Available online at www.sciencedirect.com
SCIENCE DIRECT®
Developmental and Comparative Immunology 29 (2005) 917–938
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

IMGT Locus in focus

IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN

Marie-Paule Lefranc*, Elodie Duprat, Quentin Kaas, Madeleine Tranne, Aude Thiriot, Gérard Lefranc

IMGT, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, Université Montpellier II, Institut de Génétique Humaine, IGH UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France

Received 8 February 2005; accepted 10 March 2005
Available online 27 April 2005

Abstract

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) provides a common access to expertly annotated data on the genome, proteome, genetics and structure of immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates. The NUMEROTATION concept of IMGT-ONTOLOGY has allowed to define a unique numbering for the variable domains (V-DOMAINs) and constant domains (C-DOMAINs) of the IG and TR, which has been extended to the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily (IgSF) proteins other than the IG and TR (*Dev Comp Immunol* 27:55–77, 2003; 29:185–203, 2005). In this paper, we describe the IMGT unique numbering for the groove domains (G-DOMAINs) of the MHC and for the G-LIKE-DOMAINs of the MHC superfamily (MhcSF) proteins other than MHC. This IMGT unique numbering leads, for the first time, to the standardized description of the mutations, allelic polymorphisms, two-dimensional (2D) representations and three-dimensional (3D) structures of the G-DOMAINs and G-LIKE-DOMAINs in any species, and therefore, is highly valuable for their comparative, structural, functional and evolutionary studies.

Keywords: IMGT; Colliers de Perles; Groove domain; MHC; Major histocompatibility complex; MhcSF; Superfamily; Immune system

1. Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1] is a high quality integrated knowledge resource specialized in the immunoglobulins (IG), T cell receptors (TR) and major histocompatibility complex (MHC) of human and other vertebrates, the immunoglobulin superfamily (IgSF) and MHC superfamily (MhcSF), and the related proteins of the immune system (RPI)

Abbreviations 2D, two-dimensional; 3D, three-dimensional; MHC, major histocompatibility complex; MhcSF, MHC superfamily; IG, immunoglobulin; TR, T cell receptor; IgSF, immunoglobulin superfamily; IMGT, the international ImMunoGeneTics information system®; RPI, related proteins of the immune system.

* Corresponding author. Tel.: +33 4 9961 9965; fax: +33 4 9961 9901.
E-mail address: lefranc@ligm.igh.cnrs.fr (M.-P. Lefranc).
URL: http://imgt.cines.fr.

0145-305X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2005.03.003

[1–16]. IMGT provides a common access to expertly annotated data on the genome, proteome, genetics and structure of the IG, TR, MHC, IgSF, MhcSF and RPI, according to the IMGT Scientific chart rules and to the IMGT-ONTOLOGY concepts [17]. More particularly, the IMGT unique numbering [18–21], based on the NUMEROTATION concept of IMGT-ONTOLOGY, has been set up to provide a standardized description of mutations, allelic polymorphisms, two-dimensional (2D) and three-dimensional (3D) structure representations of the IG and TR variable domains (V-DOMAINs), and constant domains (C-DOMAINs) whatever the antigen receptor, the chain type or the species [20,21]. The IMGT unique numbering for V-DOMAINs and C-DOMAINs is used in all the IMGT components [1,6], and more particularly in the databases (IMGT/LIGM-DB [22], IMGT/PRIMER-DB [23], IMGT/GENE-DB [24], IMGT/3Dstructure-DB [25]), in the tools for sequence and structure analysis (IMGT/V-QUEST [26], IMGT/JunctionAnalysis [27], IMGT/Allele-Align, IMGT/PhyloGene [28], IMGT/StructuralQuery [25]), and in the IMGT Repertoire Web resources ('IMGT Protein displays' [29,30], 'IMGT Colliers de Perles' 2D representations [31], and 'IMGT Alignments of Alleles' [32,33]; see http://imgt.cines.fr). Interestingly, the IMGT unique numbering for V-DOMAIN and for C-DOMAIN has been fully extended to the V-LIKE-DOMAINs and C-LIKE-DOMAINs of IgSF proteins other than the IG and TR [20,21,34,35]. This is particularly remarkable for the V-DOMAINs and V-LIKE-DOMAINs, the genomic structures of which are strikingly different. Indeed, the IG and TR V-DOMAINs are encoded by rearranged V-(D-)J genes [32,33], whereas the V-LIKE-DOMAINs are often encoded by a single exon [20,35].

In this paper, we define a standardized IMGT unique numbering for the MHC groove domains (G-DOMAINs) of the MHC of all jawed vertebrates. We show that this IMGT unique numbering for G-DOMAINs can be extended to the G-LIKE-DOMAINs of the MHC-like proteins (MhcSF proteins other than MHC), of any species. The IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN represents, therefore, a major step forward for the comparative analysis of the sequences and structures of these domains, and for the study of their evolution.

2. MHC chain and G-DOMAIN definition

The MHC proteins that present peptides to the T cells belong to the 'classical MHC class I' (MHC-Ia) or to the 'classical MHC class II' (MHC-IIa) [36]. The MHC-Ia comprises, in human, the HLA-A, HLA-B and HLA-C subclasses, and in mouse, the H2-D, H2-K and H2-L subclasses. The MHC-IIa comprises, in human, the HLA-DP, HLA-DQ and HLA-DR subclasses, and in mouse, the H2-A, H2-E and H2-P subclasses (H2-P being unproductive). The MHC proteins with more specific functions or which do not present peptides to the T cells belong to the 'nonclassical MHC-I' (MHC-Ib) or to the 'nonclassical MHC-II' (MHC-IIb). The MHC-Ib comprises, in human, the HLA-E, HLA-F and HLA-G subclasses (each one represented by a unique isotype), and, in mouse, the H2-Q, H2-M and H2-T subclasses (each one comprising several isotypes). The MHC-IIb comprises, in human, the HLA-DM and HLA-DO subclasses and, in mouse, the H2-DM and H2-DO subclasses (IMGT Repertoire MHC, http://imgt.cines.fr). Thus there are 11 MHC subclasses in human and in mouse (3 MHC-Ia, 3 MHC-IIa, 3 MHC-Ib, and 2 MHC-IIb).

The MHC-I proteins, expressed on the cell surface of most cells, are formed by the association of a transmembrane heavy chain (I-ALPHA chain) and a noncovalently linked light chain beta-2-microglobulin (B2M) (Fig. 1). The MHC-II proteins, expressed on the cell surface of professional antigen presenting cells (APC), are heterodimers formed by the association of two transmembrane chains, an alpha chain (II-ALPHA chain) and a beta chain (II-BETA chain) (Fig. 1).

The I-ALPHA chain of the MHC-I, and the II-ALPHA and II-BETA chains of the MHC-II proteins, comprise an extracellular region made of three domains for the MHC-I chain and of two domains for each MHC-II chain, a connecting region, a transmembrane region and an intracytoplamic region (Fig. 2). The I-ALPHA chain comprises two groove domains (G-DOMAINs), the G-ALPHA1 [D1] and G-ALPHA2 [D2] domains, and one C-LIKE domain [D3] [21,36]. The II-ALPHA chain and the II-BETA chain each comprises two domains, the G-ALPHA [D1] and C-LIKE [D2] domains, and the G-BETA [D1] and C-LIKE [D2] domains, respectively (Fig. 2). The four

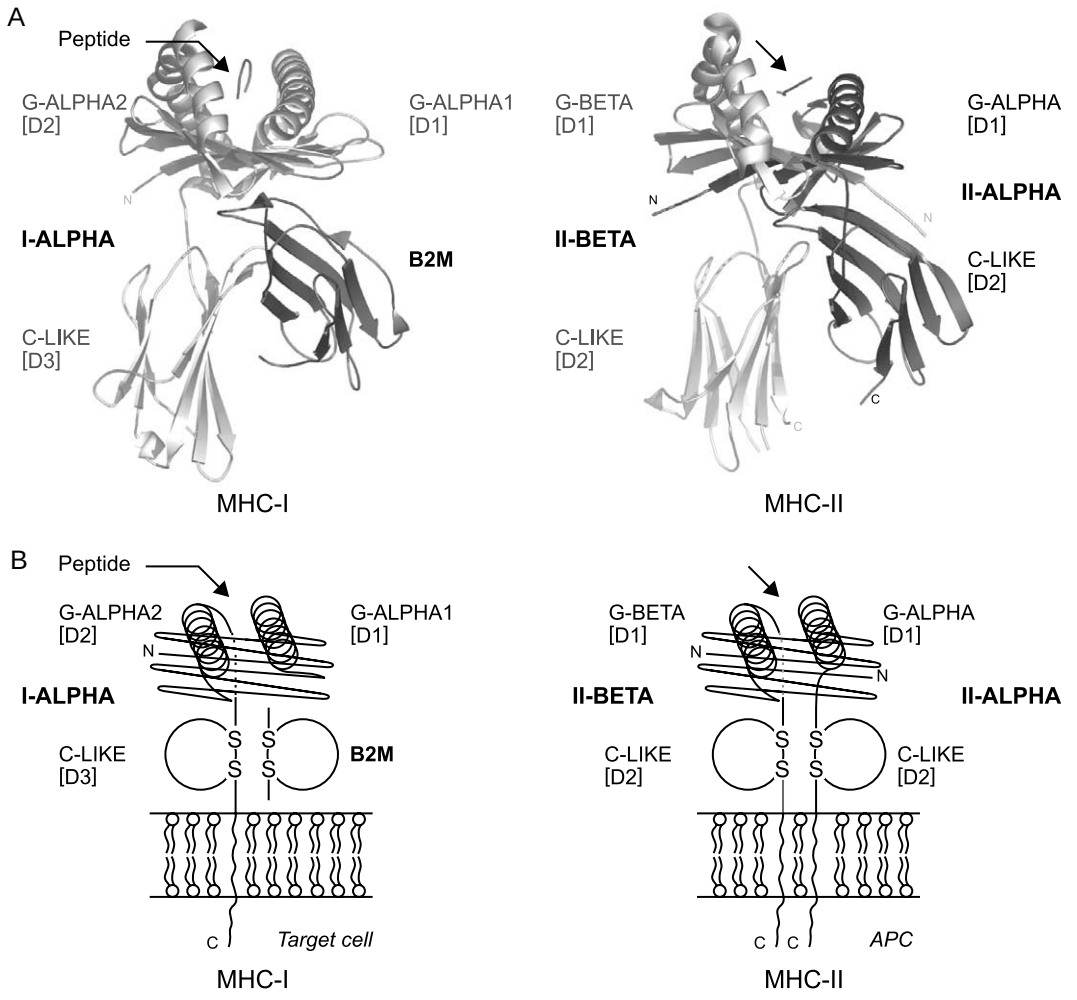

Fig. 1. 3D structures and schematic representations of the MHC class I (MHC-I) and MHC class II (MHC-II) proteins. (A) 3D structures of MHC-I and MHC-II (1oga and 1j8 h annotated coordinate files from IMGT/3Dstructure-DB [25], http://imgt.cines.fr, that include crystallographic data from the Protein DataBank PDB [37]). The MHC-I comprises the I-ALPHA and the beta-2-microglobulin (B2M) chains. The I-ALPHA chain is shown with its extracellular domains (G-ALPHA1, G-ALPHA2 and C-LIKE) [36]. The MHC-II comprises the II-ALPHA and II-BETA chains that are shown with their extracellular domains (G-ALPHA and C-LIKE for the II-ALPHA chain, G-BETA and C-LIKE for the II-BETA chain). (B) Schematic representations of the MHC-I and MHC-II proteins. The MHC-I and MHC-II are shown as transmembrane proteins, at the surface of a target cell and of an antigen presenting cell (APC), respectively. Complete MHC-I and MHC-II chains comprise the extracellular domains (shown in A) and the connecting, transmembrane and cytoplamic regions (not present in 3D structures [36], for details see Fig. 2). [D1], [D2] and [D3] indicate the position of the domains from the N-terminal end of the chains. Arrows indicate the peptide localization in the MHC groove (the N-terminal end of the peptide is in the back).

G-DOMAINs, G-ALPHA1 and G-ALPHA2 of the MHC-I proteins, and G-ALPHA and G-BETA of the MHC-II proteins have a similar groove 3D structure that consists of one sheet of four antiparallel beta strands (‘floor’ of the groove or platform) and one long helical region (‘wall’ of the groove). This groove is part of the cleft that is the peptide binding site of the classical MHC-Ia and MHC-IIa proteins [36] (Fig. 1).

Owing to the conserved structure between classical and nonclassical MHC, and according to the DESCRIPTION concept of IMGT-ONTOLOGY [5,6,17], the same labels are used, for MHC-Ia

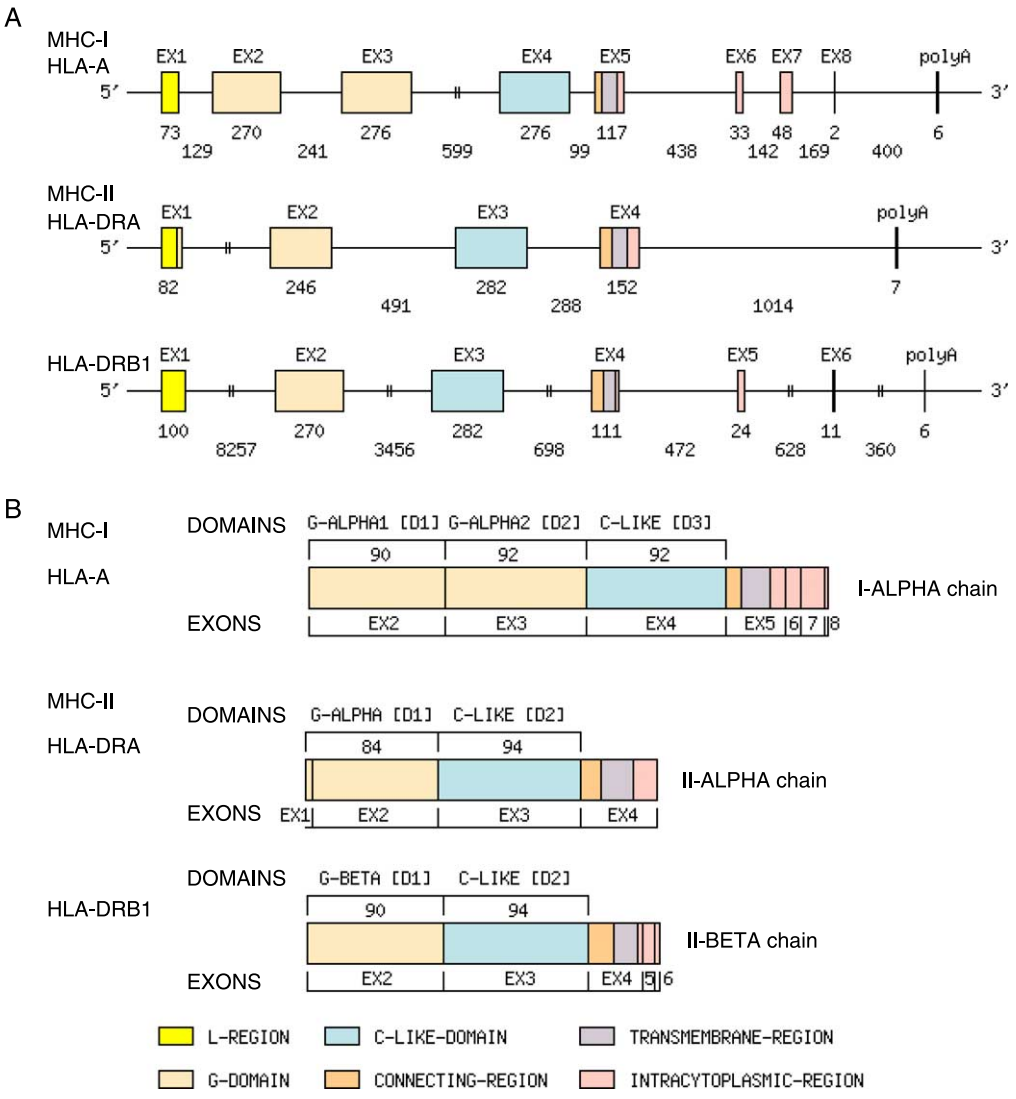

Fig. 2. Correspondence between exons and domains for the MHC-I and MHC-II. (A) Exons of the *Homo sapiens* MHC-I HLA-A gene and MHC-II HLA-DRA and HLA-DRB1 genes, shown as examples. Lengths of the exons, introns and polyA signal are in base pairs. Introns indicated with ‖ are not at scale. (B) Domains of the *Homo sapiens* MHC-I HLA-A (I-ALPHA) chain and of the MHC-II HLA-DRA (II-ALPHA) and HLA-DRB1 (II-BETA) chains, shown as examples. Lengths of the domains are in number of amino acids. The G-ALPHA domain of HLA-DRA (84 amino acids) is encoded by EX2 (82 codons) and the 3′ end of EX1 (2 codons). EMBL/GenBank/DDBJ accession numbers: HLA-A (K02883), HLA-DRA (J00203 and J00204) and HLA-DRB1 (AL137064) (Table 1).

and MHC-Ib, in the description of their heavy chain and domains: I-ALPHA chain, and G-ALPHA1 and G-ALPHA2 domains. Similarly, the same labels are used, for MHC-IIa and MHC-IIb, in the description of their respective alpha and beta chains and domains: II-ALPHA and II-BETA chains, and G-ALPHA and G-BETA domains.

3. IMGT unique numbering for G-DOMAIN

Correspondence between the four G-DOMAINs was established by extensive sequence alignment comparison of annotated MHC chains from the IMGT Repertoire [1,6] and by structural data analysis and alignment of MHC proteins with known 3D structures

from IMGT/3Dstructure-DB, http://imgt.cines.fr [25]. As each G-DOMAIN is usually encoded by a single exon, the delimitation of the domains in IMGT takes into account the limits of the exons in the genomic structure of the MHC genes (Fig. 2). In Table 1 are indicated the EMBL/GenBank/DDBJ [38–40] accession numbers of the sequences whose domains are reported in the IMGT Protein display (Fig. 3), the IMGT/3Dstructure-DB [25] entries of the representative alleles, as well as the accession numbers of genomic sequences from other alleles that were necessary to identify the splicing sites.

The IMGT Protein display (Fig. 3), based on the IMGT unique numbering for G-DOMAIN, shows, for the first time, a standardized amino acid alignment of G-DOMAINs that belong to the same or to different chain types, from the classical and nonclassical MHC-I and MHC-II, and from different species. Indeed, this IMGT Protein display includes (i) the G-ALPHA1 [D1] domains of MHC-Ia (human HLA-A*0201, HLA-B*0702, HLA-Cw*0701 alleles found in a frequent haplotype in caucasian populations, and mouse H2-D1*02, H2-K1*01, H2-L*02), (ii) the G-ALPHA1 [D1] domains of MHC-Ib (human HLA-E*01, HLA-F*01, HLA-G*01, and mouse H2-M5*02, H2-Q7*02, H2-T3*01), (iii) the G-ALPHA [D1] domains of MHC-IIa (human HLA-DPA1*0103, HLA-DQA1*0501, HLA-DRA*0101 alleles found in a frequent haplotype in caucasian populations, and mouse H2-AA*02, H2-EA*02) (iv) the G-ALPHA [D1] domains of MHC-IIb (HLA-DMA*01, HLA-DOA*01, and mouse H2-DMA*01, H2-DOA*01) (Fig. 3A), (v) the G-ALPHA2 [D2] domains of MHC-Ia and MHC-Ib (same chains as described above for the G-ALPHA1 [D1] domains), (vi) the G-BETA [D1] domains of MHC-IIa (human HLA-DPB1*0401, HLA-DQB1*0301, HLA-DRB1*1402, and mouse H2-AB*02, H2-EB1*01), and (vii) the G-BETA [D1] domains of MHC-IIb (human HLA-DMB*01, HLA-DOB*01), and mouse H2-DMB1*02, H2-DOB*01) (Fig. 3B).

For each G-DOMAIN, the positions that contribute to the groove floor comprise positions 1–49, with the A strand from positions 1–14, the AB turn positions 15–17, the B strand positions 18–28, the BC turn positions 29 and 30, the C strand positions 31–38, the CD turn positions 39–41 and the D strand positions 42–49 (Fig. 3 and Table 2). The additional position 7A represents a bulge in 3D structures and is present in some G-ALPHA domains, for instance those of the HLA-DQA1, H2-AA, HLA-DOA and H2-DOA chains. This position 7A is added if G-ALPHA sequences are introduced in G-DOMAIN alignments (Fig. 3). The gaps of the floor are localized in the turns. The AB turn (positions 15–17) comprises three amino acids in the G-ALPHA1, G-ALPHA2 and G-BETA domains but these positions are unoccupied in the G-ALPHA domains (as well as position 18 of strand B). The BC turn (positions 29 and 30) comprises two positions that are occupied in all G-DOMAINs. The CD turn (positions 39–41) is occupied by three amino acids in the G-ALPHA1 domains and only one in the other domains (G-ALPHA, G-ALPHA2 and G-BETA) (Fig. 3). Additional positions at the N-terminus of strand A or at the C-terminus of strand D can be added if necessary. Thus, two additional positions (1.2 and 1.1) are added at the N-terminus of the A strand of the G-ALPHA domains as the presence of these two amino acids was demonstrated by protein sequencing of the HLA-DRA [42] (Fig. 3A, Table 2A). In each [D1] G-DOMAIN, except the G-ALPHA domain of HLA-DMA and H2-DMA, the amino acid at position 1 (shown within parentheses in Fig. 3) is encoded by the codon that results from the splicing between the first exon (EX1) that encodes the L-REGION, and the second exon (EX2) that encodes [D1]. The two amino acids, isoleucine (I) and lysine (K) at positions 1.2 and 1.1 of HLA-DRA G-ALPHA [D1], are therefore encoded by EX1. By extrapolation, two amino acids have been added at positions 1.2 and 1.1 for the other G-ALPHA domains, but in those cases, the proteolytic cleavage site of the leader peptide (L-REGION) needs to be confirmed experimentally. It is also necessary to confirm if the amino acids at positions 1.10–1.1 of HLA-DMA and H2-DMA belong, or not, to the mature protein (Fig. 3A, Table 2A). Four additional positions (49.1–49.4) are observed at the C-terminus of strand D of the G-BETA domain of HLA-DMB and H2-DMB1 (Fig. 3B, Table 2B).

The numbering of the alpha helix starts at position 50 and ends at position 92, with five additional positions at 54A, 61A, 61B, 72A and 92A. Three of them (61A, 61B, 72A) characterize the G-ALPHA2 and/or G-BETA domains (Fig. 3B, Table 2B). Indeed, positions 61A and 72A are occupied in

Table 1
Representative MHC and MHC-I-like genes and chains

A	Species	IMGT gene and allele names[a]	gDNA[b]	Mouse strain	Mouse H2 haplotype[c]	cDNA[b]	IMGT/3D structure-DB[d]	gDNA EX1 (same allele)[b,e]	gDNA (other alleles)[b,e]
MHC-Ia	*Homo sapiens*	HLA-A*0201	K02883				1oga_A		
		HLA-B*0702	AJ292075						
		HLA-Cw*0701	Y18533					Y18499	
			Y18534						
	Mus musculus	H2-D1*02	M18523	C57BL/10	b		1juf_A		
		H2-K1*01	V00746	C57BL/10	b		1lk2_A		
		H2-L*02	L00127	BALB/c	d		(1ld9_A)		
MHC-Ib	*Homo sapiens*	HLA-E*01	AF523277				1mhe_A		
		HLA-F*01	X17093						
		HLA-G*01	J03027						
	Mus musculus	H2-M5*02	L14279	BALB/c	d				
		H2-Q7*02	X03210						
		H2-T3*01	M11742	C57BL/6	b				
MHC-I-like	*Homo sapiens*	MICA*01				U56940	1hyr_C		L29411
		MR1*01	AL356267						
		RAET1N*01	AL355497				1kcg_C		
	Mus musculus	AZGP1*01	AF281658	129/Svj	129				
		CD1D1*01	X13170				1cd1_A		
		FCGRT*01	D37872	BALB/c	d				

B		IMGT gene and allele names[a]	gDNA[b]	Mouse strain	Mouse H2 haplotype[c]	cDNA[b]	Mouse strain	Mouse H2 haplotype[c]	IMGT/3D structure-DB[d]	gDNA EX1 (same allele)[b,e]	gDNA (other alleles)[b,e]
MHC-IIa	*Hs*	HLA-DPA1*0103	X03100								
		HLA-DPB1*0401	M23907							M23906	
		HLA-DQA1*0501	Z84489						1s9v_A		
		HLA-DQB1*0301				M25325					U92032
		HLA-DRA*0101	J00204						1fv1_D	J00203	Z84814
		HLA-DRB1*1402				AJ297583					AL137064
	Mm	H2-AA*02	AY740451	B10.MOL1	(w12)	V00832	B10A	k/d2?	1iak_A		AF027865
		H2-AB*02	AY740477	B10. SNA70	(w8)	M13538	B10A	k/d2?	1iak_B		AF027865
		H2-EA*02	K00971	BALB/c	d				1fng_C		
			AY303782	BALB.K	k						
		H2-EB1*01	AF050157	129	bc						
MHC-IIb	*Hs*	HLA-DMA*01				X62744			(1 hdm_A)		X76775
		HLA-DMB*01	X76776						(1 hdm_B)		
		HLA-DOA*01	X02882								

	HLA-DOB*01	X87344					
Mm	H2-DMA*01	AF100956	129/Svj	129		1k8i_A	
	H2-DMB1*02	U35323	BALB/c	d			
	H2-DOA*01		C57BL/6J	b	AK020594		AF100956
	H2-DOB*01		C57BL/6J	b	AK053233		M11800

(A) Classical MHC-I (MHC-Ia), nonclassical MHC-I (MHC-Ib), and MHC-I-like. (B) Classical MHC-II (MHC-IIa) and nonclassical MHC-II (MHC-IIb).

[a] Owing to a lesser degree of polymorphism of the nonclassical *Homo sapiens* (*Hs*) MHC-Ib and MHC-IIb genes, compared to the classical MHC-Ia and MHC-IIa genes, a 2-digit is used for their allele description (IMGT Scientific chart, http://imgt.cines.fr): *01 refers to *0101 found in the literature. The allele nomenclature of the highly polymorphic *H. sapiens* classical MHC-Ia and MHC-IIa is according to HLA-DB [13]. *Mus musculus* (*Mm*) H2 alleles are numbered starting from allele *01 (sequence from strain C57BL/6). *H. sapiens* and *M. musculus* MHC-I-like alleles are numbered starting from allele *01 that corresponds to the IMGT reference sequence (IMGT Repertoire, http://imgt.cines.fr). The H2-PA gene (H2-PA*01 accession number D64112, strain C57BL/6) that is a pseudogene, and the H2-PB gene that has not yet been identified, are not included in the table.

[b] EMBL/GenBank/DDBJ nucleotide sequence accession numbers. gDNA: genomic DNA; cDNA: complementary DNA.

[c] From 'Mouse H2 haplotypes and polymorphisms' (IMGT Index > Strain, and IMGT Repertoire for MHC, http://imgt.cines.fr [1,6].

[d] Chain ID from IMGT/3Dstructure-DB, http://imgt.cines.fr [25]. The chain ID is shown within parentheses for 3D structures in which the chain sequence differs from the translation of the corresponding representative allele in column 3.

[e] Accession numbers that were necessary for splicing site identification: 'same allele' refers to the allele in column 3, 'other alleles' refers to alleles other than the ones in column 3.

the G-ALPHA2 domain, whereas positions 61A, 61B and 72A are occupied in G-BETA domains, at the exception of H2-AB (Table 2B). The position 92A is only occupied in the HLA-DMA and H2-DMA G-ALPHA domains (Fig. 3A, Table 2A). It is worthwhile to note that position 54A is the only additional position needed to extend the IMGT numbering for G-DOMAINs to the G-LIKE-DOMAINs of the MHC-I-like proteins (described in next paragraph and shown in Fig. 3A).

The helix (positions 50–92) seats on the beta sheet and its axis forms an angle of about 40° with the beta strands. The helix is split into two parts separated by a kink, positions 58 of G-ALPHA1, 61 of G-ALPHA2, 63 of G-ALPHA, and 62 of G-BETA being the 'highest' points on the groove floor [36].

Two cysteine, CYS-11 (in strand A) and CYS-74 (in the helix) are well conserved in the G-ALPHA2 and G-BETA domains where they participate to a disulfide bridge that fastens the helix on the groove floor (Fig. 3B). The G-ALPHA1 and G-ALPHA domains have a conserved *N*-glycosylation site at position 86 (N–X–S/T, where N is asparagine, X any amino acid except proline, S is serine and T is threonine) (Fig. 3A). A *N*-glycosylation site is also found at that position in the G-ALPHA2 domain of the mouse MHC-Ia chains (H2-D1, H2-K1 and H2-L) (Fig. 3B). The G-BETA domains (except for the HLA-DMB and H2-DMB1 chains) have a conserved potential *N*-glycosylation site at position 15 (AB turn) (Fig. 3B). Interestingly, the G-ALPHA domains of the HLA-DMA and H2-DMA chains have specific features compared to the other G-ALPHA domains and share common characteristics with the G-ALPHA2 and G-BETA domains: there is a conserved CYS-11_CYS-74 disulfide bridge, positions 61A and 61B are occupied (as in the G-BETA domains) and there is no *N*-glycosylation site at position 86 (Fig. 3).

4. IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN and sequence data analysis

The IMGT unique numbering for G-DOMAIN allows, for the first time, a standardized comparison of the amino acid (and corresponding codons) changes between the different groove domains

that belong to a same chain (G-ALPHA1 and G-ALPHA2), or to different MHC chains, and this whatever the species. Practically, the IMGT unique numbering for positions 1–39 and 73–92 of the G-ALPHA2 domains can be obtained very easily by substracting 90 from the mature protein numbering (91–129 and 163–182) (Table 3). Between these positions, the two gaps (at positions 40 and 41) and the two insertions (at positions 61A and 72A) are necessary, in the IMGT unique

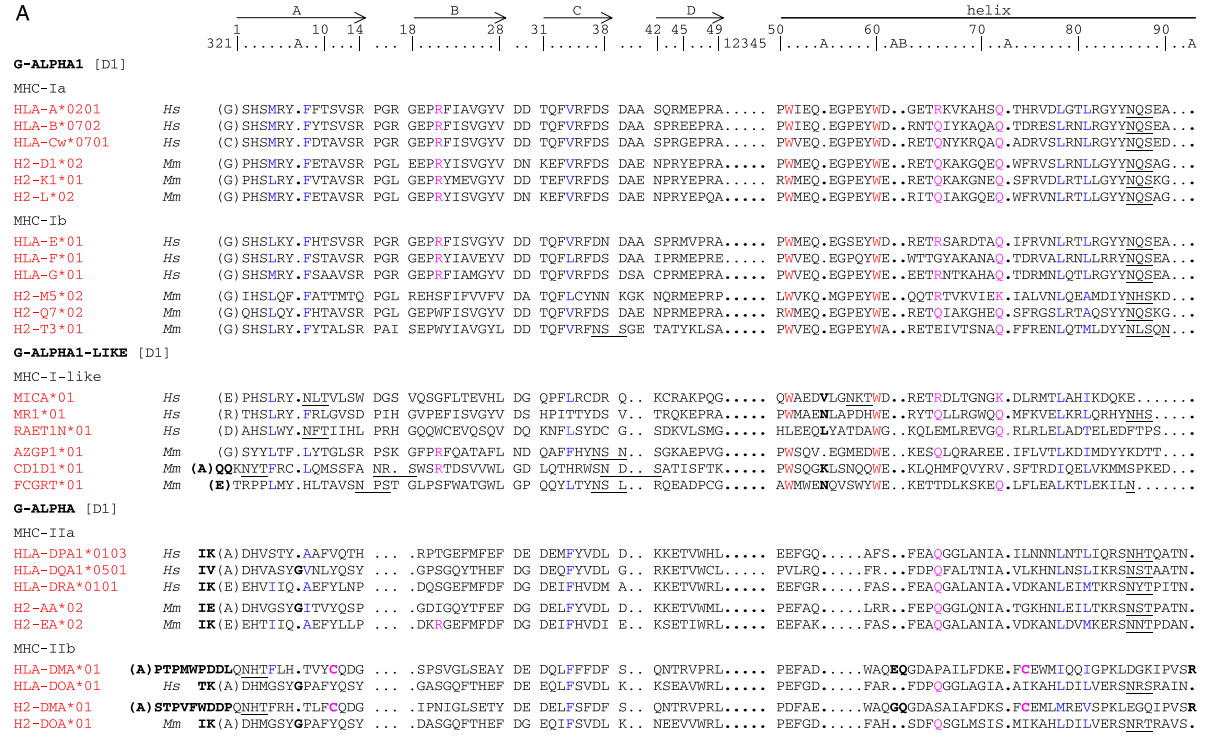

Fig. 3. IMGT Protein display of G-DOMAINs and G-LIKE-DOMAINs of representative MHC and MHC-I-like chains. (A) G-ALPHA1 [D1], G-ALPHA1-LIKE [D1] and G-ALPHA [D1] domains from classical (MHC-Ia) and nonclassical (MHC-Ib) MHC-I, from MHC-I-like, and from classical (MHC-IIa) and nonclassical (MHC-IIb) MHC-II, respectively. (B) G-ALPHA2 [D2], G-ALPHA2-LIKE [D2] and G-BETA [D1] domains from classical (MHC-Ia) and nonclassical (MHC-Ib) MHC-I, from MHC-I-like, and from classical (MHC-IIa) and nonclassical (MHC-IIb) MHC-II, respectively. [D1], [D2] and [D3] indicate the position of the domains from the N-terminal end of the chains. Membrane proteins quoted in this figure are of type I, that is with the chain N-terminal end being extracellular. Sequences are from *Homo sapiens* (*Hs*) and from *Mus musculus* (*Mm*). The IMGT Protein display is according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, based on the NUMEROTATION concept of IMGT-ONTOLOGY [17]. The G-DOMAINs and G-LIKE-DOMAINs are designated with the IMGT labels (IMGT Scientific chart, http://imgt.cines.fr). Beta strands are shown by horizontal arrows. Dots indicate missing amino acids according to the IMGT unique numbering. Amino acids resulting from a splicing with a preceding exon are shown within parentheses. EMBL/GenBank/DDBJ accession numbers are reported in Table 1. Potential *N*-glycosylation sites (N–X–S/T) are underlined. Note that the C-LIKE-DOMAIN [D2] of *M. musculus* H2-AA (NT_039649, H2-AA*01), *H. sapiens* HLA-DMA (NT_007592, HLA-DMA*01) and *M. musculus* H2-AB (NT_039649, H2-AB*01) were reported in Fig. 3 of Ref. [21]. Gene names (symbols) are according to the IMGT Nomenclature committee (IMGT-NC) [1] and to the HUGO Nomenclature Committee (HGNC) [41]. Full gene designations for the MHC genes are based on the examples shown within parentheses: human MHC-I (HLA-A: Major histocompatibility complex, class I, A), mouse MHC-I (H2-K1: histocompatibility 2, class I, K1), human MHC-II (HLA-DPA1: MHC class II, DP alpha1; HLA-DPB1: MHC class II, DP beta1), and mouse MHC-II (H2-AA: histocompatibility 2, class II, A alpha; H2-AB: histocompatibility 2, class II, A beta). Full gene designations for the MHC-I-like genes are the following: MICA, MHC class I polypeptide-related sequence A; MR1: major histocompatibility complex, class I-related; RAET1N: retinoic acid early transcript 1 N, UL16 binding protein 3; CD1D1: CD1D antigen, polypeptide 1; FCGRT: Fc fragment of IgG, receptor, transporter, alpha; AZGP1: alpha-2-glycoprotein 1, zinc. Amino acid one-letter abbreviation: A (Ala), alanine; C (Cys), cysteine; D (Asp), aspartic acid; E (Glu), glutamic acid; F (Phe), phenylalanine; G (Gly), glycine; H (His), histidine; I (Ileu), isoleucine; K (Lys), lysine; L (Leu), leucine; M (Met), methionine; N (Asn), asparagine; P (Pro), proline; Q (Gln), glutamine; R (Arg), arginine; S (Ser), serine; T (Thr), threonine; V (Val), valine; W (Trp), tryptophan; Y (Tyr), tyrosine.

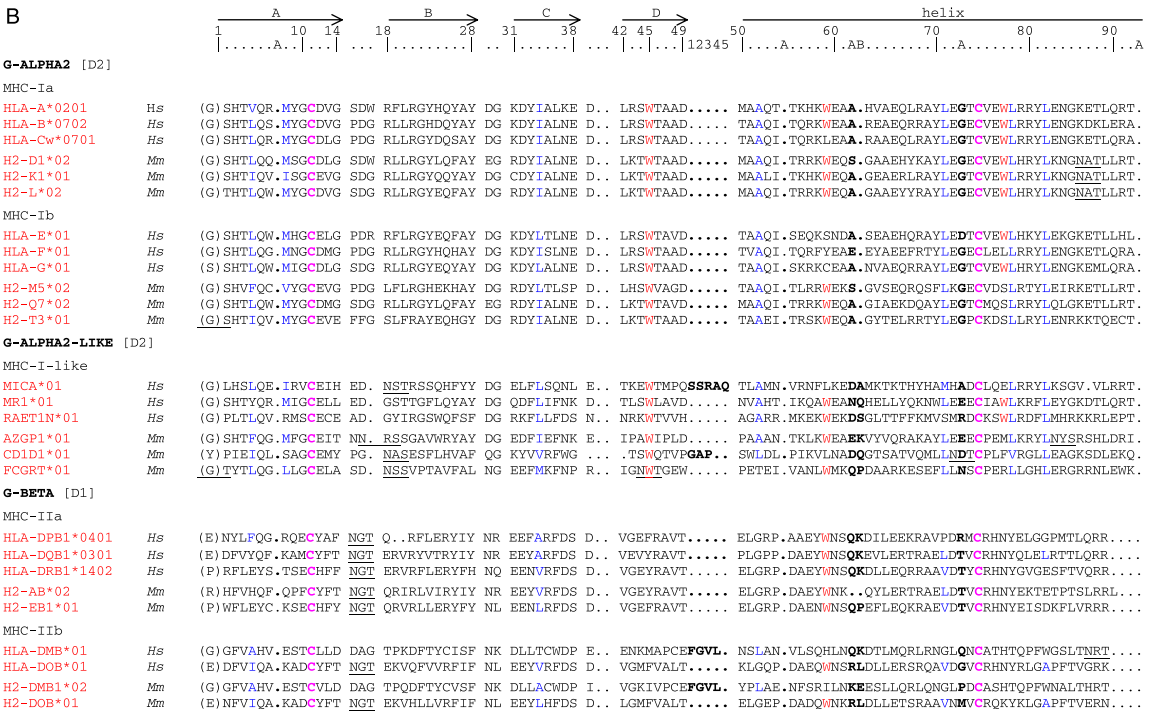

Fig. 3 (*continued*)

numbering, to allow meaningful sequence and structure alignment and comparison between the G-ALPHA1 and G-ALPHA2 sequences. Correspondence between the IMGT unique numbering for G-DOMAIN and different MHC chain numberings is shown in Table 3. Examples include the G-ALPHA1 [D1] and G-ALPHA2 [D2] domains of the classical MHC-I (MHC-Ia) HLA-A, the G-ALPHA [D1] and G-BETA [D1] domains of the classical MHC-II (MHC-IIa) HLA-DRA and HLA-DRB1, and the G-ALPHA [D1] and G-BETA [D1] domains of the nonclassical MHC-II (MHC-IIb) HLA-DMA and HLA-DMB.

The IMGT unique numbering for G-DOMAIN has been extended to the G-LIKE-DOMAINs of MhcSF proteins other than MHC. So far, only MHC-I-like proteins have been identified in the MhcSF [43–46]. The examples of MHC-I-like chains shown in Fig. 3 include the human MICA*01, MR1*01 and RAET1N*01, and the mouse AZGP1*01, CD1D1*01 and FCGRT*01. The G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] domains of these proteins show a striking structural homology with the MHC G-ALPHA1 and G-ALPHA2 domains and this, despite a high sequence divergence [46]. The implementation of the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN represents therefore a major step in the standardization of the MhcSF amino acid sequence alignments (Fig. 3). As the nucleotide positions are derived from the codon numbering, the IMGT unique numbering allows a standardized allele description and the setting up of 'Tables of alleles' and 'Alignments of alleles' for MhcSF proteins, whatever the receptor, the chain or the species (IMGT Repertoire for MHC, IMGT Repertoire for RPI, http://imgt.cines.fr). Owing to that standardization, the sequence polymorphisms of any G-DOMAIN (of any MHC gene or chain, from any vertebrate species) and the sequence polymorphisms of any G-LIKE-DOMAIN (of any MHC-I-like gene or chain, from any species) can easily be described and analysed.

Table 2
Lengths of the G-DOMAIN and G-LIKE-DOMAIN labels, according to the IMGT unique numbering

A			G-ALPHA1 [D1]				G-ALPHA1-LIKE [D1]					
			MHC-Ia		MHC-Ib		MHC-I-like					
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-A HLA-B HLA-C	H2-D1 H2-K1 H2-L	HLA-E HLA-F HLA-G	H2-M5 H2-Q7 H2-T3	MICA	MR1	RAET1N	AZGP1	CD1D1	FCGRT
			Hs	*Mm*	*Hs*	*Mm*	*Hs*	*Hs*	*Hs*	*Mm*	*Mm*	*Mm*
A-STRAND	1.3-1.1[a]	+3									+3	+1
	1-14	14	14	14	14	14	14	14	14	14	14	14
	7A	+1										
AB-TURN	15–17	3	3	3	3	3	3	3	3	3	2	3
B-STRAND	18–28	11	11	11	11	11	11	11	11	11	11	11
BC-TURN	29–30	2	2	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	3	3	3	3	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5										
HELIX	50–92	43	41	41	41	41	37	39	40	40	41	37
	54A	+1					+1	+1	+1		+1	+1
	61A	+1										
	61B	+1										
	72A	+1										
	92A	+1										
Total length		92 (+14)	90	90	90	90	86 (+1)	88 (+1)	89 (+1)	89	90 (+4)	86 (+2)

A			G-ALPHA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPA1 HLA-DRA	HLA-DQA1	H2-AA	H2-EA	HLA-DMA	HLA-DOA	H2-DMA	H2-DOA
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
A-STRAND	1.10–1.1[a]	+10	+2	+2	+2	+2	+10	+2	+10	+2
	1–14	14	14	14	14	14	14	14	14	14
	7A	+1		+1	+1			+1		+1
AB-TURN	15–17	3	0	0	0	0	0	0	0	0
B-STRAND	18–28	11	10	10	10	10	10	10	10	10
BC-TURN	29–30	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5								

HELIX	50–92	43	39	38	39	39	39	39	39	39
	54A	+1								
	61A	+1					+1		+1	
	61B	+1					+1		+1	
	72A	+1								
	92A	+1					+1		+1	
Total length		92 (+21)	82 (+2)	81 (+3)	82 (+3)	82 (+2)	82 (+13)	82 (+3)	82 (+13)	82 (+3)

B

			G-ALPHA [D2]				G-ALPHA2-LIKE [D2]					
			MHC-Ia		MHC-Ib		MHC-I-like					
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-A HLA-B HLA-C	H2-D1 H2-K1 H2-L	HLA-E HLA-F HLA-G	H2-M5 H2-Q7 H2-T3	MICA	MR1	RAET1N	AZGP1	CD1D1	FCGRT
			Hs	*Mm*	*Hs*	*Mm*	*Hs*	*Hs*	*Hs*	*Mm*	*Mm*	*Mm*
A-STRAND	1–14	14	14	14	14	14	14	14	14	14	14	14
	7A	+1										
AB-TURN	15–17	3	3	3	3	3	2	2	2	2	2	2
B-STRAND	18–28	11	11	11	11	11	11	11	11	11	11	11
BC-TURN	29–30	2	2	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1	0	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8	7	8
	49.1–49.5	+5					+5				+3	
HELIX	50–92	43	43	43	43	43	42	43	43	43	43	43
	54A	+1										
	61A	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
	61B	+1					+1	+1	+1	+1	+1	+1
	72A	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1
	92A	+1										
Total length		92 (+11)	90 (+2)	90 (+2)	90 (+2)	90 (+2)	88 (+8)	89 (+3)	89 (+3)	89 (+3)	87 (+6)	89 (+3)

B

			G-BETA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPB1	HLA-DQB1 HLA-DRB1	H2-AB	H2-EB1	HLA-DMB	HLA-DOB	H2-DMB1	H2-DOB
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
A-STRAND	1–14	14	14	14	14	14	14	14	14	14
	7A	+1								
AB-TURN	15–17	3	3	3	3	3	3	3	3	3
B-STRAND	18–28	11	9	11	11	11	11	11	11	11

(*continued on next page*)

Table 2 (*continued*)

B			G-BETA [D1]							
			MHC-IIa				MHC-IIb			
G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering	Domain maximal length	HLA-DPB1	HLA-DQB1 HLA-DRB1	H2-AB	H2-EB1	HLA-DMB	HLA-DOB	H2-DMB1	H2-DOB
			Hs	*Hs*	*Mm*	*Mm*	*Hs*	*Hs*	*Mm*	*Mm*
BC-TURN	29–30	2	2	2	2	2	2	2	2	2
C-STRAND	31–38	8	8	8	8	8	8	8	8	8
CD-TURN	39–41	3	1	1	1	1	1	1	1	1
D-STRAND	42–49	8	8	8	8	8	8	8	8	8
	49.1–49.5	+5					+4		+4	
HELIX	50–92	43	40	40	41	40	40	40	40	40
	54A	+1								
	61A	+1	+1	+1		+1	+1	+1	+1	+1
	61B	+1	+1	+1		+1	+1	+1	+1	+1
	72A	+1	+1	+1	+1	+1	+1	+1	+1	+1
	92A	+1								
Total length		92 (+11)	85 (+3)	87 (+3)	88 (+1)	87 (+3)	87 (+7)	87 (+3)	87 (+7)	87 (+3)

(A) G-ALPHA1 [D1] domains of MHC-Ia and MHC-Ib chains, G-ALPHA1-LIKE [D1] domains of MHC-I-like chains, and G-ALPHA [D1] domains of MHC-IIa and MHC-IIb chains. (B) G-ALPHA2 [D2] domains of MHC-Ia and MHC-Ib chains, G-ALPHA2-LIKE [D2] domains of MHC-I-like chains, and G-BETA [D1] domains of MHC-IIa and MHC-IIb chains. A plus (+) sign indicates additional positions (see text).

[a] The amino acids at positions 1.2 and 1.1 for HLA-DRA were identified by amino acid sequencing [42] and extrapolated for the G-ALPHA [D1] domains of the other MHC-IIa chains and for the MHC-IIb HLA-DQA and H2-DQA chains. These additional amino acids, and those at positions 1.10–1.1 of the G-ALPHA [D1] domain of the HLA-DMA and H2-DMA chains, need to be confirmed experimentally in the mature chain. This also concerns positions 1.3–1.1 of the G-ALPHA1-LIKE [D1] domain of the *M. musculus* CD1D1 and positions 1.1 of *M. musculus* FCGRT.

Table 3
Correspondence between the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, and MHC-Ia, MHC-IIa and MHC-IIb chain numberings

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
A-STRAND	1.10					1 gct ALA (A)	
	1.9					2 cct PRO P	
	1.8					3 act THR T	
	1.7					4 cca PRO P	
	1.6					5 atg MET M	
	1.5					6 tgg TRP W	
	1.4					7 cca PRO P	
	1.3					8 gat ASP D	
	1.2			1 atc ILE I		9 gac ASP D	
	1.1			2 aaa LYS K		10 ctg LEU L	
	1	1 ggc GLY (G)	91 ggt GLY (G)	3 gaa GLU (E)	1 cca PRO (P)	11 caa GLN Q	1 ggt GLY (G)
	2	2 tct SER S	92 tct SER S	4 gaa GLU E	2 cgt ARG R	12 aac ASN N	2 ggc GLY G
	3	3 cac HIS H	93 cac HIS H	5 cat HIS H	3 ttc PHE F	13 cac HIS H	3 ttc PHE F
	4	4 tcc SER S	94 acc THR T	6 gtg VAL V	4 ttg LEU L	14 aca THR T	4 gtg VAL V
	5	5 atg MET M	95 gtc VAL V	7 atc ILE I	5 gag GLU E	15 ttc PHE F	5 gcc ALA A
	6	6 agg ARG R	96 cag GLN Q	8 atc ILE I	6 tac TYR Y	16 ctg LEU L	6 cat HIS H
	7	7 tat TYR Y	97 agg ARG R	9 cag GLN Q	7 tct SER S	17 cac HIS H	7 gtg VAL V
	7A	–	–	–	–	–	–
	8	8 ttc PHE F	98 atg MET M	10 gcc ALA A	8 acg THR T	18 aca THR T	8 gaa GLU E
	9	9 ttc PHE F	99 tat TYR Y	11 gag GLU E	9 tct SER S	19 gtg VAL V	9 agc SER S
	10	10 aca THR T	100 ggc GLY G	12 ttc PHE F	10 gag GLU E	20 tac TYR Y	10 acc THR T
	11	11 tcc SER S	101 tgc CYS C	13 tat TYR Y	11 tgt CYS C	21 tgc CYS C	11 tgt CYS C
	12	12 gtg VAL V	102 gac ASP D	14 ctg LEU L	12 cat HIS H	22 cag GLN Q	12 ctg LEU L
	13	13 tcc SER S	103 gtg VAL V	15 aat ASN N	13 ttc PHE F	23 gat ASP D	13 ttg LEU L
	14	14 cgg ARG R	104 ggg GLY G	16 cct PRO P	14 ttc PHE F	24 ggg GLY G	14 gat ASP D
AB-TURN	15	15 ccc PRO P	105 tcg SER S	–	15 aat ASN N	–	15 gat ASP D
	16	16 ggc GLY G	106 gac ASP D	–	16 ggg GLY G	–	16 gct ALA A
	17	17 cgc ARG R	107 tgg TRP W	–	17 acg THR T	–	17 ggg GLY G
B-STRAND	18	18 ggg GLY G	108 cgc ARG R	–	18 gag GLU E	–	18 act THR T
	19	19 gag GLU E	109 ttc PHE F	17 gac ASP D	19 cgg ARG R	25 agt SER S	19 cca PRO P
	20	20 ccc PRO P	110 ctc LEU L	18 caa GLN Q	20 gtg VAL V	26 ccc PRO P	20 aag LYS K
	21	21 cgc ARG R	111 cgc ARG R	19 tca SER S	21 cgg ARG R	27 agt SER S	21 gat ASP D
	22	22 ttc PHE F	112 ggg GLY G	20 ggc GLY G	22 ttc PHE F	28 gtg VAL V	22 ttc PHE F
	23	23 atc ILE I	113 tac TYR Y	21 gag GLU E	23 ctg LEU L	29 gga GLY G	23 aca THR T

(*continued on next page*)

Table 3 (*continued*)

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
	24	24 gca ALA A	114 cac HIS H	22 ttt PHE F	24 gag GLU E	30 ctc LEU L	24 tac TYR Y
	25	25 gtg VAL V	115 cag GLN Q	23 atg MET M	25 aga ARG R	31 tct SER S	25 tgc CYS C
	26	26 ggc GLY G	116 tac TYR Y	24 ttt PHE F	26 tac TYR Y	32 gag GLU E	26 atc ILE I
	27	27 tac TYR Y	117 gcc ALA A	25 gac ASP D	27 ttc PHE F	33 gcc ALA A	27 tcc SER S
	28	28 gtg VAL V	118 tac TYR Y	26 ttt PHE F	28 cat HIS H	34 tac TYR Y	28 ttc PHE F
BC-TURN	29	29 gac ASP D	119 gac ASP D	27 gat ASP D	29 aac ASN N	35 gac ASP D	29 aac ASN N
	30	30 gac ASP D	120 ggc GLY G	28 ggt GLY G	30 cag GLN Q	36 gag GLU E	30 aag LYS K
C-STRAND	31	31 acg THR T	121 aag LYS K	29 gat ASP D	31 gag GLU E	37 gac ASP D	31 gat ASP D
	32	32 cag GLN Q	122 gat ASP D	30 gag GLU E	32 gag GLU E	38 cag GLN Q	32 ctg LEU L
	33	33 ttc PHE F	123 tac TYR Y	31 att ILE I	33 aac ASN N	39 ctt LEU L	33 ctg LEU L
	34	34 gtg VAL V	124 atc ILE I	32 ttc PHE F	34 gtg VAL V	40 ttc PHE F	34 acc THR T
	35	35 cgg ARG R	125 gcc ALA A	33 cat HIS H	35 cgc ARG R	41 ttc PHE F	35 tgc CYS C
	36	36 ttc PHE F	126 ctg LEU L	34 gtg VAL V	36 ttc PHE F	42 ttc PHE F	36 tgg TRP W
	37	37 gac ASP D	127 aaa LYS K	35 gat ASP D	37 gac ASP D	43 gac ASP D	37 gat ASP D
	38	38 agc SER S	128 gag GLU E	36 atg MET M	38 agc SER S	44 ttt PHE F	38 cca PRO P
CD-TURN	39	39 gac ASP D	129 gac ASP D	37 gca ALA A	39 gac ASP D	45 tcc SER S	39 gag GLU E
	40	40 gcc ALA A	–	–	–	–	–
	41	41 gcg ALA A	–	–	–	–	–
D-STRAND	42	42 agc SER S	130 ctg LEU L	38 aag LYS K	40 gtg VAL V	46 cag GLN Q	40 gag GLU E
	43	43 cag GLN Q	131 cgc ARG R	39 aag LYS K	41 ggg GLY G	47 aac ASN N	41 aat ASN N
	44	44 agg ARG R	132 tct SER S	40 gag GLU E	42 gag GLU E	48 act THR T	42 aag LYS K
	45	45 atg MET M	133 tgg TRP W	41 acg THR T	43 tac TYR Y	49 cgg ARG R	43 atg MET M
	46	46 gag GLU E	134 acc THR T	42 gtc VAL V	44 cgg ARG R	50 gtg VAL V	44 gcc ALA A
	47	47 ccg PRO P	135 gcg ALA A	43 tgg TRP W	45 gcg ALA A	51 cct PRO P	45 cct PRO P
	48	48 cgg ARG R	136 gcg ALA A	44 cgg ARG R	46 gtg VAL V	52 cgc ARG R	46 tgc CYS C
	49	49 gcg ALA A	137 gac ASP D	45 ctt LEU L	47 acg THR T	53 ctg LEU L	47 gaa GLU E
	49.1	–	–	–	–	–	48 ttt PHE F
	49.2	–	–	–	–	–	49 ggg GLY G
	49.3	–	–	–	–	–	50 gtg VAL V
	49.4	–	–	–	–	–	51 ctg LEU L
	49.5	–	–	–	–	–	–
HELIX	50	50 ccg PRO P	138 atg MET M	46 gaa GLU E	48 gag GLU E	54 ccc PRO P	52 aat ASN N
	51	51 tgg TRP W	139 gca ALA A	47 gaa GLU E	49 ctg LEU L	55 gaa GLU E	53 agc SER S
	52	52 ata ILE I	140 gct ALA A	48 ttt PHE F	50 ggg GLY G	56 ttt PHE F	54 ttg LEU L

53	53 gag GLU E	141 cag GLN Q	49 gga GLY G	51 cgg ARG R	57 gct ALA A	55 gcg ALA A
54	54 cag GLN Q	142 acc THR T	50 cga ARG R	52 cct PRO P	58 gac ASP D	56 aat ASN N
54A	–	–	–	–	–	-
55	55 gag GLU E	143 acc THR T	–	53 gat ASP D	–	57 gtc VAL V
56	56 ggt GLY G	144 aag LYS K	–	54 gcc ALA A	–	58 ctc LEU L
57	57 ccg PRO P	145 cac HIS H	–	55 gag GLU E	–	59 tca SER S
58	58 gag GLU E	146 aag LYS K	–	56 tac TYR Y	–	60 cag GLN Q
59	59 tat TYR Y	147 tgg TRP W	51 ttt PHE F	57 tgg TRP W	59 tgg TRP W	61 cac HIS H
60	60 tgg TRP W	148 gag GLU E	52 gcc ALA A	58 aac ASN N	60 gct ALA A	62 ctc LEU L
61	61 gac ASP D	149 gcg ALA A	53 agc SER S	59 agc SER S	61 cag GLN Q	63 aac ASN N
61A	–	150 gcc ALA A	–	60 cag GLN Q	62 gaa GLU E	64 caa GLN Q
61B	–	–	–	61 aag LYS K	63 cag GLN Q	65 aaa LYS K
62	62 ggg GLY G	151 cat HIS H	54 ttt PHE F	62 gac ASP D	64 gga GLY G	66 gac ASP D
63	63 gag GLU E	152 gtg VAL V	55 gag GLU E	63 ctc LEU L	65 gat ASP D	67 acc THR T
64	64 aca THR T	153 gcg ALA A	56 gct ALA A	64 ctg LEU L	66 gct ALA A	68 ctg LEU L
65	65 cgg ARG R	154 gag GLU E	57 caa GLN Q	65 gag GLU E	67 cct PRO P	69 atg MET M
66	66 aaa LYS K	155 cag GLN Q	58 ggt GLY G	66 cag GLN Q	68 gcc ALA A	70 cag GLN Q
67	67 gtg VAL V	156 ttg LEU L	59 gca ALA A	67 agg ARG R	69 att ILE I	71 cgc ARG R
68	68 aag LYS K	157 aga ARG R	60 ttg LEU L	68 cgg ARG R	70 tta LEU L	72 ttg LEU L
69	69 gcc ALA A	158 gcc ALA A	61 gcc ALA A	69 gcc ALA A	71 ttt PHE F	73 cgc ARG R
70	70 cac HIS H	159 tac TYR Y	62 aac ASN N	70 gcg ALA A	72 gac ASP D	74 aat ASN N
71	71 tca SER S	160 ctg LEU L	63 ata ILE I	71 gtg VAL V	73 aaa LYS K	75 ggg GLY G
72	72 cag GLN Q	161 gag GLU E	64 gct ALA A	72 gac ASP D	74 gag GLU E	76 ctt LEU L
72A	–	162 ggc GLY G	–	73 acc THR T	–	77 cag GLN Q
73	73 act THR T	163 acg THR T	65 gtg VAL V	74 tac TYR Y	75 ttc PHE F	78 aat ASN N
74	74 cac HIS H	164 tgc CYS C	66 gac ASP D	75 tgc CYS C	76 tgc CYS C	79 tgt CYS C
75	75 cga ARG R	165 gtg VAL V	67 aaa LYS K	76 aga ARG R	77 gag GLU E	80 gcc ALA A
76	76 gtg VAL V	166 gag GLU E	68 gcc ALA A	77 cac HIS H	78 tgg TRP W	81 aca THR T
77	77 gac ASP D	167 tgg TRP W	69 aac ASN N	78 aac ASN N	79 atg MET M	82 cac HIS H
78	78 ctg LEU L	168 ctc LEU L	70 ctg LEU L	79 tac TYR Y	80 atc ILE I	83 acc THR T
79	79 ggg GLY G	169 cgc ARG R	71 gaa GLU E	80 ggg GLY G	81 cag GLN Q	84 cag GLN Q
80	80 acc THR T	170 aga ARG R	72 atc ILE I	81 gtt VAL V	82 caa GLN Q	85 ccc PRO P
81	81 ctg LEU L	171 tac TYR Y	73 atg MET M	82 ggt GLY G	83 ata ILE I	86 ttc PHE F
82	82 cgc ARG R	172 ctg LEU L	74 aca THR T	83 gag GLU E	84 ggg GLY G	87 tgg TRP W
83	83 ggc GLY G	173 gag GLU E	75 aag LYS K	84 agc SER S	85 cca PRO P	88 gga GLY G
84	84 tac TYR Y	174 aac ASN N	76 cgc ARG R	85 ttc PHE F	86 aaa LYS K	89 tca SER S
85	85 tac TYR Y	175 ggg GLY G	77 tcc SER S	86 aca THR T	87 ctt LEU L	90 ctg LEU L
86	86 aac ASN N	176 aag LYS K	78 aac ASN N	87 gtg VAL V	88 gat ASP D	91 acc THR T
87	87 cag GLN Q	177 gag GLU E	79 tat TYR Y	88 cag GLN Q	89 ggg GLY G	92 aac ASN N
88	88 agc SER S	178 acg THR T	80 act THR T	89 cgg ARG R	90 aaa LYS K	93 agg ARG R
89	89 gag GLU E	179 ctg LEU L	81 ccg PRO P	90 cga ARG R	91 atc ILE I	94 aca THR T
90	90 gcc ALA A	180 cag GLN Q	82 atc ILE I	–	92 ccg PRO P	–

(*continued on next page*)

Table 3 (*continued*)

IMGT G-DOMAIN and G-LIKE-DOMAIN labels	IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN[a]	MHC-Ia		MHC-IIa		MHC-IIb	
		I-ALPHA chain		II-ALPHA chain	II-BETA chain	II-ALPHA chain	II-BETA chain
		HLA-A*0201		HLA-DRA*0101	HLA-DRB1*1402	HLA-DMA*01	HLA-DMB*01
		G-ALPHA1 [D1] domain	G-ALPHA2 [D2] domain	G-ALPHA [D1] domain	G-BETA [D1] domain	G-ALPHA [D1] domain	G-BETA [D1] domain
	91	–	181 cgc ARG R	83 acc THR T	–	93 gtg VAL V	–
	92	–	182 acg THR T	84 aat ASN N	–	94 tcc SER S	–
	92A	–	–	–	–	95 aga ARG R	–

Unoccupied positions according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN are shown with dashes. The codon encoding the amino acid at position 1 results from the splicing between EX1 and EX2, except for the HLA-DMA G-ALPHA domain, where it is the codon at position 1.10. The presence of I (1.2) and K (1.1) in the HLA-DRA G-ALPHA [D1] domain has been demonstrated experimentally by amino acid sequencing [42]. EMBL/GenBank/DDBJ accession numbers of (HLA-A*020101): K02883, HLA-DRA*0101: J00204, HLA-DRB1*1402: AJ297583, HLA-DMA*01: X62744, HLA-DMB*01: X76776. The allele nomenclature of the *H. sapiens* classical MHC-Ia and MHC-IIa is according to HLA-DB [13]. Owing to a lesser degree of polymorphism of the nonclassical *Homo sapiens* MHC-Ib and MHC-IIb genes, compared to the classical MHC-Ia and MHC-IIa genes, a 2-digit is used for their allele description (IMGT Scientific chart, http://imgt.cines.fr): *01 refers to *0101 found in the literature.

[a] IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN, first defined in 2002 by Marie-Paule Lefranc, Université Montpellier II, CNRS (IMGT http://imgt.cines.fr, online 15/05/2002), and this paper.

5. IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN and structural data comparison

Beyond sequence data comparison, the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN provides information on the strand, turn and helix lengths (Table 2) and allows standardized 2D graphical representations or IMGT Colliers de Perles for G-DOMAIN and G-LIKE-DOMAIN. Fig. 4 shows, as examples, the IMGT Colliers de Perles for the G-ALPHA1 and G-ALPHA2 domains of the MHC-Ia *Homo sapiens* HLA-B*0702 and *Mus musculus* H2-K1*01 (Fig. 4A), for the G-ALPHA and G-BETA domains of the MHC-IIa *Homo sapiens* HLA-DQA1*0501/HLA-DQB1*0301 and *Mus musculus* H2-AA*02/H2-AB*02 (Fig. 4B), and for the G-ALPHA1-LIKE and G-ALPHA2-LIKE domains of the MHC-I-like *Homo sapiens* MICA*01 and *Mus musculus* CD1D1*01 (Fig.4C).

Structural data comparison is straightforward using the IMGT unique numbering. Indeed, intra- and inter-domain contact analysis can be analysed and compared for any position between any G-DOMAIN or G-LIKE-DOMAIN [25,36]. This standardization not only allows the structural characterization of a position inside a domain, but also the statistical analysis of amino acid properties, position per position, between domains, as this has been demonstrated for the IG V-DOMAINs [47]. Eleven IMGT amino acid classes have been defined, based on the hydropathy, the volume and the chemical characteristics of the 20 common amino acids [47]. These classes, initially defined for a standardized comparison of the properties of the IG, TR and IgSF chains and domains will be used, with the IMGT unique numbering, for the comparison of the MHC and MhcSF chains and domains.

6. Conclusion

The IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN allows, for the first time, to compare any G-DOMAIN of MHC and G-LIKE-DOMAIN of MHC-I-like proteins, or in other words, to compare any G-set domain of MhcSF proteins. This is the third major breakthrough for domain analysis

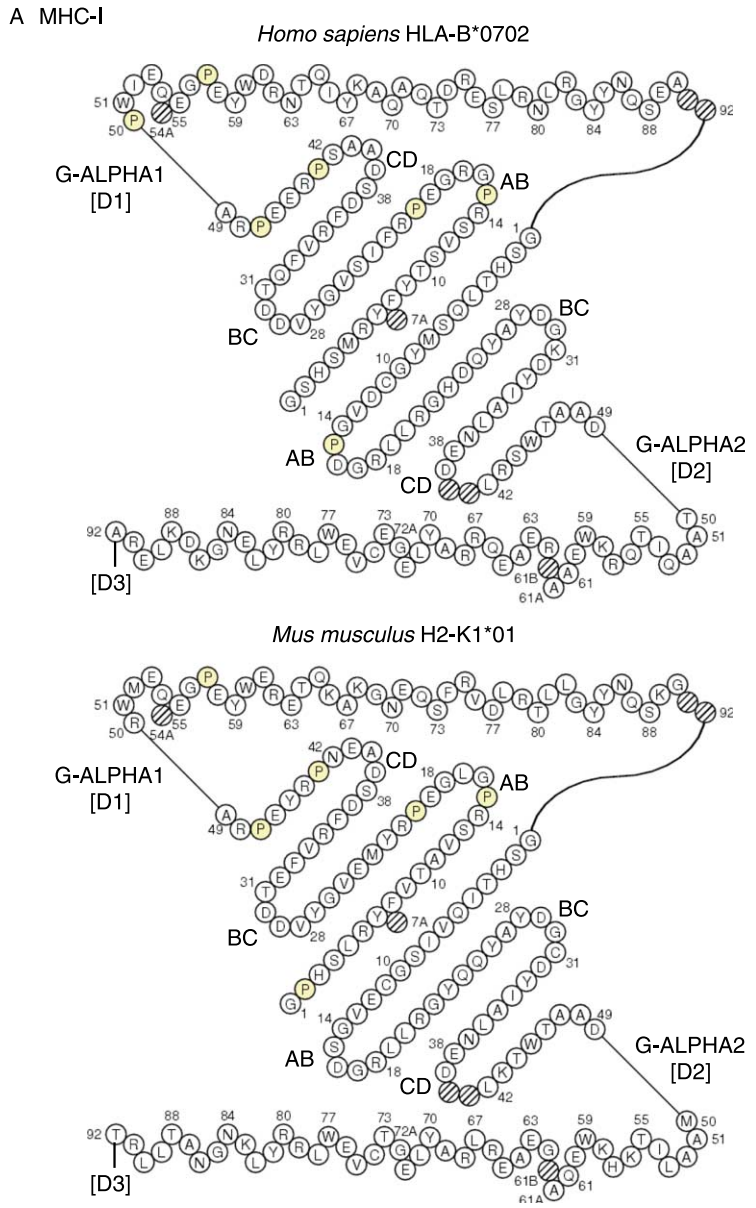

Fig. 4. IMGT Collier de Perles of G-DOMAINs and G-LIKE-DOMAINs. (A) MHC-I G-ALPHA1 [D1] and G-ALPHA2 [D2] domains of the *Homo sapiens* HLA-B*0702 and *Mus musculus* H2-K1*01 chains. (B) MHC-II G-ALPHA [D1] and G-BETA D1] domains of the *H. sapiens* HLA-DQA1*0501/HLA-DQB1*0301 and *M. musculus* H2-AA*02/H2-AA*02 chains. (C) MHC-I-like G-ALPHA1-LIKE [D1] and G-ALPHA2-LIKE [D2] domains of the *H. sapiens* MICA*01 and *M. musculus* CD1D1*01 chains. The amino acid at position 1 is encoded by a codon which results from the splicing with the preceding exon, except for the CD1D1 G-ALPHA1-LIKE domain, for which it is position 1.3. Amino acids are shown in the one-letter abbreviation. Hatched circles correspond to missing positions according to the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN. In IMGT Colliers de Perles, position 7A is only displayed in the G-ALPHA and G-ALPHA1-LIKE domains, and positions 61A and 61B in the G-BETA and G-ALPHA2-LIKE domains. As position 54A is only occupied in G-ALPHA1-LIKE of MHC-I-like proteins, this position can be omitted in IMGT Colliers de Perles, if only MHC chains are compared [36]. Position 92A is only added for MHC-DMA and H2-DMA IMGT Colliers de Perles. Note that the N-terminal end of a peptide in the cleft would be on the left hand side.

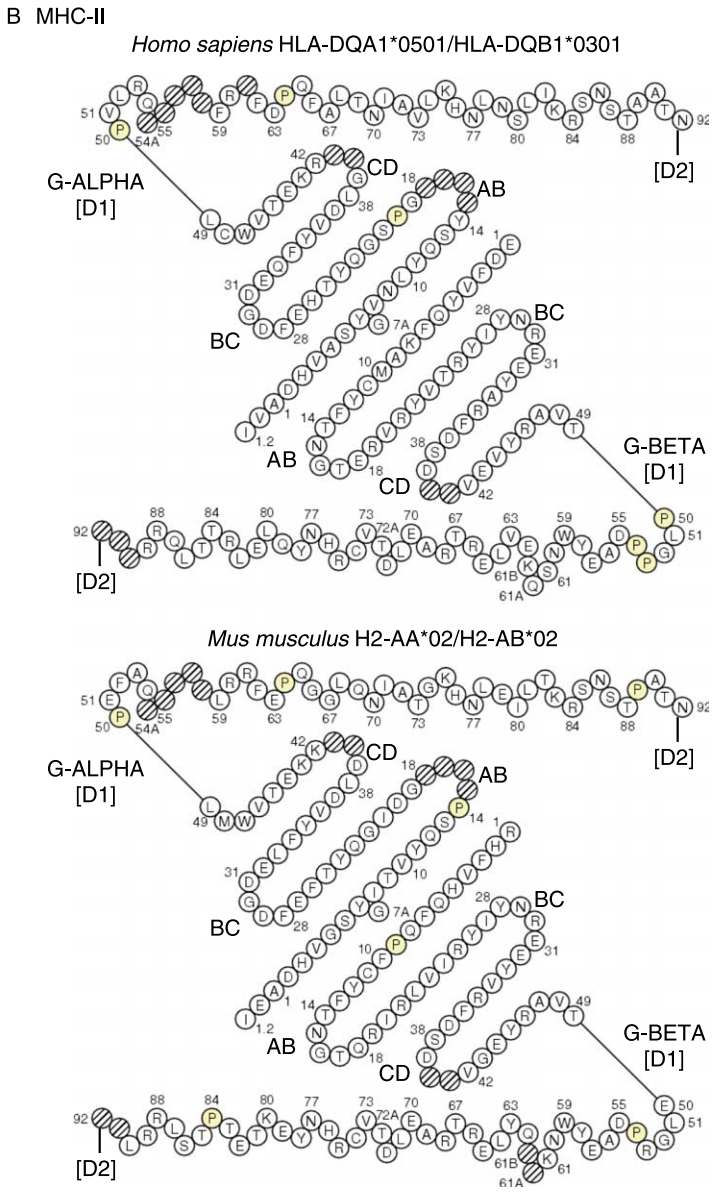

Fig. 4 (*continued*)

using the IMGT unique numbering, based on the NUMEROTATION concept of IMGT-ONTOLOGY [17]. Indeed, this completes the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN that allows to compare any V-set domain of IgSF proteins (V-DOMAIN of IG and TR, and V-LIKE-DOMAIN of IgSF proteins other than the IG or TR) [20,35], and the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN that allows to compare any C-set domain of IgSF proteins (C-DOMAIN of IG and TR, and C-LIKE-DOMAIN of IgSF other than the IG or TR) [21,35].

The IMGT unique numbering has many advantages. Three features are worth noting: (i) In IMGT,

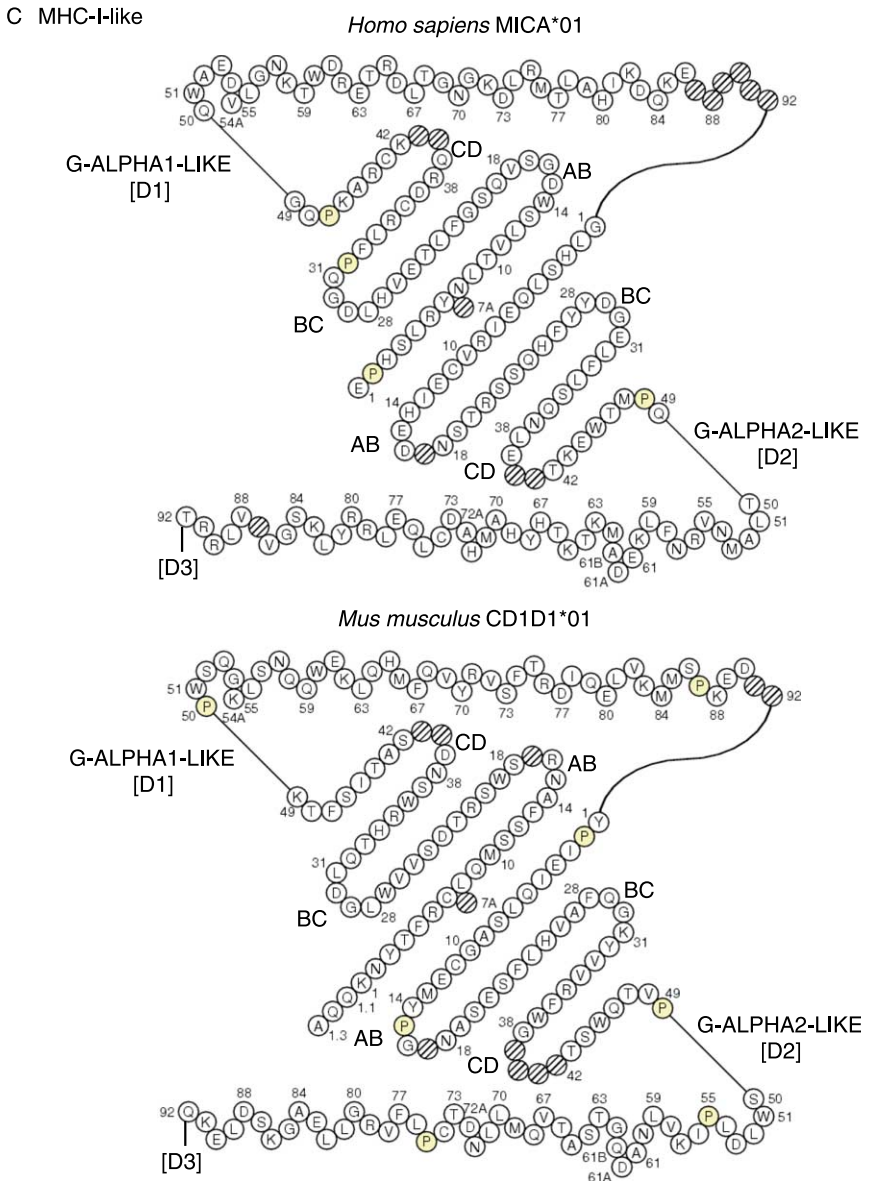

Fig. 4 (*continued*)

any domain is characterized by the length of its strands, loops and turns and, for the G-set, by the length of its helix. The strand, loop, turn or helix lengths (the number of codons or amino acids, that is the number of occupied positions) become crucial information which characterizes the domains. This first feature of the IMGT standardization based on the IMGT unique numbering allowed, for instance, to show that the distinction between the C1, C2, I1 and I2 types found in the literature and in the databases to describe the IgSF C-set domains is unapplicable when dealing with sequences for which no structural data are known (discussed in [21]). (ii) A second feature of the IMGT standardization is the comparison of cDNA

and/or amino acid sequences with genomic sequences, and the identification of the splicing sites, to delimit precisely the domains: a G-DOMAIN or a G-LIKE-DOMAIN is frequently encoded by a unique exon, as this is the case for the V-LIKE-DOMAINs [20,35], C-DOMAINs and C-LIKE-DOMAINs [21,35]. This IMGT standardization for the domain delimitation explains the discrepancies observed with the generalist Swiss-Prot database which does not take into account this criteria. (iii) At last, a third feature is the IMGT Collier de Perles which, in the absence of available 3D structures, is particularly useful to compare domains of very diverse families.

The IMGT unique numbering allows standardized representations of nucleotide and amino acid sequences in the IMGT Web resources, and more particularly in the IMGT Repertoire (http://imgt.cines.fr) (Tables of alleles, Alignments of alleles, IMGT Protein displays, IMGT Colliers de Perles, 3D structures). The IMGT unique numbering is extensively used in the IMGT databases [1,22–25] and sequence and structure analysis tools [25–28]. It is a key element of the IMGT strategy for the automatic annotation of nucleotide sequences and standardized label assignment [48]. The IMGT unique numbering represents, therefore, a major step forward in the analysis and comparison of the MhcSF domains as this was already demonstrated for the IgSF domains [20,21,35]. By providing a unique frame for the structural analysis of the V-set, C-set and G-set domains, the IMGT standardized approach opens new insight on the evolution of these domains and of their functional interactions.

Acknowledgements

We are grateful to Chantal Ginestoux, Véronique Giudicelli, Joumana Jabado–Michaloud, Géraldine Folch, Vincent Negre, Oliver Clément and Denys Chaume for helpful discussion. E.D. is holder of a doctoral grant from the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR). K.Q. was the recipient of a doctoral grant from the MENESR and is currently supported by a grant from the Association pour la Recherche sur le Cancer (ARC). IMGT is a registered mark of the Centre National de la Recherche Scientifique (CNRS). IMGT is a RIO platform since 2001 (CNRS, INSERM, CEA, INRA). IMGT was funded in part by the BIOMED1 (BIOCT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-2000-01287) programmes of the European Union and received subventions from ARC and from the Génopole-Montpellier-Languedoc-Roussillon. IMGT is currently supported by the CNRS, the MENESR (Université Montpellier II Plan Pluri-Formation, BIOSTIC-LR2004 Région Languedoc-Roussillon and ACI-IMPBIO IMP82-2004). Part of this work was carried out in the frame of the European Science Foundation Scientific Network Myelin Structure and its role in autoimmunity (MARIE).

References

[1] Lefranc M-P, Giudicelli V, Kaas Q, Duprat E, Jabado-Michaloud J, Scaviner D, et al. IMGT, the international ImMunoGeneTics information system®. Nucleic Acids Res 2005;33:D593–D7.

[2] Warr GW, Clem LW, Soderhall K. The international ImMunoGeneTics database IMGT. Dev Comp Immunol 2003;27:1.

[3] Lefranc M-P. IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 2002; 26:697–705.

[4] Lefranc M-P. IMGT-ONTOLOGY and MGT databases, tools and web resources for immunogenetics and immunoinformatics. Mol Immunol 2004;40:647–59.

[5] Lefranc M-P, Giudicelli V, Ginestoux C, Bosc N, Folch G, Guiraudou D, et al. IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biology 2003;4,0004. http://www.bioinfo.de/isb/2003/04/2004/. *In Silico* Biology 2004;4,17–29.

[6] Lefranc M-P, Clément O, Kaas Q, Duprat E, Chastellan P, Coelho I, et al. IMGT-Choreography for Immunogenetics and Immunoinformatics (http://imgt.cines.fr). *In Silico* Biology. http://www.bioinfo.de/isb/2004/05/0006/.

[7] Lefranc M-P. IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis Leukemia. 2003;17:260–6.

[8] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr<http://imgt.cines.fr/>. In: Bock G, Goode J, editors. Immunoinformatics: bioinformatics strategies for better understanding of immune function, Novartis foundation symposium 254. Chichester, UK: Wiley; 2003. p. 126–36 (discussion pp.136–142, 216–222, 250–252).

[9] Lefranc M-P. IMGT, the international ImMunoGeneTics information system® http://imgt.cines.fr <http://imgt.cines.fr/>. In: Lo BKC, editor. Antibody engineering: methods and protocols. 2nd ed. Methods in molecular biology, 2nd ed, vol. 248. Totowa, NJ: Humana Press; 2003. p. 27–49 chap 3.
[10] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1997;25:206–11.
[11] Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the International ImMunoGeneTics database. Nucleic Acids Res 1998;26:297–303.
[12] Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 1999;27:209–12.
[13] Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, et al. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2000;28:219–21.
[14] Lefranc M-P. IMGT ImMunoGeneTics Database. Int BIOforum 2000;4:98–100.
[15] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2001;29:207–9.
[16] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 2003;31:307–10.
[17] Giudicelli V, Lefranc M-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999;15:1047–54.
[18] Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol Today 1997;18:509.
[19] Lefranc M-P. The IMGT unique numbering for Immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 1999;7:132–6.
[20] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, et al. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 2003;27:55–77.
[21] Lefranc M-P, Pommié C, Kaas Q, Duprat E, Bosc N, Guiraudou D, et al. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev Comp Immunol 2005;29:185–203.
[22] Chaume D, Giudicelli V, Lefranc M-P. IMGT/LIGM-DB. In: The molecular biology database collection. Nucleic Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/504.
[23] Folch G, Bertrand J, Lemaitre M, Lefranc M-P. IMGT/PRIMER-DB. In: The molecular biology database Collection. Nucleic Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/505.
[24] Giudicelli V, Chaume D, Lefranc M-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res 2005;33:D256–D61.
[25] Kaas Q, Ruiz M, Lefranc M-P. IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. Nucleic Acids Res 2004;32:D208–D10.
[26] Giudicelli V, Chaume D, Lefranc M-P. IMGT/V-QUEST, an integrated software for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. Nucleic Acids Res 2004;32:W435–W40.
[27] Yousfi Monod M, Giudicelli V, Chaume D, Lefranc M-P. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics 2004;20:I379–I385.
[28] Elemento O, Lefranc M-P. IMGT/PhyloGene: an online software package for phylogenetic analysis of immunoglobulin and T cell receptor genes. Dev Comp Immunol 2003;27:763–79.
[29] Scaviner D, Barbié V, Ruiz M, Lefranc M-P. Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 1999;16:234–40.
[30] Folch G, Scaviner D, Contet V, Lefranc M-P. Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. Exp Clin Immunogenet 2000;17:205–15.
[31] Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 2002;53:857–83.
[32] Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London, UK: Academic Press; 2001, 458 pages.
[33] Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London, UK: Academic Press; 2001, 398 pages.
[34] Bertrand G, Duprat E, Lefranc M-P, Marti J, Coste J. Human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. Tissue Antigens 2004;64:119–31.
[35] Duprat E, Kaas Q, Garelle V, Giudicelli V, Lefranc G, Lefranc M-P. IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: Recent research developments in human genetics. Trivandrum, India: Research Signpost; 2004;2: p. 111–136.
[36] Kaas Q, Duprat E, Tourneur G, Lefranc M-P. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Immunoinformatics (Brusic V, Schoenbach C eds). Springer, The Netherlands; in press.
[37] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
[38] Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL nucleotide sequence database. Nucleic Acids Res 2005;33:D29–D33.
[39] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2005;33:D34–D8.
[40] Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T. DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res 2005;33:D25–D8.
[41] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:464–70.
[42] Larhammar D, Gustafsson K, Claesson L, Bill P, Wiman K, Schenning L, et al. Alpha chain of HLA-DR transplantation antigens is a member of the same protein superfamily as the immunoglobulins. Cell 1982;30:153–61.

[43] Maenaka K, Jones EY. MHC superfamily structure and the immune system. Curr Opin Struct Biol 1999;9:745–53.
[44] Wilson IA, Bjorkman PJ. Unusual MHC-like molecules: CD1, Fc receptor, the hemochromatosis gene product, and viral homologs. Curr Opin Immunol 1998;10:67–73.
[45] Braud VM, Allan DS, McMichael AJ. Functions of non-classical MHC and non-MHC- encoded class I molecules. Curr Opin Immunol 1999;11:100–8.
[46] Delker SL, West Jr AP, McDermott L, Kennedy MW, Bjorkman PJ. Crystallographic studies of ligand binding by Zn-alpha2-glycoprotein. J Struct Biol 2004;148:205–13.
[47] Pommié C, Levadoux S, Sabatier R, Lefranc G, Lefranc M-P. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. J Mol Recognit 2004;17:17–32.
[48] Giudicelli V, Protat C, Lefranc M-P. The IMGT strategy for the automatic annotation of IG and TR cDNA sequences: IMGT/Automat. In: Proceeding of the European Conference on Computational Biology (ECCB'2003), Paris, France, 2003 INRIA (DISC, Spid) DKB-31, p. 103–4. http://www.inra.fr/eccb2003/posters/pdf/Annot_Giudicelli_20030528_160703.pdf.