Unveiling the Diversity of Immunoglobulin Heavy Constant Gamma (IGHG) Gene Segments in Brazilian Populations Reveals 28 Novel Alleles and Evidence of Gene Conversion and Natural Selection

Verónica Calonga-Solís, Danielle Malheiros, Marcia Holsbach Beltrame, Luciana de Brito Vargas, Renata Montoro Dourado, Hellen Caroline Issler, Roseli Wassem, Maria Luiza Petzl-Erler and Danillo G. Augusto

Even though immunoglobulins are critical for immune responses and human survival, the diversity of the immunoglobulin heavy chain gene (IGH) is poorly known and mostly characterized only by serological methods. Moreover, this genomic region is not well-covered in genomic databases and genome-wide association studies due to particularities that impose technical difficulties for its analysis. Therefore, the IGH gene has never been systematically sequenced across populations. Here, we deliver an unprecedented and comprehensive characterization of the diversity of the IGHG1, IGHG2, and IGHG3 gene segments, which encode the constant region of the most abundant circulating immunoglobulins: IgG1, IgG2, and IgG3, respectively. We used Sanger sequencing to analyze 357 individuals from seven different Brazilian populations, including five Amerindian, one Japanese-descendant and one Euro-descendant population samples. We discovered 28 novel IGHG alleles and provided evidence that some of them may have been originated by gene conversion between common alleles of different gene segments. The rate of synonymous substitutions was significantly higher than the rate of the non-synonymous substitutions for IGHG1 and IGHG2 ($p = 0.01$ and 0.03, respectively), consistent with purifying selection. Fay and Wu’s test showed significant negative values for most populations ($p < 0.001$), which indicates that positive selection in an adjacent position may be shaping IGHG variation by hitchhiking of variants in the vicinity, possibly the regions that encode the Ig variable regions. This study shows that the variation in the IGH gene is largely underestimated. Therefore, exploring its nucleotide diversity in populations may provide valuable information for comprehension of its evolution, its impact on diseases and vaccine research.

Keywords: IGHG genes, immunoglobulin heavy chain, molecular characterization, genetic diversity, populations, DNA sequencing
INTRODUCTION

Immunoglobulins (Ig) are glycoproteins produced exclusively by activated B-lymphocytes and plasma cells that mediate humoral response against pathogens. Each B-cell clone presents an antigen-specific membrane-bound immunoglobulin that, together with CD79A and CD79B molecules, comprise the B-cell receptor (BCR). After stimulation by antigens, B-cells secrete immunoglobulins (antibodies) with the same antigen-binding sites than the membrane-bound molecules.

All Ig share a similar basic structure composed of four polypeptide chains: two identical heavy chains and two identical light chains. The heavy chain has one variable domain (V_H) and three or four constant domains (C_H1, C_H2, C_H3, and C_H4). Each light chain exhibits one variable domain (V_L) and one constant domain (C_L). The variable region (V_H and V_L) is responsible for antigen recognition and binding while the constant regions (C_H and C_L) primarily mediate the Ig effector functions, which includes complement activation and Fc Receptor binding (1, 2).

In humans, the immunoglobulin heavy chain gene (IGH) is located in chromosome region 14q32 and consists of four groups of gene segments: the variable heavy (IGHV), diversity heavy (IGHD), joining heavy (IGHJ), and constant heavy (IGHC). The IGH group includes IGHM, IGHD, IGHG3, IGHG1, IGHEP1,IGHA1, IGHGP, IGHG2, IGHG4, IGHG, and IGH2A gene segments and pseudogenes. Immunoglobulin light chain genes are sequenced at the nucleotide level across populations. Thus, the diversity of these gene segments is probably underestimated.

Additionally, this genomic region is not well-covered in genome-wide studies and genomic databases for two reasons: first, DNA samples used are often extracted from B-cell lines, which are not suitable for analyzing this region due to the somatic rearrangement within this locus; second, the high sequence similarity of these segments imposes technical difficulties for sequencing and genotyping (12).

Here, we analyzed the diversity of IGHG1, IGHG2, and IGHG3 in seven Brazilian populations: five Amerindian populations that have been genetically isolated for centuries and two urban populations. By analyzing deep sequencing data, we found 28 novel IGHG alleles, characterized the linkage disequilibrium of variants within these segments and analyzed the relationship among alleles. Additionally, we provided compelling evidence of the occurrence of gene conversion between different gene segments and evidence of purifying selection shaping IGHG diversity.

METHODS

Characterization of the Study Populations

This study was approved by the Brazilian National Human Research Ethics Committee (CONEP), protocol number CAAE 02727412.4.0000.0096, in accordance to the Brazilian Federal laws. We analyzed a total of 357 individuals from seven Brazilian populations, of which five are Amerindian: Guarani Kaiowa (GKW, n = 46), Guarani Nandeve (GND, n = 48), Guarani Mbya (GRC, n = 51), Kaingang from Ivaí (KIV, n = 52), and Kaingang from Rio das Cobras (KRC, n = 52); and two are urban populations: Japanese-descendants (BrJAP, n = 57) and Euro-descendants from Curitiba (CTBA, n = 51). Their detailed geographic location and sample sizes are found in Figure 1 and Table S1.

The Amerindians samples were collected between late 1980s and early 1990s. According to public data from the Brazilian Institute of Geography and Statistics (IBGE), there are approximately 900,000 Amerindians individuals in Brazil, distributed across 693 official indigenous lands (https://www.ibge.gov.br). The Guarani speak a Tupi-Guarani language, which belongs to the Tupi language family. The Kaingang speak Jê, which belongs to the Macro-Jê language stock. Analyzing mtDNA segments and the proposed time of origin of Tupi-Guarani and Jê linguistic families, Marrero and colleagues (13) estimated the Guarani population split in three partialities (Guarani Kaiowa, Guarani Nandeve and Guarani Mbya) 1,800 years ago, while the different Kaingang populations would have split more recently, around 200 years ago. Since then, they are believed to have remained isolated from each other and other urban populations due to strong cultural and language barriers (14). A former study from our group estimated that the gene flow of these Amerindian populations with non-Amerindians was low, being 0% in Guarani Kaiowa, 4% in Guarani Mbya, 14% in Guarani Nandeve and 7% in Kaingang (15).

The two urban samples were from Curitiba, the capital of Paraná State and the 5th largest city in Brazil. As a result of the Brazilian history of European colonization 500 years ago, and especially the more recent European migrations since the
19th Century, the population of Curitiba is of predominantly European ancestry. According to the public data from IBGE, 78.7% of the inhabitants of Curitiba self-declared themselves as white, 16.7% as admixed, 3% as black, 1.4% as Asian, and 0.2% as Amerindian (https://www.ibge.gov.br). The population here referred as CTBA only included Euro-descendant individuals. Therefore, we excluded all individuals with known miscegenation with Amerindian and/or other non-European ancestries. Paraná State also hosts the second largest Japanese community in Brazil, one of the largest outside Japan. The Japanese migration started in the twentieth century with the Treaty of Friendship, Commerce and Navigation between Brazil and Japan. All Japanese descendant individuals of this study (BrJAP) were born in Brazil, with either both parents or all four grandparents born in Japan. They reported no history of admixture with non-Japanese ethnicities.

DNA Extraction
Genomic DNA was extracted from peripheral blood samples by standard salting-out (16) or by the phenol-chloroform-isoamyl method (17). High-quality DNA has been stored at −80°C since the extraction. DNA integrity was evaluated by 1% agarose gel electrophoresis and purity was accessed by spectrophotometry.

Sequencing and Allele Identification
We aligned all previously known IGHG alleles and designed primers to amplify each segment specifically. To define the best set of primers we used the following approaches: (i) we ruled out unspecific amplification by verifying that all amplicons did not exhibit any variant that was specific of other segments; (ii) we certified that the genotype distribution of all single nucleotide variable sites, in each amplicon, were in accordance to Hardy Weinberg equilibrium (p > 0.05).

Polymerase chain reaction (PCR) was performed for IGHG1, IGHG2, and IGHG3 as follows: 1X Buffer (Invitrogen); 0.2 mM dNTP (Life Technologies); 1.5 mM MgCl₂ (Invitrogen, Carlsbad, CA, USA); 0.3 μM of each primer; 0.05 U/µL Taq polymerase Platinum® (Invitrogen, Carlsbad, CA, USA); and 2 ng/µL genomic DNA. The segments were amplified in a Mastercycler ep Gradient S thermocycler (Eppendorf, Hamburg, Germany), with a first step at 94°C for 2 min and 10 cycles of 94°C for 15 s, TₘA°C for 15 s and 72°C for 60 s, followed by 25 cycles of 94°C for 15 s, TₘB°C for 15 s and 72°C for 60 s, with a final extension step of 72°C for 60 s (primer sequences, location, and amplification temperatures are available in Table S2 and Figure S1). Amplicons were visualized by 1% agarose gel electrophoresis with 1% UniSafe Dye® (Uniscience, Sao Paulo, Brazil). Afterwards, PCR products were purified with 0.8 U/µL of exonuclease I enzyme (Fermentas, Waltham, MA, EUA) and 0.14 U/µL of alkaline phosphatase (ThermoFisher Scientific, Waltham, MA, EUA).

Sequencing was performed using Big Dye® Terminator Cycle Sequencing Standard v3.1 (Life Technologies, Carlsbad, CA, USA), according to manufacturer’s instructions. The sequencing reactions were performed in a Mastercycler ep Gradient S thermocycler (Eppendorf, Hamburg, Germany) with a first step at 95°C for 60 s and 25 cycles of 95°C for 10 s, 50°C for 5 s, and 60°C for 4 min, followed by capillary electrophoresis in a 350xl Genetic Analyzer Sequencer (Life Technologies, Carlsbad, CA, USA).

After sequencing, the alleles were identified according to the known alleles described at IMGT database (International ImMunoGeneTics Information System) (18). IMGT database provides public access to an integrated information system specialized in immunoglobulins (Ig), T cell receptors (TCR), and major histocompatibility complex (MHC) genes and molecules. All data submitted to the IMGT database are manually checked by experts in the field, which assure the deposit of high-quality data.

The nucleotide sequence of each individual was aligned with consensus sequences with Mutation Surveyor® DNA Variant Analysis Software v5.0.1 (Softgenetics), and their variable sites were annotated. Alleles that were different from the ones listed in the IMGT database were considered novel and were subsequently confirmed by sequencing and/or molecular cloning as described below.

The novel alleles that were observed in homozygosis (IGHG1*07, IGHG1*08, IGHG2*09, IGHG2*13, IGHG3*21, IGHG3*22, IGHG3*26) were confirmed by direct re-sequencing from a different PCR product. Novel alleles observed in
heterozygosis without phasing ambiguities due to the presence of only one heterozygous position (IGHG1*06, IGHG1*09, IGHG1*10, IGHG1*12, IGHG1*13, IGHG1*14, IGHG2*07, IGHG2*10, IGHG2*12, IGHG3*20, IGHG3*27, IGHG3*28) were also confirmed by re-sequencing. The new variants with ambiguous phasing (IGHG1*11, IGHG2*08, IGHG2*11, IGHG2*14, IGHG2*15, IGHG3*23, IGHG3*24, IGHG3*25, IGHG3*29) were confirmed by molecular cloning. In this case, the segments were re-amplified and ligated into a PTZ57R/T vector (Fermentas, Waltham, MA, EUA) with terminal deoxynucleotidyl transferase (TdT) enzyme. Afterwards, recombinant plasmids were obtained and purified from multiple transformed colonies and sequenced as described above. Novel alleles were verified based on sequences from at least two independent colonies containing each allele.

Data Analysis

Allelic frequencies were obtained by direct counting using GenALex v6.502 software (19). Hardy-Weinberg equilibrium was tested for each gene segment in all populations by Guo and Thompson’s method (20), performed in Arlequin v3.5.2 software (21). IGHG haplotypes from different gene segments were estimated via ELB algorithm and this information was used for Gm allotype haplotype inference, according to the correspondence between nucleotide variants and allotypes described by Lefranc et al. (6). Linkage disequilibrium (LD) between single nucleotide variants of each gene segment was estimated with Hapioview software (22).

Allele networks were performed with variants from each gene segment through the median-joining (MJ) algorithm (23) with Network v5.0 software. Allele frequencies were compared using the exact test of population differentiation (24) and population-pairwise FST (25, 26) with Arlequin v3.5.2 software (21). Principal component analysis (PCA) was performed using the Minitab 17 Statistics Software (27) for graphical representation of the genetic differences and similarities in the major components of variation among populations. The PCA was performed using inferred allotype haplotype frequencies to compare the frequencies from the study population with others that were previously described serologically. These haplotypes were classified according to Lefranc et al. (6), and detailed information is available in Table S3.

Neutrality tests were performed using the Tajima’s D (28), Fu and Li’s D*, F*, D, and F (29) and Fay and Wu’s H (30) in DnaSP software (31). Homologous gene segments from rhesus monkey were used as outgroup (Macaca mulatta; accession number: NW_001121238, AY292519, AY292512).

RESULTS

One Novel Single Nucleotide Variant and 28 Novel IGHG Alleles Have Been Discovered

Within all three gene segments in the seven populations analyzed, we found a total of 49 exonic variable sites, of which 26 were non-synonymous substitutions. Based on the Grantham scale (32), which ranges from 5 to 215 according to the physicochemical distance between amino acid pairs, amino acid replacements were from low to moderate (15 < D < 103) (Table 1). Of the single variable sites, 21 have not been reported in any of the previously described alleles at the IMGT database (Table 1, in bold). We also found a novel synonymous IGHG3 single nucleotide variant at the position chr14:106235856 (GRCh37.p13 primary assembly) in the CTBA population. This new variant was submitted to the dbSNP database (34) under reference SNP ID number rs155533833 (NC_000014.8:g.106235856G>A).

A total of 28 novel IGHG alleles have been found in our study: nine in IGHG1 (Table 2), nine in IGHG2 (Table 3), and ten in IGHG3 (Table 4). All novel alleles have been confirmed either by sequencing or by molecular cloning followed by sequencing. Novel alleles have been submitted to IMGT Nomenclature Committee (18), which verified the accuracy of our data and assigned official names (reports #2018-2-0824 and #2018-5-1113).

Interestingly, some new alleles of all gene segments were observed at high frequency (f > 0.10; Table 5). The highest frequencies for novel alleles were observed for IGHG1*07 in GKW (f = 0.478; 34 individuals), IGHG1*08 in BrJAP (f = 0.155; 15 individuals), IGHG2*08 in BrJAP (f = 0.202; 23 individuals), IGHG2*09 in GRC (f = 0.137; 9 individuals), IGHG3*21 in BrJAP (f = 0.158; 16 individuals), and IGHG3*22 in GRC (f = 0.157; 15 individuals).

Because most of the previous studies only described the immunoglobulin heavy chain diversity serologically, we inferred the serological Gm allotypes from our nucleotide sequence data, based on the nucleotide sequence description for each previously reported allotype (6), to allow comparison with previously reported variants. For example, the most frequent allele haplotype (alleles that are in the same chromosome and inherited together in a block) was the one comprising the gene segments IGHG1*02, IGHG2*03, IGHG3*14 (f = 0.182 to 0.740), which encodes the Gm haplotype “C” Gm21,26,27,28;17,1;(.), the most frequent IgG allotype haplotype in our populations (f = 0.21 to 0.77; Table 6). The correspondence between allele haplotype and allotype haplotypes are in the Table S4. More than one IGHG allele haplotype can define a single Gm allele haplotype, as is the case of the Gm haplotype “B” Gm5,10,11,13,14,26,27,28;17,1;(.), that is encoded in our data by the allele haplotype IGHG3*11,IGHG2*03,IGHG2*05, by IGHG3*11,IGHG1*14,IGHG2*03, or by IGHG3*11,IGHG1*03,IGHG2*08. In order to simplify the interpretation of the data, Gm haplotype identifiers (from A to M) were used as suggested by Lefranc et al. (6).

Lower IGHG Diversity Was Observed in Amerindians and Frequencies Differed Significantly Among Populations

IGHG allelic frequencies varied across populations (Table 5). A small number of highly frequent alleles were observed for all gene segments in Amerindian populations. Even though Guarani populations share a more recent common ancestor, allelic frequencies significantly differed among them (p < 0.01), with low to moderate FST values (0.02–0.10) (Table 7). Allelic frequencies did not differ between the two Kaingang populations.
| Gene segment | Exon | rsID | Location<sup>a</sup> | IMGT numbering<sup>b</sup> | Eu numbering<sup>c</sup> | Nucleic acid substitution | Amino acid substitution | Allotype | Grantham's D<sup>d</sup> | Frequency<sup>e</sup> |
|---------------|------|------|----------------------|----------------------|----------------------|----------------------|----------------------|---------|----------------------|------------------|
| IGHG1        | CH1  | rs11552998 | 106209340 | 19 | 140 | G>A | 0.006 |
|               |      | rs17850096  | 106209289 | 40 | 157 | G>C | 0.001 |
|               |      | rs1071803 | 106209119 | 120 | 214 | A>G | 0.143 |
|               |      | rs587690960 | 106208471 | 22 | 260 | A>G | 0.003 |
|               |      | rs77538050 | 106208364 | 84.3 | 296 | A>T | 0.001 |
|               |      | rs193160354 | 106208327 | 91 | 308 | C>T | 0.011 |
|               |      | rs1043109 | 106208326 | 92 | 309 | C>G | 0.011 |
|               |      | rs1043249 | 106208306 | 98 | 315 | T>C | 0.011 |
| CH3          |      | rs11557940 | 10620817 | 5 | 349 | C>T | 0.001 |
|               |      | rs1045583 | 106209086 | 12 | 356 | T>G | 0.118 |
|               |      | rs11621259 | 106209082 | 14 | 358 | C>A | 0.023 |
|               |      | rs17841087 | 106207933 | 86 | 407 | C>T | 0.114 |
|               |      | rs11904727 | 106207862 | 110 | 431 | C>G | 0.020 |
|               |      | rs370028332 | 106207858 | 112 | 432 | G>C | 0.001 |
|               |      | rs3011686 | 106207843 | 117 | 437 | G>A | 0.003 |
|               |      | rs12879979 | 106207822 | 124 | 444 | T>C | 0.106 |
|               |      | rs113679609 | 106110056 | 92 | 308 | G>C | 0.163 |
|               |      | rs587682450 | 106109825 | 9 | 353 | A>C | 0.017 |
|               |      | rs4983499 | 106109752 | 38 | 378 | G>T | 0.003 |
|               |      | rs368359789 | 106109708 | 79 | 392 | G>C | 0.054 |
|               |      | rs1049810 | 106110037 | 91 | 308 | T>C | 0.016 |
|               |      | rs113679609 | 106110005 | 92 | 309 | G>C | 0.003 |
|               |      | rs149653267 | 106207822 | 124 | 444 | T>C | 0.106 |
|               |      | rs113679609 | 106109752 | 38 | 378 | G>T | 0.003 |
|               |      | rs12879979 | 106109825 | 9 | 353 | A>C | 0.017 |
|               |      | rs4983499 | 106109752 | 38 | 378 | G>T | 0.003 |
|               |      | rs368359789 | 106109708 | 79 | 392 | G>C | 0.054 |
|               |      | rs1049810 | 106110037 | 91 | 308 | T>C | 0.016 |

<sup>a</sup> Coordinate at chromosome 14 location (GRCh37.p13 primary assembly).
<sup>b</sup> Novel single nucleotide variation submitted to dbSNP database.
<sup>c</sup> Amino acid position according to IMGT database (International ImMunoGeneTics Information System) (15).
<sup>d</sup> Physicochemical distances between amino the amino acids involved in the substitution, according Grantham (32).
<sup>e</sup> Frequency of the alternative allele, merging all the samples of this study.
IGHG1 alleles previous described and the 9 novel IGHG1 alleles identified in this study.

| Allele name | GenBank accession number | Allotype | # |
|-------------|--------------------------|----------|---|
| IGHG1*01    | -                        | 17,1     | 0 |
| IGHG1*02    | -                        | 17,1     | 391 |
| IGHG1*03    | -                        | 3        | 81 |
| IGHG1*04    | -                        | 17,1,27  | 0 |
| IGHG1*05    | -                        | 17,1     | A  |
| IGHG1*06    | MG920252                 | 3        | 2 |
| IGHG1*07    | MG920245                 | 17,1,2   | 189 |
| IGHG1*08    | MG920246                 | 3,1      | 18 |
| IGHG1*09    | MG920247                 | 17,1     | 2 |
| IGHG1*10    | MG920248                 | 17,1     | 2 |
| IGHG1*11    | MG920249                 | 17,1     | 7 |
| IGHG1*12    | MG920250                 | 17,1     | 1 |
| IGHG1*13    | MG920251                 | 17,1     | 1 |
| IGHG1*14    | MG920253                 | 17,1     | 1 |

Novel alleles (in bold) have been confirmed by sequencing and/or molecular cloning. Their official names have been assigned by IMGT nomenclature committee. IMGT, International ImMunoGeneTics Information System. aAccording to Edelman et al. (33). bAllotypes were inferred according to Lefranc et al. (6). cNumber of copies observed in this study.

(p = 0.065; Fs = 0.03). More conspicuous differences were found between the Japanese-descendant and Euro-descendant populations compared to each other, and between each of these two populations compared to the Amerindian populations, with Fs values ranging from 0.11 to 0.52, indicating moderate to high genetic differentiation.

The principal component analysis (PCA) grouping was consistent with ancestry and geography (Figure 2). Amerindi ans and Asians formed two separated groups close to each other. Europeans and admixed populations of predominantly European ancestry grouped together, while Africans were more distant.

Genotypic distributions for all gene segments were in accordance with Hardy-Weinberg equilibrium in all population samples (0.08 > p > 1).

Distinct Linkage Disequilibrium Patterns Among Populations

Linkage disequilibrium (LD) patterns differed among populations (Figure S2). Interestingly, each Guarani population exhibited a distinct LD pattern despite their close relationship. In GKW, only five variable sites were observed in all three gene segments, of which three were in absolute LD (D = 1, r2 = 1). In contrast, more variable sites (21 and 24) were observed for the other two Guarani populations. In addition, many variants that were in LD in GND were not observed in LD in GRC.

The G1m3 allotype (rs1071803) and the G2m23 allotype (rs8009156) were in strong LD in all Amerindian populations (D = 1; r2 > 0.87), as well as in CTBA (D = 1; r2 = 0.43), and BrJAP (D = 0.73; r2 = 0.92) in which fewer SNPs were observed in strong LD.

**Sequence Analysis Suggests That Gene Conversion Between Frequent Alleles of Different Gene Segments Generated Novel Alleles**

Median-Joining network (Figure 3) shows that the most frequent alleles IGHG1*01, IGHG2*03, and IGHG3*14 were central nodes in the network, with few nucleotides differing between them and the other alleles. The loops indicate possible recombination sites.

Alignment of all the known alleles of the IGHG1, IGHG2, and IGHG3 gene segments suggests that some novel alleles...
TABLE 3 | IGHG2 alleles previously described and the 9 novel IGHG2 alleles identified in this study.

| Allele name | GenBank accession number | Allotypea | IMGT unique numbering | Eu numberingb | Amino acid change | Exonic position | Consensus nucleotide | # |
|-------------|--------------------------|-----------|-----------------------|---------------|------------------|-----------------|---------------------|---|
| IGHG2*01    |                         | (-)       | 15 16 17 19 82 92 95 96 45.1 91 92 9 38 79 81 117 |               |                  |                 |                     |    |
| IGHG2*02    | MH025828                 | (-)       | 23                    |               |                  |                 |                     |    |
| IGHG2*03    | MH025827                 | (-)       | 56                    |               |                  |                 |                     |    |
| IGHG2*04    | MH025829                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*05    | MH025830                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*06    | MH025831                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*07    | MH025832                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*08    | MH025833                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*09    | MH025834                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*10    | MH025835                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |
| IGHG2*11    | MH025836                 | (-)       | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | S = G S = G A | P P > T N > S F > L V > L P A > S K > N T T | 136 137 138 140 171 189 192 193 282 308 309 353 378 392 394 437 | C A A C A C A C C G T G A G G A G | 0 |

Novel alleles (in bold) have been confirmed by sequencing and/or molecular cloning. Their official names have been assigned by IMGT nomenclature committee. IMGT, International ImMunoGeneTics Information System (18). Dots represent the consensus nucleotide.

aAccording to Edelman et al. (13).

bAlyotypes were inferred according to Lefranc et al. (6).

Number of copies observed in this study.

discovered in this study could have been generated by gene conversion between alleles of different gene segments (Figure 3).

For example, the novel allele IGHG1*11, present in BrJAP (f = 0.064), could have been generated by gene conversion between the most frequent IGHG2 allele (IGHG2*03; f = 0.579) and the most frequent IGHG1 allele (IGHG1*02; f = 0.60). In addition, gene conversion between the frequent IGHG2*03 and IGHG3*14 alleles (f = 0.735 and f = 0.833, respectively) could explain the origin of allele IGHG2*09 (f = 0.14).

**Neutrality Tests Suggest Evidence of Natural Selection Shaping IGHG Polymorphism**

Neutrality tests performed by Tajima’s D, Fu and Li’s D and F were non-significant for most populations. However, Fay and Wu’s test resulted in significant negative values for most populations, which may indicate positive selection at an adjacent site (Table 8).

Deviation of neutrality was also tested by analyzing synonymous and non-synonymous substitution rates across all the known and novel alleles of all gene segments (Tables S5–S7). Overall, the rate of synonymous substitutions (dS) was significantly higher than the rate of non-synonymous (dN) substitutions (dN/dS < 1) for IGHG1 and IGHG2 (p = 0.01 and 0.032, respectively) (Table 9), consistent with purifying selection.

**DISCUSSION**

Our main goal was to deliver an unprecedented and comprehensive nucleotide sequencing-based characterization of the IGHG gene segments in populations of different ancestries. Before this study, only 30 IGHG alleles have been described for IGHG1, IGHG2, and IGHG3 together (18). Here, we report the discovery of 28 novel alleles, of which 16 were in a single population sample of Japanese descendants (n = 57) and seven in one population sample of Euro-descendants (n = 51). It is interesting that even in Amerindian populations, which exhibited a limited diversity, seven new alleles were found. This is clear evidence that the diversity of IGHG is far from being fully described and possibly a much larger number of novel alleles will be discovered as more populations are
interrogated. We focused on the segments that code for the most abundant Ig in serum. Considering the homology and high sequence similarity, a different strategy would be needed for the precise characterization of \( \text{IGHG4} \) due to the high frequency of duplications observed for this gene segment \((35)\).

Some of the new alleles were highly frequent. The novel allele \( \text{IGHG3}^*22 \), frequent in Guarani Mbya (GRC, \( f = 0.157 \)), exhibited a lower frequency in Guarani Ñandeva (GND, \( f = 0.063 \)), and was absent in Guarani Kaiowa (GKW). These three populations share a more recent common ancestor and the differences observed can be explained by its demographic history and genetic drift. Demographic factors played a major role in shaping the diversity of other genes important for immune responses in these same Amerindian populations \((36)\). Genetic drift, particularly founder effect and bottleneck, may explain the lower diversity of \( \text{IGHG} \) in Amerindians and the fluctuation of their allelic frequencies. On the other hand, the \( \text{IGHG3}^*22 \) allele was observed only in one Kaingang individual. This fact suggests

### TABLE 4 | \( \text{IGHG3} \) alleles previous described and the 10 novel \( \text{IGHG3} \) alleles identified in this study.

| Allele name | GenBank accession number | Allotype<sup>b</sup> | # |
|-------------|--------------------------|----------------------|---|
| \( \text{IGHG3}^*01 \) | 5,10,11,13,14,26,27 | . . . . . . . . . . . . | 2 |
| \( \text{IGHG3}^*03 \) | 5,6,11,24,26 | . . . . . . abs | 0 |
| \( \text{IGHG3}^*04 \) | 5,10,11,13,14,26,27 C | . . . . . . abs abs | 0 |
| \( \text{IGHG3}^*05 \) | 5,10,11,13,14,26,27 | . . . . . . | 0 |
| \( \text{IGHG3}^*06 \) | 5,10,11,13,14,26,27 A | . . . . . . | 0 |
| \( \text{IGHG3}^*07 \) | 5,10,11,13,14,26,27 | . . . . . . | 0 |
| \( \text{IGHG3}^*08 \) | 5,14,26,27 | . . . . . . | 0 |
| \( \text{IGHG3}^*09 \) | 5,10,11,13,14,26,27 | . . . . . . G | 0 |
| \( \text{IGHG3}^*10 \) | 5,10,11,13,14,26,27 A | . . . . . . | 1 |
| \( \text{IGHG3}^*11 \) | 5,10,11,13,14,26,27 | . . . . . . T | 76 |
| \( \text{IGHG3}^*12 \) | 5,10,11,13,14,26,27 abs A C . . . . T | 7 |
| \( \text{IGHG3}^*13 \) | 5,6,10,11,14,26,27 | . . . . . . | 0 |
| \( \text{IGHG3}^*14 \) | 21,26,27,28 | . . . . . . T | 529 |
| \( \text{IGHG3}^*15 \) | 21,26,27,28 | . . . . . . T | 1 |
| \( \text{IGHG3}^*16 \) | 21,26,27,28 | . . . . . . T | 5 |
| \( \text{IGHG3}^*17 \) | 10,11,13,15,27 | . . A C abs G | 0 |
| \( \text{IGHG3}^*18 \) | 10,11,13,15,16,27 | . . . . . . abs G | 0 |
| \( \text{IGHG3}^*19 \) | 10,11,13,15,16,27 | . . . . . . abs G | 31 |
| \( \text{IGHG3}^*20 \) | MG920256 | 21,26,27,28 | . . . . . . T | 2 |
| \( \text{IGHG3}^*21 \) | MG920255 | 21,26,27,28 5,10,11,13,14,26,27 A | 18 |
| \( \text{IGHG3}^*22 \) | MG920254 | 21,27 | . . . . . . T | 26 |
| \( \text{IGHG3}^*23 \) | MH025837 | 10,11,13,16,27 abs G | . . . . . . T | 1 |
| \( \text{IGHG3}^*24 \) | MG920257 | 26,27,28 | . . . . . . | 2 |
| \( \text{IGHG3}^*25 \) | MG920258 | 21,26,27,28 T A | . . . . . . | 4 |
| \( \text{IGHG3}^*26 \) | MG920259 | 5,10,11,13,14,26,27 | . . . . . . T | 4 |
| \( \text{IGHG3}^*27 \) | MG920260 | 26,27,28 | . . . . . . T | 1 |
| \( \text{IGHG3}^*28 \) | MG786813 | 5,10,11,13,14,26,27 . . . . . . A C | . . . . . . T | 1 |
| \( \text{IGHG3}^*29 \) | MG920261 | 21,26,27,28 | . . . . . . T | 1 |

(Continued)
TABLE 4 | Continued

| Exon  | CH3 | IMGT unique numbering | Eu numbering$^a$ | Amino acid change | Exonic position | Consensus nucleotide | Allele name | GenBank accession number | Allotype$^b$ | # |
|-------|-----|----------------------|------------------|------------------|-----------------|---------------------|-------------|-------------------------|-------------|---|
| 1,4   | 4   | 10                   | 39               | 44               | 45,2            | 79                  | 81          | 84                      | 88          | 89 | 90 | 98 | 100 | 101 | 115 | 116 |
| V     | S   | S                    | V                | V                | V, M, S        | S                   | P, R        | N, K                   | SIN         | M, V | K, R | SIN | I, V | R, H | F | Y |
| 1,4   | 4   | 10                   | 39               | 44               | 45,2            | 79                  | 81          | 84                      | 88          | 89 | 90 | 98 | 100 | 101 | 115 | 116 |
| V     | S   | S                    | V                | V                | V, M, S        | S                   | P, R        | N, K                   | SIN         | M, V | K, R | SIN | I, V | R, H | F | Y |
| Exon  | CH3 | IMGT unique numbering | Eu numbering$^a$ | Amino acid change | Exonic position | Consensus nucleotide | Allele name | GenBank accession number | Allotype$^b$ | # |
| 1,4   | 4   | 10                   | 39               | 44               | 45,2            | 79                  | 81          | 84                      | 88          | 89 | 90 | 98 | 100 | 101 | 115 | 116 |
| V     | S   | S                    | V                | V                | V, M, S        | S                   | P, R        | N, K                   | SIN         | M, V | K, R | SIN | I, V | R, H | F | Y |
| Exon  | CH3 | IMGT unique numbering | Eu numbering$^a$ | Amino acid change | Exonic position | Consensus nucleotide | Allele name | GenBank accession number | Allotype$^b$ | # |
| 1,4   | 4   | 10                   | 39               | 44               | 45,2            | 79                  | 81          | 84                      | 88          | 89 | 90 | 98 | 100 | 101 | 115 | 116 |
| V     | S   | S                    | V                | V                | V, M, S        | S                   | P, R        | N, K                   | SIN         | M, V | K, R | SIN | I, V | R, H | F | Y |

Novel alleles (in bold) have been confirmed by sequencing and/or molecular cloning. Their official names have been assigned by IMGT nomenclature committee. IMGT, International ImMunoGeneTics Information System (18). Dots represent the consensus nucleotide; abs, absent.

$^a$According to Edelman et al. (33).

$^b$Allotypes were inferred according to Lefranc et al. (6).

$^#$Number of copies observed in this study.

gene flow from Guarani to Kaingang. Although GRC and KRC remain isolated due to strong cultural barriers, their immediate vicinity did result in a low degree of admixture (14).

IGHG3*11 is the most common IGHG3 allele in Euro-descendants (CTBA, $f = 0.588$) and was observed at lower frequency in Amerindians: GND ($f = 0.094$), GRC ($f = 0.010$), KIV ($f = 0.038$), KRC ($f = 0.020$), being absent in GKW. This allele corresponds to the allotype G3m5,10,11,13,14,26,27 which has been previously shown to be highly frequent in Europeans but absent in non-admixed Amerindians (37–42). Also, similar allele distribution was observed for IGHG1*03 in the study populations. These observations are consistent with previous studies from our group, which estimated the admixture rate of Guarani and Kaingang by analyzing HLA class II genes. In
TABLE 5 | One third of the novel IGHG alleles were observed in high frequencies (0.05 < \( f \) < 0.48).

| Sample size | GKW | GND | GRC | KIV | KRC | BrJAP | CTBA |
|-------------|-----|-----|-----|-----|-----|-------|------|
| **IGHG1**  |     |     |     |     |     |       |      |
| *02         | 0.522 | 0.438 | 0.725 | 0.770 | 0.606 | 0.600 | 0.228 |
| *03         | 0.094 | 0.020 | 0.020 | 0.029 |     | 0.707 |      |
| *06         |     |     |     |     |     | 0.022 |      |
| *08         | 0.478 | 0.469 | 0.235 | 0.200 | 0.365 | 0.136 | 0.033 |
| *09         |     |     |     |     |     | 0.155 |      |
| *10         |     |     |     |     |     | 0.018 |      |
| *11         |     |     |     |     |     | 0.064 |      |
| *12         |     |     |     |     |     | 0.011 |      |
| *13         |     |     |     |     |     | 0.009 |      |
| *14         |     |     |     |     |     | 0.010 |      |

| Sample size | 46  | 48  | 51  | 50  | 52  | 57   | 47   |
|-------------|-----|-----|-----|-----|-----|------|------|
| **IGHG2**  |     |     |     |     |     |      |      |
| *02         |     |     |     |     |     | 0.083 | 0.037 |
| *03         | 0.957 | 0.885 | 0.735 | 0.942 | 0.952 | 0.579 | 0.489 |
| *06         |     |     |     |     |     | 0.021 |      |
| *07         | 0.033 |     |     |     |     | 0.010 |      |
| *08         |     |     |     |     |     | 0.019 | 0.009 |
| *09         | 0.011 | 0.010 |     | 0.137 |     |      |      |
| *10         |     |     | 0.021 |     |     | 0.098 |      |
| *11         |     |     |     |     |     | 0.009 |      |
| *12         |     |     |     |     |     | 0.018 | 0.011 |
| *13         |     |     |     |     |     | 0.061 | 0.011 |
| *14         |     |     |     |     |     | 0.009 |      |
| *15         |     |     |     |     |     | 0.009 |      |

| Sample size | 46  | 48  | 51  | 52  | 52  | 57   | 51   |
|-------------|-----|-----|-----|-----|-----|------|------|
| **IGHG3**  |     |     |     |     |     |      |      |
| *01         |     |     |     |     |     | 0.020 |      |
| *02         |     |     |     |     |     | 0.010 |      |
| *03         | 0.094 | 0.010 |     | 0.038 | 0.020 | 0.588 |      |
| *04         |     |     | 0.010 |     |     | 0.069 |      |
| *05         | 1.000 | 0.844 | 0.833 | 0.952 | 0.971 | 0.474 | 0.186 |
| *06         |     |     |     |     |     | 0.009 |      |
| *07         |     |     |     |     |     | 0.010 | 0.039 |
| *08         |     |     |     |     |     | 0.272 |      |
| *09         |     |     |     |     |     | 0.018 |      |
| *10         |     |     |     |     |     | 0.158 |      |
| *11         |     |     |     |     |     |     |      |
| *12         |     |     |     |     |     | 0.322 |      |
| *13         |     |     |     |     |     | 0.018 |      |
| *14         |     |     |     |     |     | 0.018 |      |
| *15         |     |     |     |     |     | 0.035 |      |
| *16         |     |     |     |     |     | 0.009 |      |
| *17         |     |     |     |     |     | 0.039 |      |
| *18         |     |     |     |     |     | 0.010 | 0.010 |
| *19         |     |     |     |     |     | 0.010 |      |
| *20         |     |     |     |     |     | 0.009 |      |
| *21         |     |     |     |     |     | 0.010 |      |
| *22         |     |     |     |     |     | 0.009 |      |
| *23         |     |     |     |     |     | 0.010 |      |
| *24         |     |     |     |     |     | 0.010 |      |
| *25         |     |     |     |     |     | 0.010 |      |
| *26         |     |     |     |     |     | 0.010 |      |
| *27         |     |     |     |     |     | 0.010 |      |
| *28         |     |     |     |     |     | 0.010 |      |
| *29         |     |     |     |     |     | 0.010 |      |

Novel alleles (in bold) have been confirmed by sequencing and/or molecular cloning. Their official names have been assigned by IMGT nomenclature committee. IMGT, International ImMunoGeneTics Information System (18); KIV, Kaingang from Ivaí; KRC, Kaingang from Rio das Cobras; GRC, Guarani Mbyas; GKW, Guarani Kalowa; GND, Guarani Ñandeva; BrJAP, Japanese-descendant from Curitiba; CTBA, Euro-descendants from Curitiba.
TABLE 6 | Gm allotype haplotypes frequencies inferred from nucleotide sequencing.

| Sample size | GKW | GND | GRC | KIV | KRC | BrJAP | CTBA |
|-------------|-----|-----|-----|-----|-----|-------|------|
| ID\(^a\)    |     |     |     |     |     |       |      |
| A           | 0.083 | 0.01 | 0.02 | 0.02 | 0.444 |       |      |
| B           | 0.01  | 0.277 |     |     |       |       |      |
| C           | 0.522 | 0.375 | 0.588 | 0.77 | 0.363 | 0.384 | 0.211 |
| D           | 0.478 | 0.468 | 0.235 | 0.2  | 0.134 | 0.143 | 0.011 |
| I           | 0.045 |       |       |     |       |       |      |
| J           |       |       |       |     |       |       |      |
| K           |       |       |       |     |       |       |      |

\(^a\)Allotype haplotype ID are as described by Lefranc et al. \((6)\).

KIV, Kaingang from Ivaí; KRC, Kaingang from Rio das Cobras; GRC, Guarani Mbya; GKW, Guarani Kaiowa; GND, Guarani Ñandeva; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba.

TABLE 7 | Genetic differentiation for IG\(\text{HG1, IG\(\text{HG2,}\) and IG\(\text{HG3}\) among populations.}

| GKW | GND  | GRC  | KIV  | KRC  | BrJAP | CTBA |
|-----|------|------|------|------|-------|------|
|     | \(*\) | \(*\) | \(*\) | ns   | \(*\)  | \(*\) |
| GKW |      |      |      |      |       |      |
| GND | 0.02828 |      |      |      |       |      |
| GRC |      |      |      |      |       |      |
| KIV |      |      |      |      |       |      |
| KRC |      |      |      |      |       |      |
| BrJAP |      |      |      |      |       |      |
| CTBA |      |      |      |      |       |      |

Upper diagonal: the statistical significance of the exact test of population differentiation between pairs of population. \(* p < 0.001–0.01; \(*\) p < 0.001; ns p > 0.05. Lower diagonal: \(F_{ST}\) values between populations. Wright’s \(F_{ST}\) scale: 0–0.05 (in very light gray), 0.05–0.15 (in light gray), 0.15–0.25 (in medium gray), and above 0.25 (in dark gray) indicate little, moderate, great, and very great genetic differentiation between populations. KIV, Kaingang from Ivaí; KRC, Kaingang from Rio das Cobras; GRC, Guarani Mbya; GKW, Guarani Kaiowa; GND, Guarani Ñandeva; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba.

that study, the estimated admixture rate with non-Amerindians was 14.3% for GND, 3.7% for GRC, 7.2% for Kaingang, and no admixture for GKW \((15)\). The Gm allotype haplotype frequencies inferred from DNA sequencing in our study (in which the most common haplotypes were C and D) were similar to those found in former reports that characterized serologically the Guarani and Kaingang populations from Santa Catarina State, Brazil \((42)\), and other native American populations \((41, 43, 44)\).

The new allele IG\(\text{HG3}^*21\) was frequent in BrJAP \((f = 0.158)\), but absent in the other populations. According to the nucleotide sequence, it encodes the haplotype Gm5,10,11,13,14,26,27, whose frequency was previously reported as 15.2% in a study with Japanese families \((45)\). In that same study, the haplotypes C (Gm21;17,1,1;..) – 40.7%, D (Gm21;17,1,2;..) – 16.4%, I (Gm11,13,15,16,17,1,1;..) – 27.7%, and J (Gm5,11,13,3,1,23 – 15.2%) exhibited similar frequencies to the ones inferred from DNA sequencing in BrJAP, which were 38.4%, 13.4%, 28.6%, and 10.7%, respectively (Table 6). The novelty of our results is showing, for the first time, the characterization of the variants at DNA level that are responsible for the occurrence of these Gm haplotypes in Japanese populations.

Strong linkage disequilibrium (LD) (Figure S2) was observed in most Amerindian populations, as expected for these historically small populations that suffered strong genetic drift and multiple founder effects since the arrival of the first Americans to the continent and during their migration from the North to the South in the American continent. Interestingly, the patterns of LD differed among Guaranis, despite their shared ancestry. GKW exhibited a reduced number of variable sites, while GRC exhibited a reduced LD in comparison to GND. These differences could also be explained by genetic drift, as certain haplotypes that stochastically increased their frequencies in a population after their divergence may not have increased in the others.

In contrast, the Japanese-descendant and Euro-descendant populations have higher nucleotide and allele diversity and fewer
SNPs in LD. Even so, SNPs from different gene segments are in LD in these urban populations. In BrJAP, SNPs of allotypes G1m17 (rs1071803) and G2m(·) (rs8009156) are in LD ($D = 0.92; r^2 = 0.73$) and are present in the allotype haplotypes C and D, reported as the most common in Japanese populations (45).

In the MJ networks (Figure 3), IGHG1*02, IGHG2*03 and IGHG3*14 were connected with most alleles and were present at high frequencies in all populations. This pattern suggests that most of the other known alleles could have been originated from them. In the IGHG2 MJ network, one loop shows two paths where substitutions at position 161 of exon CH3 and 230 of CH2 occurred to generate the IGHG2*05, *07, and *13 alleles. It can be hypothesized that a mutation occurred in one of them, for example, IGHG2*03 at position 161 of exon CH3, generated IGHG2*05 and this allele, likewise, might have mutated at position 230 of exon CH2 originating allele IGHG2*13. As independent mutations in the same positions are extremely unlikely, the fact that the IGHG2*07 allele has a variant in the same position (230 of CH2 exon) indicates that gene conversion between alleles IGHG2*13 and IGHG2*03 originated the IGHG2*07 allele. Moreover, we suggest that the novel alleles IGHG1*11 and IGHG2*09 resulted from gene conversion between two frequent alleles of different gene segments. Overall, our data point to a major role of recombination and gene conversion originating new IGHG alleles, which is consistent with the tandem positioning and high sequence similarity of these segments, which favor unequal crossing-over (46).

Kaingang from Ival and Kaingang from Rio das Cobras presented low genetic differentiation ($F_{ST} = 0.032$), and similar allele frequencies (Table 7), most probably because of their recent common origin and gene flow due to the absence of cultural barriers, in addition to their geographical proximity. The $F_{ST}$ values between the Guarani populations were low to moderate, which is an evidence of genetic drift affecting the IGHG diversity in these populations. These results are compatible with previous reports for mtDNA in the same populations, which indicated that divergence of the three Guarani populations occurred at around 1,800 years before present (ybp), much earlier than the separation of the Kaingang populations that was estimated at of 207 ybp (13).

The PCA results (Figure 2) were consistent with geography and ancestry and showed that our data are consistent with data obtained by serologic methods, previously reported in the literature. The exception was India, which grouped with Europeans and Euro-descendants. In fact, PCA grouping does not necessarily mean common ancestry, as it can also result from migration or stochastic factors, or convergent evolution by natural selection. The grouping solely reflects the similarities of the IGHG allelic frequencies in these populations.

The results of most neutrality tests suggested that natural selection is not the major factor responsible for shaping IGHG diversity in the study populations. In other words, for IGHG the impact of genetic drift due to demographical processes is possibly stronger than the signal left by natural selection. As is known, Amerindians have a long history of migrations and isolation, and went through severe bottlenecks after the European colonization (14). Still, in GKW and KRC for IGHG2 and KRC and CTBA for IGHG3, the results of Tajima's $D$, and Fu and Li's

**FIGURE 2 |** Principal component analysis using Gm allotype haplotype frequencies was consistent with geography and ancestry. For comparisons with previously described population, we inferred the Gm allotype frequencies based on the observed nucleotide sequences, according to Lefranc et al. (6). Circles represent population data from the literature and squares represent populations from the present study. All frequencies reported in the literature are listed in Table S3. AFR, African populations; AMER, Amerindian populations; ASIA, Asian populations; EUR, European populations; EUR-BR, Euro-descendant populations from Brazil; ADM-BR, Admixed population from Brazil; KIV, Kaingang from Ival; KRC, Kaingang from Río las Cobras; GRC, Guarani Mbya; GKW, Guarani Kaiowa; GND, Guarani Ñandeua; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba.
FIGURE 3 | Relationship of IGHG alleles. Median-Joining Network of all IGHG1 (A), IGHG2 (B) and IGHG3 (C) alleles. Each circle (node) represents an allele and the size of each circle is proportional to the allele frequency. Numbers in the branches indicate the exon and the exonic position of nucleotide differences between alleles. The mv nodes (median vector) are possible unsampled or extinct ancestral sequences generated by the MJ algorithm to connect the alleles. Alleles IGHG3*11 and IGHG3*12 (C) were grouped because they do not differ in nucleotide sequence, except for the hinge size. KIV, Kaingang from Ivaí; KRC, Kaingang from Rio das Cobras; GRC, Guarani Mbya; GKW, Guarani Kaiowa; GND, Guarani Ñandeva; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba; NS, not sampled (alleles not observed in this study). The occurrence of multiple mutations in the same positions in different gene segments is extremely unlikely. In addition, sequence homology and tandem positioning favor unequal crossing over between high frequent alleles. Therefore, based on the multiple alignments, we suggest that (D) IGHG2*09 allele could be a product of gene conversion between IGHG3*03 and IGHG3*14 at position 56 (T), 58 (G), and 60 (G) of CH1 exon; and (E) IGHG1*11 could be a product of gene conversion between IGHG1*02 and IGHG2*03 at position 233 (T), 234 (G), and 254 (C) in CH2 exon.
Tajima's D – – – –
IGHG3
Kaingang from Rio das Cobras; GRC, Guarani Mbya; GKW, Guarani Kai owa; GND, Guarani Ñandeva; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba.

variation.
IGHG2
Positive selection (dN
14
Frontiers in Immunology | www.frontiersin.org

Number of alleles 15 14 29
Gene segment Test
IGHG1 Tajima’s D – – – – – – –
Fu and Li’s D – – – – – – –
Fu and Li’s F – – – – – – –
Fay and Wu’s H – – – – – – –
IGHG2 Tajima’s D –1.680*** – – – – – – –
Fu and Li’s D –2.696** – – – – – – –
Fu and Li’s F –2.793** – – – – – – –
Fay and Wu’s H –1.768** –2.745* –4.971** –5.241*** –5.518*** – – –
IGHG3 Tajima’s D – – – – – – –
Fu and Li’s D – – – – – – –
Fu and Li’s F – – – – – – –
Fay and Wu’s H – – – –4.310** –7.253*** –5.670*** –6.307* –7.372**

Statistical significance was tested by coalescent simulations with 10,000 repetitions; *p < 0.01 – 0.05; **p < 0.001 – 0.01; ***p < 0.001; -p > 0.05. KIV, Kaingang from Ivaí; KRC, Kaingang from Río das Casbas; GRC, Guarani Mbya; GKW, Guarani Kalowa; GND, Guarani Ñandeva; BrJAP, Japanese-descendants; CTBA, Euro-descendants from Curitiba.

D and F tests indicated diversity sweeps due to bottlenecks or purifying selection.

The Fay and Wu H test was significant with negative values for almost every population and gene segment analyzed. This could be interpreted as a result of an excess of derived variants at high frequencies in the gene genealogies. Fay and Wu (30) suggested that this may be a unique pattern produced by hitchhiking of variants in the vicinity that are being favored by positive selection. IGHG gene segments are located downstream of the IGHV, IGHD, and IGHL gene segments that encode the immunoglobulin variable regions, which specifically bind to antigens (2, 4). Therefore, we suggest that selection for variants in the variable region may be impacting the diversity of the constant region by hitchhiking mutations in the IGHG gene segments. This hypothesis is corroborated by the findings of Tanaka and Nei (48), who demonstrated that the non-synonymous mutation rate was higher than the synonymous rate in the gene segments that code for the Ig variable region. Their results were consistent with diversity-enhancing selection or overdominant selection driving the nucleotide diversity in the variable region.

CONCLUSION

Antibodies are pivotal for human survival, at both the individual and the population levels. It is surprising that despite decades of compelling evidence about the importance of the immunoglobulin gene variation for human immunity and the not so recent advent of sequencing technologies, most of the knowledge about IGHG is still based on serologic typing. As we see here, the fact that the regions encoded by IGHG are called “constant” does not mean these segments are not highly polymorphic. In fact, we found 16 novel alleles in a population sample of only 57 Japanese descendants. The IGHG genomic region is not well-covered in genome-wide association studies and whole genome sequencing databases. The homology and high sequence similarity of IGHG segments impose technical difficulties for sequencing, particularly at large scale. Besides,
the somatic recombination events characteristic of the IGH locus makes DNA from B-cell lines, used in so many studies, not suitable for IGH sequencing. Our study is the first to sequence systematically these segments at the nucleotide level in populations. We here present a full characterization of IGHG1-3 diversity in seven Brazilian populations, linkage disequilibrium, haplotypes and evidence of purifying selection and genetic drift. Understanding the IGH normal variation in populations and its evolution may be the key to better comprehend how the immune system fights invading organisms and non-self-antigens and also may contribute to the development of new vaccines.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Brazilian National Human Research Ethics Committee (CONEP) with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Brazilian National Human Research Ethics Committee (CONEP).

AUTHOR CONTRIBUTIONS

DA designed the study. VC-S, DA, LV, RD, and HI performed DNA sequencing and genotyping. VC-S analyzed the data. RW, VC-S, RD, HI, and LV performed molecular cloning and validation of novel alleles. MP-E, DA, DM, and RW contributed with reagents. VC-S, DA, MP-E, DM, and MB drafted the manuscript. All authors significantly contributed with ideas and critically reviewed this manuscript.

FUNDING

This work was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), PRONEX, Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Paraná, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

1. Schroeder HW, Cavacini L. Structure and function of immunoglobulins. J Allergy Clin Immunol. (2010) 125:S41–52. doi: 10.1016/j.jaci.2009.09.046

2. Beale D, Feinstein A. Structure and function of the constant regions of immunoglobulins. Q Rev Biophys. (1976) 9:135. doi: 10.1017/S003358350002390

3. Croce CM, Shander M, Martinis J, Cicurel L, D’Ancona GG, Dolby TW, et al. Chromosomal location of the genes for human immunoglobulin heavy chains. Proc Natl Acad Sci USA. (1979) 76:3416–9. doi: 10.1073/pnas.76.7.3416

4. McBride OW, Battey J, Hollis GF, Swan DC, Siebenlist U, Leder P. Localization of variable and constant region immunoglobulin heavy chain genes on subtelomeric band q32 of chromosome 14. Nucleic Acids Res. (1982) 10:8155–70. doi: 10.1093/nar/10.24.8155

5. Lefranc M-P, Lefranc G. The Immunoglobulin FactsBook. London; San Diego, CA: Academic Press (2001).

6. Lefranc M-P, Lefranc G. Human Gm, Km, and Am allotypes and their molecular characterization: a remarkable demonstration of polymorphism. Methods Mol Biol. (2012) 882:635–80. doi: 10.1007/978-1-61779-842-9_34

7. Granoff DM, Holmes SJ. G2m(23) immunoglobulin allotype and immunity to Haemophilus influenzae type b. J Infect Dis. (1992) 165(Suppl. 1):S66–9. doi: 10.1093/infdis/165.Supp-1.S66

8. Pandey JP, Kistner-Griffin E, Iwasaki M, Bu S, Deepe R, Black L, et al. Genetic markers of immunoglobulin G and susceptibility to breast cancer. Hum Immunol. (2012) 73:1155–8. doi: 10.1016/j.humimm.2012.07.340

9. Atherton A, Armour KL, Bell S, Minson AC, Clark MR. The herpes simplex virus type 1 Fc receptor discriminates between IgG1 allotypes. Hum Immunol. (2000) 30:2540–7. doi: 10.1016/S0198-8859(00)00524-0

10. Dard P, Lefranc MP, Ospitala L, Schanzer-Mazas A. DNA sequence variability of IGHG3 alleles associated to the main G3m haplotypes in human populations. Eur J Hum Genet. (2001) 9:765–72. doi: 10.1038/sj.ejhg.5200700

11. Huck S, Fort P, Crawford DH, Lefranc MP, Lefranc G. Sequence of a human immunoglobulin gamma 3 heavy chain constant region gene: comparison with the other human Cy genes. Nucleic Acids Res. (1986) 14:1779–89. doi: 10.1093/34.14.1779

12. Pandey JP, Li Z. The forgotten tale of immunoglobulin allotypes in cancer risk and treatment. Exp Hematol Oncol. (2013) 2:6. doi: 10.1186/2162-3619-2-6
13. Marrero AR, Silva-Junior WA, Bravi CM, Hutz MH, Petzl-Erler ML, Ruiz-Linares A, et al. Demographic and evolutionary trajectories of the Guarani and Kaingang natives of Brazil. Am J Phys Anthropol. (2007) 132:301–10. doi: 10.1002/aja.20515

14. Petzl-Erler ML, Lu R, Sotomaior VS. The HLA polymorphism of two distinctive South American Indian tribes: the Kaingang and the Guarani. Tissue Antigens. (1993) 41:227–37. doi: 10.1111/j.1399-0039.1993.tb0211x

15. Tsuneto LT, Probst CM, Hutz MH, Salzano FM, Rodrigues-Delfin LA, Zago MA, et al. HLA class II diversity in seven Amerindian populations. Clues about the origins of the Ach? Tissue Antigens. (2003) 62:512–26. doi: 10.1046/j.1399-0039.2003.00139.x

16. Lahiri DK, Nurnberger JJ. A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. Nucleic Acids Res. (1991) 19:5444. doi: 10.1093/19.15.5444

17. Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Lab Press (1989).

18. Lefranc M-P, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT®: the international ImMunoGeneTics information system®. 25 years on. Nucleic Acids Res. (2015) 43:D413–22. doi: 10.1093/nar/gku1056

19. Peakall R, Smouse PE. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes. (2006) 6:288–95. doi: 10.1111/j.1471-8286.2005.01155.x

20. Guo S, Thompson E. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics. (1992) 48:361–72. doi: 10.2307/2532296

21. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to analyze molecular data. Mol Ecol Resour. (2010) 10:564–7. doi: 10.1111/j.1755-0998.2010.02847.x

22. Barrett JC, Fry B, Maller J, Daly MJ. Haplovie: analysis and visualization of LD and haplotype maps. Bioinformatics. (2005) 21:263–5. doi: 10.1093/bioinformatics/bih457

23. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. (1999) 16:37–48. doi: 10.1093/oxfordjournals.molbev.a026636

24. Raymond M, Roussel F. An exact test for population differentiation. Evolution. (1995) 49:1280–3. doi: 10.1111/j.1558-5646.1995.tb04456.x

25. Reynolds J, Weir BS, Cockerham CC. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics. (1983) 105:767–79.

26. Slatkin M. A measure of population subdivision based on microsatellite allele frequencies. Genetics. (1993) 139:457–62.

27. Minitab Inc. Minitab 17 Statistical Software. (2010) Available online at: www.minitab.com (accessed November 30, 2016).

28. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. (1989) 123:585–95.

29. Fu XY, Li WH. Statistical tests of neutrality of mutations. Genetics. (1993) 133:693–709.

30. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics. (2000) 155:1405–13. Available online at: https://www.ncbi.nlm.nih.gov/pubmed/10880498

31. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. (2009) 25:1451–2. doi: 10.1093/bioinformatics/btp187

32. Grantham R. Amino acid difference formula to help explain protein evolution. Science. (1974) 185:862–4. doi: 10.1126/science.185.4154.862

33. Edelman GM, Cunningham BA, Gall WE, Gottlieb PD, Ritishubuar U, Waxdal MJ. The covariant structure of an entire gammaG immunoglobulin molecule. Proc Natl Acad Sci USA. (1969) 63:78–85. doi: 10.1073/pnas.63.1.78

34. Sherry ST, Ward M, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. (2001) 29:30–8. doi: 10.1093/nar/29.1.308

35. Brusco A, Cinque F, Savozi S, Boccazzi C, DeMarchi M, Carbonara AO. The G4 gene is duplicated in 44% of human immunoglobulin heavy chain constant region haplotypes. Hum Genet. (1997) 100:84–9. doi: 10.1007/s004390050470

36. Augusto DG, Povezan BZ, Tsuneto LT, Callegari-Jacques SM, Petzl-Erler ML. KIR gene content in amerindians indicates influence of demographic factors. PLoS ONE. (2013) 8:e56755. doi: 10.1371/journal.pone.0056755

37. Schanfield MS, Gergely J, Fudenberg HH. Immunoglobulin allotypes of European populations. Hum Hered. (1975) 25:370–7. doi: 10.1055/s-0025152748

38. Fortes-Lima C, Dugoujon JM, Hernández CL, Reales G, Calderón R. Immunoglobulin genes in andalusia (Spain). Genetic diversity in the mediterranean space. C R Biol. (2014) 337:464–56. doi: 10.1016/j.crvi.2014.08.004

39. Piazza A, van Loghem E, de Lange G, Curtoni ES, Ulizzi L, Terrenato L. Immunoglobulin allotypes in Sardinia. Am J Hum Genet. (1976) 28:77–86.

40. Pandey JP, Shannon BT, Araña-Chaves MP, Fudenberg HH. Gm and Km frequencies in a Portuguese population. Hum Genet. (1982) 61:154–6. doi: 10.1007/BF00274207

41. Lúcia M, Hamel H, Salzano FM, Freitas MMJ. The Gm polymorphism and racial admixture in six Amazonian populations. J Hum Evol. (1984) 15:517–29. doi: 10.1016/0044-2894(84)90005-6

42. Salzano FM, Steinberg AG. The Gm and Inv groups of Indians from Santa Catarina, Brazil. Am J Hum Genet. (1965) 17:273–29.

43. Johnson WE, Kohn PH, Steinberg AG. Population genetics of the human allotypes Gm, Inv, and A2m: an analytical review. Clin Immunol Immunopathol. (1977) 7:97–113. doi: 10.1016/10079-1970.01661.x

44. Schanfield MS. Immunoglobulin allotypes (Gm and KM) indicate multiple founding populations of Native Americans: evidence of at least four migrations to the New World. Hum Biol. (1992) 64:381–402.

45. Van Loghem E, Natvig JB, Matsumoto H. Genetic markers of immunoglobulins in Japanese families Inheritance of associated markers belonging to one IgA and three IgG subclasses. Ann Hum Genet. (1970) 33:351–9. doi: 10.1111/j.1469-1809.1970.tb00616.x

46. Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. (2007) 8:762–75. doi: 10.1038/nrg2193

47. Namboodiri AM, Pandey JP. The human cytomegalovirus TRL11/IRL 11-encoded FcγR binds differentially to allelic variants of immunoglobulin G1. Arch Virol. (2011) 156:907–10. doi: 10.1007/s00705-011-0937-8

48. Tanaka T, Nei M. Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol. (1989) 6:447–59.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Calonga-Solis, Malheiros, Beltrame, Vargas, Dourado, Isler, Wassem, Petzl-Erler and Augusto. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.