Developmental and Comparative Immunology 37 (2012) 279–293

Contents lists available at SciVerse ScienceDirect

Developmental and Comparative Immunology

journal homepage: www.elsevier.com/locate/dci

# New insight into the genomic structure of dog T cell receptor beta (TRB) locus inferred from expression analysis

Micaela Mineccia [a], Serafina Massari [b], Giovanna Linguiti [a], Luigi Ceci [c], Salvatrice Ciccarese [a], Rachele Antonacci [a,*]

[a] Dipartimento di Biologia, Universita' degli Studi di Bari Aldo Moro, Bari, Italy
[b] Dipartimento di Scienze e Tecnologie Biologiche ed Ambientali, Universita' del Salento, Lecce, Italy
[c] Dipartimento di Sanita' Pubblica e Zootecnia, Universita' degli Studi di Bari Aldo Moro, Bari, Italy

ARTICLE INFO

*Article history:*
Received 14 February 2012
Revised 16 March 2012
Accepted 20 March 2012
Available online 28 March 2012

*Keywords:*
T cell receptor
TRB genes
Dog
Comparative genomics
Beta chain repertoire
IMGT

ABSTRACT

Here is an updated report on the genomic organization of T cell receptor beta (TRB) locus in the domestic dog (*Canis lupus familiaris*) as inferred from comparative genomics and expression analysis. The most interesting results we found were a second TRBD–J–C cluster, which is absent from the reference genome sequence, and the annotation of two additional TRBV genes. In dogs, TRB locus consists of a library of 37 TRBV genes positioned at the 5′ end of two in tandem aligned D–J–C gene clusters, each composed of a single TRBD, 6 TRBJ and one TRBC genes, followed by a single TRBV gene with an inverted transcriptional orientation. The TRB genes are distributed in less than 300 kb, making the canine locus, one of the smaller mammalian TRB locus studied so far. The small size may be ascribed to reduced gene duplication occurrences and a lower density of total interspersed repeats compared to humans and mice. Despite the low TRBV gene content, a large and diversified beta chain repertoire is displayed in the dog peripheral blood. A full usage of TRBV and TRBJ genes, including pseudogenes, and a high level of allelic polymorphism contribute to generate diversity. Finally, this study suggests that the overall TRB locus organization is evolutionarily conserved supporting the dog as a highly suited model system for immune development and diseases.



## 1. Introduction

The domestic dog (*Canis lupus familiaris*) is added to a growing list of vertebrate species that have had their genome sequenced. The position of this species within the mammalian evolutionary tree makes it an important guide for the comparative analysis. The special degree of similarity between human and dog, in combination with the pathophysiological analogies shared between the two species, have placed the modern dog in a position of high visibility as a model system for a variety of human diseases, especially in the context of hematologic malignancies.

T cells play a central role for most of the functions within the immune system. They can be divided into two distinct subsets, αβ and γδ T cells, based on the type of T cell receptors (TR) expressed. The αβ T cells mainly recognize peptide antigens bound to class I or class II major histocompatibility (MH) proteins, whereas γδ T cells mount immune responses directly to nonpeptide antigens and superantigens (Rust et al., 1990). Canines as well as humans and mice belong to animal species in which αβ TR T lymphocytes are the predominant lymphocyte population in both the peripheral blood and lymphoid organs (Faldyna et al., 2001, 2005).

To generate a vast repertoire of TR capable of recognizing diverse peptide–MH (pMH) complexes, the loci encoding the α (TRA) and β (TRB) chain undergoes somatic rearrangements among the variable (V) and joining (J) genes (for the α chain) and V, diversity (D), and J genes (for the β chain) during development of T lymphocytes in the thymus (Gellert, 1992, 2002; Lefranc and Lefranc, 2001; Jung and Alt, 2004). After transcription, the V(D)J sequence is spliced to the constant (C) gene. The V–(D)–J rearrangement is directed by the lymphoid-specific recombinase (RAG, composed of RAG1 and RAG2) and ubiquitously expressed DNA repair proteins (Jung and Alt, 2004). RAG proteins create double-strand breaks at recombination signal (RS) sequences that flank TR V, D and J genes, and these breaks are subsequently resolved by nonhomologous end joining. Diversity at the rearrangement level is further enhanced by other processes that include the trimming of the 3′V-, 5′D- and 3′D-, and 5′J-REGION by exonuclease, the imprecise joining of nicked genes, and the addition of

* Corresponding author. Address: Dipartimento di Biologia, Universita' degli Studi di Bari Aldo Moro, Via Amendola 165/A, 70126 Bari, Italy. Tel.: +39 080 5443338; fax: +39 080 5443386.
*E-mail addresses:* mineccia14@alice.it (M. Mineccia), sara.massari@unisalento.it (S. Massari), giovanna.linguiti@uniba.it (G. Linguiti), l.ceci@veterinaria.uniba.it (L. Ceci), ciccarese@biologia.uniba.it (S. Ciccarese), r.antonacci@biologia.uniba.it (R. Antonacci).

0145-305X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.dci.2012.03.010

nongermline nucleotides (N nucleotides) by the terminal deoxynucleotidyl transferase (TdT) at the D–J, V–(D–J) and V–J junctions that can be identified in rearranged sequences (Yousfi Monod et al., 2004; Bleakley et al., 2006; Giudicelli and Lefranc, 2011). For this reason the sequence that encompasses the V–J and V–D–J junctions, called complementarity determining region (CDR) 3, is markedly polymorphic and encodes the portion of the TR V domain that directly contacts pMH (Davis and Bjorkman, 1988; Ehrenmann et al., 2010; Ehrenmann and Lefranc, 2011). Otherwise CDR1 and CDR2 loops of the TRA and TRB chains that predominantly interact with MH, are encoded within the germline V genes (Garboczi et al., 1996; Garcia et al., 1996).

Thereby, the number of V, D and J genes in the germline DNA is an important determinant of the extent of the TR diversity.

In most species of mammals (IMGT Repertoire, http://www.imgt.org; Williams et al., 1991; Jaeger et al., 1993; Cheynier et al., 1996; Schrenzel et al., 1994; Lefranc and Lefranc, 2001; Giudicelli et al., 2005; Lefranc et al., 2009), the TRB locus has the common feature of a library of V genes (TRBV) positioned at the 5′ end of two in tandem aligned D–J–C clusters, each composed of a single D (TRBD), 6-7 J (TRBJ) genes and one C (TRBC) gene, followed by a single TRBV gene with an inverted transcriptional orientation located at the 3′end. In all species, the two TRBC genes are very similar, since they differ by only a few amino acids (AA) in the coding region; conversely, they are different in their own 3′UTR regions. Moreover, trypsinogen genes (TRY) are typically interspersed among TRB genes.

The genomic structure of the dog TRB locus has recently been annotated (IMGT Repertoire, http://www.imgt.org). The TRB sequence consists of a pool of 34 TRBV genes followed by 1 TRBD, 6 TRBJ, 1 TRBC genes, suggesting the presence of only one D–J–C cluster in the dog genome. Another TRBV gene in opposite orientation lies downstream of the TRBC gene.

In this paper, we provide an upgraded annotation of the canine genome TRB locus. By means of expression analysis, we have demonstrated the existence of a second D–J–C cluster involved in the repertoire formation, while by comparative analysis we have identified and annotated two additional TRBV genes and four TRY genes already present in the genome assembly.

Overall, dog TRB represents a minimal locus compared to human and mouse, with a smaller proportion of total interspersed repeats and a lower number of genes, mainly due to a reduced number of duplication events that occurred after the separation of carnivores, primates and rodents. In spite of the presence of a lower number of germline TRBV genes, the dog TRB maintains a well-conserved structure with a high degree of homology, within coding and regulatory regions, with respect to human and mouse. In addition, the usage of TRBV and TRBJ pseudogenes as well as the presence of a large number of alleles seems to compensate for the reduced availability of germline TRB genes and contribute to generating the beta chain repertoire diversity.

## 2. Materials and methods

### 2.1. Genome analyses

To analyze the dog TRB locus, a sequence of 324056 pb (gaps excluded) was retrieved directly from the reference sequence NW_876260[1] (*Canis lupus familiaris* chromosome 16 genomic contig) of the Build 2.1 whole genome shotgun sequence available at GenBank. Particularly, the analyzed region comprises contigs from AAEX02035464 to AAEX02035468.

MOXD2 (monoxygenase, dopamine-beta-hydroxylase-like 2, DBH-like2) and EPHB6 (ephrin type-B receptor 6) genes, flanking respectively, the 5′ and 3′ ends of TRB locus, were included in the analysis. MOXD2 was present in AAEX02035468 contig, while the first seven exons of the EPHB6 gene were recovered in AAEX02035464 contig. Both genes were predicted to be functional in dog (GenBank ID: XM_539871 and XM_532743).

Computational analysis of the dog TRB locus was conducted with the following programs: RepeatMasker for the identification of genome-wide repeats and low complexity regions (Smit, A.F.A., Hubley R., Green P., RepeatMasker at http://www.repeatmasker.org) and Pipmaker (Schwartz et al., 2000; http://www.pipmaker.bx.psu.edu/pipmaker/) for the alignment of the dog sequence with the human and mouse counterpart. The human sequence, from MOXD2 to EPHB6 genes, derived from the GenBank and IMGT/LIGM-DB (Giudicelli et al., 2006) ID: L36092 (684973 pb) plus Ensembl *Homo sapiens* chromosome 7 VEGA41 partial sequence (36754 pb) corresponding to the first seven exons of the EPHB6 gene. The mouse counterpart sequence (739138 pb) encompasses three contigs (GenBank ID: AE000663-664-665) plus 38178 pb of the EPHB6 gene derived from GenBank ID: AF336378. RepeatMasker screens DNA sequences for interspersed repeats and low complexity DNA sequences. The RepeatMasker analysis has been carried out using the Carnivora Repbase division section.

### 2.2. Animals (source of tissue)

Peripheral blood samples were obtained from three different crossbred unrelated animals: one healthy and two sick individuals suffering from visceral leishmaniasis, one of which with regression of the disease after treatment with multiple daily doses of *N*-methylglucamine antimoniate (unpublished). All animal manipulations were carried out with the approval of the Bari Animal Ethics Commitee and in compliance with Institutional Animal Care and Use Comittee (IACUC) requirements.

### 2.3. 5′ RACE

Total RNA was extracted from peripheral blood leukocytes (PBL) of three unrelated animals using the Trizol method according to the manufacturer's protocol (Invitrogen, Carlsbad, CA). Four RNA samples were obtained: 5R from the healthy, 5Rm from a sick, and 5RCG as well as 5RD from another sick animal. Four different experiments were set up. About 5 μg of total RNA were reverse transcribed with Superscript II (Invitrogen, Carlsbad, CA) by using the oligodT (5R and 5Rm series clones) or a specific primer, CB2L (Supplementary Table S1) (5RCG and 5RD series clones), designed on the sequence of the first exon of the dog TRBC gene sequence. After linking a poly-C tail at the 5′-end of the cDNAss, the cDNAds was performed with Platinum Taq Polymerase (Invitrogen) by using CB2L as lower primer and an anchor oligonucleotide as upper primer (AAP) provided from the supplier (Invitrogen). The PCR conditions were the following: 30 s at 94 °C, 1 min at 55 °C, 1 min at 72 °C for 35 cycles. The final cycle was extended for 7 min at 72 °C. The products were then amplified in a subsequent nested PCR experiment by using B8L/AUAP primer pair. B8L (Supplementary Table S1) is designed upstream CB2L while AUAP oligonucleotide was provided from the supplier (Invitrogen). The PCR conditions were the following: 30 s at 94 °C, 45 s at 58 °C, 45 s at 72 °C for 35 cycles. The final cycle was extended for 30 min at 72 °C.

[1] This record exists as outdated version since it has been recently replaced by the NCBI ID: NW_003726086.1 derived from the latest CanFam3.1 whole genome shotgun sequence. The TRB sequence retrieved and analysed in this study ranges from base 6706213 to 7035672, gaps included.

### 2.4. Cloning and sequence analysis

The RACE products were then gel-purified and cloned using StrataClone PCR Cloning Kit (Statagene). 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 W2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/) and IMGT® tools [IMGT/V-QUEST (Brochet et al., 2008; Giudicelli et al., 2011) with integrated IMGT/JunctionAnalysis tools (Yousfi Monod et al., 2004; Giudicelli and Lefranc, 2011)] and the IMGT unique numbering for V domain (Lefranc et al., 2003) (http://www.imgt.org/). The sequence analysis revealed that 75% of the 5R, 33% of the 5Rm, 30% of the 5RCG and 68% of the 5RD series selected clones contained rearranged V–D–J–C transcripts, with a total of 60%, 66%, 80% and 30% of different clones with a correct open reading frame, respectively. All cDNA clones were registered in EMBL database with the Accession numbers from HE653930 to HE653965.

### 2.5. Long-PCR

Genomic DNA was extracted from PBL of a healthy individual by Easy-DNA Kit (Invitrogen) and 50–100 ng used for the amplifications. Expand High Fidelity DNA polymerase (Roche) and TaKaRa LA Taq polymerase (TAKARA BIO INC.) were used to eliminate possible PCR errors and to increase the chances of getting the right DNA fragment. Primers were designed based on known genomic sequences or on the cDNA sequence of rearranged sequences containing TRBJ2. Genomic TRBD2, TRBJ2 and TRBC2 sequences were acquired from six PCR clones. Primer combinations used were JB1.6U and JB8L for the JB1-2gen clone, JB8U and B8L for JB8Ugen3 clone, CBEXIdogU and 3UTRdogL for the C2gen11 clone. Additional PCR experiments were performed to minimize any possible PCR bias and/or error, using JB1.6U/CIIIEXL, CB1U/JB8L, and CIIIEXU/JB8L primer pairs for the JB1CIIIgen11, CBJB8gen and CIIIEXUgen7 clone, respectively. The nucleotide sequences and the positions of all primers used in this study are listed in Supplementary Table S1. Every PCR was performed following the manufacturer's instructions for each DNA polymerase. Amplified DNA fragments were purified and cloned using TOPO XL PCR Cloning Kit (Invitrogen).

### 2.6. Phylogenetic analysis

The TRBV, TRBJ and TRBC genes used for the phylogenetic analysis were retrieved from the following genomic sequences deposited in the GenBank database: NG_001333 (human TRB locus contig); NG_006980 (mouse TRB locus contig); NW_876260[1] and HE653929 (this work) (dog TRB locus contig).

Multiple alignments of the sequences under analysis were carried out with the MUSCLE program (Edgar, 2004), while we used a manual alignment for TRBV sequences from FR1 to CDR3.

Phylogenetic analyses were conducted in MEGA5 (Tamura et al., 2011). We used the neighbour-joining (NJ) method to obtain TRBJ and TRBV phylogenetic trees; while the maximum parsimony (MP) method was used to reconstruct the TRBC phylogenetic trees.

## 3. Results

### 3.1. Analysis of the dog TRB locus retrieved from the genome assembly: identification of non-annotated TRB and no-related TRB genes

The dog genome has been sequenced (Lindblad-Toh et al., 2005), and assemblies are continually being refined. Recently the TRB locus annotation has been updated. The canine TRB locus has been recognized in four continuous contigs (GenBank ID: AAEX02035465, AAEX02035466, AAEX02035467 and AAEX02035468) of the *Canis lupus familiaris* chromosome 16 genomic sequence (GenBank ID: NW_876260[1]). In these contigs, the international ImMunoGeneTics information system® (IMGT®, http://www.imgt.org) have annotated a pool of 35 TRBV genes, grouped in 25 subgroups, positioned at the 5′ of a single D–J–C cluster composed of a sole TRBD, six TRBJ and one TRBC gene. A single TRBV gene with an inverted transcriptional orientation located at the 3′ end of the locus has also been identified. Fifteen of the 35 TRBV genes (43%) were classified as pseudogenes, while all TRBJ genes except one appear to be functional.

Thus, the general structural organization of the dog TRB locus seems to be similar to that found in other mammalian species (i.e. human and mouse), except for the presence of a single D–J–C cluster.

To provide further information on the dog TRB locus genomic architecture, we retrieved directly from the genome assembly and joined the sequences of the four contigs (gaps excluded). We checked within the sequence the presence of MOXD2 and EPHB6 genes, which flank, respectively, the 5′ and 3′ ends of all mammalian TRB loci studied so far. The MOXD2 gene was present in AAEX02035468 contig; while the EPHB6 gene was recovered in the subsequent AAEX02035464 contig (see Section 2.1). The assembly of the contigs produced a sequence of a little more than 320 kb that we compared with the human and mouse corresponding region (detailed in Section 2.1). In both cases the regions are about 700 kb long.

We first analyzed the compositional properties (G + C content) and identified interspersed repeated sequences in all three species. The analysis carried out through the Repeat Masker program provided the results summarized in Supplementary Table S2. The dog GC content of 44.30% is almost identical to human (42.45%). The density of total interspersed repeats ranges from 24.93% in dogs to 46.18% in mice with a 33.83% mean value in humans. LINEs, predominantly of the LINE1 family, are the most abundant repeat elements in dogs (11.98%) as well as in humans (16.86%) and mice (30.71%).

The dog masked sequence was then aligned against the human and mouse counterparts, using the PipMaker program (Schwartz et al., 2000), and the alignment expressed as percentage identity plot (pip) (Supplementary Fig. S1). The position of all canine genes, together with the location and the orientation of the interspersed repeats, is represented in the pip. The presence in the pip of superimposed lines indicates the occurrence of redundant matches along the entire region. The clearest matches are in correspondence of all dog TRBV genes, except TRBV1 and TRBV30, due to the homology among genes. The TRBV1 is the only variable gene that lacks significant similarity to any of the human TRBV genes but displays 76% identity with the murine TRBV1 gene. The TRBV30 lies at the 3′ end of the TRBC gene in an inverted transcriptional orientation relative to the other TRB genes. This location, in common with all mammalian species studied so far (IMGT Repertoire, http://www.imgt.org), may have promoted its evolutionary maintenance. Moreover, the sequence analysis revealed in the presence of two unclassified TRBV genes as indicated by duplicated matches at the coordinates 97531–97980 and 98890–99353, respectively, in the dog sequence (arrows in Supplementary Fig. S1).

The first TRBV gene belongs to the TRBV4 subgroup and was predicted to be functional since the leader sequence at the 5′, a correct recombination signal (RS) sequence located at 3′ end, and the functionality of the coding region (absence of frameshifts and stop signals) were determined. On the contrary, the second TRBV represents a remnant of a gene since it lacks most of the canonical

parameters. However, it showed a high level of identity (66%) with the human TRBV2.

Moreover, the pip recognizes four additional identity regions. We identified these regions as the TRYX3 (coordinates 30672–34334), TRYX5 (coordinates 40870–45877), TRY7 (coordinates 56759–58861) and TRYPA (coordinates 253852–257018) trypsinogen genes. The redundant matches in correspondence of TRY7 and TRYPA genes indicate that in the human and mouse counterparts these are present as duplicate genes.

The gene positions of the new identified related and no related TRB genes of the dog genome are reported in Supplementary Table S3.

Two other regions with duplicate matches of identity were identified in the pip and they do not coincide with any known coding regions (arrowheads in Supplementary Fig. S1) but they are intergenic repeated units present within the human and mouse TRB locus. High percentage identities coincide also with non-coding regions especially between dog and human sequences. Finally, the absence of identity correlated to the presence of repeat elements in the dog sequence indicate that the repetitive sequences, LINE1 in particular, significantly contribute to the architecture of the dog TRB locus.

To delineate the genomic structure of the dog TRB locus, the masked sequence was also aligned against itself with the same program. Inspection of the obtained dot-plot matrix allowed us to identify portion of the sequence that align with more regions within the sequence itself (Fig. 1). In particular we observed, beside the perfect main diagonal line for the match of each base with itself, three parallel lines, longer than 2 kb, identifying a duplicated region of about 30 kb in which the TRBV2, TRBV3 and TRBV4 subgroup genes have arisen through a series (3–4) of tandem duplication events. The homology unit is interrupted, due to insertions or deletions, subsequent to the initial duplication of the whole region (Fig. 1b). The dot-plot matrix confirms the high level of nucleotide identity between TRBV genes as indicated by dots and diagonal lines in the central part of the dot, with TRBV1 at the 5′, and TRBV30 at the 3′ outside of the identity region. No duplicate regions were present at the D–J–C gene cluster, nor were any duplicate regions present at the no-TRB related genes, with the exception of a homology unit between TRY7 and TRYPA (gray arrowhead in Fig. 1a).

Finally, the dotplot matrix shows three intergenic regions (coordinate 31172–31257, 171004–171093 and 180783–180959, respectively) with a simple repetitive pattern scattered along the entire dog TRB locus (gray arrows in Fig. 1a). The analysis of the sequence alignment reveals that in all cases they are tRNA-SINEs, the most abundant interspersed repeats in the canine genome (Bentolila et al., 1999).

### 3.2. cDNA data identify an additional D–J cluster missing from the genome assembly

To evaluate the functional competency, predicted by computational analysis, of the dog TRB genes in the genome assembly and to identify possible novel, expressed TRB genes we performed 5′ rapid amplification of cDNA ends (RACE) experiments.

For this purpose we set up four different 5′ RACE experiments. In the first, reverse transcription was started from RNA samples prepared from peripheral blood leukocytes (PBL) of a healthy dog using oligodT (5R cDNAss sample). In the second, reverse transcription was performed from RNA of a sick dog suffering from visceral leishmaniasis using oligodT (5Rm cDNAss sample). The last two reverse transcriptions were generated from RNA of a sick animal suffering from visceral leishmaniasis and subjected to medical treatment (see Section 2.2), using oligodT (5RCG cDNAss sample) or a TRBC gene specific primer (5RD cDNAss sample).

We chose the RACE approach to avoid the potential bias associated with the use of the multiple primer sets to obtain TR beta chain transcript sequences required to amplify from all TRBV sequences and to take advantage of the unique TRBC sequence present in the dog genome reference sequence. All RACE products were then gel-purified and cloned into the TA-vector. Random selected positive clones for each cloning were sequenced and a total of 36 diverse clones of different length containing rearranged V–D–J–C transcripts with a correct open reading frame were obtained and named in accordance with the cDNAss sample by which they were derived.

Each sequence was analyzed by IMGT/V-QUEST program (Brochet et al., 2008; Giudicelli et al., 2011) that identifies the variable (V), diversity (D) and joining (J) genes and alleles by alignment with the germline dog TR gene and allele sequence set from the IMGT reference directory. In our cDNA clones, we identified 14 different TRBV genes belonging to 13 different subgroups. In 22 out of 36 clones, the TRBV gene perfectly matched the corresponding germline sequence. The remaining 14 sequences showed a nucleotide identity from 97 to 99% with respect to the reference germline gene sequence. Based on the assumption that sequences sharing $\leqslant$97% nucleotide identity represent distinct genes (Arden et al., 1995a, 1995b), we referred to these as new alleles. In this respect it should be noted the high degree of allelic polymorphism of the TRBV genes, as we found 25 alleles for a total of 14 different genes, although only three individuals were analyzed. Interestingly, the TRBV26 and TRBV28 genes both qualified as pseudogenes because of a stop codon in the germline CDR3-IMGT at position 108 (IMGT unique numbering, Lefranc et al., 2003) (IMGT Repertoire, Alignments of alleles, http://ww.imgt.org), originated a productive transcript (5R130 and 5R88 clones) due to deletion of the last codons (stop codon included) during rearrangement. In the same way, the TRBJ1-3 gene that has a stop codon in its 5′J-REGION was found expressed in productive transcripts, following the trimming of the stop codon during rearrangement. In addition, the IMGT/V-QUEST program identified the TRBD1 gene in 17 clones and 5 TRBJ sequences (TRBJ1-1, TRBJ1-2, TRBJ1-3, TRBJ1-4, TRBJ1-6) in 30 clones. However, 16 TRBJ sequences showed a high nucleotide identity (above 80%) with the corresponding germline ones of the IMGT reference directory; whereas, 14 TRBJ sequences showed a high level of similarity with the human and mouse TRBJ genes belonging to the second TRB D–J–C cluster (data not shown).

With regard to the C portion, the nucleotide sequence was the same in all cDNAs and matched well with the known TRBC gene.

These data suggest that the repertoire of dog TRBV genes seems to be comprehensive with respect to the genome assembly but there is evidence for the presence of further TRBJ as well as TRBD genes missing from the genome assembly.

### 3.3. Molecular characterization of the D–J–C region and analysis of the gene content

To isolate the additional D–J–C cluster, we performed long-PCR experiments on dog genomic DNA. First, we set up a PCR by using a gene-specific primer pair (JB1.6U and JB8L) designed respectively on the coding sequence of the TRBJ1-6 gene derived from the genomic reference assembly, and on the previously unclassified J sequence of the 5R20 cDNA clone. A second amplification reaction was performed by using an upper primer, designed in the same unclassified J sequence (JB8U), and a lower primer (B8L) designed on the first exon of the TRBC gene (for the list of primers see Supplementary Table S1).

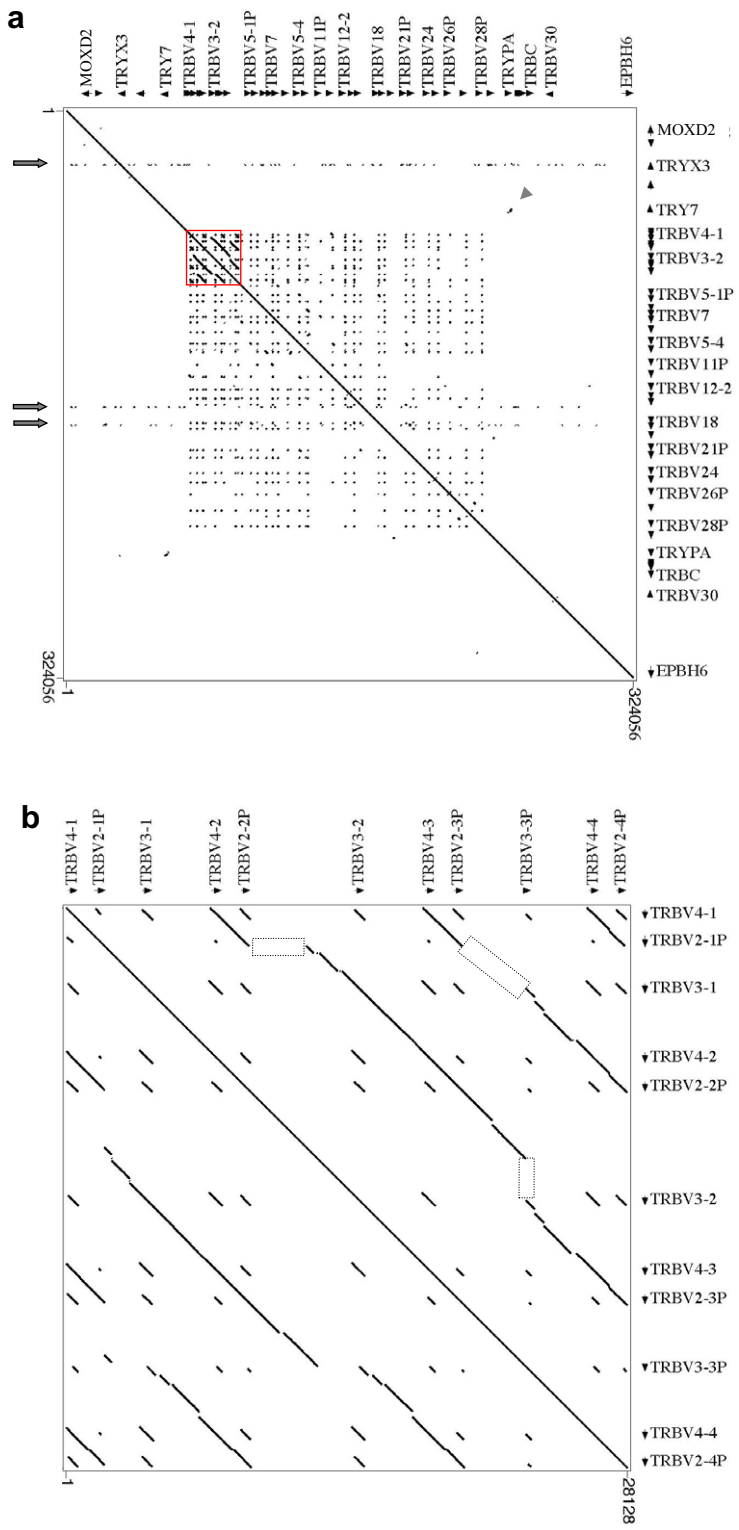


**Fig. 1.** Dot-plot of the dog TRB sequence against itself. (a) With the exception of the main diagonal, dots and diagonal lines indicate internal homology units in the sequence. The red box shows the TRBV region underwent to duplication events enlarged in (b). The arrows indicate tRNA-SINEs scattered along the dog TRB locus. The gray arrowhead points to the homology between TRY7 and TRYPA genes. (b) The pattern of parallel lines indicates the duplicated region of about 30 kb containing the TRBV2, TRBV3 and TRBV4 genes. The shaded boxes indicate gaps (insertions or deletions) between the homology units. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The PCR products (of 7789 bp and 4080 bp) were cloned and named JB1-2gen and JB8Ugen3 clone, respectively (Fig. 2). The sequence analysis showed the presence in JB1-2gen clone, downstream the TRBJ1-6 gene, of a TRBC gene with the third intron and the 3′UTR region different with respect to the known gene, indicating the existence of two TRBC genes. Downstream the new TRBC gene, a TRBD and three novel TRBJ genes were found. In the second clone JB8Ugen3, we isolated three more novel TRBJ genes and the first exon of the second TRBC gene.

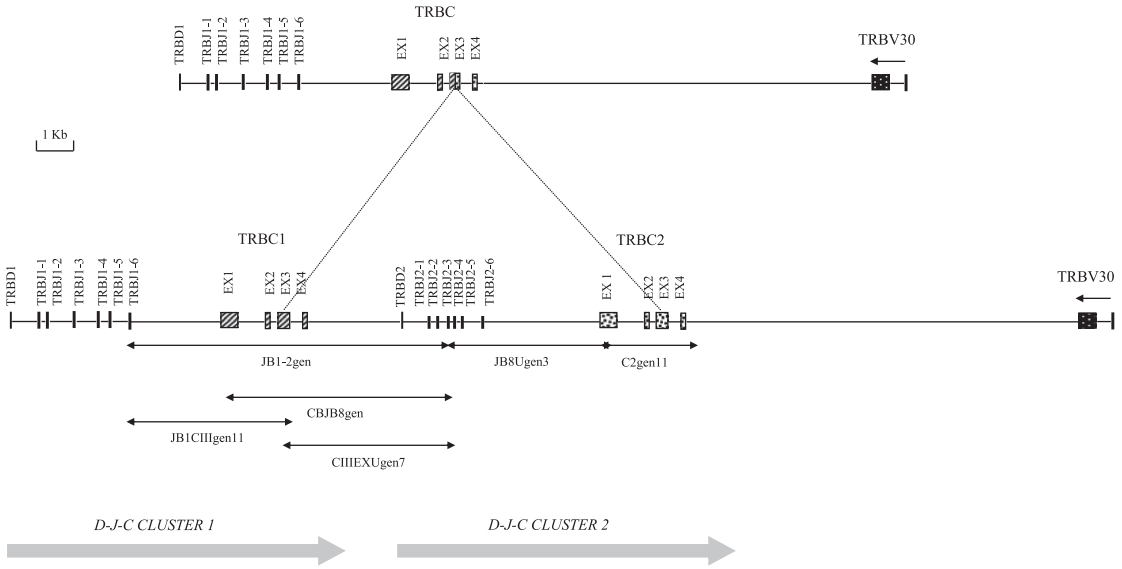


**Fig. 2.** Schematic map of the D–J–C region of the dog TRB locus. Boxes indicate the positions of the TRB genes identified within the region, with the cluster names indicated below. The genes are not to scale. The sequenced region in this study (EMBL ID: HE653929) was assembled from six PCR products (arrowed solid lines) and inserted within the TRBC third exon of the retrieved genomic assembly (GenBank ID: NW_876260[1]). The name of each PCR clone is also indicated.

**Table 1**
Percentage nucleotide identities of each exon/intron between dog, human and mouse TRBC genes. The length for each exon and intron is also reported.

| | EX 1 | INT 1 | EX 2 | INT 2 | EX 3 | INT 3 | EX 4 (CDS) | EX 4 (3′ UTR) |
|---|---|---|---|---|---|---|---|---|
| *TRBC1 versus TRBC2* | | | | | | | | |
| Dog | 99 | 99 | 100 | 98 | 99 | 8 | 94 | 30 |
| Human | 100 | 70 | 88 | 75 | 92 | 6 | 93 | 8 |
| Mouse | 99 | 11 | 80 | 23 | 85 | 8 | 80 | 14 |
| *Length (nt)* | | | | | | | | |
| Dog TRBC1 | 387 | 697 | 18 | 157 | 107 | 323 | 18 | 224 |
| Dog TRBC2 | 387 | 700 | 18 | 157 | 107 | 295 | 18 | 205 |
| Human TRBC1 | 387 | 442 | 18 | 153 | 108 | 323 | 18 | 384 |
| Human TRBC2 | 387 | 517 | 18 | 144 | 108 | 292 | 24 | 520 |
| Mouse TRBC1 | 375 | 552 | 18 | 98 | 108 | 310 | 18 | 575 |
| Mouse TRBC2 | 375 | 506 | 18 | 145 | 108 | 283 | 18 | n.d. |

The sequence of the last TRBC gene was completed by another PCR experiment with CBEXIdogU/3UTRdogL primer pair derived from the first exon and the 3′UTR of the reference TRBC gene, respectively (clone C2gen11 in Fig. 2).

Additional PCR experiments were performed to minimize any possible PCR bias and/or error, using primer sets as indicated in Section 2.5. The final sequence obtained from six PCR products was assembled and it spans a total of 13539 bp (Fig. 2).

Hence, the analysis revealed the presence of a gap within the current canine TRB assembly, encompassing a second D–J–C cluster. Moreover, the comparison between our data and the public database sequence locates the gap in the genomic assembly downstream the exon 3 of the single TRBC gene (Fig. 2).

The obtained sequence (EMBL ID: HE653929) was inserted between the position 265278 and 278816 of the previously retrieved genomic assembly in order to reconstruct and analyze the entire dog TRB region that spans 333327 pb, MOXD2 and EPHB6 genes included.

Hence, similarly to human and mouse, dog TRB locus consists of two in tandem aligned D–J–C clusters distributed over about 14 kb positioned downstream a large region of TRBV genes, each spans about 7 kb and consists of one TRBD, six TRBJ and one TRBC gene. At less than 11 kb, separated from the last TRBC gene, in an inverted transcriptional orientation, a TRBV gene lies.

The genes of the second D–J–C cluster were annotated as TRBD2, TRBJ2 and TRBC2, followed for the TRBJ2, by a hyphen and a number corresponding to their position in the cluster (Supplementary Table S4). The genes within the two D–J–C clusters were analyzed in detail.

The dog TRBC genes are composed as in all known mammalian species of four exons and three introns. The four exons measure 387, 18, 107 and 18 bp, respectively. The 3′UTR region that is 224 bp long in TRBC1 gene and 205 bp long in TRBC2 gene is also present in the fourth exon. When we compared the nucleotide sequences of TRBC1 and TRBC2 coding region, we observed that they differ for only three nucleotides (nt) (one in EX1, one in EX3 and one in EX4). On the contrary, the 3′UTR differs extensively between the two genes. Minimal length differences were detected in the first (697 bp and 700 bp for TRBC1 and TRBC2 gene, respectively), in the second (157 bp in both cases) and in the third (323 and 295 bp for TRBC1 and TRBC2 gene, respectively) introns. In contrast, the third intron differs significantly in nucleotide composition. Table 1 summarizes the percentage of nucleotide identities for each exon and intron calculated in pair-wise combinations between the two TRBC genes in dog as well as in human and mouse. It can be noticed the higher level of similarity between dog TRBC genes with respect to human and mouse.


**Fig. 3.** IMGT Protein display of the dog, human and mouse TRBC genes. Description of the strands and loops is according to the IMGT unique numbering for C-DOMAIN (Lefranc et al., 2005).

The dog TRBC exon/intron splicing sites are in accordance with human and mouse (IMGT Repertoire, http://www.imgt.org). Splicing frame 1 occurs between L-PART1 and L-PART2, between the V–D–J-EXON and the TRBC1 EX1 as well as between TRBC EX1 and EX2, EX2 and EX3, with at each splicing the classical creation of a codon. Splicing frame 0 occurs between TRBC genes third and fourth exons without the addition of a codon.

The dog TRBC genes encode a similar protein of 176 AA. Particularly, the differences of three nt between them result in two AA changes, one (S15 > P) located at the AB turn of the extracellular C-BETA domain and the other one (V15 > I) in the transmembrane region (TM) (Fig. 3). The C-BETA domain encoded by EX1 is 129 AA long as in human. The C region also comprises the connecting region (CO) of 21 AA (encoded by EX2 and the 5′ part of EX3) with a cysteine involved in the interchain disulfide binding, the TM of 21 AA (encoded by the 3′ part of EX3 and the first codon of EX4) and the cytoplasmic region (CY) of 5 AA (encoded by EX4).

The nucleotide and deduced amino acid sequences of the TRBJ2 together with the previously identified TRBJ1 genes in the region are reported in Fig. 4a. All genes are typically 46–53 bp long and

**(a)**

| TRBJ gene name | J-NONAMER GGTTTTTGT | J-SPACER ********** | J-HEPTAMER CACTGTG | J-REGION | 5'splice donor |
|---|---|---|---|---|---|
| TRBJ1-1 | gtttgcctc | cttgccccatgt | cactgtg | TGAACACTGAAGTGTTCTTCGGAAAAGGCACCAGGCTCACAGTTATAG<br>N T E V F F G K G T R L T V I | gtaaga |
| TRBJ1-2 | cattttaga | gtggccatattc | tgatgtg | TTATGATTTTAACTTTGGCCCAGGGACCAAGCTGACAGTTGTAG<br>Y D F N F G P G T K L T V V | gtaagg |
| TRBJ1-3 | ggttctgaa | gtggatctggga | ggctgtg | CTTTTGAAACACCTTGCACTTTGGGGACGGGAGCCGGCTCACTGTTGTAG<br>F * N T L H F G D G S R L T V V | gtaagt |
| TRBJ1-4 | agttttcct | accagtctgcag | ggttgtg | TAACAATGAAAAATTGTATTTTGCAAGTGGAACCAAGCTTTCTGTCTTGG<br>N N E K L Y F A S G T K L S V L | gtaagt |
| TRBJ1-5 | gggtttgtc | acactcttgctg | tgctgtg | GAACAACCAGGCACAGCACTTTGGAGACGGGACTCGACTCTCTGTCCTGG<br>N N Q A Q H F G D G T R L S V L | gtaagg |
| TRBJ1-6 | ggttttacc | acagctgcctgc | agctgtg | TTTCTATAACTCGCCCCTCTACTTTGGAACAGGCACCAGGCTCACCGTGACAG<br>F Y N S P L Y F G T G T R L T V T | gtatgg |
| TRBJ2-1 | aaattcttg | gcagcccctgcc | cactgtg | CTAATATGGGGAGCAGCACTTTGGGCCAGGGACCCGGCTCACCGTCCTAG<br>* Y G E Q H F G P G T R L T V L | gtaaga |
| TRBJ2-2 | ggtttgcgc | ctgggccccctg | agctgtg | CCAACACGGGGCAGCTGTACTTCGGGGCCGGTTCCAAGCTGGCCGTGCTGG<br>N T G Q L Y F G A G S K L A V L | gtaagg |
| TRBJ2-3 | tttttttgt | gctggggctccg | ggctgtg | AGCACAGAAACTCAGTATTTCGGCGGGGGCACCCGGCTGACCGTGCTCG<br>S T E T Q Y F G G G T R L T V L | gtaagg |
| TRBJ2-4 | tgctgcgca | gggcgggggggc | ggctgtg | AGCCTGGACACCCAGTACTTCGGGGCGGGCACCCGGCTGACCGTGCTCG<br>S L D T Q Y F G A G T R L T V L | gtaatc |
| TRBJ2-5 | tgtttctgt | gctgcgcctggg | ggctgtg | AGCCAAAATACCCAGTACTTCGGGGCGGGCACCCGGCTGACCGTGCTAG<br>S Q N T Q Y F G A G T R L T V L | gtaagc |
| TRBJ2-6 | ggtttgcgc | gcggggctgggc | ctctgtg | CGCTACGAGCAGTATTTCGGCGCCGGCACCAGGCTCACGGTCCTCG<br>R Y E Q Y F G A G T R L T V L | gtgaga |

**(b)**

| TRBD gene name | 5′D-NONAMER GGTTTTTGT | 5′D-SPACER ********** | 5′D-HEPTAMER CACTGTG | D-REGION | 3′D-HEPTAMER CACAGTG | 3′D-SPACER ********** | 3′D-NONAMER ACAAAAACC |
|---|---|---|---|---|---|---|---|
| TRBD1 | ggtttttgt | acaaagctgtaa | cattgtg | GGTACAGGGGGC<br>G T G G<br>V Q G<br>Y R G | cacggtg | attcaactctatgggaaagcttt | acaaaaacc |
| TRBD2 | catttttgt | atcctgctgtaa | cattgtg | GGAACTGGGGGACT<br>G T G G<br>E L G D<br>N W G T | caccatg | attcaggtggaggggtccttttt | acaaaaagc |

**Fig. 4.** Nucleotide and deduced amino acid sequences of the dog TRBJ (a) and TRBD (b) genes. The consensus sequences of the heptamer and nonamer (Hesse et al., 1989) are provided at the top of the figure and underlined. The numbering adopted for the gene classification is reported on the left of each gene. In (a), the donor splice site for each TRBJ is shown. The canonical FGXG amino acid motifs are underlined. The unusual TRBJ1-4 gene motif is underlined in red. In (b), the inferred amino acid sequences of the TRBD genes in the three coding frames are reported. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conserve the canonical FGXG amino acid motif, whose presence defines the functionality of the J genes. The only exception is the TRBJ1-4 gene, where the motif is slightly altered. This gene was identified within three cDNA clones (5R29, 5R45 and 5R5). It is noteworthy that the TRBJ2-1 is a pseudogene owing due to the presence of a STOP-CODON at its 5′ end. Each TRBJ gene is flanked by a 12-spacer RS at its 5′ end and has a donor splice at its 3′ end. Most of the positions of the heptamer as well as the nonamer match with the canonical consensus sequence (Hesse et al., 1989; Akamatsu et al., 1994; Lefranc and Lefranc, 2001). In particular, the last three nt of the J-HEPTAMER, gtg, which flank the J-REGION, are highly conserved. These three nt have been demonstrated to be essential for RS recognition, and changes in those positions preclude rearrangement. Within the J-NONAMER, most individual positions are less conserved. However, a stretch of three 't', that appear to be a key feature of the sequence, is almost perfectly conserved, with the exception of the TRBJ2-4 gene.

Fig. 4b shows the nucleotide and deduced amino acid sequences of the two TRBD genes. They consist of a 12 bp (TRBD1) or 14 bp (TRBD2) g-rich stretch that can be productively read in its three coding phases and encoded 1–3 glycine (G), depending on the phase. The first coding phase is common between the two TRBD genes. The RS that flank the 5′ and 3′ sides of the D-REGION are well conserved.

### 3.4. Analysis of the beta chain transcripts

The availability of the updated genomic sequence of the dog TRB locus allowed us to pinpoint the V–D–J–C arrangements within the cDNA clones. Fig. 6 shows the deduced amino acid sequences of the V–D–J regions of all 36 cDNA clones according to the IMGT unique numbering for the V-REGION and V-DOMAIN (Lefranc et al., 2003). All TR β transcripts exhibit typical features, i.e. a leader region of 14–20 AA, the conserved AA CYS 23 in FR1, CYS 104 in FR3, and TRP 41 in FR2. As mentioned above, the TRBV gene subgroups were identified by IMGT/V-QUEST program, and the new TRBV alleles were labeled by consecutive numbers with respect to the reference germline gene sequence (marked as *01).

All TRBJ regions within the set of cDNA sequences were assigned unambiguously to the distinct germline genes derived from the reference sequence NW_876260[1] (for TRBJ1 genes) and EMBL ID: HE653929 (for TRBJ2 genes). As in the case of the TRBV genes, we identified new alleles, that were labeled by consecutive numbers with respect to the germline allele *01 (Fig. 5).

Although the number is too low to be statistical relevant, the TRBV16 gene seems to be preferentially used (8/36 = 22%) and a slight increase in the use of the TRBJ2-5 (5/36 = 14%) and the TRBJ2-6 (7/36 = 19%) genes can be observed (Fig. 5). Moreover, the analysis of associated V–D–J region and C region in transcripts allowed us to establish that the proline (P15) of the genomic TRBC2 sequence (Fig. 3) is a polymorphism since the serine is present in all cDNA containing TRBJ2 genes and presumably the TRBC2 gene.

More complex is the determination and contribution of the TRBD genes involved in the CDR3 formation. The rearranged CDR3-IMGT extends from codon 105 (codon following the 2nd-CYS codon 104 of the V-REGION) to codon 117 (codon preceding the J-PHE 118) – or, in other loci, J-TRP 118 – that belongs to the F/W-G-X-G motif characteristic of the J-REGION (IMGT Repertoire, http://www.imgt.org, Lefranc et al., 2003; Lefranc, 2011a). The mean length of the CDR3-IMGT was approximately the same in the dog (mean 12.9 AA, range 8–20 AA) (Fig. 5) with respect to human (mean 12.7 AA; Hall and Lanchbury, 1995) and the mouse (mean 11.9 AA; Rock et al., 1994) peripheral blood. For a close inspection of the CDR3, the nucleotide sequences have been excised from each cDNA sequence and analyzed in detail (Fig. 6). TRBD genes sustain substantial base deletion and overall transformation, so that the genes are usually unrecognizable without ambiguity. By the comparison of the TRBD genomic sequences, the nt located in the CDR3 regions were considered to belong to a TRBD gene if they constituted a stretch of at least six consecutive nt (50% of the TRBD1 germline nt length) corresponding to the TRBD1 or TRBD2 germline sequences. In this way the TRBD was identified unambiguously and accurately in 14 out of the 36 sequences, with the TRBD1 (Fig. 6a) present in 10 (71.4%) and the TRBD2 (Fig. 6b) in 4 (28.6%) clones. Particular features of the CDR3 region of the clones are the presence within the D region

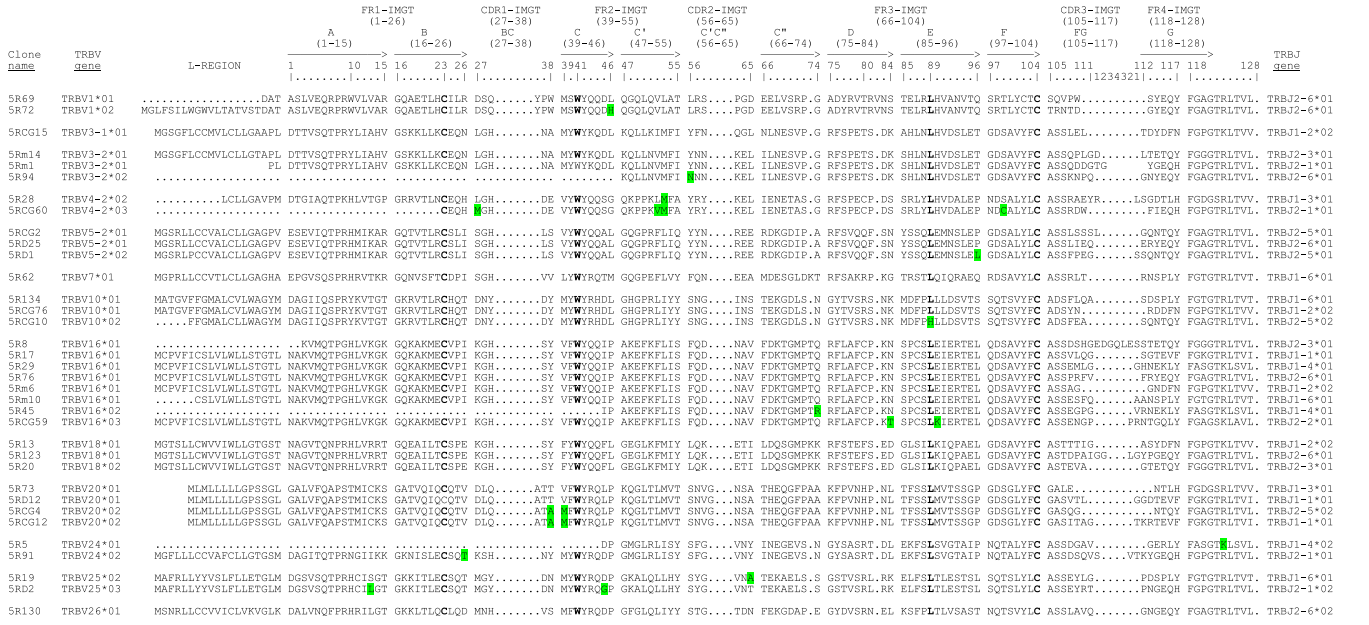


**Fig. 5.** Protein display of the cDNA clones. The TRBV and TRBJ genes, named in accordance with the criteria specified in the text, are listed respectively at the left and the right of the figure. Leader region (L-REGION), complementarity determining regions (CDR-IMGT) and framework regions (FR-IMGT) are also indicated, according to the IMGT unique numbering for V-DOMAIN (Lefranc et al., 2003). The five conserved AA of the V-DOMAIN (1st-CYS 23, CONSERVED-TRP 41, hydrophobic AA 89, 2nd-CYS 104 and J-PHE 118) are indicated in bold. The TRBV allele amino acid changes, if any, are boxed. The name of the clones is also reported.

**a**

| CLONE | TRBV | -nt V | TRBD1 ggtacaggggc<br>CDR3 SEQUENCE | -nt J | TRBJ | CDR3 nt/AA |
|---|---|---|---|---|---|---|
| 5R17 | TRBV16*01 | -4 | GCC AGC AGT Gta c**ta cag gg**a tcg ggt ACT GAA GTG TTC | -5 | TRBJ1-1 | 39/13 |
| 5R28 | TRBV4-2*02 | -1 | GCT AGC AGC AGG Gct gaa **tac agg** ctt agt ggg gac ACC TTG CAC | -10 | TRBJ1-3 | 45/15 |
| 5R29 | TRBV16*01 | -2 | GCC AGC AGT GAA atg tt**a ggg gg**g cAC AAT GAA AAA TTG TAT | -2 | TRBJ1-4 | 42/14 |
| 5R94 | TRBV3-2*02 | -5 | GCC AGC AGC aaa aat ccg **cag ggg** aaC TAC GAG CAG TAT | -2 | TRBJ2-6 | 39/13 |
| 5R130 | TRBV26*01 | -4 | GCC AGC AGT ctc gcg **gta cag gg**a aat ggC GAG CAG TAT | -5 | TRBJ2-6 | 39/13 |
| 5Rm1 | TRBV3-2*01 | 0 | GCC AGC AGC CAA GAc gac **ggt aca ggg** TAT GGG GAG CAG CAC | -4 | TRBJ2-1 | 42/14 |
| 5Rm6 | TRBV16*01 | -6 | GCC AGC AGc g**ca ggg gg**a aAT GAT TTT AAC | -2 | TRBJ1-2 | 30/10 |
| 5Rm10 | TRBV16*01 | 0 | GCC AGC AGT GAA TCa t**tt cag g**ct gcg AAC TCG CCC CTC TAC | -7 | TRBJ1-6 | 42/14 |
| 5RCG12 | TRBV20*02 | -2 | GGT GCT AGT at**t acg gcg gg**g act aaa cgC ACT GAA GTG TTC | -4 | TRBJ1-1 | 42/14 |
| 5RCG76 | TRBV10*01 | -2 | GCC GAC AGT TAC a**ac agg g**AT GAT TTT AAC | -2 | TRBJ1-2 | 30/10 |

**b**

| CLONE | TRBV | -nt V | TRBD2 ggaactgggggact<br>CDR3 SEQUENCE | -nt J | TRBJ | CDR3 nt/AA |
|---|---|---|---|---|---|---|
| 5R8 | TRBV16*01 | -3 | GCC AGC AGT GAc agc cat gga gaa gac ggc c**aa ctg g**ag tct AGC ACA GAA ACT CAG TAT | 0 | TRBJ2-3 | 60/20 |
| 5R123 | TRBV18*01 | -3 | GCC AGC ACC GAc cca gcc at**t ggg gga ct**a gga tac cct ggc GAG CAG TAT | -6 | TRBJ2-6 | 51/17 |
| 5Rm14 | TRBV3-2*01 | -2 | GCC AGC AGC CAA cct **ctg ggg gac** ctC ACA GAA ACT CAG TAT | -2 | TRBJ2-3 | 42/14 |
| 5RD12 | TRBV20*01 | -3 | GGT GCT AGc gtc act t**tg ggg g**gg gaC ACT GAA GTG TTC | -4 | TRBJ1-1 | 39/13 |

**c**

| CLONE | TRBV | -nt V | TRBD?<br>CDR3 SEQUENCE | -nt J | TRBJ | CDR3 nt/AA |
|---|---|---|---|---|---|---|
| 5R5 | TRBV24*01 | -2 | GCC AGC AGC GAc ggc gcg gtg ggc GAA AGA TTG TAT | -7 | TRBJ1-4 | 36/12 |
| 5R13 | TRBV18*01 | -6 | GCC AGC ACt aca acg atc gga gcg tca TAT GAT TTT AAC | -1 | TRBJ1-2 | 39/13 |
| 5R19 | TRBV25*02 | 0 | GCC AGC AGT GAA TAc ctg gga ccg gAC TCG CCC CTC TAC | -8 | TRBJ1-6 | 39/13 |
| 5R20 | TRBV18*02 | -1 | GCC AGC ACC GAA Gtc gct gGC ACA GAA ACT CAG TAT | -1 | TRBJ2-3 | 36/12 |
| 5R45 | TRBV16*02 | -2 | GCC AGC AGT GAA ggt ccg ggg gta cgg AAT GAA AAA TTG TAT | -4 | TRBJ1-4 | 42/14 |
| 5R62 | TRBV7*01 | -4 | GCC AGC AGT Cgc cta act cga AAC TCG CCC CTC TAC | -7 | TRBJ1-6 | 36/12 |
| 5R69 | TRBV1*01 | -6 | AGc cag gta c**c**a tgg aGC TAC GAG CAG TAT | -1 | TRBJ2-6 | 30/10 |
| 5R72 | TRBV1*02 | -11 | ACC cga aac acg gac gGC TAC GAG CAG TAT | -1 | TRBJ2-6 | 30/10 |
| 5R73 | TRBV20*01 | -6 | GGT GCc ctg gaa AAC ACC TTG CAC | -7 | TRBJ1-3 | 24/8 |
| 5R76 | TRBV10*01 | -6 | GCC AGC AGc cca agg ttt gtc ttc CGC TAC GAG CAG TAT | 0 | TRBJ2-6 | 39/13 |
| 5R88 | TRBV28*01 | -7 | GCC AGT ttg ggg cgt ggt ttg gAT ACC CAG TAC | -7 | TRBJ2-5 | 33/11 |
| 5R91 | TRBV24*02 | 0 | GCC AGC AGC GAT Tct cag gtt tca gtc acg aAA TAT GGG GAG CAG CAC | -2 | TRBJ2-1 | 48/16 |
| 5R134 | TRBV10*01 | -4 | GCC GAC AGT Ttt ctg cag gct tcc gAC TCG CCC CTC TAC | -8 | TRBJ1-6 | 39/13 |
| 5RCG2 | TRBV5-2*01 | -4 | GCC AGC AGC ctt tcg agc agt ctc gGC CAA AAT ACC CAG TAC | -1 | TRBJ2-5 | 42/14 |
| 5RCG4 | TRBV20*02 | 0 | GGT GCT AGT CAg ggA AAT ACC CAG TAC | -5 | TRBJ2-5 | 27/9 |
| 5RCG10 | TRBV10*02 | -6 | GCC GAC AGc ttc gag gct AGC CAA AAT ACC CAG TAC | 0 | TRBJ2-5 | 36/12 |
| 5RCG15 | TRBV3-1*01 | 0 | GCC AGC AGC CTA GAa ctg acg gaT TAT GAT TTT AAC | 0 | TRBJ1-2 | 36/12 |
| 5RCG59 | TRBV16*03 | -2 | GCC AGC AGT GAA aac ggc ccc cct aga AAC ACG GGG CAG CTG TAC | -2 | TRBJ2-2 | 45/15 |
| 5RCG60 | TRBV4-2*03 | 0 | GCT AGC AGC AGG GAc tgg ttt att GAG CAG CAC | -10 | TRBJ2-1 | 33/11 |
| 5RD1 | TRBV5-2*02 | -2 | GCC AGC AGC TTc ccg gag gga tct AGC CAA AAT ACC CAG TAC | 0 | TRBJ2-5 | 42/14 |
| 5RD2 | TRBV25*03 | 0 | GCC AGC AGT GAA TAc agg aca ccg aAT GGG GAG CAG CAC | -5 | TRBJ2-1 | 39/13 |
| 5RD25 | TRBV5-2*02 | -1 | GCC AGC AGC TTG att gaa cag gag CGC TAC GAG CAG TAT | 0 | TRBJ2-6 | 39/13 |

**Fig. 6.** CDR3 nucleotide sequences retrieved from the cDNA clones. CDR3-IMGT sequences are shown from codon 105 (codon after the 2nd-CYS 104 of the V-REGION) to codon 117 (codon before J-PHE 118 of the J-REGION) according to the IMGT unique numbering (Lefranc et al., 2003). They are grouped on the basis of the TRBD1 (a), TRBD2 (b) usage or no TRBD usage (c). The CDR3 nt/AA length, 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 (Bleakley et al., 2006) are indicated in upper cases. The germline region of the TRBD1 and TRBD2 gene are indicated at the top of each figure. The sequences considered to present recognizable TRBD genes (see Section 3.4) are indicated in bold and lower cases and nucleotide substitutions are underlined. Nucleotides that cannot be attributed to any V, D or J regions (N-nucleotides) are indicated in lower cases on the left and on the right sides of the TRBD regions. Numbers in the left and right columns indicate the number of nt that are trimmed from the 3′V-REGION and 5′J-REGION, respectively.

of nt substitutions (5Rm10 and 5RCG12 in Fig. 6a) with respect to the germline sequences.

The remaining 22 sequences (61.1%), with no recognizable TRBD genes, were grouped separately (Fig. 6c). These last sequences could be interpreted as direct V–J junctions. However, it is also possible that nucleotide trimming masked the initial participation of D gene during the rearrangement.

The comparison with the germline sequences allowed us also to examine the trimming activity of the exonuclease during rearrangement at the 3′-V and 5′-J boundaries of the cDNA and, as a

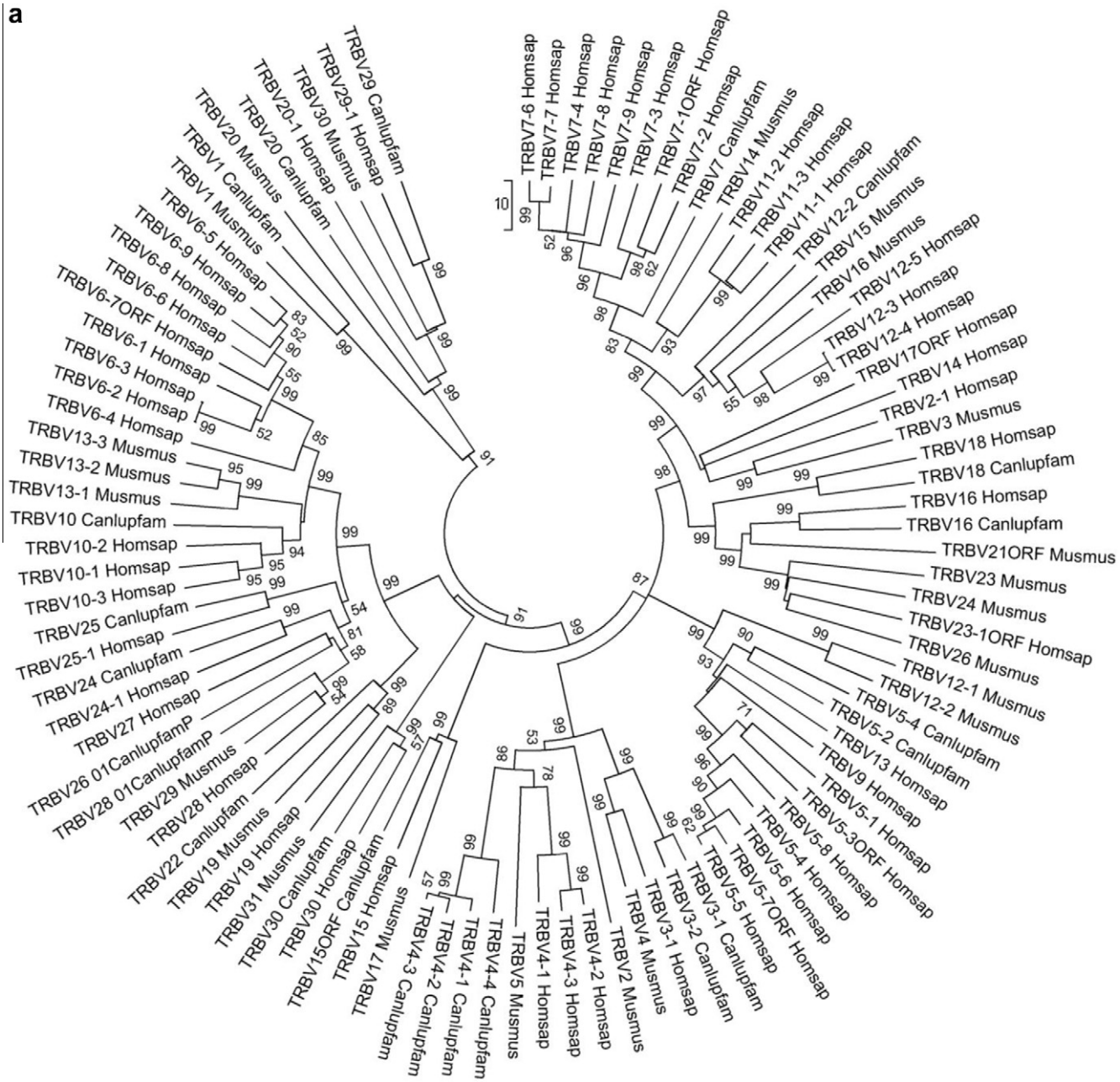


**Fig. 7.** Phylogenetic trees of the dog, human and mouse TRBV (a), TRBJ (b) and TRBC (c) genes. The unrooted trees, inferred from nucleotide sequences, were reconstructed using NJ (a, b) and MP (c) methods. Numbers along the branches show the percentage occurrence of nodes in 100 replicates of simulation. The gene nomenclature is according to IMGT-ONTOLOGY concepts of classification (Lefranc, 2011b). *Canis lupus familiaris* TRBV, TRBD, TRBJ and TRBC gene names have been approved by the IMGT nomenclature committee. The IMGT standardized abbreviation for taxon is used: six letters for species (Musmus, Homsap) and nine letters for subspecies (Canlupfam).

consequence, the length of the (N-D)-REGION. The degree of germline nucleotide trimming is similar in all groups, with the mean value of 6.9, 5.8 and 6.7 nt respectively for the first, the second and the third cDNA group (Fig. 6). Nevertheless, the degree of germline nucleotide trimming is different when comparing cDNA derived from animals suffering from visceral leishmaniasis (5Rm, 5RCG and 5RD clones) with respect to the healthy individual (5R clones), with 4.9 nt versus 8.0 nt mean value, respectively. Similarly, the mean length for (N-D)-REGION is 12.2 nt for the sick and 16.4 nt for the healthy individuals. However, it is noteworthy that the mean length of the CDR3 in AA is comparable when calculated separately in the healthy (series 5R clones) with respect to the sick individuals (series 5Rm, 5RCG and 5RD clones) (mean 13.1 AA versus 12.6 AA).

Finally, although the TRBD gene was unambiguously recognizable in only 39% of the cDNA clones, the interpretation of these rearrangements revealed that the intracluster rearrangements represent the substantial portion of the repertoire (71%), with 7 TRBD1–TRBJ1 and 3 TRBD2–TRBJ2 rearrangements (Fig. 6a and b). Three cDNA clones show an intercluster rearrangement by direct 5′- to 3′ joining across TRBD1–TRBJ2 clusters (5R94, 5R130 and 5Rm1 clones). Interestingly, we also observed one TRBD2-TRBJ1 joining (5RD12 clone in Fig. 6b) that may only be explained by chromosomal inversions, or with more

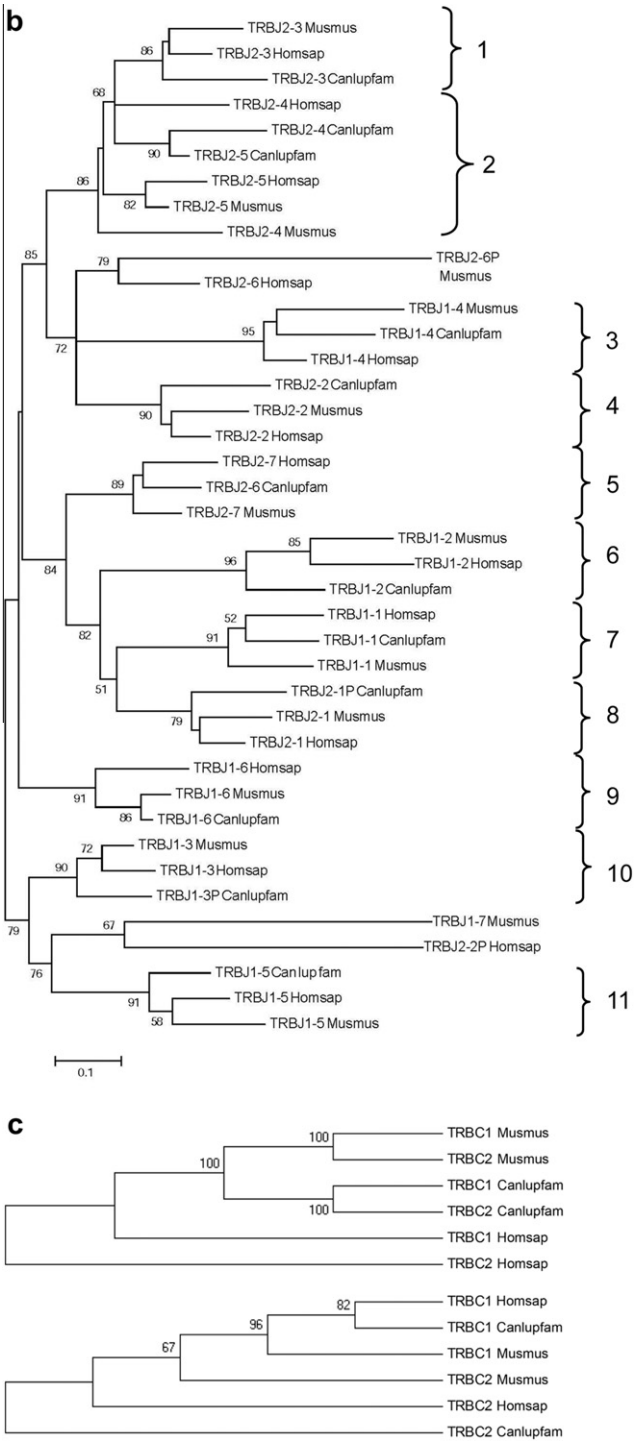


**Fig. 7** (*continued*)

probability, by trans-rearrangement occurring during TRB locus recombination.

### 3.5. Identification of the regulatory elements

The comparison of the dog TRB genomic sequence with the human and mouse counterparts allowed us to interpret two regions as regulatory elements on the basis of their evolutionary constraint.

The first of these regions, spanning from 94 bp upstream through 60 bp downstream of the TRBD1 gene (coordinate 6751291–6751456 in NW_876260[1] dog assembly), displays 86% and 87% of nucleotide identity with the mouse and human corresponding region, respectively. This region has been experimentally determined in the mouse as part of the TRBD1 promoter (PDβ1), which together with the beta enhancer element (Eβ) regulate the tissue and stage specific assembly of the TRB genes (Carvajal and Sen, 2000).

The tree-species alignment (Supplementary Fig. S2a) shows that the transcription factor binding site for Ikaros/Lyf-1 at the −35 GC-rich motif, demonstrated to be essential for the murine promoter activity, is well conserved, whereas others such as the TATA box and the GATA binding sequence at −74 are absent from both dog and human sequences.

Intriguingly, we also found a CpG site in the dog heptamer of the 23-spacer RS shared with mice but not with humans. Analysis of methylation levels surrounding RS before, during and after recombination have demonstrated that demethylation of this CpG site appears to be essential for TRBD1 cleavage by the V–(D)–J recombinase in mouse (Whitehurst et al., 2000).

The second identity region was identified in a 167 bp sequence located 6441 bp from the TRBC2 gene (coordinate 6737925–6738083 in NW_876260[1] dog assembly) and it displays 85% and 89% of nucleotide identity respectively with the mouse and human sequences (Supplementary Fig. S2b). This region corresponds to the Eβ element fully characterized in mice and humans (Tripathi et al., 2000; Gottschalk and Leiden, 1990).

Among the protein binding motifs termed βE1-βE7 in mouse and Tβ1-Tβ5 in human enhancer, the most consistent conservation was found in the 98 bp core-binding domain comprising the well-defined nuclear factor binding βE3 and βE4/Tβ3 motifs and the 3′-flanking sequence at 3′end of mouse βE4. This core-binding motif contains a bHLH ligand E box motif (in βE3) and the Ets-CBF composite motif (in βE4/Tb3), both of which are commonly found within the enhancers of antigen receptor genes. Another typical E-box and a second Ets/CBF interacting element are present in 3′- flanking and βE6 motif, respectively. A conserved motif outside the core-binding fragment is the βE1 motif, which was proposed to represent the main GATA-3 factor binding site within the enhancer. Overall similarity in βE5 motif is less conserved.

### 3.6. Phylogenetic analysis of the dog TRBV, TRBJ and TRBC genes

The evolutionary relationship of the dog TRBV, TRBJ and TRBC genes was investigated by comparing them with all available human and mouse counterparts. First, we combined the coding region nucleotide sequences of all functional TRBV genes from dogs, humans and mice in the same alignment and a phylogenetic tree was made using neighbour-joining (NJ) method (Fig. 7a). The dog TRBV26 and TRBV28 pseudogenes were also included since they still can originate productive transcripts (Fig. 5). The extensive intermingling of the dog, human and mouse TRBV subgroups is consistent with the occurrence of distinct subgroups prior to mammalian radiation. The phylogenetic groupings were supported by high bootstrap values. Conversely, the phylogenetic clustering of orthologous multimember subgroups from different species indicates that duplications within subgroups have occurred post-speciation. As expected, the dog and mouse TRBV1 are grouped together and lack a human orthologue.

The same phylogenetic analysis was also performed on the dog TRBJ genes. In this case the nucleotide sequences of all dog germline TRBJ genes (coding region plus RS), with the homologous human and mouse, were included in the alignment. The alignment was then used to construct a phylogenetic tree by using NJ method (Fig. 7b). The results of the analysis support a distribution of the

dog TRBJ genes into 11 distinct phylogenetic groups, each composed of orthologues as indicated by their conserved germline position with respect to that of the constant genes within the corresponding TRB locus (IMGT Repertoire, http://www.imgt.org). The only exception is the closer phylogenetic relationship of the dog TRBJ2-4 and TRBJ2-5 genes with respect to the human and mouse orthologues. Moreover, a TRBJ2-6 orthologous lacks in the dog locus.

A phylogenetic tree of the coding region of the TRBC genes of dogs, humans and mice was also constructed using the maximum parsimony (MP) method (upper part of Fig. 7c). The tree resolved the TRBC genes, differently from the TRBV and TRBJ genes, into species-specific branches thus related with the evolution of each species. We performed a further phylogenetic analysis based on the nucleotide sequences of the 3′ UTR portions of the TRBC genes. All the sequence portions from the stop codon to the poly-A site were aligned and a tree was constructed by using the same method (lower part of Fig. 7c). In the tree, significant bootstrap values support a clear-cut subdivision according to the different TRBC isotype. This different branching order with respect to the above tree is in accordance with previous studies (Antonacci et al., 2008) that revealed a substantially different mechanism of evolution that occurred on the coding and non-coding regions of the mammalian TRBC genes.

## 4. Discussion

The present study clearly refines the organization of the dog TRB locus and contributes to gain a better understanding of the canine TRB gene repertoire and evolution. The most striking finding was the annotation of two additional TRBV genes (Supplementary Table S3), and the identification of a second D–J–C cluster (Fig. 2), missing in the genome assembly, bearing the canine TRB locus to more resemble the genomic structure of most of mammalian TRB loci (i.e. human and mouse TRB) (IMGT Repertoire, http://www.imgt.org; Williams et al., 1991; Jaeger et al., 1993; Cheynier et al., 1996; Schrenzel et al., 1994; Bosc and Lefranc, 2000; Lefranc and Lefranc, 2001). All together our results and sequence data previously published (Lindblad-Toh et al., 2005; IMGT Repertoire, http://www.imgt.org), indicate that the dog TRB locus consists of a pool of 37 TRBV genes, grouped in 25 subgroups, positioned upstream of two in tandem aligned D–J–C gene clusters, each composed of a single TRBD, six TRBJ and one TRBC genes, followed by a single TRBV gene with an inverted transcriptional orientation located at the 3′ end (Fig. 8). Also the regulatory architecture of the dog TRB locus, with the PDβ1 at the 5′ of D–J–C cluster 1 and the Eβ element downstream the TRBC2 gene, seems to be preserved. The conservation pattern of motifs (Supplementary Fig. S2),

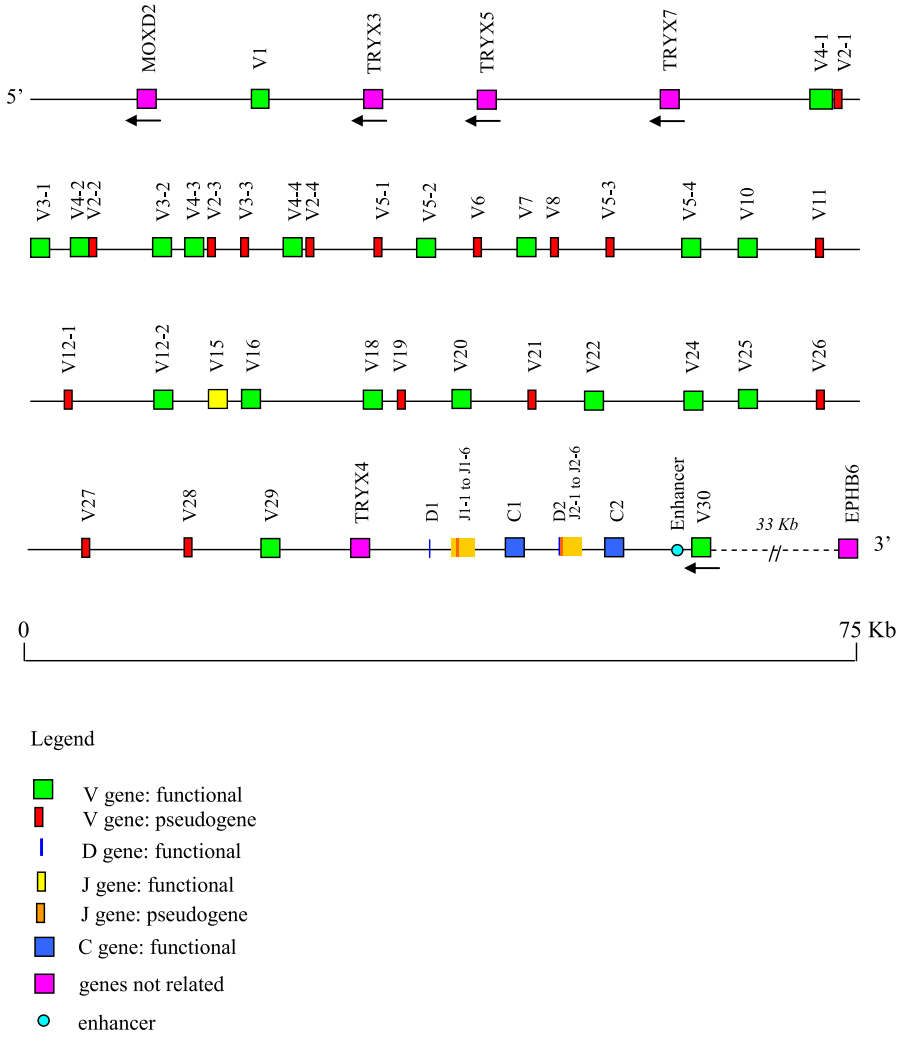


**Fig. 8.** Schematic representation of the genomic organization of the dog TRB locus as deduced from the genome assembly Build 2.1 (gaps excluded) and this work. The diagram shows the position of all V, J and C TRB genes according to IMGT nomenclature. Boxes representing genes are not to scale. Exons are not shown.

especially within the Eβ core, across the genomes, reflects their functions in the TR chain biosynthesis.

Similarly to the other mammals, the MOXD2 and EPHB6 genes flank, respectively, the 5′ and 3′ ends of the TRB locus. Moreover, four TRY genes are intercalated with the TRB genes, three (TRYX3, TRYX5, and TRY7) are located at the 5′ end of the locus, downstream the TRBV1 gene, and a single TRY (TRY-PA) is intermixed between the last TRBV gene and the first D–J–C cluster (Fig. 8).

The canine TRB locus spans less than 300 kb making it smaller in size than its human and mouse homologues that are about 650 kb each (Rowen et al., 1996; Glusman et al., 2001). The lower density of total interspersed repeats compared to human and mouse corresponding regions may have been affected the size of the dog TRB locus (Supplementary Fig. S1; Supplementary Table S2). This is in line with comparative analysis data according to which the smaller size of the dog genome is primarily due to the presence of a lower proportion of total repetitive elements than human and mouse (Lindblad-Toh et al., 2005). Moreover, the percentage of repetitive elements within the dog TRB locus (24.92%) is even lower if compared with the entire genome (35.15%) mainly due to the minor presence of SINEs (7.19% versus 10.57%) as well as LINEs (11.98% versus 18.74%). On the contrary, the analysis previously carried out on the dog TRG locus (Massari et al., 2009), shows that the density of total interspersed repeats (32.94%, of which 8.59% are SINEs and 20.59% are LINEs) is well correlated to that of the overall genome.

However, despite their small proportion, both LINEs and SINEs, contribute to the dog TRB structure for their prevalence and distribution throughout the entire region (Supplementary Fig. S1). In this regard, it is noteworthy in the dot-plot matrix that there is a repetitive pattern distributed along the TRB locus of tRNA-SINEs which have been found with a high frequency and scattering in the genome of several species of the Canoidea, including the domestic dog (Bentolila et al., 1999).

More properly, the size of the TRB loci can be correlated with the extent and complexity of gene duplications. The human and mouse loci (GenBank and IMGT/LIGM-DB ID: L36092 and AE000663-664-665), with the same size, have about the same number of TRBV subgroups (32 and 31, respectively); but in mice, only two subgroups have multiple members with three genes for each, embedded in a single homology unit. Conversely, an extensive duplication involves the mouse TRY gene family with 20 genes. In human, TRB locus consists of nine multimember TRBV subgroups with different number of genes for each, incorporated in six homology units with a total of 67 TRBV genes. Eight TRY genes were identified in this locus.

Even more complex and extensive gene duplications characterize the bovine TRB locus, extending beyond 700 kb (Connelley et al., 2009). The sequence contains 134 TRBV genes distributed in 24 subgroups. 11 subgroups are multimembers with a massive expansion of some of them up to 40 members. Only five TRY genes were identified.

The genomic analysis of the dog TRB locus indicates that very few gene expansions have occurred in it since only five out of 25 subgroups are multimembers with a limited number of genes (from 2 to 3). The most remarkable duplication event occurred within the region in which TRBV2, TRBV3 and TRBV4 genes are located (Fig. 1). As generally occurs, the members of these expanded subgroups are intercalated with each other in a recurrent pattern composed of alternating TRBV2, TRBV3 and TRBV4 genes. Only four dog TRY genes were found. Therefore the duplication events that increase the germline repertoire and undertake more specialized functions are reduced in dogs.

Despite the substantial disparity in the number, the phylogenetic clustering of the TRBV genes between the different species reveals that the birth of the distinct TRBV subgroups occurred prior to the radiation of mammals, while duplications within subgroups in dog as well as in human and mouse have occurred post-speciation from common progenitors of different subgroups (Fig. 7a). Therefore, gene duplication followed by subsequent diversification is the major mode of evolution of the TRBV genes in the different species. Interestingly, the dog TRB locus shares with mouse and probably bovine (Connelley et al., 2009) a TRBV gene that is functional at the locus 5′ end, whereas there is a pseudogene in the human locus.

In addition, the dog germline repertoire seems to be further reduced considering the higher number of V pseudogenes (16 of the 37 TRBV, 43%), compared with human (19%) and mouse (34%).

In contrast to TRBV genes, the organization of the dog TRBD, TRBJ and TRBC genes in two tandem cassettes is well conserved. Except for two (TRBJ1-3 and TRBJ2-1), all the TRBJ genes appear to be functional and our phylogenetic studies showed a strict homology among TRBJ genes with a conserved germline position within its own cluster in the different TRB loci (Fig. 7b; IMGT Repertoire, http://www.imgt.org).

Unlike the TR V-BETA domain (TRBV–TRBD–TRBJ region) that binds to diverse peptide-MH complexes, the C-BETA domain interacts with components of the CD3, which are non-polymorphic. Consequently, due to structural restrictions, TRBC genes are subject to concerted evolutionary pressures with intra-species homogenization (Glusman et al., 2001; Rudikoff et al., 1992; Antonacci et al., 2008). Also dog TRBC genes were found to encode nearly identical products (only two AA are different) (Fig. 3; Table 1) likely because of minimal divergence after duplication. Conversely, a strong selective pressure preserved the sequence of the 3′ UTR of each TRBC gene in dogs just as in all mammalian species (Fig. 7c; Antonacci et al., 2008).

To evaluate the effect of the contraction in the germline TRBV collection on the dog beta chain repertoire we performed 5′ RACE experiments on the peripheral blood from three diverse animals in different state of health. In our cDNA collection (Fig. 5), we found that all TRBV and TRBJ genes within cDNAs were identified in the dog germline sequence. In addition 20 out of 36 clones (55%) retain a TRBJ2 gene suggesting that the new identified cluster is relevant for the production of the beta chain transcripts. Interestingly, our cDNA show that some pseudogenes, both TRBV and TRBJ, contribute to generating the beta chain repertoire. These pseudogenes have a single stop codon that is deleted during the somatic rearrangement. We also observed a high rate of TRBV and TRBJ allelic polymorphism consistent with the analysis conducted on the dog genome by which it contains a limited gene pool with a significant heterogeneity responsible to the enormous phenotypic diversity within breeds (Lindblad-Toh et al., 2005). Therefore we speculate that the usage of pseudogenes and the high number of alleles (39% of TRBV and 25% of TRBJ genes) would compensate for the decrease of the germline repertoire ensuring diversity, as is the case for other species.

The present knowledge on the genomic sequence of the dog TRB locus has also allowed us a detailed analysis of the CDR3 (Fig. 6). The CDR3 is heterogeneous for amino acid composition and its mean length is approximately the same in dog with respect to human and mouse. The CDR3 loop heterogeneity, besides the random somatic rearrangements of TRB genes, is further enhanced by the trimming at the 3′-V and 5′-J gene boundaries and the addition of N-nucleotides at the V–D and D–J junctions.

These processes appear to be regulated in order to generate a CDR3 of appropriate length, indicating that the CDR3 length in TR beta chain is essential for T cell function. In fact, when we analyze separately the cDNA clones from the healthy individual, we observe that the conservation of the CDR3 length was achieved by a greater deletion at the 3′-V and 5′-J gene boundaries and a

concomitant increase in N-nucleotide addition at the V–D and D–J junctions during rearrangement; on the other hand, a decreased trimming activity of the exonuclease as well as the addition of a smaller number of N-nucleotides by the TdT seeming to maintain an appropriate length of the CDR3 in animals suffering from visceral leishmaniasis. Thus, the junctional diversity due to exonuclease and TdT during rearrangement may work differently and depending on the state of the animal. Additional studies in healthy and sick dogs are necessary to support these findings and to know the functional significance.

The available of the complete sequence of the dog TRB locus allows researchers to concentrate on functional studies without the uncertainly deriving from a partial knowledge of the genomic repertoire of the TRB genes and their organization.

## Acknowledgements

This work was supported by funding from Fondazione Cassa di Risparmio di Puglia (Research Project entitled "*Immunogenetica applicata al controllo della Leishmania infantum in Puglia*"). The financial support of the University of Bari, University of Salento is also gratefully acknowledged.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.dci.2012.03.010.

## References

Akamatsu, Y., Tsurushita, N., Nagawa, F., Matsuoka, M., Okazaki, K., Imai, M., Sakano, H., 1994. Essential residues in V(D)J recombination signals. J. Immunol. 153, 4520–4529.

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

Arden, B., Clark, S.P., Kabelitz, D., Mak, T.W., 1995a. Mouse T-cell receptor variable gene segment families. Immunogenetics 42, 501–530.

Arden, B., Clark, S.P., Kabelitz, D., Mak, T.W., 1995b. Human T-cell receptor variable gene segment families. Immunogenetics 42, 455–500.

Bentolila, S., Bach, J.M., Kessler, J.L., Bordelais, I., Cruaud, C., Weissenbach, J., Panthier, J.J., 1999. Analysis of major repetitive DNA sequences in the dog (*Canis familiaris*) genome. Mamm. Genome 10, 699–705.

Bleakley, K., Giudicelli, V., Wu, Y., Lefranc, M.-P., Biau, G., 2006. IMGT standardization for statistical analyses of T cell receptor junctions: the TRAV-TRAJ example. In Silico Biol. 6, 573–588.

Bosc, N., Lefranc, M.-P., 2000. The mouse (*Mus musculus*) T cell receptor beta variable (TRBV), diversity (TRBD) and joining (TRBJ) genes. Exp. Clin. Immunogenet. 17, 216–228.

Brochet, X., Lefranc, M.-P., Giudicelli, V., 2008. 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. 36, W503–508.

Cheynier, R., Henrichwark, S., Wain-Hobson, S., 1996. Sequence of the rhesus monkey T-cell receptor beta chain diversity and joining loci. Immunogenetics 43, 83–87.

Connelley, T., Aerts, J., Law, A., Morrison, W.I., 2009. Genomic analysis reveals extensive gene duplication within the bovine TRB locus. BMC Genomics 10, 192.

Davis, M.M., Bjorkman, P.J., 1988. T-cell antigen receptor genes and T-cell recognition. Nature 334, 395–402.

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

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. 38, D301–307.

Ehrenmann, F., Lefranc, M.-P., 2011. 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. 6, 750–761.

Faldyna, M., Levá, L., Knötigová, P., Toman, M., 2001. Lymphocyte subsets in peripheral blood of dogs – a flow cytometric study. Vet. Immunol. Immunopathol. 82, 23–37.

Faldyna, M., Sinkora, J., Knotigova, P., Leva, L., Toman, M., 2005. Lymphatic organ development in dogs: major lymphocyte subsets and activity. Vet. Immunol. Immunopathol. 104, 239–247.

Garboczi, D.N., Ghosh, P., Utz, U., Fan, Q.R., Biddison, W.E., Wiley, D.C., 1996. Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 384, 134–141.

Garcia, K.C., Degano, M., Stanfield, R.L., Brunmark, A., Jackson, M.R., Peterson, P.A., Teyton, L., Wilson, I.A., 1996. An alphabeta T cell receptor structure at 2.5 A and its orientation in the TCR-MHC complex. Science 274, 209–219.

Gellert, M., 1992. Molecular analysis of V(D)J recombination. Annu. Rev. Genet. 26, 425–446.

Gellert, M., 2002. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu. Rev. Biochem. 71, 101–132.

Giudicelli, V., Chaume, D., Lefranc, M.-P., 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 33, D256–D261.

Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., Lefranc, M.-P., 2006. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 34, D781–D784.

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

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

Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J.C., Smit, A.F., Wang, K., Koop, B.F., Hood, L., 2001. Comparative genomics of the human and mouse T cell receptor loci. Immunity 15, 337–349.

Gottschalk, L.R., Leiden, J.M., 1990. Identification and functional characterization of the human T-cell receptor beta gene transcriptional enhancer: common nuclear proteins interact with the transcriptional regulatory elements of the T-cell receptor alpha and beta genes. Mol. Cell. Biol. 10, 5486–5495.

Hall, M.A., Lanchbury, J.S., 1995. Healthy human T-cell receptor beta-chain repertoire. Quantitative analysis and evidence for J beta-related effects on CDR3 structure and diversity. Hum. Immunol. 43, 207–218.

Hesse, J.E., Lieber, M.R., Mizuuchi, K., Gellert, M., 1989. V(D)J recombination: a functional definition of the joining signals. Genes Dev. 3, 1053–1061.

Carvajal, I.M., Sen, R., 2000. Functional analysis of the murine TCRβ-Chain gene enhancer. J. Immunol. 164, 6332–6339.

Jaeger, E.E., Bontrop, R.E., Lanchbury, J.S., 1993. Nucleotide sequences, polymorphism and gene deletion of T cell receptor beta-chain constant regions of *Pan troglodytes* and *Macaca mulatta*. J. Immunol. 151, 5301–5309.

Jung, D., Alt, F.W., 2004. Unraveling V(D)J recombination; insights into gene regulation. Cell 116, 299–311.

Lefranc, M.-P., Lefranc, G., 2001. The T cell Receptor FactsBook. Academic Press, London, UK.

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

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., 2005. IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. Dev. Comp. Immunol. 29, 185–203.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G., Duroux, P., 2009. IMGT®, the international ImMunoGeneTics information system®. Nucleic Acids Res. 37, D1006–1012.

Lefranc, M.-P., 2011a. IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. Cold Spring Harb. Protoc. 6, 633–642.

Lefranc, M.-P., 2011b. From IMGT-ONTOLOGY CLASSIFICATION Axiom to IMGT standardized gene and allele nomenclature: for immunoglobulins (IG) and T cell receptors (TR). Cold Spring Harb. Protoc. 6, 627–632.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kama, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., Evan Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., deJong, P.J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H.G., Pollinger, J.P., Stephen, M.J., Searle, S.M.J., Sutter, N.B., Thomas, R., Webber, C.Broad Institute Genome Sequencing Platform, Lander, E.S., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803–819.

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

Rock, E.P., Sibbald, P.R., Davis, M.M., Chien, Y.H., 1994. CDR3 length in antigen-specific immune receptors. J. Exp. Med. 179, 323–328.

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

Rudikoff, S., Fitch, W.M., Heller, M., 1992. Exon-specific gene correction (conversion) during short evolutionary periods: homogenization in a two-gene family encoding the beta-chain constant region of the T-lymphocyte antigen receptor. Mol. Biol. Evol. 9, 14–26.

Rust, C.J., Verreck, F., Vietor, H., Koning, F., 1990. Specific recognition of staphylococcal enterotoxin A by human T cells bearing receptors with the V gamma 9 region. Nature 346, 572–574.

Schrenzel, M.D., Watson, J.L., Ferrick, D.A., 1994. Characterization of horse (*Equus caballus*) T-cell receptor beta chain genes. Immunogenetics 40, 135–144.

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

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.

Tripathi, R.K., Mathieu, N., Spicuglia, S., Payet, D., Verthuy, C., Bouvier, G., Depetris, D., Mattei, M.-G., Hempel, W.M., Ferrier, P., 2000. Definition of a T-cell receptor b gene core enhancer of V(D)J recombination by transgenic mapping. Mol. Cell. Biol. 20, 42–53.

Whitehurst, C.E., Schlissel, M.S., Chen, J., 2000. Deletion of germline promoter PD beta 1 from the TCR beta locus causes hypermethylation that impairs D beta 1 recombination by multiple mechanisms. Immunity 13, 703–714.

Williams, C.B., Blankenhorn, E.P., Byrd, K.E., Levinson, G., Gutman, G.A., 1991. Organization and nucleotide sequence of the rat T cell receptor beta-chain complex. J. Immunol. 146, 4406–4413.

Yousfi Monod, M., Giudicelli, V., Chaume, D., Lefranc, M.-P., 2004. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V–J and V–D–J JUNCTIONs. Bioinformatics 20 (Suppl. 1), i379–i385.