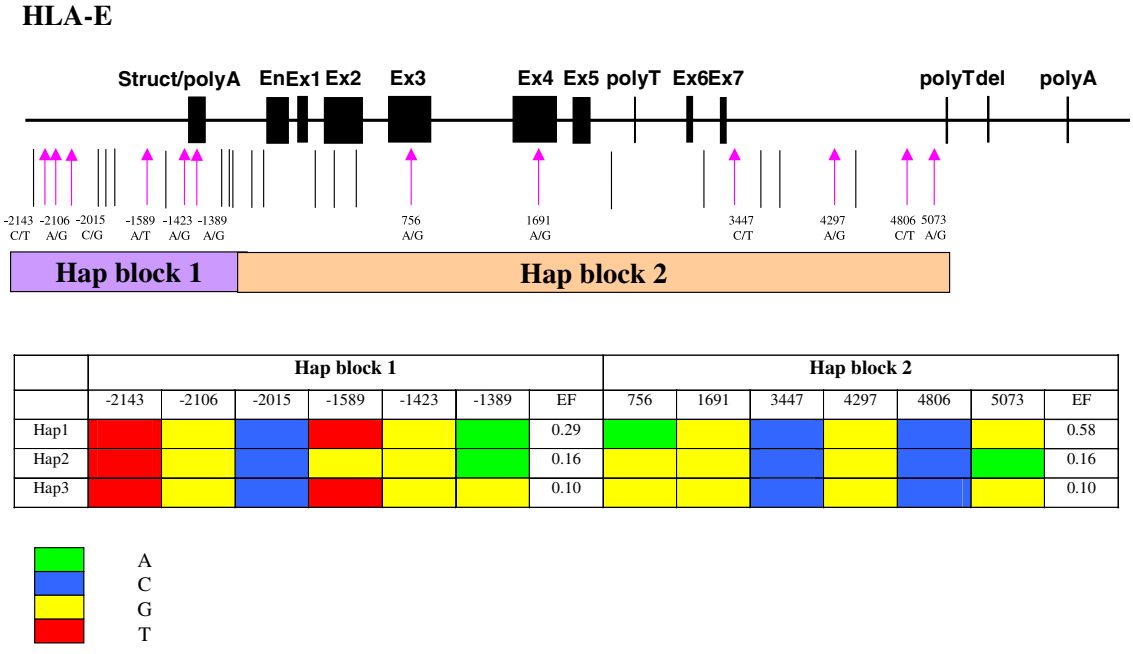
substantially over previously developed concepts and software (Geraghty et al. 2000).

## SNP base-calling from sequence trace data

Both homozygous and heterozygous traces from each amplicon were first combined and analyzed for sequence content using Phred and Phrap (Ewing and Green 1998), and data were viewed using Consed (Gordon et al. 1998). A significant step towards making a sequencing project of this size manageable in a small lab was the in-lab development of a program for heterozygous sequence analysis (called heterozygous trace resolution or HTR), which was able to perform most of the analysis of sequence-based SNP discovery and typing for heterozygous DNAs. This software is similar in intent to Polyphred and Mutation Surveyor by Soft Genetics, LLC, and is specifically adapted for high-throughput analysis [Database of Essential Genes (DEG) unpublished data]. A subset of the data was confirmed manually, which demonstrated 99.8% accuracy of the software interpretations. Detailed information about these SNPs and indels is included as an appendix.

## Post sequencing analysis

Although contiguous sequence was obtained over most of the regions, due to the fact that sequencing reactions were performed on PCR amplified products, minor areas within each sequence displayed motifs that inhibited high-quality sequence interpretations as indicated in Fig. 1 and detailed as follows: HLA-E positions −802 to −687, gap bounded by 5' secondary structure and 3' polynucleotide A tract; 2,424 to 2,445, polynucleotide T tract; 5,487 to 5,506, polynucleotide T tract; 5,885 to 5,897, region with multiple deletions and substitutions; 6,815 to 6,824, polynucleotide A tract; HLA-F positions −960 to −787, gap bounded by 5' polynucleotide A and 3' polynucleotide G; −746 to −590, AG repetitive region; HLA-G positions 3,585 to 3,622, Polynucleotide T tract. Portions of these regions were ambiguous and were submitted to Genbank with *N*'s in the ambiguous positions. Genbank accession numbers for these sequences are AF523274 to AF523311 and AY645724 to AY645776.




| | Hap block 1 | | | | | | | Hap block 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | -2143 | -2106 | -2015 | -1589 | -1423 | -1389 | EF | 756 | 1691 | 3447 | 4297 | 4806 | 5073 | EF |
| Hap1 | | | | | | | 0.29 | | | | | | | 0.58 |
| Hap2 | | | | | | | 0.16 | | | | | | | 0.16 |
| Hap3 | | | | | | | 0.10 | | | | | | | 0.10 |

**Fig. 2** Deep resequencing of the HLA-E gene reveals new allelic variants, novel genomic variation, and an underlying haplotype block structure. The HLA-E gene is represented by a *cartoon* with the exon–intron structures indicated by *black boxes*. Immediately beneath, *vertical lines* and *arrows* indicate the positions of SNPs identified, the latter of which (*purple*) identify tagged SNPs. Beneath the *lines* and *arrows*, the positions and the allelic variants are indicated. Positions are designated according to accepted convention, with the antithymocyte globulin start codon being designated position 1 and the base immediately preceding it position −1 using Genbank accession NT_007592 as a source for genomic sequence. Haplotype blocks that subdivide the gene into two blocks are depicted beneath. At the bottom are indicated the sequences of the haplotypes with more than 0.1 expected frequency, comprising each block with colors indicating the tagged nucleotides that represent each position as indicated. Columns designated *EF* contain the expected frequency of the respective haplotype to the immediate left calculated from the data reported here. For example, the expected frequency of an individual in the population with *hap1* from *block 1* combined with *hap3* from *block 2* would be calculated as 0.29×0.10

### Linkage disequilibrium index, haplotype blocks, and identification of tagged amplicons

The linkage disequilibrium index (LDI) was used to quantify genetic diversity at multiple loci (Zhao, unpublished), and may be thought of as an extension to the traditional measurement of D'. LDI varies within the range (0,1) with value 0 under linkage equilibrium (LE) and value 1 under perfect linkage disequilibrium (LD). To detect haplotype blocks, a chi-square test was used to guide the detection, rather than using pairwise measurement D' (Gabriel et al. 2002). In that case each SNP consisting of two alleles is measured for LDI with neighboring SNPs and grouped into a single block if the LD, as computed by the chi-square test, is calculated to be high.

### Construction and expression of HLA-E, HLA-F, and HLA-G allelic variants

Allelic variants HLA-G*0103, HLA-G*0104, and HLA-G*0106 harbor a single amino acid substitution compared with the G*010103 allele, in addition to one or two synonymous changes. Both membrane-bound and soluble forms of these alleles were artificially generated by mutating all the polymorphic positions and subsequently cloned into a mammalian expression vector pNS using standard methods. Allelic forms of HLA-F were similarly constructed and cloned in expression vectors also as described. The membrane and soluble forms of HLA-G*010103, HLA-G*0103, HLA-G*010401, and HLA-G*0106 in pNS were introduced into B-LCL 721.221 by electroporation as described (Lee et al. 1998a). The expression of the membrane-bound HLA-G molecules was evaluated by flow cytometry 48 h after the transfection. Briefly, $2\times10^5$ cells were incubated with mAb 87G [specific for HLA-G (Lee et al. 1995)], W6/32 (pan HLA), or control mAb 16G1 (isotype matched) as described (Lee et al. 1998a). For the soluble form of the HLA-G allele, expression was evaluated by enzyme-linked immunosorbent assay on days 6 and 14 after transfection as described (Fujii et al. 1994).

## Results

Low allelic polymorphism is a hallmark of the human nonclassical class I antigens. Indeed, excepting one gross change introduced by a null mutation (HLA-G*0105N) and one subtle change by an amino acid substitution (E*0101 vs E*0103), there are no functional differences known among the few remaining allelic variants. Given the potential of HLA-E function to be qualitatively affected by

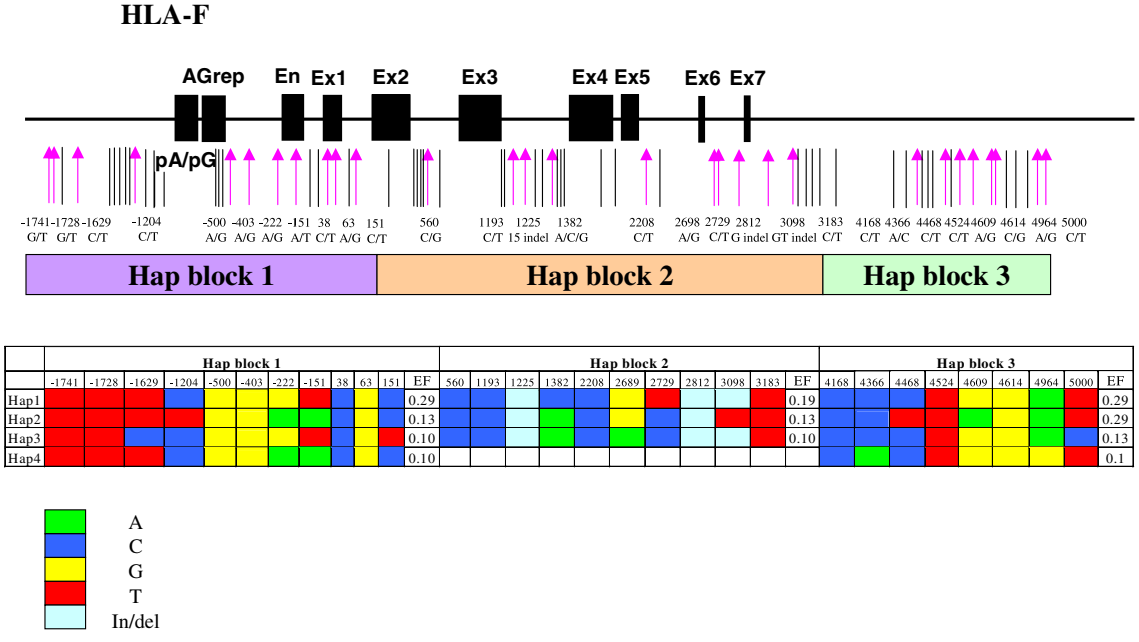



**Fig. 3** Deep resequencing of the HLA-F gene reveals new allelic variants, novel genomic variation, and an underlying haplotype block structure. The HLA-F gene is represented by a *cartoon* with the exon–intron structures indicated by *black boxes*. Immediately beneath, *vertical lines* and *arrows* indicate the positions of SNPs identified, the latter of which (*purple*) identify tagged SNPs. Beneath the *lines* and *arrows*, the positions and the allelic variants are indicated. Positions are designated according to accepted convention, with the antithymocyte globulin start codon being designated position 1 and the base immediately preceding it position −1 using Genbank accession NT_007592 as a source for genomic sequence. Haplotype blocks that subdivide the gene into three blocks are depicted beneath. At the bottom are indicated the sequences of the haplotypes with more than 0.1 expected frequency, comprising each block with colors indicating the tagged nucleotides that represent each position as indicated. Columns designated *EF* contain the expected frequencies of the respective haplotype to the immediate left calculated from the data reported here (e.g., frequency of individuals with *hap1 block1*, *hap2 block2*, *hap4-block3*=0.29×0.13×0.1=0.0038)

quantitative changes in surface levels (Lee et al. 1998b), the potential of alternative splicing of HLA-G (Ishitani and Geraghty 1992) to be affected by noncoding substitutions (Hunt et al. 2005), and the potential of the uniquely regulated expression of HLA-F (Lee and Geraghty 2003) to be affected by promoter or enhancer sequences, it would appear that any genetic association study targeting the nonclassical antigens might consider including a comprehensive analysis of relevant SNPs. Using long-range PCR protocols, DNA from 33 individuals was directly sequenced for each of these genes (Fig. 1). A summary of the origins of these DNAs is contained in Table 1.

This sequencing strategy was successful in general to yield contiguous data over most of each of the three nonclassical class I genes, despite the high homology among them and with the 15 other MHC class I genes and pseudogenes (Geraghty et al. 1992a). In addition, the protocol defined here yielded uniformly reproducible results over a large panel of DNAs including over 100 tested. Because of the nature of the PCR, however, portions of the genes could not be sequenced due to polynucleotide tracts or other complex sequences that were differentially amplified during the PCR. For the HLA-E gene, there were five such regions (detailed in Materials and methods) at positions indicated in Fig. 2. Most of these were polynucleotide stretches where an absolute base count was not possible. Altogether, 8,325 bp of DNA sequence was obtained from each of 33 examples of the HLA-E gene. From these data, 29 SNPs and 1 indel were identified.

Excepting similar regions that could not be sequenced, 6,600 bp of sequence was sampled from each of the 33 DNAs yielding a substantially increased number of 66 SNPs and 9 indels over that seen for the HLA-F locus (Fig. 3). The HLA-G locus yielded 5,591 bp of sequence for each of the 33 DNAs (Fig. 4). This locus was the most variable of the three containing 87 SNPs and 7 indels. Altogether, from the three loci, 20,516 bp of genomic sequence data was obtained from each of the 33 DNAs yielding 182 SNPs and 17 indels as positioned in Figs. 2, 3, and 4.

Of the SNPs identified in the HLA-F and HLA-G sequences, several were found within established promoter sequences (Table 2). One SNP occurred in the kB2 site of enhancer A (Gobin et al. 1998), while a second relatively frequent SNP occupied a central position in the SXY module (Gobin and van den Elsen 2000; Rousseau et al. 2004). In the latter case, all samples with −144 T had a 12-bp deletion in intron 3 (at position 1,193), and all of these

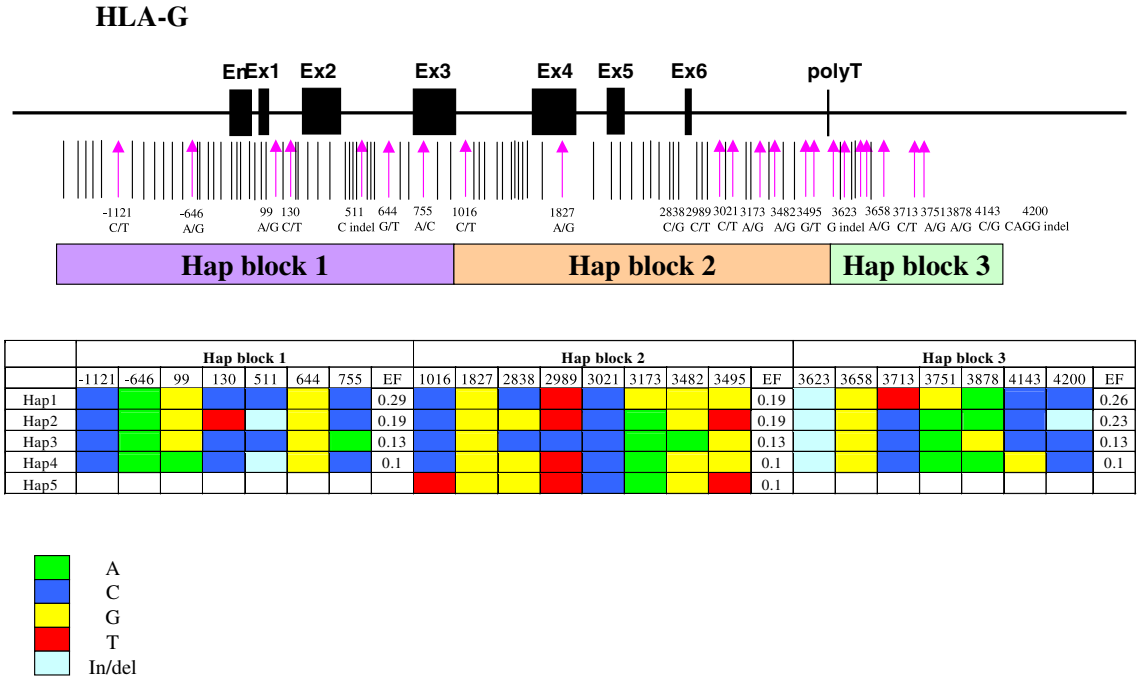



**Fig. 4** Deep resequencing of the HLA-G gene reveals new allelic variants, novel genomic variation, and an underlying haplotype block structure. The HLA-G gene is represented by a *cartoon* with the exon–intron structures indicated by *black boxes*. Immediately beneath, *vertical lines* and *arrows* indicate the positions of SNPs identified, the latter of which (*purple*) identify tagged SNPs. Beneath the *lines* and *arrows*, the positions and the allelic variants are indicated. Positions are designated according to accepted convention with the antithymocyte globulin start codon being designated position 1 and the base immediately preceding it position −1 using Genbank accession NT_007592 as a source for genomic sequence. Haplotype blocks that subdivide the gene into three blocks are depicted beneath. At the bottom are indicated the sequences of the haplotypes with more than 0.1 expected frequency, comprising each block with colors indicating the tagged nucleotides that represent each position as indicated. Columns designated EF contain the expected frequencies of the respective haplotype to the immediate left calculated from the data reported here (e.g., frequency of individuals with *hap1 block1*, *hap5 block2*, *hap4-block3*=0.29×0.1×0.1=0.0029)

samples were the HLA-F*010102 allele. Five SNPs in the HLA-G promoter were located within what have been defined as trophoblast specific response elements (Ober et al. 2003; Solier et al. 2001). Five additional variable positions were contained within the interferon-stimulated response element/gamma (interferon) activated site flanking region and the heat shock protein element in the HLA-G 5' flanking region. About 140 bp of the HLA-E promoter have been defined, and none of the SNPs identified in that gene were found within these elements.

To make the most effective use of these data for mapping studies and to gain an understanding of some of the evolutionary constraints among the three genes, the LDI was computed for all of the SNPs. From each of the derivative haplotypes, it was possible to identify haplotype-tagged SNPs (Chapman et al. 2003) within a haplotype block that, when queried, have a high probability of yielding the derivative information about all the SNPs in that block (Takeuchi et al. 2005) (blocks indicated in Figs. 2, 3, 4).

Given that all of the data were derived from phased chromosomes either because of homozygous MHCs or family data, and given subhaplotype structures for each gene, it was possible to investigate the linkage between the haplotype blocks among the three genes (Table 3). In general, the linkage or lack thereof reflected the respective distances between each locus, with HLA-E showing essentially no linkage to HLA-F, while exhibiting some linkage to portions of the HLA-G locus. Interestingly, while HLA-F is relatively close to HLA-G and while some blocks were highly linked between the two loci, block 1 of the HLA-G locus showed no significant linkage with any portion of the HLA-F block structure. Although it is not yet clear with regard to functional consequence, a specific breakdown of linkage between these loci, and indeed, other loci in the region, may be essential towards a complete understanding of any genetic linkage detected as associated with disease phenotype.

Part of the motivation for this study, and part of the motivation behind the recent IHWG workshop with regard to these loci, was to establish a definitive data set that could verify previously described data and possibly identify new allelic variants. Because the established database and nomenclature for HLA currently assigns allele names to consider only messenger RNA (mRNA) sequence variation (Marsh et al. 2005), we consider the variation discovered here under a similar division. Three new HLA-E alleles were found: one encoding a synonymous substitution (E*010304), another encoding two alterations in the 5' untranslated region (E*01010102), and intron 5 (E*01 030102) (Table 4). Similarly, three new coding sequence variants were identified, all of which contained nonsynonymous substitutions. HLA-F*0102 contained a C (Ala) to T (Val) substitution in the signal sequence at position 38 (codon 13), while HLA-F*0103 and HLA-F*0104 contained C (Pro) to A (Gln) at position 212 (codon 71) in exon 2 and the T (Ser) at position 814 (codon 272) was changed to C (Pro) in exon 4, respectively. At the HLA-G locus we identified three new coding sequence variants, one of which introduced a Ser to Phe substitution in exon 2 (nucleotide 110; codon 37, G*010401, Table 4). This work generated a total of nine new coding allelic variants and importantly gathered negative data that several reported variants may not exist. All of these data were submitted to, and the names listed were officially assigned by, the World Health Organization Nomenclature Committee and to the MHC database and the SNP database.

**Table 2** Promoter sequence variation in nonclassical class I

| | Regulator site | Position[a] | SNPs | References |
|---|---|---|---|---|
| HLA-F | | | | |
| Enhancer A | *k*B2 | −226 to −216 | −222 A to G | (Gobin et al. 1998; Gobin and van den Elsen 2000) |
| | *k*B1 | −209 to −199 | – | |
| SXY module | S | −171 to −165 | – | (Gobin and van den Elsen 2000; Rousseau et al. 2004) |
| | X1X2 | −149 to −131 | −144 A to T[b] | |
| | Y | −112 to −108 | – | |
| HLA-G | | | | |
| Trophoblast-specific response elements | TSRE | −1201 to −1100 | −1179 G to A<br>−1155 G to A<br>−1140 T to A<br>−1138 A to G<br>−1121 C to T | (Ober et al. 2003; Solier et al. 2001) |
| Interferon-stimulated response element/gamma (interferon) activated site flanking region | ISRE/GAS | −754 to −735 | −762 T to C<br>−725 C to G<br>−716 G to T | (Ober et al. 2003; Solier et al. 2001) |
| Heat shock protein element | HSE | −487 to −475 | −477 G to C | (Ober et al. 2003; Solier et al. 2001) |
| | HSE | | −486 C to A | |

[a]Positions defined according to Figs. 3 and 4
[b]All samples with −144 T have 12 bp (1,193) deletion in intron 3 and HLA-F*010102. Some of them have 15 bp (1,225) deletion in intron 3

**Table 3** Gene–gene independence test using Fisher exact test (*p* value)

| | | HLA-F | | | | HLA-G | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Block 1 | Block 2 | Block 3 | Total | Block 1 | Block 2 | Block 3 | Total |
| HLA-E | Block 1 | 0.30 | 0.20 | 0.71 | | 0.34 | 0.03 | 0.28 | |
| | Block 2 | 0.92 | 0.48 | 0.20 | | 0.02 | 0.05 | 0.06 | |
| | Total | | | | 0.85 | | | | 0.18 |
| HLA-F | Block 1 | | | | | 0.07 | 6*e-5 | 0.01 | |
| | Block 2 | | | | | 0.26 | 0.003 | 0.21 | |
| | Block 3 | | | | | 0.10 | 8*e-4 | 0.003 | |
| | Total | | | | | | | | 0.001 |

Although allelic variation at the amino acid level among the HLA-E, HLA-F, and HLA-G loci is limited, there remains interest in testing those differences that do exist for functional consequence. As part of the IHWG workshop and of this study, as we were identifying new variants, we used that information to assemble expression constructs that allowed for protein expression in mammalian cell lines and in bacteria useful for direct refolding and protein studies (Strong et al. 2003). There are four HLA-G amino acid variants and the HLA-G*0105N null mutant; all five constructs have been assembled and used in transfection studies (Table 5). Both membrane bound forms and soluble

**Table 4** Summary of previously known and new HLA-E, HLA-F, and HLA-G allelic variants

| HLA-E | | HLA-F | | HLA-G | |
|---|---|---|---|---|---|
| **E*01010101** | [a] | **F*01010101** | [a,c] | G*01010101 | [a,c] |
| *E*01010102* | [e,i,j] | *F*01010102* | [e,i,j] | *G*01010102* | [f,i,j] |
| **E*01030101** | [a,b] | *F*01010103* | [f,i,j] | *G*01010103* | [f,i,j] |
| *E*01030102* | [e,i,j] | *F*01010104* | [f,i,j] | *G*01010104* | [e,i,j] |
| E*010302 | [a,b] | *F*01010105* | [f,i,j] | *G*01010105* | [f,i,j] |
| E*010303 | unverified[d] | *F*01010106* | [e,i,j] | G*01010201 | [a,c] |
| *E*010304* | [f,h,j] | *F*01010107* | [f,i,j] | *G*01010202* | [f,i,j] |
| E*0104 | unverified[d] | F*01010201 | [a] | **G*010103** | [a,c] |
| | | *F*01010202* | [e,i,j] | G*010104 | unverified[d] |
| | | *F*01010203* | [f,i,j] | G*010105 | [c] |
| | | *F*01010204* | [e,i,j] | G*010106 | [a] |
| | | *F*01010205* | [f,i,j] | G*010107 | [c] |
| | | F*01010301 | [a] | G*010108 | [c] |
| | | *F*01010302* | [f,i,j] | *G*010109* | [f,h,j] |
| | | *F*01010303* | [e,i,j] | *G*010110* | [f,h,j] |
| | | *F*01010304* | [f,i,j] | G*0102 | unverified[d] |
| | | *F*0102* | [f,g,j] | **G*0103** | [a,c] |
| | | *F*01030101* | [f,g,j] | **G*010401** | [a,c] |
| | | *F*01030102* | [f,i,j] | G*010402 | unverified[d] |
| | | ***F*0104*** | [f,g,j] | G*010403 | unverified[d] |
| | | | | **G*0105N** | [a] |
| | | | | **G*0106** | [c] |
| | | | | *G*0107* | [e,g,j] |

Italics indicate the names of new variants. Bold lettering indicates cDNA is available from the IHWG.org as detailed in Table 5

[a]Verified in this report

[b]Verified in Grimsley et al. (2002)

[c]Identified in two or more independent reports

[d]Identified in a single report and resequencing from this study suggest these variants may not exist

[e]New variant allele found in single cell line in this study

[f]New variant allele confirmed in two or more cell lines in this study

[g]Nonsynonymous new variant

[h]Synonymous new variant

[i]New variant in intron and untranslated region

[j]The name listed for this sequence has been officially assigned by the WHO Nomenclature Committee. This follows the agreed policy that, subject to the conditions stated in the most recent Nomenclature Report (Marsh et al. 2005), names will be assigned to new sequences as they are identified. Lists of such new names will be published in the following WHO Nomenclature Report

**Table 5** Available HLA-E, HLA-F, and HLA-G cDNAs cloned into expression vectors [a, b]

| Alleles | Expression vectors |
|---|---|
| HLA-E*01010101 | pHN1, pNS, pLNCx, pHEBo |
| HLA-E*01030101 | pHN1, pNS, pLNCx |
| HLA-F*01010101 | pHN1, pNS, pcDNA, pHEBo |
| HLA-F*0104 | pNS |
| S·& M·HLA-G*010103 | pHN1, pNS, pLNCx, pHEBo |
| S·& M·HLA-G*0103 | pNS |
| S·& M·HLA-G*010401 | pNS |
| S·& M·HLA-G*0106 | pNS |

[a] pHN1 = bacterial expression; pNS, pLNCx, pcDNA, and pHEBo = eucaryotic expression

[b] All constructs are available from the IHWG cell and gene bank (http://www.ihwg.org/shared/cbankover.htm)

forms for all of these variants have been used effectively in transfection studies (Fig. 5). All constructs are available from the IHWG website (http://ihwg.org/).

## Discussion

In this work we set out to build a comprehensive framework of sequence data that can be used in the identification of genetic factors affecting HLA-E, HLA-F, and HLA-G expression and function. Both the expression and function of the three nonclassical antigens and the LD among the haplotype blocks between each gene suggest that genetic studies that focus on clinical outcomes relevant to pregnancy might benefit by examining the potential for contributions by all three loci. First, all three class I antigens are expressed by placental trophoblasts that are in direct contact with maternal immune cells, suggesting functional interactions that may be important to maternal immunity (Ishitani et al. 2003). A second consideration is clearly evident from the LD between the defined haplotype blocks. As might be expected from their close physical distance, HLA-F and HLA-G show significant linkage between portions of the genes, while other structural blocks showed no significant linkage. While HLA-E showed little linkage to HLA-F, there was significant linkage of the 3' block at HLA-E with the 5' block within the HLA-G locus. Therefore, a biological rationale for collaboration among these loci exists, as does a physical rationale for considering that any linkage effect of one locus may be detecting an effect of another. Of course, any such physical linkage may be taken as evidence of functional selection at both loci.

A further consideration for genetic association studies is a focus on not only coding variation but also on promoter and other regulatory variation. HLA-G mRNA is alternatively spliced (Ishitani and Geraghty 1992), and although only two of these alternative splicing variants have been shown to produce protein in vivo (Fujii et al. 1994), the possibility of genetic variation altering relative levels of these two proteins provides an intriguing potential for

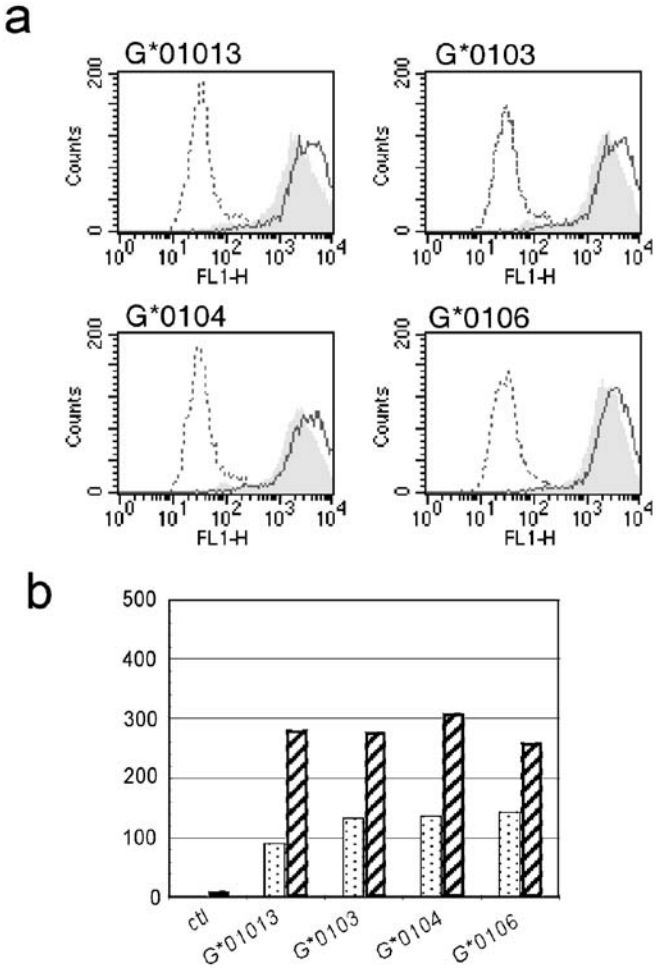



**Fig. 5** Expression of alternative nonsynonymous alleles of soluble and membrane HLA-G1. **a** cDNAs encoding the membrane forms of HLA-G*01013, HLA-G*0103, HLA-G*0104, and HLA-G*0106 and cloned into expression vector pNS were transfected into B-LCL 721.221 cells. Cell surface expression of HLA-G was evaluated 48 h after transfection by flow cytometry with 87G (*gray histograms*) and olG (*solid lines*), along with the isotype matched control antibody (*dashed lines*). **b** cDNAs encoding the soluble form of the same HLA-G alleles as indicated were introduced into 721.221, and the secretion of HLA-G was measured by enzyme-linked immunosorbent assay using w6/32 and 87G on supernatant sampled on day 3 (*dotted bars*) and day 11 (*hatched bars*) after transfection. The y-axis measures arbitrary units of fluorescence corresponding approximately to an equivalent number of nanograms per milliliter of the soluble form of HLA-G1 protein. *Bars* above position *ctl* report results from control supernatant similarly analyzed from untransfected 721.221 cells

functional effect. Further, as different cell-surface levels of HLA-E can affect the signaling through CD94-NKG2 (Lee et al. 1998b; Llano et al. 1998), this opens the possibility that variation in promoter or enhancer sequences that alter expression levels could have direct functional consequences. While the function of HLA-F is unknown, the overall patterns of surface expression provide evidence of a fundamental role in the immune response in both normal and maternal immunity (Ishitani et al. 2003). Whether amino acid substitutions or expression levels will affect this function may also be worth considering.

Although allelic variation (mRNA) at the nonclassical class I loci is low, a number of unconfirmed variations have accumulated over the years, and part of the intent of this study was to confirm as much of that as possible. Of the eight variants in the HLA-G*0101 group listed on the Anthony Nolan HLA database (Marsh et al. 2005), only four were identified in the chromosomes analyzed in this report, and further examination of over 100 DNAs in other studies also failed to identify them (data not shown). All of those variants that we could not verify were described in a single report (Kirszenbaum et al. 1999). It was also not possible to identify any examples of the HLA-G*0102 allele, the only nonsynonymous variant that could not be verified. All of the data we have assembled is available as raw sequence trace data in an accessible form for verification.

Regarding variation outside of coding sequences, it was possible to identify several common variants that alter experimentally established promoter elements (Table 2). Although limited, such variation underscores the potential for altered transcriptional levels and further emphasizes the rationale for identifying, through genetic analysis, additional new variants that affect transcription or transcript processing. Several specific promoter elements have been defined, but it is unlikely that all regulatory sequences in these genes have been identified. For example, multispecies conserved sequences are predicted to be relevant to controlling gene expression due to their extreme conservation over long separated lineages, while having no evident function based on prior knowledge (Margulies and Green 2003; Murphy et al. 2003). These and other relevant sequences may first be established as functional in affecting gene expression by genetic analysis alone.

One point made from these data derives from the amount of new SNP data we identified in this small study. Despite the extensive genomic sequencing and polymorphism analysis that the MHC has been subjected to (Geraghty et al. 2002, 2000; Horton et al. 2004; Stewart et al. 2004), and the number of studies published directly of coding sequence variation within HLA-E, HLA-F, and HLA-G, there were, nonetheless, a significant number of new SNPs identified in the small populations sampled in this study. In these three samples, 30% of the common variation identified was not present in any of the public databases, a number slightly lower than that found for genes elsewhere in the human genome (37%), but still highly significant in a search for causative functional variants. In that regard, 18 of the tagged SNPs were not included in public databases, making a prior compilation of haplotype blocks substantially incomplete relative to the present data.

On a more general note, most linkage mapping protocols for complex traits consist of a primary genome scan with low marker density followed by high-density genotyping around linkage peaks (or complete sequencing) (Wiltshire et al. 2005). In studies focused on the MHC, such peaks will include a few or several of the well-characterized genes in the region, and once target regions of the MHC have been identified, deep resequencing of those regions will be necessary to precisely localize causative genetic components. In that regard, this work demonstrates the feasibility of an approach to analyze any of the more than 140 genes that lie within the MHC, because resequencing HLA class I genes, among 17 highly similar class I sequences contained with this region, is likely the most challenging example.

**Acknowledgements** The expert technical assistance of Scott Medley is gratefully acknowledged. This work was supported by National Institute of Health grants AI33484 and AI49245 to DEG. A.I. was supported in part by Grant in Aid number 05671391 from the Ministry of Education, Science, and Culture in Japan and by National Institutes of Health grant AI49213. L.P.Z. was supported by CA106320.

## References

Agrawal S, Pandey MK (2003) The potential role of HLA-G polymorphism in maternal tolerance to the developing fetus. J Hematother Stem Cell Res 12:749–756

Chapman JM, Cooper JD, Todd JA, Clayton DG (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31

Daza-Vamenta R, Glusman G, Rowen L, Guthrie B, Geraghty DE (2004) Genetic divergence of the rhesus macaque major histocompatibility complex. Genome Res 14:1501–1515

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res 8:186–194

Fujii T, Ishitani A, Geraghty DE (1994) A soluble form of the HLA-G antigen is encoded by a messenger ribonucleic acid containing intron 4. J Immunol 153:5516–5524

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

Gazit E, Slomov Y, Goldberg I, Brenner S, Loewenthal R (2004) HLA-G is associated with pemphigus vulgaris in Jewish patients. Hum Immunol 65:39–46

Geraghty DE (1993) Structure of the HLA class I region and expression of its resident genes. Curr Opin Immunol 5:3–7

Geraghty DE, Koller BH, Orr HT (1987) A human major histocompatibility complex class I gene that encodes a protein with a shortened cytoplasmic segment. Proc Natl Acad Sci U S A 84:9145–9149

Geraghty DE, Wei XH, Orr HT, Koller BH (1990) Human leukocyte antigen F (HLA-F). An expressed HLA gene composed of a class I coding sequence linked to a novel transcribed repetitive element. J Exp Med 171:1–18

Geraghty DE, Koller BH, Hansen JA, Orr HT (1992a) The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. J Immunol 149:1934–1946

Geraghty DE, Stockschleader M, Ishitani A, Hansen JA (1992b) Polymorphism at the HLA-E locus predates most HLA-A and -B polymorphism. Hum Immunol 33:174–184

Geraghty DE, Fortelny S, Guthrie B, Irving M, Pham H, Wang R, Daza R, Nelson B, Stonehocker J, Williams L, Vu Q (2000) Data acquisition, data storage, and data presentation in a modern genetics laboratory. Rev Immunogenet 2:532–540

Geraghty DE, Daza R, Williams LM, Vu Q, Ishitani A (2002) Genetics of the immune response: identifying immune variation within the MHC and throughout the genome. Immunol Rev 190:69–85

Gobin SJ, van den Elsen PJ (2000) Transcriptional regulation of the MHC class Ib genes HLA-E, HLA-F, and HLA-G. Hum Immunol 61:1102–1107

Gobin SJ, Keijsers V, van Zutphen M, van den Elsen PJ (1998) The role of enhancer A in the locus-specific transactivation of classical and nonclassical HLA class I genes by nuclear factor kappa B. J Immunol 161:2276–2283

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Grimsley C, Kawasaki A, Gassner C, Sageshima N, Nose Y, Hatake K, Geraghty DE, Ishitani A (2002) Definitive high resolution typing of HLA-E allelic polymorphisms: identifying potential errors in existing allele data. Tissue Antigens 60:206–212

Hirankarn N, Kimkong I, Mutirangura A (2004) HLA-E polymorphism in patients with nasopharyngeal carcinoma. Tissue Antigens 64:588–592

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899

Hunt JS, Petroff MG, McIntire RH, Ober C (2005) HLA-G and immune tolerance in pregnancy. FASEB J 19:681–693

Hylenius S, Andersen AM, Melbye M, Hviid TV (2004) Association between HLA-G genotype and risk of pre-eclampsia: a case-control study using family triads. Mol Hum Reprod 10:237–246

Ishitani A, Geraghty DE (1992) Alternative splicing of HLA-G transcripts yields proteins with primary structures resembling both class I and class II antigens. Proc Natl Acad Sci U S A 89:3947–3951

Ishitani A, Kishida M, Sageshima N, Yashiki S, Sonoda S, Hayami M, Smith AG, Hatake K (1999) Re-examination of HLA-G polymorphism in African Americans. Immunogenetics 49:808–811

Ishitani A, Sageshima N, Lee N, Dorofeeva N, Hatake K, Marquardt H, Geraghty DE (2003) Protein expression and peptide binding suggest unique and interacting functional roles for HLA-E, F, and G in maternal–placental immune recognition. J Immunol 171:1376–1384

Kaiser BK, Barahmand-Pour F, Paulsene W, Medley S, Geraghty DE, Strong RK (2005) Interactions between NKG2x immunoreceptors and HLA-E ligands display overlapping affinities and thermodynamics. J Immunol 174:2878–2884

Kirszenbaum M, Djoulah S, Hors J, Prost S, Dausset J, Carosella ED (1999) Polymorphism of HLA-G gene and protein. J Reprod Immunol 43:105–109

Koller BH, Geraghty DE, Shimizu Y, DeMars R, Orr HT (1988) HLA-E. A novel HLA class I gene expressed in resting T lymphocytes. J Immunol 141:897–904

Kovats S, Main EK, Librach C, Stubblebine M, Fisher SJ, DeMars R (1990) A class I antigen, HLA-G, expressed in human trophoblasts. Science 248:220–223

Le Bouteiller P, Rodriguez AM, Mallet V, Girr M, Guillaudeux T, Lenfant F (1996) Placental expression of HLA class I genes. Am J Reprod Immunol 35:216–225

Lee N, Geraghty DE (2003) HLA-F surface expression on B cell and monocyte cell lines is partially independent from tapasin and completely independent from TAP. J Immunol 171:5264–5271

Lee N, Malacko AR, Ishitani A, Chen MC, Bajorath J, Marquardt H, Geraghty DE (1995) The membrane-bound and soluble forms of HLA-G bind identical sets of endogenous peptides but differ with respect to TAP association. Immunity 3:591–600

Lee N, Goodlett DR, Ishitani A, Marquardt H, Geraghty DE (1998a) HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. J Immunol 160:4951–4960

Lee N, Llano M, Carretero M, Ishitani A, Navarro F, Lopez-Botet M, Geraghty DE (1998b) HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. Proc Natl Acad Sci U S A 95:5199–5204

Llano M, Lee N, Navarro F, Garcia P, Albar JP, Geraghty DE, Lopez-Botet M (1998) HLA-E-bound peptides influence recognition by inhibitory and triggering CD94/NKG2 receptors: preferential response to an HLA-G-derived nonamer. Eur J Immunol 28:2854–2863

Margulies EH, Green ED (2003) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. Cold Spring Harb Symp Quant Biol 68:255–263

Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Hurley CK, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI, Trowsdale J (2005) Nomenclature for factors of the HLA system, 2004. Tissue Antigens 65:301–369

Murphy WJ, Bourque G, Tesler G, Pevzner P, O'Brien SJ (2003) Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. Hum Genomics 1:30–40

Nicolae D, Cox NJ, Lester LA, Schneider D, Tan Z, Billstrand C, Kuldanek S, Donfack J, Kogut P, Patel NM, Goodenbour J, Howard T, Wolf R, Koppelman GH, White SR, Parry R, Postma DS, Meyers D, Bleecker ER, Hunt JS, Solway J, Ober C (2005) Fine mapping and positional candidate studies identify HLA-G as an asthma susceptibility gene on chromosome 6p21. Am J Hum Genet 76:349–357

Ober C, Aldrich CL, Chervoneva I, Billstrand C, Rahimov F, Gray HL, Hyslop T (2003) Variation in the HLA-G promoter region influences miscarriage rates. Am J Hum Genet 72:1425–1435

Pfeiffer KA, Fimmers R, Engels G, van der Ven H, van der Ven K (2001) The HLA-G genotype is potentially associated with idiopathic recurrent spontaneous abortion. Mol Hum Reprod 7:373–378

Rousseau P, Masternak K, Krawczyk M, Reith W, Dausset J, Carosella ED, Moreau P (2004) In vivo, RFX5 binds differently to the human leucocyte antigen-E, -F, and -G gene promoters and participates in HLA class I protein expression in a cell type-dependent manner. Immunology 111:53–65

Shiroishi M, Tsumoto K, Amano K, Shirakihara Y, Colonna M, Braud VM, Allan DS, Makadzange A, Rowland-Jones S, Willcox B, Jones EY, van der Merwe PA, Kumagai I, Maenaka K (2003) Human inhibitory receptors Ig-like transcript 2 (ILT2) and ILT4 compete with CD8 for MHC class I binding and bind preferentially to HLA-G. Proc Natl Acad Sci U S A 100:8856–8861

Solier C, Mallet V, Lenfant F, Bertrand A, Huchenq A, Le Bouteiller P (2001) HLA-G unique promoter region: functional implications. Immunogenetics 53:617–625

Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, Humphray SJ, Hunt S, Mungall AJ, Osoegawa K, Palmer S, Roberts AN, Rogers J, Sims S, Wang Y, Wilming LG, Elliott JF, de Jong PJ, Sawcer S, Todd JA, Trowsdale J, Beck S (2004) Complete MHC haplotype sequencing for common disease gene mapping. Genome Res 14:1176–1187

Strong RK, Holmes MA, Li P, Braun L, Lee N, Geraghty DE (2003) HLA-E allelic variants. Correlating differential expression, peptide affinities, crystal structures, and thermal stabilities. J Biol Chem 278:5082–5090

Takeuchi F, Yanai K, Morii T, Ishinaga Y, Taniguchi-Yanai K, Nagano S, Kato N (2005) Linkage disequilibrium grouping of SNPs reflecting haplotype phylogeny for efficient selection of tag SNPs. Genetics 170:291-304

Tripathi P, Abbas A, Naik S, Agrawal S (2004) Role of 14-bp deletion in the HLA-G gene in the maintenance of pregnancy. Tissue Antigens 64:706–710

Wiltshire S, Morris AP, McCarthy MI, Cardon LR (2005) How useful is the fine-scale mapping of complex trait linkage peaks? Evaluating the impact of additional microsatellite genotyping on the posterior probability of linkage. Genet Epidemiol 28:1–10