Available online at www.sciencedirect.com
SCIENCE DIRECT®
Developmental and Comparative Immunology 29 (2005) 185–203
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains

Marie-Paule Lefranc*, Christelle Pommié, Quentin Kaas, Elodie Duprat, Nathalie Bosc, Delphine Guiraudou, Christelle Jean, Manuel Ruiz, Isabelle Da Piédade, Mathieu Rouard, Elodie Foulquier, Valérie Thouvenin, Gérard Lefranc

IMGT, the International ImMunoGeneTics Information System®, LIGM, Laboratoire d'ImmunoGénétique Moléculaire, Université Montpellier II, UPR CNRS 1142, IGH, 141 rue de la Cardonille, 34396 Montpellier cedex 5, France

Received 19 May 2004; accepted 16 July 2004
Available online 1 September 2004

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 for the V-LIKE-DOMAINs. In this paper, this standardized characterization is extended to the constant domains (C-DOMAINs), and to the C-LIKE-DOMAINs, leading, for the first time, to their standardized description of mutations, allelic polymorphisms, two-dimensional (2D) representations and tridimensional (3D) structures. The IMGT unique numbering is, therefore, highly valuable for the comparative, structural or evolutionary studies of the immunoglobulin superfamily (IgSF) domains, V-DOMAINs and C-DOMAINs of IG and TR in vertebrates, and V-LIKE-DOMAINs and C-LIKE-DOMAINs of proteins other than IG and TR, in any species.

Keywords: IMGT; Immunoglobulin; T cell receptor; Constant domain; Immunoglobulin superfamily; V-set; C-set; Colliers de Perles

1. Introduction

IMGT, the international ImMunoGeneTics information system® (http://imgt.cines.fr) [1] is a high quality integrated knowledge resource specialized in immunoglobulins (IG), T cell receptors (TR), major

Abbreviations: 2D, two-dimensional; 3D, tridimensional; IG, immunoglobulin; IgSF, immunoglobulin superfamily; MHC, major histocompatibility complex; RPI, related proteins of the immune system; TR, T cell receptor.

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

0145-305X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2004.07.003

histocompatibility complex (MHC), and related proteins of the immune system (RPI) of human and other vertebrates [1–14]. IMGT provides a common access to expertly annotated data on the genome, proteome, genetics and structure of the IG, TR, MHC, and RPI, based on the IMGT Scientific chart rules and on the IMGT-ONTOLOGY concepts [15]. More particularly, the IMGT unique numbering [16–18], 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 tridimensional (3D) structure representations of the IG and TR variable domains (V-DOMAINs), whatever the antigen receptor, the chain type or the species [18]. The IMGT unique numbering for V-DOMAINs is used in all the IMGT components [3–8]: databases (IMGT/LIGM-DB [19], IMGT/PRIMER-DB [20], IMGT/GENE-DB [21], IMGT/3Dstructure-DB [22]), tools for sequence and structure analysis (IMGT/V-QUEST [23], IMGT/JunctionAnalysis [24], IMGT/Allele-Align, IMGT/PhyloGene [25], IMGT/StructuralQuery [22]), and Web resources ('IMGT Protein displays' [26,27], 'IMGT Colliers de Perles' 2D representations [28], and 'IMGT Alignments of Alleles' [29,30]; see IMGT Repertoire, http://imgt.cines.fr). Interestingly, the IMGT unique numbering for V-DOMAIN was fully applicable to the V-LIKE-DOMAINs of proteins other than IG and TR [18], although their genomic structure (V-LIKE-DOMAINs are often encoded by one exon) is different from that of the IG and TR V-DOMAINs (encoded by the rearranged V–(D–)J genes).

In this paper, the standardized IMGT unique numbering is extended to the IG and TR constant domains (C-DOMAINs) of the IG and TR of all jawed vertebrates, and to the C-LIKE-DOMAINs of proteins other than IG and TR of any species. The IMGT unique numbering represents therefore a major step forward for the comparative analysis and for the 2D and 3D structure and evolution studies of the immunoglobulin superfamily (IgSF) domains, V-DOMAINs and C-DOMAINs of IG and TR in vertebrates, and V-LIKE-DOMAINs and C-LIKE-DOMAINs of proteins other than IG and TR, in any species.

2. C-DOMAIN definition and relations with C-REGION

The C-DOMAIN of an IG or TR chain is a 3D structural unit comprising about 100 amino acids in seven antiparallel beta strands, on two sheets [31–33]. The seven strands of the C-DOMAIN, designated as A, B, C, D, E, F and G, have a topology and 3D structure similar to that of a V-DOMAIN without its C′ and C″ strands. Indeed, the C-DOMAIN beta sandwich fold is built up from the seven beta strands arranged so that four strands form one beta sheet, and three strands form a second sheet [31–33]. Depending from the topology of the D strand, which can be in one or the other sheet, the beta sandwich comprises ABE and GFCD, or ABED and GFC. ABE and GFC are closely packed against each other and joined together by a disulfide bridge from strand B in the first sheet with strand F in the second sheet [31], conserved in both the C-DOMAINs and V-DOMAINs. Amino acids with conserved physico-chemical characteristics form and stabilize the framework by packing the beta sheets through hydrophobic interactions giving a hydrophobic core [31,34].

Whereas the C-DOMAIN is a structural unit of an IG or TR chain, the C-REGION represents the part of an IG or TR chain encoded by the C-GENE [29,30]. Depending on the IG or TR chain type, the C-DOMAIN may correspond to a complete C-REGION, or to most of the C-REGION, or to only part of the C-REGION (if the C-REGION comprises several C-DOMAINs). As respective examples, (i) the domains C-KAPPA, C-LAMBDA and C-IOTA correspond to the complete C-REGION of an IG kappa, lambda, and iota chains, respectively (if the IG light chain type is not specified, the domain is designated as CL); (ii) the domains C-ALPHA, C-BETA, C-GAMMA and C-DELTA correspond to most (but not to the entirety) of the C-REGION of the TR alpha, beta, gamma and delta chains, respectively (indeed, the TR chains are transmembrane proteins whose C-REGION encoded by the C-GENE comprises, in addition to the C-DOMAIN, the CONNECTING-REGION, the TRANSMEMBRANE-REGION, and the INTRACYTOPLASMIC-REGION) [29,30]; (iii) the domains CH1, CH2, CH3 and,

if present, CH4, correspond to only part of the C-REGION of the IG heavy chains (e.g. the domains CH1, CH2 and CH3 of the human IGHG1 represent 98, 110 and 105 amino acids, respectively, on a total of 399 or 330 amino acids for the complete C-REGION of a membrane gamma 1 chain or that of a secreted gamma 1 chain, respectively [29]) (Fig. 1). It is worth to note that these relations between C-DOMAIN and C-REGION are quite different from those between V-DOMAIN and V-REGION, since the V-DOMAIN of an IG or TR chain results from the junction of two or three different regions: V and J (V–J-REGION of the IG light chains, and of the TR alpha and gamma chains), or V, D and J (V–D–J-REGION of the IG heavy chains, and of the TR beta and delta chains) [29,30] (Fig. 1).

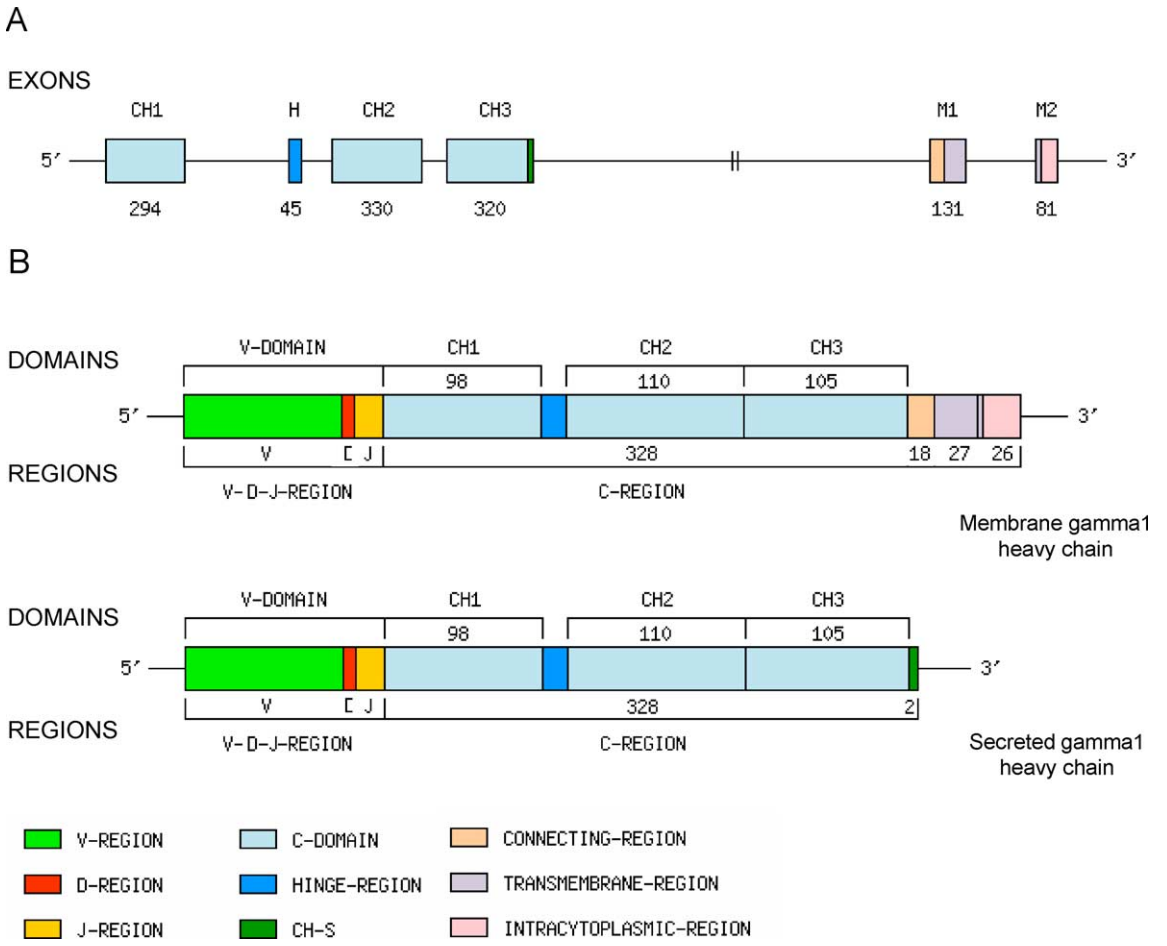

Fig. 1. Correspondence between exons, domains and regions. (A) Exons of the *Homo sapiens* IGHG1 gene shown as an example. Length of the exons are in base pairs. Introns are not at scale (see Ref. [29] for a representation at scale). (B) Domains and regions of a *Homo sapiens* membrane and secreted IG gamma 1 heavy chain shown as examples. Lengths of the domains and regions are in number of amino acids. The CH3 exon (320 nucleotides) encodes 107 amino acids (105 amino acids of the CH3 domain and 2 amino acids of CH-S, only present in the secreted gamma 1 chain). Exon M1 (131 nucleotides) encodes 44 amino acids (the 18 amino acids of the CONNECTING-REGION and 26 of the 27 amino acids of the TRANSMEMBRANE-REGION), exon M2 (81 nucleotides) encodes 27 amino acids (the last amino acid of the TRANSMEMBRANE-REGION and the 26 amino acids of the INTRACYTOPLASMIC-REGION). (Human IGHC 'Alignments of Alleles' and Protein displays in IMGT Repertoire, http://imgt.cines.fr and [29]).

3. IMGT unique numbering for C-DOMAIN

Owing to the high conservation of the structure of the immunoglobulin fold, the IMGT unique numbering for the C-DOMAINs of the IG and TR chains is derived from the IMGT unique numbering for V-DOMAIN [16–18], based on the NUMEROTATION concept of IMGT-ONTOLOGY. In the IMGT unique numbering, the conserved amino acids always have the same position, for instance Cysteine 23 (1st-CYS), Tryptophan 41 (CONSERVED-TRP), hydrophobic amino acid 89, Cysteine 104 (2nd-CYS). The hydrophobic amino acids of the antiparallel beta strands (framework regions) are also found in conserved positions [29,30].

In order to set up the IMGT unique numbering for C-DOMAIN, we first identified the amino acid positions, which correspond to equivalent positions in the V-DOMAIN. This correspondence was established by sequence alignment comparison of annotated IG and TR from the IMGT/LIGM-DB database [5,8,19] and by structural data analysis of IG and TR with known 3D structures from IMGT/3Dstructure-DB, http://imgt.cines.fr [22]. Seventy-two positions were identified as structurally equivalent between the C-DOMAIN and the V-DOMAIN, when the strands A to G were compared. They comprise: positions 1–15 (strand A), 16–26 (strand B), 39–45 (strand C), 77–84 (strand D), 85–96 (strand E), 97–104 (strand F), 118–128 (strand G) (Table 1). These positions are boxed in Fig. 2 and are indicated in the header upper line of the IMGT Protein display (Fig. 3) (Ruiz M., Martinez-Jean C. and Lefranc M.-P. IMGT Repertoire (for IG and TR) > Protein displays, on-line 27/02/2001, http://imgt.cines.fr).

We then identified the C-DOMAIN characteristic positions. These positions, shown in the header lower line of the IMGT Protein display (Fig. 3), comprise either additional or missing amino acid positions in the C-DOMAIN compared to the V-DOMAIN. Thirty-seven additional positions are characteristic of the C-DOMAIN numbering and are designated by a number followed by a dot and a number (Table 1): 1.1–1.9 at the N-terminal end of the C-DOMAIN, 15.1–15.3 at the AB turn, 45.1–45.9 which represent a characteristic transversal strand CD, 84.1–84.7, 85.7–85.1 at the DE loop (these positions correspond to longer antiparallel D and E strands in the C-DOMAIN), 96.1 and 96.2 at the EF turn. Interestingly, the additional positions 45.1–45.9 in the C-DOMAIN, compared to the V-DOMAIN, correspond to structural differences between the V- and C-DOMAINs. Indeed, these positions 45.1–45.9 represent a transversal strand between C and D in the C-DOMAIN whereas, in contrast, C′ and C″ in the V-DOMAIN are antiparallel strands. Thirty-three positions are missing in the C-DOMAIN compared to the V-DOMAIN. Thirty-one of these missing positions (46–76) correspond to the last amino acid of the C strand, to the two C′ and C″ strands, and to the C′C″ (or CDR2-IMGT) loop of the V-DOMAIN [18]. The last two missing positions (37 and 38) are in the BC loop. The C-DOMAIN BC loop has a maximum length of 10 amino acids (positions 27–36) compared to the maximum length of 12 amino acids for the equivalent V-DOMAIN CDR1-IMGT.

Table 1
Rules for gaps and additional positions in C-DOMAIN and C-LIKE-DOMAIN

A-Length of the A strand (positions 1.9-1.1, 1-15) and B strand (positions 16–26), and gaps at the AB turn								
Examples			A strand 1.9-1.1[a], 1–15			B strand 16–26		Number of gaps at the AB turn
Species	Gene	Domain	Number of additional positions at the N-terminus	Gap positions[b]	A strand length (16–24)[a]	Gap positions[b]	B strand length (8–11)	
Homo sapiens	IGHG1	CH1, CH3	4		19		11	0
Homo sapiens	IGHG1	CH2	6		21		11	0
Mus musculus	CD1D	C-LIKE [D3]	1	15	15	16	10	2
Homo sapiens	HLA-B	C-LIKE [D3]	1	14 15	14	16	10	3
Homo sapiens	TRAC	C-ALPHA	5	10 13 14 15	16	16	10	4
Homo sapiens	TRDC	C-DELTA	6	10 12 13 14 15	16	16 17 18	8	7

(*continued on next page*)

Table 1 (*continued*)

B-Length of the AB turn (positions 15.1–15.3)

Examples			AB turn 15.1–15.3	
Species	Gene	Domain	Additional positions[c]	AB turn length (0–3)[c]
Homo sapiens	IGHG1	CH1, CH3		0
Homo sapiens	TRBC2	C-BETA2	15.1	1
Homo sapiens	CD4	C-LIKE [D2]	15.1	1
Homo sapiens	IGHG1	CH2	15.1 15.2	2

C-Length of the BC loop (positions 27–36)

Examples			BC loop 27–36		
Species	Gene	Domain	Number of gaps	Gap positions[d]	BC loop length (6–10)[e]
Homo sapiens	IGHG1	CH2	0		10
Homo sapiens	IGHE	CH3	0		10
Mus musculus	IGHE	CH3	1	32	9
Homo sapiens	IGHG1	CH1, CH3	2	31 32	8
Homo sapiens	TRAC	C-ALPHA	3	31 32 33	7
Mus musculus	IGHD	CH1	4	30 31 32 33	6

D-Length of the C strand (positions 39–45) and CD transversal strand (positions 45.1–45.9)

Examples			C strand 39–45	CD transversal strand[g] 45.1–45.9	
Species	Gene	Domain	C strand length (7)[g]	Additional positions[f]	CD length (0–9)[f]
Homo sapiens	TRAC	C-ALPHA	7		0
Homo sapiens	CD4	C-LIKE [D4]	7		0
Homo sapiens	TRDC	C-DELTA	7	45.1	1
Homo sapiens	FCGR1A	C-LIKE [D1]	7	45.1,45.2	2
Homo sapiens	FCGR1A	C-LIKE [D2]	7	45.1–45.3	3
Homo sapiens	IGHG1	CH1	7	45.1–45.3	3
Homo sapiens	IGHG1	CH2, CH3	7	45.1–45.4	4
Homo sapiens	TRBC2	C-BETA2	7	45.1–45.5	5
Homo sapiens	TRGC1	C-GAMMA1	7	45.1–45.5	5
Homo sapiens	HLA-B	C-LIKE [D3]	7	45.1–45.5	5
Homo sapiens	IGHA1	CH3	7	45.1–45.6	6
Sus crofa	IGHE	CH3	7	45.1–45.7	7
Anarhichas minor	IGHC1S1	C-IOTA	7	45.1–45.9	9
Seriola quinqueradiata	IGIC1S22	C-IOTA	7	45.1–45.9	9
Siniperca chuatsi	IGIC1S3	C-IOTA	7	45.1–45.9	9
Siniperca chuatsi	IGIC1S5	C-IOTA	7	45.1–45.9	9

E-Length of the D strand (positions 77–84) and E strand (positions 85–96), and gaps at the DE

Examples			D strand 77–84[h]		E strand 85–96		Number of gaps at the DE turn
Species	Gene	Domain	Gap positions[i]	D strand length (5–8)[h]	Gap positions	E strand length (8–12)[j]	
Homo sapiens	IGHG1	CH2, CH3		8		12	0
Homo sapiens	FCGR1A	C-LIKE [D2]	83 84	6	85 86	10	4
Homo sapiens	FCGR1A	C-LIKE [Dl]	82 83 84	5	85 86	10	5

F-Length of the DE turn (positions 84.1–84.7, 85.7–85.1)

Examples			DE turn 84.1–84.7, 85.7–85.1	
Species	Gene	Domain	Additional positions[k]	DE turn length (6–14)[k]
Meleagris gallopavo	Telokin	C-LIKE [D]	84.1	1
Homo sapiens	CD4	C-LIKE [D4]	84.1	1
Homo sapiens	CD3E	C-LIKE [D]	84.1, 84.2, 85.1, 85.2	4

(*continued on next page*)

Table 1 (*continued*)

Examples			DE turn 84.1–84.7, 85.7–85.1	
Species	Gene	Domain	Additional positions[k]	DE turn length (6–14)[k]
Homo sapiens	ICAM1	C-LIKE [D1]	84.1–84.3, 85.1, 85.2	5
Mus musculus	IGHE	CH1	84.1–84.3, 85.1–85.3	6
Homo sapiens	TRDC	C-DELTA	84.1–84.4, 85.1–85.4	8
Homo sapiens	TRGC1	C-GAMMA1	84.1–84.4, 85.1–85.4	8
Homo sapiens	IGHG1	CH1, CH2, CH3	84.1–84.4, 85.1–85.4	8
Mus musculus	TRBC1	C-BETA1	84.1–84.5, 85.1–85.4	9
Canis familiaris	IGHA	CH3	84.1–84.5, 85.1–85.5	10
Homo sapiens	IGHA1	CH3	84.1–84.6, 85.1–85.5	11
Homo sapiens	IGHM	CH2	84.1–84.6, 85.1–85.6	12
Homo sapiens	TRBC2	C-BETA2	84.1–84.7, 85.1–85.6	13
Homo sapiens	TRAC	C-ALPHA	84.1–84.7, 85.1–85.7	14

G-Length of the E strand (positions 85–96) and F strand (positions 97–104), and gaps at the EF turn

Examples			E strand 85–96		F strand 97–104		Number of gaps at the EF turn
Species	Gene	Domain	Gap positions	E strand length (8–12)[l]	Gap positions	F strand length (4–8)[m]	
Homo sapiens	IGHG1	CH2, CH3		12		8	0
Homo sapiens	FCGR2A	C-LIKE [D1]	*85 86* 96	9		8	1
Homo sapiens	IGHG1	CH1	96	11		8	1
Bos taurus	IGHG1	CH1	96	11	97	7	2
Homo sapiens	TRGC1	C-GAMMA1	96	11	97	7	2
Homo sapiens	HLA-B	C-LIKE [D3]	95 96	10	97	7	3
Rattus norvegicus	IGHG2B	CH1	91 92[n]	10	97	7	3
Homo sapiens	TRDC	C-DELTA	95 96	10	97 98	6	4
Oryctolagus cuniculus	IGHG	CH1	95 96	10	97 98	6	4
Homo sapiens	TRAC	C-ALPHA	93 94 95 96	8	97 98 99 100	4	8

H-Length of the EF turn (positions 96.1–96.2)

Examples			EF turn 96.1–96.2	
Species	Gene	Domain	Additional positions[o]	EF turn length (0.2)[o]
Homo sapiens	IGHG1	CH1, CH2, CH3		0
Homo sapiens	TRBC2	C-BETA2	96.1	1
Homo sapiens	IGHM	CH1	96.1 96.2	2

I-Length of the FG loop (positions 105–117) (except for the C-BETA domains)

Examples			FG loop 105–117		
Species	Gene	Domain	Number of gaps	Gap positions	FG loop length (7–13)[p]
Homo sapiens	IGHE	CH1	0		13
Mus musculus	TRAC	C-ALPHA	1	111	12
Homo sapiens	IGHG1	CH3	1	111	12
Homo sapiens	IGHG1	CH1, CH2	2	111 112	11
Homo sapiens	FCGR2A	C-LIKE [D2]	3	110 111 112	10
Ovis aries	IGHA	CH2	4	110 111 112 113	9
Homo sapiens	FCGR2A	C-LIKE [D1]	6	109 110 111 112 113 114	7

J-Length of the FG loop of the C-BETA domains (positions 105–111.6, 112.6–117)

Examples			FG loop 105–111.6, 112.6–117		
Species	Gene	Domain	Number of additional positions	Additional positions[q]	FG loop length (25)[q]
Homo sapiens	TRBC2	C-BETA2	12	111.1–111.6, 112.1–112.6	25

(*continued on next page*)

K-Length of the G strand (positions 118–128)				
Examples			G strand positions 118–128	
Species	Gene	Domain	C-terminal positions	G strand length (4–11)[r]
Homo sapiens	IGHG1	CH1	121	4
Homo sapiens	HLA-A	C-LIKE [D3]	121	4
Mus musculus	IGHE	CH1	122	5
Homo sapiens	IGHG1	CH2, CH3	125	8
Mus musculus	IGHE	CH2, CH3, CH4	125	8
Mus musculus	TRBC1	C-BETA1	125	8
Homo sapiens	IGKC	C-KAPPA	126	9
Homo sapiens	FCGR2A	C-LIKE [D2]	126	9
Homo sapiens	IGLC1	C-LAMBDA1	127	10
Homo sapiens	FCGR1A	C-LIKE [D1]	127	10
Homo sapiens	FCGR2A	C-LIKE[D1]	127	10

Rules are described by comparison to the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [18]. Gaps and additional positions were confirmed for proteins with known 3D structures (PDB codes in IMGT Repertoire > Protein displays, http://imgt.cines.fr). Strand, turn and loop lengths are shown, in number of amino acids between parentheses, in the table headers.

[a] Up to nine additional positions (numbered 1.1–1.9, starting from position next to 1 towards the N-terminal end) may be found at the N-terminal end of the A strand. The maximal length of the A strand is 24 amino acids.

[b] Gap positions are based on 3D structures, and if 3D structures are not known, gaps are equally distributed on strands A and B with, for an odd number, one more gap on strand A.

[c] C-DOMAIN and C-LIKE-DOMAIN may have additional amino acids (potentially 3) at the AB turn, which define the AB turn length.

[d] Gap positions start with 32, then 31, 33, 30.

[e] The maximum length of the BC loop is 10 amino acids. If the number of amino acids is odd, there is one more amino acid position on the left. There are no positions 37 and 38 in C-DOMAINs and C-LIKE-DOMAINs.

[f] The maximum length of the C strand is seven amino acids. There is no position 46 in C-DOMAINs and C-LIKE-DOMAINs.

[g] The CD transversal strand is characteristic of the C-DOMAIN and C-LIKE-DOMAIN. The maximum length of the CD strand is 9 amino acids, found in Teleostei IGIC (available online in IMGT Repertoire http://imgt.cines.fr): IGIC1S1 (AF137397) gene of Spotted wolffish (*Anarhichas minor*), IGIC1S22 (AB062662) gene of Five-ray yellowtail (*Seriola quinqueradiata*), IGIC1S3 (AF454470) and IGIC1S5 (AY013294) genes of Chinese perch (*Siniperca chuatsi*). However, since this length is exceptional, usual IMGT Collier de Perles only display seven positions. Amino acid positions are added from left to right in sequence alignments.

[h] The maximal length of the D strand is eight amino acids. There are no positions 75 and 76 in C-DOMAINs and C-LIKE-DOMAINs.

[i] Gap positions are based on 3D structures, and if 3D structures are not known, gaps at the DE turn are equally distributed on strands D and E with, for an odd number, one more gap on strand D.

[j] The maximum length of the E strand is 12 amino acids. Note that the E strand length also depends from gaps found at the EF turn, which is reflected by E strand lengths of eight and nine amino acids, as described in Table 1G.

[k] Most of the C-DOMAINs and C-LIKE-DOMAINs have additional amino acids (potentially 14) at the DE turn which extend the D and E antiparallel beta strands. The number of additional positions defines the DE turn length. The numbering of the additional positions starts from positions next to 84 and 85, respectively, towards the top of the DE turn. If the number of additional amino acids is odd, there is one more position on the left.

[l] The maximal length of the E strand is 12 amino acids. Note that the E strand length also depends from gaps found at the DE turn (gap positions 85, 86 shown in italics, for the *Homo sapiens* FCGR2A) as described in Table 1E.

[m] The maximal length of the F strand is eight amino acids. Gap positions are based on 3D structures, but if 3D structures are not known, gaps are based on sequence alignments.

[n] Gaps were assigned by sequence alignment with the CH1 of the IGHG2A and IGHG2C.

[o] C-DOMAIN and C-LIKE-DOMAIN may have additional positions (potentially 2) at the EF turn, which define the EF length.

[p] Except for the C-BETA domains (TRBC sequences) described in Table 1J, the maximal length of the FG loop is 13 amino acids. For an odd number of gaps, there is one more gap on the left (starting with position 111).

[q] C-BETA domains (TRBC sequences) have an insertion of 12 positions between 111 and 112. The length of the FG loop in C-BETA domains is 25 amino acids. The numbering of the additional positions starts from positions next to 111 and 112, respectively, towards the top of the FG loop.

[r] If longer G strands are found, positions will be numbered consecutively.

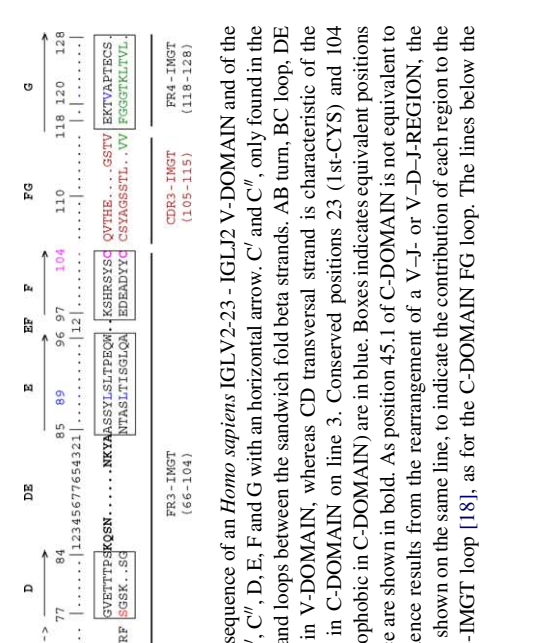

Fig. 2. Correspondence between the C-DOMAIN and V-DOMAIN IMGT unique numbering. Amino acid sequence of an *Homo sapiens* IGLV2-23 - IGLJ2 V-DOMAIN and of the *H. sapiens* IGLC1 C-DOMAIN are shown as examples. The upper line indicates the beta strands A, B, C, C′, C″, D, E, F and G with an horizontal arrow. C′ and C″, only found in the V-DOMAIN, are shown with dashed arrows. AB, BC, CD, C′C″, DE, EF and FG correspond to the turns and loops between the sandwich fold beta strands. AB turn, BC loop, DE and EF turns, and FG loop are found in both V-DOMAIN and C-DOMAIN. C′C″ loop is only found in V-DOMAIN, whereas CD transversal strand is characteristic of the C-DOMAIN. The IMGT unique numbering is shown on lines 2 and 3 with additional positions found in C-DOMAIN on line 3. Conserved positions 23 (1st-CYS) and 104 (2nd-CYS) are in magenta. Conserved positions 41 (CONSERVED-TRP), 89 (hydrophobic) and 121 (hydrophobic in C-DOMAIN) are in blue. Boxes indicates equivalent positions in both the V-DOMAIN and C-DOMAIN. Amino acids at additional positions in the C-DOMAIN sequence are shown in bold. As position 45.1 of C-DOMAIN is not equivalent to position 46 in V-DOMAIN, both positions 45.1 and 46 are shown in this figure. As a V-DOMAIN sequence results from the rearrangement of a V–J- or V–D–J-REGION, the sequence of the germline V-REGION and that of the germline J-REGION (in green) taken as examples are shown on the same line, to indicate the contribution of each region to the V-DOMAIN. Note that in a 'true' V–J rearrangement, the gaps would be placed at the top of the CDR3-IMGT loop [18], as for the C-DOMAIN FG loop. The lines below the sequences indicate the V-DOMAIN FR-IMGT and CDR-IMGT and their delimitations.

Rules for gaps and additional positions in C-DOMAIN and C-LIKE-DOMAIN are described in details in IMGT Scientific chart (http://imgt.cines.fr) and in Table 1. Correspondence between the C-DOMAIN and V-DOMAIN IMGT unique numbering is shown in Fig. 2 with an *Homo sapiens* IGLV2-23 - IGLJ2 V-DOMAIN and the *Homo sapiens* IGLC1 C-DOMAIN taken as examples.

It is worth to note that it is the analysis of structural data and sequence comparisons that we were carrying out to apply the IMGT unique numbering for the description of the IG and TR C-DOMAIN, which showed us that the standardized numbering of the FG loop of the C-DOMAIN could be applied to the CDR3-IMGT of rearranged IG and TR sequences (Fig. 3 in Ref. [18]). More precisely, the hydrogen bonds between second-CYS 104 and position 119 in the C-DOMAIN correspond structurally to the hydrogen bonds between second-CYS 104 and the Glycine which follows the J-TRP or J-PHE in the J-REGION. This Glycine was therefore numbered as 119, and as a consequence, J-TRP and J-PHE as 118.

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

The IMGT unique numbering for C-DOMAIN allows for the first time a standardized comparison of the nucleotide substitutions and amino acid changes between different constant domains of a same gene or chain, or between constant domains of different IG and TR genes or chains, from either the same species, or from different species (Fig. 3). The IMGT unique numbering is also crucial for the standardization of the allele description. Indeed, in IMGT, the polymorphisms are described by comparison to the sequences from the IMGT reference directory. All the human IG and TR gene names from this IMGT reference directory [29,30], including the C-GENEs, were approved by the Human Genome Organisation (HUGO) Nomenclature Committee (HGNC) in 1999 [39], and entered in IMGT/GENE-DB [5,21], GDB [40], LocusLink at NCBI (USA) [41] and GeneCards [42]. This standardized nomenclature, based on the CLASSIFICATION concept of IMGT-ONTOLOGY [15], represented

a major step in the setting up of the 'Tables of alleles' and 'Alignments of alleles' of the IG and TR genes (http://imgt.cines.fr). V-GENE polymorphisms in IMGT [29,30] have been described from the start according to the IMGT unique numbering for V-REGION set up in 1997 [16–18]. In contrast, C-GENE polymorphisms were initially described according to the exon numbering [29,30] and sequence comparison between exons of different lengths was not easy. The implementation of the IMGT unique numbering for the C-DOMAIN represents therefore a new major step in the setting up of standardized 'Alignments of alleles' whatever the receptor, the chain, or the species (IMGT Repertoire, http://imgt.cines.fr) (Fig. 3). Owing to that standardization, the sequence polymorphisms of any C-DOMAIN of any IG or TR can very easily be compared and analysed.

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

Beyond sequence data comparison, the IMGT unique numbering for C-DOMAIN provides information on the strand and loop lengths (Table 2) and allows standardized IMGT Protein displays (Fig. 3) and IMGT Colliers de Perles (Fig. 4) for the IG and TR C-DOMAINs of any chain type from any species. Practically, structural data comparison of strand or loop of the same length can be done directly using the IMGT unique numbering. For example, all codons (or amino acids) at position 28 can be compared between domains with a BC loop of a given length. This standardization allows the structural characterization of a position inside a domain, and the statistical analysis of amino acid properties, position per position, between domains, as this has been demonstrated for the V-DOMAIN [34]. Fig. 4 shows the IMGT Colliers de Perles for the *Homo sapiens* IGHG1 CH1, C-KAPPA, C-LAMBDA1 and *Mus musculus* C-BETA1 domains, as examples.

As soon as the first IMGT Collier de Perles was set up on the Web site in December 1997, the enormous potential of the IMGT unique numbering as a means to control data coherence was obvious. For new sequences, for which no 3D structures are available, the IMGT Colliers de Perles allow to precisely delimit the strands and loops and give information on the topological organization of the domain. The IMGT unique numbering is also used in more sophisticated representations of the IMGT Colliers de Perles on two layers (Fig. 4) which allow, when 3D structures are available, the visualisation of the hydrogen bonds between amino acids belonging to beta strands from the same sheet or from different sheets (IMGT/3Dstructure-DB, http://imgt.cines.fr) [22].

6. IMGT unique numbering for C-LIKE-DOMAIN

A C-LIKE-DOMAIN is a domain of similar structure to a C-DOMAIN, found in chains other than IG and TR [43–52]. The IMGT unique numbering for the C-LIKE-DOMAIN follows exactly the same rules as those of the C-DOMAIN (Table 1). Strand and loop lengths of 40 examples of C-LIKE-DOMAINs are given in Table 2. The IMGT Protein display of the corresponding C-LIKE-DOMAIN sequences are shown in Fig. 3. The IMGT Colliers de Perles of four representative C-LIKE-DOMAINs (*Homo sapiens* HLA-B [D3], B2M [D], FCGR2A [D1], and *Meleagris gallopavo* telokin [D]) are represented in Fig. 5. Detailed IMGT Alignment of alleles of *Homo sapiens* FCGR3B and IMGT Colliers de Perles of [D1] and [D2] of the FCGR3B*02 allele [53] further highlight the importance of the IMGT unique numbering standardization for the polymorphism and structure analysis and comparison of the C-LIKE-DOMAINs.

The IMGT unique numbering provides, for the first time, a standardized approach to analyse the sequences and structures of any domain belonging to the C-set of the IgSF [32] (the C-set comprises the IMGT C-DOMAINs and C-LIKE-DOMAINs). Three features are worth noting: (i) In IMGT, any C-DOMAIN or C-LIKE-DOMAIN is characterized by its strand and loop lengths (Table 2). Examples are shown in Figs. 3–5. This first feature of the IMGT standardization based on the IMGT unique numbering shows 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 [32,33,54–56] is unapplicable when dealing with sequences for which no structural data are known. Indeed, the four domain

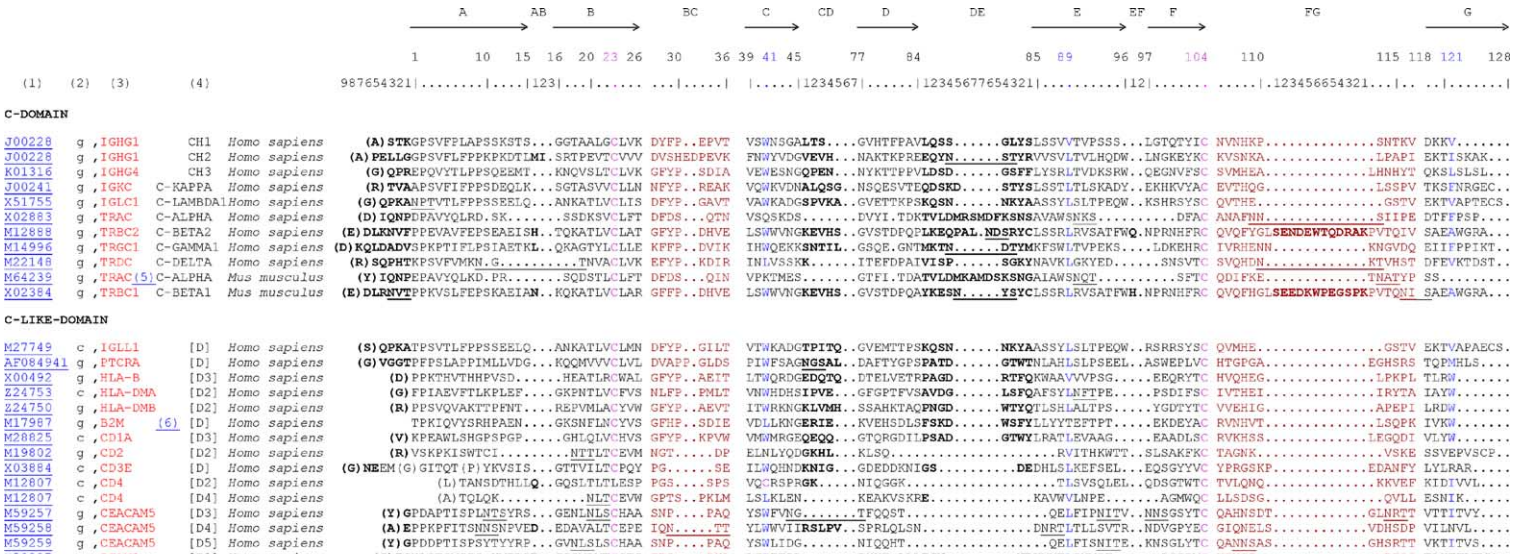

Fig. 3. IMGT Protein display of examples of C-DOMAINs (IG and TR) and C-LIKE-DOMAINs (proteins other than IG and TR). The protein display is according to the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN, based on the NUMEROTATION concept of IMGT-ONTOLOGY [15]. Sandwich fold beta strands are shown 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 between parentheses (for *Homo sapiens* FCGR3B [D2], the information is from M90745, for *H. sapiens* HLA-DMA [D2] from NT_007592, for *H. sapiens* CD3E [D] from NT_033899, for *H. sapiens* CD4 [D2] and [D4] from NT_009759, for *H. sapiens* CEACAM5 [D3], [D4] and [D5] from NT_011109, and for *Mus musculus* H2-Aa [D2] and H2-Ab [D2] from NT_039649). Putative N-glycosylation sites (N-X-S/T) are underlined. (1) Accession numbers are from IMGT/LIGM-DB (http://imgt.cines.fr) [5,19] for IG and TR and from EMBL/GenBank/DDBJ [35–37] for proteins other than IG and TR; Telokin identifier is from PDB [38] and IMGT/3Dstructure-DB [22]. (2) Molecule type. c: cDNA; g: genomic DNA; p: protein. (3) Gene names (symbols) for IG and TR are according to the IMGT Nomenclature committee (IMGT-NC) [29,30] and the HUGO Nomenclature Committee (HGNC) [39]. Full gene designations are the following: IGHG1: Immunoglobulin heavy constant gamma 1; IGHG4: Immunoglobulin heavy constant gamma 4; TRAC: T cell

receptor alpha constant; TRBC2: T cell receptor beta constant 2; TRGC1: T cell receptor gamma constant 1; TRDC: T cell receptor delta constant; IGLL1: Immunoglobulin lambda-like polypeptide 1; PTCRA: pre T-cell antigen receptor alpha; HLA-B: Major histocompatibility complex, class I, B; HLA-DMA: MHC class II, DM alpha; HLA-DMB: MHC class II, DM beta; B2M: Beta-2-microglobulin; CD1A: CD1A antigen, a polypeptide; CD2: CD2 antigen (p50), sheep red blood cell receptor; CD3E: CD3E antigen, epsilon polypeptide (TiT3 complex); CD4: CD4 antigen (p55); CEACAM5: Carcinoembryonic antigen-related cell adhesion molecule 5; ICAM1: intercellular adhesion molecule 1 (CD54), human rhinovirus receptor; ICAM2: intercellular adhesion molecule 2; VCAM1: vascular cell adhesion molecule 1; FCGR1A: Fc fragment of IgG, high affinity Ia, receptor for (CD64); FCGR2A: Fc fragment of IgG, low affinity IIa, receptor for (CD32); FCGR2B: Fc fragment of IgG, low affinity IIb, receptor for (CD32); FCGR3B: Fc fragment of IgG, low affinity IIIb, receptor for (CD16); FCER1A: Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide; H2-Aa: histocompatibility 2, class II antigen A, alpha; H2-Ab: histocompatibility 2, class II antigen A, beta; H2-K: histocompatibility 2, K region; CD1D: CD1D antigen, d polypeptide. (4) Domain name. The C-DOMAINs are designated with the IMGT labels (IMGT Scientific chart, http://imgt.cines.fr) The C-LIKE-DOMAINs are designated by the letter D between brackets with a number, corresponding to the position of the domain from the N-terminal end of the protein, and relative to the other domains. Membrane proteins quoted in this figure are of type I, that is with the N-terminal end being extracellular. There is no number if there is a unique C-LIKE domain in the chain. (5) Q at position 6 is according to M64239 (and replace A in PDB and IMGT/3Dstructure-DB entries 1tcr_A). (6) B2M [D] is encoded by EX2 and is preceded at the N-terminal end by SGLEGIQR or TGLYAIQK encoded by the EX1 end in *Homo sapiens* or *Mus musculus*, respectively (EX1 encodes 23 amino acids, including the L-REGION). The splicing site is between the last EX1 codon (encoding R in *Homo sapiens*, K in *Mus musculus*) and the first EX2 codon (encoding T in both species). The proteolytic cleavage site of the L-REGION is not known. (7) [D1] is encoded by EX3 and is preceded at the N-terminal end by seven amino acids (A)SADSQA encoded by EX2. The proteolytic cleavage site of the L-REGION is not known. (8) The N-terminal and C-terminal ends of Telokin [D] need to be confirmed by genomic sequence. 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.

types C1, C2 (also known as H [57]), I1 and I2, are based on structural differences [55,58]. C1 and C2 differ by the location of the fourth strand D, which is in the ABE sheet for C1, and in the GFC sheet for C2 (the D strand is then designated as C′ by some authors) (Fig. 6). Thus C1 contains ABED and GFC sheets, whereas C2 is characterized by ABE and GFCD (or GFCC′) sheets (Table 3). The C1 type initially defined by the topology observed in the IG constant domain was described as being shared by the TR constant domain, and the C-like domain of MHC class I, MHC class II and B2M, and thought to be a characteristic of the immunoreceptor of the adaptative immune response. In the literature, the assignment of C-set domains to C1 or C2 is often done by extrapolation, without available 3D structures. Moreover, the 3D structure of the 2C T cell receptor C-ALPHA domain (PDB and IMGT/3Dstructure-DB: 1tcr_A) displays a quite different conformation than that of the C1 type with a beta sheet containing the GABED strands [59]. Therefore, the distinction between C1 and C2 is not so straightforward and as more 3D structures become available, more intermediate or variant structures will be described.

The structural analysis of telokin, the C-terminal domain of myosin light chain kinase [60] revealed a new domain type designated as 'I' [55], which was defined as an intermediate between the V and C1 types because, as frequently found in the V domain type, the A strand, in its second part, is located on the GFC sheet (and therefore designated as A′ strand), and as C1, it lacks the V-DOMAIN C′ and C″ strands. This I type was later divided in I1 and I2 [58] which differ by the location of the fourth strand, which is in the ABE sheet for I1, and in the A′GFC sheet for I2. Thus I1 contains ABED and A′GFC sheets, whereas I2 is characterized by ABE and A′GFCD (or A′GFCC″) sheets (I1 and I2 are described as I and E types in [61]). In the absence of 3D structures, a classification of the constant domains based on such structural differences cannot be taken into account, and the IMGT description based on strand and loop lengths and IMGT Colliers de Perles is more pertinent. (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 limits of the C-LIKE-DOMAIN,

Table 2
Strand and loop lengths of examples of C-DOMAINs and C-LIKE-DOMAINS

Species	Gene name	PDB code	Domain	A 1.9–1.1 1–15 (6–23)		AB 15.1–15.3 (0–2)	B 16–26 (7–11)	BC 27–36 (4–10)	C 39–45 (7)	CD 45.1–45.7 (0–7)	D 77–84 (4–8)	DE 84.1–84.7 85.7–86.1 (0–14)	E 85–96 (8–12)	EF 96.1–96.2 (0–2)	F 97–104 (4–8)	FG 105–117 111.1–111.6, 112–1–112.6 (7–25)	G 118–128 (4–10)	Total
C-DOMAINS																		
Homo	IGHG1	1hzh_H	CH1	+4	19		11	8	7	3	8	8	11		8	11	4	98
sapiens	IGHG1	1hzh_H	CH2	+6	21	2	11	10	7	4	8	8	12		8	11	8	110
	IGHG4	1adq_A	CH3	+4	19		11	8	7	4	8	8	12		8	12	8	105
	IGKC	1dfb_L	C–KAPPA	+4	19		11	8	7	5	8	9	12		8	11	9	107
	IGLC1	1a8j_L	C–LAMBDA1	+5	20		11	8	7	5	8	8	12		8	9	10	106
	TRAC	1qrn_D	C–ALPHA	+5,− 4	16		10	7	7		7	14	8		4	11	7	91
	TRBC2	1qm_E	C–BETA2	+7	22	1	11	8	7	5	8	13	12	1	8	25	8	129
	TRGC1	1hxm_B	C–GAMMA1	+8	23	1	11	8	7	5	7	8	11		7	13	9	110
	TRDC	1hxm_A	C–DELTA	+6,− 5	16		8	8	7	1	8	8	10		6	12	9	93
Mus mus-	TRAC	1tcr_A	C–ALPHA	+5,− 4	16		10	7	7		7	14	8		4	12	2	87
culus	TRBC1	1tcr_B	C–BETA1	+7	22	1	11	8	7	5	8	9	12	1	8	25	8	125
C-LIKE-DOMAINs																		
Homo	IGLL1		[D]	+5	20		11	8	7	5	8	8	12		8	9	10	106
sapiens	PTCRA		[D]	+5	20		11	9	7	5	8	8	12		8	12	7	107
	HLA–B	1a1m–A	[D3]	+1,− 2	14		10	8	7	5	8	8	10		7	11	4	92
	HLA–DMA	1hdm_A	[D2]	+1,− 1	15		11	8	7	4	8	8	10		7	11	4	93
	HLA–DMB	1hdm_B	[D2]	+1,− 1	15		11	8	7	5	8	8	10		7	11	4	94
	B2M	1lds_A	[D]	− 1	14		11	8	7	4	8	8	10		7	11	4	92
	CD1A	1onq_A	[D3]	+1,− 1	15		10	8	7	4	8	8	10		7	12	4	93
	CD2	1hnf	[D2]	+1,− 4	12		9	5	7	4	4		9		8	9	9	76
	CD3E		[D]	+3	18		11	4	7	4	8	4	11		8	13	6	94
	CD4	1wio_A	[D2]	− 5	10	1	11	6	7	2	6		9		8	11	7	78
	CD4	1wio_A	[D4]	− 5,− 4	6		7	8	7		8	1	9		6	10	5	67
Homo	CEACAM5		[D3]	+2	17		11	6	7		6		10		8	13	7	85
sapiens	CEACAM5		[D4]	+2	17	1	11	5	7	5	8		12		8	13	6	93
	CEACAM5		[D5]	+2	17		11	6	7		6		10		8	13	7	85
	ICAM1	1d3l_A	[D1]	+3,− 1	17	1	11	6	7		7	5	10		6	9	9	88
	ICAM2	1zxq	[D1]	+5,− 1	19	1	11	6	7		7	4	10		6	9	9	89
	VCAMI	1vsc_A	[D1]	+2	17	1	11	6	7	4	6	4	11		7	9	9	92
	ICAM1	1d3l_A	[D2]	+2	17		11	7	7	4	8		12	2	8	16	10	102
	ICAM2	1zxq	[D2]	+2	17		11	7	7	4	8	7	12		8	16	10	107
	VCAM1	1vsc_A	[D2]	+1	16		11	7	7	4	8	9	12		8	16	9	107
	FCGR1A		[D1]	+2	17	2	11	7	7	2	5		10		8	7	10	86
	FCGR2A	1fcg_A	[D1]	+2	17	2	11	7	7	2	6		9		8	7	10	86
	FCGR2B	2fcb_A	[D1]	+2	17	2	11	7	7	2	6		9		8	7	10	86
	FCGR3B	1e4k_C	[D1]	+2	17	2	11	7	7	2	5		10		8	7	10	86
	FCER1A	1f6a_A	[D1]	+3	18		11	7	7	2	5		10		8	7	10	85
	FCGR1A		[D2]	− 1	14		11	7	7	3	6		10		8	9	10	85
	FCGR2A	1fcg_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	9	85
	FCGR2B	2fcb_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	9	85
	FCGR3B	le4k_C	[D2]	− 1	14		11	7	7	3	6		10		8	10	10	86
	FCER1A	1f6a_A	[D2]	− 1	14		11	7	7	3	6		10		8	10	10	86
	KIR2DL2	1efx_D	[D1]	+2	17	1	11	5	7	5	8	3	12		8	14	9	100
	KIR2DL2	1efx_D	[D2]	+2	17	1	11	5	7	5	8	3	11		7	14	9	98

Musmuscu-	IGLL1		[D]	+5	20	11	8	7	5	8	8	12	8	9	10	106
lus	PTCRA		[D]	+5	20	11	9	7	5	8	8	12	8	12	7	107
	H2–Aa	1d9k_C	[D2]	+1, −1	15	11	8	7	5	8	8	10	7	11	4	94
	H2–Ab	1d9k_D	[D2]	+1, −1	15	11	8	7	5	8	8	10	7	11	4	94
	H2–K	2vaa_A	[D3]	+1, −2	14	10	8	7	5	8	8	10	7	11	4	92
	B2M	2vaa_B	[D]	−1	14	11	8	7	4	8	8	10	7	11	4	92
	CD1D	1cd1_A	[D3]	+1, −1	15	10	8	7	4	8	8	10	7	12	4	93
Meleagris	Telokin	1fhg_A	[D]		15[a]	11	6	7	7	8	1	12	8	9	9[a]	93
gallopavo																

The range of lengths observed in the selected examples of C-DOMAINs and C-LIKE-DOMAINs are shown between parentheses in the header. The total number of amino acids of each domain is indicated in the column 'Total'.
[a] The N-terminal and C-terminal ends of telokin [D] need to be confirmed by genomic sequence.

a C-LIKE-DOMAIN being frequently encoded by a unique exon, as this is the case for the C-DOMAIN (Fig. 1). This IMGT standardization for the domain delimitations 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 C-LIKE-DOMAIN IMGT Collier de Perles, which, in the absence of available 3D structures, is particularly useful to compare domains of very diverse families, and to characterize them by their strand and loop lengths.

7. Conclusion

The IMGT unique numbering allows, for the first time, to compare any C-DOMAIN of IG and TR and C-LIKE-DOMAIN of proteins other than IG or TR, between them, and to any V-DOMAIN and V-LIKE-DOMAIN [18], that is to compare any domain belonging to the IgSF C-set or V-set. Sequences and 3D structures can be analysed whatever the domain (C-DOMAIN, C-LIKE-DOMAIN, V-DOMAIN, V-LIKE-DOMAIN), the receptor (IG, TR, or more generally IgSF), the chain type (heavy or light for IG; alpha, beta, gamma or delta for TR; or more generally IgSF chain), or the species. The IMGT unique numbering has many advantages. The strand and loop lengths (the number of codons or amino acids, that is the number of occupied positions) become crucial information, which characterizes the domains (V-DOMAINs, V-LIKE-DOMAINs, C-DOMAINs and C-LIKE-DOMAINs). The IMGT unique numbering allows standardized representations of nucleotide and amino acid sequences in IMGT Repertoire (http://imgt.cines.fr): Tables of strand and loop lengths, Tables of alleles, Alignments of alleles, Protein displays, IMGT Colliers de Perles, 3D structures. The IMGT unique numbering is applied through all the components (databases, tools and Web resources) of the IMGT information system® (http://imgt.cines.fr) [5], for the standardized label annotations [62] and database queries [19–22]. The IMGT unique numbering represents, therefore, a major step forward in the analysis and comparison of the sequence evolution and structure of the IgSF domains.

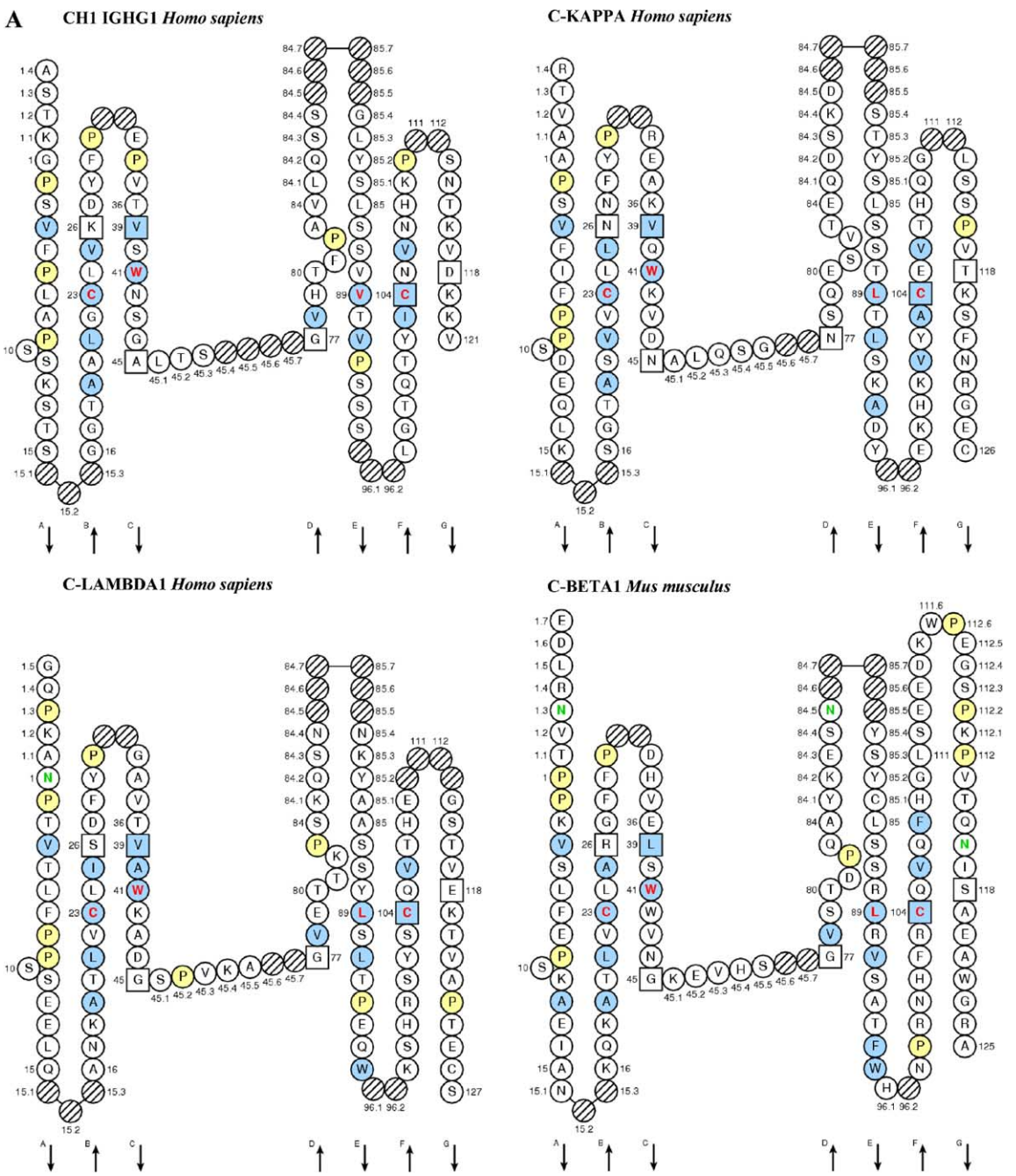

Fig. 4. IMGT Collier de Perles of C-DOMAINs. (A) on one layer (B) on two layers. CH1 of *Homo sapiens* IGHG1 (IMGT/LIGM-DB: J00228); C-KAPPA of *H. sapiens* IGKC (IMGT/LIGM-DB: J00241); C-LAMBDA1 of *H. sapiens* IGLC1 (IMGT/LIGM-DB: X51755); C-BETA1 of *Mus musculus* TRBC1 (IMGT/LIGM-DB: X02384). The first amino acids (A1.4, R1.4, G1.5 and E1.7) are encoded by a codon which results from the splicing with an IGHJ, IGKJ, IGLJ and TRBJ, respectively [29,30]. Amino acids are shown in the one-letter abbreviation. Positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and Tryptophan (W) are found in more than 50% of analysed IG and TR sequences are shown in blue. All Proline (P) are shown in yellow. The positions 26, 39, 104 and 118 are shown in squares by

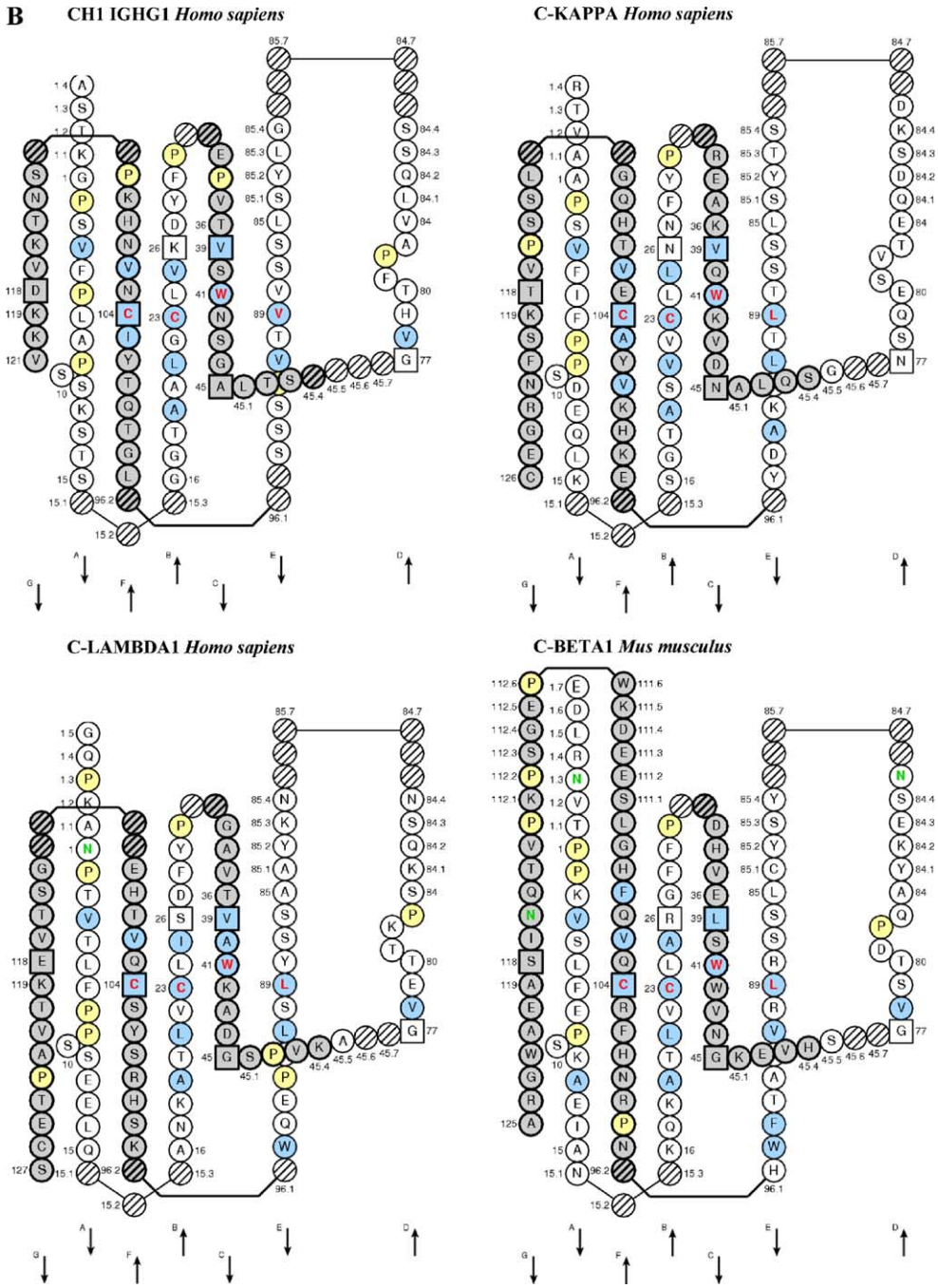

Fig. 4 (*continued*)

homology with the corresponding positions in the V-DOMAINs. Positions 45 and 77 which delimit the characteristic transversal 'CD' strand of the C-DOMAINs are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C-DOMAINs. Arrows indicate the direction of the beta strands (IMGT Repertoire, http://imgt.cines.fr).

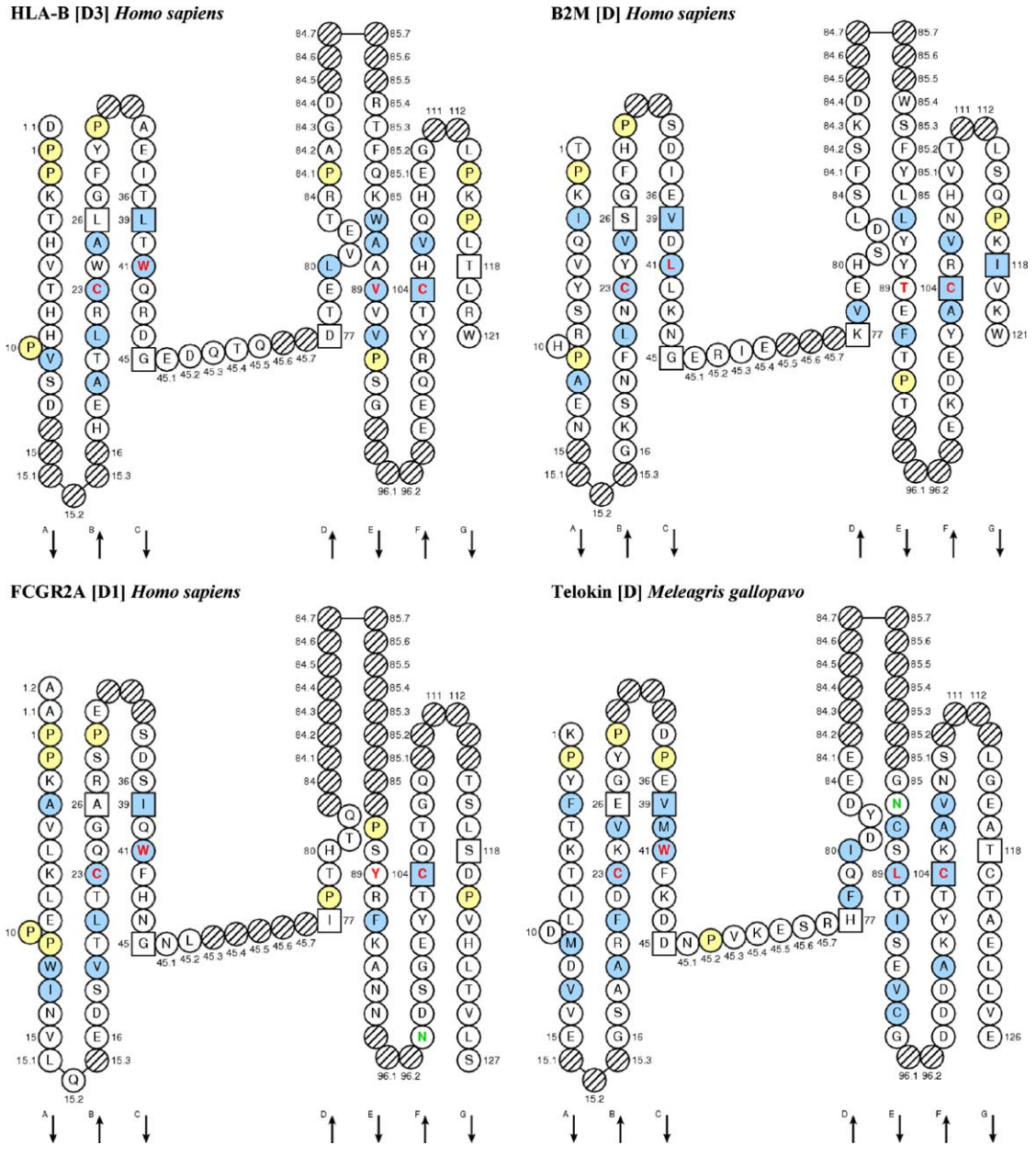

Fig. 5. IMGT Colliers de Perles of C-LIKE-DOMAINs. [D3] of *Homo sapiens* HLA-B (EMBL/GenBank/DDBJ: AB088084; PDB and IMGT/3Dstructure-DB: 1alm_A); [D] of *H. sapiens* B2M (EMBL/GenBank/DDBJ: M17987; PDB and IMGT/3Dstructure-DB: 1alm_B); [D1] of *H. sapiens* FCGR2A (EMBL/GenBank/DDBJ: M90723; PDB and IMGT/3Dstructure-DB: 1fcg_A); [D] of *Meleagris gallopavo* (turkey) telokin (PDB and IMGT/3Dstructure-DB: 1fhg_A). Amino acids are shown in the one-letter abbreviation. Positions at which hydrophobic amino acids (hydropathy index with positive value: I, V, L, F, C, M, A) and Tryptophan (W) are found in more than 50% of analysed IG and TR sequences are shown in blue. All Proline (P) are shown in yellow. The positions 26, 39, 104 and 118 are shown in squares by homology with the corresponding positions in the V-DOMAINs. Positions 45 and 77 which delimit the characteristic transversal 'CD' strand of the C-LIKE-DOMAINs are also shown in squares. Hatched circles correspond to missing positions according to the IMGT unique numbering for C-DOMAINs and C-LIKE-DOMAINs. Arrows indicate the direction of the beta strands (IMGT Repertoire, http://imgt.cines.fr).

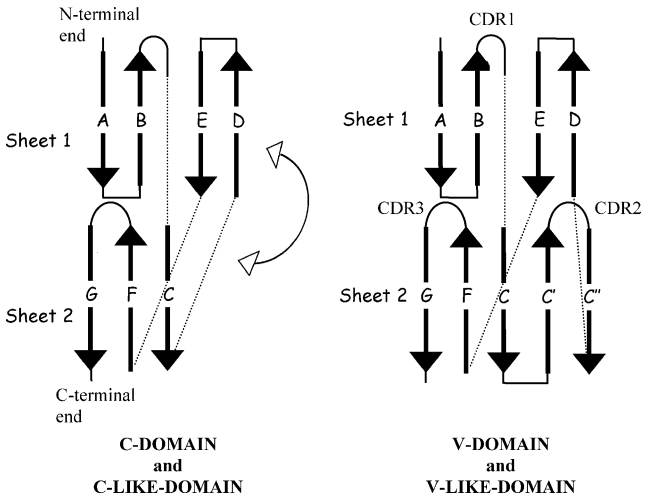

Fig. 6. Schematic representations of the C-DOMAIN and C-LIKE-DOMAIN, and of the V-DOMAIN and V-LIKE-DOMAIN. A double-headed arrow shows that the D strand can be localized in sheet 1 (on the back) or sheet 2 (on the front) depending from the length of the CD transversal strand. The second part of the A strand can be located in sheet 2 and is then designated as A′. This feature, described as 'strand A switching' in the literature, is not shown in IMGT Colliers de Perles, as this can be determined or verified only if 3D structures are available.

Table 3
Correspondence between the IMGT classification of the IgSF domains (C-DOMAIN, C-LIKE-DOMAIN, V-DOMAIN and V-LIKE-DOMAIN) and the diverse designations found in the literature

IMGT IgSF domains	C-DOMAIN (for IG and TR)	C-LIKE-DOMAIN (for proteins other than IG and TR)				V-DOMAIN (for IG and TR)	V-LIKE-DOMAIN (for proteins other than IG and TR)
Literature	C1[a] [32]	C1[a] [32]	C2 [32] or H [57]	I1 [58] or I [55,61]	I2 [58] or E [61]	V [32]	V [32]
Sheet 1	ABED	ABED	ABE	ABED	ABE	ABED	ABED
Sheet 2	GFC	GFC	GFCD (or GFCC′)	A′GFC	A′GFCD (or A′GFCC′)	A′GFCC′C″	A′GFCC′C″

[a] In the literature C1 is used for the IG and TR C-DOMAINs and for the MHC C-like domains. The diverse designations found in the literature for the C-set domain (C1, C2, Il, 12) are based on strand localisation in sheets 1 and 2. These designations are not used for the assignment of IMGT C-DOMAIN and C-LIKE-DOMAIN sequences, for the following reasons: (i) they are frequently extrapolated to sequences for which no 3D structures are available, therefore leading to misinterpretation, (ii) the designation 'strand C′' used in the literature for the C2 (or H) and I2 (or E) domain types is not the equivalent of strand C′ of the V-DOMAINs and V-LIKE-DOMAINs, (iii) the distinction made in the literature between C1 (as found in proteins of the immune adaptative response) and C2 (as found in other proteins) may suffer exceptions, (iv) as more 3D structures become available, and given the heterogeneity of the loop and strand lengths for C-DOMAINs and C-LIKE-DOMAINs, more intermediate or variant structures may be described (as shown by the 2C T cell receptor 3D structure IMGT/3Dstructure-DB: 1tcr_A).

Acknowledgements

We are grateful to Géraldine Folch, Chantal Ginestoux, Véronique Giudicelli, Joumana Jabado-Michaloud, Céline Protat, Dominique Scaviner and Denys Chaume for their helpful discussions. We thank Laurent Douchy and Bertrand Monnier for contribution to the tables, and Nora Bonnet-Saidali and Anita Gomez for editorial work. IMGT is funded by the European Union's 5th PCRDT (QLG2-2000-01287) program, the Centre National de la Recherche Scientifique (CNRS), and the Ministère de l'Education Nationale et de de la Recherche.

References

[1] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2003;31:307–10.

[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 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, Jabado-Michaloud J, Magris S, Scaviner D, Thouvenin V, Combres K, Girod D, Jeanjean S, Protat C, Yousfi Monod M, Duprat E, Kaas Q, Pommié C, Chaume D, Lefranc G, IMGT-ONTOLOGY for Immunogenetics and Immunoinformatics. In Silico Biology 2003;4,0004. http://www.bioinfo.de/isb/2003/04/0004/. *In Silico* Biology 2004;4: 17–29.

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

[7] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®. 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 p. 136–42, 216–22, 250–52].

[8] Lefranc M-P. IMGT, the international ImMunoGeneTics information system®. http://imgt.cines.fr. In: B.K.C. Lo Antibody engineering: methods and protocols. Methods in Molecular 248 Biology, 2nd ed. Totowa, NJ: Humana press; 2003;248:27–49 [chapter 3].

[9] Giudicelli V, Chaume D, Bodmer J, Müller W, Busin C, Marsh S, Bontrop R, Lemaitre M, Malik A, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1997;25:206–11.

[10] Lefranc M-P, Giudicelli V, Busin C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Chaume D. IMGT, the International ImMunoGeneTics database. Nucl Acids Res 1998;26:297–303.

[11] Lefranc M-P, Giudicelli V, Ginestoux C, Bodmer J, Müller W, Bontrop R, Lemaitre M, Malik A, Barbié V, Chaume D. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 1999;27:209–12.

[12] Ruiz M, Giudicelli V, Ginestoux C, Stoehr P, Robinson J, Bodmer J, Marsh SG, Bontrop R, Lemaitre M, Lefranc G, Chaume D, Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2000;28:219–21.

[13] Lefranc M-P. IMGT ImMunoGeneTics Database. International BIOforum 2000;4:98–100.

[14] Lefranc M-P. IMGT, the international ImMunoGeneTics database. Nucl Acids Res 2001;29:207–9.

[15] Giudicelli V, Lefranc M-P. Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 1999;15:1047–54.

[16] Lefranc M-P. Unique database numbering system for immunogenetic analysis. Immunol Today 1997;18:509.

[17] Lefranc M-P. The IGMT unique numbering for immunoglobulins. T cell receptors and Ig-like domains. The Immunologist 1999;7:132–6.

[18] Lefranc M-P, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 2003;27:55–77.

[19] Chaume D, Giudicelli V, Lefranc M-P. IMGT/LIGM-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32 http://www3.oup.co.uk/nar/database/summary/504.

[20] Folch G, Bertrand J, Lemaitre M, Lefranc M-P. IMGT/PRIMER-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32 http://www3.oup.co.uk/nar/database/summary/505.

[21] Giudicelli V, Lefranc M-P. IMGT/GENE-DB. In: The molecular biology database collection. Nucl Acids Res 2004;32. http://www3.oup.co.uk/nar/database/summary/503.

[22] 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. Nucl Acids Res 2004;32:D208–D210.

[23] 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. Nucl Acids Res 2004; 32:W435–W440.

[24] Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. 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.

[25] 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.

[26] 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.

[27] 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.

[28] Ruiz M, Lefranc M-P. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 2002;53:857–83.

[29] Lefranc M-P, Lefranc G. The immunoglobulin FactsBook. London, UK: Academic Press; 2001, pp. 458.

[30] Lefranc M-P, Lefranc G. The T cell receptor FactsBook. London, UK: Academic Press; 2001, pp. 398.

[31] Lesk AM, Chothia C. Evolution of proteins formed by beta-sheets II. The core of the immunoglobulin domains. J Mol Biol 1982;160:325–42.

[32] Williams AF, Barclay AM. The immunoglobulin family: domains for cell surface recognition. Annu Rev Immunol 1988;6:381–405.
[33] Bork P, Holm L, Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. J Mol Biol 1994;242:309–20.
[34] 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 Recogn 2004;17:17–32.
[35] Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, Harte N, Kanz C, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Stoehr P, Stoesser G, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R, The EMBL. Nucleotide sequence database. Nucl Acids Res 2004;32:D27–D30.
[36] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank: update. Nucl Acids Res 2004;32: D23–D26.
[37] Miyazaki S, Sugawara H, Ikeo K, Gojobori T, Tateno Y. DDBJ in the stream of various biological data. Nucl Acids Res 2004;32:D31–D34.
[38] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28:235–42.
[39] Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. Genomics 2002;79:464–70.
[40] Letovsky SI, Cottingham RW, Porter CJ, Li PW. GDB: the human genome database. Nucl Acids Res 1998;26:94–9.
[41] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucl Acids Res 2001;29:137–40.
[42] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 1997;13:163.
[43] Garrett TPJ, Wang J, Yan Y, Liu J, Harrison SC. Refinement and analysis of the structure of the first two domains of human CD4. J Mol Biol 1993;234:763–78.
[44] Bajorath J, Peach RJ, Linsley PS. Immunoglobulin fold characteristics of B7-1 (CD80) and B7-2 (CD86). Protein Sci 1994;3:2148–50.
[45] Huang Z, Li S, Korngold R. Immunoglobulin superfamily proteins: structure, mechanisms, and drug discovery. Biopolymers 1997;43:367–82.
[46] Samaridis J, Colonna M. Cloning of novel immunoglobulin superfamily receptors expressed on human myeloid and lymphoid cells: structural evidence for new stimulatory and inhibitory pathways. Eur J Immunol 1997;27:660–5.
[47] Chretien I, Marcuz A, Courtet M, Katevuo K, Vainio O, Heath JK, White SJ, Du Pasquier L. CTX, a Xenopus thymocyte receptor, defines a molecular family conserved throughout vertebrates. Eur J Immunol 1998;28:4094–104.
[48] Halaby DM, Mornon JP. The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity. J Mol Evol 1998;46:389–400.
[49] Ioerger TR, Du C, Linthicum DS. Conservation of cys–cys trp structural triads and their geometry in the protein domains of immunoglobulin superfamily members. Mol Immunol 1999; 36:373–86.
[50] Davis RS, Dennis Jr G, Odom MR, Gibson AW, Kimberly RP, Burrows PD, Cooper MD. Fc receptor homologs: newest members of a remarkably diverse Fc receptor gene family. Immunol Rev 2002;190:123–36.
[51] Guethlein LA, Flodin LR, Adams EJ, Parham P. NK cell receptors of the orangutan (*Pongo pygmaeus*): a pivotal species for tracking the coevolution of killer cell Ig-like receptors with MHC-C. J Immunol 2002;169:220–9.
[52] Guselnikov SV, Ershova SA, Mechetina LV, Najakshin AM, Volkova OY, Alabyev BY, Taranin AV. A family of highly diverse human and mouse genes structurally links leukocyte FcR, gp42 and PECAM-1. Immunogenetics 2002;54:87–95.
[53] 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.
[54] Jones EY. The immunoglobulin superfamily. Curr Opin Struct Biol 1993;3:846–52.
[55] Harpaz Y, Chothia C. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J Mol Biol 1994;238: 528–39.
[56] Smith DK, Xue H. Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 1997;274:530–45.
[57] Hunkapiller T, Hood L. Diversity of the immunoglobulin gene superfamily. Adv Immunol 1989;44:1–63.
[58] Casasnovas JM, Stehle T, Liu JH, Wang JH, Springer TA. A dimeric crystal structure for the N-terminal two domains of intercellular adhesion molecule-1. Proc Natl Acad Sci USA 1998;95:4134–9.
[59] Garcia KC, Degano M, Stanfield RL, Brunmark A, Jackson MR, Peterson PA, Teyton L, Wilson IA. An alphabeta T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. Science 1996;274:209–19.
[60] Holden HM, Ito M, Hartshorne DJ, Rayment I. X-ray structure determination of telokin, the C-terminal domain of myosin light chain kinase, at 2.8 Å resolution. J Mol Biol 1992;227: 840–51.
[61] Sun P, Boyington J. Overview of protein folds in the immune system In: Curr protocols immunology. New York, NY: Wiley; 2000, A.1N.1–A.1N.45.
[62] 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-31,103–104. http://www.inra.fr/eccb2003/posters/pdf/Annot_Giudicelli_20030528_160703.pdf.