Human Immunology 77 (2016) 841–853

Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

HLA-F coding and regulatory segments variability determined by massively parallel sequencing procedures in a Brazilian population sample

Thálitta Hetamaro Ayala Lima [a], Renato Vidal Buttura [a], Eduardo Antônio Donadi [b], Luciana Caricati Veiga-Castelli [b], Celso Teixeira Mendes-Junior [c], Erick C. Castelli [a,d,*]

[a] *Molecular Genetics and Bioinformatics Laboratory, Experimental Research Unit (UNIPEX), Sector 5, School of Medicine, Unesp – Univ. Estadual Paulista, Botucatu, State of São Paulo, Brazil*
[b] *School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, State of São Paulo, Brazil*
[c] *Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil*
[d] *Pathology Department, School of Medicine, Unesp – Univ. Estadual Paulista, Botucatu, State of São Paulo, Brazil*

A R T I C L E I N F O

Article history:
Received 11 February 2016
Revised 18 July 2016
Accepted 19 July 2016
Available online 20 July 2016

Keywords:
HLA-F
MHC
Variability
Polymorphisms
Haplotypes
Next Generation Sequencing
NGS

A B S T R A C T

Human Leucocyte Antigen F (*HLA-F*) is a non-classical HLA class I gene distinguished from its classical counterparts by low allelic polymorphism and distinctive expression patterns. Its exact function remains unknown. It is believed that *HLA-F* has tolerogenic and immune modulatory properties. Currently, there is little information regarding the *HLA-F* allelic variation among human populations and the available studies have evaluated only a fraction of the *HLA-F* gene segment and/or have searched for known alleles only. Here we present a strategy to evaluate the complete *HLA-F* variability including its 5′ upstream, coding and 3′ downstream segments by using massively parallel sequencing procedures. *HLA-F* variability was surveyed on 196 individuals from the Brazilian Southeast. The results indicate that the *HLA-F* gene is indeed conserved at the protein level, where thirty coding haplotypes or coding alleles were detected, encoding only four different HLA-F full-length protein molecules. Moreover, a same protein molecule is encoded by 82.45% of all coding alleles detected in this Brazilian population sample. However, the *HLA-F* nucleotide and haplotype variability is much higher than our current knowledge both in Brazilians and considering the *1000 Genomes Project* data. This protein conservation is probably a consequence of the key role of *HLA-F* in the immune system physiology.

1. Introduction

HLA-F is a non-classical gene of the Human Leukocyte Antigen (HLA) complex located within the Major Histocompatibility Complex (MHC). Most of the HLA genes are related to antigen presentation, mainly those known as classical HLA genes. HLA genes are considered the most polymorphic ones among vertebrates [1], but this variability is usually explored and reported within the classical HLA genes.

Although surrounded by the most variable human genes, current *HLA-F* knowledge suggests low allelic and protein polymorphism and distinctive expression patterns. While the pivotal role of classical class I HLA genes (such as *HLA-A*, *HLA-B* and *HLA-C)* in antigen presentation is well understood and documented, the importance and function of non-classical genes (such as *HLA-G*, *HLA-E* and *HLA-F*) are not completely understood. So far, most of the studies evaluated *HLA-G*, and several pieces of evidence point to its important role in immune tolerance [2–15].

The exact function of *HLA-F* remains unknown, but it is suggested that its encoded molecule has tolerogenic and immune modulatory features following the same pattern observed for other

Abbreviations: VCF, variant call format; μL, microliters (10^{-6} litro); DNA, deoxyribonucleic acid; GATK, Genome Analysis Toolkit; HLA, human leukocyte antigens; IMGT, international immunogenetics database; ILT2, immunoglobulin-like transcript 2; ILT4, immunoglobulin-like transcript 4; LD, linkage disequilibrium; MHC, Major Histocompatibility Complex; MAF, minor allele frequency; NGS, Next Generation Sequencing; NK, natural killer; UTR, untranslated region; °C, celsius degree; PCR, polymerase chain reaction; SNP, single nucleotide polymorphism; SP, São Paulo State, Brazil; STR, short tandem repeats; β2M, β2-microglobulin.

* Corresponding author at: Departamento de Patologia, Faculdade de Medicina, Unesp, Botucatu, SP CEP 18618970, Brazil.
E-mail address: castelli@fmb.unesp.br (E.C. Castelli).

http://dx.doi.org/10.1016/j.humimm.2016.07.231
0198-8859/© 2016 American Society for Histocompatibility and Immunogenetics. Published by Elsevier Inc. All rights reserved.

non-classical genes. This is based mostly on a few *HLA-F* features. Firstly, at the current time, *HLA-F* has a low variability, with only 22 different coding sequences described so far by the IPD-IMGT/HLA database, version 3.24.0 [16]. This same pattern of variation is observed for other non-classical HLA genes such as *HLA-E* and *HLA-G* [10,12,17–21], which have acknowledged roles on immune modulation [22–29]. Secondly, *HLA-F* (or *MHC-F* in primates) presents high sequence conservation among primates, suggesting an important role in cellular physiology, therefore supporting a critical role in the immune response [2,30,31]. Thirdly, analyses of the predicted amino acid sequences and their comparison among MHC-F molecules of different primates indicate that they present a typical HLA class I structure, i.e., a heavy chain (α domain) with five domains: two forming the peptide binding groove (α1 and α2), an immunoglobulin-like domain (α3), a transmembrane domain and a cytoplasmic domain [15,31,32]. This heavy domain may be associated with a beta-2-microglobulin (β2M) molecule. Fourthly, the interaction of HLA-F molecules with inhibitory receptors ILT2 and ILT4, usually at monocytes, was reported, suggesting that they might influence immune effector cells activation [33].

On the other hand, it is possible that *HLA-F* might influence HLA antigen presentation. HLA class I molecules lacking β2M or their associated peptides usually do not present the typical HLA class I molecule structure, being expressed as open conformers [34]. These atypical HLA molecules are unstable and do not remain on the cell surface. However, recent evidence indicates that the HLA-F molecule might bind to these HLA open conformers, and physical interaction appears to play a key role on the stabilization and transportation of these class I molecules in open conformation to the cell surface [31,32,35].

HLA-F expression was detected so far at B cells, T cell line HUT37 and lymphoid tissues such as thymus, spleen and tonsil [32,33,36,37]. Although this molecule apparently presents a typical HLA class I structure, it is not clear whether they are able to associate with peptides and present them on the cell surface [2,38], once this molecule was mainly found as an intracellular protein [15,32,33,36,39].

Although the HLA-F molecule seems to remain mainly in an intracellular environment, some studies found conflicting results when considering trophoblast cells. For instance, while HLA-F was formerly detected on the trophoblast cells surface [40], latter studies reported its expression only in the cytoplasm of decidual trophoblasts [41] or no HLA-F expression at all [42]. Whether this is an indication that HLA-F plays an immunomodulatory role on pregnancy or not remains to be investigated.

The HLA-F molecule is expressed in a range of tumors and can apparently influence the patient prognosis, but the findings are contradictory or lesion dependent. Its expression in gastric adenocarcinoma, for instance, is related to poor prognosis and an aggressive tumor behavior, probably related to the immunosuppression of anti-tumor cells [43]. Modifications in the *HLA-F* expression pattern in patients with breast cancer, esophageal squamous cell carcinoma and lung cancer is also related with patient prognosis, since overexpression of this molecule seems to be associated with poor outcomes [44–46]. Furthermore, the HLA-F molecule expression was associated with poor outcome in patients with hepatocellular carcinoma, and also with cell invasion and metastasis [47]. In addition, even low levels of HLA-F were associated with a worse prognosis in patients with neuroblastoma [48]. However, Zhang and colleagues didn't find any clinical significance for *HLA-F* gene expression in gastric cancer [49]. Polymorphisms at non-classical genes such as *HLA-G*, *HLA-E* and *HLA-F* have been associated with susceptibility to hepatitis B or hepatocellular carcinoma [50]. Finally, HLA-F expression was recently associated with Systemic Lupus Erythematosus disease activity, and patients with high levels of this molecule seems to present low disease activity [51].

Another feature of non-classical HLA genes that remains to be investigated for *HLA-F* is a possible role on transplantation. Several studies reported the influence of non-classical genes such as *HLA-G* and *HLA-E* on graft acceptance [23,29,52–55], but no study was conducted regarding *HLA-F* [56]. Notwithstanding that, due to the putative HLA-F tolerogenic and immune modulatory roles, following the same pattern observed for other non-classical genes, a possible role on pregnancy and transplantation should not be ruled out.

It has been considered that *HLA-F* presents few polymorphisms. As previously introduced, only 22 different sequences are presented on the IPD-IMGT/HLA database (version 3.24.0). Nevertheless, *HLA-F* variability has been poorly explored in worldwide populations. In fact, only one study sequenced the entire *HLA-F* segment in some cell lines [57], while the few remaining studies searched for variable sites that were already described at that moment [50,58–60], or evaluated only a partial *HLA-F* coding sequence using sequencing procedures [61]. Thus, it is not clear whether *HLA-F* truly presents low diversity, or if it is a bias driven by its poor exploration. Independently, considering the region in which *HLA-F* is located, i.e., the most variable in the human genome, this low variability might be a consequence of the putative key role of *HLA-F* on the regulation of immune responses.

Here we present a strategy to evaluate the variability of the complete *HLA-F* gene segment, including its regulatory sequences, by using massively parallel sequencing (or Next Generation Sequencing – NGS) procedures. This strategy was applied to evaluate the *HLA-F* variability on an urban Brazilian population sample from the State of São Paulo (Southeaster Brazil), which is characterized by high inter-ethnic admixture [62] and is considered a large repository of genetic variation.

2. Materials and methods

2.1. HLA-F gene structure definition

The annotations at the human genome draft versions hg19 or hg38 consider that the *HLA-F* gene presents at least three transcript variants, NM_018950.2, NM_001098478.1 and NM_001098479.1, indicating that exon 4 might be present or absent from transcripts and that two different segments may characterize the *HLA-F* 3′ untranslated region (UTR) (Fig. 1). When the IPD-IMGT/HLA *HLA-F* structure is taken into account, the *HLA-F* coding sequence (CDS) is compatible with NM_001098479.1 and NM_018950.2. In addition, it presents 300 nucleotides upstream the first translated ATG and about 305 nucleotides downstream the stop codon, with no information regarding where the *HLA-F* transcription starts and ends (Fig. 1). Nevertheless, the identified transcripts do suggest that at least two different segments might be considered as 3′UTR, the first with 118 nucleotides (NM_018950.2 and NM_001098478.1) and the second one, completely different, with 120 nucleotides (NM_001098479.1) (Fig. 1).

Since there is no consensus regarding the *HLA-F* exon and intron structure, here we chose to adopt the structure defined by the transcript variant NM_018950.2 (Fig. 1), which is the closest to the structure considered by the IPD-IMGT/HLA database [16]. All the variable sites detected were positioned considering the Adenine of the first translated ATG as nucleotide +1 (as used by the IPD-IMGT/HLA database). Because of the structure highlighted in NM_018950.2, the segment upstream nucleotide −124 was considered as the 5′ upstream segment, between position −124 and −1 as the 5′UTR segment, between +1 and +2944 as the coding region, between +2945 and +3063 (the transcription end point at NM_018950.2) as the 3′UTR, and all base pairs downstream position +3063 as the 3′ downstream segment (Fig. 1). It should be

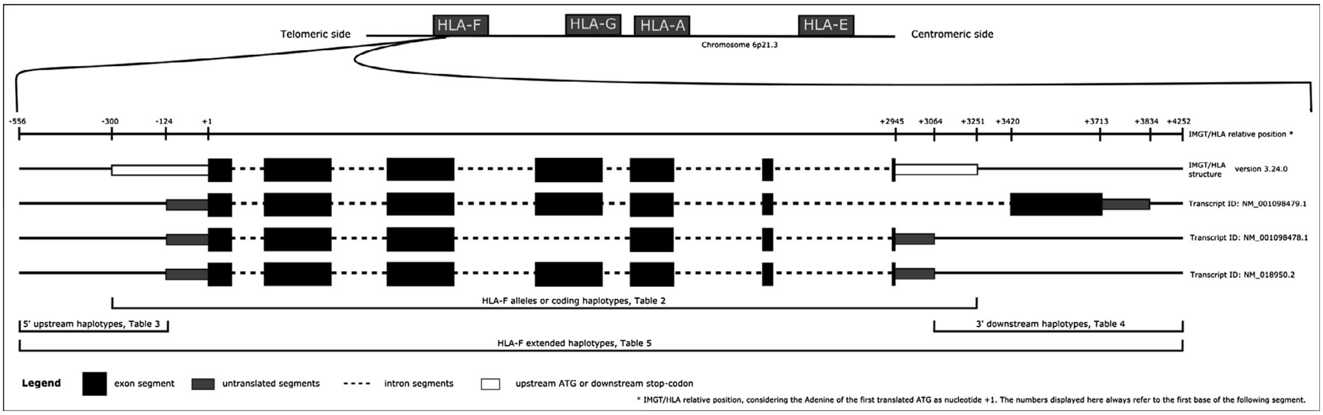
Telomeric side
HLA-F
HLA-G
HLA-A
HLA-E
Centromeric side
Chromosome 6p21.3
-556
-300
-124
+1
+2945
+3064
+3251
+3420
+3713
+3834
+4252
IMGT/HLA relative position *
IMGT/HLA structure version 3.24.0
Transcript ID: NM_001098479.1
Transcript ID: NM_001098478.1
Transcript ID: NM_018950.2
HLA-F alleles or coding haplotypes, Table 2
5' upstream haplotypes, Table 3
3' downstream haplotypes, Table 4
HLA-F extended haplotypes, Table 5
Legend
exon segment
untranslated segments
intron segments
upstream ATG or downstream stop-codon
* IMGT/HLA relative position, considering the Adenine of the first translated ATG as nucleotide +1. The numbers displayed here always refer to the first base of the following segment.

Fig. 1. *HLA-F* gene structure and transcripts.

emphasized that part of this 5′ upstream segment might represent both the proximal and distal *HLA-F* promoter segments, and also that part of the 3′ downstream segment is considered as 3′UTR by the *HLA-F* transcript variant NM_001098479.1.

2.2. DNA samples and amplification

HLA-F variability was evaluated in 196 non-related individuals from the state of São Paulo, Southeastern Brazil. All participants signed an informed consent before blood withdrawal, and this study protocol was reviewed and approved by the local Human Research Ethics Committee. Brazilian individuals from the state of São Paulo are highly admixed. For instance, the inter-ethnic admixture estimates for a population sample of white individuals of this same geographic region resulted in 79% European, 14% African, and 7% Amerindian contributions [63]. For Afro-Brazilians, estimates of interethnic admixture showed 62%, 26% and 12% of European, African and Amerindian contribution, respectively. For the mulattos, 37% and 63% of European and African contribution, respectively [64]. These proportions reinforce the admixed nature of sampled individuals. It should be emphasized that samples were randomly collected, regardless of the skin color and ancestry background of each individual.

The *HLA-F* amplification encompassed 5′ upstream, 5′UTR, complete coding region (including introns), 3′UTR and 3′ downstream segments. In general, an amplicon of approximately 6245 nucleotides was produced, from nucleotides 29,689,321 to 29,695,535 considering the sequence available for chromosome 6 (human genome assembly hg19). However, in order to avoid extremely repetitive regions, only the variability between nucleotides 29,690,685 (position −556) and 29,695,490 (position +4252) is reported (Fig. 1). Amplification was carried out using primers HFPR.F1 (5′-GGAGAGAACACTCAGGTGGC-3′) and HFUT.R1 (5′-CCACTAAA CACCCAGCCCAT-3′) in a final volume of 50 μL, containing 0.2 mM of dNTPs (Invitrogen – Carlsbad, CA USA), 15 pmol of each primer, 1.5 units of DNA polymerase (Long PCR Enzyme Mix, Thermo Fisher Scientific Inc., Waltham, MA, USA) and 0.8X the PCR buffer supplied with the DNA polymerase. Cycling conditions were performed as follows: 94 °C for 5 min, 10 cycles of 94 °C for 30 s, 63 °C for 45 s and 68 °C for 7 min, 22 cycles of 94 °C for 30 s, 61 °C for 45 s and 68 °C for 7 min. Amplification was evaluated on 1% agarose gel stained with GelRed® (Biotium™, Hayward, CA, USA), purified with ExoProStar (GE Healthcare Life Sciences), quantified using Qubit dsDNA High-Sensitivity (Thermo Fisher Scientific Inc., Waltham, MA, USA) assays and normalized to 0.2 ng/μL, which is the recommended concentration for sequencing library preparation.

2.3. Library preparation and sequencing

Amplicons were sequenced by massively parallel sequencing (or Next Generation Sequencing – NGS) procedures using the MiSeq Platform (Illumina, Inc., San Diego, CA USA). Sequencing libraries were prepared using Nextera XT Sample Preparation Kit and multiplexed with the Nextera XT Index Kit (both from Illumina, Inc.). Libraries were quantified by qPCR using Kapa (Kapa Biosystems, Wilmington, USA) and normalized to 4 nM, which is the recommended concentration. Fragmentation size was estimated using Bioanalyzer High Sensitivity DNA chips (Agilent Technologies, Santa Clara, CA, USA). Sequencing was performed using MiSeq Reagent Kit (V2, 500 cycles – Illumina, Inc.). The paired-end sequencing data was evaluated using freely available and locally developed software, as described below and detailed at the Supplementary File.

2.4. The use of Nextera kits for HLA class I gene sequencing

Although the use of Nextera kits is straightforward, we detected some difficulties regarding its use and bias in the template fragmentation pattern produced by them. This bias was mainly characterized by a sub representation of reads on certain GC-rich gene segments, typical of regulatory regions and some introns [65,66]. In the case of *HLA-F*, we detected such bias on two short segments, the first at the 5′ upstream segment around position −291, and the second at intron 2 around position +650. The coverage detected was calculated considering the number of reads covering a specific base. Although we aimed to achieve a minimum coverage of 500 for the entire segment, the coverage varied widely considering all samples for *HLA-F* gene and the segments mentioned above presented coverage reduction sometimes as low as 20 reads, even if the average coverage was higher than 1000. This appears to be a bias of enzymatic DNA fragmentation produced by the Nextera Kit [66]. Those segments presenting low coverage would be prone to genotyping errors, especially when a homozygous genotype is inferred at a position presenting low coverage. This issue has been dealt with *hla-mapper* and *vcfx*, and also by the imputation steps that were performed, as described further. Nevertheless, approaches such as this one should be considered when analyzing HLA genes by NGS and when using enzymatic fragmentation kits such as Nextera. It should be mentioned that we detected such bias for all HLA class I genes, including *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F* and *HLA-G*. For *HLA-F*, the use of vcfx introduced only 0.55% of missing alleles and the program PHASE [67] was used to impute them. This strategy is detailed further.

2.5. Raw data processing (mapping)

Prior to mapping, all short DNA segments produced by NGS, referenced to as 'reads', were trimmed on both ends for primer and adapter sequences. A major challenge when performing HLA sequencing by NGS procedures is to get a reliable mapping of the produced reads. This is particularly true when more than one HLA gene is sequenced together.

Because of the polymorphic and repetitive nature of most of the HLA genes due to its paralogous origins, several issues arise when raw data (the reads) are mapped to any reference human genome.

First, mainly because of the polymorphic nature of classical HLA genes, the sequences (or reads) obtained by NGS procedures usually present too many nucleotide differences when compared to any human genome draft. Thus, aligners (such as the well established Burrows-Wheeler Aligner – BWA) cannot correctly map HLA-derived sequences unless the acceptable mismatch level is increased. Usually, by using default parameters, many reads simply do not find a match in the reference genome leading to a mapping bias that underestimates HLA variability and overestimates reference allele frequencies [68]. On the other hand, increasing mismatch levels may eventually lead to misplacements of reads, as explained below.

Second, mainly because HLA genes are very similar to each other due to a common paralogous origin, the presence of differences that diminish the identity of a given HLA read with its corresponding haplotype sequence at the reference genome may increase its similarity with a homologous HLA gene. Thus when the acceptable mismatch level is increased to avoid the aforementioned mapping bias, this second issue leads to a great number of reads mapping to more than one HLA gene into the reference genome, or simply not mapping properly. In this scenario, an overestimation of genetic diversity in segments presenting a great number of mismapped reads may arise.

Third, depending on the NGS strategy used, a great number of short reads can be produced and short reads could reinforce the issues presented above and a larger amount of incorrectly mapped reads could be obtained.

All the issues introduced above would be circumvented if each HLA gene is tagged with different indexed adapters and mapped against a reference of this gene, but this is cost ineffectively. Usually, the ultimate goal in this field is to sequence all HLA genes (or most of them) from a given individual in a single sequencing run, tagging them with a single index, a strategy that would keep the possibility of using the limited number of available indexes to multiplex many samples.

Finally, although many companies have introduced different NGS HLA-typing kits and specific applications to deal with HLA typing, these products are mainly focused on reporting predefined HLA alleles, for clinical purposes. In this context, they are usually restricted to the segments tracked by the IPD-IMGT/HLA database [16], which does not include regulatory promoters and complete 3′ untranslated segments. In addition, available applications are expensive and usually work pretty well when the company's HLA typing kits are used, but they do not handle the data properly when homemade or other alternative solutions to characterize HLA genes are applied.

To overcome these matters and to achieve a better evaluation of HLA genes when (a) several HLA loci are sequenced together using NGS procedures, (b) when other genes outside the HLA complex are also included, or (c) even when homemade HLA sequencing solutions are applied, we developed a tool named *hla-mapper* to assist the read mapping procedure. *hla-mapper* is a tool created for research laboratories and it is not intended to be used for clinical purposes or for HLA typing, although its outputs might be embedded in pipelines for HLA allele typing procedures. Instead, it is a tool designed to help researchers to perform a reliable mapping of their HLA-related NGS data. Under the hood, it uses the power and convenience of the BWA aligner [69] and the Samtools package [70] to create a series of mapping and filtering procedures in order to separate reads for each of the supported HLA genes and provide a reliable map of those sequences, as described in the Supplementary Material (Fig. S1). The *hla-mapper* software is available at www.castelli-lab.net and detailed at the Supplementary File.

By using *hla-mapper*, reads were filtered to keep only pairs in which both sequences were larger than 80 nucleotides and presented at least 80% of bases with a minimum quality of 25 (Phred ⩾ 25), which is associated with a 99.7% accuracy. After the removal of small or low quality sequences, PCR duplicates, i.e., pairs with the same forward and reverse sequences, were removed. Then, *hla-mapper* addressed each sequence to its specific HLA gene (in this case, *HLA-F*). The final step of data processing by *hla-mapper* generates BAM files with all sequences mapped to the reference genome version hg19.

2.6. Genotype calling and processing

The Genome Analysis Toolkit (GATK, version 3.4) UnifiedGenotyper and HaplotypeCaller routines [71–73] were used to infer genotypes, and VCF (variant call format) files were generated considering all samples together, using the following parameters: dcov set to 1000 (only for UnifiedGenotyper), stand_call_conf set to 200, stand_emit_conf set to 50, and the chromosome 6 sequence (hg19) as reference. A mixed VCF file was then manually generated, in which most of the variable sites came from HaplotypeCaller, but some low frequency variants came from UnifiedGenotyper. This procedure was applied because HaplotypeCaller is better to infer genotypes, especially in regions presenting indels, but UnifiedGenotyper is better to detect low frequency variants (e.g., singletons).

Some level of mismapped reads is expected on HLA genes and might bias genotype inference, particularly in situations characterized by low coverage. This is particularly true when more than one HLA gene is sequenced at the same time, mainly because of the polymorphic nature and the high level of sequence similarity among HLA genes. This is the case of the present work, since it is part of a larger study in which all HLA class I genes were sequenced together.

Although this occurrence was minimized by *hla-mapper*, this must be considered when analyzing any HLA gene by NGS procedures. To circumvent these issues, the mixed VCF file generated by GATK was handled by *vcfx* (available at www.castelli-lab.net), which applied a series of rules in order to get a reliable genotype calling and assure that only high quality homozygous and heterozygous genotypes are passed forward to the phasing and imputation procedures to be applied afterwards, as detailed in the Supplementary File. After applying *vcfx*, the *HLA-F* variation data presented approximately 0.55% of missing alleles that were imputed latter by the PHASE algorithm [67], as further described. The PHASE algorithm [67] consists in a coalescent-based statistical method for haplotype reconstruction that deals with multi allelic *loci* and missing alleles, thus being suitable to manage HLA genotypes derived from long-range sequences.

The *HLA-F* gene presents some short tandem repeats (STR) along the regulatory and coding regions. One of these STRs lays on the *HLA-F* 3′UTR, approximately at position +3097 within the segment that is tracked by the IPD-IMGT/HLA database [16]. The genotypes regarding such STR were inferred manually by taking into account the number of reads detected for each allele by the GATK UnifiedGenotyper routine and the rates of stutter formation for unquestionable heterozygous genotypes. Since it was observed that real alleles produce stutter amplification (with the subtraction of one or more repeat unities) that varies according to the size of the original allele (i.e., smaller alleles present less stutter than larger alleles) [74], putative heterozygous genotypes composed by alleles that differ by a single repeat unity were considered as true heterozygotes only when the number of reads of the smaller allele divided by the number of reads of the larger allele was greater than a previously established empirical value for stutter amplification generated by the larger allele.

The association between each variable site was inferred using GATK (routine ReadBackedPhasing) using a minimal Phase Quality Threshold of 1000, dcov set to 1000 and minimum base quality set to 25. This assures that only alleles from variable sites that are close enough to be present in a same fragment (same read) should be phased. Considering the low *HLA-F* polymorphism described so far and the fact that, sometimes, variable sites are quite distant from each other in a same sample, not all variable sites were straightforwardly phased. In the present series, 61.7% of the heterozygous sites were phased using GATK. The remaining 38.3% were phased using the PHASE algorithm [67], which also imputed the 0.55% of missing alleles.

To proceed with the PHASE algorithm analysis, all singletons were removed, i.e., variable sites detected in only one individual and in a heterozygous state. Then, the partially GATK-phased VCF file was converted into an input file for PHASE and an accessory file containing the known phases between variable sites. In some situations, blocks of known phase, but with unknown phase among them, were generated and the PHASE algorithm was used to infer phase between these blocks until a complete pair of haplotypes is defined. After the final PHASE run, it was verified that all inferred haplotype pairs were compatible with the known phases (the ones determined by GATK). The most probable haplotype pair of each sample was then converted into complete *HLA-F* sequences using the hg19 reference sequence as draft and replacing the correct nucleotide in each position. In addition, singletons that were properly phased by GATK were introduced at their proper sequences. By using a local BLAST server with a database containing all known *HLA-F* alleles described so far (files in Fasta format downloaded from IPD-IMGT/HLA database, version 3.24.0), the closest known *HLA-F* coding allele was defined for each haplotype. The aligned files (BAM format), phased genotypes and the *HLA-F* sequences for each individual are available for download at the www.castelli-lab.net website.

2.7. Statistical analysis

The frequency of each *HLA-F* haplotype was computed by the direct counting method and adherences of diplotype proportions to expectations under Hardy–Weinberg equilibrium were tested by the exact test of Guo and Thompson [75] using the ARLEQUIN 3.5.1.3 software [76,77]. ARLEQUIN was also used to calculate gene and nucleotide diversity. The LD pattern was evaluated by calculating *D*′, LOD scores, and LD plots were visualized using Haploview 4.2 [78], considering only variable sites with a minor allele frequency (MAF) of 1% and the haplotypes defined as addressed earlier. Data from the *1000 Genomes Project* [79] was downloaded directly from the project browser (http://browser.1000genomes.org/) and the complete sequences were obtained using *vcfx*.

3. Results

The strategy proposed in the previous section revealed the presence of 70 variable sites on the evaluated *HLA-F* segment and surrounding sequences considering the 196 samples from Brazil (Table 1). Of those, 62 variable sites (88.6%) detected in this Brazilian population sample have also been detected and reported by the *1000 Genomes Project*, that evaluated 2504 individuals from worldwide populations (Table 1).

We detected ten variable sites located upstream the first translated ATG, between segment −556 and −1. Of those, 5 variable sites presented a minor allele frequency (MAF) higher than 10% and 9 presented a MAF higher than 1%. In addition, considering the segment tracked by the IPD-IMGT/HLA database [16], which is indicated in shades of gray at Table 1, two variable sites detected here may be considered as new mutations (one of them even considering the *1000 Genomes Project* database) since they have not been detected and described in that database (Table 1).

Considering the segment between the first translated codon until the stop codon, 45 variable sites were detected (Table 1). Nine variable sites were found on exonic segments, 5 of them representing non-synonymous mutations. A non-synonymous mutation located at exon 2 (position +342) has been previously associated with the F∗01:04 allele. Two other non-synonymous mutations at exon 3 (positions +923 and +982) configure new variable sites according to the IPD-IMGT/HLA database and the 1000 Genomes Project. The two remaining non-synonymous mutations occur at exon 4 (positions +1570 and +1771), the former configuring a variable site previously detected by the *1000 Genome Project* but not described at the IPD-IMGT/HLA database, while the latter has been previously associated with alleles *F*∗*01:03:01:01* or *F*∗*01:03:01:02*.

Considering the segment downstream the stop codon, 15 variable sites were detected, most of them presenting high frequencies for the minor allele (Table 1). In addition, taking into account the *HLA-F* structure highlighted in NM_018950.2, no variable sites were detected at the *HLA-F* 3′UTR segment, and the variable sites indicated above encompass a segment downstream the *HLA-F* transcription end point. Finally, only five out of 70 variable sites did not fit the Hardy-Weinberg expectations (positions +1943, +2698, +2726, +3097 and +3942).

These 70 variable sites were arranged into 37 different extended haplotypes. In order to characterize these extended haplotypes, we will present haplotypes of each *HLA-F* segment separately (Fig. 1), starting from the variable sites laying in the segment that is tracked by the IPD-IMGT/HLA database, i.e., from −300 to +3250 (Table 2), the segment upstream the transcription start site (5′ upstream), i.e., from −556 to −125 (Table 3), the 3′ downstream segment, i.e., after the transcription termination site (Table 4), and finally, the extended set of haplotypes (Table 5).

The haplotypes encompassing the segment tracked by the IPD-IMGT/HLA database were named according to the closest known *HLA-F* allele reported at the IPD-IMGT/HLA database version 3.24.0 followed by the positions concerning any differences eventually detected (Table 2). This segment was selected in order to allow us to name the haplotypes following the pattern proposed by the IPD-IMGT/HLA database and also to compare our data with information already reported in the aforementioned database.

Altogether, 30 different haplotypes (or *HLA-F* alleles) were detected in this Brazilian population sample. Of those, five were identical to known *HLA-F* alleles, including alleles *F*∗*01:01:01:09*, *F*∗*01:01:01:08*, *F*∗*01:01:01:01*, *F*∗*01:03:01:01* and *F*∗*01:01:01:11*, listed in order of frequency (Table 2). The remaining 25 coding haplotypes represents new *HLA-F* sequences or new *HLA-F* alleles determined by the combination of previously known mutations (at least by the *1000 Genomes Project*) that were associated with other known IPD-IMGT/HLA alleles. For instance, all five copies of the sequence carrying adenine at position +342 (rs17875380, Table 1), which is associated with the IPD-IMGT/HLA allele *F*∗*01:04*, also carried guanine at position +1943 (rs17184813) and Thymine at position +3189, which does not match the described IPD-IMGT/HLA sequence, resulting in a new *HLA-F* allele. This new *F*∗*01:04*-like allele [alternatively named as F∗01:01:01:09 (342A) at Tables 2 and 4] is also described by the *1000 Genomes*

Table 1
List of variable sites detected at the *HLA-F* gene considering the segment between nucleotide −556 and +4252, considering the Adenine of the first translated ATG as +1.

Chr6 Position	SNPid	*HLA-F* segment [a]	Notes[b]	IMGT/HLA position [c]	Reference allele [d]	Frequency (2*n*=392)	First alternative	Frequency (2*n*=392)	Second alternative	Frequency (2*n*=392)
29690694	rs3998799	5' Upstream	F	-547	C	0.6250	G	0.3750	-	-
29690715	rs114795943	5' Upstream	F	-526	T	0.9898	A	0.0102	-	-
29690741	rs17875377	5' Upstream	F	-500	G	0.9745	A	0.0255	-	-
29690838	rs17875378	5' Upstream	F	-403	G	0.9821	A	0.0179	-	-
29691003	rs56044823	5' Upstream	C,F	-238	C	0.9464	T	0.0536	-	-
29691019	rs1362126	5' Upstream	F	-222	G	0.6250	A	0.3750	-	-
29691090	rs1362125	5' Upstream	F	-151	T	0.5612	A	0.4388	-	-
29691097	rs2075682	5' Upstream	F	-144	A	0.8367	T	0.1633	-	-
29691140	rs2072896	5'UTR	F	-101	C	0.8367	G	0.1633	-	-
29691197	rs771810518	5'UTR	C,E	-44	A	0.9974	AT	0.0026	-	-
29691303	rs2076183	Exon 1	A,F	63	G	0.8367	A	0.1633	-	-
29691390	rs563604282	Intron 1	F	150	C	0.1607	G	0.8393	-	-
29691391	rs530577894	Intron 1	F	151	T	0.1607	C	0.8393	-	-
29691445	rs61739958	Exon 2	A,C,E,F	205	C	0.9974	A	0.0026	-	-
29691582	rs17875380	Exon 2	B,F	342	C	0.9872	A	0.0128	-	-
29691634	rs117540842	Exon 2	A,C,F	394	C	0.9923	G	0.0077	-	-
29691713	rs2076182	Intron 2	F	473	A	0.8367	C	0.1633	-	-
29691744	rs2076181	Intron 2	F	504	C	0.8367	G	0.1633	-	-
29691772	rs2076179	Intron 2	F	532	T	0.5612	C	0.4388	-	-
29691774	rs2076178	Intron 2	F	534	A	0.8010	C	0.1990	-	-
29691800	rs17875381	Intron 2	F	560	C	0.9490	G	0.0510	-	-
29691857	rs2072895	Intron 2	F	617	C	0.5612	G	0.4388	-	-
29691897	rs764147863	Intron 2	C,E	657	G	0.9974	C	0.0026	-	-
29692163	.	Exon 3	B,C,E	923	G	0.9974	A	0.0026	-	-
29692222	rs750557004	Exon 3	B,C,E	982	G	0.9974	T	0.0026	-	-
29692305	rs1632953	Intron 3	F	1065	A	0.1607	G	0.8393	-	-
29692334	rs9258186	Intron 3	F	1094	A	0.8367	G	0.1633	-	-
29692382	.	Intron 3	C,E	1142	G	0.9974	C	0.0026	-	-
29692429	rs140848774	Intron 3	F	1193-1204	AAATTTCTGAGGG	0.8367	A	0.1633	-	-
29692433	rs534575062	Intron 3	F	1193	T	0.3240	C	0.6760	-	-
29692462	rs142275427	Intron 3	F	1225-1239	TGGAATACCGATCCGC	0.9694	T	0.0306	-	-
29692465	rs368156595	Intron 3	C,D,F	1225	A	0.9974	C	0.0026	-	-
29692470	rs563176182	Intron 3	F	1230	C	0.8673	A	0.1327	-	-
29692553	rs193250743	Intron 3	C,F	1313	A	0.9949	G	0.0051	-	-
29692562	rs1736926	Intron 3	F	1322	A	0.1607	G	0.8393	-	-
29692618	rs144376525	Intron 3	C,F	1378	C	0.9949	T	0.0051	-	-
29692622	rs2072899	Intron 3	F	1382	C	0.3980	A	0.4388	G	0.1633
29692623	rs17178385	Intron 3	F	1383	C	0.9490	G	0.0510	-	-
29692634	rs1736925	Intron 3	F	1394	T	0.1607	G	0.8393	-	-
29692710	rs776718577	Intron 3	C,E	1470	A	0.9974	G	0.0026	-	-
29692729	rs2072898	Intron 3	F	1489	T	0.8367	G	0.1633	-	-
29692737	rs181298082	Intron 3	C,E,F	1497	T	0.9974	C	0.0026	-	-
29692810	rs142311857	Exon 4	B,C,D,F	1570	C	0.9974	A	0.0026	-	-
29693011	rs1736924	Exon 4	B,F	1771	C	0.1607	T	0.8393	-	-
29693019	rs201844059	Exon 4	A,C,F	1779	G	0.9949	A	0.0051	-	-
29693113	rs2076177	Intron 4	F	1873	C	0.8367	T	0.1633	-	-
29693183	rs17184813	Intron 4	F	1943	G	0.9566	A	0.0434	-	-
29693448	rs17875383	Intron 5	F	2208	C	0.9745	T	0.0255	-	-
29693499	rs2235383	Intron 5	F	2259	A	0.8367	G	0.1633	-	-
29693579	.	Intron 5	C,E	2339	C	0.9974	G	0.0026	-	-
29693726	rs114325231	Intron 5	C,F	2486	G	0.9464	T	0.0536	-	-
29693938	rs2735061	Intron 6	F	2698	G	0.8240	A	0.1760	-	-
29693969	rs1736922	Intron 6	F	2729	C	0.7628	T	0.2372	-	-
29694044	rs1736922	Intron 6	F	2811	T	0.8546	TG	0.1454	-	-
29694046	rs201715782	Intron 6	C,E,F	2805	G	0.9974	GA	0.0026	-	-
29694337 [e]	rs56321148	3' Downstream[e]	F	3097 [e]	$[TG]_{12}$	0.2602	$[TG]_{10}$	0.2372	$[TG]_{13}$	0.3878
29694427	rs1059174	3' Downstream	F	3189	C	0.1607	T	0.8393	-	-
29694443	rs62391801	3' Downstream	C,F	3205	A	0.9464	G	0.0536	-	-
29694526	.	3' Downstream	E	3288	C	0.9974	T	0.0026	-	-
29694570	rs3734813	3' Downstream	F	3332	A	0.8367	G	0.1633	-	-
29694680	rs3734814	3' Downstream	F	3442	A	0.8367	C	0.1633	-	-
29694681	rs3734815	3' Downstream	F	3443	A	0.8367	T	0.1633	-	-
29694777	rs17875384	3' Downstream	F	3539	G	0.9490	A	0.0510	-	-
29694832	rs111228498	3' Downstream	E,F	3594	C	0.9974	T	0.0026	-	-
29695176	rs371631016	3' Downstream	F	3938	TTA	0.9949	T	0.0051	-	-
29695180	rs1419696	3' Downstream	F	3942	G	0.7628	A	0.2372	-	-
29695199	rs143726784	3' Downstream	F	3961	A	0.9847	G	0.0153	-	-
29695305	rs2523405	3' Downstream	F	4067	T	0.5612	G	0.4388	-	-
29695412	rs17875386	3' Downstream	F	4174	C	0.9923	T	0.0077	-	-
29695466	rs2735060	3' Downstream	F	4228	T	0.5612	A	0.4388	-	-

The segment that is tracked by the IMGT/HLA database is indicated in gray.

[a] *HLA-F* segment. Variable sites were positioned following the Adenine of the first translated ATG as nucleotide +1. Because of the structure highlighted in NM_018950.2, variable sites upstream nucleotide −124 were considered as encompassing the 5′ upstream segment; between position −124 and −1 as encompassing the 5′UTR segment; between +1 and +2944 encompassing the coding region, with each exon or intron indicated; between +2945 and +3063 as encompassing the 3′UTR, and all variable sites further +3063 as encompassing the 3′ downstream segment (please refer to Fig. 1).

[b] NOTES: (A) Synonymous mutation on exon; (B) Non-synonymous mutation on exon; (C) New variable site on a region covered by the IPD-IMGT/HLA database, considering version 3.24.0; (D) Singleton with a defined haplotype; (E) Singleton with haplotype not inferred; (F) Variable site also detected by the *1000 Genomes Project*, Phase 3, considering all the 2504 individuals.

[c] IPD-IMGT/HLA relative position, considering the Adenine of the first translated ATG as nucleotide +1.

[d] The reference allele at the human genome draft version hg19.

[e] Variable site at position 29694337 (+3097). This variable site corresponds to an STR at position +3097. It presented five alternative alleles, differing from each other by the number of dinucleotide TG repeats. Only the two most frequent alternative alleles (besides de reference one at the hg19) are depicted at the table. Other alternative alleles are $[TG]_{11}$ with a frequency of 0.1020, $[TG]_{14}$ with a frequency of 0.0102 and only $[TG]_{9}$ with a frequency of 0.0026.

Table 2
List of *HLA-F* haplotypes found in Brazil, considering the whole segment that is present at the IPD-IMGT/HLA database.

HLA-F allele or haplotype[a]	Encoded molecule[b]	Brazilian samples frequency (2*n* = 392)	1000 Genomes samples frequency[c] (2*n* = 5008)
F∗01:01:01:09	F∗01:01	0.1735	0.1957
F∗01:01:01:08	F∗01:01	0.1709	0.0861
F∗01:01:01:01	F∗01:01	0.1454	0.1258
F∗01:03:01:01	F∗01:03	0.1020	0.0667
F∗01:03:01:01 (1383G)	F∗01:03	0.0459	0.0437
F∗01:01:01:01 ([TG]$_{12}$)	F∗01:01	0.0459	0.0667
F∗01:01:01:05 (1943G, [TG]$_{12}$)	F∗01:01	0.0434	0.0495
F∗01:01:01:11	F∗01:01	0.0408	0.0124
F∗01:01:02:03 (1943G, 2208C, [TG]$_{11}$)	F∗01:01	0.0408	0.0044
F∗01:01:02:01 (new-238T, 1230A, 1943G, new2486T, new3205G, [TG]$_{13}$)	F∗01:01	0.0383	0.0002
F∗01:01:02:04 (3189T)	F∗01:01	0.0306	0.0018
F∗01:01:02:03 (1943G, [TG]$_{11}$)	F∗01:01	0.0179	0.0010
F∗01:01:01:09 (342A) or F∗01:04 (1943G, 3189T)	F∗01:04	0.0128	0.0038
F∗01:01:01:08 (-222G, 2698G, [TG]$_{11}$)	F∗01:01	0.0128	0.0220
F∗01:01:02:02 (new-238T, 1943G, 2208C, new2486T, new3205G, [TG]$_{13}$)	F∗01:01	0.0128	–
F∗01:01:01:09 (new394G)	F∗01:01	0.0077	0.0114
F∗01:01:02:03 (1943G, [TG]$_{13}$)	F∗01:01	0.0077	0.0024
F∗01:01:01:05 (1943G, [TG]$_{14}$)	F∗01:01	0.0077	–
F∗01:01:02:03 (1943G, 2208C, [TG]$_{12}$)	F∗01:01	0.0077	0.0006
F∗01:03:01:01 (new1378T, 1383G, [TG]$_{11}$)	F∗01:03	0.0051	0.0026
F∗01:01:01:08 (new1779A)	F∗01:01	0.0051	0.0002
F∗01:01:01:01 (new1313G)	F∗01:01	0.0051	0.0002
F∗01:01:02:02 (new-238T, 1943G, 2208C, new2486T, new3205G)	F∗01:01	0.0026	–
F∗01:01:02:03 (new1570A, 1943G, 2208C, [TG]$_{11}$)	Unnamed	0.0026	–
F∗01:01:02:02 (1943G, 2208C, [TG]$_{11}$)	F∗01:01	0.0026	–
F∗01:03:01:01 ([TG]$_{11}$)	F∗01:03	0.0026	0.0006
F∗01:03:01:01 (new1225C)	F∗01:03	0.0026	0.0012
F∗01:01:01:01 ([TG]$_{14}$)	F∗01:01	0.0026	–
F∗01:03:01:01 ([TG]$_{13}$)	F∗01:03	0.0026	0.0004
F∗01:01:01:11 ([TG]$_{9}$)	F∗01:01	0.0026	–
Nucleotide diversity		0.00454 ± 0.00224	0.00557 ± 0.00272
Gene diversity		0.89540 ± 0.00650	0.91470 ± 0.00180

[a] Haplotype names were given according to the closest known *HLA-F* allele (IPD-IMGT/HLA) followed by the differences/divergences that were observed for this given haplotype. The segment that is considered by the IPD-IMGT/HLA database starts on nucleotide −300 up to nucleotide +3250. The word "new" refers to variable sites that are not currently described at the IMGT/HLA database, version 3.24.0. The total number of dinucleotide TG repeats on the *HLA-F* STR at position +3097 (IPD-IMGT/HLA) is indicated when different from the known *HLA-F* allele. The mutation at position +1570 generates a new HLA-F molecule, referred as "unnamed" because only the IPD-IMGT/HLA database can proper name this new allele and its consequent product. Haplotypes are listed in order of frequency.

[b] The encoded HLA-F molecule considering this haplotype.

[c] *1000 Genomes Project* samples, phase 3. There are 134 additional haplotypes not represented in the table, with a summed frequency of 0.3007.

Project. Other coding haplotypes followed this same pattern, such as *F∗01:01:01:05* and *F∗01:01:02:03*. New *HLA-F* alleles determined by variable sites that have not been described at the IPD-IMGT/HLA database (version 3.24.0) were also detected in this present series. They were placed at Table 2 with the "new" mutations indicated after the closest known *HLA-F* allele. Some of these haplotypes presented high frequencies, such as *F∗01:01:02:01* (new-238T, 1230A, 1943G, new2486T, new3205G, [TG]$_{13}$), carrying three new mutations and accounting for more than 3.8% of the haplotypes identified in this Brazilian population sample. In addition, at least 72% of the new *HLA-F* alleles detected here are identical to the ones described by the *1000 Genomes Project* considering data from phase 3 (Table 2). Among these new alleles determined by new variable sites, only one encodes an HLA-F molecule different from those already described (i.e., F∗01:01, F∗01:02, F∗01:03 and F∗01:04) and it was indicated as "unnamed" at Table 2 and detected in just one individual. Since most of the variable sites occur at intron segments or represent synonymous mutations, the *HLA-F* sequences detected in this Brazilian sample encode only four different HLA-F protein molecules, with molecule F∗01:01 presenting a frequency of about 82.45%, followed by F∗01:03 with 16.08% and F∗01:04 with 1.28%.

Considering the 5′ upstream segment evaluated (between nucleotides −556 and −125 as described earlier), eight haplotypes were detected (Table 3). These haplotypes were named sequentially according to their frequencies as F∗upstream-A to F∗upstream-H. Two haplotypes were quite frequent, the first one (F∗upstream-A) found with a frequency of 39.8%, and the second one (F∗upstream-B) with a frequency of 37.5%. All these haplotypes found in this Brazilian population sample were identical to the ones described by the *1000 Genomes Project* (phase 3).

The 10 haplotypes detected for the *HLA-F* 3′ downstream segment (variable sites after the last position tracked by the IPD-IMGT/HLA database, i.e., position +3250) are illustrated at Table 4. As for the 5′ upstream segment, haplotypes were named following their frequencies. Most of these 3′ downstream haplotypes were identical to haplotypes described by the *1000 Genomes Project*, phase 3.

It should be mentioned that we have detected 14 variable sites with the alternative allele in only one heterozygous individual, as singletons (Table 1). Of those, only two singletons had their phase with other variable sites inferred by the GATK and were included in the haplotype analysis (positions +1225 and +1570). The *HLA-F* segment between the first translated ATG and the stop codon presented 11 singletons, in which three represent non-synonymous mutations (positions +923, +982 and +1570), certainly determining three new *HLA-F* protein molecules. Thus, the *HLA-F* nucleotide variability at the coding region is probably higher than the one presented at Table 2.

The 5′ upstream, coding and 3′ downstream haplotypes were found arranged into 37 extended haplotypes as presented in Table 5. The frequencies of resultant diplotypes did fit the Hardy-Weinberg expectations ($P = 0.38858 \pm 0.01311$). Usually, each different *HLA-F* coding allele (or new *HLA-F* coding alleles derived

Table 3
List of haplotypes at the _HLA-F_ 5′ upstream segment considering the segment between positions −556 and −125 found in a Brazilian population sample.

Chr6 position [a]	29690694	29690715	29690741	29690838	29691003	29691019	29691090	29691097	Brazilian Samples Frequency	1000 Genomes Samples Frequency[c]
SNPid	rs3998799	rs114795943	rs17875377	rs17875378	rs56044823	rs1362126	rs1362125	rs2075682		
IMGT/HLA [b]	-547	-526	-500	-403	-238	-222	-151	-144	(2n = 392)	(2n = 5008)
F*upstream-A	C	T	G	G	C	G	T	A	0.3980	0.3728
F*upstream-B	G	T	G	G	C	A	A	A	0.3750	0.3125
F*upstream-C	C	T	G	G	C	G	T	T	0.0842	0.1254
F*upstream-D	C	T	G	G	C	G	A	A	0.0536	0.0579
F*upstream-E	C	T	G	G	T	G	T	T	0.0536	0.0755
F*upstream-F	C	T	A	A	C	G	T	T	0.0179	0.0102
F*upstream-G	C	A	G	G	C	G	A	A	0.0102	0.0222
F*upstream-H	C	T	A	G	C	G	T	T	0.0077	0.0196
Nucleotide diversity									0.00443±0.00280	0.00455±0.00285
Gene diversity									0.68950±0.01410	0.73770±0.00360

Minor alleles are marked in shades of gray. Since the _HLA-F_ 5′ upstream segment is not yet characterized, haplotypes were named as F∗upstream-A to -H, according to their frequencies in Brazil.

[a] Chromosome 6 position according to the human genome draft version hg19.

[b] IMGT/HLA relative position, considering the Adenine of the first translated ATG as nucleotide +1.

[c] _1000 Genomes Project_ samples, phase 3. There are seven additional haplotypes not represented in the table, with a summed frequency of 0.0039.

Table 4
List of haplotypes at the _HLA-F_ 3′downstream segment found in a Brazilian population sample.

Chr6 position [a]	29694570	29694680	29694681	29694777	29695176	29695180	29695199	29695305	29695412	29695466	Brazilian Samples Frequency	1000 Genomes Samples Frequency [c]
SNPid	rs3734813	rs3734814	rs3734815	rs17875384	rs371631016	rs1419696	rs143726784	rs2523405	rs17875386	rs2735060		
IMGT/HLA [b]	+3332	+3442	+3443	3539	+3938	+3942	+3961	+4067	+4174	+4228	(2n=392)	(2n = 5008)
F*downstream-A	A	A	A	G	TTA	G	A	**G**	C	**A**	0.4158	0.3632
F*downstream-B	A	A	A	G	TTA	**A**	A	T	C	T	0.2372	0.2532
F*downstream-C	**G**	**C**	**T**	G	TTA	G	A	T	C	T	0.1531	0.2115
F*downstream-D	A	A	A	G	TTA	G	A	T	C	T	0.1097	0.0705
F*downstream-E	A	A	A	**A**	TTA	G	A	T	C	T	0.0510	0.0497
F*downstream-F	A	A	A	G	TTA	G	**G**	**G**	C	**A**	0.0153	0.0200
F*downstream-G	**G**	**C**	**T**	G	TTA	G	A	T	**T**	T	0.0077	0.0164
F*downstream-H	A	A	A	G	**T**	G	A	**G**	C	**A**	0.0051	0.0056
F*downstream-I	**G**	**C**	**T**	G	TTA	G	A	T	C	**A**	0.0026	-
F*downstream-J	A	A	A	G	TTA	G	A	**G**	C	T	0.0026	-
Nucleotide diversity											0.00233±0.00141	0.00260±0.00154
Gene diversity											0.73430±0.01390	0.75120±0.00310

Minor alleles are marked in shades of gray. Since the _HLA-F_ 3′ downstream segment is not yet characterized, haplotypes were named as F∗downstream-A to -J, according to their frequencies.

[a] Chromosome 6 position according to the human genome draft version hg19.

[b] IMGT/HLA relative position, considering the Adenine of the first translated ATG as nucleotide +1.

[c] _1000 Genomes Project_ samples, phase 3. There are 21 additional haplotypes not represented in the table, with a summed frequency of 0.0099.

from them) is preferentially associated with the same 5′ upstream and 3′downstream haplotype, with rare exceptions. For instance, the coding allele _F∗01:01:01:01_ and new sequences that were probably derived from it are mainly associated with F∗upstream-B and F∗downstream-A; _F∗01:03:01:01_ and new sequences that were probably derived from it are mainly associated with F∗upstream-A and F∗downstream-D or –E; and the F∗01:04 $^{(1943G,\ 3189T)}$ [or F∗01:01:01:09 $^{(342A)}$] sequence detected in Brazil is associated with F∗upstream-A and F∗downstream-B. In fact, the entire _HLA-F_ segment presented a strong LD (Fig. 2), which justifies the pattern of association observed among the 5′ upstream, coding and 3′ downstream haplotypes (Table 5).

4. Discussion

NGS procedures are useful to characterize the variability of any DNA segment, but its use is particularly challenging when HLA genes are the main focus. This issue arises mainly due to the high level of sequence similarity among paralogous HLA genes. Also, the high level of polymorphisms, that characterizes most of the HLA genes, might bias the read mapping procedure mainly overestimating variants that are close to the reference used [68]. To circumvent this issue, we used _hla-mapper_, which applies a series of filters to assign each read to its proper gene, as described in the Supplementary Material.

The strategy used here to infer haplotypes took advantage of the straightforward phase observed in many NGS reads (evaluated by the GATK ReadBackedPhasing). In addition, it is also supported by probabilistic inferences provided by the PHASE algorithm. Both strategies were combined in order to produce reliable haplotypes, in which both fragment-inferred and probabilistic-inferred haplotypes were in agreement. Although the strategy used here is quite different from the one applied by the _1000 Genomes Project_ to acquire genotypes and haplotypes, in general, most of the haplotypes detected in this Brazilian population sample were also detected in the samples used by the aforementioned project.

To date, there are 22 coding _HLA-F_ alleles (or haplotypes) described by the IPD-IMGT/HLA database version 3.24.0. These alleles encode only 4 different HLA-F protein molecules, defined as HLA-F∗01:01, F∗01:02, F∗01:03 and F∗01:04. The overall

Table 5
HLA-F extended haplotypes for the region between nucleotide −556 and +4252, considering the Adenine of the first translated ATG as +1.

5′ upstream[a]	*HLA-F* allele [b]	3′ downstream [c]	Brazilian Samples (2*n* = 392)	1000 Genomes Samples [d] (2*n* = 5008)
F*upstream-B	F*01:01:01:01	F*downstream-A	0.1429	0.1252
F*upstream-B	F*01:01:01:01 ($[TG]_{12}$)	F*downstream-A	0.0332	0.0477
F*upstream-B	F*01:01:01:01 ($[TG]_{12}$)	F*downstream-F	0.0128	0.0190
F*upstream-B	F*01:01:01:01 (new1313G)	F*downstream-A	0.0051	0.0002
F*upstream-B	F*01:01:01:01 ($[TG]_{14}$)	F*downstream-A	0.0026	-
F*upstream-B	F*01:01:01:01	F*downstream-F	0.0026	-
F*upstream-D	F*01:01:01:05 (1943G, $[TG]_{12}$)	F*downstream-A	0.0434	0.0491
F*upstream-D	F*01:01:01:05 (1943G, $[TG]_{14}$)	F*downstream-A	0.0077	-
F*upstream-B	F*01:01:01:08	F*downstream-A	0.1684	0.0853
F*upstream-B	F*01:01:01:08 (new1779A)	F*downstream-A	0.0051	0.0002
F*upstream-G	F*01:01:01:08 (-222G, 2698G, $[TG]_{11}$)	F*downstream-A	0.0051	0.0158
F*upstream-G	F*01:01:01:08 (-222G, 2698G, $[TG]_{11}$)	F*downstream-H	0.0051	0.0054
F*upstream-B	F*01:01:01:08	F*downstream-J	0.0026	-
F*upstream-D	F*01:01:01:08 (-222G, 2698G, $[TG]_{11}$)	F*downstream-A	0.0026	-
F*upstream-E	F*01:01:02:01 (new-238T, 1230A, 1943G, new2486T, new3205G, $[TG]_{13}$)	F*downstream-C	0.0383	0.0002
F*upstream-E	F*01:01:02:02 (new-238T, 1943G, 2208C, new2486T, new3205G, $[TG]_{13}$)	F*downstream-C	0.0128	-
F*upstream-E	F*01:01:02:02 (new-238T, 1943G, 2208C, new2486T, 3205G)	F*downstream-C	0.0026	-
F*upstream-C	F*01:01:02:02 (1943G, 2208C, $[TG]_{11}$)	F*downstream-C	0.0026	-
F*upstream-C	F*01:01:02:03 (1943G, 2208C, $[TG]_{11}$)	F*downstream-C	0.0332	0.0024
F*upstream-F	F*01:01:02:03 (1943G, $[TG]_{11}$)	F*downstream-C	0.0179	0.0010
F*upstream-H	F*01:01:02:03 (1943G, $[TG]_{13}$)	F*downstream-C	0.0077	0.0024
F*upstream-C	F*01:01:02:03 (1943G, 2208C, $[TG]_{11}$)	F*downstream-G	0.0077	0.0020
F*upstream-C	F*01:01:02:03 (1943G, 2208C, $[TG]_{12}$)	F*downstream-C	0.0051	0.0006
F*upstream-C	F*01:01:02:03 (1943G, 2208C, $[TG]_{12}$)	F*downstream-I	0.0026	-
F*upstream-C	F*01:01:02:03 (new1570A, 1943G, 2208C, $[TG]_{11}$)	F*downstream-C	0.0026	-
F*upstream-C	F*01:01:02:04 (3189T)	F*downstream-C	0.0306	0.0018
F*upstream-A	F*01:01:01:09	F*downstream-B	0.1735	0.1951
F*upstream-A	F*01:01:01:09 (342A) or F*01:04 (1943G, 3189T)	F*downstream-B	0.0128	0.0038
F*upstream-A	F*01:01:01:09 (new394G)	F*downstream-B	0.0077	0.0114
F*upstream-A	F*01:01:01:11	F*downstream-B	0.0408	0.0124
F*upstream-A	F*01:01:01:11 ($[TG]_{9}$)	F*downstream-B	0.0026	-
F*upstream-A	F*01:03:01:01	F*downstream-D	0.1020	0.0667
F*upstream-A	F*01:03:01:01 ($[TG]_{11}$)	F*downstream-D	0.0026	0.0004
F*upstream-A	F*01:03:01:01 ($[TG]_{13}$)	F*downstream-D	0.0026	0.0004
F*upstream-A	F*01:03:01:01 (new1225C)	F*downstream-D	0.0026	0.0012
F*upstream-A	F*01:03:01:01 (1383G)	F*downstream-E	0.0459	0.0435
F*upstream-A	F*01:03:01:01 (new1378T, 1383G, $[TG]_{11}$)	F*downstream-E	0.0051	0.0026
Nucleotide diversity			0.00396 ±0.00194	0.00457±0.00223
Gene diversity			0.90150 ±0.00670	0.91820±0.00190

Haplotypes presenting the same *HLA-F* allele or new *HLA-F* alleles named following the same known *HLA-F* allele were grouped together (shades of gray) and ordered according to their frequencies.

[a] Upstream haplotypes, please refer to Table 3 for a detailed haplotype sequence.

[b] *HLA-F* allele based on known alleles from the IPD-IMGT/HLA database, version 3.24.0. Please refer to Table 2 for details.

[c] Downstream haplotypes, please refer to Table 4 for a detailed haplotype sequence.

[d] 1000 Genomes Project samples, phase 3. There are 158 additional haplotypes not represented in this table, with a summed frequency of 0.3042.

HLA-F variability has been minimally explored among human populations. Few studies have evaluated it, some of them using methodologies that detected only known variable sites [50,57–61]. Altogether, these studies indicated that only two HLA-F protein molecules are mainly detected, named F*01:01 and F*01:03, with F*01:01 corresponding for more than 90% of the HLA-F encoded molecules. These two molecules differ by the exchange of a Serine for a Proline at codon 251. Notwithstanding that, considering the limitations of previous studies, i.e., the search for previously known variable sites only, or the characterization of just a part of the *HLA-F* gene, the *HLA-F* variability might have been underestimated. Therefore, it was not clear whether *HLA-F* is really

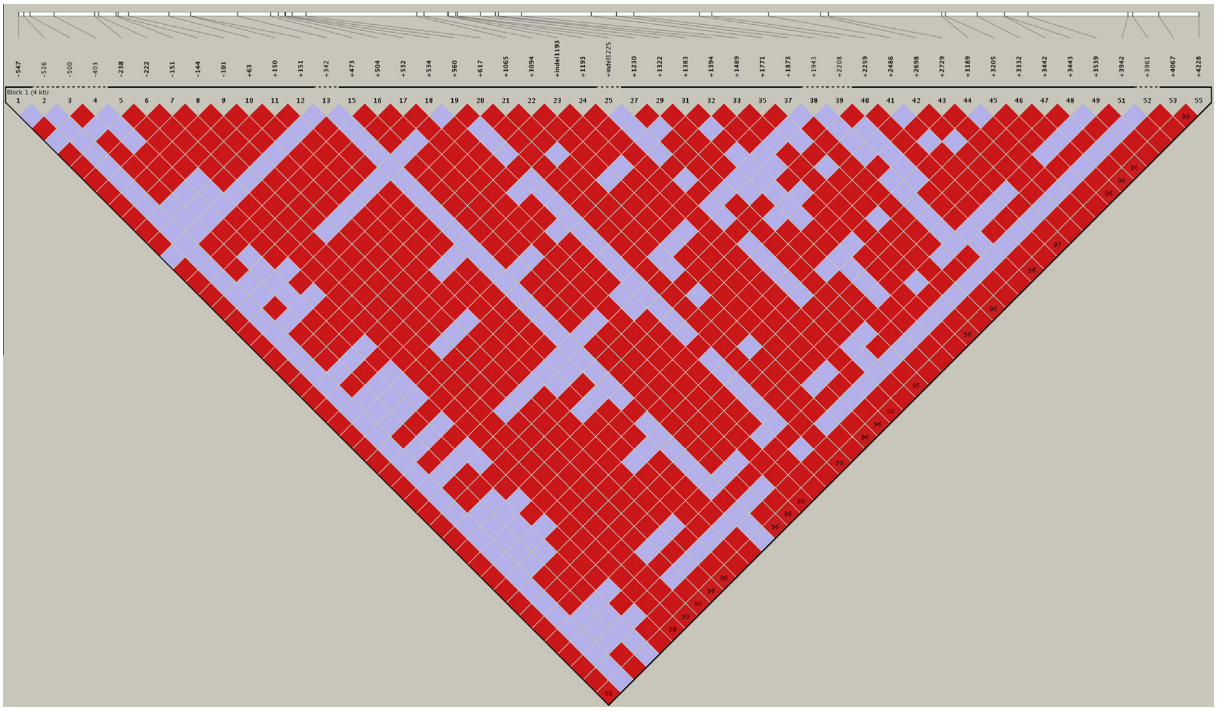

Fig. 2. Linkage Disequilibrium patterns considering variable sites at the *HLA-F* gene and surrounded sequences in a Brazilian population sample. LD plot generated by Haploview. Areas in dark red or dark gray indicate strong LD (LOD ⩾ 2, *D*′ = 1), shades of pink or shades of gray indicate moderate LD (LOD ⩾ 2, *D*′ < 1), blue or light gray indicates weak LD (LOD < 2, *D*′ = 1), and white indicates no LD (LOD < 2, *D*′ < 1). *D*′ values different from 1.00 are represented inside the squares as percentages. LOD, log of the odds; *D*′, pairwise correlation between single-nucleotide polymorphisms. The LD plot was generated using SNPs with a minimum allele frequency of 1%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conserved at the nucleotide and protein level, or if it has not been properly explored.

We evaluated the entire *HLA-F* variability by using a qualified methodology for the detection of every variable site present in the sample. In addition, the 5′ upstream and 3′ downstream regions were also evaluated. As observed in former studies performed in China and Brazil, the *HLA-F* gene is associated with low protein polymorphism. Considering the variable sites that were phased and included in the haplotype analysis (Tables 1 and 2), only four different HLA-F encoded molecules were found in the present sample, including F∗01:01 (82.45%), F∗01:03 (16.08%) and F∗01:04 (1.28%), following the same pattern observed in China [59] and in Southern Brazil [61]. We did not detect the *F∗01:02* allele in the present sample, corroborating the former *HLA-F* study in Brazil [61]. In addition, a new HLA-F molecule (due to the presence of a novel non-synonymous mutation) was also detected in a single individual. However, it should be mentioned that two non-synonymous mutations that occurred just once (singletons) were not included in the haplotype analysis (positions +923 and +982) excluding therefore two new HLA-F protein molecules.

Besides this protein conservation, we detected 70 variable sites comprehending the entire *HLA-F* gene. They were located between the 5′ upstream segment (from −547) and the 3′ downstream segment (up to +4228). These variable sites were arranged into 37 extended haplotypes, and 30 coding haplotypes considering the segment tracked by the IPD-IMGT/HLA database. These coding haplotypes improve the current knowledge depicted by the IMGT/HLA database version 3.24.0, since only five (16.6%) were identical to previously known *HLA-F* alleles (*F∗01:01:01:01*, *F∗01:01:01:08*, *F∗01:01:01:09, F∗01:01:01:11* and *F∗01:03:01:01*). Altogether, these five *HLA-F* alleles account for 63.3% of all the *HLA-F* sequences in this sample. Therefore, the *HLA-F* haplotype variability (or number of alleles) would be much higher than the last one described by the IPD-IMGT/HLA database version 3.24.0.

The 25 remaining coding haplotypes (or *HLA-F* alleles) did present nucleotide differences when compared to known IPD-IMGT/HLA haplotypes, and most of them were found at least twice in the present series. Moreover, they are also identical to the haplotypes described by the *1000 Genomes Project* phase 3. These haplotypes configure new *HLA-F* alleles, increasing the number of known alleles for this gene. Nevertheless, besides possible errors regarding genotype calling at the microsatellite (as described in the Methods section), the number of new *HLA-F* sequences reported here (25 new sequences, Table 2) is higher than the number of *HLA-F* sequences reported so far (22 known *HLA-F* sequences considering release 3.24.0). However, it is important to note that the existence of such new *HLA-F* alleles should be confirmed by cloning and Sanger sequencing, especially those that diverge only by the number of dinucleotide repeats in the STR at position +3097, and a proper name must be assigned for them by the Who Nomenclature Committee for Factors of the HLA System. Therefore, these new sequences described here were named considering the closest known allele followed by the differences detected.

Some of the new sequences detected were very similar to known *HLA-F* sequences, differing by few variable sites. One example is the sequence encoding the F∗01:04 molecule, which was different from the *F∗01:04* allele described in the IPD-IMGT/HLA database up to version 3.24.0. In all five cases, the *F∗01:04* allele detected in Brazil lacks the guanine deletion located at position +1943 and the cytosine at position +3189, as described by the IPD-IMGT/HLA database. In fact, the aforementioned deletion is associated with several *HLA-F* alleles as described by the IPD-IMGT/HLA database up to version 3.24.0, but we did not detect it in the present sample. Another example is the eighteen copies of the *F∗01:03:01:02* allele detected in Brazil, which are different from the IPD-IMGT/HLA sequence in at least two positions, including the lack of the guanine deletion at position +1943 and a different nucleotide at position +3189 (Table 2). In fact, these sequences are close to the one described for the F*01:03:01:01 allele, but with guanine at position +1383 (Table 2). These alternative sequences for *F*01:04* and *F*01:03:01:02* (as well as for other new *HLA-F* alleles) were also detected by the *1000 Genomes Project* (Table 2). It should be noted that the *1000 Genomes Project* (phase 3) did not detect this guanine deletion either, and we have manually evaluated all our samples to search for possible genotype errors at this position and no deletion was found, leading us to speculate whether this deletion really exists. Nevertheless, despite this increased number of *HLA-F* coding alleles, exception made for two sequences, these new alleles present synonymous mutations in exons or variable sites in introns, thus encoding the same F*01:01 or F*01:03 molecules as previously described. This protein conservation is probably a consequence of the key role of *HLA-F* in the immune system physiology.

The mechanisms underlying this lack of protein variability detected for *HLA-F* are not understood, mainly because *HLA-F* is located within the most variable region of the human genome. This phenomenon had already been observed for other non-classical class I genes such as *HLA-G* and *HLA-E* [14,17–20,80–82], which present tolerogenic and immune modulatory properties [22–25,27,83,84]. The *HLA-F* gene is also very conserved among primates evolution (including bonobo, gorilla, orangutan and chimpanzee), supporting the hypothesis of a critical role in the immune response [31]. It is possible that the MHC-F molecule (or HLA-F in humans) is involved together with the MHC-G and MHC-E (HLA-G and HLA-E in humans) in the regulation of T and NK cells activity [58]. According to the analyses of the predicted HLA-F amino acid sequence and its comparison with other primate MHC-F proteins, HLA-F has a typical classical class I structure, i.e., the presence of a heavy chain formed by the peptide binding (α1 and α2), immunoglobulin like (α3), transmembrane and cytoplasmic domains, associated with β2 microglobulin [32]. Finally, HLA-F has a restricted tissue expression pattern [33,38,41,85,86]. Because of these characteristics and the similarities with other non-classical genes such as *HLA-G* and *HLA-E*, it is believed that the HLA-F molecule has tolerogenic properties and a key role in immune modulation.

In addition, similar to what has been observed for the HLA-G 5′ upstream segment [20,87,88], *HLA-F* presents at least two highly frequent divergent 5′upstream haplotypes, with similar frequencies (F*upstream-A and F*upstream-B, Table 3). These haplotypes might be associated with differential expression patterns due to the presence of variable sites influencing the binding of transcription factors. However, different from what was observed for the aforementioned gene [20,81,89], in this Brazilian population sample, no variable site was detected on both *HLA-F* 3′UTR segments defined on Fig. 1.

Although the expression of HLA-F on grafted tissues, during pregnancy, and on pathological conditions such as tumors, are poorly explored, we cannot rule out a *HLA-F* role in the outcome of such situations, due to its putative tolerogenic and immune modulatory properties, following the same pattern observed for other non-classical genes.

To the best of our knowledge, this is the first study to fully characterize the *HLA-F* gene variability considering both regulatory segments and the entire coding sequence in Brazil, which is considered one of the most admixed populations in the world and a significant repository of genetic variation [63]. Protein conservation could be considered a requirement for all proposed *HLA-F* functions previously addressed. For example, high levels of protein diversity would impair HLA-F capacity to bind to the respective ligands, since all of them also present conserved regions that would bind to HLA-F. In addition, protein conservation may enhance the HLA-F capacity to bind to class I molecules presenting an open conformation. This same rationale applies to the HLA-F binding to class I molecules for recycling purpose, or even to its interaction with inhibitory receptors such as ILT2 and ILT4. Whatever the HLA-F function is, protein conservation observed in the present study supports the hypothesis that this function might be critical to the immune system and cellular physiology.

Conflict of Interests

None.

Acknowledgements

This work was supported by grants obtained from the São Paulo Research Foundation (FAPESP/Brazil – grant # 2013/17084-2), and from CNPq/Brazil (grants # 304471/2013-5, 304931/2014-4 and 309572/2014-2).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.humimm.2016.07.231.

References

[1] D.E. Geraghty, B.H. Koller, J.A. Hansen, H.T. Orr, The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments, J. Immunol. 149 (1992) 1934.

[2] R. Rojo, M.J. Castro, J. Martinez-Laso, J.I. Serrano-Vela, P. Morales, J. Moscoso, et al., MHC-F DNA sequences in bonobo, gorilla and orangutan, Tissue Antigens 66 (2005) 277.

[3] P.J. Bjorkman, P. Parham, Structure, function, and diversity of class I major histocompatibility complex molecules, Annu. Rev. Biochem. 59 (1990) 253.

[4] E.D. Carosella, B. Favier, N. Rouas-Freiss, P. Moreau, J. Lemaoult, Beyond the increasing complexity of the immunomodulatory HLA-G molecule, Blood 111 (2008) 4862.

[5] M. Cao, S.M. Yie, J. Liu, S.R. Ye, D. Xia, E. Gao, Plasma soluble HLA-G is a potential biomarker for diagnosis of colorectal, gastric, esophageal and lung cancer, Tissue Antigens 78 (2011) 120.

[6] K. Dunker, G. Schlaf, J. Bukur, W.W. Altermann, D. Handke, B. Seliger, Expression and regulation of non-classical HLA-G in renal cell carcinoma, Tissue Antigens 72 (2008) 137.

[7] T. Twito, J. Joseph, A. Mociornita, V. Rao, H. Ross, D.H. Delgado, The 14-bp deletion in the HLA-G gene indicates a low risk for acute cellular rejection in heart transplant recipients, J. Heart Lung Transplant. 30 (2011) 778.

[8] R. Brown, K. Kabani, J. Favaloro, S. Yang, P.J. Ho, J. Gibson, et al., CD86+ or HLA-G+ can be transferred via trogocytosis from myeloma cells to T cells and are associated with poor prognosis, Blood 120 (2012) 2055.

[9] N. Rouas-Freiss, R.M. Goncalves, C. Menier, J. Dausset, E.D. Carosella, Direct evidence to support the role of HLA-G in protecting the fetus from maternal uterine natural killer cytolysis, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 11520.

[10] T.V. Hviid, R. Rizzo, L. Melchiorri, M. Stignani, O.R. Baricordi, Polymorphism in the 5' upstream regulatory and 3' untranslated regions of the HLA-G gene in relation to soluble HLA-G and IL-10 expression, Hum. Immunol. 67 (2006) 53.

[11] T.V. Hviid, N. Milman, S. Hylenius, K. Jakobsen, M.S. Jensen, L.G. Larsen, HLA-G polymorphisms and HLA-G expression in sarcoidosis, Sarcoidosis Vasc. Diffuse Lung Dis. 23 (2006) 30.

[12] T.V. Hviid, S. Hylenius, C. Rorbye, L.G. Nielsen, HLA-G allelic variants are associated with differences in the HLA-G mRNA isoform profile and HLA-G mRNA levels, Immunogenetics 55 (2003) 63.

[13] E.A. Donadi, E.C. Castelli, A. Arnaiz-Villena, M. Roger, D. Rey, P. Moreau, Implications of the polymorphism of HLA-G on its function, regulation, evolution and disease association, Cell Mol. Life Sci. 68 (2011) 369.

[14] E.C. Castelli, J. Ramalho, I.O. Porto, T.H. Lima, L.P. Felicio, A. Sabbagh, et al., Insights into HLA-G genetics provided by worldwide haplotype diversity, Front. Immunol. 5 (2014) 476.

[15] C.A. O'Callaghan, J.I. Bell, Structure and function of the human MHC class Ib molecules HLA-E, HLA-F and HLA-G, Immunol. Rev. 163 (1998) 129.

[16] J. Robinson, J.A. Halliwell, J.D. Hayhurst, P. Flicek, P. Parham, S.G. Marsh, The IPD and IMGT/HLA database: allele variant databases, Nucleic Acids Res. 43 (2015) D423.

[17] L.P. Felicio, I.O. Porto, C.T. Mendes-Junior, L.C. Veiga-Castelli, K.E. Santos, R.P. Vianello-Brondani, et al., Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3' untranslated regions, Tissue Antigens 83 (2014) 82.

[18] L.C. Veiga-Castelli, E.C. Castelli, C.T. Mendes Jr., W.A. da Silva Jr., M.C. Faucher, K. Beauchemin, et al., Non-classical HLA-E gene variability in Brazilians: a nearly invariable locus surrounded by the most variable genes in the human genome, Tissue Antigens 79 (2012) 15.

[19] E.C. Castelli, C.T. Mendes-Junior, A. Sabbagh, I.O. Porto, A. Garcia, J. Ramalho, et al., HLA-E coding and 3' untranslated region variability determined by next-generation sequencing in two West-African population samples, Hum. Immunol. 76 (2015) 945.

[20] E.C. Castelli, C.T. Mendes-Junior, L.C. Veiga-Castelli, M. Roger, P. Moreau, E.A. Donadi, A comprehensive study of polymorphic sites along the HLA-G gene: implication for gene regulation and evolution, Mol. Biol. Evol. 28 (2011) 3069.

[21] A. Antoun, S. Jobson, M. Cook, P. Moss, D. Briggs, Ethnic variability in human leukocyte antigen-E haplotypes, Tissue Antigens 73 (2009) 39.

[22] S. Aractingi, N. Briand, C. Le Danff, M. Viguier, H. Bachelez, L. Michel, et al., HLA-G and NK receptor are expressed in psoriatic skin: a possible pathway for regulating infiltrating T cells?, Am J. Pathol. 159 (2001) 71.

[23] M. Biedron, J. Rybka, T. Wrobel, I. Prajs, R. Poreba, K. Kuliczkowski, The role of soluble HLA-G and HLA-G receptors in patients with hematological malignancies after allogeneic stem cell transplantation, Med. Oncol. 32 (2015) 664.

[24] C.V. Brenol, T.D. Veit, J.A. Chies, R.M. Xavier, The role of the HLA-G gene and molecule on the clinical expression of rheumatologic diseases, Rev. Bras. Rheumatol. 52 (2012) 82.

[25] E.D. Carosella, P. Moreau, J. Le Maoult, M. Le Discorde, J. Dausset, N. Rouas-Freiss, HLA-G molecules: from maternal-fetal tolerance to tissue acceptance, Adv. Immunol. 81 (2003) 199.

[26] G.K. da Silva, P. Vianna, T.D. Veit, S. Crovella, E. Catamo, E.A. Cordero, et al., Influence of HLA-G polymorphisms in human immunodeficiency virus infection and hepatitis C virus co-infection in Brazilian and Italian individuals, Infect. Genet. Evol. 21 (2014) 418.

[27] D.D. Dong, S.M. Yie, K. Li, F. Li, Y. Xu, G. Xu, et al., Importance of HLA-G expression and tumor infiltrating lymphocytes in molecular subtypes of breast cancer, Hum. Immunol. 73 (2012) 998.

[28] T.V. Hviid, HLA-G in human reproduction: aspects of genetics, function and pregnancy complications, Hum. Reprod. Update 12 (2006) 209.

[29] E. Hosseini, A.P. Schwarer, A. Jalali, M. Ghasemzadeh, The impact of HLA-E polymorphisms on relapse following allogeneic hematopoietic stem cell transplantation, Leuk. Res. 37 (2013) 516.

[30] N. Otting, R.E. Bontrop, Characterization of the rhesus macaque (Macaca mulatta) equivalent of HLA-F, Immunogenetics 38 (1993) 141.

[31] J.P. Goodridge, A. Burian, N. Lee, D.E. Geraghty, HLA-F complex without peptide binds to MHC class I protein in the open conformer form, J. Immunol. 184 (2010) 6199.

[32] S.D. Wainwright, P.A. Biro, C.H. Holmes, HLA-F is a predominantly empty, intracellular, TAP-associated MHC class Ib protein with a restricted expression pattern, J. Immunol. 164 (2000) 319.

[33] E.J. Lepin, J.M. Bastin, D.S. Allan, G. Roncador, V.M. Braud, D.Y. Mason, et al., Functional characterization of HLA-F and binding of HLA-F tetramers to ILT2 and ILT4 receptors, Eur. J. Immunol. 30 (2000) 3552.

[34] F.A. Arosa, S.G. Santos, S.J. Powis, Open conformers: the hidden face of MHC-I molecules, Trends Immunol. 28 (2007) 115.

[35] J.P. Goodridge, A. Burian, N. Lee, D.E. Geraghty, HLA-F and MHC class I open conformers are ligands for NK cell Ig-like receptors, J. Immunol. 191 (2013) 3553.

[36] D.S. Allan, E.J. Lepin, V.M. Braud, C.A. O'Callaghan, A.J. McMichael, Tetrameric complexes of HLA-E, HLA-F, and HLA-G, J. Immunol. Methods 268 (2002) 43.

[37] N. Lee, A. Ishitani, D.E. Geraghty, HLA-F is a surface marker on activated lymphocytes, Eur. J. Immunol. 40 (2010) 2308.

[38] P. Paul, N. Rouas-Freiss, P. Moreau, F.A. Cabestre, C. Menier, I. Khalil-Daher, et al., HLA-G, -E, -F preworkshop: tools and protocols for analysis of non-classical class I genes transcription and protein expression, Hum. Immunol. 61 (2000) 1177.

[39] V.M. Braud, D.S. Allan, A.J. McMichael, Functions of nonclassical MHC and non-MHC-encoded class I molecules, Curr. Opin. Immunol. 11 (1999) 100.

[40] A. Ishitani, N. Sageshima, N. Lee, N. Dorofeeva, K. Hatake, H. Marquardt, et al., 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 (2003) 1376.

[41] T. Shobu, N. Sageshima, H. Tokui, M. Omura, K. Saito, Y. Nagatsuka, et al., The surface expression of HLA-F on decidual trophoblasts increases from mid to term gestation, J. Reprod. Immunol. 72 (2006) 18.

[42] R. Apps, L. Gardner, A. Moffett, A critical look at HLA-G, Trends Immunol. 29 (2008) 313.

[43] S. Ishigami, T. Arigami, H. Okumura, Y. Uchikado, Y. Kita, H. Kurahara, et al., Human leukocyte antigen (HLA)-E and HLA-F expression in gastric cancer, Anticancer Res. 35 (2015) 2279.

[44] A. Lin, X. Zhang, Y.Y. Ruan, Q. Wang, W.J. Zhou, W.H. Yan, HLA-F expression is a prognostic factor in patients with non-small-cell lung cancer, Lung Cancer 74 (2011) 504.

[45] X. Zhang, A. Lin, J.G. Zhang, W.G. Bao, D.P. Xu, Y.Y. Ruan, et al., Alteration of HLA-F and HLA I antigen expression in the tumor is associated with survival in patients with esophageal squamous cell carcinoma, Int. J. Cancer 132 (2013) 82.

[46] A. Harada, S. Ishigami, Y. Kijima, A. Nakajo, T. Arigami, H. Kurahara, et al., Clinical implication of human leukocyte antigen (HLA)-F expression in breast cancer, Pathol. Int. 65 (2015) 569.

[47] Y. Xu, H. Han, F. Zhang, S. Lv, Z. Li, Z. Fang, Lesion human leukocyte antigen-F expression is associated with a poor prognosis in patients with hepatocellular carcinoma, Oncol. Lett. 9 (2015) 300.

[48] F. Morandi, G. Cangemi, S. Barco, L. Amoroso, M. Giuliano, A.R. Gigliotti, et al., Plasma levels of soluble HLA-E and HLA-F at diagnosis may predict overall survival of neuroblastoma patients, Biomed. Res. Int. 2013 (2013) 956878.

[49] J.G. Zhang, X. Zhang, A. Lin, W.H. Yan, Lesion HLA-F expression is irrelevant to prognosis for patients with gastric cancer, Hum. Immunol. 74 (2013) 828.

[50] J. Zhang, L. Pan, L. Chen, X. Feng, L. Zhou, S. Zheng, Non-classical MHC-Iota genes in chronic hepatitis B and hepatocellular carcinoma, Immunogenetics 64 (2012) 251.

[51] V. Jucaud, M.H. Ravindranath, P.I. Terasaki, L.E. Morales-Buenrostro, F. Hiepe, T. Rose, et al., Serum antibodies to HLA-E, HLA-F and HLA-G in patients with SLE during disease flares: clinical relevance of HLA-F autoantibodies, Clin. Exp. Immunol. (2015).

[52] E. Hosseini, A.P. Schwarer, M. Ghasemzadeh, Do human leukocyte antigen E polymorphisms influence graft-versus-leukemia after allogeneic hematopoietic stem cell transplantation?, Exp Hematol. 43 (2015) 149.

[53] J. Krongvorakul, S. Kantachuvesiri, A. Ingsathit, S. Rattanasiri, T. Mongkolsuk, P. Kitpoka, et al., Association of soluble human leukocyte antigen-G with acute tubular necrosis in kidney transplant recipients, Asian Pac. J. Allergy Immunol. 33 (2015) 117.

[54] G.I. Mossallam, R.A. Fattah, A. El-Haddad, H.K. Mahmoud, HLA-E polymorphism and clinical outcome after allogeneic hematopoietic stem cell transplantation in Egyptian patients, Hum. Immunol. 76 (2015) 161.

[55] J.C. Crispim, C.T. Mendes-Junior, I.J. Wastowski, G.M. Palomino, L.T. Saber, D.M. Rassi, et al., HLA polymorphisms as incidence factor in the progression to end-stage renal disease in Brazilian patients awaiting kidney transplant, Transpl. Proc. 40 (2008) 1333.

[56] M.A. Pabon, C.E. Navarro, J.C. Osorio, N. Gomez, J.P. Moreno, A.F. Donado, et al., Impact of human leukocyte antigen molecules E, F, and G on the outcome of transplantation, Transpl. Proc. 46 (2014) 2957.

[57] C.W. Pyo, L.M. Williams, Y. Moore, H. Hyodo, S.S. Li, L.P. Zhao, et al., HLA-E, HLA-F, and HLA-G polymorphism: genomic sequence defines haplotype structure and variation spanning the nonclassical class I genes, Immunogenetics 58 (2006) 241.

[58] J. Moscoso, J.I. Serrano-Vela, A. Arnaiz-Villena, MHC-F polymorphism and evolution, Tissue Antigens 69 (Suppl. 1) (2007) 136.

[59] F.H. Pan, X.X. Liu, W. Tian, Characterization of HLA-F polymorphism in four distinct populations in Mainland China, Int. J. Immunogenet. (2013).

[60] S.K. Kim, M.S. Hong, M.K. Shin, Y.K. Uhm, J.H. Chung, M.H. Lee, Promoter polymorphisms of the HLA-G gene, but not the HLA-E and HLA-F genes, is associated with non-segmental vitiligo patients in the Korean population, Arch. Dermatol. Res. 303 (2011) 679.

[61] L.F. Manvailer, P.F. Wowk, S.B. Mattar, J.S. da Siva, Bicalho M. da Graca, V.M. Roxo, HLA-F polymorphisms in a Euro-Brazilian population from Southern Brazil, Tissue Antigens 84 (2014) 554.

[62] F.C. Parra, R.C. Amado, J.R. Lambertucci, J. Rocha, C.M. Antunes, S.D. Pena, Color and genomic ancestry in Brazilians, Proc. Natl. Acad. Sci. U. S. A. 100 (2003) 177.

[63] L.B. Ferreira, C.T. Mendes-Junior, C.E. Wiezel, M.R. Luizon, A.L. Simoes, Genomic ancestry of a sample population from the state of Sao Paulo, Brazil, Am. J. Hum. Biol. 18 (2006) 702.

[64] Y.C. Muniz, L.B. Ferreira, C.T. Mendes-Junior, C.E. Wiezel, A.L. Simoes, Genomic ancestry in urban Afro-Brazilians, Ann. Hum. Biol. 35 (2008) 104.

[65] T. Shiina, S. Suzuki, Y. Ozaki, H. Taira, E. Kikkawa, A. Shigenari, et al., Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers, Tissue Antigens 80 (2012) 305.

[66] W. Wang, Z. Wei, T.W. Lam, J. Wang, Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions, Sci. Rep. 1 (2011) 55.

[67] M. Stephens, N.J. Smith, P. Donnelly, A new statistical method for haplotype reconstruction from population data, Am. J. Hum. Genet. 68 (2001) 978.

[68] D.Y. Brandt, V.R. Aguiar, B.D. Bitarello, K. Nunes, J. Goudet, D. Meyer, Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data, G3 (Bethesda) 5 (2015) 931.

[69] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754.

[70] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078.

[71] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491.

[72] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, et al., The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297.

[73] G.A. Van der Auwera, M.O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, et al., From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline, Curr. Protoc. Bioinformatics (2013). 11:11 10 1.

[74] C. Brookes, J.A. Bright, S. Harbison, J. Buckleton, Characterising stutter in forensic STR multiplexes, Forensic Sci. Int. Genet. 6 (2012) 58.

[75] S.W. Guo, E.A. Thompson, Performing the exact test of Hardy-Weinberg proportion for multiple alleles, Biometrics 48 (1992) 361.

[76] L. Excoffier, G. Laval, S. Schneider, Arlequin (version 3.0): an integrated software package for population genetics data analysis, Evol. Bioinform. Online 1 (2005) 47.

[77] L. Excoffier, H.E. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, Mol. Ecol. Resour. 10 (2010) 564.

[78] J.C. Barrett, B. Fry, J. Maller, M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps, Bioinformatics 21 (2005) 263.

[79] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, et al., An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56.

[80] C.T. Mendes-Junior, E.C. Castelli, D. Meyer, A.L. Simoes, E.A. Donadi, Genetic diversity of the HLA-G coding region in Amerindian populations from the Brazilian Amazon: a possible role of natural selection, Genes Immun. 14 (2013) 518.

[81] A. Sabbagh, P. Luisi, E.C. Castelli, L. Gineau, D. Courtin, J. Milet, et al., Worldwide genetic variation at the 3' untranslated region of the HLA-G gene: balancing selection influencing genetic diversity, Genes Immun. 15 (2014) 95.

[82] K. van der Ven, S. Skrablin, G. Engels, D. Krebs, HLA-G polymorphisms and allele frequencies in Caucasians, Hum. Immunol. 59 (1998) 302.

[83] H. Zheng, R. Lu, S. Xie, X. Wen, H. Wang, X. Gao, et al., Human leukocyte antigen-E alleles and expression in patients with serous ovarian cancer, Cancer Sci. 106 (2015) 522.

[84] B.R. Zanetti, D.F. Carvalho-Galano, N.L. Feitosa, M.K. Hassumi-Fukasawa, F.A. Miranda-Camargo, L.M. Maciel, et al., Differential expression of immune-modulatory molecule HLA-E in non-neoplastic and neoplastic lesions of the thyroid, Int. J. Immunopathol. Pharmacol. 26 (2013) 889.

[85] A. Ishitani, N. Sageshima, K. Hatake, The involvement of HLA-E and -F in pregnancy, J. Reprod. Immunol. 69 (2006) 101.

[86] D.E. Geraghty, X.H. Wei, H.T. Orr, B.H. Koller, 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 (1990) 1.

[87] Z. Tan, A.M. Shon, C. Ober, Evidence of balancing selection at the HLA-G promoter region, Hum. Mol. Genet. 14 (2005) 3619.

[88] L. Gineau, P. Luisi, E.C. Castelli, J. Milet, D. Courtin, N. Cagnin, et al., Balancing immunity and tolerance: genetic footprint of natural selection in the transcriptional regulatory region of HLA-G, Genes Immun. 16 (2015) 57.

[89] E.C. Castelli, C.T. Mendes-Junior, N.H. Deghaide, R.S. de Albuquerque, Y.C. Muniz, R.T. Simoes, et al., The genetic structure of 3'untranslated region of the HLA-G gene: polymorphisms and haplotypes, Genes Immun. 11 (2010) 134.