Article

The T cell receptor (TRB) locus in *Tursiops truncatus*: from sequence to structure of the alpha/beta heterodimer in the human/dolphin comparison

Giovanna Linguiti[1,*], Sofia Kossida[2], Ciro Leonardo Pierri[3], Joumana Jabado-Michaloud[2], Geraldine Folch[2], Serafina Massari[4], Marie-Paule Lefranc[2], Salvatrice Ciccarese[1,*]and Rachele Antonacci[1]

[1]University of Bari, "Aldo Moro", Department of Biology, Bari, Italy
[2]IMGT®, the international ImMunoGeneTics information system®, Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), Montpellier, France
[3] Department of   Bioscience, Biotechnologie and Biopharmaceutics, Bari, Italy
[4]Department of Biological and Environmental Science and Technologies, University of Salento, Lecce, Italy
* Correspondence: salvatricemaria.ciccarese@uniba.it and giovanna.linguiti@uniba.it

Abstract: Bottlenose dolphin (*Tursiops truncatus*) belongs to the Cetartiodactyla and similarly to the other cetaceans represent the most successful mammalian colonization of the aquatic environment. Here we report a genomic, evolutionary and expression study of *Tursiops truncatus* T cell receptor beta (TRB) genes. Although the organization of the dolphin TRB locus is similar to that of the other artiodactyl species, with three in tandem D-J-C clusters located at its 3' end, its uniqueness is given by the reduction of the total length due essentially to the absence of duplications and to the deletions that have drastically reduced the number of the germline TRBV genes. We have analyzed the relevant mature transcripts from two subjects. The simultaneous availability of rearranged TRA and TRB cDNA from the peripheral blood of one of the two specimens, and the human/dolphin amino acids multi sequence alignments, allowed us to calculate the most likely interactions at the protein interface between the alpha/beta heterodimer in complex with major histocompatibility class I (MH1) protein. Interacting amino acids located in the CDR-IMGT of the dolphin variable V alpha and beta domains were identified. According to comparative modelisation, the atome pair contact sites analysis between the human MH1 grove (G) domains and the TR V domains confirms conservation of the structure of the dolphin TR/pMH.

Keywords: T cell receptor, dolphin genome , TRB locus, TRBV, TRBJ, TRBD and TRBC genes, TRA and TRB gene expression analysis, multiple sequence alignments (MSA), 3D modelisation, IMGT

1. Introduction

Bottlenose dolphin (*Tursiops truncatus*) and the other cetaceans phylogenetically belong to the Cetartiodactyla clade, which includes Cetacea and Artiodactyla [1,2]. The divergence of cetaceans from their terrestrial ancestors about 54 million years ago [3] resulted in a complete adaptation to aquatic life. In addition to anatomical and physiological innovations required for life in water, cetaceans must have been confronted with challenges from changing environmental pathogens while they transitioned from land to sea [4,5]. These challenges exerted intensified selection pressure on the whole genomes of cetacean lineage, including the Delphinidae family [6], as well as on gene families related to the immune system [7].

αβ or γδ T cell receptor (TR) heterodimers are members of the Ig superfamily. Each gene family encoding for the α, β, γ and δ chains is organized in a distinct genomic locus that are named TRA, TRB, TRG or TRD based on the different TR chain. To obtain an effective repertoire of TR, each gene family encodes the variable domain of its own chain through a random process of DNA rearrangement that involves somatic recombination

of *V* (*variable*), *D* (*diversity*) and *J* (*joining*) genes for the β and δ chains, and *V* and *J* genes for the α and γ chains. After transcription, the resulting rearranged V(D) J region, is spliced to the *C* (*constant*) gene that encodes the constant domain of the receptor. The resulting chain is a protein with the variable domain composed of seven distinguishable regions: three hypervariable loops or complementarity-determining regions (CDR-IMGT) and four framework regions (FR-IMGT). Two of the CDR loops, CDR1-IMGT and CDR2-IMGT, are encoded by the *V* gene. The third CDR loop (CDR3-IMGT) reflects the ability of the *V* gene to rearrange to any (*D*) *J* gene [8]. The *V*, *D*, *J* and *C* genes exist within each TR locus as multigene subfamilies, and the total number of their gene members varies in different species. The identification of the TR germline repertoire within a given species provides information about its potential for the adaptive immune response against antigens.

Recent studies enabled us to determine and characterize the genomic organization of the dolphin TRG, TRA and TRD loci [9]. Comparative studies showed that the dolphin TRG locus is the simplest and the smallest among the mammalian TRG loci identified to date [10-12]and is organized in a single V-J-C cassette comprising only 2 TRGV, 3 TRGJ genes and a single TRGC gene [9].

In particular, comparative analysis highlighted the overall organization of the dolphin TRG locus resembles the structure of the TRGC5 cassette, the ancestral unit that gave rise to reiterated duplications in ruminants TRG locus [13,14]. Dolphin genome assembly, although fragmented and incomplete, confirmed that the TRD genes are clustered within the TRA locus, as in eutherians and birds, and revealed a conserved structure and organization with respect to the artiodactyls and the others terrestrial mammals analyzed so far [15,16,17]. Moreover, expression data on the dolphin TRG and TRA/TRD potential repertoire provided evidence for an unusual ratio of productive/unproductive transcripts which arise from the TRG V-J gene rearrangement and allowed the prediction of the most likely and most stable γδ pairing for a "public" gamma delta TR repertoire [9].

The aim of this work was to provide a scenario on the T cell receptor loci in *Tursiops truncatus* analyzing the germline and expressed TRB repertoires. Thus, with the release of the most recent dolphin genome assembly (mTurTru1.mat.Y), we deduced the genomic structure of the dolphin TRB locus, we characterized and annotated its genes and we have highlighted its uniqueness in comparative analysis with other artiodactyls. Expression analysis allowed us to evaluate the contribution of each *V*, *D*, *J*, single gene to the dolphin TRB repertoire. Starting from the TRA and TRB cDNA sequences, we were able to build a 3D human/dolphin comparative modeling of the *T.truncatus* TR alpha/beta in complex with major histocompatibility class I (MH1) protein and beta-2 microglobulin (B2M) and to identify the resulting interaction energy between aminoacids sidechains in the CDR-IMGT [IMGT® http://www.imgt.org] of the dolphin TR/MH structure.

2. Materials and Methods

2.1. Genome analyses

To analyse the TRB locus of the bottlenose dolphin (*Tursiops truncatus*), the mTurTru1.mat.Y assembly of the whole genome shotgun sequence (BioProject ID: PRJNA625792) obtained by Illumina NovaSeq sequencing technology, was searched using the BLAST algorithm.

A sequence of 277,931 bp (gaps included) was retrieved directly from a unique genomic scaffold mapped on chromosome 9 (NW_022983121.1) of the reference sequence NC_047042.1, identical to CM022282.1, available at NCBI from 93263345 to 93541276.

The *Tursiops truncatus* TRB locus is in forward (FWD) orientation on chromosome 9. In particular, the analysed region comprises the MOXD2 (monooxygenase DBH like 2,

pseudogene) and EPHB6 (EPH receptor B6) genes, already annotated within the scaffold and flanking respectively, the 5' and 3' ends of the TRB locus.

The genomic sequences of the TRB locus of humans (*Homo sapiens*) [8,18-20] and of artiodactyla I.e. Arabian camel or dromedary (*Camelus dromedarius*), pig (*Sus scrofa*) and sheep (*Ovis aries*) [21-24] were used against the dolphin genome sequence to identify, based on homology by the BLAST program, the corresponding genomic TRBV, TRBD, TRBJ and TRBC genes.

The beginning and the end of each coding exon were identified with accuracy by the presence of splice sites or of recombination signal (RS) sequences (V-RS, 5'D-RS and 3'D-RS, J-RS for TRBV, TRBD and TRBJ genes, respectively). The locations of the TRB genes are provided in Supplementary Table S1. The sequence comparison has also allowed for the identification and characterization of the dolphin protease genes (TRY). The locations of the TRY genes and the flanking TRB locus, MOXD2 and EPBH6 genes, are provided in Supplementary Table S2.

Computational analysis of the dolphin TRB locus was conducted with the following programs: RepeatMasker for the identification of genome-wide repeats and low complexity regions (RepeatMasker at http://repeatmasker.org), Pipmaker (25, http://pipmaker.bx.psu.edu/ pipmaker/) for the alignment of the dolphin sequence with human (GenBank accession number L36092), goat (GenBank accession number NC_030811.1), pig (GenBank accession number NC_0101460.4) and Arabian camel (GenBank accession number NW_011591622.1) counterparts. RepeatMasker screens DNA sequences for interspersed repeats and low complexity DNA sequences.

2.2. Classification of the dolphin TRB genes

Considering the percentage of nucleotide identity of the genes with respect to human [8,18,26] and the other mammalian species and based on the genomic position within the locus, each TRB gene was classified, and the nomenclature was established according to the IMGTconcepts of classification, approved by HGNC, VGNC, NCBI and the IUIS Nomenclature Committee for IG and TR of all vertebrate species with jaws (gnathostomata) from fish to humans [8,20,27] (Supplementary Table S1). The functionality of the *V, D, J* and *C* genes, based on the IMGT Scientific chart rules [20], was predicted through the manual alignment of sequences adopting the following parameters: (a) identification of the leader sequence at the 5' of the TRBV genes; (b) determination of proper RS located at 3' of the TRBV (V-RS), 5' and 3' ends of the TRBD (5'D-RS and 3'D-RS) and 5' of the TRBJ (J-RS), respectively; (c) determination of conserved acceptor and donor splicing sites (IMGT® http://www.imgt.org, IMGT Education > IMGT Aide-mémoire > Splicing sites); (d) estimation of the expected length of the coding regions; (e) absence of frameshifts and stop codons in the coding regions of the genes (IMGT® http://www.imgt.org, IMGT Scientific chart > 1. Sequence and 3D structure identification and description > IMGT functionality).

The TRBV genes were assigned to 20 different subgroups, based on the percentage of nucleotide identity by using Clustal Omega alignment tool, which is available at EMBL-EBI website (http://www.ebi.ac.uk/), adopting the criterion that sequences with a nucleotide identity of more than 75% in the V-region belong to the same subgroup.

The TRBD, TRBJ and TRBC genes were annotated, according to the similarity with the other artiodactyl species and their specific organization in three D-J-C clusters, numbered C1, C3 and C2 from 5' to 3' in the locus [22-24]. Each TRBJ gene of the TRBJ1, TRBJ2 and TRBJ3 sets was designed by a hyphen and a number corresponding to their position in the cluster. They were all predicted to be functional, except for the TRBJ1-5 and TRBJ3-6 which are pseudogenes owing to STOP-CODON and TRBJ1-2 (non canonical J- NONAMER), TRBJ1-4 (non canonical J-HEPTAMER) and TRBJ3-2 (nonconserved J-MOTIF) classified as ORF (Supplementary Table S1).

2.3.   Phylogenetic analyses

The TRBV genes used for the phylogenetic analysis were retrieved from the following sequences deposited in the GEDI (for GenBank/ENA/DDBJ/IMGT/LIGM-DB) databases: L36092 (*Homo sapiens* TRB locus as annotated in IMGT/LIGM-DB database [8,18], NW_011591622, NW_011593440, NW_011591151, NW_011620189, NW_011616084, NW_011607149, NW_011601111 and LT837971 (*Camelus dromedarius* TRB locus contig as characterized by [23,24], NC_010460 (*Sus scrofa* TRB locus contig as characterized by [21]; and NC_030811.1 (*Capra hircus* TRB locus contig characterized by [28].

Multiple alignments of the gene sequences under analysis were carried out with the MUSCLE program [29]. The evolutionary analyses were conducted in MEGA7 [30]. We used the neighbour-joining (NJ) method to reconstruct the phylogenetic tree [31]. The evolutionary distances were computed using the p-distance method [32].

2.4. 5′ rapid amplification of cDNA ends (RACE) PCR

Blood samples were provided by Zoomarine Italia S.p.a. (Rome, Italy) and were collected from two unrelated dolphins, a male (Marco) and a female (Leah), identified with M and L letters, respectively.

Total RNA was isolated from peripheral blood leukocytes (PBL) using the Trizol method according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). About 5 μg of RNA was reverse transcribed with Superscript II (Invitrogen, Carlsbad, CA) by using a specific primer, TCB1L1 (5′ GCTGGGGTCCTCCTTGTC 3′). After linking a poly-C tail at the 5′end of the cDNAss, the cDNAds was performed with Platinum Taq Polymerase (Invitrogen) by using a specific primer, TCB1L2 (5′ TTCCCGTTCACCCACCAG 3′) as lower primer and an anchor oligonucleotide as upper primer (AAP) provided from the supplier (Invitrogen). PCR conditions were the following: one cycle at 94 °C for 1 min; 35 cycles at 94 °C for 30 s, 58 °C for 45 s, 72 °C for 1 min; a final cycle of 30 min at 72 °C. The products were then amplified in a subsequent nested PCR experiment by using a specific lower primer, TCB1L3 (5′ GTGGTGACGGGGTAGAAG 3′) and AUAP oligonucleotide as upper primer, provided from the supplier (Invitrogen). Nested PCR conditions were the following: one cycle at 94 °C for 1 min; 30 cycles at 94 °C for 30 s, 58 °C for 35 s, 72 °C for 30 s; a final cycle of 30 min at 72 °C. All the specific primers, TCB1L1, TCB1L2 and TCB1L3 were designed on the sequence of the first exon of the TRBC gene. The RACE products were then gel-purified and cloned using the StrataClone PCR Cloning Kit (Stratagene). Random selected positive clones for each cloning were sequenced by a commercial service. cDNA sequence data were processed and analyzed using the Blast program (http://www.blast.ncbi.nlm.nih.gov/Blast.cgi), Clustal Omega alignment tool (http://www.ebi.ac.uk) and IMGT_ tools (IMGT/V-QUEST [33,34] with integrated IMGT/JunctionAnalysis tools [35,36] and the IMGT unique numbering for the V domain [17,20]. All cDNA clones were registered in EMBL database with the Accession numbers from HG764428 to HG764459.

2.5. Crystal structure sampling via folding recognition and multiple sequence alignments (MSA)

The folding recognition methods implemented in pGenThreader and i-Tasser were used for highlighting TR alpha and TR beta homologous protein-crystallized structures. With this aim, the deduced amino acid sequences of the *Tursiops truncatus* TRA (TRAV20-S2*01, 5RAL27 clone, GenBank: LN610732.1; TRAV18-1*02-J20, 5RAL11 clone, GenBank: LN610716.1) and TRB (TRBV30-D2-J2-2, L5RBL27 clone, clone 5RBL27 in GenBank: HG764454.1; TRBV30-D2-J3-3, L5RBL13 clone, clone 5RBL13 in GenBank: HG764440.1) cDNAs were used as query sequences for running pGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/) and i-Tasser (https://zhanglab.ccmb.med.umich.edu/I-TASSER/) to screen the PDB, searching for the most similar deposited crystallized structures [37-41]. SPDBV was used for building a 3D all atom model of all the investigated *T. truncatus* TR alpha or TR beta chains by using the human TR alpha or TR beta available in the human crystallized TR in complex

with major histocompatibility class I (MH1) protein made of an I-ALPHA chain noncovalently associated with beta-2-microglobulin (B2M) [20], available under the pdb code "3hg1.pdb".

The human major histocompatibility class I (MH1 I-ALPHA) protein and beta-2 microglobulin (B2M) sequences were used as query sequences for identifying their counterparts in *T. truncatus*, through blastp searches. The identified protein sequences were modelled by using the structures of the human MH1 I-ALPHA and B2M available under the pdb code "3hg1.pdb", used as a protein template, by SPDB according to our validated protocols [40-42]. All the generated 3D all-atom models were energetically minimized by using the Yasara Minimization server [41-43].

2.6. 3D modeling of T.truncatus TR alpha and TR beta in complex with T.truncatus MH1I-ALPHA and B2M

The proposed 3D comparative protein complex consisting of *T. truncatus* TR alpha and TR beta in complex with *T. truncatus* MH1 I-ALPHA and B2M was obtained by superimposing the above-cited single chains domains (*T. truncatus* TR alpha, TR beta, MH1 I-ALPHA and B2M) on the corresponding 3D atomic coordinates of each corresponding chain within the crystallized human TR in complex with human MH1 I-ALPHA and B2M, available under the PDB code 3hg1.pdb [44], by PyMOL as previously described [41,42].

It should be noted that CANCER/MART-1 decapeptide was removed from the 3hg1 before starting the comparative modeling analysis. Superimposition operations were performed through the "super" command implemented in PyMOL, starting from the structural alignment of the analysed backbones. The "super" command allows aligning the selected proteins under investigation for performing a comparative structural analysis, due to its ability in providing a sequence-independent structure-based pairwise alignment. Notably, the "super" command is more robust than the "align" command because it successfully performs also superimposition of proteins with a lower sequence similarity. Then, it was possible to model/relax missing/buried residues located at the protein-protein interface, solving clashes and putative breaks in the backbone [39,41,42] . All the generated 3D all-atom models were energetically minimized by using the Yasara Minimization server [41-43]. The obtained final models were examined in VMD, PyMOL, and SPDBV by visual inspection searching for putative unsolved clashes [40-42]. Protein-protein binding regions were highlighted by selecting residues within 4 Å at the protein-protein interface, in the superimposed structures.

The FoldX AnalyseComplex assay was performed to determine the interaction energy between the four generated *T. truncatus* TR alpha and TR beta models and between each of the four combined TR alpha/beta in complex with the *T. truncatus* MH1 I-ALPHA and B2M, but also for determining the interaction energy between the human counterparts of the crystallized 3hg1.pdb as a reference system, a validation strategy and for comparative purposes.

The way the FoldX AnalyseComplex operates is by unfolding the selected targets and determining the stability of the remaining molecules and then subtracting the sum of the individual energies from global energy. More negative energies indicate a better binding. Positive energies indicate no binding [45,46].

3. Results

#### 3.1. Genomic structure of the Tursiops truncatus TRB locus

We employed the latest version of the whole genome assembly (mTurTru1.mat.Y) of the bottlenose dolphin (*Tursiops truncatus*) submitted by Vertebrate Genomes Project, to NCBI (BioProject ID: PRJNA625792) to identify the TRB locus in this species. We retrieved a sequence approximately 278 kb in length (gaps included), comprising the

MOXD2 and the EPHB6 genes that flank the 5' and 3' ends, respectively, of all mammalian TRB loci studied to date.

A standard BLAST search of the genomic sequence was then performed by using human [18,20] and artiodactyl sequences i.e. *Sus scrofa*, *Camelus dromedarius* and *Ovis aries* [21-24] to identify and annotate all dolphin TRB genes.

The sequence analysis showed a conserved general structural organization of the dolphin TRB locus with a library of TRBV genes positioned at the 5' end of the D-J-C clusters, followed by a single TRBV gene located at the 3' end in an inverted transcriptional orientation (Figure 1 A).
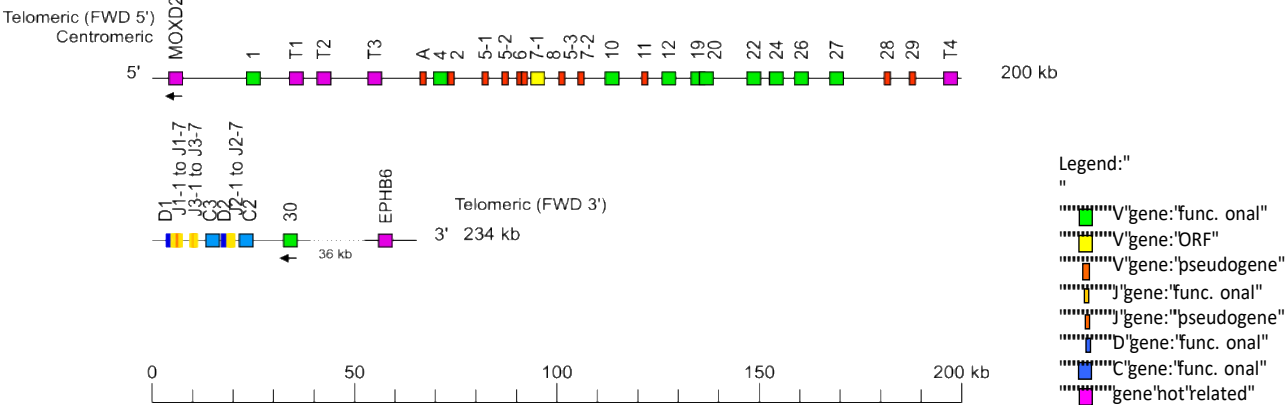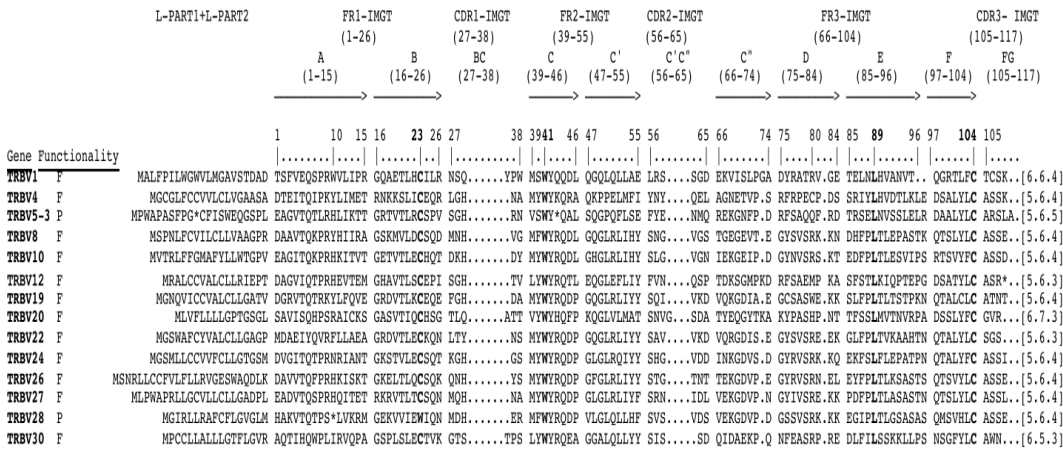

Figure 1. (A) Schematic representation of the genomic organization of the bottlenose dolphin (*Tursiops truncatus*) TRB locus deduced from the genome assembly mTurTru1.mat.Y (CM022282.1). The diagram shows the positions of all the related and unrelated TRB genes according to IMGT nomenclature and locus representation [19]. The boxes representing the genes are not to scale. The exons are not shown. The arrows indicate the transcriptional orientation of the genes. **(B)** The IMGT Protein Display of the dolphin TRBV genes. Only functional genes and in-frame pseudogenes are shown. The description of the strands and loops and of the FR-IMGT and CDR-IMGT is according to the IMGT unique numbering for V-REGION [47]. The five conserved amino acids of the V-DOMAIN (1st-CYS 23, except for the TRBV28 pseudogene, CONSERVED-TRP 41, hydrophobic AA 89, 2nd-CYS 104 and J-PHE 118) are indicated in bold. The amino acid length of the CDR-IMGT AA is also indicated in square brackets.

Moreover, it revealed that the D-J-C cluster is organized in three D-J-C sets similar to those found in ruminants [24,48], pig [21] and camel species [22,23,49], confirming the tight evolutionary relationship between Cetacea and Artiodactyla. D-J-C cluster 1 contains one TRBD, six TRBJ genes but lacks the TRBC gene, due to a gap in the genomic sequence. Likewise, the second set (D-J-C cluster 3) lacks the TRBD gene and it includes seven TRBJ and one TRBC genes. D-J-C cluster 2 is the only complete set with one TRBD,

seven TRBJ and one TRBC genes (Figure 1A). Approximately 10 kb away from the TRBC2 gene lies the TRBV30 gene, with an inverted transcriptional orientation. The classification, position and predicted functionality of all TRB genes are reported in Supplementary Table S1.

The comparison of the entire dolphin TRB locus sequence with those previously characterized in mammals, allowed us to identify and partially annotate four unrelated TRB genes consisting of a group of trypsin-like serine protease (TRY) genes that are typically interspersed among the TRB genes. Three TRY genes are located downstream of TRBV1, and one (TRY4) is located upstream of the D-J-C region (Figure 1A). The gene predicted functionality (only the TRY3 gene is functional) is reported in Supplementary Table S2, together with the position of all TRY genes. The classification, position and predicted functionality of the MOXD2 and EPHB6 genes, which delimit the TRB locus, are also reported.

3.2. Description of the D-J-C clusters region

The structure of the dolphin D-J-C clusters region is similar to that of other artiodactyl species [21-24,28,48,49], with the TRBD, TRBJ and TRBC genes distributed within three in tandem D-J-C clusters positioned at the 3' end of the TRB locus, with a number 1, 3 and 2 corresponding to their position from 5' to 3' (Figure 1A). We have attributed the name D-J-C cluster 3 to the central cluster as in sheep, goat, dromedary and pig [21-24,28,49].

The nucleotides and deduced amino acid sequences of all the TRBJ genes identified in the region are reported in Figure 2A. Due to the lack of the TRBC1, the TRBJ1 genes was assessed by comparison with the human and artiodactyl corresponding gene sequences. All TRBJ genes are typically 40-53 bp in length and are predicted to be functional, except for TRBJ1-5 and TRBJ3-6 pseudogenes for a STOP-CODON in the coding region (Figure 2A, Supplementary Table S1) and TRBJ1-2, TRBJ1-4 and TRBJ3-2 classified as open reading frame (ORF). Each TRBJ is flanked by a 12 RS at the 5' end and by a donor splice site at the 3' end. All the RSs are well conserved with respect to the consensus. The nucleotide and deduced amino acid sequences of the only two TRBD genes are shown in Figure 2B. They consist of a 13 bp (TRBD1) and 15 bp (TRBD2) G-rich stretches that can be productively read through their three coding phases and encode 1-3 glycine residues, depending on the phase. The RSs that flank the 5' and 3' sides of the coding region are well conserved with respect to the consensus.

The two dolphin available TRBC genes are composed of four exons and three introns and encode a protein of 177 amino acids. The difference of ten nucleotides between them result in three amino acid (AA) changes, two in the extracellular and one in cytoplasmic regions (Figure 2C). The C-domain encoded by EX1 is 129 AA in length as in humans and pig because of the identical length of the FG loop with respect to other artiodactyl species, based on the IMGT unique domain for C-DOMAIN [50]. The C region also comprises a connecting region (CO) of 21 AA (encoded by EX2 and the 5' part of EX3) with a cysteine involved in the interchain disulphide binding, a TM of 21 AA (encoded by the 3' part of EX3 and the first codon of EX4) and a cytoplasmic region (CY) of 5 AA (encoded by EX4).

A

```
           J-NONAMER    J-SPACER    J-HEPTAMER
           GGTTTTTGT  ***********  CACTGTG
Gene name  GGTTTTTGT  ***********  CACTGTG                                    J-REGION                                      5'splice donor
TRBJ1-1    gctttcctc  cttactccgcgt cactgtg      CAAACACTGACGTCTTCTTCGGAGAAGGCACCAGGCTCACGGT CGTAG     gtaaga
                                                  N   T   D   V   F  F G E G T R L T V V
TRBJ1-2    cattgtaag  gcggctgtattc tgatgtg      CTATGACTATACCTTCGGCCCAGGGACCAAGCTGACTGTT GTAG       gtaggt
                                                  Y   D   Y   T  F G P G T K L T V V
TRBJ1-3    ggttttgaa  gtgaatctggga ggctgtg      CTTTGCAGACACCTACTATTTTGGGGAAGGAAGCCGGCTCACT GTCATAG  gtcata
                                                  F   A   D   T   Y   Y  F G E G S R L T V I
TRBJ1-4    gtttcccta  ccgggctttcgt gttctgc      GACTAATGAAAGCCTGTATTTCGGCAGCGGAACGAAGCTTTCTGGGCTGG   gtaagt
                                                  T   N   E   S   L   Y  F G S G T K L S G L
TRBJ1-5    gggtttgcc  acgcccgtgtta tactgtg      GGGCAACCGGGGGCAGCACTTTGGAGGTGGGACTTGACTCCCT GTCCTGG  gtcctg
                                                  G   N   R   G   Q   H  F G G G T * L P V L
TRBJ1-6    ggttttcct  acagcctcctgc agctgtg      TTCCTATAATTCGCCCCTTCCACTTTGGGACCGGCACCAGGCTCACCGTGACAG gtatca
                                                  S   Y   N   S   P   L   H  F G T G T R L T V T
TRBJ3-1    gaattcttg  gtagccccttcc cactgtg      CTCCTACAGTGAGCAGTACTTCGGGCAGG GACCAGGCTCACGGTGCTAA     gtaagg
                                                  S   Y   S   E   Q   Y  F G P G T R L T V L
TRBJ3-2    gtttgctcc  tgggtcccaagg ctgcgtg      GATACCACACAGCTGTATTCTGGGGATGGTTCCAAGCTGACTGTGCT GG    gtcagg
                                                  D   T   T   Q   L   Y   S  G D G S K L T V R
TRBJ3-3    cagttttgt  cctgggcctcag ggctgtg      AGCCAAAACACTCAGTACTTCGGCGGGGGCACCCGGCTGACTGTGCTAG      gtgagc
                                                  S   Q   N   T   Q   Y  F G G G T R L T V L
TRBJ3-4    cctttttgt  gctgagccctga agctgtg      AGCACTGACACTCAGTACTTCGGGCCGGGCACGCGGCTCCTG GTGCTAG     gtgagc
                                                  S   T   D   T   Q   Y  F G P G T R L L V L
TRBJ3-5    ggtttttgc  gcggcgctgggg ggccgtg      ACTCAGCCCCTGTACTTCGGGCCGGGCACGCGGCTCCTGGTGCTAG        gtgagc
                                                  T   Q   P   L   Y  F G P G T R L L V L
TRBJ3-6    ggtttgggt  ggggagccccgg ggctgtg      CTTTACAGGCGCCGCCCTGCCCTTCGGGGCCGGCAGCCCCTGACCGTGGTG G  gtgagt
                                                  F   T   G   A   A   L   P  F G A G S * P W W
TRBJ3-7    ggtttgcgt  gctggactggga ctctgtg      CTCCAATGAGCGGTATTTCGGCCCAGGCACCAGGCTCACGGTCACAG       gtgaga
                                                  S   N   E   R   Y  F G P G T R L T V T
TRBJ2-1    gaattcttg  gtagcccctttcc cactgtg     CTCCTACAATGAGCTGCACTTCGGGCAGGGACCAGGCTCACGGT GCTAA    gtaagg
                                                  S   Y   N   E   L   H  F G P G T R L T V T
TRBJ2-2    agtttgtgc  ctgagttcccag ggctgtg      TGAACACCGAGCTGTATTTTGGCAATGGTTCCAAGCTGACTGTGCTGG      gtaagg
                                                  N   T   E   L   Y  F G N G S K L T V L
TRBJ2-3    ggttttttgt cctgggcctcag ggctgtg      AGCAACGACCCTCTGTTCTTCGGCGAGGGCACCCGGCTGACTGTGCTAG      gtgagc
                                                  S   N   D   P   L   F  F G E G T R L T V L
TRBJ2-4    cctttttgt  gctgagcccaga agctgtg      AGCACTGAAACTCAGTACTTCGGCCGCGGCACCCGGCTGTCG GTGCTAG     gtgcta
                                                  S   T   E   T   Q   Y  F G A G T R L S V L
TRBJ2-5    ggttttttgc gcggcgctgggg ggccgtg      ACTCTCCCCCGGTTTTCGGGCCGGGCACGCGGCTCCTGGTGCTA G        gtgagc
                                                  S   P   P   V  F G P G T R L L V L
TRBJ2-6    ggtttgggt  ggggagccccgg ggctgtg CTTTACAGGCGCCGCCCTGCCCTTCGGGGCCGGCAGCCGGCTGACCGTGGTGG      gtgagt
                                                  F   T   G   A   A   L   P  F G A G S R L T V V
TRBJ2-7    gggtttgag  tgcggggcttct cactgtg      CTCCTATGAGCAGTATTTTGGGCCCAGGCACCAGGCTCACGGTCATAG      gtaaga
                                                  S   Y   E   Q   Y  F G P G T R L T V I
```

B

```
          5'D-NONAMER  5'D-SPACER  5'D-HEPTAMER               3'D-HEPTAMER  3'D-SPACER               3'D-NONAMER
          GGTTTTTGT    ***********  CACTGTG                    CACAGTG      ***********************  ACAAAAACC
gene name GGTTTTTGT    ***********  CACTGTG     D-REGION       CACAGTG      ***********************  ACAAAAACC
  TRBD1   cgttttttgt   ataaagctgtaa cgttgtg     AGGACAGGGGGGC  cacagtg      attcaaccctacgggaatccttt   acaaaaacc
                                                  R   T  G  G
                                                  G   Q  G  G
                                                  D   R  G
  TRBD2   cattttttgt   ttcatagtgtaa cattgtg     GGGACTTCGGGGGGG cacgat g     attcagttagaggaagggctttt   acaaaacgc
                                                  G   T  S  G  G
                                                  G   L  R  G
                                                  D   F  G  G
```

C

```
                                                                                            C-DOMAIN
                                                                                             [EX1]

                A        AB          B           BC         C       CD        D            DE              E        EF   F           FG                G
              (1-15)   (16-26)    (27-38)    (39-45)   (77-84)              (85-96)       (97-104)  (105-117)         (118-128)
              -------->          ----------->  ------------>    ------------------>                  -------->       ----------------->    ----------------->
              1    10 15  16      23 26 27      36  3941 45      77 80 84              85 89    96  97  104 105              117 118121   128
              87654321|......|....|123|......|..|! |........|..|....|12345677654321|..|.....|12|......| |......12345677654321......|    |.........|
Turtru TRBC2 (D)DLSKVTPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKQVET..GVSTDPQPHKEDPSR.NDSSYCLSSRLRVSASFWY.NPRNLFRC QAWFHGLSDKDEW..KQDRKPVTQNI SAEVRGRA...
Turtru TRBC3 (D)DLSKVTPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKQVET..GVSTDPQPHKEDPSR.NDSSYCLSSRLRVSASFWY.NPRNLFRC QAWFHGLSDEDEW..KQDRKPVTQII SAEVRGRA...
Homsap TRBC1 (E)DLNKVFPPEVAVFEPSEAEISH..TQKATLVCLAT GFFP..DHVE LSWWVNGKEVHS..GVSTDPQPLKEQPAL.NDSRYCLSSRLRVSATFWQ.NPRNHFRC QVQFYGLSENDEW..TQDRAKPVTQIV SAEAWGRA...
Homsap TRBC2 (E)DLKNVFPPEVAVFEPSEAEIS H..TQKATLVCLAT GFFP..DHVE LSWWVNGKEVHS..GVSTDPQPLKEQPAL.NDSRYCLSSRLRVSATFWQ.NPRNHFRC QVQFYGLSENDEW..TQDRAKPVTQIV SAEAWGRA...
Caphir TRBC1 (D)NLSRVHPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LTWWVNRKQVTT..GVSTDPEPYKEDLTQ.NDSRYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTDQDQW..EEQDRAKPVTQNI SAETWGRA...
Caphir TRBC2 (D)DLRQVHPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LTWWVNRKQVTT..GVSTDPEPYKEDLTQ.NDSRYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTDQDQW..EEQDRDRKPVTQNI SAETWGRA...
Caphir TRBC3 (D)DLSRVHPPKVAVFEPSEAEIS Q..TQKATLVCLAT GFYP..DHVE LTWWVNRKQVTT..GVSTDPEPYKEDLTQ.NDSRYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTDQDQW..DEQDRDKPVTQNI SAETWGRA...
Camdro TRBC1 (E)DLSQVKPPKVVVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKEVQS..GVSADPQPYREDPNL.SDSTYCLSSR LRVSATFWH.NPRNHFRC QVHFHGLTDKDRDQWKEGWPMPISQNV SAETWGRA...
Camdro TRBC2 (E)DLSQVKPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKEVQS..GVSTDPQPYREDPSL.SDSTYCLSSRLRVSATFWH.NPRNHFRC QVHFHGLTDKDRDQWKEDWPMPVSQNV SAEAWGRA...
Camdro TRBC3 (E)DLSQVKPPKVAVFEPSAAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKEVQS..GVSTDP QPYREDPSL.SDSTYCLSSRLRVSATFWH.NPRNHFRC QVHFHGLTDKDRDQWKE DWPMPVSQNV SAEAWGRA...
Susscr TRBC1 (E)DLQQVRPPKVAVFEPSEAEISR..TQKA TLVCLAT GFYP..DHVE LSWWVNGKQVQS..GVSTDLQPYREDPSR.NDSSYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTEDDEW..EYNWTKPITQNI SAEAWGKA...
Susscr TRBC3 (E)DLQQVRPPKVAVFEPSEAEISR..T QKATLVCLAT GFYP..DHVE LSWWVNGKQVQS..GVSTDLQPYREDPSR.NDSSYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTEDDEW..EYNWTKPITQNI SAEAWGKA...
Susscr TRBC2 (E)DLQQVRPPKVAVFEPSEAEISR..TQKATLVCLAT GFYP..DHVE LSWWVNGKQVQS..GVSTDLQPYREDPSR.NDSSYCLSSRLRVTAAFWH.NPRNHFRC QVQFYGLTEDDEW..EYNWTKPITQNI SAEAWGKA...

          CONNECTING-REGION  | TRANSMEMBRANE-REGION |CYTOPLASMIC-REGION

             [EX2]                 [EX3]                   [EX4]
Turtru TRBC2 (D)CGLTS   (A)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKKKDS
Turtru TRBC3 (D)CGLTS   (A)SYQQGVLSATILYEILLGKATLYAVLVSALVLMA M  VKKKDS
Homsap TRBC1 (D)CGFTS   (V)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKDF
Homsap TRBC2 (D)CGFTS   (V)SYQQ GVLSATILYEILLGKATLYAVLVSALVLMAM  VKRKDSRG
Caphir TRBC1 (D)CGVTS   (A)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKDS
Caphir TRBC2 (D)CGVTS   (A)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKDS
Caphir TRBC3 (D)CGVTS   (A)SYQQGVLSATILYEILLGKAT LYAVLVSALVLMAM  VKRKDS
Camdro TRBC1 (H)CGFTS   (V)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKDS
Camdro TRBC2 (D)CGFTS   (A)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKDS
Camdro TRBC3 (D)CGFTS   (V)SYQQGVLSATILYEILLGKATLYAVLVSALVLMAM   VKRKD
Susscr TRBC1 (D)CGFSS   (A)SYQQGVLSATLLYEILLGKAALYAVLVSALVLMAM   VKRKGA
Susscr TRBC3 (D)CGFSS   (A)SYQQGVLSATLLYEILLGKAALYAVLVSALVLMAT   VKRKGA
Susscr TRBC2 (D)CGFSS   (A)SYQQGVLSATLLYEILLGKATLYA VLVSALVLMAT  VKKKDS
```

Figure 2. Nucleotide and deduced amino acid sequences of the dolphin TRBD **(A)**, TRBJ **(B)** and TRBC **(C)** genes. The consensus sequence of the heptamer and nonamer is provided at the top of the figure and is underlined . The numbering adopted for the gene classification is reported on the left of each gene. In **(A)**, the inferred amino acid sequence of the TRBD genes in the three coding frames are reported. In **(B)**, the donor splice site for each TRBJ is shown. The canonical FGXG amino acid motifs are underlined. In **(C)**, IMGT Protein display of the dolphin TRBC genes as derived from the alignment by Clustal W with the human, sheep, dromedary and pig TRBC amino acid sequences. The strands and loops are according to the IMGT unique numbering for the C-DOMAIN [50].

3.3. Classification and phylogenetic analysis of the dolphin TRBV genes

In the retrieved sequence, we annotated 23 TRBV germline genes grouped into 20 distinct subgroups according to the criterion that sequences with a nucleotide identity of more than 75% in the V region belong to the same subgroup. Only two subgroups have more than one member, TRBV5 and TRBV7 with three and two genes, respectively.

Twelve out of 23 TRBV genes (approximately 50%) are predicted to be functional as defined by the IMGT rules (see the "Materials and methods"), and eleven are pseudogenes (Table 1 and Supplementary Table S3).

The classification and nomenclature of the dolphin TRBV gene subgroups were established first by the IMGT/V-QUEST tool [33,34], comparing each V region sequence with the germline human, pig and sheep TRBV sequence set from the IMGT reference directory. In the Cetartiodactyla superorder the gene nomenclature based on IMGT standardized rules [20] has been previously well defined [21-23,28].

In Figure 1B the TRBV potential functional genes and in-frame pseudogenes are shown. All sequences exhibit the typical framework regions (FR-IMGT) and complementarity determining regions (CDR-IMGT) as well as four amino acids: cysteine 23 (1st-CYS) in FR1-IMGT, tryptophan 41 (CONSERVED-TRP) in FR2-IMGT, hydrophobic amino acid 89, and cysteine 104 (2nd-CYS) in FR3-IMGT (except for the TRBV6 pseudogene) [47]. Conversely, CDR-IMGT varies in amino acid composition and length, indicated by numbers between brackets, separated by dots.

The phylogenetic analysis was performed comparing all dolphin TRBV functional genes, ORF and pseudogenes with the available corresponding pig, dromedary and goat genes whose sequences were chosen by adopting two selection criteria: (1) only potential functional genes and in-frame pseudogenes were included; and (2) only one gene per each of the subgroups was selected.

The V-REGION nucleotide sequences of all selected TRBV genes were combined in the same alignment, and an unrooted phylogenetic tree was made using the NJ method [31] The tree shows that each of the 20 dolphin subgroups form a monophyletic group, when present, with corresponding pig, dromedary and goat genes, this being consistent with the occurrence of distinct subgroups prior to the divergence of the mammalian species. The phylogenetic clustering confirmed the classification of the dolphin TRBV genes derived from the sequence analysis. Therefore, each dolphin TRBV subgroup was classified as orthologous to the corresponding artiodactyl subgroups (Figure 3).
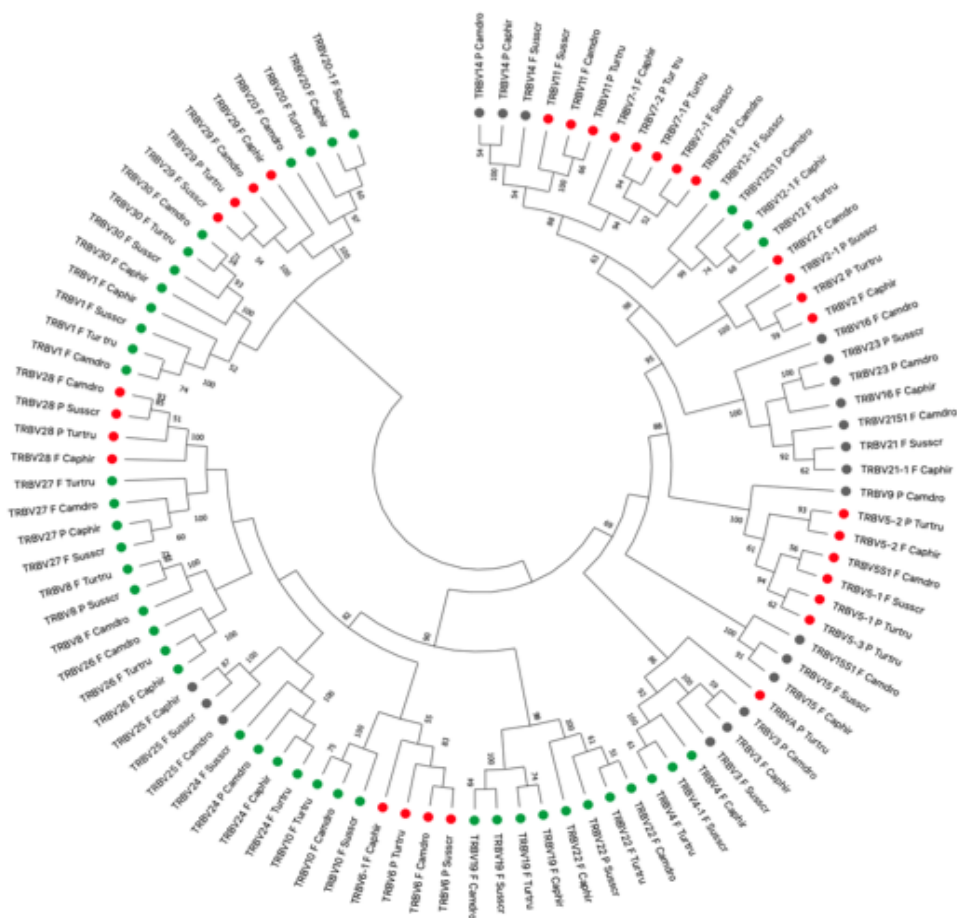
Figure 3. The NJ tree inferred from the dolphin (*Tursiops truncatus*), goat (*Capra hircus*), pig (*Sus scrofa*) and dromedary (*Camelus dromedarius*) TRBV gene sequences. The evolutionary analysis was conducted in MEGAX [30]. The optimal tree with the sum of branch length = 16.02748379 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) is shown next to the branches [51]. The tree is drawn to scale with branch lengths in the same units as those of the evolutionary distances used to infer phylogenetic trees. The evolutionary distances were computed using the Maximum Composite Likelihood method [52]and are in the units of the number of base substitutions per site.   This analysis involved 95 nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). There were a total of 473 positions in the final dataset.   The dolphin TRBV subgroup classification is performed according to the clustering with the orthologous mammalian TRBV subgroups.   The different colors highlight the distribution of the phylogenetic groups (green balls for dolphin functional genes; red balls for dolphin pseudogenes; dark balls for missing genes in dolphin). The gene functionality according to IMGT rules (F: functional, ORF: open reading frame, P: pseudogene) is indicated. The IMGT 6-letter for species (Turtru, Camdro, Susscr and Caphir) standardized abbreviation for taxon is used.

As summarized in Table 1, the total number of the dolphin TRBV genes is considerably lower if compared with the artiodactyl species, with humans and carnivores too, mainly due to the lack of several TRBV subgroups. Particularly, eight (TRBV3, TRBV9, TRBV14, TRBV15, TRBV16, TRBV21, TRBV23 and TRBV25) subgroups were not found in dolphin (dark balls in Figure 3). TRBV13 and ORF TRBV17 genes, present in human [8], are missing in cetartiodactyls (dolphin, goat, pig and dromedary) and carnivores (dog) [53] genomes. The lack of the dolphin TRBV18 is shared with pig and dromedary; pig TRB locus lacks TRBV16 and TRBV26, while goat and dog loci lack TRBV23 subgroups. TRBV9 subgroup was not found in pig and dog genomes whereas the TRBV14 is absent from the dog locus (Table 1).

Table 1. Characterization of the TRBV subgroups (number of genes, functionality and CDR-IMGT lengths) in dolphin (*Tursiops truncatus*), goat (*Capra hircus*), pig (*Sus scrofa*), camel (*Camelus dromedarius*), and dog (*Canis lupus familiaris*).

TRBV Subgroup	*Tursiops truncatus*	CDR-IMGT lengths	*Capra hircus*	CDR-IMGT lengths	*Sus scrofa*	CDR-IMGT lengths	*Camelus dromedarius*	CDR-IMGT lengths	*Canis lupus familiaris*	CDR-IMGT lengths
TRBV1	1F	6.6.4	1F	6.6.2	1F	6.6.4	1F	6.6.2	1F	6.6.4
TRBV2	1P	–	1F	5.6.4	1F,4P	5.6.4	1F	5.6.4	3P	–
TRBV3	–	–	1F	5.6.3	1F	5.6.4	1P	5.6.4	2F,1P	5.6.4
TRBV4	1F	5.6.4	1F	5.6.4	4F,1P	5.6.4	–	–	4F	5.6.4
TRBV5	3P	5.6.5	18F,12P	5.6.4	2F,1P	5.6.4	3F	5.6.4	2F,2P	5.6.4
TRBV6	1P	–	11F,14P 4O	5.6.4 5.6.3	1P	5.6.4	1F	5.6.4	1P	–
TRBV7	2P	–	2F	5.6.4	2F	5.6.4	2F	5.6.4	1F	5.6.4
TRBV8	1O	5.6.4	1P	–	1P	–	1F	5.6.4	1P	–
TRBV9	–	–	1P	–	–	–	1P	–	–	–
TRBV10	1F	5.6.4	1P	–	1F	5.6.4	1F	5.6.4	1F	5.6.4
TRBV11	1P	–	1P	–	1F	5.6.4	1F	5.6.4	1P	–
TRBV12	1F	5.6.3	2F	5.6.4	1F,1P	5.6.4	2P	–	1F,1P	5.6.4
TRBV13	–	–	–	–	–	–	–	–	–	–
TRBV14	–	–	1P	5.6.4	1F	5.6.4	1P	–	–	–
TRBV15	–	–	1F	5.6.4	1F	5.6.4	2F	5.6.4	1O	5.6.4
TRBV16	–	–	1F	5.6.4	–	–	1F	5.6.4	1F	5.6.4
TRBV17	–	–	–	–	–	–	–	–	–	–
TRBV18	–	–	1P	–	–	–	–	–	1F	5.6.4
TRBV19	1F	–	1F	5.6.4	1F	5.6.4	1F	5.6.4	1P	5.6.4
TRBV20	1F	6.7.3	1F	6.7.3	3F	6.7.3	1F	6.7.3	1F	6.7.3
TRBV21	–	–	6F	5.6.4 5.5.3	1F	5.6.4	2F,1P	5.6.4	1P	5.6.4
TRBV22	1F	5.6.3	1F	5.6.3	1P	–	1F	5.6.4	1F	5.6.4
TRBV23	–	–	–	–	1P	–	1P	5.6.4	–	–
TRBV24	1F	5.6.4	1F	5.6.4	1F	5.6.5	1P	–	1F	5.6.5
TRBV25	–	–	1F	5.6.4	1F	5.6.4	1F	5.6.4	1F	5.6.4
TRBV26	1F	5.6.4	1F	5.6.4	–	–	1F	5.6.4	1F	5.6.4
TRBV27	1F	5.6.4	1P	5.6.4	1F	5.6.4	1F	5.6.4	1P	5.6.4
TRBV28	1P	5.6.4	1F	5.6.4	1F	5.6.4	1F	5.6.4	1F	5.6.4
TRBV29	1P	5.6.4	1F	5.6.3	1F	5.7.3	1F	5.7.2	1F	5.7.3
TRBV30	1F	6.5.3	1F	6.5.3	1F	6.5.3	1F	6.5.3	1F	6.5.3
TRBVA	1P	–	–	–	–	–	–	–	–	–
Total per Fct	11F+11P + 1O		54F+33P +4O		27F+11P		25F+8P		22F+1O+13P	
Total genes	23		91		38		33		36	

3.4. Genomic structure and evolution of the dolphin TRBV locus

Genomic comparison of the dolphin TRB locus with that of the human, dromedarius, pig and goat, highlighted the uniqueness of the structure of the dolphin TRBV locus. Twenty-three dolphin TRBV genes, grouped into 20 distinct subgroups, lie in a region of approximately 145 kb (Figure 1A). The dot-plot matrix of the dolphin versus human TRB locus confirms the high level of nucleotide identity between TRBV genes as indicated by dots and diagonal lines that correspond with gene location (Figure 4).
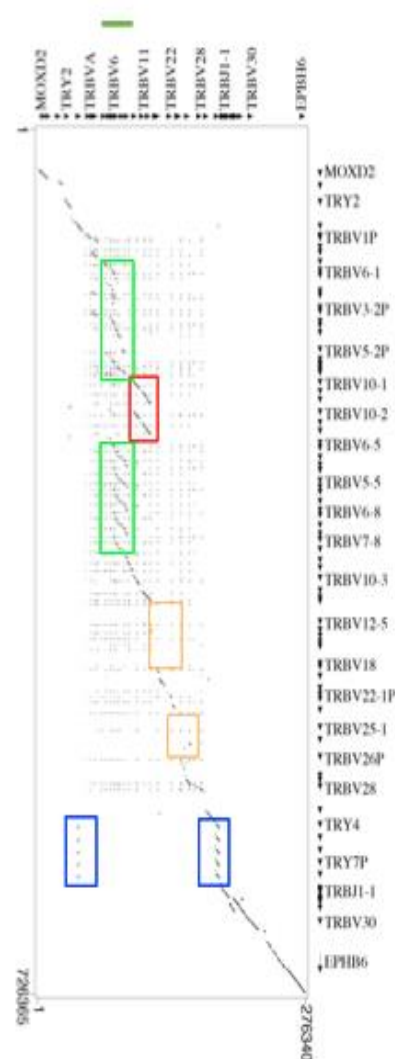
Figure 4. Dot-plot matrix of dolphin/human TRB sequence genomic comparison. The transcriptional orientation of each gene is indicated by arrow heads. Colored rectangles (light green, red and blue) enclose TRBV duplicated regions in human and (orange) deleted TRBV regions in dolphin as referred in the text. The green line indicates the only region where gene duplication took place in *T. truncatus* TRB locus.

The only region where gene duplication took place in *T. truncatus* TRB locus concerns the TRBV5 and TRBV7 genes, that are inserted in a group consisting of duplications of individual TRBV5 (from 5-1 to 5-3), and TRBV7 (from 7-1 to 7-2) pseudogenes, and are interspersed with the TRBV6 pseudogene and with the TRBV8 ORF (Figure 1A and green line in Figure 4). This group of six pseudogenes and a single ORF is found in the matrix showing the dot-plot of dolphin against the dromedarius TRBV cluster (green line in Supplementary Figure S1) and against the pig TRBV cluster sequences (green line in Supplementary Figure S2). This group of genes whose sequential order on the genome is conserved in cetaceans, swine and Tylopoda, occupies about 15 kb of the TRB locus in the dolphin genome.

On the contrary, the corresponding group in human went through duplications up to eight copies for TRBV5 (TRBV5-8) and up to nine copies for TRBV6 and TRBV7 (TRBV6-9, TRBV7-9). It should be noted that the latter are all functional genes, with the exception of a single pseudogene (TRBV7-5). Such duplications occupy a region extending for 200 kb (green rectangles in Figure 3) which is interrupted by a portion of 50 kb (red rectangle in Figure 4) containing the only copy of TRBV9 gene and the duplication of the group consisting of TRBV10, TRBV11 and TRBV12 genes (Figure 4) In *Capra hircus* TRB locus [28], in correspondence of the dolphin TRBV5-1, TRBV5-2 and

TRBV6 genes, is housed a series of repeated duplicative events that lead to the existence of 30 copies for the TRBV5 and 29 for TRBV6, for a total of 59 (29 functional, 26 pseudogenes and 4 ORF) genes. This region seat of repeated duplicative events extends without interruption for 180 Kb (red rectangle in Supplementary Figure S3).

Furthermore, in the human-dolphin genome comparative analysis, the region including seven genes (from 12-4 to 18) in humans is deleted in the dolphin (orange rectangle in Figure 4) where TRBV12 and TRBV19 genes are neighbors in the same region (Figure 1A). If compared to the other cetartiodactyls, this deletion (orange rectangle) in dromedary (Supplementary Figure S1), in pig (Supplementary Figure S2) and in goat (Supplementary Figure S3) matrices, is unique and typical of the dolphin TRB locus. The deleted region marked with an orange square in Figure 4 is due to the presence of the interspersed deletions of TRBV23 and TRBV25 dolphin genes in comparison with the human contig. Finally, the dotplot matrix in Figure 4 shows five lines of similarity indicating the duplications in human of TRY4 (from TRY4 to TRY8) (blue rectangles) and five dots in correspondence of the only functional TRY gene (TRY3) (blue rectangles) present in dolphin (Supplementary Table S2).

3.5. 5' RACE PCR assay

To evaluate the functional competency of the three D-J-C clusters, we performed three 5' RACE experiments on total RNA isolated from the peripheral blood of two unrelated animals, a male (M) and a female (L), using a TRBC-specific primer. Each RACE product was gel-purified and cloned into the TA-vector and randomly selected positive clones for each cloning were sequenced [9]. A total of 45 diverse clones of different lengths containing rearranged V-(D)-J-C transcripts with a correct open reading frame were obtained. In particular, 22 out of 45 cDNA clones belonged to the male subject (M5RBL series), while the remaining 23 cDNAs derived from the female subject (L5RBL series) (Figure 5). Five clones of the L5RBL series (L5RBL12, L5RBL13, L5RBL20, L5RBL2 and L5RBL27), and three clones of the M5RBL series (M5RBL15, M5RBL22 and M5RBL30) showed an identical coding sequence for a total of 13 redundant clones.

Figure 5. Protein display of the cDNA clones. The TRBV and TRBJ genes, named according to the criteria specified in the text, are listed, respectively, at the left and the right of the figure. Leader region (L-REGION), complementarity determining regions (CDR-IMGT) and framework regions (FR-IMGT) are also indicated according to the IMGT unique numbering for the V-DOMAIN [47]. The five conserved amino acids of the V-DOMAIN (1st-CYS 23, CONSERVED-TRP 41, hydrophobic AA 89, 2nd-CYS 104 and J-PHE 118) are colored (IMGT color menu). The TRBV allele amino acid

changes, if any, are boxed. The name of the clones is also reported.

Each of the remaining 32 cDNA sequences was manually analyzed to identify the TRBV, TRBD and TRBJ genes through alignment with the germline dolphin TR genes. Figure 5 shows the deduced amino acid sequences of the V-(D)-J region of all cDNA clones according to IMGT unique numbering for the V-REGION and V-DOMAIN [47]. In the cDNA clones, we identified 7 different TRBV genes belonging to 7 out of 12 different subgroups consisting of predicted functional germline genes. In 16 out of 32 clones, the TRBV gene perfectly matched the corresponding germline sequence. The remaining 16 TRBV sequences showed a nucleotide identity from 97 to 99% with respect to the reference germline gene sequence. We referred to these as new alleles, although we cannot rule out sequencing errors or that duplicated genes have been missed. The TRBV30 subgroup gene represented the most expressed (17 unique different clones, 53%) followed by the TRBV12 (7 clones) and the TRBV4 (4 clones) genes. Only one clone was found for four (TRBV10, TRBV20, TRBV22 and TRBV27) subgroups. Moreover, the functionality of the dolphin TRBV12 gene was confirmed since originated seven productive transcripts despite the presence of a stop codon in the germline CDR3-IMGT at position 108 (Figure 1B) [47] (IMGT Repertoire, Alignments of alleles, http://ww.imgt.org), that was trimmed during rearrangement (Figures 5, 6). In both subjects analyzed, of the 12 functional genes present in the locus (Figure 1A, B) and described in Table 1, five of them (TRBV1, TRBV8, TRBV19, TRBV24 and TRBV26) were not found expressed in peripheral blood (Figure 5).

3.6. Analysis of the V-D-J rearrangements

A certain difference in the two analyzed individuals was observed in the usage of the TRBV genes in the recombinative events. In fact, all TRBV subgroups were found in the cDNA from male subject (M5RBL series) while only 3 out 7 TRBV subgroups were identified in the female (L5RBL series) cDNA. It is interesting to note that the TRBV30 gene is almost equally recurrent (9 out of 23 L5RBL series and 8 out of 22 M5RBL series) in the transcripts of the two animals, suggesting that this gene may be involved for the expression of a public TRB repertoire (allele*02 shared in both individuals).

All TRBJ genes within the set of cDNA sequences were assigned unambiguously to the distinct germline genes derived from the reference sequence and only in two cases (TRBJ2-7 and TRBJ3-1) we identified new alleles, that were labeled by consecutive numbers with respect to the germline allele *01 (Figure 5). The TRBJ2 set genes with a frequency of 20/32 clones appear to be preferentially expressed over the TRBJ3 set genes with a frequency of 11/32 clones. The surprising result concerns the almost absence of the use of the TRBJ1-1 gene with a single (L5RBL4) clone that contains it (Figures 5 and 6).

clone	TRBV	CDR3	TRBD	TRBJ
L5RBL1	TRBV4*01	GCCAGCAGCAggat**ACAGGG**ttccACCCAGCCCCTGTAC A S S R I **Q** **G** S T Q P L Y	TRBD1	TRBJ3-5*02
L5RBL2	TRBV4*01	GCCAGCAGCActgggtactt**GGGGC**gaagggccACTCAGCCCCTGTAC A S S T G Y L **G** R R A T Q P L Y	TRBD1	TRBJ3-5*01
M5RBL3	TRBV4*01	GCCAGCAGCgcttcccgc**GGACAGGGGGG**gCAGCCCCTGTAC A S S A S R **G** **Q** **G** **G** Q P L Y	TRBD1	TRBJ3-5*01/*02
L5RBL25	TRBV4*02	GCCAGCAGCaaagctcaa**GACAG**ccgcAACGACCCTCTGTTC A S S K A Q D S R N D P L F	TRBD1	TRBJ2-3*01
M5RBL11	TRBV10*02	GCCAGCAGTGAgcgggagtcgaggGAGCAGTAT A S S E R E S R E Q Y	–	TRBJ2-7*01
L5RBL4	TRBV12*03	GCCAGCAGAcccgacctcgccgcggtcAACACTGACGTCTTC A S R P D L A A V N T D V F	–	TRBJ1-1*01
M5RBL5	TRBV12*02	GCCAGC**GGACAGG**cggcccgtCTGTTC A S **G** **Q** A A R L F	TRBD1	TRBJ2-3*01
M5RBL6	TRBV12*01	GCCAGCAGgctttctggaagCTCCTATGAGCAGTAT A S R L S G S S Y E Q Y	–	TRBJ2-7*01
M5RBL26	TRBV12*01	GCCAGCAGgggg**ACAGGGGGG**ACCGAGCTGTAT A S R G **T** **G** **G** T E L Y	TRBD1	TRBJ2-2*01
M5RBL7	TRBV12*01	GCCAGCgattctgaagccagaggtggccctACTCAGCCCCTGTAC A S D S E A R G G P T Q P L Y	–	TRBJ3-5*01
M5RBL8	TRBV12*01	GCCAGCAGtagggcatc**GGGGGC**gAGGTCTGATcccccggtt A S S R A S **G** A R S D P P V	TRBD1	TRBJ2-5*01
M5RBL9	TRBV12*01	GCCcggagtgggctgtctTCTCCCCCGGTT A R S G L S S P P V	–	TRBJ2-5*01
M5RBL10	TRBV20*02	GTGTCAGAGtcctatctg**CAGGGGG**aGAGCAGTAT G V R V L S A **G** **G** E Q Y	TRBD1	TRBJ2-7*01
M5RBL31	TRBV22*02	TCCGGTAGCGcc**AGGAC**tggggagagcGACCCTCTGTTC S G S A **R** **T** G E S D P L F	TRBD1	TRBJ2-3*01
M5RBL32	TRBV27*02	GCCAGCAGTTTggga**TCGGGGGG**cctCAACGACCCTCTGTTC A S S L G **S** **G** **G** L N D P L F	TRBD2	TRBJ2-3*01
L5RBL12	TRBV30*01	GCCTGGAATCTt**GGGACTTCGGGGGGG**cccttaAATGAGCTGCAC A W N L **G** **T** **S** **G** **G** P L N E L H	TRBD2	TRBJ2-1*01
L5RBL27	TRBV30*01	GCCtcccctaa**TTCGGGGG**agACCGAGCTGTAT A S P N **S** **G** E T E L Y	TRBD2	TRBJ2-2*01
L5RBL13	TRBV30*01	GCCTGGgga**TCGGGGG**tggcgaACCCTCTGTTC A W G **S** **G** V A N P L F	TRBD2	TRBJ2-3*02
L5RBL14	TRBV30*01	GCCTGGagtcccgat**GGGGG**aGACCCTCTGTTC A W S P D **G** **G** D P L F	TRBD1/D2	TRBJ3-3*02
M5RBL15	TRBV30*02	GCCTGGtcccgaagccggggaggaaagactAGCCAAAACACTCAGTAC A W S R S R G G K T S Q N T Q Y	–	TRBJ3-3*01
L5RBL16	TRBV30*01	GCCTGGgtcaacgggta**CAGGGGGG**ccgAGCCCCTGTAC A W V N G Y **R** **G** A E P L Y	TRBD1	TRBJ3-5*01
L5RBL17	TRBV30*01	GCCggagaac**TTCGGGGGG**acgAGTCTCCCCCGGTT A G E L **R** **G** D E S P P V	TRBD2	TRBJ2-5*01
L5RBL18	TRBV30*01	GCCTGGAATctagt**AGGGG**CTGACACTCAGTAC A W N L V **G** A D T Q Y	TRBD1	TRBJ3-4*01
M5RBL19	TRBV30*02	GCCTGGcgggg**CAGGGGG**acgcatagtACTCAGCCCCTGTAC A W R G **R** **G** T H S T Q P L Y	TRBD1	TRBJ3-5*01
L5RBL20	TRBV30*02	GCCTGGAATCTatcgggttcagcggggtctttgCAGCCCCTGTAC A W N L S G S A G S L Q P L Y	–	TRBJ3-5*01/*02
L5RBL21	TRBV30*01	GCCTGGgttttcg**GGGGGGC**cgggGAGCAGTAT A W V F **G** **G** A G E Q Y	TRBD1	TRBJ2-7*01
M5RBL22	TRBV30*02	GCCTGGgcacgagtgctaaACAGTGAGCAGTAC A W A R V L N S E Q Y	–	TRBJ3-1*02
M5RBL28	TRBV30*02	GCCTGGtc**CAGGG**tgaTGAACACCGAGCTGTAT A W S **R** V M N T E L Y	TRBD1	TRBJ2-2*01
M5RBL29	TRBV30*02	GCCTGGAAgaa**AGGGG**cgggaACCGAGCTGTAT A W K K **G** A G T E L Y	TRBD1	TRBJ2-2*01
M5RBL30	TRBV30*02	GCCTGGAAgccaggacgggcgAACACCGAGCTGTAT A W K P G R A N T E L Y	–	TRBJ2-2*01
M5RBL23	TRBV30*02	GCCggggccctgaaggcatg**GGACAGGGG**acCTCAGTAC A G A L K A W **D** **R** **G** P Q Y	TRBD1	TRBJ3-4/J3-3
M5RBL24	TRBV30*02	GCCttggatCAGTAT A L D Q Y	–	TRBJ2-7*01

Figure 6. CDR3-IMGT nucleotide and predicted amino acid sequences retrieved from the TRB cDNA clones. CDR3-IMGT sequences are shown from codon 105 (the codon after the 2nd-CYS 104 of the V-REGION) to codon 117 (the codon before J-PHE 118 of the J-REGION) according to the unique numbering [47]. The CDR3-IMGT nucleotide/amino acid sequence, and the classification of the TRBV and TRBJ genes of each clone are also listed. Nucleotides of the 3'V-REGION and of the 5'J-REGION are indicated in uppercase letters. The sequences considered to present recognizable coding regions of the TRBD genes are indicated in red uppercase letters. The amino acids belonging to a TRBD region are in bold. Nucleotides that cannot be attributed to any V, D or J region (N-nucleotides), are indicated in lower cases on the left and on the right sides of the TRBD regions. The name of each clone is also reported. Colored clones are referred in the text.

More complex is the determination of the contribution of the TRBD genes involved in the CDR3 formation. For a close inspection, the nucleotide sequences corresponding to the CDR3 have been excised from each cDNA and analyzed in detail (Figure 6). By comparison with the TRBD genomic sequences, the nucleotides located in the CDR3 were considered belonging to a TRBD gene if they constituted a stretch of at least five consecutive nucleotides. We were able to assess the contribution of only the TRBD1 and TRBD2 germline sequences but was not possible to identify the presence of a third TRBD gene because this is missing in the current genomic assembly. In this way, the TRBD was unambiguously identified in 21 out of 32 sequences (62.5%) with the TRBD1 present in 16 and TRBD2 in 5 clones, respectively. The remaining 11 sequences either do not have an identifiable TRBD gene maybe for the lack of the TRBD3 reference sequence, or it was not possible to distinguish between TRBD1 and TRBD2 genes. However, the absence of a TRBD region could be also interpreted as direct V-J junction and it is even possible that nucleotide trimming masked the initial participation of the TRBD gene during the rearrangement. The corresponding amino acid sequence of the CDR3 loop deduced from the nucleotide sequences is heterogeneous for amino acid composition and length (Figures 5 and 6). The mean length of the CDR3 loop was approximately the same in the dolphin (mean 12,3 amino acids in a range of 5-16 amino acids) with respect to the human (mean 12,7 amino acids; [54]), pig (mean 12,2 amino acids; [21]), camel (mean 12,8 amino acids; [22]) and goat (mean 12,3 amino acids; [28]).

Finally, although the TRBD gene was not unambiguously recognizable in all of the cDNA clones, the interpretation of these rearrangements revealed that only five clones show the intracluster rearrangements with TRBD2-TRBJ2 (yellow in Figure 6). Intercluster rearrangements represent the substantial portion of the repertoire, with 9 TRBD1–TRBJ2 (light blue in Figure 6) and 7 TRBD1–TRBJ3 (green in Figure 6) rearrangements.

3.7. 3D comparative modeling of the T.truncatus TR alpha/beta chain in complex with MH1 I-ALPHA and B2M

To calculate the most likely interactions between the putative alpha/beta pairing in complex with MH1 I-ALPHA and B2M, we analysed the amino acid sequences of the two types of rearranged TRA cDNA, originated by TRAV20S2*01–J17 (5RAL27) and TRAV18-1*02–J20 (5RAL11) rearrangements, respectively found in the peripheral blood of the animal identified with the letter L in our previous study [9], and of relevant TRB cDNA, originated by TRBV30–D2–J3 (L5RBL13) and TRBV30–D2–J2 (L5RBL27) rearrangements among the nine found in this study in the peripheral blood of the same animal (TRBV30 clones initialed by the initial letter L in Figures 5 and 6).

The 3hg1.pdb crystallized structure consisting of the human TR alpha/beta chains in complex with MH1 I-ALPHA and B2M was identified along pGenThreader and I-tasser searches and used as a protein template for guiding the 3D comparative modeling analysis (Figure 7). *T.truncatus* TR alpha and TR beta chains were modelled by using as a protein template 3hg1.pdb chain D, the human TR alpha, sharing with the modelled *T.truncatus* TR alpha putatively coded by 5RAL27 or 5RAL11 clonotypes, 46 or 41 % of identical amino acids, respectively, or 3hg1.pdb chain E, the human TR beta, sharing with *T.truncatus* TR beta, putatively coded by L5RBL13 or L5RBL27 clonotypes, 71-59 % of identical residues, respectively, and 100% coverage. All *T. truncatus* TRA and TRB sequences showed 100 coverage with the corresponding human TR alpha or TR beta crystallized chains.

The pairwise sequence structure alignment between the investigated *T.truncatus* TR alpha and TR beta and the human crystallized counterpart (from 3hg1.pdb) was obtained by using ClustalW [55] implemented in the Jalview package [56]. The obtained sequence-structure pairwise alignment (Supplementary Figure S4 A and B) was reported in SPDBV alignment panel for guiding/building the 3D comparative model of both TRA and TRB chains, according to validated protocols [37-41].

For studing interactions between *T.truncatus* TR alpha and TR beta with MH1 I-ALPHA in *T.truncatus* we searched for *T.truncatus* MH1 I-ALPHA and we identified the sequence XP_033706382.1 sharing with the human MH1 I-ALPHA (3hg1.pdb, chain A) sequence more than 79% of identical amino acids and 100% coverage. Similarly, we searched for *T.truncatus* best hit B2M counterpart and the sequence XP_033707513.1 was identified, sharing with the human B2M sequence more than 74% of identical amino acids and 100% coverage. A 3D comparative model of the *T.truncatus* MH1 I-ALPHA and B2M was built by using as a protein template the corresponding counterparts taken from 3hg1.pdb starting from the pairwise sequence structure alignment built by CulstalW, as proposed for TR alpha and TR beta (Figure 7 and Supplementary Figure S4 C and D). The root mean square deviation (RMSD) between the coordinates of the built 3D comparative models and the crystallized 3hg1.pdb ranged between 0.15 and 0.5 Å.
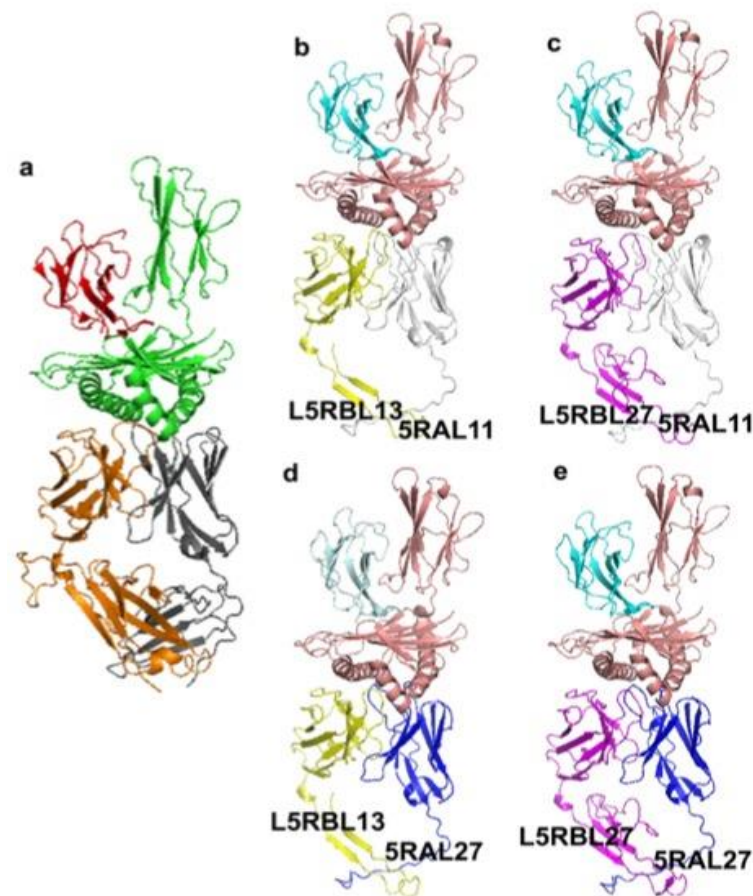
Figure 7. *T. truncatus* TR alpha and TR beta in complex with MH1 I-ALPHA and B2M. **Panel a**. Human TR alpha and TR beta are reported in grey (TR alpha) and orange (TR beta) cartoon, respectively, in complex with MH1 I-ALPHA (green cartoon) and B2M (red cartoon). **Panel b-e**. Different complex poses showing the investigated combination of *T. truncatus* TR alpha (5RAL11 clonotype, white cartoon; 5RAL27 clonotype, blue cartoon) and TR beta (L5RBL13 clonotype, yellow cartoon; L5RBL27 clonotype, blue cartoon) in complex with MH1 I-ALPHA (salmon cartoon) and of B2M (cyan cartoon).

3.8. Binding energy calculations at the protein interface between TR alpha and TR beta chains: computational analyses predict the pairing of the TRAV20-J17 (or TRAV18-J20) and TRBV30-D2-J3-3 variable domains

The interaction energies calculated between the TR alpha and TR beta and between the TR alpha/beta and MH1 I-ALPHA resulted in a negative value (Table 2), confirming that there might be a binding interaction in all the investigated cases. This result is

encouraging, also due to the strategy validated by obtaining negative binding energies for the interactions between the human TR alpha and TR beta chains and the MH1 I-ALPHA chain within the crystallized complex available under the PDB code "3hg1.pdb".

Table 2. Interaction energies estimated at the protein interface between TR alpha and TR beta chains and between the reported TR chains and MH1 I ALPHA.

Pdb	3hg1.pdb		L5RBL13_5RAL11		L5RBL13_5RAL27		L5RBL27_5RAL11		L5RBL27_5RAL27	
Group1	A	AB	A	AB	A	AB	A	AB	A	AB
Group2	B	C	B	C	B	C	B	C	B	C
IntraclashesGroup 1	8,81	21,96	10,74	19,12	10,24	18,07	10,51	19,99	8,14	19,16
IntraclashesGroup 2	12,39	10,93	7,48	13,04	7,24	13,43	8,66	13,22	10,63	15,06
Interaction Energy	-54,67	-5,58	-21,02	-7,23	-21,91	-5,73	-17,32	-8,35	-17,75	-9,26
Backbone Hbond	-10,05	-0,44	-3,95	-0,36	-3,41	-1,94	-3,96	-0,91	-2,91	-2,59
Sidechain Hbond	-15,85	-2,94	-9,53	-5,44	-6,10	-4,44	-9,00	-7,16	-6,20	-11,53
Van der Waals	-40,11	-11,63	-22,14	-10,44	-19,44	-11,84	-21,66	-11,09	-19,93	-14,02
Electrostatics	-6,23	-1,62	-2,43	-4,48	-1,59	-2,17	-3,25	-4,42	-2,07	-2,14
Solvation Polar	48,22	16,23	26,43	15,44	22,71	18,22	26,39	17,93	24,38	21,15
Solvation Hydrophobic	-53,35	-14,63	-30,35	-11,58	-27,13	-13,32	-29,88	-11,64	-27,49	-14,33
Van der Waals clashes	0,63	0,93	0,10	0,80	0,48	1,24	0,24	0,86	0,37	0,56
entropy sidechain	15,44	7,15	13,18	7,45	7,76	6,92	15,16	8,16	10,23	11,72
entropy mainchain	10,62	1,39	7,84	1,95	5,31	1,97	8,44	0,62	6,74	2,44
torsional clash	0,14	0,05	0,80	0,06	0,10	0,13	0,58	0,06	0,01	0,13
backbone clash	2,49	2,71	1,35	1,33	0,65	3,45	0,98	1,57	0,72	2,86
helix dipole	0,01	-0,01	0,00	0,05	0,00	-0,04	0,00	0,02	0,00	-0,02
electrostatic kon	-1,19	-0,05	-0,97	-0,68	-0,60	-0,47	-1,51	-0,78	-0,87	-0,63
Entropy Complex	2,38	2,38	2,38	2,38	2,38	2,38	2,38	2,38	2,38	2,38
Number of Residues	814,00	814,00	657,00	657,00	655,00	655,00	695,00	695,00	693,00	693,00
Interface Residues	108,00	44,00	57,00	42,00	52,00	43,00	64,00	40,00	56,00	46,00

Notably, the strongest binding interactions between the investigated *T. truncatus* TR alpha and TR beta 3D models is observed in the L5RBL13_5RAL27 clonotypes containing protein complex and in the L5RBL13_5RAL11 clonotypes containing protein complex (in terms of interaction energies calculated by FoldX Analyse complex assay) (Table 2, Figure 8 A and B), whereas the strongest interaction between TR alpha/beta and MH1 I-ALPHA is observed in the L5RBL27_5RAL27 clonotypes containing protein complex (Table 2).

While interaction energies estimated at the protein interface within the modelled *T. truncatus* TR alpha and TR beta appear weaker than their interacting counterparts in the crystallized structure (3hg1.pdb), interaction energies calculated at the protein interface between the modelled *T. truncatus* TR alpha/beta and MH1 I-ALPHA appear stronger than their counterpart in the human 3hg1.pdb in all the cases (Table 2).

Interacting amino acids located in the CDR-IMGT of the variable alpha (V-ALPHA) and beta (V-BETA) domains of the TRA/TRB chains are shown in Figure 8 A and B. It is interesting to note the recurring position in the DE loop of the FR3-IMGT regarding the L5RBL13 clonotype (red arrow indicating E84 in Figure 8 B), the 5RAL27 clonotype (red arrow indicating K 85 in Figure 8 A) and the 5RAL11 clonotype (red arrow indicating T 82 in Figure 8 B).
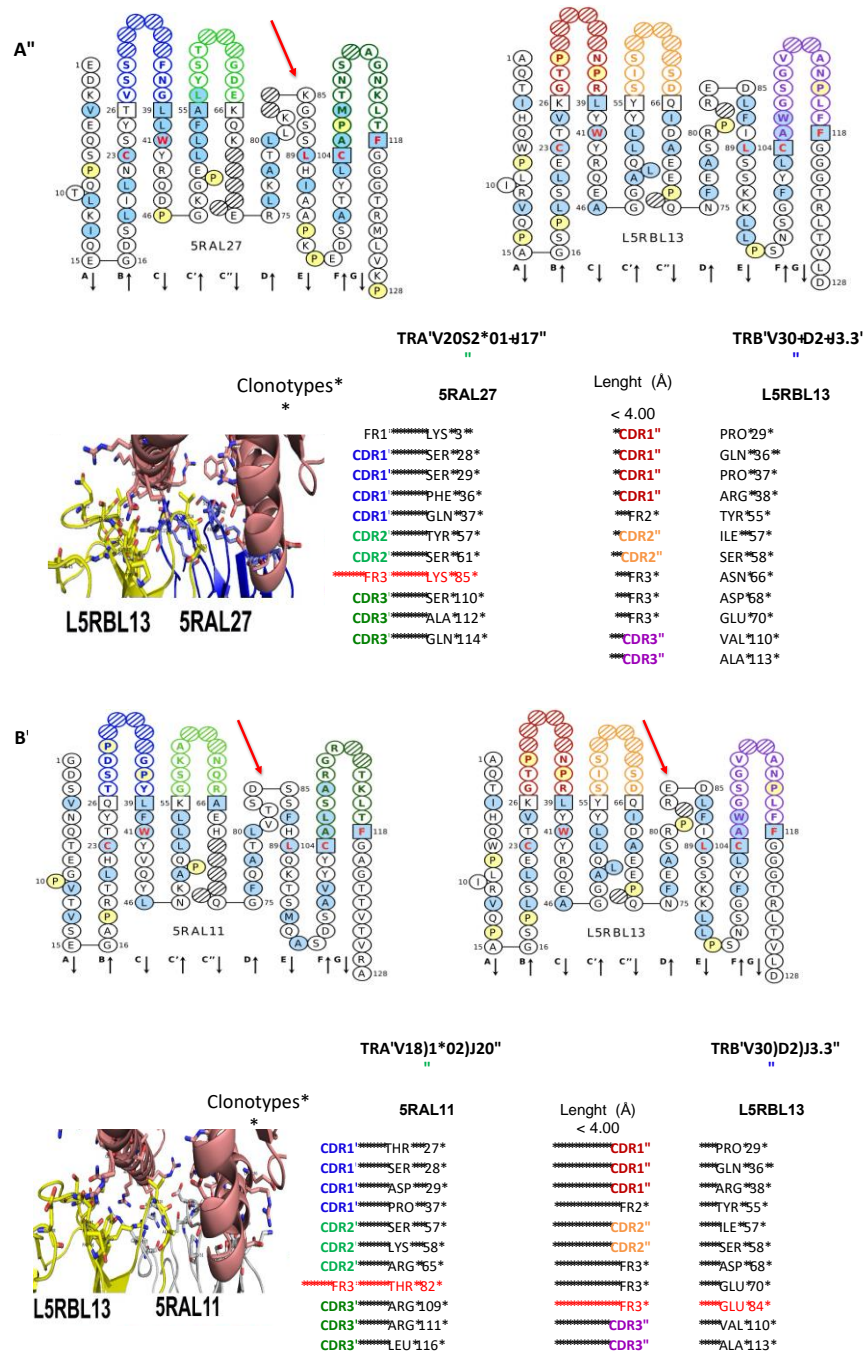
Figure 8. Computationally inferred interaction between 5RAL27 V-alpha domain (TRAV20S2*01-J17) and 5RBL13 V-beta domain (TRB V30-D2-J3.3). (**A**) and between 5RAL11 V-alpha domain (TRA V18-1*02-J20) and L5RBL13 V-beta domain (TRBV30-D2-J3-3). (**B**) cDNA clonotypes. IMGT Collier de Perles of 5RAL27/5RBL13 and 5RAL11/5RBL13 clones are shown [50,57]. In 5RAL11 and 5RAL27 V-alpha domain CDR-IMGT are blue-green-green; in L5RBL13 V-beta domain CDR-IMGT are red-orange-purple. Interacting residues at the interface of the alpha/beta TR 3D models with MH1 I-ALPHA in *T. truncatus* are reported in sticks representation (lower left). The protein complex interface were computed by the pGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/) and i-Tasser (https://zhanglab.ccmb.med.umich.edu/I-TASSER/).

3. Discussion

In the superorder of Cetartiodactyla, bottlenose dolphin (*Tursiops truncatus*) and the other cetaceans represent the most successful mammalian colonization of the aquatic environment and have undergone a radical transformation from the original mammalian bodyplan. In a previous work [9] we reported in dolphin genomic and expression studies on TR gamma (TRG) and alpha/delta (TRA/TRD) loci. In this study

we report in dolphin an extensive analysis of the genomic organization, evolution and expression of the TRB genes and the 3D comparative modeling of the TR alpha/beta heterodimer in complex with MH1 I-ALPHA and B2M.

According to comparative analyses with human and Artiodactyla, i.e. dromedarius [22,23], pig [21] and goat [28], the organization of the dolphin TRB locus is similar to that of the other artiodactyl species, with three in tandem D-J-C clusters located at its 3' end. In its genomic structure, the dolphin TRB locus is the smallest in length (gaps included), comprising the MOXD2 and the EPHB6 genes that flank the 5' and 3' ends, respectively. The overall length di 276 Kb corresponds to less than half of that of *Homo sapiens* (620 Kb), to about half of that of *Capra hircus* (558 Kb) [28], to a little more than half of that of *Sus scrofa* (407 Kb) [21,58] and it is slightly smaller than that of *Camelus dromedarius* (302 Kb) [49]. The uniqueness of the dolphin TRB locus when compared to the other artiodactyls and to human is given both by the almost complete absence of duplications in the TRBV region and by the presence of deletions that drastically reduce the number of the variable genes (see orange rectangles in Figure 4).

An in-depth analysis of the genomic structure of the TRBV genes highlights a ratio of 1/1 of functional genes to pseudogenes against a ratio of about 2/1 in the other examined artiodactyls (*Capra hircus* and *Sus scrofa*) and in carnivores (*Canis lupus familiaris*). A 3/1 ratio is present in *Camelus dromedarius* (Table 1). The phylogenetic analysis shows that each of the 20 dolphin TRBV subgroups form a monophyletic group, when present, with corresponding pig, dromedary and goat genes, this being consistent with the occurrence of distinct subgroups prior to the divergence of the mammalian species. The phylogenetic clustering (Figure 3) confirmed the classification of the dolphin TRBV genes derived from the sequence analysis.

In the analysis of the TRB V-D-J rearrangements, intercluster rearrangements were observed to constitute the substantial portion of the repertoire, with 10 TRBD1–TRBJ2 and 6 TRBD1–TRBJ3. The fact that the 5'RACE assay resulted from a primer designed on a region shared by all the constant genes lead us to assume that the 19 clones containing the TRBJ2 set genes and the 12 clones containing the TRBJ3 set genes have the TRBJ gene linked to the corresponding TRBC2 and TRBC3, respectively (Figure 6). Moreover, since TRBC3 gene is located upstream TRBJ2 cluster in the germline DNA (Figure 1 A and Supplementary Figure S5), we cannot rule out that TRBJ2 can be joined to TRBC3 as a product of a trans-splicing between a transcript with TRBJ2-TRBC2 genes and a transcript containing TRBC3 gene [59].

The expression data highlight the preferential usage of TRBV30 in both the analyzed subjects. Out of a total of 45 cDNA clones obtained from the peripheral blood of the two subjects, TRBV30 is expressed with a percentage of 66.6% (30/45 clones). The redundancy of the clones (from +1 to +3) of the series containing the TRBV30, is highlighted in Figure 5. The preferential usage of the TRBV30 gene, that lies at the 3' end of the TRBC2 gene in an inverted transcriptional orientation in comparison with other artiodactyls represents a condition of diversity in dolphin. In fact, considering the total number of cDNA clones from peripheral lymphoid organs (spleen and blood), against 30/45 clones in dolphin, 2/27 in goat [28], 1/26 in pig [21], 0/35 in camel [22] and 0/36 in dog [53] are reported in literature.

The fact that both the unrelated subjects show a biased usage of the TRBV30 could be explained by the sharing by the two subjects of the same acquatic environment. Then the presence of common antigens may have stimulate T cells with a particular type of beta chain to expand, suggesting the existence of a basic "public" repertoire of a given alpha/beta TR [60]. Therefore we used the rearranged TRA [9] and TRB cDNA clones from the peripheral blood of one of the two analyzed subjects (Leah in this work), to build a 3D comparative modeling of the *T.truncatus* TR alpha/beta in complex MH1 I-ALPHA and B2M. The four possible 3D stuctures (b, c, d, e panels) of TRAV20S2*01–J17 (5RAL27) and TRAV18-1*02–J20 (5RAL11) in combination with TRBV30–D2–J3 (L5RBL13) and TRBV30–D2–J2 (L5RBL27) clonotypes (Figure 7) retain a very high

similarity with the 3hg1.pdb structure of the human TR alpha/beta chains in complex with MH1 I-ALPHA and B2M.

The negative value resulted from the interaction energies calculated between the TR alpha and TR beta and between the TR alpha/beta and MH1 I-ALPHA (Table 2), allowed the identification of the resulting interacting aminoacids located in the CDR-IMGT at a distance of length < 4,00 (Å). In particular in L5RBL13 clonotype, Tyr 55 located on the edge of CDR2 (Figure 8 A and 8 B) and in 5RAL27 clonotype, Tyr 57 on CDR2 (Figure 8 A) correlate well with recent literature [61,62]. Piepenbrink et al., in their article reported that TR and MH proteins were refolded from bacterially expressed inclusion bodies. In this paper, results of the interaction free energy between two amino acid sidechains obtained in double-mutant cycle, pointed out how the engagement of Tyr at the centre of the interface by sidechains of CDR TRA and TRB, contributes, depending on its location, significantly to TR affinity [61].

Our results show that interaction energies calculated at the protein interface between the modelled *T. truncatus* TR alpha/beta and MH1 I-ALPHA appear stronger than their counterpart in the human 3hg1.pdb in all the examined cases (Table 2).
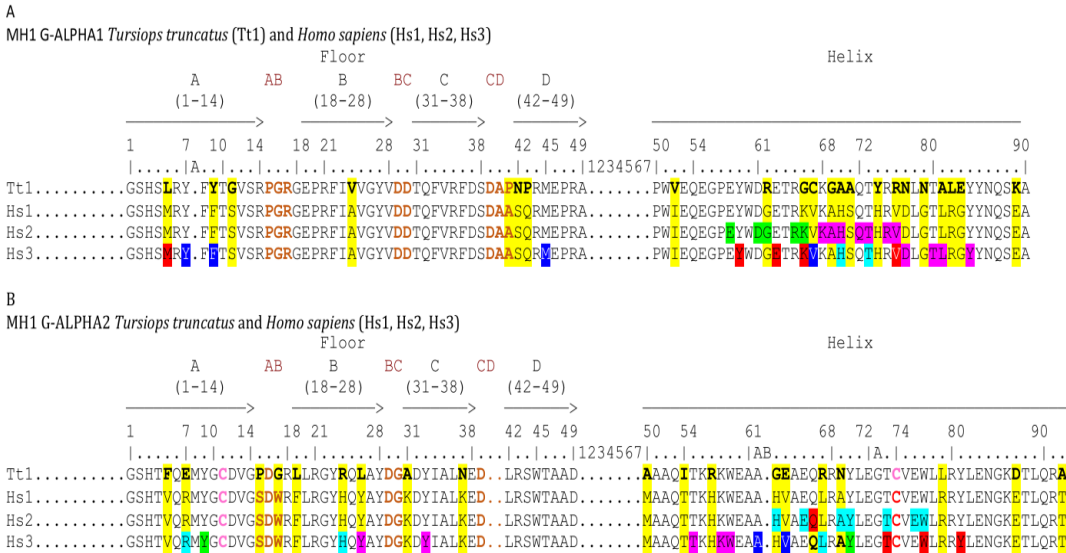
Figure 9. Alignments of the G-ALPHA1 (**A**) and G-ALPHA2 (**B**) domains of *Tursiops truncatus* MH1 I-ALPHA (acc. number: XP_033706382.1) with *Homo sapiens* MH1 I-ALPHA HLA-A*0201. Alignments are according to the IMGT unique numbering for G domain [63]. Amino acid differences between species are highlighted in yellow (and further in bold in *Tursiops truncatus* lines Tt1). The contacts between the Homsap G domains and the TR V domains are displayed in lines Hs2: in (**A**), G-ALPHA1 contacts with V-ALPHA in green and V-BETA in pink, in (**B**) G-ALPHA2 contacts with V-ALPHA in blue and V-BETA in red (arbitrary colors). The IMGT pMH contact sites [64-68] between the Homsap G domains and the peptide are displayed in lines Hs3 (colors evoking the contact sites). Positions 61A, 61B and 72A are characteristic of the G-ALPHA2 and are not reported in G-ALPHA1. Hs1, Hs2 and Hs3 are from 3hg1 in IMGT/3Dstructure-DB, http://www.imgt.org.
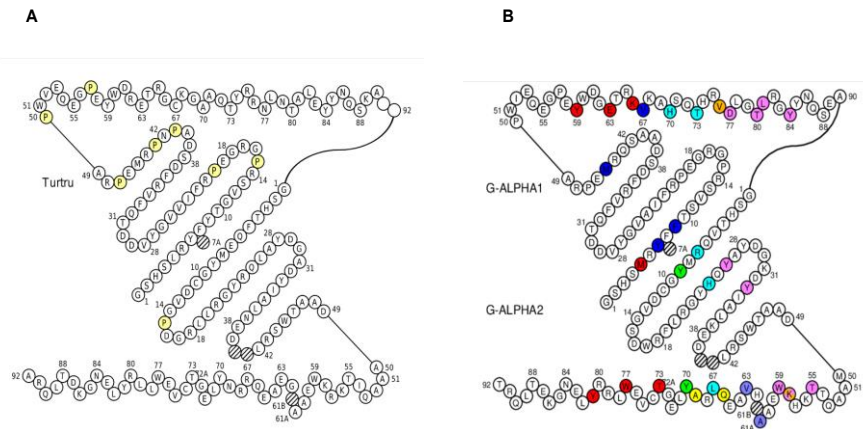
Figure 10. IMGT Collier de Perles of the G-ALPHA1 and G-ALPHA2 of the MH1 I-ALPHA. (**A**), IMGT Collier de Perles of *Tursiops Truncatus*, obtained using the IMGT/Collier-de-Perles tool [66] and translation of (acc. number: XP_033706382.1). Amino acid (AA) and gaps (hatched positions) are according to the IMGT unique numbering [63]. AA 91 and 92 are absent in G-ALPHA1 (empty circles (left) or not shown (right)). Positions 61A, 61B and 72A are characteristic of the G-ALPHA2 and are not reported in the G-ALPHA1 IMGT Collier de Perles. (**B**) IMGT Collier de Perles of 3hg1, available in IMGT/3Dstructure-DB, http://www.imgt.org. The IMGT pMH contact sites [64-67] between the Homsap G domains and the peptide MLANA (melan-A, MART1) Pr26-35 (ELAGIGILTV A2>L), also known as melanoma antigen recognized by T cells 1 or MART-1, are displayed with the C1-C11 pMH contact site colors (see also Hs3 in Figure 9).

For a complete picture, according to comparative modelisation, a detailed atome pair contact sites analysis in the human 3hg1.pdb based on the IMGT unique numbering is reported in Supplementary Table S4, Figures 9 and 10. Considering the features of contacts between the *Homo sapiens* G domains and the TR V domains displayed in lines Hs2 of the alignments of the G-ALPHA1 (Figure 9A) and G-ALPHA2 (Figure 9B) domains of *Tursiops truncatus* MH1 I-ALPHA with *Homo sapiens* MH1 I-ALPHA HLA-A*0201, the probability that the same positions are involved in atom pair contacts in dolphins are very high (in Figure 9, in yellow the amino acids in dolphin compared to 3hg1). Therefore, the results of the comparative human/dolphin modelisation, integrated with the atome pair contact sites analysis between the human MH1 grove (G) domains and the TR V domains, confirm conservation of the structure of the dolphin TR/pMH

Supplementary Materials: Table S1: Description of the TRB genes in the *Tursiops truncatus* genome assembly NC_047042.1 (NCBI Reference Sequence). The position of all genes and their classification and functionality are reported. **Table S2:** Description of the unrelated TRB genes in the *Tursiops truncatus* genome assembly NC_047042.1 (NCBI Reference Sequence). The position of all genes and their classification and functionality are reported. **Table S3:** Description of the *Turtru* TRBV pseudogenes. **Table S4:** Contact analysis 3Hg1. **Figure S1:** Dot-plot matrix of dolphin/dromedarius TRB sequence genomic comparison. The transcriptional orientation of each gene is indicated by arrow heads. Colored rectangle (orange) encloses TRBV deleted regions in dolphin as referred in the text. The green line indicates the only region where gene duplication took place in *T. truncatus* TRB locus. **Figure S2:** Dot-plot matrix of dolphin/pig TRB sequence genomic comparison. The transcriptional orientation of each gene is indicated by arrow heads. Colored rectangle (orange) encloses TRBV deleted regions in dolphin as referred in the text. The green line indicates the only region where gene duplication took place in *T. truncatus* TRB locus. **Figure S3:** Dot-plot matrix of dolphin/goat TRB sequence genomic comparison. The transcriptional orientation of each gene is indicated by arrowheads. Colored rectangles (red) enclose TRBV duplicated regions in goat and (orange) TRBV deleted regions in dolphin as referred in the text. **Figure S4 A:** Sequence-structure alignment of the *T.truncatus* TR alpha chains with human TR alpha chains from the crystallized structures 3hg1.pdb and 5d2l.pdb. **Figure S4 B:** Sequence-structure alignment of the *T.truncatus* TR beta chains with the human TR beta chains from the crystallized structures 3hg1.pdb and 5d2l.pdb. **Figure S4 C:** Sequence-structure pairwise

alignment of the human MH1 I-ALPHA (from 3hg1.pdb) with its closest counterpart in *T.truncatus*. **Figure S4 D:** Sequence-structure pairwise alignment of the human B2M (from 3hg1.pdb) with its closest counterpart in *T.truncatus*. **Figure S5:** *Tursiops truncatus* TRB locus D-J-C Region.

Author Contributions: Author Contributions: Conceptualization, G.L., M-P.L., S.C. and R.A.; Methodology, G.L., C.L.P., J.J-M., G.F. and S.M.; Investigation, J. J-M., G.F., and S.M. Writing – Original Draft Preparation, G.L. and R.A.; Writing – Review & Editing, R.A., C.L.P., M-P.L. and S.C.; Supervision, S.K., M-P.L. and S.C. Resources and Funding Acquisition, G.L., R.A. and S.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gingerich, P.D.; Haq, Mu.; Zalmout, I.S.; Khan, I.H.; Malkani, M.S. Origin of whales from   early artiodactyls: hands and feet of Eocene Protocetidae from Pakistan. *Science* **2001,** *293*,2239-42.

2. Price, S.A.; Bininda-Emonds, R,P.; Gittleman, J.L. A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla). *Biol Rev Camb Philos Soc.* **2005,** *80*, 445-73.

3. Dawkins, R. The Ancestor's tale: A Pilgrimage to the Dawn of Evolution. Collana Biologia.Arnoldo Mondadori Editore **2004**, pp. 673.

4. 4.Shen, Y.; Yue, F.; McCleary, D.F.; Ye, Z.; Edsall, L.; Kuan, S.; Wagner, U.; Dixon, J.; Lee, L.; Lobanenkov, V.V.; Ren, B. A map of the cis-regulatory sequences in the mouse genome. *Nature* **2012**, *488*, 116-20.

5. Ishengoma, E.; Agaba, M. Evolution of toll-like receptors in the context of terrestrial ungulates and cetaceans diversification. *BMC Evol. Biol.* **2017**, *17*, 54.

6. Ciccarese S.; Carlucci R.; Ciani E.; Corcella E.; Cosentino A.; Fanizza C.; Linguiti G.; Antonacci R. Cytochrome b marker reveals an independent lineage of Stenella coeruleoalba in the Gulf of Taranto. *PLoS One* **2019**, *14*, 3

7. Dias, B.C.; Nerey, M.F. Analyses of RAG1 and RAG2 genes suggest different evolutionary rates in the Cetacea lineage. *Mol. Immunol*. **2020**, *117*, 131-138.

8. Lefranc, M.-P.; Lefranc, G. *The T cell receptor FactsBook;* Academic Press, London, **2001**, 1-398.

9. Linguiti, G.; Antonacci, R.; Tasco, G.; Grande, F.; Casadio, R.; Massari, S.; Castelli, V.; Consiglio, A.;Lefranc, M.P.; Ciccarese, S. Genomic and expression analyses of Tursiops truncatus T cell receptor gamma (TRG) and alpha/delta (TRA/TRD) loci reveal a similar basic public   repertoire in dolphin and human. *BMC Genomics* **2016**, *17*, 634.

10. Massari, S.; Bellahcene, F.; Vaccarelli, G.; Carelli, G.; Mineccia, M.; Lefranc, M.-P.; Antonacci, R.; Ciccarese, S. The deduced structure of the T cell receptor gamma locus in Canis lupus familiaris. *Mol. Immunol*. **2009**, *46*, 2728–2736.

11. Ciccarese, S.; Burger, P.A.; Ciani, E.; Castelli, V.; Linguiti, G.; Plasil, M.; Massari, S.; Horin, P.; Antonacci, R. The Camel Adaptive Immune Receptors Repertoire as a Singular Example of Structural and Functional Genomics. *Front. Genet*. **2019**, *10*, 997.

12. 12.Antonacci, R.; Linguiti, G.; Burger, P.A.; Castelli,V.; Pala, A.; Fitak, R.; Massari, S.; Ciccarese, S. Comprehensive genomic analysis of the dromedary T cell receptor gamma (TRG) locus and identification of a functional TRGC5 cassette. *Dev. Comp. Immunol*. **2020**, *106*, 103614.

13. 13. Vaccarelli, G.; Miccoli, M.C.; Antonacci, R.; Pesole, G.; Ciccarese, S. Genomic organization   and recombinational unit duplication-driven evolution of ovine and bovine T cell receptor gamma loci. *BMC Genomics* **2008**, *9*, 81.

14. Miccoli, M.C.; Antonacci, R.; Vaccarelli, G.; Lanave, C.; Massari, S.; Cribiu, E.; Ciccarese, S. Evolution of TRG clusters in cattle and sheep genomes as drawn from the structural analysis of the ovine TRG2@ locus. *J. Mol. Evol*. **2003**, *57*, 52–62.

15. Antonacci, R.; Lanave, C.; Del Faro, L.; Vaccarelli, G.; Ciccarese, S.; Massari, S. Artiodactyl emergence is accompanied by the birth of an extensive pool of diverse germline TRDV1 genes. *Immunogenetics* **2005**, *57*, 254–66.

16. Piccinni, B.; Massari, S.; Caputi Jambrenghi, A.; Giannico, F.; Lefranc, M.-P.; Ciccarese, S.; Antonacci, R. Sheep (Ovis aries) T cell receptor alpha (TRA) and delta (TRD) genes and genomic organization of the TRA/TRD locus. *BMC Genomics* **2015**, *16*, 709.

17.   Pégorier, P.; Bertignac, M.; Nguefack Ngoune, V.; Folch, G.; Jabado-Michaloud, J.; Giudicelli, V.; Duroux, P.; Lefranc, M. P.; Kossida, S. IMGT® Biocuration and Comparative Analysis of *Bos taurus* and *Ovis aries* TRA/TRD Loci. *Genes* **2020**, *12*,1-30.

18.   Rowen, L.; Koop, B.F.; Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **1996**, *272*, 1755–1762.

19.   Lefranc, M.-P. Locus maps and genomic repertoire of the human T cell receptor genes. *The Immunologist* **2000**, *8/3*, 72-79.

20.   Lefranc, M.-P. Immunoglobulin (IG) and T cell receptor genes (TR): IMGT and the birth and rise of immunoinformatic. *Front. Immunol*. **2014**, *5*, 22.

21.   Massari, S.; Bellini, M.; Ciccarese, S.; Antonacci, R. Overview of the germline and expressed repertoires of the TRB genes in *Sus scrofa*. *Front. Immunol.* **2018**, *9*, 2526.

22.   Antonacci, R.; Bellini, M.; Pala, A.; Mineccia, M.; Hassanane, M. S.; Ciccarese, S.; Massari, S. The occurrence of three D-J-C clusters within the dromedary TRB locus highlights a shared evolution in Tylopoda, Ruminantia and Suina. *Dev. Comp. Immunol*. **2017**, *76*, 105–119.

23.   Antonacci, R.; Bellini, M.; Castelli, V.; Ciccarese, S.; Massari, S. Data charactering the genomic structure of the T cell receptor (TRB) locus in *Camelus dromedarius*. Data Brief **2017**, *14*, 507–514.

24.   Antonacci, R.; Di Tommaso, S.; Lanave, C.; Cribiu, E. P.; Ciccarese, S.: Massari, S. Organization, structure and evolution of 41 kb of genomic DNA spanning the D-J-C region of the sheep TRB locus. *Mol. Immunol.* **2008**, *45*, 493–509.

25.   Schwartz, S.; Zhang, Z.; Frazer, K.A.; Smit, A.; Riemer, C.; Bouck, J.; Gibbs, R.; Hardison, R.; Miller, W. PipMaker- a web server for aligning two genomic DNA sequences. *Genome Res*. **2000**, *10*, 577-586.

26.   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.

27.   Lefranc, M.-P. WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T cell receptors**.** WHO-IUIS Nomenclature Subcommittee for immunoglobulins and T receptors report August 2007, 13th International Congress of Immunology, Rio de Janeiro, Brazil. *Dev Comp Immunol*. **2008**, *32*(5),461-3.

28.   Giannico F.; Massari S.; Caputi Jambrenghi A.; Soriano A.; Pala A.; Linguiti G.; Ciccarese S.; Antonacci R. The expansion of the TRB and TRG genes in domestic goats (Capra hircus) is characteristic of the ruminant species. *BMC Genomics* **2020**, *21*, 623.

29.   Edgar, R.C. MUSCLE: a multiple sequence alignment with reduced time and space complexity. *BMC   Bioinformatics* **2004**, *5*,113.

30.   Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis   version 7.0 for bigger datasets. *Molecular Biology and Evolution* **2016**, *33*, 1870-1874.

31.   Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. **1987**, *4*, 406-425.

32.   Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: New York,NY, USA, 2000.

33.   Brochet, X.; Lefranc, M.P.; Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res*. **2008**, *36*, W503-508.

34.   Giudicelli, V.; Brochet, X.; Lefranc, M.-P. IMGT/V-QUEST: IMGT standardized analysis of the immunoglobulin (IG) and T celle receptor (TR) nucleotide sequences. *Cold Spring Harb*. *Protoc*. **2011**, *6*, 695-715.

35.   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.

36.   Giudicelli, V.; Lefranc, M.-P. IMGT/JunctionAnalysis: IMGT standardized analysis of the V-J and V-D-J junction of the rearranged immunoglobulins (IG) and T cell receptors(TR). *Cold Spring Harb*. *Protoc*. **2011**, *6*, 716-725.

37.   Lobley, A.; Sadowski, M.I.; Jones, D.T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics* **2009**, *25*, 1761–1767.

38.   Yang, J.; Zhang, Y.   Protein Structure and Function Prediction Using I-TASSER. *Curr. Protoc. Bioinforma* **2015**, *52*,5.8.1-15.

39.   Trisolini, L.; Gambacorta, N.; Gorgoglione, R.; Montaruli, M.; Laera, L.; Colella, F.; Volpicella, M.; De Grassi, A.; Pierri, C.L. FAD/NADH Dependent Oxidoreductases: From Different Amino Acid Sequences to Similar Protein Shapes for Playing an Ancient Function. *J. Clin. Med*. **2019**, *8*, 2117.

40.   Pierri, C.L.; Parisi, G.; Porcelli, V.   Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening. Biochim Biophys Acta - *Proteins Proteomics* **2010**, *1804*, 1695–1712.

41.   Mercurio, I.; Tragni, V.; Busto, F.; De Grassi, A.; Pierri, C. L.   Protein structure analysis of the interactions between SARS-CoV-2 spike protein and the human ACE2 receptor: from conformational changes to novel neutralizing antibodies. *Cell Mol Life Sci* **2020**, 1–22.

42.   Pierri, CL.; Bossis, F.; Punzi. G; De Grassi, A.; Cetrone, M.; Parisi, G.; Tricarico, D.   Molecular modeling of antibodies for the treatment of TNFα-related immunological diseases. *Pharmacol Res Perspect* **2016,** *4*, e00197.

43.   Krieger, E.; Joo, K.; Lee, J.; Lee, J.; Raman, S.; Thompson, J.; Tyka, M.; Baker, D.; Karplus, K. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins Struct. Funct. Bioinforma*. **2009**, *77*, 114-22.

44.   Cole, DK.; Yuan, F.; Rizkallah, PJ.; Milles, J.J.; Gostik, E.; Price, D.A.; Gao, G, F.; Jakobsen, B. K.; Sewell, A. K.   Germ line-governed recognition of a cancer epitope by an immunodominant human T-cell receptor. *J Biol Chem*. **2009**, *284*, 27281-9.

45. Van Durme, J.; Delgado, J.; Stricher, F.; Serrano, L.; Schymkowitz, J.; Rousseau, F. A graphical interface for the FoldX forcefield. *Bioinformatics* **2011**, *27*, 1711–1712.

46. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Res* **2005**, *33*, W382-8.

47. 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.

48. Connelley, T.; Aerts, J.; Law, A.; Morrison, W. I. Genomic analysis reveals extensive gene duplication within the bovine TRB locus. *BMC Genomics* **2009**, 10.

49. Antonacci, R.; Bellini, M.; Ciccarese, S.; Massari, S. Comparative analysis of the TRB locus in the Camelus genus. *Front. Genet.* **2019**, *10*, 482.

50. Lefranc, M.-P.; Pommié, C.; Kaas, Q.; Duprat, E.; Bosc, N.; Guiraudou, D.; Jean, C.; Ruiz, M.; Da Piédade, I.; Rouard, M.; Foulquier, E.; Thouvenin, V.; Lefranc, G. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Dev. Comp. Immunol*. **2005**, *29*, 185-203.

51. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **1985**, *39*, 783-791.

52. Tamura, K.; Nei, M.; & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences (USA)* **2004**, *101*, 11030-11035.

53. Mineccia, M.; Massari, S.; Linguiti, G.; Ceci, L.; Ciccarese, S.; Antonacci, R. New insight into the genomic structure of dog T cell receptor beta (TRB) locus inferred from expression analysis. *Dev. Comp. Immunol.* **2012**, *37*, 279–293.

54. Hall, M.A.; Lanchbury, J. S. Healthy human T-cell receptor beta- chain repertoire. Quantitative anlysis and evidence for J beta-related effects on CDR3 structure and diversity. *Hum. Immunol*. **1995**, *43*, 207-18.

55. Persson, B. Bioinformatics in protein analysis. *EXS* **2000**, *88*, 215–231.

56. Waterhouse, A.M.; Procter, J.B.; Martin, D.M.; Clamp, M.; Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **2009**, *25*,1189–1191.

57. Lefranc, M.-P. IMGT Collier de Perles for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb. Protoc.* **2011**, *6*, 643–51.

58. Pégorier, P.; Bertignac, M.; Chentli, I.; Nguefack Ngoune, V.; Folch, G.; Jabado-Michaloud, J.; Hadi-Saljoqi, S.; Giudicelli, V.; Duroux, P.; Lefranc, M.P.; Kossida, S. IMGT Biocuration and Comparative Study of the T Cell Receptor Beta Locus of Veterinary Species Based on Homo sapiens TRB. *Front. Immunol.* **2020**, *11*, 821.

59. Di Tommaso, S.; Antonacci, R.; Ciccarese, S.; Massari, S. Extensive analysis of D-J-C arrangements allows the identification of different mechanisms enhancing the diversity in sheep T cell receptor β-chain repertoire**.** *BMC Genomics* **2010**, *11*, 3.

60. Venturi, V.; Kedzierska, K.; Tanaka, M.M.; Turner, S.J.; Doherty, P.C.; Davenport, M.P. Method for assessing the similarity between subsets of the T cell receptor repertoire. *J Immunol Methods* **2008**, *329*(1–2), 67–80.

61. Piepenbrink, K.H.; Blevins, S.J.; Scott, D.R.; Baker, B.M. The basis for limited specificity and MHC restriction in a T cell receptor interface. *Nat Commun.* **2013**, *4*, 1948.

62. Castro, C.C.; Luoma, A.M.; Adams, E.J. Coevolution of T-cell receptors with MHC and non-MHC ligands. *Immunol Rev.* **2015**, *267*(1), 30–55.

63. Lefranc, M.-P.; Duprat, E.; Kaas, Q.; Tranne, M.; Thiriot, A.; Lefranc, G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKEDOMAIN. *Dev Comp Immunol* **2005**, *29*, 917-938.

64. Kaas, Q. and Lefranc, M-P. T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. *In Silico Biol.* **2005**, *5*, 505-528.

65. Kaas, Q.; Duprat, E.; Tourneur, G.; Lefranc M-P. IMGT standardization for molecular characterization of the T cell receptor/peptide/MHC complexes. In: Schoenbach C, Ranganathan S and Brusic V (ed), Immunoinformatics. *Immunomics Reviews* **2008** Series of Springer Science and Business Media LLC, Springer, New York, USA chap 2, pp 19-49.

66. Ehrenmann, F.; Kaas, Q.; Lefranc, M.P. (2010) IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res* **2010,** *38*, D301-307.

67. Ehrenmann, F.; Lefranc, M.P. IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb. Protoc.* **2011**, *2011*, 750–761.

68. Ehrenmann, F.; Giudicelli, V.; Duroux, P.; Lefranc, M.P. IMGT/Collier de Perles: IMGT standardized representation of domains (IG, TR, and IgSF variable and constant domains, MH and MhSF groove domains). *Cold Spring Harb. Protoc.* **2011**, *2011*, 726–736.