Received: 15 April 2020 | Revised: 18 June 2020 | Accepted: 9 July 2020

DOI: 10.1111/tan.13996

ORIGINAL ARTICLE

HLA Immune Response Genetics WILEY

Hla-C genetic diversity and evolutionary insights in two samples from Brazil and Benin

Andreia S. Souza[1,2] | Paulin Sonon[3] | Michelle A. Paz[1,4] | Léonidas Tokplonou[5,6,7] | Thálitta H. A. Lima[1,2] | Iane O. P. Porto[1,4] | Heloisa S. Andrade[1,2] | Nayane dos S. B. Silva[1,4] | Luciana C. Veiga-Castelli[8] | Maria Luiza G. Oliveira[8] | Ibrahim Abiodoun Sadissou[3] | Juliana Doblas Massaro[3] | Kabirou A. Moutairou[9] | Eduardo A. Donadi[10] | Achille Massougbodji[6] | André Garcia[5] | Moudachirou Ibikounlé[7] | Diogo Meyer[11] | Audrey Sabbagh[5] | Celso T. Mendes-Junior[12] | David Courtin[5] | Erick C. Castelli[1,2,4]

[1]Molecular Genetics and Bioinformatics Laboratory—Experimental Research Unity, School of Medicine, São Paulo State University (UNESP), Botucatu, São Paulo, Brazil

[2]Genetics Program, Institute of Biosciences of Botucatu, São Paulo State University (UNESP), Botucatu, São Paulo, Brazil

[3]Laboratório de Biologia Molecular, Programa de Imunologia Básica e Aplicada (IBA), Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

[4]Pathology Program, School of Medicine, São Paulo State University (UNESP), Botucatu, São Paulo, Brazil

[5]Institut de Recherche pour le Développement (IRD), UMR 261 MERIT, Université de Paris, Paris, France

[6]Centre d'Etude et de Recherche sur le Paludisme Associé à la Grossesse et à l'Enfance, Cotonou, Benin

[7]Département de Zoologie, Faculté des Sciences et Techniques, Université d'Abomey-Calavi, Cotonou, Benin

[8]Department of Genetics, School of Medicine of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

[9]Laboratoire de Biologie et Physiologie Cellulaire, Université d'Abomey-Calavi, Cotonou, Benin

[10]Department of Medicine, School of Medicine of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

[11]Department of Genetics and Evolutionary Biology, University of São Paulo (USP), São Paulo, Brazil

[12]Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

Correspondence
Erick C. Castelli, Departamento de Patologia, Faculdade de Medicina de Botucatu, Unesp—Botucatu, SP, CEP: 18618970, Brazil.
Email: erick.castelli@unesp.br

Funding information
Brazil-France Research Cooperation Program USP/COFECUB, Grant/Award Number: Uc Me 169-17; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: Finance Code 001; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Numbers: 2013/17084-2, 2017/19223-0; Institut de Recherche pour le Développement

Human leukocyte antigen-C (HLA-C) is a classical HLA class I molecule that binds and presents peptides to cytotoxic T lymphocytes in the cell surface. HLA-C has a dual function because it also interacts with Killer-cell immunoglobulin-like receptors (KIR) receptors expressed in natural killer and T cells, modulating their activity. The structure and diversity of the *HLA-C* regulatory regions, as well as the relationship among variants along the *HLA-C locus*, are poorly addressed, and few population-based studies explored the *HLA-C* variability in the entire gene in different population samples. Here we present a molecular and bioinformatics method to evaluate the entire *HLA-C* diversity, including regulatory sequences. Then, we applied this method to survey the *HLA-C* diversity in two population samples with different demographic histories, one highly admixed from Brazil with major European contribution, and one from Benin with major African contribution. The *HLA-C* promoter and 3′UTR

were very polymorphic with the presence of few, but highly divergent haplotypes. These segments also present conserved sequences that are shared among different primate species. Nucleotide diversity was higher in other segments rather than exons 2 and 3, particularly around exon 5 and the second half of the 3′UTR region. We detected evidence of balancing selection on the entire *HLA-C locus* and positive selection in the HLA-C leader peptide, for both populations. HLA-C motifs previously associated with KIR interaction and expression regulation are similar between both populations. Each allele group is associated with specific regulatory sequences, reflecting the high linkage disequilibrium along the entire *HLA-C locus* in both populations.

KEYWORDS

Beninese population, Brazilian population, *HLA-C*, natural selection, NGS, variability

1 | INTRODUCTION

The human leukocyte antigen-C (*HLA-C*) gene encodes an important molecule for antigen presentation and natural killer (NK) cell modulation. *HLA-C*, together with the other classical HLA class I genes, *HLA-A* and *HLA-B*, are among the most polymorphic human genes. *HLA-C* presents about 5709 alleles reported in the IPD-IMGT/HLA database, version 3.39, while *HLA-B* and *HLA-A* present 7126 and 5907 alleles, respectively. This variability is mainly associated with their function because classical HLA class I molecules bind diverse intracellular peptides and present them at the cell surface to T CD8+ cells. The HLA-peptide complex on the cell surface allows the recognition of self and nonself antigens by T lymphocytes, triggering an immune response against cells presenting foreign or abnormal peptides, such as virally infected or tumor cells.[1] Classical HLA class I molecules can also interact with activating and inhibitory killer-cell immunoglobulin-like receptors (KIR) of NK cells, modulating their activity.[2-5] However, *HLA-C* shows some peculiar features that make it an unusual classical HLA gene.

HLA-C presents lower cell surface expression levels when compared with *HLA-A* and *HLA-B* and, thus, it has been less associated with restricted CTL responses.[6] Nonetheless, the HLA-C expression levels have been shown to influence many clinical parameters, such as the HIV viral load,[7-10] the outcome of unrelated hematopoietic cell transplantation,[11,12] among others, suggesting that the expression levels might influence the HLA-C-restricted CTL response.

Polymorphism within *HLA-C* regulatory regions are known to influence HLA-C expression levels and disease phenotypes. For instance, a single-nucleotide polymorphism (SNP) located 35 kb upstream *HLA-C* (rs9264942) and a variant at the 3′ untranslated region (3′UTR, rs67384697) are associated with increased HLA-C expression levels and decreased HIV viral load.[7-10] The 3′UTR variant affects miR-148a binding and thus HLA-C expression. Moreover, it was also associated with a deleterious effect in psoriasis and Crohn's disease.[13,14] The influence on the binding of transcription factors has previously been described for some *HLA-C* promoter variants. It has been suggested that the variant rs2395471, located about 800 bp upstream of the transcription start site may influence the binding of Oct1,[15] while variants rs2524094 and rs10657191 could affect TNF-α and IFN-γ responses, respectively.[16] Some *HLA-C* allele groups, such as *HLA-C*03*, *-C*07*, and *-C*17*, present variations in one or more transcription factor binding sites at the core promoter.[17] These alleles also carry *rs2395471*G* (negatively influencing Oct1 binding) and an intact miR-148a binding site, both associated with lower HLA-C cell surface expression levels.[15,17]

HLA-C is the only classical HLA class I gene expressed at the maternal-fetal interface. Along with the nonclassical genes *HLA-G* and *HLA-E*, it plays a pivotal immunomodulatory role for placentation and pregnancy success.[18,19] Among classical loci, HLA-C is the most important in NK cell activity regulation because all HLA-C proteins bind to activating and inhibitory KIR ligands.[20] Considering that both *KIR* and *HLA-C* are highly polymorphic, pregnancy success may depend on the combination of fetal *HLA-C* and maternal *KIRs* [18,21,22]. Specific *KIR/HLA-C* combinations have also been associated with susceptibility to autoimmune and inflammatory diseases, outcomes in a series of infectious diseases (reviewed in References 23-25), and alloreactivity in hematopoietic stem cell transplantation (reviewed in Reference 25). Given the role of HLA-C in

immune defense and reproduction, selective pressures are likely to have shaped its genetic variability in different populations.[26,27]

High polymorphism levels across classical HLA *loci* are related to the presentation of a wide range of peptides within a population.[28] Nonetheless, HLA-C is also associated with the modulation of NK cells activity, such as performed by the highly conserved nonclassical HLA genes, *HLA-G* and *HLA-E*. Thus, it is not clear (i) how HLA-C presents both features considering its polymorphic nature and expression pattern, and (ii) how natural selection has shaped this gene's variability because the coding region of classical and nonclassical HLA genes reflect different selection profiles. Both *HLA-G* and *HLA-E* coding regions are under the influence of purifying selection,[29,30] while balancing selection acting on the segments encoding the peptide-binding groove of classical HLA molecules is well documented.[31-36]

The structure and diversity of the *HLA-C* regulatory regions, as well as the relationship among variants along the *HLA-C locus* (including the promoter and 3′ UTR) have been poorly addressed, and no population-based study explored the *HLA-C* haplotypic structure and relationship between variation in regulatory and coding regions. Furthermore, most of the known *HLA-C* alleles in the IPD-IMGT/HLA database lack the proximal promoter, 3′UTR, and intron sequences. Here we propose a second-generation sequencing and bioinformatic approach to evaluate the complete *HLA-C* variability encompassing at least 1500 bp of the promoter segment, the coding region, and the entire 3′UTR. We have applied this method to survey the genetic diversity and haplotype structure of two highly diverse population samples with different demographic histories: an interethnic admixed Brazilian sample with major European contribution and an autochthonous population from Benin, with major African ancestry. This comparison allowed us to discriminate whether the genetic and haplotype patterns observed in this study were a consequence of specific demographic histories or were due to shared evolutionary aspects. The bioinformatics workflow optimizes sequence mapping at *HLA-C* and infers haplotypes combining the phase information obtained directly from the sequencing data and also using probabilistic models. We found that the regulatory segments present few promoter and 3′UTR haplotypes, but these are highly divergent from each other, and that each coding allele group is associated with similar regulatory sequences. Our results suggest that while most exons of HLA-C show hallmarks of balancing selection, a different scenario arises when exon 1 (the leader peptide), the promoter, and initial 3′UTR segments are taken into account.

2 | MATERIALS AND METHODS

2.1 | Brazilian samples

We analyzed 418 unrelated samples from the state of São Paulo, Southeastern Brazil, from two different cities (Botucatu and Ribeirão Preto). Each individual signed an informed consent term before blood withdrawal. This study protocol was reviewed and approved by the Human Research Ethics Committee of the School of Medicine (UNESP/Brazil)—Protocol #24157413.7.0000.5411.

To estimate the ancestry of the population sample, we randomly selected 205 individuals and used Ancestry Informative Markers (AIMs). A panel with 34 AIMs[37] was used to estimate the contribution of European, African, and Amerindian/Asian ancestry for this sample. We used STRUCTURE v2.3.4 [38] to infer the population structure and individual ancestry rates, with $K = 3$. For parental samples, we used data from the 1000Genomes Project for European, African, and East Asian populations (404 Europeans from populations TSI, FIN, GBR, and IBS, 504 Africans from YRI, LWK, GWD, MSL, and ESN, and 504 East Asians from CHB, JPT, CHS, CDX, and KHV). Native American markers are a challenge to be detected because of the brief time of separation between this group and the East Asian, besides the admixture with the European group.[39] For this reason, the East Asian group was used as a representative for Amerindians.[40] Ancestry estimates for the general sample were 75.5% European, 16% African, and 8.5% Amerindian/Asian; however, individual ancestries ranged considerably, with individuals with major (>50%) European, African, or Asian/Amerindian ancestries.

2.2 | Beninese samples

We evaluated *HLA-C* variability of 108 unrelated individuals of the Toffin ethnic group from Sô-Ava, an area located in the Southern region of Benin, 12 km North of Cotonou, the economic capital of Benin. Toffins are etymologically known as “people of water.” Informed consent was obtained from all participants included in the study before blood collection. The study was approved by the Ethics Committee of the “Faculté des Sciences de la Santé (FSS)” of Cotonou, Benin and registered at No.12/03/2012/CEIFSS/UAC.

2.3 | *HLA-C* gene amplification and sequencing libraries

Genomic DNA was extracted by a salting-out procedure or QIAamp DNA Blood Midi kit (Qiagen, Hilden,

Germany), according to manufacturer's instructions. DNA samples were quantified with Qubit dsDNA Broad Range Assays (Thermo Fisher Scientific Inc., Waltham, Massachusetts) and normalized to 50 ng/μL.

HLA-C was amplified as a single amplicon of approximately 5671 nucleotides (not including primers sequences), spanning nucleotides 31 268 667-31 274 338 (chromosome 6 assembly hg38). Polymerase chain reaction (PCR) was carried out with primers HCPR.F1 (5′-TGAAGA ACTGAACAGCAACTA-3′) and HCUT.R1 (5′-GTCTGAG GGATAAGGGGCA-3′) in a final volume of 50 μL, containing 0.30 μM of each primer, 0.20 mM of each dNTP (Invitrogen, Carlsbad, California), 1.25 units of DNA polymerase (PrimeSTAR GXL, TaKaRa Bio Company), 1X PCR buffer solution supplied with the DNA polymerase and 50 ng of genomic DNA. Cycling conditions were 30 cycles of 98°C for 10 seconds, 60°C for 15 seconds, and 68°C for 6 minutes. Amplicons were evaluated on 1% agarose gel stained with GelRed (Biotium, Hayward), purified using Illustra ExoProStar (GE Healthcare), quantified using Qubit dsDNA High-Sensitivity Assays (ThermoFisher Scientific), and normalized to 0.2 ng/μL.

To prepare sequencing libraries, we used the Nextera XT Library Preparation Kit and Nextera XT Index Kit (both from Illumina, Inc.). Libraries were quantified by qPCR using Kapa (Kapa Biosystems, Wilmington). The fragmentation pattern was assessed with High-Sensitivity DNA Bioanalyzer chips (Agilent Technologies, California), and we normalized samples based on the quantification and the fragmentation pattern. Sequencing was performed using MiSeq Reagent Kit (V2, 500 cycles, 2 × 250 bp) in a MiSeq Platform, as recommended by Illumina Inc.

2.4 | Raw data processing, mapping, and genotyping

The polymorphic nature of HLA genes and the high sequence similarity loci may bias read mapping[41] in assays where all class I genes were sequenced together (as is the case in the present study). We, therefore, used *hla-mapper dna* to optimize read mapping at the *HLA-C locus* [42], version 3.0.5, with the following parameters: trimming error threshold set to 0.05, minimum read size set 70, and tolerance set to 0.05.

After the mapping procedure, we observed the underrepresentation of reads at intron 2. This low sequence depth is not related to mapping bias because further investigations using other library preparation kits circumvented this issue, indicating the underrepresentation was due to a fragmentation bias of Nextera XT kit, which affects especially CG-rich regions, such as intron 2. This was previously reported for the *HLA-A locus* [32].

We used the Genome Analysis Toolkit (*GATK*, version 4.1) *HaplotypeCaller* in the GVCF mode to infer genotypes, using hg38 as a reference (Mckenna et al.[43]). The multisample G.VCF file was created with *GATK CombineGVCFs* and the multisample VCF file with genotype likelihoods [created with *GATK GenotypeGVCFs*] was annotated considering the dbSNP version 150.

We processed the multisample VCF to introduce missing alleles on genotypes with low likelihoods or unbalanced genotypes, using *vcfx version 2.0*, available at www.castelli-lab.net/apps/vcfx. The missing alleles were introduced on genotypes presenting a likelihood lower than 99.999% (using *vcfx checkpl*), and on unbalanced genotypes (using *vcfx checkad*, with default parameters). Next, we refined variant calls using *vcfx evidence*, with an additional step in which we manually checked the variants that were excluded. Since *vcfx* introduced a large number of missing alleles at intron 2 because of the low sequencing depth and a large number of highly unbalanced genotypes in this segment, we opted to exclude the entire intron 2 from the analysis. However, this issue only affects studies using Nextera, and we have not observed a similar underrepresentation of intron 2 in WGS or using other fragmentation strategies.

2.5 | Phasing and *HLA-C* allele calling

We removed singletons before phasing. Next, we used *GATK ReadBackedPhasing* to phase sites occurring on the same read,[43] using a haplotype quality threshold of 2000. Nonetheless, this algorithm ignores indels and multiallelic *loci*. Then, we used *phasex* (available upon request) to perform the haplotyping analysis. This software considers the phasing sets detected by RBP and creates several VCF replicates, each of them, including a random phase set defined by RBP, for each sample. Then, *phasex* uses *Beagle 4.1* [44] to phase each VCF replicate, and the results are compared. For *HLA-C*, we used 100 replicates. When a sample presents the same pair of haplotypes in at least 95/100 of the replicates, this pair is fixed and passed forward to the next round of replicates. This strategy is repeated until no new sample achieves the 95/100 threshold. After the final round, all samples presenting the same haplotype pair in at least 70% of the replicates are considered phased and these haplotypes are used in the forthcoming analysis. Considering all heterozygous sites, 77.51% were directly phased using RBP. The combination *phasex/Beagle 4.1* phased the remaining 22.49%, most referring to indels or multiallelic *loci*, and also imputed the 0.37% of missing alleles observed after the *vcfx* treatment. Since we removed intron 2 as discussed earlier, the association between the segment

upstream and downstream intron 2 was entirely obtained using probabilistic models by *Beagle 4.1* [44].

After haplotyping, singletons were manually reintroduced whenever possible (ie, when a singleton is observed in a read that encompasses a neighboring heterozygous variation site). Then, we used *vcfx fasta* to create complete genomic sequences and *vcfx transcript* to create the CDS sequences (coding sequence, all exons)—two for each individual. Since *HLA-C* is encoded at the GRCh38 chromosome 6 reverse strand, sequences were reversed and complemented using *emboss revseq* [45].

We designed Perl scripts coupled with a local BLAST server with a database containing all known HLA alleles described by the IPD-IMGT/HLA database (version 3.36.0) to identify the closest known *HLA-C* coding allele for each different sequence we have detected, and the mutations observed when compared with it.[46] We inferred the encoded proteins with *emboss transeq* [45]. Promoters and 3′ UTR haplotypes were named according to sequence similarities and their phylogenetic relationship.

2.6 | Other analyses

We used Arlequin 3.5 to estimate haplotype frequencies, departures from Hardy-Weinberg equilibrium (HWE) expectations, nucleotide diversity, gene diversity, exact test of population differentiation, F_{ST} between populations, and Tajima's *D* for each *HLA-C* exon, intron, and regulatory segment, and *HLA-C* as a whole. The significance of the *D* statistic was tested by generating random samples under the hypothesis of selective neutrality and population equilibrium, using 5000 simulations.[47] The input files for Arlequin were created using *vcfx arlequin*. The nucleotide diversity and Tajima's *D* plots were calculated by using *VariScan* [48] in a sliding window approach of 150 bp and a step size of 3. We used *Pypop* to perform the Ewens-Watterson test.[49] The d_N/d_S ratio test, which evaluates the ratio of synonymous and nonsynonymous nucleotide substitution, was calculated using *MEGA* version 7.0.20 [50] for each *HLA-C* exon and using FUBAR[51] implemented in the *HYPHY 2.5* [52] package to evaluate positive selection for each codon. Because the allele diversity in a population is highly dependent on the sample size, we performed a rarefication procedure resulting in allelic richness estimates using *HP-RARE* [53] and 108 samples (the number of individuals in Benin). We also estimated the mean number of different haplotypes and the haplotype diversity observed in 10 000 replicates of 50 randomly selected individuals of each sample using a local Perl script. The frequency of each promoter, coding, and 3′UTR sequence was compared between samples using the Fisher exact test, using Bonferroni for correction of the significance level (α_c), considering the number of different sequences observed in each case.

Linkage disequilibrium (LD) among SNPs within the *HLA-C locus* was assessed using *Haploview* 4.2,[54] considering only sites with a minor allele frequency (MAF) higher than 1%. The PED and MAP files for Haploview were generated using *vcfx haploview*.

As a quality control for *HLA-C* typing, we used *Optitype* [55] to infer *HLA-C* coding alleles, and compared the results with those from our workflow.

3 | RESULTS

We evaluated *HLA-C* variability encompassing at least 1500 nucleotides from the upstream promoter to the end of the last *HLA-C* exon, with the exception of intron 2, in two different population samples. There were 359 variable sites across *HLA-C*, reported in Table S1, together with their chromosome positions, SNPid, the reference allele frequencies, and other information regarding each site. Genotype frequencies were consistent with the HWE expectations for more than 96% of the variants ($P > .05$), and the remaining are randomly distributed across *HLA-C* in both population samples. This supports our high-confidence genotype calls because biased mappings would lead to many variants not fitting HWE.

The region upstream from the *HLA-C* start codon, with approximately 1500 bases, which includes the distal promoter, the proximal promoter, and the 5′UTR, has 86 variants arranged into 33 different sequences (Table 1 and Alignment S1). The *HLA-C* CDS has 47 different sequences (Table 2) encoding 40 HLA-C protein molecules (Table 2). The *HLA-C* 3′UTR is encoded in the last exon and has 40 variable sites arranged into 25 different sequences (Table 3 and Alignment S2).

3.1 | *HLA-C* genetic diversity

HLA-C promoters were named following sequence similarities and their phylogenetic relationships (Table 1 and Alignment S1). There were 13 different promoter groups, with many shared variants within each group. In both samples, the most frequent promoter sequence was *P01:01*, with a frequency of 16.7% among Brazilians and 27.3% among Beninese (Table 1). This promoter is linked with alleles from the *HLA-C*04* group. Promoter frequencies differ between populations (considering $\alpha_c = 0.0015$), as observed for P01:01 ($P = .0006$), P04:02 ($P < .00001$), P05:01 ($P = .0002$), and P10:01 ($P < .00001$). Each *HLA-C* promoter group is associated with a specific *HLA-C*

coding allele group, supporting the high LD observed across the gene (Figure S1).

The *HLA-C* coding segment (all exons and introns, except intron 2) presented 82 different sequences, named as described in methods (Table S2). Among them, 67 sequences (81.71%) are identical to a known IPD-IMGT/HLA alleles, and the 15 represent new alleles (grouped as unknown sequences at Table S2), with a summed frequency of 2.09% when both populations are pooled. Some of these new alleles were detected more than once. For instance, there were six copies of a sequence that encodes *C*18:01*, with the same sequence as described in the IPD-IMGT/HLA database, but with a different nucleotide at the 3′UTR. Because of that, these sequences are placed under unknown in Table S2. As observed for the promoter segment, allele frequencies

TABLE 1 List of HLA-C promoter sequences in two population samples from Brazil and Benin, their frequencies and the HLA-C molecules associated with them

HLA-C promoters[a]	Brazilian frequency (2n = 836)	Benin frequency (2n = 216)	Associated HLA-C allele groups	Associated HLA-C molecules
P01:01	0.167	0.273	*C*04*	*C*04:01, C*04:07, C*04:09N*
P01:02	0.002	—	*C*04*	*C*04:01*
P01:03	0.001	—	*C*04*	*C*04:01*
P02:01	0.002	—	*C*01*	*C*01:02*
P02:02	0.022	0.014	*C*01*	*C*01:02*
P03:01	0.004	—	*C*14*	*C*14:02*
P03:02	0.029	—	*C*14*	*C*14:02, C*14:03*
P04:01	0.001	—	*C*06*	*C*06:02*
P04:02	0.140	0.028	*C*06, C*12*	*C*06:02, C*12:02, C*12:03*
P04:03	0.001	—	*C*06*	*C*06:02*
P04:04	0.007	—	*C*06*	*C*06:02*
P04:05	0.002	—	*C*12*	*C*12:03*
P04:06	0.001	—	*C*12*	*C*12:02*
P04:07	0.004	—	*C*12*	*C*12:02*
P05:01	0.062	0.144	*C*16*	*C*16:01, C*16:02, C*16:04*
P06:01	0.037	0.005	*C*15*	*C*15:02, C*15:05, C*15:08, C*15:09, C*15:13*
P07:01	0.001	—	*C*08*	*C*08:02*
P07:02	0.001	—	*C*08*	*C*08:01*
P07:03	0.093	0.069	*C*05, C*08*	*C*05:01, C*08:02, C*08:04*
P07:04	0.001	—	*C*08*	*C*08:03*
P08:01	0.047	0.028	*C*02*	*C*02:02, C*02:10, C*02:14*
P08:02	0.012	0.046	*C*02*	*C*02:10*
P08:03	—	0.005	*C*02*	*C*02:10*
P09:01	0.097	0.056	*C*03*	*C*03:02, C*03:03, C*03:04*
P10:01	0.030	0.130	*C*17*	*C*17:01, C*17:03, C*17:38*
P11:01	0.011	—	*C*07*	*C*07:04*
P11:02	0.001	—	*C*07*	*C*07:04*
P12:01	0.016	0.032	*C*18*	*C*18:01, C*18:02*
P13:01	0.011	—	*C*07*	*C*07:01*
P13:02	0.108	0.125	*C*07*	*C*07:01, C*07:06, C*07:18, C*07:35*
P13:03	0.005	—	*C*07*	*C*07:01*
P13:04	0.083	0.046	*C*07*	*C*07:02*
P13:05	0.001	—	*C*07*	*C*07:02*

[a]Please refer to the Alignment S1 for the promoter sequences.

TABLE 2 List of *HLA-C* CDS sequences detected in two population samples from Brazil and Benin, and their frequencies

HLA-C CDS[a]	Brazilian frequency (2n = 836)	Beninese frequency (2n = 216)
*C*01:02:01*	0.0239	0.0139
*C*02:02:02*	0.0383	0.0139
*C*02:10:01*	0.0191	0.0648
*C*02:14:02*	0.0012	—
*C*03:02:02*	0.0084	0.0324
*C*03:03:01*	0.0431	–
*C*03:03:04*	—	0.0046
*C*03:04:01*	0.0335	—
*C*03:04:02*	0.0096	0.0185
*C*03:04:58*	0.0012	—
C*04:01:01	0.1675	0.2731
*C*04:07:01*	0.0012	—
C*04:09 N	0.0024	—
*C*05:01:01*	0.0419	0.0046
C*06:02:01	0.0909	0.0139
*C*06:02:03*	—	0.0046
*C*07:01:01*	0.0897	0.0602
*C*07:01:02*	0.0156	—
*C*07:01:09*	0.0012	—
*C*07:02:01*	0.0837	0.0463
*C*07:04:01*	0.0120	–
*C*07:06:01*	—	0.0231
*C*07:18:01*	0.0167	0.0324
C*07:35	—	0.0093
*C*08:01:01*	0.0012	—
*C*08:02:01*	0.0514	0.0417
*C*08:03:01*	0.0012	–
*C*08:04:01*	0.0012	0.0231
*C*12:02:02*	0.0084	—
*C*12:03:01*	0.0562	0.0093
*C*14:02:01*	0.0287	—
*C*14:03:01*	0.0036	—
*C*15:02:01*	0.0287	—
*C*15:05:02*	0.0024	0.0046
*C*15:08:01*	0.0012	—
C*15:09	0.0012	—
*C*15:13:01*	0.0036	—
C*16:01:01	0.0431	0.1435
*C*16:02:01*	0.0120	—
*C*16:04:01*	0.0072	—
C*17:01:01	0.0239	0.1296
*C*17:03:01*	0.0048	—
C*17:38	0.0012	—
*C*18:01:01*	0.0084	—
*C*18:02:01*	0.0072	0.0324
Unknown1[b]	0.0012	—
Unknown2[b]	0.0012	—

Note: The symbol (—) represents the absence of this allele.

[a]*HLA-C* coding alleles according to the IPD-IMGT/HLA database version 3.36.0.

[b]There were two rare new CDS sequences, but encoding known HLA-C protein molecules. The first encodes *C*12:02*, and the second encodes *C*03:96*.

vary between both population samples (considering $\alpha_c = 0.001$) (Tables 2 and S2), as observed for CDS alleles *C*04:01:01* ($P = .0006$), *C*06:02:01* ($P = .0001$), *C*16:01:01* ($P = .00001$), and *C*17:01:01* ($P = .00001$). Only two CDS sequences are not identical to any sequence described in the IPD-IMGT/HLA database, but they encode previously described proteins (Table 2).

The haplotypes within the 3′UTR segment are arranged in four distinct groups separated by many mutational steps (Alignment S2). The most frequent 3′UTR haplotype in both populations was *U01:05, in LD* with *HLA-C*04* alleles, and others. As observed for the promoter region, each 3′UTR lineage (U01, U02, U03, and U04) is in LD with specific *HLA-C* allele groups (Table 3).

When we combine the promoter, CDS, and 3′UTRs sequences as extended haplotypes, there are 74 different *HLA-C* sequences (Table 4), and each *HLA-C* allele group is related to specific promoter and 3′UTR sequences, resulting in a high LD across the gene (Figure S1) and the single segregation block observed when all samples are pooled together.

HLA-C extended haplotypes do not deviate from HWE expectations in both populations ($P = .6608 \pm .0126$ in Benin; $P = .4940 \pm .0139$ in Brazil). The population differentiation exact test based on haplotype frequencies showed significant differences between these population samples ($P < .000001$), which is in agreement with the frequency shifts observed when considering any *HLA-C* segment (promoter, CDS, and 3′UTR). Population differentiation measured by F_{ST} was low ($F_{ST} = 0.0173$, gametic phase unknown, and $F_{ST} = 0.0290$, with known gametic phase), both statistically significant ($P < .0001$), and also low for each *HLA-C* exon, intron, and regulatory segments (Table 5). Allelic richness and private allelic richness are higher among Brazilians (2.01 and 0.09, respectively), than among Beninese (1.95 and 0.03, respectively). Moreover, when calculating the mean number of different haplotypes observed in 10 000 resamplings of 50 individuals, Beninese present 21.74 haplotypes with haplotype diversity of

0.8749, while Brazilians present 34.66 with haplotype diversity of 0.9472. Similar results are also observed when considering CDS sequences and encoded proteins.

3.2 | *HLA-C* evolutive aspects

Except for exon 6, all *HLA-C* segments presented high nucleotide diversity (π), positive Tajima's *D*, and negative normalized *F* values (Table 5). Both populations presented a significant positive Tajima's *D* for exon 2, exon 3, exon 4, exon 5, the 3′UTR, and the entire CDS. Likewise, they presented significant negative normalized *F* (Ewens-Watterson) for exon 1, exon 4, and exon 5 (Table 5). Using a 150 bp sliding window with a step size of three nucleotides to calculate nucleotide diversity and Tajima's *D*, both populations presented the same pattern, so we only plot results for the largest sample (Brazilians) in this section. Nucleotide diversity varies across the promoter segment, the region between positions −500 and −800 showing high conservation and negative Tajima's *D* (Figure 1). Nucleotide diversity was high throughout *HLA-C* CDS, except for two segments: the middle of both exon 2 (windows from 127 to 306) and 4 (windows from 655 to 828) (Figure 1). The highest nucleotide diversity coincides with exon 5 (windows from 868 to 1002), which presents frequent nonsynonymous mutations, including the insertion of six amino acids related to the *C*17* alleles (amino acid 301 to 306, Table S3). The first half of the 3′UTR segment is very conserved in both samples, but the second half presented nucleotide diversity higher than the observed for exons

TABLE 3 List of HLA-C 3′UTR sequences detected in two population samples from Brazil and Benin, and their frequencies

HLA-C 3′UTR sequence[a]	Brazilian frequency (2n = 836)	Beninese frequency (2n = 216)	Associated HLA-C molecules
U01:01	0.075	0.056	*C*03:02, C*03:03, C*03:04*
U01:02	0.019	—	*C*03:04*
U01:03	0.002	—	*C*03:04*
U01:04	0.032	—	*C*14:02, C*14:03*
U01:05	0.208	0.315	*C*01:02, C*04:01, C*04:07, C*04:09 N, C*18:01, C*18:02*
U01:06	0.001	—	*C*04:01*
U01:07	0.001	—	*C*18:01*
U01:08	—	0.005	*C*04:01*
U02:01	0.029	0.125	*C*17:01, C*17:03, C*17:38*
U02:02	0.001	—	*C*17:03*
U02:03	—	0.005	*C*17:01*
U03:01	0.106	0.069	*C*07:01, C*07:35*
U03:02	0.012	—	*C*07:04*
U03:03	0.081	0.046	*C*07:02*
U03:04	0.002	—	*C*07:02*
U03:05	0.017	0.056	*C*07:06, C*07:18*
U04:01	0.193	0.032	*C*06:02, C*12:02, C*12:03, C*15:02, C*15:05, C*15:08, C*15:09, C*15:13*
U04:02	0.001	—	*C*06:02*
U04:03	0.061	0.144	*C*16:01, C*16:02, C*16:04*
U04:04	0.001	—	*C*16:01*
U04:05	0.097	0.069	*C*05:01, C*08:01, C*08:02, C*08:03, C*08:04*
U04:06	0.055	0.074	*C*02:02, C*02:10, C*02:14*
U04:07	0.002	—	*C*02:02*
U04:08	0.001	—	*C*02:02*
U04:09	—	0.005	*C*02:10*

[a]Please refer to the Alignment S2 for the 3′UTR sequences.

TABLE 4 List of *HLA-C* extended haplotypes detected in two population samples from Brazil and Benin, and their frequencies

HLA-C haplotypes				
Promoter[a]	**CDS[b]**	**UTR[c]**	**Brazilian frequency (2n = 836)**	**Beninese frequency (2n = 216)**
P01:01	*C*04:01:01*	U01:05	0.1627	0.2685
P01:01	*C*04:01:01*	U01:06	0.0012	—
P01:01	*C*04:01:01*	U01:08	—	0.0046
P01:01	*C*04:07:01*	U01:05	0.0012	—
P01:01	*C*04:09N*	U01:05	0.0024	—
P01:02	*C*04:01:01*	U01:05	0.0024	—
P01:03	*C*04:01:01*	U01:05	0.0012	—
P02:01	*C*01:02:01*	U01:05	0.0024	—
P02:02	*C*01:02:01*	U01:05	0.0215	0.0139
P03:01	*C*14:02:01*	U01:04	0.0036	—
P03:02	*C*14:02:01*	U01:04	0.0251	—
P03:02	*C*14:03:01*	U01:04	0.0036	—
P04:01	*C*06:02:01*	U04:02	0.0012	—
P04:02	*C*06:02:01*	U04:01	0.0813	0.0139
P04:02	*C*06:02:03*	U04:01	—	0.0046
P04:02	*C*12:02:02*	U04:01	0.0036	—
P04:02	Unknown2	U04:01	0.0012	—
P04:02	*C*12:03:01*	U04:01	0.0538	0.0093
P04:03	*C*06:02:01*	U04:01	0.0012	—
P04:04	*C*06:02:01*	U04:01	0.0072	—
P04:05	*C*12:03:01*	U04:01	0.0024	—
P04:06	*C*12:02:02*	U04:01	0.0012	—
P04:07	*C*12:02:02*	U04:01	0.0036	—
P05:01	*C*16:01:01*	U04:03	0.0419	0.1435
P05:01	*C*16:01:01*	U04:04	0.0012	—
P05:01	*C*16:02:01*	U04:03	0.0120	—
P05:01	*C*16:04:01*	U04:03	0.0072	—
P06:01	*C*15:02:01*	U04:01	0.0287	—
P06:01	*C*15:05:02*	U04:01	0.0024	0.0046
P06:01	*C*15:08:01*	U04:01	0.0012	—
P06:01	*C*15:09*	U04:01	0.0012	—
P06:01	*C*15:13:01*	U04:01	0.0036	—
P07:01	*C*08:02:01*	U04:05	0.0012	—
P07:02	*C*08:01:01*	U04:05	0.0012	—
P07:03	*C*05:01:01*	U04:05	0.0419	0.0046
P07:03	*C*08:02:01*	U04:05	0.0502	0.0417
P07:03	*C*08:04:01*	U04:05	0.0012	0.0231
P07:04	*C*08:03:01*	U04:05	0.0012	—
P08:01	*C*02:02:02*	U04:06	0.0347	0.0139
P08:01	*C*02:02:02*	U04:07	0.0024	—
P08:01	*C*02:02:02*	U04:08	0.0012	—
P08:01	*C*02:10:01*	U04:06	0.0072	0.0139

(Continues)

TABLE 4 (Continued)

HLA-C haplotypes				
Promoter[a]	**CDS[b]**	**UTR[c]**	**Brazilian frequency (2n = 836)**	**Beninese frequency (2n = 216)**
P08:01	*C*02:14:02*	U04:06	0.0012	—
P08:02	*C*02:10:01*	U04:06	0.0120	0.0463
P08:03	*C*02:10:01*	U04:09	—	0.0046
P09:01	*C*03:02:02*	U01:01	0.0084	0.0324
P09:01	*C*03:03:01*	U01:01	0.0431	—
P09:01	*Unknown1*	U01:01	0.0012	—
P09:01	*C*03:03:04*	U01:01	—	0.0046
P09:01	*C*03:04:02*	U01:01	0.0096	0.0185
P09:01	*C*03:04:01*	U01:01	0.0120	—
P09:01	*C*03:04:01*	U01:02	0.0191	—
P09:01	*C*03:04:01*	U01:03	0.0024	—
P09:01	*C*03:04:58*	U01:01	0.0012	—
P10:01	*C*17:01:01*	U02:01	0.0239	0.1250
P10:01	*C*17:01:01*	U02:03	—	0.0046
P10:01	*C*17:03:01*	U02:01	0.0036	—
P10:01	*C*17:03:01*	U02:02	0.0012	—
P10:01	*C*17:38*	U02:01	0.0012	—
P11:01	*C*07:04:01*	U03:02	0.0108	—
P11:02	*C*07:04:01*	U03:02	0.0012	—
P12:01	*C*18:01:01*	U01:05	0.0072	—
P12:01	*C*18:01:01*	U01:07	0.0012	—
P12:01	*C*18:02:01*	U01:05	0.0072	0.0324
P13:01	*C*07:01:02*	U03:01	0.0108	—
P13:02	*C*07:01:01*	U03:01	0.0897	0.0602
P13:02	*C*07:01:09*	U03:01	0.0012	—
P13:02	*C*07:06:01*	U03:05	—	0.0231
P13:02	*C*07:18:01*	U03:05	0.0167	0.0324
P13:02	*C*07:35*	U03:01	—	0.0093
P13:03	*C*07:01:02*	U03:01	0.0048	—
P13:04	*C*07:02:01*	U03:04	0.0024	—
P13:04	*C*07:02:01*	U03:03	0.0801	0.0463
P13:05	*C*07:02:01*	U03:03	0.0012	—

[a]*HLA-C* CDS sequences. Please refer to Table 2 for their frequencies.
[b]*HLA-C* 3′UTR sequences. Please refer to the Alignment S2 for the promoter sequences and Table 3 for their frequencies.
[c]*HLA-C* promoter sequences. Please refer to the Alignment S1 for the promoter sequences and Table 1 for their frequencies.

2 and 3 (Figure 1). The high nucleotide diversity in exons 1, 2, 3, and 5 is also demonstrated when we considered each segment separately in both samples (Table 5). The highest Tajima's *D* observed across the *HLA-C* CDS ($D = 3.71$ for Brazilians and $D = 2.98$ for Beninese) is related to windows from position 193 to 342, that encodes residues between positions 41 and 90 (alpha 1 domain) of the HLA-C mature protein and it is important for both antigen presentation and KIR binding. In this segment, there were six highly frequent amino acid exchanges, including the C1/C2 dimorphism at position 80, which is important for KIR interaction (Table S3). The second half of the 3′UTR presents the highest Tajima's *D* across all *HLA-C* exons.

TABLE 5 Nucleotide diversity, neutrality tests, and F_{ST} across the *HLA-C* locus

		Brazilian samples			Beninese samples			
Region	**Length**	**Nucleotide diversity**	**Tajima's D[a]**	**Ewens-Watterson[b]**	**Nucleotide diversity**	**Tajima's D[a]**	**Ewens-Watterson[b]**	F_{ST}
Promoter and 5′UTR	1525	0.013 ± 0.006	1.175, $P = .895$	Fo = 0.095, Fe = 0.132, F = −0.803, $P = .178$	0.014 ± 0.007	1.044, $P = .888$	Fo = 0.143, Fe = 0.232, F = −1.061, $P = .071$	**0.020, $P < .01$**
Exon 1	73	0.024 ± 0.015	1.511, $P = .924$	**Fo = 0.275, Fe = 0.543, F = −1.444, $P = .030$**	0.031 ± 0.019	**1.887, $P = .961$**	**Fo = 0.234, Fe = 0.475, F = −1.474, $P = .013$**	**0.026, $P = .026$**
Intron 1	130	0.019 ± 0.011	1.109, $P = .889$	**Fo = 0.130, Fe = 0.359, F = −1.658, $P < .001$**	0.019 ± 0.011	1.022, $P = .861$	Fo = 0.196, Fe = 0.348, F = −1.195, $P = .050$	**0.011, $P < .001$**
Exon 2	270	0.026 ± 0.013	**3.178, $P = .997$**	**Fo = 0.086, Fe = 0.214, F = −1.557, $P < .001$**	0.025 ± 0.013	**2.256, $P = .983$**	Fo = 0.146, Fe = 0.232, F = −1.026, $P = .086$	**0.011, $P < .001$**
Intron 2[c]	250	—	—	—	—	—	—	—
Exon 3	276	0.029 ± 0.015	**1.949, $P = .968$**	**Fo = 0.092, Fe = 0.164, F = −1.184, $P = .028$**	0.029 ± 0.015	**1.679, $P = .954$**	Fo = 0.156, Fe = 0.202, F = −0.648, $P = .274$	**0.021, $P < .001$**
Intron 3	587	0.014 ± 0.007	1.665, $P = .949$	**Fo = 0.106, Fe = 0.235, F = −1.412, $P = .004$**	0.016 ± 0.008	1.406, $P = .932$	Fo = 0.148, Fe = 0.232, F = −0.999, $P = .096$	**0.017, $P < .001$**
Exon 4	276	0.021 ± 0.011	**2.072, $P = .975$**	**Fo = 0.161, Fe = 0.314, F = −1.241, $P = .031$**	0.023 ± 0.012	**2.111, $P = .978$**	**Fo = 0.159, Fe = 0.292, F = −1.238, $P = .031$**	**0.026, $P < .001$**
Intron 4	124	0.046 ± 0.024	1.196, $P = .891$	**Fo = 0.218, Fe = 0.417, F = −1.270, $P = .042$**	0.050 ± 0.026	0.779, $P = .825$	Fo = 0.202, Fe = 0.348, F = −1.146, $P = .065$	**0.008, $P = .036$**
Exon 5	120	0.041 ± 0.022	**1.828, $P = .961$**	**Fo = 0.152, Fe = 0.387, F = −1.587, $P < .001$**	0.068 ± 0.035	**1.803, $P = .966$**	**Fo = 0.157, Fe = 0.348, F = −1.502, $P = .003$**	**0.028, $P < .001$**
Intron 5	440	0.017 ± 0.009	**2.272, $P = .984$**	**Fo = 0.153, Fe = 0.296, F = −1.238, $P = .028$**	0.016 ± 0.008	1.315, $P = .909$	Fo = 0.183, Fe = 0.319, F = −1.153, $P = .058$	**0.012, $P = .009$**
Exon 6	33	0.001 ± 0.003	**−1.218, $P = .044$**	Fo = 0.962, Fe = 0.672, F = 1.461, $P = .933$	0.003 ± 0.006	−0.372, $P = .286$	Fo = 0.895, Fe = 0.833, F = 0.371, $P = .469$	**0.026, $P = .026$**
Intron 6	107	0.013 ± 0.009	0.858, $P = .830$	Fo = 0.415, Fe = 0.496, F = −0.459, $P = .392$	0.014 ± 0.009	0.953, $P = .850$	Fo = 0.296, Fe = 0.475, F = −1.098, $P = .109$	**0.007, $P = .027$**
Exon 7	48	0.014 ± 0.012	1.709, $P = .935$	Fo = 0.652, Fe = 0.672, F = −0.101, $P = .488$	0.012 ± 0.011	0.941, $P = .850$	Fo = 0.716, Fe = 0.833, F = −0.700, $P = .262$	0.004, $P = .144$
Intron 7	164	0.016 ± 0.010	1.473, $P = .932$	Fo = 0.228, Fe = 0.387, F = −1.073, $P = .097$	0.015 ± 0.009	0.909, $P = .842$	Fo = 0.331, Fe = 0.383, F = −0.375, $P = .433$	**0.012, $P = .009$**
Exon8/3′UTR	425	0.023 ± 0.012	**1.850, $P = .963$**	Fo = 0.123, Fe = 0.194, F = −0.967, $P = .103$	0.025 ± 0.013	**1.739, $P = .959$**	Fo = 0.160, Fe = 0.250, F = −0.978, $P = .110$	**0.015, $P < .001$**
CDS	1101	0.026 ± 0.012	**2.462, $P = .989$**	Fo = 0.070, Fe = 0.095, F = −0.838, $P = .163$	0.030 ± 0.014	**2.211, $P = .988$**	Fo = 0.129, Fe = 0.143, F = −0.294, $P = .469$	—

Note: Significant *P*-values are marked in boldface (P < 5%).

[a]Tajima's *D* values computed by Arlequin 3.5 software. *D* statistic tested by generating random samples under the hypothesis of selective neutrality and population equilibrium, and the ***P*** was obtained as the proportion of simulated *D* statistics less than or equal to the D observed (No. of simulation = 5.000).

[b]Ewens-Watterson's F normalized values computed by Pypop software (numReplicates = 20.000).

[c]Intron 2 was not evaluated in this study.

The d_N/d_S ratio test indicated an excess of nonsynonymous changes at exon 1, which is consistent with positive selection in both populations (Table 6). This is also supported by the positive and significant normalized *F* and Tajima's *D* for exon 1. All variable sites detected in exon 1 are frequent nonsynonymous mutations (Table S3). Furthermore, we found the same profile in all populations from the 1000Genomes project, and also in the sequences available at the IPD-IMGT/HLA database (data not shown). When we estimate d_N/d_S for each *HLA-C* codon separately, there is evidence of positive selection for the codons encoding amino acids 24, 80, 116, 156, and 163 (mature protein, not considering the leader peptide, posterior probability >0.95), in both population samples, amino acids 9 and 95 Among Brazilians, and amino acid 73 Among Beninese. There was no evidence of positive selection in any codon encoding the leader peptide.

4 | DISCUSSION

Here we present a bioinformatics approach to evaluate *HLA-C* variability when using second-generation sequencing, providing accurate genotypes and haplotypes from the upstream promoter up to the complete 3′UTR segment.

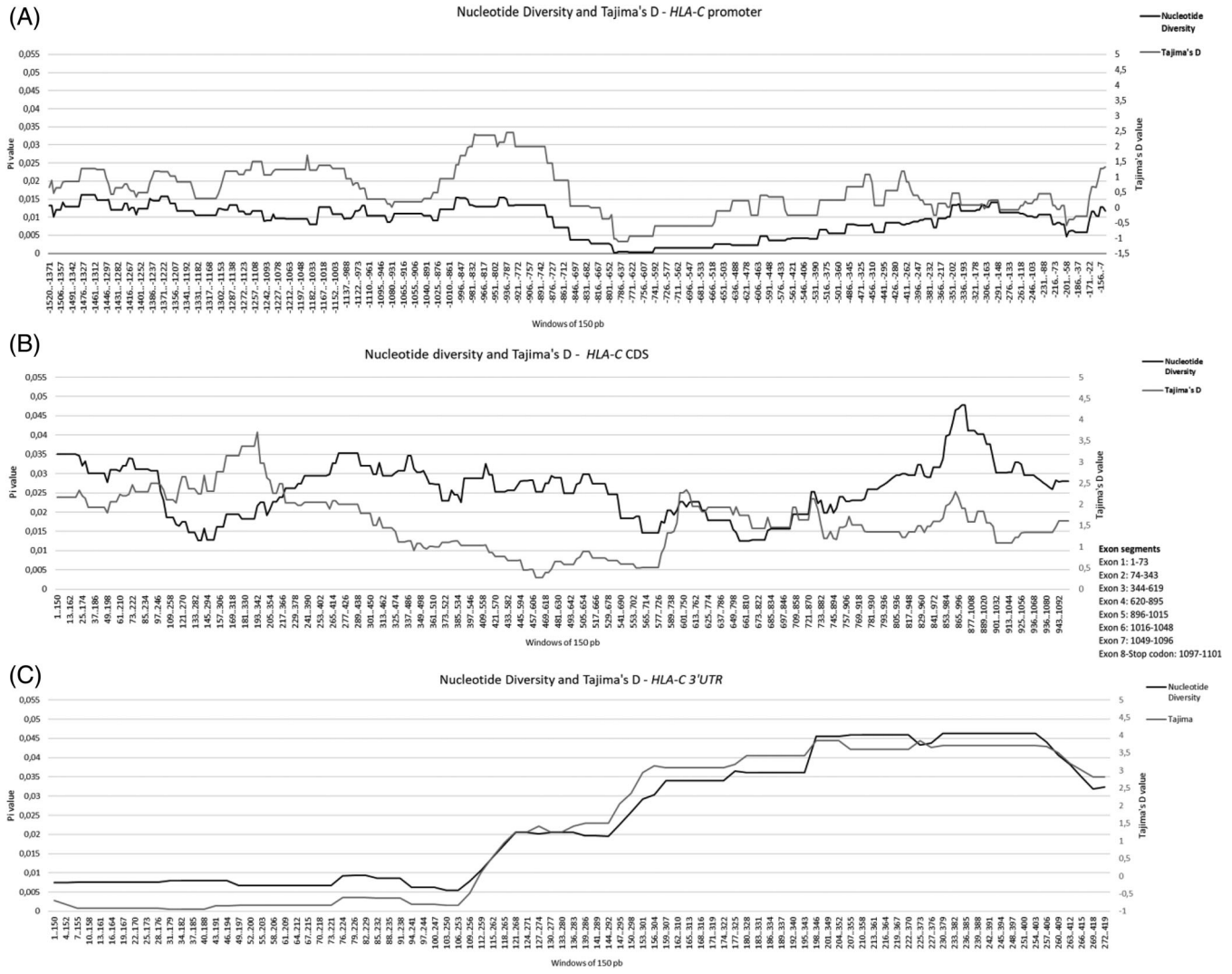

FIGURE 1 Nucleotide diversity and Tajima's D at the *HLA-C* promoter, CDS and 3′UTR. A, Nucleotide diversity and Tajima's D of *HLA-C* 5′ upstream regulatory segment examined in all aligned sequences using sliding window of 150 bp with a step size of 3. The x-axis represents the windows of 150pb comprehending the *HLA-C* segments and the nucleotide position relative to the alignment using as reference the position −1 from IMGT database (genomic data); B, nucleotide diversity of *HLA-C* exons. The position 1(from window 1..150) in the x-axis corresponds to first base from exon 1, the subsequent positions was given by alignment position (see caption); and **C**, nucleotide diversity of *HLA-C 3′UTR*

TABLE 6 d_N/d_S ratio test of the *HLA-C* exons and CDS

Region	Size	Codons	Brazilian sample			Beninese sample		
			Ha = *Neutrality* ($d_N \neq d_S$)	*Ha* = Positive Selection ($d_N > d_S$)	*Ha* = Purifying selection ($d_N < d_S$)	*Ha* = *Neutrality* ($d_N \neq d_S$)	*Ha* = Positive selection ($d_N > d_S$)	*Ha* = Purifying selection ($d_N < d_S$)
Exon 1	72	24	**d_N-d_S = 2.031, *P* = .044**	**d_N-d_S = 2.034, *P* = .022**	d_S-d_N = −2.051, *P* = 1.000	**d_N-d_S = 2.387, *P* = .019**	**d_N-d_S = 2.378, *P* = .009**	d_S-d_N = −2.363, *P* = 1.000
Exon 2	270	90	d_N-d_S = 0.382, *P* = .703	d_N-d_S = 0.381, *P* = .352	d_S-d_N = −0.376, *P* = 1.000	d_N-d_S = 0.537, *P* = .592	d_N-d_S = 0.554, *P* = .290	d_S-d_N = −0.541, *P* = 1.000
Exon 3	276	92	d_N-d_S = 0.113, *P* = .910	d_N-d_S = 0.118, *P* = .453	d_S-d_N = −0.115, *P* = 1.000	d_N-d_S = −0.028, *P* = .978	d_N-d_S = −0.028, *P* = 1.000	d_S-d_N = 0.028, *P* = .489
Exon 4	276	92	d_N-d_S = −1.403, *P* = .163	d_N-d_S = −1.368, *P* = 1.000	d_S-d_N = 1.402, *P* = .082	d_N-d_S = −1.135, *P* = .259	d_N-d_S = −1.140, *P* = 1.000	d_S-d_N = 1.159, *P* = .124
Exon 5	120||138	40||46	d_N-d_S = −0.021, *P* = .983	d_N-d_S = −0.022, *P* = 1.000	d_S-d_N = 0.022, *P* = .491	d_N-d_S = 0.340, *P* = .735	d_N-d_S = 0.339, *P* = .368	d_S-d_N = −0.339, *P* = 1.000
Exon 6	33	11	d_N-d_S = 0.649, *P* = .518	d_N-d_S = 0.631, *P* = .265	d_S-d_N = −0.625, *P* = 1.000	d_N-d_S = 1.025, *P* = .307	d_N-d_S = 1.037, *P* = .151	d_S-d_N = −1.039, *P* = 1.000
Exon 7	48	16	d_N-d_S = 1.505, *P* = .135	d_N-d_S = 1.472, *P* = .072	d_S-d_N = −1.506, *P* = 1.000	d_N-d_S = 1.508, *P* = .134	d_N-d_S = 1.504, *P* = .068	d_S-d_N = −1.489, *P* = 1.000
CDS	1101||1119	367||373	d_N-d_S= 0.016, *P* = .987	d_N-d_S = 0.016, *P* = .494	d_S-d_N = −0.016, *P* = 1.000	d_N-d_S = 0.409, *P* = .683	d_N-d_S = 0.409, *P* = .342	d_S-d_N = −0.410, *P* = 1.000

Note: d_N/d_S ratio test was performed using all sequences found for each segment of the *HLA-C* locus, using bootstrap method (5000 replicates). Significant *P*-values are marked in boldface.

This methodology relies entirely on publicly available software. The majority of variants and sequences here detected had already been described by the IPD-IMGT/HLA database. However, the preexisting sequences do not influence the detection of similar or identical sequences given our pipeline's properties, since preexisting sequences are not considered by our genotyping and phasing procedure. Allele imputation did not exceed 0.37%, and the GATK *ReadBackedPhasing* phased directly 77.51% of the heterozygous sites. We found at least one allele from each main *HLA-C* allele group (ie, one field resolution level). For the Brazilian sample, the CDS (and genomic) allele frequencies are similar to the ones reported in previous studies using different typing techniques[56,57] (http://www.allelefrequencies.net/). For Beninese, the CDS frequencies are similar to the ones reported in another sub-Saharan African population sample from Senegal, West Africa, named Mandenka, as reported elsewhere.[58] This is the first survey addressing Beninese *HLA-C* genetic diversity. Most 3′UTR haplotypes here detected are compatible with partial or full 3′UTR sequences reported by IPD-IMGT/HLA. Moreover, 90% of the coding alleles reported here are identical to the ones called by the Optitype software,[55] and the majority of the differences between methods include alleles that differ in segments not covered by the Optitype database. Unfortunately, the Optitype database available for download is outdated.

As observed for other HLA genes, such as *HLA-A* and *HLA-G* [29,32], there was a strong LD across the entire *HLA-C locus*, with a single segregation block (Figure S1) and many variants in complete LD. This LD profile supports the close relationship among the coding and regulatory sequences. Since each allele group is related to specific regulatory sequences, different transcription factors and microRNAs may modulate their expression levels. Previous studies have demonstrated heterogeneous expression levels among different *HLA-C* allele groups.[7,9,59]

Since *HLA-C* is expressed together with classical HLA presenting peptides in somatic tissues and together with nonclassical HLA in placenta, a distinct pattern from other loci,[6] we evaluated natural selection signatures across the *HLA-C locus* to better understand HLA-C biology and function, and also how natural selection has shaped its variability and haplotype structure. In this regard, the presence of divergent haplotypes is compatible with balancing selection, as observed across the entire *HLA-C locus* with few exceptions. Positive Tajima's *D* values are consistent with an excess of alleles with an intermediate frequency due to balancing selection directly on the promoter, or due to selection on the protein-coding portion of *HLA-C* and mediated by linkage.

Our results indicate balancing selection not only at exons 2 and 3, as expected for an antigen presentation

gene, but at other *HLA-C* segments, such as part of the promoter, exon 1, exon 4 and 5, and the second half of the *HLA-C* 3′UTR, and also some intronic segments. The Ewens-Watterson test for neutrality also indicates a low homozygosity rate and supports the evidence of balancing selection. It is not clear whether these results reflect a hitchhiking effect due to the high LD throughout *HLA-C* or direct selection on these regions. Moreover, Goeury and colleagues also reported positive Tajima′s *D* values for all exons except exon 6 in another sub-Saharan sample, with the highest value located at exon 2, as detected here in our Brazilian and Beninese samples.[58] Balancing selection was described for classical HLA coding regions at exons 2 and 3, enhancing antigen presentation capabilities, and hardly any demographic or genetic factors can explain the high degree of polymorphism and excess of nonsynonymous variants at these genes (reviewed by[60]). However, since *HLA-C* and *HLA-B* are only 82Kb apart, the LD between these two loci might be influencing these results.

4.1 | An immunomodulatory role may be shaping the variability at *HLA-C* exon 1

We detected an excess of nonsynonymous variants in exon 1, with evidence of positive selection in this exon in both population samples (Tables 5 and 6) when considering the entire segment. No evidence was detected when specific codons within exon 1 are evaluated using FUBAR. Similar results have been described in Mandenka, in which all exon 1 variants configure nonsynonymous mutations.[58] This is intriguing because exon 1 encodes the HLA-C leader peptide and it is primarily involved in targeting HLA-C to the cellular membrane, and it is not part of the mature HLA-C molecule. This pattern was not observed for *HLA-A* in this same Brazilian sample.[32]

The leader peptide plays other roles besides addressing the molecule to the secretory pathway. After being cleaved, a leader peptide may act as a hormone, neurotransmitter, and as a self-antigen.[61] This evidence of positive selection may be related to possible coevolution between *HLA-C* and *HLA-E*. Residues 3-11 of the HLA-C leader peptide can bind to HLA-E molecules as self-antigens, stabilizing HLA-E in the cell surface and modulating NK cell activity[62] by interacting mainly with the CD94/NKG2A inhibitory receptor.[63] Both HLA-C and HLA-E are co-expressed in the placenta during pregnancy.[19] In our sample, there are four major HLA-C leader peptides considering residues 3-11, VMAPRTLIL (now denominated LP1, with a mean frequency of 62.8%), VMAPRALLL (LP2, 22.8%), VMAPRTLLL (LP3, 9.3%), and VMAPQALLL (LP4, 5.1%). LP1 and LP3 are the ancestral protein sequences and are encoded by *MHC-C* in many primates, including *Pongo abelii*, *Gorilla gorilla*, and *Pan troglodytes*, as well as other MHC genes among primates (The IPD-MHC database, Release 3.4.0.0, from 19 December 2019). LP2 and LP4 can be found only in humans and are encoded only by *HLA-C* in the MHC complex. Thus, LP2 and LP4 are human-specific HLA-C leader peptides, and they have been maintained in high frequency in both these populations.

HLA-E among humans is mainly dimorphic (Arg107Gly), with two major proteins, *E*01:01* (Arg) *and E*01:03* (Gly), with frequencies around 50% in worldwide populations. This HLA-E dimorphism causes an alteration of the HLA-E peptide repertoire[64] and seems to be maintained by balancing selection.[65,66] *E*01:03* is the ancestral allele and can be found among many primates, *while E*01:01* raised among humans,[67] reaching frequencies around 50%. Previous studies indicated that the ancestral *HLA-C* leader peptide LP1 binds preferentially to *E*01:03* (also the ancestral allele), while the human-exclusive LP2 binds preferentially to the human-exclusive *E*01:01*.[64] Thus, *there* is a functional connection between *E*01:01* and LP2 among humans and both have raised among humans and present high frequencies in worldwide populations. It is not clear why *E*01:01* frequency has increased, but previous studies indicated that *E*01:01* is the least effective against NK cell lysis,[64] and one possible explanation would involve reaching a balance between immunosuppression and immune defense. Balancing selection may be maintaining different alleles in intermediate frequencies, some of them favoring immune responses against pathogens, and some favoring immune tolerance, avoiding either exacerbated responses and autoimmunity.

However, other scenarios are possible, including positive selection to improve secretion efficiency,[68] linkage disequilibrium (LD) between *HLA-E*01:01* and LP2, and also the high LD across *HLA-C* (Figure S1), because balancing selection in the coding region maintains many *HLA-C* alleles in worldwide populations. Moreover, there is evidence of positive selection in codons that influences the HLA-C peptide repertoire and KIR binding.

4.2 | Peptide presentation may be shaping the variability at *HLA-C* exons 2 to 5

As expected for HLA genes,[60] we detected balancing selection at exons 2 and 3, which encodes the antigen-binding domain, in both populations (Table 5). We also detected positive selection in amino acid positions that influence the HLA-C peptide-binding repertoire,[69] including positions 9, 73, 80, 95, 116, 156, and 163 (mature protein, not considering the leader peptide, FUBAR posterior probability >0.95). Moreover, there was

evidence of positive selection in position 80, related to the HLA-C dimorphism C1/C2 that influence the binding of KIR receptors,[2] which coincides with the highest Tajima's *D* value across the *HLA-C* CDS. HLA-C presents many conserved motifs in the α1 and α2 domains that are exclusive to *HLA-C* [6]. This gene also presents a reduced diversity at the antigen recognition site when compared with *HLA-A* and *HLA-B*, and therefore a reduced set of self-peptides able to bind HLA-C.[6,58,70,71] Buhler and colleagues studied a large set of samples (second field level resolution) to estimate pairwise molecular distances among classical HLA class I alleles and predict pairwise peptide-binding distances between all alleles and corresponding encoded molecules.[71] *HLA-C* presented lower allele diversity and their molecules exhibit lower pairwise peptide-binding distances when compared with other classical HLA class I loci.[71] Moreover, *HLA-C* alleles seem to be subdivided into two broad groups regarding peptide presentation properties. These observations support the minor role of *HLA-C* to the diversification of class I peptide presentation.[71] However, it does not rule out the influence of balancing selection on the antigen-presenting region, as it was already observed for the C1/C2 allotypes that are important for KIR interaction.[4]

Here we found evidence of balancing/positive selection in the antigen presentation region and also in codons related to KIR interaction. However, some segments in the region encoding the peptide-binding groove are conserved, especially at the end of exon 2 (Figure 1B, windows from position 109 to 342). This feature also supports the minor role of *HLA-C* in antigen presentation when compared with other classical HLA class I genes. Thus, evolution in *HLA-C* may favor KIR interaction conserving important residues in this matter, and also diversifying others such as residue 80 [the C1/C2 dimorphism], but, in a lesser extent, also diversifying the region related to antigen presentation.

We also detected a high nucleotide diversity and evidence of balancing selection for exons 4 (α3 domain) and 5 (transmembrane domain) in both population samples. This pattern was also observed among Mandenka.[58] Since the α3 domain interacts with the CD8 co-receptor to facilitate TCR signaling, we expected conservation of this *HLA-C* segment as was observed for *HLA-A* [32]. In spite of that, the HLA-C α3/CD8 interaction may be preserved because amino acids located within residues 222-245 that are important to HLA/CD8 interaction are preserved[72-74] with only one rare variant (Glu229Gln) within this region (Table S3) that is associated with allele *HLA-C*15:13*. Many variants surround this segment and present high heterozygosis, but it is uncertain whether these variants influence α3/CD8 interaction. The segment encoding the transmembrane domain presented the highest nucleotide diversity among all *HLA-C* segments, with frequent nonsynonymous mutations and a frequent inframe indel.

4.3 | Interaction with regulatory molecules may be shaping the variability at *HLA-C* promoter and 3′UTR regions

The promoter region presents a conserved segment around position −700, with a monomorphic region of 115 nucleotides from position 6:31272722 to 6:31272836 (hg38). Ramsuran et al[17] had previously detected greater conservation in this segment studying homozygous cell lines,[17] and here we detected the same conservation in 526 individuals from two different populations. This segment is also conserved among different primates. Notwithstanding, some frequent variations at the promoter region (such as rs2395471 A/G) appear to be crucial for gene regulation, as have been shown by the different mRNA expression levels in cells that present promoter haplotypes that differ by few mutations.[15,17]

Nucleotide diversity and Tajima′s *D* were low at the beginning of the 3′UTR segment but very high at the second half (Figure 1), in both population samples. Usually, the beginning of the 3′UTR segment is important for miRNA binding and post-transcription regulation.[75-77] This conservation contrasts with the excess of heterozygosis detected across HLA-C. The first half of the 3′UTR might have been maintained because of its critical role for *HLA-C* post-transcriptional regulation. This region might be under purifying selection. MicroRNA binding analysis indicate that the second half of the 3′UTR is not an important target for miRNAs with few exceptions, such as miR-148a (data not shown). It is well established that variants modifying the miR-148a binding site influence HLA-C expression levels.[9,13] Thus, this region might be under a relaxed purifying selection, and it coincides with the presence of many frequent variants.

We may find both conserved and highly polymorphic segments throughout *HLA-C*. While variants that may affect gene regulation, such as the one within the miR-148a binding site or other variants within the promoter are maintained at high heterozygosis in both populations, there are highly conserved segments in the regulatory regions. This may contribute to a complex mechanism of transcriptional and post-transcriptional regulation of gene expression in tissues with different microenvironments. Thus, we suggest that HLA-C expression is finely regulated due to its dual function, that is, antigen presentation and T and NK cell modulation. Besides, the interaction between HLA-C and T and NK cells receptors, as also with the CD8 co-receptor, is apparently conserved

considering the variability here detected in a population-based study in two samples from different continents.

4.4 | LD, coupled with regulation of gene expression and *HLA-C* dual function, may be influencing the entire *HLA-C* variability landscape

HLA-C diversity is high in both populations considering the entire *locus* and each segment separately. Some haplotypes are more frequent in one population than the other, but they are usually present in both. For instance, *HLA-C*16* and *HLA-C*17*, alleles and their associated regulatory sequences are frequent in Ghana,[78] Among Mandenka[58] and Beninese, but not in Brazil, while the opposite is observed for *HLA-C*05* and *HLA-C*06*. Nevertheless, they are all represented in both populations. Because of that, population differentiation measured by F_{ST}, although significant (probably due to the large sample sizes), was low ($F_{ST} = 0.0173$). Interestingly, when the amino acid sequence is taken into account (Table S3), there are similar frequencies in important motifs for the HLA-C function and expression regulation. For instance, motif KYRV (amino acids 66, 67, 69, and 76), which is associated with lower cell surface HLA-C expression,[79] presents a frequency of 84% in Brazil and 87% in Benin. The dimorphism C1/C2 (amino acid 80) important for KIR interaction is also common in both populations (C1 frequency of 45.8% in Benin and 54.2% in Brazil). Despite the lower frequency of C1 in Benin ($P = .027$), we found no statistical significance after correction probably due to the small sample size of the Benin sample. However, previous reports have indicated lower C1 frequency in malaria high-endemic populations.[80] Additionally, KIR2DL3 and its cognate HLA-C1 ligand were significantly associated with the development of cerebral malaria and thus, natural selection may explain this reduced frequency of the KIR2DL3-HLA-C1 combination observed in malaria high-endemic populations.[80] The Guanine deletion influencing miR-148a binding (and HIV outcome) presents a frequency of 41.15% in Brazil and 32.41% in Benin, and it is included in the U4 3′UTR haplotypes (Alignment S2 and Table 4). Variant rs2395471 at the promoter segment, in which allele Adenine influences the binding of Oct1 and is associated with higher HLA-C expression[15] presents the same frequency in both populations (Table S1). The same for rs10657191 that influences the response to IFN-γ.[16] The only exception in this scenario would be variant rs2524094, which may influence the response to TNF-α[16] because the variant associated with a functional TNF response is much more frequent in Benin than Brazil.

The frequency of amino acids influencing peptide-binding and thus the HLA-C peptide repertoire[81] are different between samples, as observed for amino acid 9, 114, 116, and 156 (Table S3). Because of that, the peptide repertoire presented by these populations might be different. Moreover, the frequency of the amino acids that compose the transmembrane segment is also different between samples, but it is not clear whether these modifications (which includes a large in-frame indel) modifies HLA-C function and stability.

5 | CONCLUDING REMARKS

Here we present a molecular and bioinformatic approach to evaluate the entire *HLA-C* variability using NGS and freely available software. We applied this method to survey *HLA-C* diversity in two population samples with different demographic histories, Brazil and Benin. We also present the haplotypes and frequencies for the complete *HLA-C* regulatory regions. The *HLA-C* promoter was very polymorphic, with the presence of few haplotypes presenting many mutational steps apart, but also presenting a monomorphic segment of 115 nucleotides around position −700 that is shared among different primates. Nucleotide diversity was very high at exons 2 and 3 and higher in other exonic segments and the second half of the 3′UTR region. We detected evidence of balancing selection on the entire *HLA-C locus* (exception made to the promoter region) and positive selection in exon 1/leader peptide, for both populations. The frequencies of HLA-C motifs previously associated with KIR interaction and expression regulation are similar between both populations, while we detected differences in the frequency of amino acids that influence the peptide-binding repertoire and the transmembrane region. Linkage disequilibrium along the *HLA-C locus* is high, with many variants in complete LD. Because of that, each allele group is associated with specific regulatory sequences. The same patterns of nucleotide diversity, natural selection signatures, regulatory and extended haplotypes, were observed in both samples, Brazil with a major European contribution, and Benin with a major African contribution. This emphasizes the role of shared evolutionary aspects rather than specific demographic histories in shaping *HLA-C* genetic diversity.

ACKNOWLEDGMENTS

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo—FAPESP/Brazil (Grants

2013-17084-2 and 2017/19223-0); by Institut de Recherche pour le Développement, France; by Brazil-France Research Cooperation Program USP/COFECUB (Grant numbers Uc Me 169-17). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001. A PhD scholarship was awarded by Institut de Recherche pour le Développement to Léonidas Tokplonou and to Paulin Sonon by CAPES/PROEX/Brazil (Finance Code 001).

CONFLICT OF INTEREST

The authors confirm that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data availability statement The data that support the findings of this study are available from the corresponding author upon request

ORCID

Andreia S. Souza https://orcid.org/0000-0002-6853-1559
Luciana C. Veiga-Castelli https://orcid.org/0000-0002-2652-8562
Erick C. Castelli https://orcid.org/0000-0003-2142-7196

REFERENCES

1. Rock KL, Reits E, Neefjes J. Present yourself! by MHC class I and MHC class II molecules. *Trends Immunol*. 2016;37(11):724-737.
2. Parham P. Killer cell immunoglobulin-like receptor diversity: balancing signals in the natural killer cell response. *Immunol Lett*. 2004;92(1–2):11-13.
3. Yawata M, Yawata N, Draghi M, Little AM, Partheniou F, Parham P. Roles for HLA and KIR polymorphisms in natural killer cell repertoire selection and modulation of effector function. *J Exp Med*. 2006;203(3):633-645.
4. Parham P, Moffett A. Variable NK cell receptors and their MHC class I ligands in immunity, reproduction and human evolution. *Nat Rev Immunol*. 2013;13(2):133-144.
5. Penman BS, Moffett A, Chazara O, Gupta S, Parham P. Reproduction, infection and killer-cell immunoglobulin-like receptor haplotype evolution. *Immunogenetics*. 2016;68(10):755-764.
6. Blais M-E, Dong T, Rowland-Jones S. HLA-C as a mediator of natural killer and T-cell activation: spectator or key player? *Immunology*. 2011;133(1):1-7.
7. Apps R, Qi Y, Carlson JM, et al. Influence of HLA-C expression level on HIV control. *Science*. 2013;340(6128):87-91.
8. Thomas R, Apps R, Qi Y, et al. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet*. 2009;41(12):1290-1294.
9. Kulkarni S, Savan R, Qi Y, et al. Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*. 2011;472(7344):495-498.
10. Fellay J, Shianna KV, Ge D, et al. A whole-genome association study of major determinants for host control of HIV-1. *Science*. 2007;317(5840):944-947.
11. Petersdorf EW, Gooley TA, Malkki M, et al. HLA-C expression levels define permissible mismatches in hematopoietic cell transplantation. *Blood*. 2014;124(26):3996-4003.
12. Tiercy J-M. HLA-C incompatibilities in allogeneic unrelated hematopoietic stem cell transplantation. *Front Immunol*. 2014;5:216-216.
13. Kulkarni S, Qi Y, O'HUigin C, et al. Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. *Proc Natl Acad Sci U S A*. 2013;110(51):20705-20710.
14. Chen H, Hayashi G, Lai OY, et al. Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. *PLoS Genet*. 2012;8(2):e1002514.
15. Vince N, Li H, Ramsuran V, et al. HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. *Am J Hum Genet*. 2016;99(6):1353-1358.
16. Hundhausen C, Bertoni A, Mak RK, et al. Allele-specific cytokine responses at the HLA-C locus: implications for psoriasis. *J Invest Dermatol*. 2012;132(3) Pt 1:635-641.
17. Ramsuran V, Hernandez-Sanchez PG, O'HUigin C, et al. Sequence and phylogenetic analysis of the untranslated promoter regions for HLA class I genes. *J Immunol*. 2017;198(6):2320-2329.
18. Chazara O, Xiong S, Moffett A. Maternal KIR and fetal HLA-C: a fine balance. *J Leukoc Biol*. 2011;90(4):703-716.
19. Hackmon R, Pinnaduwage L, Zhang J, Lye SJ, Geraghty DE, Dunk CE. Definitive class I human leukocyte antigen expression in gestational placentation: HLA-F, HLA-E, HLA-C, and HLA-G in extravillous trophoblast invasion on placentation, pregnancy, and parturition. *Am J Reproductive Immunology*. 2017;77(6):e12643-n/a.
20. Sharkey AM, Gardner L, Hiby S, et al. Killer Ig-like receptor expression in uterine NK cells is biased toward recognition of HLA-C and alters with gestational age. *J Immunol*. 2008;181(1):39-46.
21. Moffett-King A. Natural killer cells and pregnancy. *Nat Rev Immunol*. 2002;2(9):656-663.
22. Hiby SE, Walker JJ, O'Shaughnessy KM, et al. Combinations of maternal KIR and fetal HLA-C genes influence the risk of preeclampsia and reproductive success. *J Exp Med*. 2004;200(8):957-965.
23. Kuśnierczyk P. Killer cell immunoglobulin-like receptor gene associations with autoimmune and allergic diseases, recurrent spontaneous abortion, and neoplasms. *Front Immunol*. 2013;4(8):1-11.
24. Kulkarni S, Martin MP, Carrington M. The Ying and Yang of HLA and KIR in human disease. *Semin Immunol*. 2008;20(6):343-352.
25. Parham P. MHC class I molecules and Kirs in human history, health and survival. *Nat Rev Immunol*. 2005;5(3):201-214.
26. Parham P, Norman PJ, Abi-Rached L, Guethlein LA. Human-specific evolution of killer cell immunoglobulin-like receptor recognition of major histocompatibility complex class I molecules. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1590):800-811.
27. Augusto DG, Petzl-Erler ML. KIR and HLA under pressure: evidences of coevolution across worldwide populations. *Hum Genet*. 2015;134(9):929-940.

28. Meyer D, Thomson G. How selection shapes variation of the human major histocompatibility complex: a review. *Ann Hum Genet*. 2001;65(1):1-26.
29. Castelli EC, Gerasimou P, Paz MA, et al. HLA-G variability and haplotypes detected by massively parallel sequencing procedures in the geographicaly distinct population samples of Brazil and Cyprus. *Mol Immunol*. 2017;83:115-126.
30. Ramalho J, Veiga-Castelli LC, Donadi EA, Mendes-Junior CT, Castelli EC. HLA-E regulatory and coding region variability and haplotypes in a Brazilian population sample. *Mol Immunol*. 2017;91:173-184.
31. Sanchez-Mazas A. An apportionment of human HLA diversity. *Tissue Antigens*. 2007;69(Suppl 1):198-202.
32. Lima THA, Souza AS, Porto IOP, et al. HLA-A promoter, coding, and 3′UTR sequences in a Brazilian cohort, and their evolutionary aspects. *HLA*. 2019;93(2–3):65-79.
33. Meyer D, Single RM, Mack SJ, Erlich HA, Thomson G. Signatures of demographic history and natural selection in the human major histocompatibility complex loci. *Genetics*. 2006; 173(4):2121-2142.
34. Hedrick PW, Thomson G. Evidence for balancing selection at HLA. *Genetics*. 1983;104(3):449-456.
35. Hedrick PW, Whittam TS, Parham P. Heterozygosity at individual amino acid sites: extremely high levels for HLA-A and -B genes. *Proc Natl Acad Sci U S A*. 1991;88(13):5897-5901.
36. Garrigan D, Hedrick PW. Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. *Evolution*. 2003;57(8):1707-1722.
37. Fondevila M, Phillips C, Santos C, et al. Revision of the SNPforID 34-plex forensic ancestry test: assay enhancements, standard reference sample genotypes and extended population studies. *Forensic Sci Int Genet*. 2013;7(1):63-74.
38. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2): 945-959.
39. Phillips C, Salas A, Sánchez JJ, et al. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*. 2007;1(3–4):273-280.
40. Posth C, Nakatsuka N, Lazaridis I, et al. Reconstructing the deep population history of central and South America. *Cell*. 2018;175(5):1185-1197.e1122.
41. Brandt DY, Aguiar VR, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3 (Bethesda)*. 2015;5(5):931-941.
42. Castelli EC, Paz MA, Souza AS, Ramalho J, Mendes-Junior CT. Hla-mapper: an application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. *Hum Immunol*. 2018;79:678-684.
43. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
44. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007;81(5):1084-1097.
45. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16(6): 276-277.
46. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015;43(Database issue):D423-D431.
47. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. *Mol Ecol Resour*. 2010;10(3):564-567.
48. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*. 2005;21(11):2791-2793.
49. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G. PyPop update—a software pipeline for large-scale multilocus population genomics. *Tissue Antigens*. 2007;69 (Suppl 1):192-197.
50. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33(7):1870-1874.
51. Murrell B, Moola S, Mabona A, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol Biol Evol*. 2013;30(5):1196-1205.
52. Kosakovsky Pond SL, Poon AFY, Velazquez R, et al. HyPhy 2.5-a customizable platform for evolutionary hypothesis testing using phylogenies. *Mol Biol Evol*. 2020;37(1):295-299.
53. Kalinowski ST. Hp-rare 1.0: a computer program for performing rarefaction on measures of allelic richness. *Mol Ecol Notes*. 2005;5(1):187-189.
54. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005; 21(2):263-265.
55. Szolek A, Schubert B, Mohr C, Sturm M, Feldhahn M, Kohlbacher O. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics*. 2014;30(23):3310-3316.
56. Rodrigues C, Macedo LC, Bruder AV, et al. Allele and haplotype frequencies of HLA-A, B, C, DRB1 and DQB1 genes in polytransfused patients in ethnically diverse populations from Brazil. *Int J Immunogenet*. 2015;42(5):322-328.
57. González-Galarza FF, Takeshita LY, Santos EJ, et al. Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res*. 2015;43(Database issue):D784-D788.
58. Goeury T, Creary LE, Brunet L, et al. Deciphering the fine nucleotide diversity of full HLA class I and class II genes in a well-documented population from sub-Saharan Africa. *HLA*. 2018;91(1):36-51.
59. Aguiar VRC, César J, Delaneau O, Dermitzakis ET, Meyer D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet*. 2019;15(4):e1008091.
60. Meyer D, C Aguiar VR, Bitarello BD, C Brandt DY, Nunes K. A genomic perspective on HLA evolution. *Immunogenetics*. 2018; 70(1):5-27.
61. Hegde RS. Targeting and beyond: new roles for old signal sequences. *Mol Cell*. 2002;10(4):697-698.
62. Lemberg MK, Bland FA, Weihofen A, Braud VM, Martoglio B. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. *J Immunol*. 2001;167(11): 6441-6446.
63. Borrego F, Ulbrecht M, Weiss EH, Coligan JE, Brooks AG. Recognition of human histocompatibility leukocyte antigen (HLA)-E complexed with HLA class I signal sequence–derived

peptides by CD94/NKG2 confers protection from natural killer cell–mediated lysis. *J Exp Med*. 1998;187(5):813-818.
64. Celik AA, Kraemer T, Huyton T, Blasczyk R, Bade-Döding C. The diversity of the HLA-E-restricted peptide repertoire explains the immunological impact of the Arg107Gly mismatch. *Immunogenetics*. 2016;68(1):29-41.
65. Felicio LP, Porto IOP, Mendes-Junior CT, et al. Worldwide HLA-E nucleotide and haplotype variability reveals a conserved gene for coding and 3′ untranslated regions. *Tissue Antigens*. 2014;83(2):82-93.
66. Veiga-Castelli LC, Castelli EC, Mendes CT, 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*. 2012;79(1):15-24.
67. Grimsley C, Ober C. Population genetic studies of HLA-E: evidence for selection. *Hum Immunol*. 1997;52(1):33-40.
68. Li YD, Xie ZY, Du YL, et al. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene*. 2009;436(1–2):8-11.
69. Fan QR, Wiley DC. Structure of human histocompatibility leukocyte antigen (HLA)-Cw4, a ligand for the KIR2D natural killer cell inhibitory receptor. *J Exp Med*. 1999;190(1):113-123.
70. Bitarello BD, Francisco Rdos S, Meyer D. Heterogeneity of dN/dS ratios at the classical HLA class I genes over divergence time and across the allelic phylogeny. *J Mol Evol*. 2016;82(1):38-50.
71. Buhler S, Nunes JM, Sanchez-Mazas A. HLA class I molecular variation and peptide-binding properties suggest a model of joint divergent asymmetric selection. *Immunogenetics*. 2016;68(6–7):401-416.
72. Salter RD, Benjamin RJ, Wesley PK, et al. A binding site for the T-cell co-receptor CD8 on the alpha 3 domain of HLA-A2. *Nature*. 1990;345(6270):41-46.
73. Wesley PK, Clayberger C, S-c L, Krensky AM. The CD8 coreceptor interaction with the α3 domain of HLA class I is critical to the differentiation of human cytotoxic t-lymphocytes specific for HLA-A2 and HLA-Cw4. *Hum Immunol*. 1993;36(3):149-155.
74. Salter RD, Norment AM, Chen BP, et al. Polymorphism in the α3 domain of HLA-A molecules affects binding to CD8. *Nature*. 1989;338:345-347.
75. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*. 2007;8:69.
76. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*. 2007;27(1):91-105.
77. Majoros WH, Ohler U. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics*. 2007;8:152.
78. Norman PJ, Hollenbach JA, Nemat-Gorgani N, et al. Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLoS Genet*. 2013;9(10):e1003938.
79. Sibilio L, Martayan A, Setini A, et al. A single bottleneck in HLA-C assembly. *J Biol Chem*. 2008;283(3):1267-1274.
80. Hirayasu K, Ohashi J, Kashiwase K, et al. Significant association of KIR2DL3-HLA-C1 combination with cerebral malaria and implications for co-evolution of KIR and HLA. *PLoS Pathog*. 2012;8(3):e1002565.
81. Kaur G, Gras S, Mobbs JI, et al. Structural and regulatory diversity shape HLA-C protein expression levels. *Nat Commun*. 2017;8:15924.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Souza AS, Sonon P, Paz MA, et al. *Hla-C* genetic diversity and evolutionary insights in two samples from Brazil and Benin. *HLA*. 2020;1–19. https://doi.org/10.1111/tan.13996