HFE gene polymorphism defined by sequence based typing of the Brazilian population and a standardized nomenclature for *HFE* allele sequences

Running head: *HFE* new nomenclature

Campos, W.N.[1], Massaro, J.D.[1], Martinelli, A.L.C.[2], Halliwell, J.A.[3], Marsh, S.G.E.[3,4], Mendes-Junior, C.T.[5], Donadi, E.A.[1].

[1] Division of Clinical Immunology, Department of Medicine, School of Medicine of Ribeirao Preto, University of Sao Paulo, Brazil.

[2] Division of Gastroenterology, Department of Medicine, School of Medicine of Ribeirao Preto, University of Sao Paulo, Brazil.

[3] Anthony Nolan Research Institute, Hampstead, London, NW3 2QG, UK.

[4] UCL Cancer Institute, Royal Free Campus, Hampstead London, NW3 2QG

[5] Departamento de Química, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901, Ribeirão Preto-SP, Brazil.

Key words: *HFE* gene; nomenclature**;** Brazilian; hereditary hemochromatosis, iron overload.

Correspondence

Wagner Narciso de Campos

Division of Clinical Immunology

Department of Medicine School of Medicine of Ribeirao Preto

University of Sao Paulo

Av. Bandeirantes, 3900 14049-900, Ribeirão Preto SP, Brazil.

Tel: +55 16 3602 3373

Fax: +55 16 3602 0229

e-mail: wncampos@yahoo.com.br

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/tan.13097

Abstract

The HFE molecule controls iron uptake from gut, and defects in the molecule have been associated with iron overload, particularly in hereditary hemochromatosis. The *HFE* gene including both coding and boundary intronic regions were sequenced in 304 Brazilian individuals, encompassing healthy individuals and patients exhibiting hereditary or acquired iron overload. Six sites of variation were detected: i) H63D C>G in exon 2, ii) IVS2 (+4) T>C in intron 2, iii) a C>G transversion in intron 3, iv) C282Y G>A in exon 4, v) IVS4 (-44) T>C in intron 4, and vi) a new Guanine deletion (G>del) in intron 5, which were used for haplotype inference. Nine *HFE* alleles were detected and six of these were officially named on the basis of the HLA Nomenclature, defined by the WHO Nomenclature Committee for Factors of the HLA System, and published via the IPD-IMGT/HLA website. Four alleles, *HFE**001, 002, 003 and 004 exhibited variation within their exon sequences.

Introduction

The *HFE* gene (abbreviation of High Fe) is located at the 6p21.3 telomeric chromosomal region of the extended Major Histocompatibility Complex (MHC), and it is considered to be a non-classical *HLA* class I-like gene (1, 2). *HFE* is composed of seven exons, translating in to 348 amino acids of the α-chain of the HFE molecule, which comprises three extracellular domains, a transmembrane region and a cytoplasmic tail. Exon 1 encodes the signal peptide, exons 2, 3 and 4 encode the α_1, α_2 and α_3 domains respectively, exon 5 encodes the transmembrane domain and exon 6 the cytoplasmic tail (3). The

HFE molecule contains four highly conserved Cysteine residues, which form two intramolecular disulfide bridges (C225 and C282) in the α_3 domain (4). The α_2 and α_3 domains are essential for the tertiary structure of the protein and for the interaction of the α_3 domain with β2-microglobulin, enabling HFE expression on the cell surface (5). The protein encoded by the *HFE* gene is strongly related to the regulation of intestinal iron absorption (6, 7).

The *HFE* gene exhibits many previously described single nucleotide polymorphisms (SNPs), as illustrated in Table 1. However, only three of these have been frequently studied: C282Y, H63D and S65C (5, 8, 9). As the relationship between these SNPs has not been systematically evaluated, we undertook a study of the coding and boundary intronic regions of the *HFE* gene in a selected Brazilian population. As there was no standardized nomenclature for the *HFE* gene, we worked with the WHO Nomenclature Committee for Factors of the HLA System, to provide official allele names for a number of HFE alleles (10).

The *HFE* coding region (exons 2, 3, 4, 5) and boundary introns were sequenced (Figure 1) in a Brazilian population sample consisting of 100 healthy individuals and 204 patients exhibiting hereditary (hereditary hemochromatosis, n=14) or acquired diseases (hepatitis C, n=130 and hepatocellular carcinoma, n=60) associated with liver iron overload.

DNA samples were extracted from peripheral blood cells using a salting-out procedure. Each exon was individually amplified by polymerase-chain reaction (PCR) using the sets of primers shown in Table 2. Primers were designed using the GENNERUNNER v.3.05 software (Hastings Software, Inc., Hastings, NY)

and the tools of the UCSC Genome Bioinformatics group (http://genome.ucsc.edu). An illustration of the amplified *HFE* segments is shown in Figure 1. Exons were amplified in a final volume of 50 µL, containing 1X PCR buffer [70 mM Tris-HCl pH 8.8, 20 mM $(NH_4)_2SO_4$], 3 mM $MgCl_2$, 1mM dithiothreitol, 0.2 mM of each dNTP, 10 pmol of each primer, 1.1 units of Taq DNA-polymerase (Platinum-Invitrogen, Carlsbad, CA), and 200 ng of genomic DNA. The initial denaturation cycle was carried out at 94°C for 5 minutes, followed by 30 cycles of 45 seconds at 94°C, 45 seconds at annealing temperature and 1 minute at 72°C, and by a final extension step at 72°C for 7 minutes. Amplification products were directly sequenced (ABI 3500 Genetic Analyzer, Applied Biosystems, Foster City, CA), using the same primers designed for amplification, according to manufacturer's instructions.

Observed SNPs were used to assemble the genetic profile of these samples and to reconstruct *HFE* haplotypes by two computational procedures: i) the ELB algorithm implemented by the ARLEQUIN v3.1 (11) software (alpha value: 0.01; epsilon value: 0.01; gamma value: 0; sampling interval: 500; burn-in steps: 100,000; recombination steps: 0), and ii) the PHASE method [number of iterations: 1000 (iterations in the final run increased 10 times); thinning interval: 1; burn-in value: 100; 5 value for each locus: 1; number of independent runs: 10], implemented by the PHASE v.2. software (12). To obtain consensual and reliable haplotypes for each sample, haplotype estimates were performed using several simulations, in which population samples were input as a single group.

Working with the WHO Nomenclature Committee for factors of the HLA

System, we established a standardized nomenclature for *HFE* alleles (Figure 2), using as reference the HLA Nomenclature described by Marsh *et al* (10), as described below:

1. Each *HFE* allele name has a unique number that corresponds up to three fields. The first field contains three digits and the following two fields contain two digits, each field is separated by a colon.
2. The first field is used to list the allele group, the number is assigned sequentially in the order in which the allele was determined, each number corresponding to a specific protein sequence. Alleles whose names differ in the first field must differ by one or more non-synonymous nucleotide substitutions altering the amino acid sequence of the protein.
3. The second field is used to indicate synonymous nucleotide differences between alleles within the coding region.
4. The third field is used to designate alleles that differ only by intronic nucleotide variation.
5. All allele names are a minimum of one field in length, additional fields are added only when necessary due to the discovery of related allele sequences.

To validate sequences, extended amplifications (introns 2-5) were performed for genomic (double stranded DNA) and cloned (single stranded DNA) samples. To obtain representatives single strand DNA samples, the following procedures were performed: i) purification of the extended (introns 2-5) PCR products (Jetx PCR-DNA and Gel Band Purification Kit, GE Healthcare, Buckinghamshire, UK), ii) ligation of each PCR product to pGEM®-T easy vector system (Promega, Madison, WI), iii) insertion of the vector into bacteria

by thermal shock, iv) growth of transformed bacteria in culture medium (LB Agar, Fisher-Scientific, Hampton, New Hampshire), containing 200 mg/mL ampicillin and 40 mg/mL Xgal (Ludwig Biotec, Alvorada, Brazil), and v) purification of each cloned DNA samples using the Illustra plasmid prep mini spin Kit (GE Healthcare). The products obtained were directly sequenced (ABI 3500 Genetic Analyzer), using the same primers designed for amplification, according to manufacturer's instructions.

Five recognized SNPs and a new mutation were observed in the *HFE* gene: i) H63D C>G at exon 2 (rs1799945), ii) IVS2 (+4) T>C at intron 2 (rs2071303), iii) the intron 3 C>G (rs807209), iv) C282Y G>A at exon 4 (rs1800562), v) IVS4 (-44) T>C at intron 4 (rs1800708), and vi) a new Guanine deletion (G>del) in intron 5. The sequence that included this new mutation is shown in the electropherogram in Figure 3. This fragment was cloned and amplified by EXON5 pair of primers, covering the whole exon 5 and a partial cover of introns 4 and 5, since the last intron holds the recently annotated mutation. The identification of SNPs was retrieved from the NCBI (http://www.ncbi.nlm.nih.gov) and Ensembl (http://www.ensembl.org) databases.

The haplotype inference analysis yielded nine alleles, of which four *HFE**001, *002, *003, and *004 allele groups, encompassing only SNPs at exonic sequences (Table 3), were validated as official *HFE* alleles and published via the IPD-IMGT/HLA website (http://www.ebi.ac.uk/ipd/imgt/hla/), release number 3.27.0 (20170120) (13, 14). A single sample showed a discordant result by the ELB method in only one simulation by ARLEQUIN software v.3.1 (10), while all other simulations provided identical results,

regardless of the method or population structure evaluated. The discordant haplotype showed low probability of occurrence (0.092), when recalculated by PHASE v.2 software (12).

This study enabled us to define a number of *HFE* allele sequences and receive official names for these alleles. By so doing we were able to link a number of polymorphic sites that has previously been observed within the *HFE* gene, and link these within the allele sequences. In addition, the study of Brazilian healthy individuals, which constitutes a highly genetically diverse population and the study of an array of patients exhibiting hereditary or acquired liver iron overload (associated with the *HFE* gene) may permit a wide recognition of the *HFE* coding and boundary intron region variability. Finally, the standardization of the *HFE* gene nomenclature will help on the definition of new *HFE* alleles on future studies.

Acknowledgments

We acknowledge CNPq - National Council for Scientific and Technological Development for the financial support, Drs Eduardo Luiz Rachid Cançado and Andrea Evangelista da Silva for providing hereditary hemochromatosis samples, and Drs Aguinaldo Luiz Simões and Claudia E. Vieira Wiezel for helping on cloning procedures.

Conflict of interest

The authors confirm that there are no conflicts of interest.

References

1. Feder J, Gnirke A, Thomas W, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature genetics*. 1996:399-408.
2. Horton R, Wilming L, Rand V, et al. Gene map of the extended human MHC. *Nature Reviews Genetics*. 2004;**5**:889-99.
3. Rhodes D, Trowsdale J. Alternate splice variants of the hemochromatosis gene Hfe. *Immunogenetics*. 1999;**49**:357-9.
4. Le Gac Gé, Dupradeau F-Y, Mura C, et al. Phenotypic expression of the C282Y/Q283P compound heterozygosity in HFE and molecular modeling of the Q283P mutation effect. *Blood cells, molecules & diseases*. 2002;**30**:231-7.
5. Merryweather-Clarke AT, Pointon JJ, Jouanolle AM, Rochette J, Robson KJ. Geography of HFE C282Y and H63D mutations. *Genetic testing*. 2000;**4**:183-98.
6. Krause A, Neitz S, Mägert H-J, et al. LEAP-1, a novel highly disulfide-bonded human peptide, exhibits antimicrobial activity. *FEBS letters*. 2000;**480**:147-50.
7. Park CH, Valore EV, Waring AJ, Ganz T. Hepcidin, a urinary antimicrobial peptide synthesized in the liver. *Journal of biological chemistry*. 2001;**276**:7806-10.
8. Franchini M, Veneri D. Recent advances in hereditary hemochromatosis. *Annals of hematology*. 2005;**84**:347-52.
9. Anderson G, Ramm G, Subramaniam V, Powell L. HFE gene and hemochromatosis. *Journal of gastroenterology and hepatology*. 2004;**19**:712-.
10. Marsh SGE, Albert ED, Bodmer WF, et al. Nomenclature for factors of the HLA system, 2010. *Tissue antigens*. 2010;**75**:164.
11. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*. 2005;**1**:47.
12. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*. 2001;**68**:978-89.
13. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic acids research*. 2015: 43:D423-431.
14. Marsh SGE. Nomenclature for factors of the HLA system, update August 2015. *Tissue Antigens*. 2015;**86**:463-8.

TABLES

Table 1. Previous names of the most important variation sites observed at the *HFE* gene, and their reference SNP numbers (rsIDs#). The most studied *HFE* SNPs are in bold type.

Current name used in literature	rsIDs #	Gene position*
V53M	rs28934889	+8641G>A (exon 2)
V59M	rs111033557	+8659G>A (exon 2)
H63D	**rs1799945**	**+8671C>G (exon 2)**
S65C	rs1800730)	+8677A>T (exon 2)
IVS2 (+4) T>C	rs2071303	+8828T>C (intron 2)
Q127H	rs28934595	+9074A>C (exon 3)
E168Q	rs146519482	+9195G>C (exon 3)
C282Y	**rs1800562**	**+10633G>A (exon 4)**
Q283P	rs111033563	+10636A>C (exon 4)
E168Q	rs146519482	+9195G>C (exon 4)
W169T	rs797045145er	+9199G>A (exon 4)
IVS4 (-44) T>C	rs1800708	+10795T>C (intron 4)

* According to the *HFE* RefSeq gene sequence NG_008720.2, which is the *HFE**001:01:01 allele

Table 2. Description of *HFE* gene target regions, including the sequences of the primers used to amplify the complete coding and boundary introns region and those used for partial segment amplification. The annealing temperatures and the amplicon lengths are also shown.

Amplified Region	Primer sequences	Annealing temperature	Amplicon length
Intron 1 to Intron 5	*HFE*- complete coding region Forward – 5´- GTGGCAGAGAAAAGCACACAAG – 3´ *HFE*- complete coding region Reverse– 5´- GAGACTTTCATTCTGGGGAG – 3´	60ºC	2,757 kbp
Intron 1 (partial) Exon 2 Intron 2 (partial)	*HFE*exon2F: 5´-GCTGATGGTATGAGTTGATGC-3´ *HFE*exon2R: 5´-CTCAGACTTCCAGCTGTTTC-3´	60ºC	444 bp
Intron 2 (partial) Exon 3 Intron 3 (partial)	*HFE*exon3F: 5´-GCTTCCTGAGATCATTTGGTCC-3´ *HFE*exon3R: 5´-CAGAATTTGGAGAGGCACACAG-3´	59ºC	485 bp
Intron 3 (partial) Exon 4 Intron 4 (partial)	*HFE*exon4F: 5´- CTGATCTGACTGCTCTCCAAGTG-3´ *HFE*exon4R: 5´-CATAATTACCTCCTCAGGCAC-3´	53ºC	545 bp
Intron 4 (partial) Exon 5 Intron 5 (partial)	*HFE*exon5F: 5´-GTGCCTGAGGAGGTAATTATGG-3´ *HFE*exon5R: 5´-CTTCCCATGGATGCCAGATC-3´	58ºC	308 bp

Table 3. Allele identification and haplotype composition of nucleotide variations observed along the coding and boundary introns region of the *HFE* gene observed in Brazilian individuals. Mutated bases are indicated in boldface and variation sites at exonic regions are highlightened in grey.

Allele identification	Haplotype composition						Previous nomenclature
	+3511[a]	+3668[b]	+5197[c]	+5473[d]	+5635[e]	+5811[f]	
*HFE**001:01:01[†]	C	T	C	G	T	G	
*HFE**001:01:unofficial2[††]	C	**C**	C	G	T	G	
*HFE**001:01:unofficial3[††]	C	T	**G**	G	T	G	
*HFE**001:01:unofficial 4[††]	C	T	C	G	**C**	G	
*HFE**001:01:unofficial5[††]	C	**C**	C	G	**C**	G	
*HFE**001:01:unofficial6[††]	C	**C**	**G**	G	T	**del**	
*HFE**002 [†]	**G**	**C**	C	G	T	G	H63D
*HFE**003[†]	C	T	C	**A**	T	G	C282Y
*HFE**004[†]	**G**	**C**	C	**A**	T	G	H63D/C282Y

Identification of SNPs and most frequent allele according to NCBI (http://www.ncbi.nlm.nih.gov) and Ensembl (http://www.ensembl.org). Nucleotide locations were assigned considering the first transcribed base as +1: a - rs1799945 (H63D); b - rs2071303 (IVS2 (+4) T>C); c - rs807209; d - rs1800562 (C282Y); e - rs1800708 (IVS4 (-44) T>C); f - deletion of a Guanine at intron 5.

[†]Alleles submited by our research group, and published via the IPD-IMGT/HLA Database (http://www.ebi.ac.uk/ipd/imgt/hla/) (1, 14). Validation covering the whole gene region (exonic and intronic) by next generation sequence is in course.

[††]These alleles are designated as unofficial, since they do not present official names assigned by IPD-IMGT/HLA. Intronic regions harboring these mutations must be cloned and sequenced in order to obtain an official nomenclature.

Figure legends

Figure 1. *HFE* gene structure at chromosome (6p21.3) according to the NCBI RefSeq gene sequence NG_008720.2, which is the *HFE**001:01:01 allele, SNPs ID# (NCBI Database - http://www.ncbi.nlm.nih.gov/snp/; Ensembl Database - http://www.ensembl.org/). Shaded grey areas indicate the sequenced area.

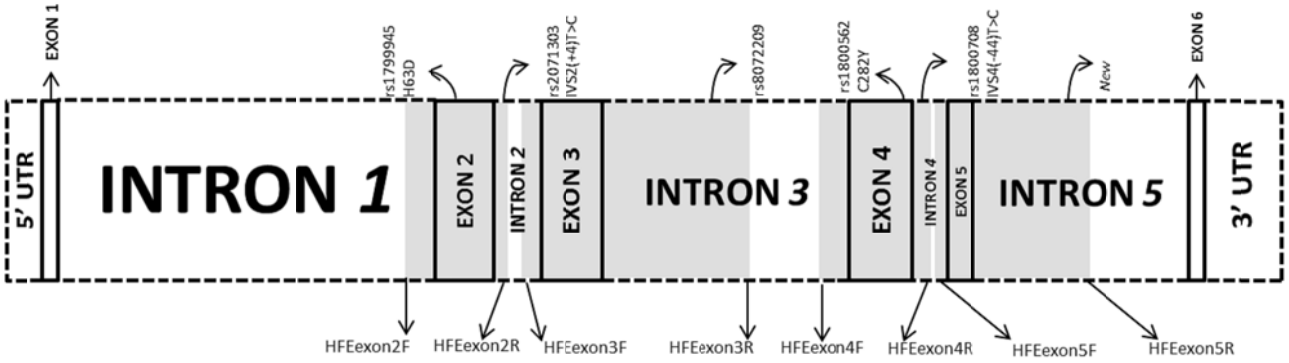

Figure 2. Nomenclature for the *HFE* alleles. The first field of three digits represents the number of the encoded HFE proteins. The second field of two digits represents the number of synonymous exonic mutations for each allele. The third field of two digits represents the number of intronic mutations for each allele.

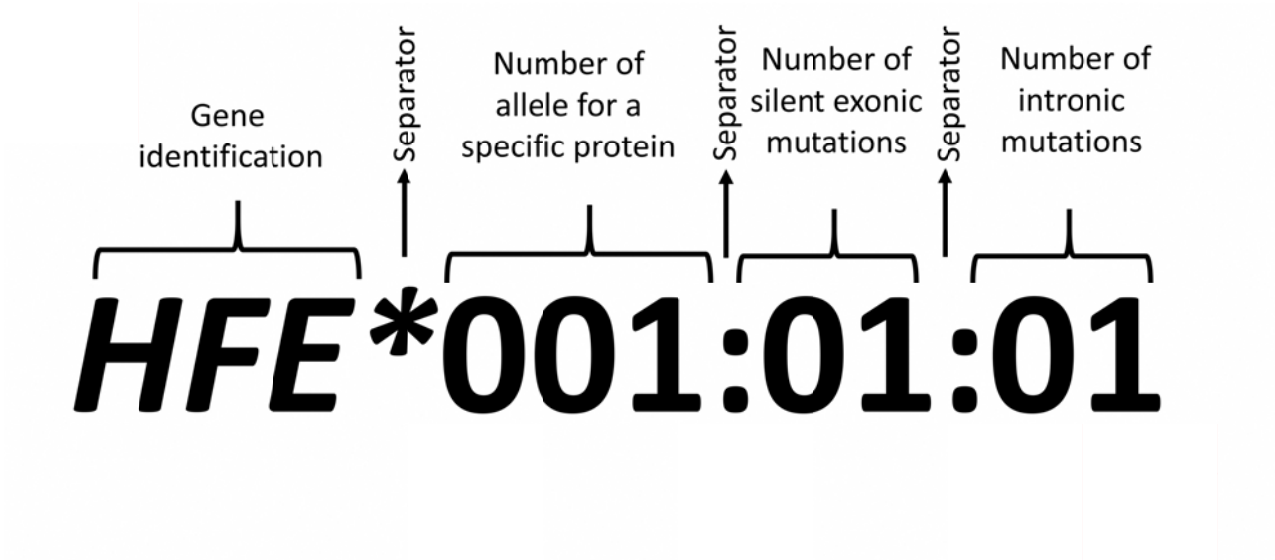

Figure 3. Electropherogram indicating a deletion of a Guanine at intron 5, observed in just one out of the 304 individuals evaluated in this study. The deletion was observed in heterozygosis, causing a shift in the reading frame and shuffling the bases in downstream sequencing as seen in the heterozygous sample (double strand non-cloned sequence). The sequence was cloned and amplified by EXON5 pair of primers covering the whole exon 5 and a partial cover of introns 4 and 5, since the last intron holds a recently annotated new mutation.