A Comprehensive Study of Polymorphic Sites along the HLA-G Gene: Implication for Gene Regulation and Evolution

Erick C. Castelli,*1 Ceulso T. Mendes-Junior,2 Luciana C. Veiga-Castelli,3 Michel Roger,4 Philippe Moreau,5 and Eduardo A. Donadi3

1Laboratório de Genética Molecular e Citogenética, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Brasil
2Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto-SP, Brasil
3Divisão de Imunologia Clínica, Departamento de Clínica Médica, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto-SP, Brasil
4Département de Microbiologie et Immunologie, Centre Hospitalier de L’Université de Montréal, Montréal, Québec, Canada
5Commissariat à l’Energie Atomique et aux Energies Alternatives, Université Paris VII/DSV/l2BM/Service de Recherches en Hématologie-Immunologie, Institut Universitaire d’Hématologie, Hôpital Saint-Louis, Paris, France

*Corresponding author: E-mail: erick.castelli@gmail.com.

Associate editor: Yoko Satta

Abstract

HLA-G molecule plays an important role on immune response regulation and has been implicated on the inhibition of T and natural killer cell cytolytic function and inhibition of allogeneic T-cell proliferation. Due to its immune-modulator properties, the HLA-G gene expression has been associated with the outcome of allograft and of autoimmune, infectious, and malignant disorders. Several lines of evidence indicate that HLA-G polymorphisms at the 5′-upstream regulatory region (5′ URR) and 3′-untranslated region (3′ UTR) may influence the HLA-G expression levels. Because Brazilians represent one of the most heterogeneous populations in the world with the widest HLA-G coding region variability already detected among the studied populations, a high level of variability and haplotype diversity would be expected in Brazilians. On this basis, the 5′ URR, coding, and 3′ UTR variability were evaluated in a Brazilian series consisting of 100 healthy bone marrow donors, as well as the linkage disequilibrium pattern along the gene and the extended haplotypes encompassing several gene segment variations. The HLA-G locus seems to present six different HLA-G lineages showing functional variations mainly in nucleotides of the regulatory regions. Differences were observed at the 5′ URR in positions that either coincide with or are close to transcription factor–binding sites and at the 3′ UTR mainly in positions that have already been reported to influence HLA-G mRNA availability. We report several lines of evidence for balancing selection acting on the regulatory regions, which may indicate that these HLA-G lineages may be related to the differential HLA-G expression profiles.

Key words: HLA-G, 5′ upstream regulatory region (5′ URR), promoter, 3′ untranslated region (3′ UTR), polymorphism, haplotypes, Brazilians.

Introduction

The HLA complex is a 3.6 Mb high-density gene region located at the 6p21.3 chromosome region, encompassing more than 200 genes (Klein and Sato 2000a, 2000b). The HLA-G gene is a nonclassical class I HLA locus, which, like its classical counterparts, is composed of eight exons and seven introns. In contrast to classical class I loci, HLA-G has a stop codon at exon 6, leading to a short cytoplasmic tail, exhibits a 5′ upstream regulatory (or promoter) region (5′ URR) extending at least 1.4 kb from the initial ATG (Solier et al. 2001; Moreau et al. 2009) and presents an extended 3′ untranslated region (3′ UTR).

HLA-G is predominantly expressed at the maternal–fetal interface, particularly in the fetal extravillous cytotrophoblast cells, placental macrophages, and mesenchymal chronic villi (Berger et al. 2010) and has primarily been associated with maternal–fetal tolerance (Carosella, Moreau, et al. 2008). HLA-G is believed to protect the fetus against trophoblast damage caused by maternal natural killer (NK) (Rouas-Freiss et al. 1997) and T-cytotoxic cells (CTLs) (Le Gal et al. 1999) during pregnancy (Rouas-Freiss et al. 1997; Berger et al. 2010), also preventing proliferation of CD4+ T cells (LeMaoult et al. 2004) and tolerizing dendritic cells (Ristic et al. 2005). Apart from pregnancy, HLA-G expression in nonpathological conditions is restricted, detected in the thymus, cornea, proximal nail matrix, pancreas, and hematopoietic stem cells (Crisa et al. 1997; Mallet et al. 1999; Le Discorde et al. 2003; Menier et al. 2004; Ito et al. 2005; Cirulli et al. 2006). In pathological situations, HLA-G expression is observed in numerous tumors, viral infections, inflammatory and autoimmune diseases, and engrailed tissues (Carosella, Moreau, et al. 2008; Donadi et al. 2011).

HLA-G molecules may protect cells from CD8 and NK cell–mediated cytolysis through direct binding to the...
ILT-2 (IL1R1 and CD85j), ILT-4 (IL1R2 and CD85d), and KIR2DL4 (CD158d) leukocyte receptors (Colonna et al. 1997; Cosman et al. 1997; Colonna et al. 1998; Allan et al. 1999; Ponte et al. 1999; Rajagopalan and Long 1999). Dendritic cell function is also inhibited by interaction with ILT-2 and ILT-4 receptors (Carosella and Horuzsko 2007), which also interact with other class I HLA molecules, but the highest affinity is for HLA-G (Brown et al. 2004; Carosella, Favier et al. 2008; Donadi et al. 2011).

In contrast to the classical HLA class I loci, limited HLA-G coding region variability has been observed in worldwide populations (Castelli, Mendes-Junior et al. 2007). Based on the gametic phase (haplotypes) of 73 single-nucleotide polymorphisms (SNP) observed between exon 1 and intron 6, 46 HLA-G alleles have been currently recognized by the International Immunogenetics Information System (IMGT, database 3.1.0, July 2010). Most of these variation sites are either intronic or synonymous substitutions, yielding only 15 different full-length proteins (two null alleles), of which only 5 are frequently observed in worldwide populations (Castelli, Mendes-Junior et al. 2007; Castelli et al. 2010).

A relatively higher degree of variation is observed in the 5′ UTR (Tan et al. 2005) and 3′ UTR (Castelli et al. 2010). To date, 29 SNPs have been identified in the HLA-G promoter region (Hvid et al. 2004; Tan et al. 2005; Hvid et al. 2006; Berger et al. 2010), which may be implicated in the regulation of HLA-G expression, considering that many of these polymorphisms are within or close to known or putative regulatory elements. In addition, polymorphic sites at the 5′ UTR may be in linkage disequilibrium (LD) with the eight polymorphic sites identified at the 3′ UTR (Nicolae et al. 2005; Castelli et al. 2010), some of them influencing alternative splicing (Aubouef et al. 2002), and mRNA stability (Rousseau et al. 2003). At least three polymorphic sites at the 3′ UTR are associated with the regulation of HLA-G expression levels: the 14-bp deletion/insertion polymorphism (Hiby et al. 1999; O’Brien et al. 2001; Hvid et al. 2003; Rousseau et al. 2003), the presence of guanine in the position +3142, which increases the affinity of specific microRNAs (miRNA) for HLA-G mRNA, decreasing HLA-G expression (Tan et al. 2007; Veit and Chies 2009), and the presence of an adenine at position +3187, modifying an AU-rich motif in the HLA-G mRNA and decreasing its stability (Yie et al. 2008).

In a previous study evaluating the 3′ UTR polymorphic sites associated with posttranscriptional HLA-G regulation in the Brazilian population, we showed that their frequencies were higher than 5% and were in LD, suggesting that their influence on HLA-G expression is not independent (Castelli et al. 2010). Although several studies have demonstrated the importance of both the 5′ URR and the 3′ UTR in the HLA-G expression profile, a complete LD and haplotype evaluation along the entire HLA-G gene has not been performed. In analogy to what has been observed for the 3′ UTR, a similar pattern of LD would be expected in the promoter region or even in the entire gene. On this basis, we analyzed the variability of the promoter, coding, and 3′ UTR of the HLA-G gene and its haplotype structure in a series of Southeastern Brazilians.

**Materials and Methods**

**Subjects**

The Local Research Ethics Committee approved the protocol of the study (Protocol HCRP #12398/2004), and all participants gave written informed consent before blood withdrawal. A total of 100 healthy unrelated bone marrow donors from different regions of the State of São Paulo, Southeastern Brazil, were randomly selected and their DNA was obtained using a salting-out procedure (Miller et al. 1988).

**HLA-G Promoter Region Polymorphism Assignment**

For the evaluation of the HLA-G promoter region variation, a polymerase chain reaction (PCR)-amplified fragment of approximately 1752-bp encompassing the −1446 (5′ URR) and +388 (exon 2) nucleotides was produced using the GPRMO.S—ACATTCTAGAGCTTCACAAGATG (Ober et al. 2003; Tan et al. 2005) and GPRMO.R—GCCCTTGGTGTCCGTGTC primers. Amplification was performed in a final volume of 50 μl containing 1× PCR buffer (70 mM Tris–HCL pH 8.8, 20 mM (NH4)2SO4, 3 mM MgCl2, 1 mM dithiothreitol, 1% Triton X 100, 50 μg bovine serum albumin), 0.2 μM of each dNTP, 12 pmol of each primer, 1.5 units of platinum DNA polymerase (Invitrogen, Carlsbad, CA) and 200 ng of genomic DNA. The initial denaturation cycle was carried out at 94 °C for 3 min, followed by 32 cycles at 94 °C for 30 s, 59 °C for 30 s, and 72 °C for 135 sec and by a final extension step at 72 °C for 5 min. The amplification product was evaluated using 1% agarose gel.

PCR products were directly sequenced using the primers: G-908R (5′-TTCTACCTCAGTGTGTAAGTGTC-3′), G-830F (5′-CACACGAAAATCTAGGGCTAC-3′), GPR-247 (5′-CTCAAGCGTGGCTCTCGGTC-3′), and G-304R (5′-GCCAACGTCTCTGTCAGT-3′) (Ober et al. 2003); HLA-G AS-FW-G (5′-AACAGTGCTAGACCGCAC-3′), HLA-GAS-RE-G (5′-GAAGAGGCTTCCGGGC-3′), HLA-GAS-FW-A (5′-AACAGTGCTAGACCGCA-3′), and HLA-GAS-RE-A (5′-GAAGAGGCTTCCGGGT-3′) (Tan et al. 2005); HGO1F (5′-TAAAAGTCCCGTCTCAACCCCA-3′) (Castelli et al. 2010); and the same primers used for the promoter amplification (5′-exon 1, intron 1, and part of exon 2), which are not considered to be a part of the HLA-G promoter region. All the sequences obtained were aligned with each other and each variation site detected was individually annotated using the SNPex software (http://www.fmrp.usp.br/immunogen/snpex/), specifically designed to store and manage the variation sites in the HLA-G locus. Although the promoter region between −1050 and −770 nucleotides presents variation as described elsewhere (Tan et al. 2005), three of four variation sites described in this region did not reach polymorphic frequencies exceeding 1% [just one occurrence already detected among 268

**Downloaded from** https://academic.oup.com/mbe/article-abstract/28/11/3069/1045931 by guest on 03 March 2019
sampled chromosomes (Tan et al. 2005); thus, these variation sites were not included in our study.

**HLA-G Coding and 3’ UTR Characterization**

For the HLA-G coding region polymorphism assignment, all individuals had their HLA-G alleles defined by nucleotide sequence variations encompassing exons 1–4 (approximately 1975 bp), as described elsewhere (Castelli et al. 2010). The complete sequence between the first base of exon 1 and the last base of exon 2 (hence including intron 1) and the complete sequence of exons 3 and 4 were evaluated, encompassing all the variation sites already described for these regions (IMGT database version 3.10, July 2010). Additionally, three variation sites at the beginning of intron 3 were also evaluated. This procedure allowed the analysis of 43 previously described variation sites: 2 at exon 1 (positions 15 and 36), 5 at intron 1 (positions 99, 126, 130, 147, and 188), 13 at exon 2 (positions 234, 239, 280, 292, 293, 297, 306, 324, 361, 362, 366, 372, and 408), 12 at exon 3 (positions 706, 726, 727, 738, 741, 748, 755, 778, 814, 871, 902, and 934), 3 at intron 3 (positions 1016, 1019, and 1054), and 8 at exon 4 (positions 1590, 1659, 1681, 1682, 1734, 1799, 1800, and 1827).

The 3’ UTR of the HLA-G locus (also designated as exon 8) was evaluated by nucleotide sequence variations encompassing +2945 and +3259 nucleotides (314 bp between the 3’ ends of the primers), by using a methodology described elsewhere (Castelli et al. 2010). At least eight variation sites were described in this region, including the 14-bp deletion/insertion polymorphism and the +3142 and +3187 SNPs, known to be relevant for HLA-G expression control (Castelli et al. 2010). For both coding and 3’ UTR amplification, PCR products were directly sequenced using the primers proposed in the original manuscript (Castelli et al. 2010).

The amplification for the coding and promoter regions has an overlap of about 400 nucleotides, encompassing a highly polymorphic region that includes exon 1, intron 1, and most of exon 2. This strategy increases the probability of identifying a potential preferential amplification occurring in one of the two amplified regions.

It should be mentioned that part of the data concerning the coding and the 3’ UTR of the HLA-G locus presented in this manuscript have been previously published (Castelli et al. 2010) because the samples used in both studies overlap by about 70%.

**Haplotyping Procedures**

Currently, the sequencing of PCR products does not allow the determination of the phase of any variation sites detected or haplotypes, except in cases of homozygous samples or in cases in which all but one variation site is homozygous. Thus, computational inferences using probabilistic models are necessary to obtain haplotypes.

The presence of a significant association between all HLA-G variation sites was evaluated by means of a likelihood ratio test of LD (Excoffier and Slatkin 1998) using the ARLEQUIN version 3.1 software (http://cmpg.unibe.ch/software/arlequin3/) (Excoffier et al. 2005) and Haploview 4.2 (http://www.broad.mit.edu/mpg/haploview/) (Barrett et al. 2005). Given the positive association but unknown gametic phase, the PHASE (http://stephenslab.uchicago.edu/software.html) method (Stephens et al. 2001), implemented by the PHASE v2 package (Mac OS version) (Stephens et al. 2001), was used to assign the most probable haplotype constitution of each sample.

The database containing the variation sites was loaded into the PHASE software and the analyses were performed together with the following parameters using six independent runs with different seed values: 1) number of interactions: 1000, 2) thinning interval: 1, and 3) burn-in value: 100. Independent runs yielded the same results for all samples, except for four samples, in which a rare SNP allele with just one occurrence was present, such as the one present in the G*01:09 allele. Such SNPs occur in the coding region and these alleles were only represented once in the present sample, leading the PHASE algorithm to produce haplotypes with very low probability values (50%). To circumvent this problem, we opted to include in the database two additional heterozygous samples for these rare SNPs, informing the phase of the haplotypes bearing these rare SNPs based on the IMGT coding region official sequences. We performed six additional independent runs with this new data set, specifying the known phase for those two additional samples and using the same parameters described above. Independent runs yielded the same results for all samples, matching the results with the previous six runs, but improving the inference quality for those four samples with rare alleles. Additionally, we used polymorphic sites of each of the evaluated regions separately (promoter, coding, and 3’ UTR) as well as a small subset of polymorphic sites spread across these three regions, to infer haplotypes by the same method described above. The results obtained in these additional analyses were compatible with the ones obtained using all polymorphic sites altogether.

To better evaluate the reliability of the PHASE method, two other approaches were used. First, several HLA-G promoter haplotypes were confirmed by directional sequencing of heterozygous samples by using the primers described in the promoter assignment section HLA-GAS-FW-G and HLA-GAS-FW-A, both specific for the variations found in the −1305 position, and HLA-GAS-RE-G and HLA-GAS-RE-A, both specific for the variations found in the +15 position. Second, samples carrying some common haplotypes had their HLA-G gene amplified from the promoter to the 3’ UTR region in a single PCR reaction (by using the primer GPROMO.S from the promoter and HG08R from the 3’ UTR) using a high-fidelity DNA polymerase (Platinum Taq DNA Polymerase High Fidelity; Invitrogen). The amplification product was cloned into an appropriate vector (TOPO XL PCR Cloning Kit; Invitrogen) and sequenced, in order to define the phase of all the variation sites in the evaluated region. No discrepancies between the sequences obtained in the cloning procedures and by the computational inference methods were found.

**Statistical Analysis**

Allelic frequencies and observed heterozygosity (h O) were computed by the direct counting method. Adherences of
genotypic proportions to expectations under Hardy–Weinberg equilibrium were tested by the exact test of Guo and Thompson (Guo and Thompson 1992) employing the GENEPOP 3.4 software (http://genepop.curtin.edu.au/) (Raymond and Rousset 1995). The expected heterozygosity values (hₚ) and haplotype diversity as well as their standard deviations were estimated by the ARLEQUIN 3.1 program (Excoffier et al. 2005). Haplotype networks were constructed by using the median joining algorithm implemented in the Network 4.6.0.0 program (http://www.fluxus-engineering.com/sharenet.htm) (Bandelt et al. 1995). Tajima’s D and Fu and Li’s F neutrality tests were performed using DNAsp 5.1 (http://www.ub.edu/dnasp/) (Librado and Rozas 2009). The Ewens–Watterson neutrality test was performed using the PyPop 0.7.0 software (Lancaster et al. 2007).

Results

HLA-G Variation and LD Patterns

We evaluated the frequency of 76 previously reported variation sites along the HLA-G locus, including 25 SNPs at the promoter region, 43 SNPs at the coding region, and 8 variation sites at the 3’ UTR. Because 21 SNPs in the coding region were monomorphic, they were not considered for further analyses, that is, SNPs at the 234, 239, 280, 293, 306, 324, 361, 362, 366, 408, 726, 727, 738, 741, 778, 934, 1659, 1681, 1682, 1734, and 1800 positions. Of the remaining 55 variation sites, 51 were polymorphic (frequencies higher than 1%) and 4 were only represented once, achieving a frequency of 0.5% (positions +297 allele A, +871 allele A, +902 allele C, and +1016 allele T). These SNPs are associated with the HLA-G alleles known as G*01:04:05, G*01:01:09, G*01:09, and G*01:01:02:02, respectively. With the exception of the SNP at the −369 position (P = 0.0338) and the 14-bp polymorphism at the 3’ UTR (P = 0.0127), the genotypes of the 53 remaining polymorphic sites fit Hardy–Weinberg expectations. The nucleotide diversity at the HLA-G locus was 0.00643, about eight times higher than the human genome average (0.00075) (Sachidanandam et al. 2001; Tan et al. 2005).

The presence of a significant association between all HLA-G variation sites was evaluated by means of a likelihood ratio test of LD (Excoffier and Slatkin 1998) using the ARLEQUIN version 3.1 program (Excoffier et al. 2005) and Haploview 4.2 (Barrett et al. 2005). Figure 1 illustrates the pattern of LD along the HLA-G locus. It can be noticed that the majority of pairs of variation sites did present LD. Figure 1A illustrates the LD pattern by using polymorphisms with a minimum allele frequency (MAF) of 2%. Three haplotype blocks were defined using the confidence interval method implemented by the Haploview software. The first block is composed of the first four SNPs of the promoter region (−1306, −1179, −1155, and −1140), the second block is composed of variation sites of both promoter and coding regions (from −1138 to +1590), and the last block is composed of variation sites present in the 3’ UTR (from 14 bp to +3196). In addition, another LD plot was designed by using variation sites with MAF of 10% (fig. 18). It can be noticed that, with the exception of the SNP at the position +3035 in the 3’ UTR region, most of the remaining variations sites (11 for coding, 12 for promoter, and 7 for the 3’ UTR region) did present strong LD with each other. The lack of LD concerning the +3035 SNP could be due to the occurrence of two independent C > T mutations at such position in very different HLA-G lineages (HG0103 and HG010103, at opposite sides of fig. 2). Therefore, it can be observed that a strong LD is maintained throughout the whole gene. The promoter SNP at the −725 position (C/G/T) was removed from both plots because Haploview only accepts biallelic SNPs. Nevertheless, the LD test performed by Arlequin did corroborate the Haploview findings, indicating with few exceptions a strong LD between the −725 SNP and others (data not shown).

Given the positive association between pairs of SNPs (fig. 1) but unknown gametic phase, the PHASE method (Stephens et al. 2001) was used to determine the most probable haplotype constitution of each sample, as described earlier. All variation sites evaluated were included in this approach, encompassing SNPs of the regulatory, coding, and 3’ UTR. Twenty-eight different HLA-G extended haplotypes were found, with frequencies ranging from 0.5% to 21.0% and a haplotype diversity of 0.906. To evaluate the similarity among these haplotypes and cluster them into very similar haplotype groups, a haplotype network was created by using these 28 haplotype sequences (fig. 2). It can be noticed that these haplotypes are arranged in six haplotype lineages. Members of a same lineage share: 1) very similar promoter haplotypes that only differ in a maximum of two variation sites, 2) coding region variations that produce the same HLA-G protein or rare alleles that originated from point mutation in an ancestral high-frequency allele, and 3) a specific 3’ UTR haplotype (table 2).

The first lineage (HG010101) is isolated from the others in a single cluster. Inside this group, it can be noticed three main subdivisions, in which the first subdivision (HG010101a) is composed of three haplotypes presenting the same coding and very similar 3’ UTR and promoter sequences (fig. 3); the second subdivision (HG010101b) is represented by two haplotypes that present the same promoter and 3’ UTR sequences and very similar coding haplotypes (fig. 3); the third subdivision (HG010101c) is separated from the others in a single cluster and they all present the same coding (the G*01:01:01:05 allele or alleles originated from it by single mutations) and 3’ UTR sequences as well as a very similar promoter haplotype. The second lineage (HG0103) is isolated in a cluster and it presents the coding sequence of the G*01:03 allele and the same 3’ UTR sequence as well as very similar promoter haplotypes. The third lineage (HG0104) presents very similar promoter and coding sequences (all related to the G*01:04 allele group) and the same 3’ UTR haplotype. The fourth lineage (HG010102) is isolated from the others and it presents the same promoter sequence and very similar coding and 3’ UTR sequences (usually alleles originated from G*01:01:02:01 by single mutations). The fifth lineage, HG010103, presents the same promoter and 3’ UTR sequences and coding
alleles that share the A → T mutation at the position +748 at exon 3 (G*01:01:03:01, G*01:01:03:02, and G*01:01:05). The last lineage, HG010108, was found in two individuals and it presents the promoter of the HG010102 lineage, the G*01:01:08 coding allele and the 3’ UTR haplotype of the HG0103 lineage. The sequence of each haplotype, the proposed consensus sequence for each HLA-G main lineage and its frequencies are illustrated in figure 3. One haplotype found (H27) did not fit the pattern of haplotype clusters observed in the network (fig. 2), and it was considered in

FIG. 1. LD between pairs of SNPs at the HLA-G locus. The image was generated in the Haplovie program using SNPs with frequency >2% (A) and >1% (B). Areas in dark red indicate strong LD (LOD ≥ 2, D’ = 1), shades of pink indicate moderate LD (LOD ≥ 2, D’ < 1), blue indicates weak LD (LOD < 2, D’ = 1), and white indicates no LD (LOD < 2, D’ < 1). The haplotype blocks were defined by the confidence intervals method implemented by Haplovie. D’ values different from 1.00 are represented inside the squares as percentages. LOD, log of the odds; D’, pairwise correlation between SNPs.
figure 3 and table 1 as the results of possible crossing-over events between haplotypes from different lineages. In addition, although haplotypes H07 and H09 were included in the sublineages HG010101a and HG010101b, respectively, they seem to be the results of crossing-over events between haplotypes from different HG010101 sublineages.

In table 2, it is evident that 13 haplotypes were found just once in the present series. However, some of these haplotypes were confirmed by cloning (H07 and H18) or were present in allele combinations that facilitate the phase inference, such as H25, that was found in a sample with homozygosis in all the variation sites except for the position -443. In addition, at least five haplotypes were compatible with rare alleles that were found in some populations and now in Brazil, such as G*01:01:09 and G*01:04:05, all five with sequences that are compatible with the ones described in IMGT. The remaining five haplotypes that occurred just once were found in samples carrying very common haplotypes such as H01 and H10, which facilitate the PHASE inference, which resulted in probabilities ranging from 98.0% to 99.8%. Given these evidences, we believe that these haplotype inferences are quite robust, even for these 13 haplotypes that occurred just once.

HLA-G Promoter Sequence Variation and Haplotype Structure

By taking together the known transcripts for the full-length HLA-G protein available in the Ensembl Project database (ENST00000360323, ENST00000428952, ENST00000437467, ENST00000383621, and ENST00000411831), the HLA-G 5’ UTR sequence (transcribed but not translated) vary between transcripts, most of them presenting a 5’ UTR of 24 nucleotides before the ATG in exon 1 (positions -25 to -1).

Given that, all the variation sites evaluated in the present manuscript in the promoter sequence must be considered as variations of a 5’ URR sequence (upstream regulatory, not transcribed).

In order to understand the variability and the haplotype structure of the 5’ URR of the HLA-G locus, the haplotypes of the promoter region were extracted from the 28 extended haplotypes illustrated in figure 3. Considering only the promoter region, we identified 25 variation sites [23 SNPs and 2 insertion/deletion (indel)] in the present sample (table 1), encompassing the 1100 promoter nucleotides evaluated as described above. All these 25 variation sites have been already described in previous reports (Tan et al. 2005; Rizzo et al. 2008; Berger et al. 2010). It is noteworthy to mention that, exception made for the SNP at the -443 position, all variation sites did present allele frequencies higher than 2% in the present series. The two indels consist of an insertion of an additional guanine in a series of five between the -542 and -538 positions and a deletion of an adenine in a series of eight adenines between the -537 and -530 positions (Hviid et al. 2006; Rizzo et al. 2008). To avoid conflict with previous data, we
used the nomenclature proposed by Rizzo and colleagues, in which the guanine insertion was considered at the \(-540\) position and the adenine deletion was considered at the \(-533\) position. The nucleotide diversity of the promoter region was 0.00692, almost the same as for the entire HLA-G locus (0.00643).

The haplotyping analysis using the PHASE method revealed the presence of 11 different HLA-G promoter haplotypes (considering all the variation sites and the first SNP of exon 1 at the \(+15\) position), with frequencies ranging from 0.5% to 32.0% (table 1) and a haplotype diversity of 0.816. Because these promoter haplotypes were similar to those described by Tan et al. (2005), we used the same nomenclature as proposed by these authors. As stated by Tan and colleagues, the 12 haplotypes seem to be arranged in four promoter lineages with very few variations inside each group of lineages. It is interesting to note that two major promoter lineages (PROMO-G010101 and PROMO-G010102) account for 77.5% of the promoter haplotypes, with a distribution of 32.0% for PROMO-G010102 and of 45.5% for PROMO-G010101. The haplotypes from these two promoter lineages differ from each other in at least 12 variation sites (table 1), most of them located near to or within transcription factor–binding sites (Tan et al. 2005; Moreau et al. 2009).

Two novel HLA-G promoter haplotypes were found, named PROMO-G010101f and PROMO-G0103e. PROMO-G010101f presents an adenine deletion at the \(-533\) position and PROMO-G0103e presents a guanine insertion at the \(-540\) position (table 1). Although we have not found known haplotypes such as PROMO-G0103b and G0103c, it is possible that our new promoter haplotype PROMO-G0103e is
Table 1. HLA-G Promoter Polymorphisms and Haplotypes in the Brazilian Population.

<table>
<thead>
<tr>
<th>Promoter Haplotype</th>
<th>HLA-G Promoter SNP Position</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROMO-G0104a</td>
<td>A G A A A C T G G T A A G A C G G G A A C A</td>
<td>0.135</td>
</tr>
<tr>
<td>PROMO-G0104b</td>
<td>A G A A A C T G G T A A G A C C A G G A A C A</td>
<td>0.005</td>
</tr>
<tr>
<td>PROMO-G010102a</td>
<td>A G G T A C T G G T A A G A C C A G G G A A C A</td>
<td>0.320</td>
</tr>
<tr>
<td>PROMO-G010101a</td>
<td>A G A A A C C T A G A G G A C A A C G G C G C G</td>
<td>0.225</td>
</tr>
<tr>
<td>PROMO-G010101b</td>
<td>A G A A A C C G T A G A G G A C A A C G G C G C G</td>
<td>0.070</td>
</tr>
<tr>
<td>PROMO-G010101c</td>
<td>A G A A A T C G T A G A G G A C A A C G G C G C G</td>
<td>0.045</td>
</tr>
<tr>
<td>PROMO-G010101d</td>
<td>A G A A A C C C T A G A G G A C A G C G G C G C G</td>
<td>0.045</td>
</tr>
<tr>
<td>PROMO-G010101f</td>
<td>A G A A A C C C T A G A G G C A A C G G C G C G</td>
<td>0.070</td>
</tr>
<tr>
<td>PROMO-G0103a</td>
<td>G G G A G C C T T A G A G I A C A A G G A A A G T G</td>
<td>0.020</td>
</tr>
<tr>
<td>PROMO-G0103d</td>
<td>G G G A G C C T T A G A G I A C A A A A A G T G</td>
<td>0.040</td>
</tr>
<tr>
<td>PROMO-G0103e</td>
<td>G G G A G C C T T A G A G I A C A A A A A G T G</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Note: Four different lineages of haplotypes were found, grouped by their similarity. Differences among the haplotypes are given in shades of gray. "I" denotes an insertion of a guanine at position 540. "-" denotes a deletion of adenine at position 533.

* The variation at the position +15 is in the coding sequence in exon 1 and must not be considered as 5' UTR. This SNP was used in this table to better characterize the haplotypes accordingly to Tan and Colleagues, 2005.

As mentioned above, the 3' UTR haplotypes were named according to previous nomenclature (Castelli et al. 2007). For the HLA-G expression control (Kouesse et al. 2010), five other SNPs were detected: +3003T/C, +3010C/G, +3027A/C, +3035C/G, and +3042/C/G. For the +3142C/G and +3187G/C SNPs, known to be relevant for the HLA-G expression control, only eight variations were detected in the 3' UTR, including the 14 bp insertion/deletion polymorphism and the +3142C/G and +3187G/C SNPs, known to be relevant for the HLA-G expression control. (Kouesse et al. 2010). Five other SNPs were detected: +3003T/C, +3010C/G, +3027A/C, +3035C/G, and +3042/C/G.
T, and \(3196 \text{ C/G}\). All the variation sites did present allele frequencies higher than 4.5%. The present haplotype analysis revealed the same haplotypes as described in the previous manuscript, named UTR-1 to UTR-8, with frequencies ranging from 0.5% for UTR-8 to 25.5% for UTR-1 (table 2). It is interesting to note that 52% of the 3\(\text{’UTR}\) haplotypes were represented by UTR-1 and UTR-2, both with a frequency of 26.0%. Both haplotypes differ in five of eight variation sites found at the 3\(\text{’UTR}\), a scenario that is reminiscent to the one observed for the promoter region. The nucleotide diversity of the 3\(\text{’UTR}\) is the highest one among the HLA-G regions evaluated (0.01073), about 14.2 times the diversity of the entire human genome (Sachidanandam et al. 2001), with a haplotype diversity of 0.818, higher than that observed for the promoter region.

### Relationship Between Promoter, Coding, and 3\(\text{’UTR}\) Haplotypes

In order to investigate the relationship between the 5\(\text{’URR}\), coding, and 3\(\text{’UTR}\) haplotypes found in the present series, the extended haplotypes described in figure 3 were stratified into promoter, coding, and 3\(\text{’UTR}\) haplotypes. The promoter haplotypes were named according to the data proposed in table 1. The coding HLA-G haplotypes were named according to the official HLA-G alleles described in the IMGT database. The 3\(\text{’UTR}\) haplotypes were named according to a previous manuscript evaluating this same region (Castelli et al. 2010). It should be emphasized that the HLA-G 3\(\text{’UTR}\) and coding region sequence variation and haplotype structure were thoroughly evaluated in previously published manuscripts. Table 2 illustrates the relationship between the promoter, coding, and 3\(\text{’UTR}\) haplotypes.

<table>
<thead>
<tr>
<th>Extended Haplotype</th>
<th>Frequency (2(n = 200))</th>
<th>Promoter Haplotype</th>
<th>Coding Haplotype</th>
<th>3(\text{’UTR}) Haplotype</th>
<th>HLA-G Lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>H01</td>
<td>0.210</td>
<td>PROMO-G010101a</td>
<td>01:01:01:01</td>
<td>UTR-1</td>
<td>HG010101a</td>
</tr>
<tr>
<td>H05</td>
<td>0.045</td>
<td>PROMO-G010101d</td>
<td>01:01:01:01</td>
<td>UTR-1</td>
<td>HG010101a</td>
</tr>
<tr>
<td>H03</td>
<td>0.065</td>
<td>PROMO-G010101f</td>
<td>01:01:01:04</td>
<td>UTR-6</td>
<td>HG010101b</td>
</tr>
<tr>
<td>H02</td>
<td>0.070</td>
<td>PROMO-G010101b</td>
<td>01:01:01:05</td>
<td>UTR-4</td>
<td></td>
</tr>
<tr>
<td>H04</td>
<td>0.045</td>
<td>PROMO-G010101c</td>
<td>01:01:01:05</td>
<td>UTR-4</td>
<td></td>
</tr>
<tr>
<td>H06</td>
<td>0.005</td>
<td>PROMO-G010101a</td>
<td>01:01:01:05</td>
<td>UTR-4</td>
<td></td>
</tr>
<tr>
<td>H08</td>
<td>0.005</td>
<td>PROMO-G010101a</td>
<td>01:01:01:09</td>
<td>UTR-4</td>
<td>HG010101c</td>
</tr>
<tr>
<td>H10</td>
<td>0.155</td>
<td>PROMO-G010102a</td>
<td>01:01:02:01</td>
<td>UTR-2</td>
<td></td>
</tr>
<tr>
<td>H11</td>
<td>0.060</td>
<td>PROMO-G010102a</td>
<td>01:06</td>
<td>UTR-2</td>
<td></td>
</tr>
<tr>
<td>H12</td>
<td>0.035</td>
<td>PROMO-G010102a</td>
<td>01:05N</td>
<td>UTR-2</td>
<td></td>
</tr>
<tr>
<td>H13</td>
<td>0.005</td>
<td>PROMO-G010102a</td>
<td>01:01:02:02</td>
<td>UTR-2</td>
<td></td>
</tr>
<tr>
<td>H14</td>
<td>0.005</td>
<td>PROMO-G010102a</td>
<td>01:09</td>
<td>UTR-2</td>
<td></td>
</tr>
<tr>
<td>H15</td>
<td>0.005</td>
<td>PROMO-G010102a</td>
<td>01:06</td>
<td>UTR-8</td>
<td>HG010102</td>
</tr>
<tr>
<td>H16</td>
<td>0.035</td>
<td>PROMO-G010102a</td>
<td>01:01:03:01</td>
<td>UTR-7</td>
<td></td>
</tr>
<tr>
<td>H17</td>
<td>0.005</td>
<td>PROMO-G010102a</td>
<td>01:01:01:05</td>
<td>UTR-7</td>
<td></td>
</tr>
<tr>
<td>H18</td>
<td>0.005</td>
<td>PROMO-G010102a</td>
<td>01:03:01:0b</td>
<td>UTR-7</td>
<td>HG010103</td>
</tr>
<tr>
<td>H19</td>
<td>0.040</td>
<td>PROMO-G0103d</td>
<td>01:03</td>
<td>UTR-5</td>
<td></td>
</tr>
<tr>
<td>H20</td>
<td>0.020</td>
<td>PROMO-G0103a</td>
<td>01:03</td>
<td>UTR-5</td>
<td></td>
</tr>
<tr>
<td>H21</td>
<td>0.020</td>
<td>PROMO-G0103e</td>
<td>01:03</td>
<td>UTR-5</td>
<td></td>
</tr>
<tr>
<td>H22</td>
<td>0.005</td>
<td>PROMO-G0103e</td>
<td>01:03b</td>
<td>UTR-5</td>
<td>HG0103</td>
</tr>
<tr>
<td>H23</td>
<td>0.085</td>
<td>PROMO-G0104a</td>
<td>01:04:01</td>
<td>UTR-3</td>
<td></td>
</tr>
<tr>
<td>H24</td>
<td>0.040</td>
<td>PROMO-G0104a</td>
<td>01:04:04</td>
<td>UTR-3</td>
<td></td>
</tr>
<tr>
<td>H25</td>
<td>0.005</td>
<td>PROMO-G0104b</td>
<td>01:04:01</td>
<td>UTR-3</td>
<td></td>
</tr>
<tr>
<td>H26</td>
<td>0.005</td>
<td>PROMO-G0104a</td>
<td>01:04:05</td>
<td>UTR-3</td>
<td>HG0104</td>
</tr>
<tr>
<td>H28</td>
<td>0.010</td>
<td>PROMO-G010102a</td>
<td>01:01:08</td>
<td>UTR-5</td>
<td>HG010108</td>
</tr>
<tr>
<td>H07</td>
<td>0.005</td>
<td>PROMO-G010101a</td>
<td>01:01:01:01</td>
<td>UTR-6</td>
<td>Possible crossing-over</td>
</tr>
<tr>
<td>H08</td>
<td>0.005</td>
<td>PROMO-G010101f</td>
<td>01:01:01:05</td>
<td>UTR-6</td>
<td>events between haplotypes from different lineages</td>
</tr>
</tbody>
</table>
alleles mainly represented by G*01:01:02:01, the null G*01:05N, G*01:06, and the rare G*01:09. G*01:09 is supposed to be originated by a single mutation in the G*0106 allele because both shared a thymine at the +1790 position and the G*01:05N and G*01:06 alleles are supposed to be derived from G*01:01:02:01 by single mutations. The same pattern is observed for the other HLA-G lineages, in which a very close promoter and coding haplotypes are observed, usually coding alleles that produce the same HLA-G protein (as in the HG0103 and HG0104 lineages) and a specific 3′ UTR haplotype.

Only 3 haplotypes of 28 did not fit this pattern (table 2; figs. 2 and 3). One example is H27, in which the pattern of the HG0104 lineage is observed for the promoter and coding haplotypes, but the pattern of the HG010102 lineage is observed for the 3′ UTR. These samples are probably cases of crossing-over among frequent HLA-G extended haplotypes (table 2).

Haplotyping Procedures
The present study used the PHASE method to infer haplotypes (Stephens et al. 2001), evaluating more stringent parameter values than those used by the method as ‘default’. We increased 10 times the interaction value and we performed a total of 12 independent runs using different seed values, obtaining the same results for 96 samples in all runs, and at least 10 runs indicated the same haplotype for the remaining four samples carrying rare alleles such as G*01:09. Directional sequencing analysis was also employed to verify some of the haplotype inferences in heterozygous individuals, confirming the PROMO-G010101f and PROMO-G0103a haplotypes. Moreover, several haplotypes were found in homozygosis in some individuals, including the PROMO-G010101a, PROMO-G0102a, and PROMO-G0104a as well as their extended haplotypes H01, H10, and H23 (table 2). The PROMO-G0104b haplotype was confirmed in a homozygous individual for all the SNPs evaluated except the −443 position (haplotypes H23 and H25, table 2). The PROMO-G010101d haplotype was confirmed by its presence in a homozygous individual for all the SNPs evaluated except the −483 position (H01 and H05 haplotypes, table 2). Haplotypes were inferred obtaining inferences with a mean probability of 0.9687 ± 0.1013 for the first six runs considering all samples (such average probability increases to 0.9858 ± 0.0494 if one does not consider the four samples with rare HLA-G alleles). These six runs did not include extra individuals with genotypes composed of the IMGT sequence of rare alleles, as performed in the six remaining runs. Additionally, the common extended haplotypes H01, H02, H03, H10, H11, and H28 and one possible crossing-over event (H07) was confirmed by cloning and sequencing. On this basis, it is reasonable to propose that the haplotype inference performed in the present study is quite robust and reliable.

Selection Acting on the HLA-G Locus
The Ewens–Watterson neutrality test was performed to investigate whether natural selection may exert significant pressure on the allele and genotype frequencies in the 5′ URR,

<table>
<thead>
<tr>
<th>Table 3. Neutrality Tests Performed on the HLA-G Locus.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Region</strong></td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>HLA-G promoter</td>
</tr>
<tr>
<td>HLA-G coding</td>
</tr>
<tr>
<td>HLA-G 3′ UTR</td>
</tr>
<tr>
<td>HLA-G gene</td>
</tr>
</tbody>
</table>

* In the software input files, all indels were converted into SNPs to prevent loss of data because DNAsp do not support indels to calculate Tajima’s D and Fu and Li’s F and D.
coding region, and 3’ UTR of the HLA-G locus (table 3). Only the 3’ UTR presented a statistically significant (P = 0.0095) negative normalized F value (observed homozygosity lower than expected homozygosity); however, the 5’ URR did present a negative normalized F value, exhibiting a low P-value (P = 0.1276). Considering the HLA-G extended haplotypes, which encompasses the 5’ URR, coding region, and 3’ UTR, the Ewens–Watterson neutrality test did reveal a nonsignificant negative normalized F value higher than the values obtained for each region separately (P = 0.5218). The second neutrality test employed, Tajima’s D test, which takes into account sequence diversity, resulted in the observation that all estimated D values were positive, three of them reaching significance (table 3). In fact, the 5’ URR, 3’ UTR, and the complete gene did present statistically significant positive estimated D values (P < 0.05, respectively). Taken together, the negative Ewens–Watterson’s normalized F values and positive Tajima’s D values reflect an excess of heterozygosity that may be due to balancing selection (Tan et al. 2005). The third and fourth neutrality tests used (Fu and Li’s F and Fu and Li’s D) are based on the frequency spectrum and positive values correspond to an excess of old mutations, which is compatible with balancing selection (Tan et al. 2005). In the present series, all the values for Fu and Li’s F were positive with statistically significant P-values for the 5’ URR, 3’ UTR and for all the regions evaluated together.

Discussion

In this study, we carried out an analysis of the genetic variability of the 5’ URR, coding, and 3’ UTR of the HLA-G locus in a randomly selected series of bone marrow donors from southeastern Brazil. Although several reports have demonstrated the importance of both 5’ URR and 3’ UTR for the expression profile of the HLA-G locus, a complete LD and haplotype evaluation among these 5’ URR, coding, and 3’ UTR variation sites has not been simultaneously performed, and the same pattern of variation observed in the 3’ UTR could be present in the promoter region or even in the entire gene. Because Brazilians represent one of the most heterogeneous populations in the world (Parra et al. 2003) and show the largest HLA-G variability already detected (Castelli, Mendes-Junior et al. 2007; Castelli, Mendes-Junior, Wiezel, et al. 2007; Castelli, Mendes-Junior, et al. 2008) among the populations studied at a high resolution level, including North India, Polish, German, and Chinese Han (Donadi et al. 2011), a high level of variability and haplotype diversity would be expected in a Brazilian sample.

The promoter haplotype frequency distribution in the Brazilian population resembles those described for both African American and European American populations (Tan et al. 2005), which is compatible with the Brazilian ancestry background (Parra et al. 2003). The association between promoter haplotypes and HLA-G alleles presented by Ober’s group was confirmed by the present data, exception made for the association of the PROMO-G010102a haplotype and the G*01:01:08 allele.

Although almost all promoter haplotypes found in the present study are shared by European Americans, African Americans, Chinese, Danes (Ober et al. 2003; Hviid et al. 2004; Tan et al. 2005), and Brazilians, some inconsistencies regarding the HLA-G promoter diversity and haplotypes were observed in recent studies. Berger et al. have addressed the variability of the HLA-G promoter region and compared the frequency of all the variation sites and haplotypes detected in European American (EA) women with recurrent spontaneous abortion and healthy EA fertile women (Berger et al. 2010). Although three of seven frequent promoter haplotypes found in Berger’s study are shared by the Brazilian population, most of the haplotypes with a frequency higher than 0.5% found in EA were not detected among Brazilians. In Berger’s study, the haplotypes were obtained using 12 variation sites and the three most frequent haplotypes found in EA were compatible with the PROMO-G010101, PROMO-G010102, PROMO-G0103, and PROMO-G0104 promoter lineages described here. Given the background of the Brazilian population, the larger number of haplotypes described by Berger et al. is quite unexpected and may be explained by the different sample sizes in the two studies and the large number of promoter region haplotypes exhibiting very low frequencies.

Another study addressed the variability of the HLA-G promoter region regarding its influence on the expression level of soluble HLA-G (Hviid et al. 2006), evaluating the 5’ URR between −762 and −400 nucleotides as well as the 14-bp polymorphism in 61 Caucasian healthy individuals, but only the frequent haplotypes were shared between samples.

Recently, Rizzo et al. have addressed the variability of the promoter and coding region in patients with systemic lupus erythematosus and healthy controls. These authors also used the PHASE method to infer promoter and coding haplotypes in 14-bp insertion homozygous individuals. With the exception of a promoter haplotype that is compatible with both PROMO-G0102 and PROMO-G0104 lineages, considering only 13 promoter SNPs, none of the haplotypes found in this particular Italian population is compatible with the ones observed in the present series, including the haplotypes associated with the guanine insertion at −540 and the adenine deletion at −533 (Rizzo et al. 2008).

The discrepancies observed between the above mentioned studies may reflect the features and the size of the populations used in each work (Hviid: 61 individuals (Hviid et al. 2006) and Rizzo: 36 individuals (Rizzo et al. 2008)), which may directly influence haplotyping accuracy (Bettencourt et al. 2008).
HLA-G lineage (table 2). The HG010108 lineage do not follow the same pattern observed in the remaining lineages because it has a promoter of the HG010102 or HG010103 lineage and a 3’ UTR of the HG0103 lineage, but a coding sequence that apparently do not have an origin in the HG010102 or HG010103 lineages. Curiously, the HG010102, HG010103, and HG010108 promoter is the most compatible with primate sequences (Tan et al. 2005). Moreover, a small sample of 3’ UTR sequence data from nonhuman primates (Castro et al. 2000) revealed that Old World Monkeys and Great Apes present exclusively UTR-5 and UTR-3, respectively, which are not so frequent in humans (Castelli et al. 2010). Given that the promoter and the 3’ UTR haplotypes of this HG010108 lineage are compatible with primates, it is possible that this haplotype is in fact much older than the other ones, resembling an ancestral haplotype that has been lost by genetic drift or selection. In the present series, three chromosomes could not be properly assigned to a specific lineage because they probably resulted from crossing-over between frequent haplotypes. One such example is represented by the H07 and H27 haplotypes (fig 3; table 2), which seem to be a product of crossing-over between haplotypes from the HG010101a and HG010101b lineages in the first case and from HG0104 and HG010102 in the second case (table 2).

The six HLA-G haplotype lineages do present functional variation mainly in their regulatory regions. Considering the promoter region (5’ URR), the minor allele frequency in 24 of 25 variation sites is higher than 2%, with a nucleotide diversity higher than the coding region (table 3) and an average of one variation site per 52 nucleotides. The 3’ UTR presents the minor allele frequency higher than 2% in all eight variation sites, as well as the highest nucleotide diversity, with an average of one variation site per 45 nucleotides. Although the coding region presented the lowest nucleotide diversity, with an average of one variation site per 62 nucleotides and with only 18 of 22 variation sites presenting minor allele frequencies higher than 2%, it did reveal the highest haplotype diversity. However, it should be emphasized that most of the polymorphic sites in the coding region are in fact synonymous substitutions. Therefore, it is plausible to propose that the regulatory regions (5’ URR and 3’ UTR) are more functionally variable than the coding region.

Genetic Diversity and Functional Aspects of HLA-G Extended Haplotypes

The mRNA level of a particular gene is usually regulated by its rate of synthesis, mainly driven by its 5’ URR, transcription factors that are produced and microenvironmental agents, as well as by the rate of mRNA decay, specially driven by the 3’ UTR influencing mRNA stability and degradation (Kuersten and Goodwin 2003). Although many factors may affect transcriptional and post-transcriptional control of HLA-G production (Moreau et al. 2009), the reasons for HLA-G expression in some but not in other tissues have not been fully elucidated. Several lines of evidence indicate that numerous nucleotide variations at 5’ URR and 3’ UTR of the HLA-G locus may influence HLA-G expression and consequently tissue distribution in physiological and pathological conditions. In addition, variation sites observed in introns may be involved in HLA-G regulation processes, such as alternative splicing.

The HLA-G 5’ URR is unique among the HLA genes (Solier et al. 2001). Due to the presence of a modified enhancer A (enhA) and a deleted interferon-stimulated response element (ISRE), the proximal HLA-G promoter is unresponsive to NF-kB (Gobin et al. 1998) and IFN-γ (Gobin et al. 1999). Among the regulatory elements known to stimulate HLA-G, a 244-bp region located −1.2 kb from exon 1 has been shown to be important for its spatiotemporal expression in transgenic mice and was proposed to have a locus control region (LCR) function (Schmidt et al. 1993; Yelavarthi et al. 1993). This LCR exhibits a binding site for CREB1 factor (−1380/−1370), which also binds to two additional cAMP response elements at −934 and −770 positions from the ATG. CREB1 allows promoter transactivation with the coactivators CREB-binding protein (CBP)/P300 (Gobin et al. 2002). In addition, a binding site (Interferon Sequence Responsive Element/ISRE) for IFN response factor-1 (IRF-1) is located at the −744 bp position (Lefebvre et al. 2001), beside a nonfunctional GAS-like element (−734) (Chu et al. 1999) and is involved in HLA-G transactivation following IFN-β treatment (Lefebvre et al. 2001). The HLA-G promoter also contains a heat shock element at the −459/−454 position that binds heat shock factor-1 (HSF-1) (Ibrahim et al. 2000) and a progesterone receptor binding site at −37 bp from ATG (Yie et al. 2006). On the other hand, the ras-responsive element binding factor 1 (RREB-1) downregulates HLA-G promoter activity through three ras response elements located at positions −1356, −142, and −53 (Flajollet et al. 2009) and is likely to act through C-terminal–binding protein (CtBP) implicated in chromatin remodeling (Shi et al. 2003).

Many of the promoter region polymorphisms (table 1; fig. 4) either coincide with or are close to known or putative regulatory elements and thus may affect the binding of HLA-G regulatory factors. For instance, this phenomenon was demonstrated by Ober’s group regarding the −725 G/C/T SNP, a variation site that is very close to ISRE, in which the −725G allele was associated with a significantly higher expression level compared with the others (Ober et al. 2006). In addition, besides the variation sites influencing the binding of transcription factors per se, the methylation status of the HLA-G gene promoter is crucial to the transcriptional activity of the gene (Moreau et al. 2003; Mouillot et al. 2005) and the promoter methylation status may also be affected by polymorphisms located at CpG sites (Ober et al. 2006).

The HLA-G 3’ UTR contains several regulatory elements (Kuersten and Goodwin 2003) including polyadenylation signals and AU-rich elements (Yie et al. 2008; Alvarez et al. 2009), and polymorphic sites that may potentially influence HLA-G transcription, translation, or both by several different mechanisms, particularly targets for microRNAs (miRNAs) (Castelli et al. 2009). Among them, it is worth mentioning the 14-bp polymorphism, which has been associated with the magnitude of HLA-G production (Rebmann et al. 2001), particularly
by modulating HLA-G mRNA stability (Hiby et al. 1999; O’Brien et al. 2001; Hviid et al. 2003; Rousseau et al. 2003). Although the mechanisms implicated have not been elucidated, HLA-G alleles presenting the 14-bp (5’-ATTGGTT-CATGCCCT-3’) sequence (Harrison et al. 1993) have been associated with lower mRNA production for most membrane bound and soluble isoforms in trophoblast samples (Hviid et al. 2003; Hviid 2006). On the other hand, a fraction of HLA-G mRNA transcripts presenting the 14-base insertion can be further processed (alternatively spliced) by the removal of 92 bases from the mature HLA-G mRNA (Hiby et al. 1999; Hviid et al. 2003), yielding smaller HLA-G transcripts, reported to be more stable than the complete mRNA forms (Rousseau et al. 2003). Besides the 14-bp polymorphism, two variation sites at the 3’ UTR have been reported to influence HLA-G expression. The presence of an adenine at the +3187 position, which is 4-bp upstream of an AU-rich motif, mediates mRNA degradation, leading to a decreased HLA-G expression (Yie et al. 2008). The presence of a guanine at the +3142 position may increase the affinity of the miR-148a, miR-148b, and miR-152 microRNAs for the HLA-G mRNA, therefore decreasing mRNA availability by mRNA degradation and translation suppression (Tan et al. 2007). In addition to the +3142 SNP, a recent in silico analysis revealed that several human miRNAs have the potential to bind to the HLA-G mRNA 3’ UTR and may influence HLA-G expression; however, the binding affinity might be influenced by polymorphisms present in their target regions, emphasizing the role of the 14-bp polymorphism and the +3003, +3010, +3027, and +3035 SNPs (fig. 3), which encompass a region of only 32 nucleotides (Castelli et al. 2009; Donadi et al. 2011). Alleles from these three major 3’ UTR polymorphic sites associated with HLA-G production are in strong LD with each other, illustrating a scenario in which their influence may not be mutually exclusive (Castelli et al. 2010). It is noteworthy that the 14-bp insertion is always accompanied by the +3142G and +3187A alleles (fig. 3), both previously associated with low mRNA availability, indicating that the low mRNA production associated with the 14-bp insertion (Hviid et al. 2003) may also be a consequence of the presence of these polymorphisms associated with the 14-bp polymorphism (Castelli et al. 2010). Besides, these three polymorphisms present a high LD with promoter variation sites (fig. 3 and table 2), which indicates that, in an in vivo scenario, given a proper microenvironment stimulus for HLA-G expression, the rate of transcription, and translation may be influenced both by promoter and by 3’ UTR polymorphisms, that is, each variation site exerts its influence in a coordinated and dependent manner.

**Genetic Diversity and Evolutionary Aspects of HLA-G Extended Haplotypes**

Given the immune tolerance property of HLA-G and its benign or harmful presence depending on the situation, the expression of this gene must be under very tight control (Donadi et al. 2011). It was previously shown that the pattern of variation at the HLA-G promoter region is characterized by two divergent lineages of promoter haplotypes that are maintained by balancing selection in worldwide human populations. These two divergent lineages may have different promoter activity and might be involved in a fine balance between high-expressing and low-expressing HLA-G haplotypes (Tan et al. 2005). Our data do corroborate the presence of these two main HLA-G lineages, in which the extended HLA-G lineages HG010102, HG010103, HG010108, and HG0104 (left part of fig. 2) correspond to the first one of Ober’s lineage, whereas extended lineages HG010101 and HG0103 correspond to the second Ober’s lineage (table 1).

Regarding the HLA-G locus as a whole, the two most frequent lineages, that is, sublineage HG010101a (H01 and H05) and lineage HG010102 (H10 to H15), which are equally frequent (about 26%) and account for more than 52% of the HLA-G haplotypes, are very different from each other. In fact, they differ in more than 50% of the 55 segregating sites analyzed, especially in the 5’ URR and 3’ UTR, with 11 and 4 fixed differences, respectively (fig. 4). Most of these nucleotide differences coincide with or are close to known or putative transcription factor-binding sites at the 5’ URR or are even present at the 3’ UTR sites that have been reported to influence HLA-G mRNA availability (14-bp polymorphism, +3142 and +3187 SNPs). Additionally, both lineages do present several differences in the coding region, but the same G*01:01 HLA-G protein is encoded in approximately 79.8% of cases.

---

**Fig. 4.** Nucleotide differences between the two most frequent HLA-G lineages, G010101a and HG010102.
The coding region suffers a strong selective pressure for invariance (purifying selection), that is, preservation of nucleotide and amino acid sequences, reduced variability, and lower than expected nonsynonymous mutation rate (Castro et al. 2000). In fact, strong evidence of purifying selection at the coding region was disclosed by the performance of a synonymous and nonsynonymous nucleotide substitution test considering all HLA-G alleles available at the IMGT database, which revealed an excess of synonymous changes in all exons (1—5), which is consistent with purifying selection (Mendes-Junior CT et al., unpublished data). Because HLA-G presents several biological effects related to the control of immune response (Lee et al. 1995; Diehl et al. 1996; Ishitani et al. 2003), one may expect that an invariable mechanism of maternal tolerance should be more effective, conferring a higher reproductive fitness. It is possible that such purifying selection would result from this tolerogenic feature of the HLA-G molecule. For example, LILRB1 and LILRB2 are receptors expressed on the surface of several leukocytes and they bind to the α3 domain of the HLA-G molecule. Given that LILRBs are considered to be the major HLA-G receptors, it is noteworthy that only one worldwide HLA-G allele do present a nonsynonymous polymorphic site at exon 4 (G*01:06), which codes the α3 domain of the molecule. One might expect that polymorphic residues observed in this domain may negatively influence LILRB interactions, modulating the inhibitory intracellular signaling. It is interesting to observe that the only frequent allele with a nonsynonymous mutation at exon 4 (G*01:06) has been associated with preeclampsia in several populations (Donadi et al. 2011).

In contrast to the coding region, the regulatory regions (5′ URR and 3′ UTR) suffer selection toward heterozygosis (balancing selection) (Aldrich et al. 2002; Tan et al. 2005; Mendes-Junior et al. 2007; Castelli et al. 2010). It could be possible that the balancing selection signature at the 3′ UTR region results from a hitchhiking effect of balancing selection at the 5′ URR. However, this scenario is not straightforward, given that an in silico study revealed that most of the polymorphic sites of the 3′ UTR region are miRNA-binding sites and their alleles presumably affect the miRNA-binding affinity (Castelli et al. 2009). These results suggest that these miRNAs might play a relevant role on the HLA-G expression pattern and, hence, the 3′ UTR region would be a direct target for balancing selection.

The coexistence of different selective pressures over a given gene would be favored by intragenic recombination. In fact, the existence of three recombining haplotypes in the present sample support the existence of crossing-over events throughout the HLA-G gene, particularly inside the HLA-G coding-region. It should be emphasized that different patterns of natural selection shaping variability of different parts of a same gene have been previously observed. Various class I and II MHC genes have provided evidence consistent with both balancing and purifying selection in humans, mice, and elephants. Although HLA class I antigen-binding sites are suffering overdominant selection, coding regions that are not involved in antigen presentation appear to have experienced purifying selection (Hughes and Nei 1989; Archie et al. 2010). Outside the MHC, there is strong evidence that balancing selection has shaped the pattern of variation of the 5′ URR region of the CCR5 gene (Bamshad et al. 2002; Ramalho et al. 2010), whereas its coding region has been subject to positive selection or neutral evolution (Sabeti et al. 2005).

In order to better evaluate which polymorphic sites are in fact driven by these balancing selection signatures, windows of 150 bp in the promoter, coding, and 3′ UTR regions were established and the Tajima’s D were calculated in each one of them. The promoter region did present three windows with significantly positive Tajima’s D values: one window with the variation −1306, another window with variations −762, −725, −616, −689, and −666, and a third window with the variation −201. The 3′ UTR region did present significantly positive Tajima’s D only in the window with the variations +3142, +3187, and +3196. The coding region presented two windows with significantly positive Tajima’s D values, one with the variations +15 and +36 at exon 1 and +99, +126, +130, and +147 at intron 1, and the other with the variation +372. Interestingly, these windows are not adjacent and may indicate polymorphisms with a greater functional relevance. For example, the position −1306 is in the LCR of the HLA-G gene and may influence the binding of several transcriptional factors, including the RREB-1. The positions −762, −725, and −716 are close to the ISRE motif present around position −744, a binding site for IRF-1. The position −201 is in the nonfunctional enhancer A, which may influence its functionality. In addition, positions −201, +15, and +36 are in complete LD with position −1306, which may also explain such high Tajima’s D value by a hitchhiking effect. However, the window with the polymorphic sites in intron 1 (+99 to +147) was in fact intriguing. It is not possible to evaluate the impact of such polymorphisms in the HLA-G function, unless they influence HLA-G splicing patterns, the mRNA secondary structure, or its stability. The polymorphic site +372 may be not relevant because it is a synonymous exchange, but it is in elevated LD with all promoter variations discussed above. The 3′ UTR region presented balancing selection signature only in the window with polymorphic sites that were evaluated regarding their functional relevance (+3142 influencing miRNA binding and +3187 influencing mRNA stability). Taking these evidences, we believe that indeed balancing selection may act primarily in the regulatory regions in the most functionally relevant polymorphic sites.

The HLA-G trend toward heterozygosity may assure a fine balance between high-expressing and low-expressing HLA-G haplotypes, that is, during pregnancy, high-expressing haplotypes would be favored in the absence of infection, whereas low-expressing haplotypes would be favored in the presence of infection. Apparently, the presence of both a high-expressing and a low-expressing haplotypes in an individual may have been an advantage during evolution, proning individuals to face situations when a high or low HLA-G expression is profitable. This is strongly reinforced by all the neutrality tests performed (Ewens–
Watterson, Tajima’s D and Fu and Li’s F and D), which revealed evidences of balancing selection acting only on the regulatory regions (5’ URR and 3’ UTR) and on the HLA-G locus as whole, probably by a hitchhiking effect due to the regulatory regions surrounding the coding region. These results may be, however, obscured by the population history (admixture) that characterizes this Brazilian urban population. Population stratification, for example, may add to selection, overestimating the balancing selection signatures obtained here. However, these same evidences have been found in other populations, such as Han Chinese, African American, and European American (Tan et al. 2005). Nevertheless, analyses of other worldwide autochthonous population samples should be carried out to reinforce this hypothesis.

Due to the observation of very divergent HLA-G lineages, as illustrated in figure 4, accounting for more than 52% of the HLA-G haplotypes, one may argue that the six main HLA-G lineages (HG010101 and their sublineages, HG010102, HG010103, HG010108, HG0103, and HG0104) as well as the very divergent HLA-G lineages (fig. 2, sides A and B) are also suffering balancing selection toward heterozygosis and that these very divergent HLA-G lineages might be associated with different HLA-G expression profiles. The Ewens–Watterson neutrality test was used to evaluate this matter, and three additional tests were performed considering 1) the main HLA-G lineages of each sample, with all the HG010101 lineages stratified into their sublineages and each crossing considered to involves different haplotypes; 2) the main HLA-G lineages of each sample, with all HG010101 lineages considered as a whole, the possible crossing-overs between sequences of the same lineage considered as an allele of this lineage and the possible crossing-overs between different lineages considered as different haplotypes; and 3) the side (A or B) of figure 2 in which each HLA-G haplotype of each sample was placed, in order to evaluate the heterozygosis between very divergent HLA-G lineages. All these tests revealed negative normalized F values, but only the last test (divergent lineages) did reveal a significant negative normalized F value (F = -1.9637, P = 0.0205). A closer evaluation of the HLA-G lineages from sides A and B (fig. 2) reveals that each side is composed exclusively by HLA-G haplotypes harboring one of the promoter lineages proposed by Tan et al. (2005). Therefore, although the inclusion of both the 3’ UTR and the coding regions enhance the resolution of the HLA-G network, the present results corroborate previous findings (Tan et al. 2005), that is, the existence of two highly divergent lineages of haplotypes (each one with sublineages) in diverse human populations, which may have very different transcriptional activity (determined by both the promoter and the 3’ UTR variability) and might result in precise and adequate protein levels.

In conclusion, the HLA-G locus seems to present six different HLA-G lineages showing functional variations mainly in nucleotides of the regulatory regions. These include differences in the 5’ URR at positions that either coincide with or are close to known transcription factor–binding sites and differences in the 3’ UTR mainly at positions that have already been reported to influence HLA-G mRNA stability and degradation rate. The evidence of balancing selection acting on the regulatory regions indicates that these HLA-G lineages are probably related to different expression profiles, depending on microenvironmental factors and on physiological or pathological conditions of the individual.

Acknowledgments
This study was supported by the Brazilian National Research Council (CNPq/Brazil—Grants 475670/2007-8 and 558476/2008-0) and the binational collaborative research program CAPES-COFECUB (project # 653/09). E.C.C. was supported by a postdoctoral fellowship (07/58420-4) from FAPESP/Brazil. L.C.V.C is supported by a postdoctoral fellowship (150175/2009-4) from CNPq/Brazil. M.R. is recipient of Research Scholar Award from the Fonds de la Recherche en Santé du Québec. We thank Marie-Claude Faucher, Karine Beauchemin, Sandra Rodrigues, and Flavia Tresmichin de Almeida for invaluable help. The authors declare no competing financial interests.

References


