Reassignment of the murine 3'TRDD1 recombination signal sequence.

Cédric Touvrey, Lindsay Cowell, Ann Lieberman, Patrice Marche, Evelyne Jouvin-Marche, Serge Candéias

Re-assignment of the murine 3'*TRDD1* recombination signal sequence

C. Touvrey · L. G. Cowell · A. E. Lieberman · P. N. Marche · E. Jouvin-Marche · S. M. Candéias

C. Touvrey · P. N. Marche · E. Jouvin-Marche · S.M. Candéias (@)

CEA, DSV, DRDC, Laboratoire d'Immunochimie, Grenoble, F-38054 France;

INSERM U548, Grenoble, France.

e-mail: serge.candeias@cea.fr

Tel : +33-4-38-78-92-49

Fax : +33-4-38-78-98-03

L. G. Cowell · A. E. Lieberman

Departments of Biostatistics & Bioinformatics and Immunology

Duke University Centre for Computational Immunology,

Duke University,

Durham, NC, USA.

C. Touvrey (present address)

Ludwig Institute for Cancer Research / CHUV,

University Hospital Lausanne,

Lausanne, Switzerland

Abstract

T cell receptor genes are assembled in developing T lymphocytes from discrete *V*, *D* and *J* genes by a site-specific somatic rearrangement mechanism. A flanking recombination signal, composed of a conserved heptamer and a semi-conserved nonamer separated by 12 or 23 variable nucleotides, targets the activity of the rearrangement machinery to the adjoining *V*, *D* and *J* genes. Following rearrangement of *V*, *D* or *J* genes, their respective recombination signals are ligated together. Although these signal joints are allegedly invariant, created by the head-to-head abuttal of the heptamers, some do exhibit junctional diversity. Recombination signals were initially identified by comparison and alignment of germ-line sequences with the sequence of rearranged genes. However, their overall low level of sequence conservation makes their characterization solely from sequence data difficult. Recently, computational analysis unravelled correlations between nucleotides at several positions scattered within the spacer and recombination activity, so that it is now possible to identify putative recombination signals and determine and predict their recombination efficiency. In this paper, we analyzed the variability introduced in signal joints generated after rearrangement of the *TRDD1* and *TRDD2* genes in murine thymocytes. The recurrent presence of identical nucleotides inserted in these signal joints led us to reconsider the location and sequence of the TRDD1 recombination signal. By combining molecular characterization and computational analysis, we show that the functional TRDD1 recombination signal is shifted inside the putative coding sequence of the *TRDD1* gene, and consequently that this gene is shorter than indicated in the databases.

Keywords:

T cell receptor gene; V(D)J recombination; recombination signal; signal joint; junctional diversity

Introduction

T cell receptor (TR) and immunoglobulin (Ig) *V*, *D* and *J* genes need to be rearranged by a site-specific recombinase before they are expressed in T and B lymphocytes, respectively. This recombinase complex is composed of two lymphoid specific proteins, the products of the recombination activating genes (RAG)-1 and -2, and ubiquitous DNA repair factors involved in DNA repair by non-homologous end joining (Jung and Alt 2004). The recombination activity is specifically targeted to TR and Ig *V*, *D* and *J* genes by the presence at their borders of a short DNA motif called recombination signal (RS) (Tonegawa 1983). After synapsis of the genes to be rearranged (Hiom and Gellert 1998), the introduction of a double strand break by the RAG proteins exactly at the border of the genes and their RS generates hairpin sealed coding ends and blunt, phosphorylated signal ends (McBlane et al. 1995). Processing of the coding ends requires opening of the hairpin structure, which can generate P nucleotides, and includes the eventual addition of non-templated N nucleotides by the terminal nucleotidyl transferase (TdT) and/or the removal of a few nucleotides from the coding ends before they are ligated together to create a novel exon (Jung and Alt 2004). The result is a highly diverse repertoire of coding joints. In sharp contrast, it is generally held that signal ends are ligated together without any processing, so that signal joints result from a perfect head to head juxtaposition of the RSs.

RSs were first identified by sequence alignment of *Ig* genes (Tonegawa 1983), and later functionally characterized in recombination assays using artificial extrachromosomal substrates (Akira et al. 1987; Hesse et al. 1989). They consist of a conserved heptamer (CACAGTG) separated from a semi conserved nonamer (ACAAAAACC) by 12 or 23 bases. The first bases of the heptamer are conserved in almost all the human and murine functional RSs. The A/T rich nonamer on the other hand is less well conserved, although it acts as the docking site for RAG-1, which then recruits RAG-2 onto the RS (Difilippantonio et al. 1996).

Analysis of the spacer sequence was more complex because both 12 and 23 spacers are highly heterogeneous (Cowell et al. 2004; Glusman et al. 2001). However, the recent development of new computational tools unravelled correlations between nucleotides at several positions within the spacer, and correlations between the identity of nucleotides at key positions and the overall recombination efficiency of the RS (Cowell et al. 2002; Cowell et al. 2003). This mutual information model allows the calculation, for each RS, of a "RS information content" (RIC) score that accurately reflects and efficiently predicts the signal's recombination activity (Cowell et al. 2004).

In spite of the generally accepted models for V(D)J recombination, a large body of evidence indicates that signal ends can be processed and modified before they are ligated to form a signal joint (Candéias et al. 1996; Kanari et al. 1998; Touvrey et al. 2006). Recently, we undertook the analysis of the junctional diversity in signal joints created when the *TRDD1* and *TRDD2* genes are rearranged together in murine thymocytes. Three quarters of the signal joints were found to exhibit variability created by the insertion of N nucleotides in between the RSs' heptamers. The recurrence of a di-nucleotide in these modified signal joints prompted us to re-consider the location of the break point at the 3' end of the *TRDD1* gene. Our results identified a new functional RS23 3' of the murine *TRDD1* gene.

Materials and methods

Mice-Thymocyte DNA preparation

CD3ε$^{\Delta 5/\Delta 5}$ mice onto a BALB/c background (Gallagher et al. 1998) were bred in the animal facility of the Commissariat à l'Energie Atomique-Grenoble, France. They were sacrificed by CO_2 inhalation and their thymus removed at 6-8 wks of age. The thymi were crushed, and thymocyte DNA was prepared from single cell suspensions with the Nucleospin Tissue Kit (Macherey-Nagel, Hoerdt, France) according to the manufacturer's instructions.

Signal Joint amplification, analysis and cloning

TR DD1/DD2 signal joints were amplified with the 3'DD1 and 5'DD2 oligonucleotide primers (Table 1) from total thymocyte DNA with the AmpliTaq Gold system (Applied Biosystems, Courtaboeuf, France) in 37 cycles consisting of 30 sec at 94°C, 30 sec at 58°C and 30 sec at 72°C followed by an incubation of 10 min at 72°C. Before the first cycle, the enzyme was activated at 94°C for 6 min. To analyze junctional diversity, one half of the product was incubated for 3 hrs with or without ApaL1 (10 U, New England Biolabs, St Quentin en Yvelines, France). Digests were then separated on an agarose gel, transferred onto a nylon membrane and hybridized with 3'DD1p (Table 1), a radiolabelled oligonucleotide specific for the signal end moiety of the amplicon. Analysis was performed on a Personnal Molecular Imager FX with the Quantity One software (Bio-Rad, Marnes la Coquette). For sequencing, after purification on an agarose gel, the PCR products were cloned into the PGEM-T Easy vector (Promega, Charbonnières, France) according to the manufacturer's instructions and transformed into competent bacteria. Following plating, signal joint-containing colonies were identified by hybridization with the 5'DD2p radio-labelled probe. Plasmid DNA was prepared from positive colonies, and digested with the restriction enzyme ApaL1 to discriminate between ApaL1-sensitive signal joints, resulting from the perfect

ligation of the TRDD1 and TRDD2 RSs (ApaL1-S) and those in which one or both of the RS ends have been processed/modified, resulting in ApaL1-resistant (ApaL1-R) signal joints, as described (Candéias et al. 1996; Touvrey et al. 2006). ApaL1-R signal joints were then sequenced with the 5'DD2p (Table 1) primer using BigDye Terminator reagent (Applied Biosystems, Courtaboeuf, France) to determine the molecular nature of the modifications.

Identification of the 3'DD1 break point

Ligation mediated-PCR (LM-PCR) was performed as described (Hempel et al. 1998) using CD3$\varepsilon^{\Delta 5/\Delta 5}$ total thymocyte DNA as template and the DR19/DR20 double stranded linker (Roth et al. 1993). Amplification was performed with the 3'DD1 and DR20 oligonucleotide primers. The resulting PCR products were digested to completion with ApaL1, separated on an agarose gel, transferred onto a nylon membrane and hybridized with 3'DD1p as described above. To determine the exact location of the double strand break, this amplicon was cloned and individual colonies sequenced.

p53 amplification.

DNA input in PCR reactions was verified by amplification of a fragment of the p53 gene with the oligonucleotides X6.5 and X7 (Table 1) for 35 cycles consisting of 30 sec at 94°C, 30 sec at 55°C and 1 min at 72°C followed by an incubation of 10 min at 72°C. Before the first cycle, DNA was denatured for 5 min at 94°C. PCR products were separated on an agarose gel in the presence of ethidium bromide and visualized on a Vistra Systems Fluorimager SI with the Image Quant 5.2 software (Molecular Dynamics, Bondoulfe, France).

Results

We analyzed the structure of TR DD1/DD2 signal joints in thymocytes prepared from CD3ε-deficient CD3$\varepsilon^{\Delta 5/\Delta 5}$ mice. CD3ε-deficiency prevents TR expression on developing thymocytes and arrests T cell development at an early stage, but does not impede V(D)J recombination activity. Consequently, *TRD* gene rearrangement is normal in these mice (Gallagher et al. 1998; Malissen et al. 1995). Signal joints are believed to be created by the perfect juxtaposition of RS heptamers, thereby generating a restriction site for the enzyme ApaL1 if these heptamers conform to the consensus sequence and begin by a CAC or GTG triplet. However, following amplification and ApaL1 digestion, only a minor fraction of TR DD1/DD2 signal joints contain an ApaL1 restriction site and most of the PCR product is not digested (Fig 1). Thus, it appears that in the majority of the recombination events, TR DD1/DD2 signal joints do exhibit junctional diversity. After amplification, cloning and ApaL1 digestion, only about 26% (11 out of 42) of the individual signal joints were found to contain a site for this restriction enzyme. Therefore, in almost three quarters of TR DD1/DD2 signal joints, the TRDD1 and/or the TRDD2 signal ends have been processed before ligation. To determine the exact nature of these modifications, ApaL1-R plasmids were sequenced (Fig 2C). We referred to published sequences (Chien et al. 1987) to delineate the boundaries of both *TRDD1* and *TRDD2* genes and their RSs (Fig 2A). The sequence of ApaL1-R TR DD1/DD2 signal joints were aligned with the theoretical signal and coding ends generated by RAG cleavage (Fig 2B) in order to identify the nucleotides that have been lost and/or added prior to the ligation. None of these joints were found to exhibit nucleotide loss. All were found to contain non-templated N nucleotides, as described for TRB, TRA and other TRD signal joints (Candéias et al. 1996; Kanari et al. 1998; Touvrey et al. 2006). Interestingly, each of the ApaL1-R junctions contained at least 3 added nucleotides. This feature is unusual when compared to our previous results where additions of one or two nucleotides were

frequent (Candéias et al. 1996; Touvrey et al. 2006). In addition, all the stretches of added bases contain a "CA" dinucleotide immediately adjacent to the TRDD1 heptamer (Fig 2C). As the *TRDD1* gene sequence ends with "CA", this observation raises the possibility that the position of the double strand break introduced by the RAG-1/RAG-2 complex 3' of the *TRDD1* gene is 2 nucleotides (nt) inside the published gene sequence.

To determine the exact location of this break point, we performed LM-PCR on DNA prepared from CD3ε-deficient thymocyte. This assay allows the amplification of blunt ended DNA molecules, such as V(D)J recombination-generated signal ends, following ligation of a ds linker (Fig 3A and 3B). The product detected in RAG-$2^{-/-}$ thymocyte DNA (Fig 3A) was considered to result from the presence of a non specific DNA break because it migrates slightly slower than the products amplified from CD3ε-deficient thymocytes, is totally resistant to ApaL1 digestion (fig 3A), and because no TR DD1/DD2 signal joints could be amplified from this sample (Fig1). As expected for signal ends, (Roth et al. 1993), the PCR products obtained from CD3ε-deficient DNA after linker ligation are almost fully digestible by ApaL1 (Figs 3A and 3B). The creation of a restriction site for this enzyme upon ligation with the double stranded linker indicates that most of the 3'TRDD1 signal ends end in a CAC triplet. However, this reconstitution does not allow any discrimination between the two possible cleavage sites. Therefore, the LM-PCR products were cloned and individual plasmids sequenced. All of the TRDD1 3' signal ends identified in this experiment (20 clones from 2 mice) were found to contain an additional CA dinucleotide (Fig 3C). Because the presence of these "extra" bases would not prevent the creation of an ApaL1 restriction site upon formation of a signal joint, we then sequenced two putatively un-modified ApaL1-S signal joints identified during our initial screening procedure. They were also found to contain this CA dinucleotide (Fig 2D). We concluded from these experiments that this CA dinucleotide results

from the introduction of a double strand break 2 bases upstream of the described 3'TRDD1 RS, within the *TRDD1* gene and is present in all the TR DD1/DD2 signal joints.

Alignment with a consensus RS23 sequence shows (Fig 4A) that this shift results in an increase in the percentage of nucleotide identity from 36% (14 bases out of 39) to 56 % (22 matches out of 39). The numbers of identical bases in the heptamer (5) and in the nonamer (6) motifs do not change, but they are distributed differently. The main difference resides in the spacer, where the number of positions in the RS matching with the consensus sequence raises from 3 to 11. However, simple sequence comparison and alignment do not provide any information regarding the RS functionality, except for a few essential positions such as the CAC triplet in the heptamer. (sentence deleted). When the spacer sequence containing the maximum mutual information (Cowell et al. 2004) was used in place of the consensus spacer sequence, the proposed 2nt shift again induces an increase in the percentage of identity (from 41% to 53%) of the 3'TRDD1 RS with this model RS23 sequence (Fig 4B). As before, this increase results from a higher number of identical bases in the 23 nt spacer sequence. More importantly, this 2 nt shift raises the RIC score of the 3'TRDD1 RS from -70 to -52. This increased RIC score indicates that the shifted RS has the higher recombination efficiency when compared to the published sequence. This finding is in agreement with the fact that the only molecular species found during the LM-PCR analysis of the 3'TRDD1 signal ends identified a breakpoint 2 nt inside the *TRDD1* gene.

Discussion

TRD gene rearrangement is complex. This locus encompasses two *D* genes that can be used concurrently. The determination of *TRDD1* and *TRDD2* gene/RS limits were initially made solely by comparing their germ line sequences with the sequence of rearranged genes and attributing RS properties to the sequence most similar to published RS sequences (Chien et al. 1987). The short length of the *TRDD1* and *TRDD2* genes, together with a high level of end processing (exonucleolytic nibbling, P and N nucleotides) makes the exact assignment of their boundaries in coding joints difficult and sometimes imprecise (Elliott et al. 1988; Lacy et al. 1989; Takagaki et al. 1989). By analyzing the diversity of the reciprocal TR DD1/DD2 signal joints, we identified a pattern suggesting that bases from the *TRDD1* gene were specifically retained in these signal joints. Imprecise cutting leading to the retention of coding flank has been reported only within the context of V(D)J recombination performed with mutated RAG-1 and/or RAG-2 proteins (Lee and Desiderio 1999; Talukder et al. 2004). Thus, it seems highly unlikely that this phenomenon can account for the systematic inclusion of the CA dinucleotide in all the TR DD1/DD2 signal joints in mice possessing normal RAG proteins and V(D)J recombination activity, as was previously suggested by (Carroll et al. 1993). We therefore considered the possibility that the location of the double-strand break generating the inclusion of these two bases in the signal joints is not random but programmed, dictated by the position of the 3'TRDD1 RS. To verify this hypothesis, we cloned and sequenced the corresponding signal ends. The only molecular species found was terminated by this "extra" CA dinucleotide. Once loaded onto a RS, the RAG-1/RAG-2 complex creates a double strand break exactly at the gene/RS boundary (McBlane et al. 1995; Roth et al. 1993). This result therefore defines a new RS23 at the 3' end of the *TRDD1* gene, where the heptamer sequence reads CACACAG in place of CACAGGT (Fig 4C). This new RS23 presents more homology with a consensus RS23 sequence and with a model RS23 in which the spacer sequence

contains the highest mutual information possible. At the very least, our results indicate that there are two overlapping RSs downstream of the murine *TRDD1* gene, which differ in their position by 2 nt. However, out of the 20 signal ends sequenced after LM-PCR, none was found to correspond to the published 3'TRDD1 RS. Thus, if it is at all functional, the published 3'TRDD1 RS sequence, ending with the CACAGGT heptamer, is not used in more than 5% of the recombination events. This experimental conclusion is strongly supported by the fact that the calculated RIC score for the new 3' TRDD1 RS23, as defined in this study, is much higher than the score for the published TRDD1 RS23. This difference denotes a difference in recombination efficiency and correlates with our observations: out of two overlapping putative RS competing for the binding of the RAG proteins and the recombination machinery, only one is found in signal joints. Thus, our work defines a new sequence for the functional RS23 downstream of the *TRDD1* gene (Fig 4C). Interestingly, in this revised version, the nucleotides in positions 5-8 of the nonamer (AAAC) now match the consensus sequence. These nucleotide positions are those that, of positions within the RS nonamer, show the lower position-wise entropy, i.e. the most conserved, in a collection of 155 physiologic RS23 elements (Cowell et al. 2002). This conservation may reflect their involvement in base-specific DNA-protein contact, as shown by interference footprinting assays (Swanson 2002). Finally, it should be noted that this proposed revision of the murine 3'TRDD1 RS strikingly increases its similarity to its human counterpart, downstream of the *TRDD2* gene. The orthologous murine and human RSs now have an identical heptamer, and they differ at only 6 positions, 5 in the spacer and 1 in the nonamer. Without this revision, the sequences of these RSs differ at 24 of the 39 position (Fig 5).

In conclusion, we show by a combination of molecular biology and computational analysis that the *3'TRDD1* gene/RS boundary is shifted 2 nt inside the *TRDD1* gene when compared with the published sequence, and therefore that when used in a recombination event, the *TRDD1* gene is shortened by 2 nt. This was probably not recognized earlier because of the

very high level of diversity in TR DD1/DD2 coding joints, in which only fragments of the *TRDD1* sequence are retained and/or identifiable (Elliott et al. 1988; Lacy et al. 1989; Takagaki et al. 1989). In this context, it is formally possible that the phenomenon we are describing in this paper for the TRDD1 gene also applies to other TR or Ig genes but was never detected because it is masked by the exonucleolytic nibbling of coding ends that takes place during V(D)J recombination. Indeed, examination of 59 murine TRAJ genes from the IMGT database revealed for 17 of them the presence, adjacent to the heptamer, of a TG dinucleotide. These 17 genes could therefore have 2 overlapping heptamers. However, RIC score calculation predicted in each case that a 2 bp shift would result in a less efficient RS element (Table 2). Thus, it appears that the mere presence of a TG or CA dinucleotide at the border of a heptamer is not sufficient to modify the functional RS element, and that the revision that we are proposing for the 3'TRDD1 RS is specific for this gene. This study illustrates the contribution of computational modelling to the difficult task of RS identification and characterization, even in the case of already described functional genes.

Acknowledgements

CT was supported in part by grant from “l’Association pour la Recherche contre le Cancer”. LGC is supported by a Career Award from the Burroughs Wellcome Fund. We are thankful to Dr C. Aude-Garcia for her comments on the manuscript. We are grateful to Irène Maréchal and Soumalamya Bama for excellent animal care.

References

Akira, S., Okazaki, K., and Sakano, H.: Two pairs of recombination signals are sufficient to cause immunoglobulin V-(D)-J joining. *Science 238:* 1134-8, 1987

Candéias, S., Muegge, K., and Durum, S. K.: Junctional diversity in signal joints from T cell receptor beta and delta loci via terminal deoxynucleotidyl transferase and exonucleolytic activity. *J.Exp.Med. 184:* 1919-1926, 1996

Carroll, A. M., Slack, J. K., and Mu, X.: V(D)J recombination generates a high frequency of nonstandard TCR D delta-associated rearrangements in thymocytes. *J Immunol 150:* 2222-30, 1993

Chien, Y. H., Iwashima, M., Wettstein, D. A., Kaplan, K. B., Elliott, J. F., Born, W., and Davis, M. M.: T-cell receptor delta gene rearrangements in early thymocytes. *Nature 330:* 722-7, 1987

Cowell, L. G., Davila, M., Kepler, T. B., and Kelsoe, G.: Identification and utilization of arbitrary correlations in models of recombination signal sequences. *Genome Biol 3:* RESEARCH0072, 2002

Cowell, L. G., Davila, M., Ramsden, D., and Kelsoe, G.: Computational tools for understanding sequence variability in recombination signals. *Immunol Rev 200:* 57-69, 2004

Cowell, L. G., Davila, M., Yang, K., Kepler, T. B., and Kelsoe, G.: Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. *J Exp Med 197:* 207-20, 2003

Difilippantonio, M. J., McMahan, C. J., Eastman, Q. M., Spanopoulou, E., and Schatz, D. G.: RAG1 mediates signal sequence recognition and recruitment of RAG2 in V(D)J recombination. *Cell 87:* 253-62, 1996

Elliott, J. F., Rock, E. P., Patten, P. A., Davis, M. M., and Chien, Y. H.: The adult T-cell receptor delta-chain is diverse and distinct from that of fetal thymocytes. *Nature 331:* 627-31, 1988

Gallagher, M., Candéias, S., Martinon, C., Borel, E., Malissen, M., Marche, P. N., and Jouvin-Marche, E.: Use of TCR ADV gene segments by the delta chain is independent of their position and of CD3 expression. *Eur.J.Immunol. 28:* 3878-3885, 1998

Glusman, G., Rowen, L., Lee, I., Boysen, C., Roach, J. C., Smit, A. F., Wang, K., Koop, B. F., and Hood, L.: Comparative genomics of the human and mouse T cell receptor loci. *Immunity 15:* 337-49, 2001

Hempel, W. M., Stanhope-Baker, P., Mathieu, N., Huang, F., Schlissel, M. S., and Ferrier, P.: Enhancer control of V(D)J recombination at the TCRbeta locus: differential effects on DNA cleavage and joining. *Genes Dev. 12:* 2305-2317, 1998

Hesse, J. E., Lieber, M. R., Mizuuchi, K., and Gellert, M.: V(D)J recombination: a functional definition of the joining signals. *Genes Dev 3:* 1053-61., 1989

Hiom, K. and Gellert, M.: Assembly of a 12/23 paired signal complex: a critical control point in V(D)J recombination. *Mol Cell 1:* 1011-9, 1998

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

Kanari, Y., Muto, M., and Yamagishi, H.: TCR delta gene rearrangements revealed by fine structure of the recombination junction in mice. *Microbiol Immunol 47:* 883-94, 2003

Kanari, Y., Nakagawa, R., Arakawa, H., and Yamagishi, H.: Variable gene segment-specific N-insertions at the signal joint of T-cell receptor Vbeta-Dbeta recombinations. *Immunol Lett 61:* 151-5, 1998

Lacy, M. J., McNeil, L. K., Roth, M. E., and Kranz, D. M.: T-cell receptor delta-chain diversity in peripheral lymphocytes. *Proc Natl Acad Sci U S A 86:* 1023-6, 1989

Lee, J. and Desiderio, S.: Cyclin A/CDK2 regulates V(D)J recombination by coordinating RAG- 2 accumulation and DNA repair. *Immunity. 11:* 771-781, 1999

Malissen, M., Gillet, A., Ardouin, L., Bouvier, G., Trucy, J., Ferrier, P., Vivier, E., and Malissen, B.: Altered T cell development in mice with a targeted mutation of the CD3-epsilon gene. *EMBO.J. 14:* 4641-4653, 1995

McBlane, J. F., van Gent, D. C., Ramsden, D. A., Romeo, C., Cuomo, C. A., Gellert, M., and Oettinger, M. A.: Cleavage at a V(D)J recombination signal requires only RAG1 and RAG2 proteins and occurs in two steps. *Cell 83:* 387-95, 1995

Roth, D. B., Zhu, C., and Gellert, M.: Characterization of broken DNA molecules associated with V(D)J recombination. *Proc Natl Acad Sci U S A 90:* 10788-92., 1993

Swanson, P. C.: Fine structure and activity of discrete RAG-HMG complexes on V(D)J recombination signals. *Mol Cell Biol 22:* 1340-51, 2002

Takagaki, Y., Nakanishi, N., Ishida, I., Kanagawa, O., and Tonegawa, S.: T cell receptor-gamma and -delta genes preferentially utilized by adult thymocytes for the surface expression. *J Immunol 142:* 2112-21, 1989

Talukder, S. R., Dudley, D. D., Alt, F. W., Takahama, Y., and Akamatsu, Y.: Increased frequency of aberrant V(D)J recombination products in core RAG-expressing mice. *Nucleic Acids Res 32:* 4539-49, 2004

Tonegawa, S.: Somatic generation of antibody diversity. *Nature 302:* 575-81, 1983

Touvrey, C., Borel, E., Marche, P. N., Jouvin-Marche, E., and Candéias, S. M.: Gene specific signal joint modifications during V(D)J recombination of *TCRAD* locus genes in murine and human thymocytes. *Immunobiology in press*, 2006

Figure legends

Fig 1: TR DD1/DD2 signal joints are diverse.

Signal joints generated following rearrangement of *TRDD1* and *TRDD2* genes were amplified from thymocyte DNA prepared from two CD3ε-deficient and one RAG-2-deficient mice. They were then were migrated on an agarose gel without (-) or with (+) ApaL1 digestion and analyzed as described in Materials and Methods. The arrowhead indicates the position of the 3' TRDD1 signal end. The bottom panel shows amplification of the *p53* gene as a DNA input control.

Fig 2: Sequences of the ApaL1-R TR DD1/DD2 signal joints

(A): Germline sequences of the murine *TRDD1* and *TRDD2* genes and their RS elements. TRDD1 and TRDD2 coding regions are in bold characters. RS heptamers and nonamers, separated by 12 or 23 bases, are underlined. The identification of the different elements is based on the published sequences (Chien et al. 1987). **(B):** Representation of the putative coding and signal ends. The boundaries of the putative coding and signal ends (CE and SE, respectively) are depicted for both TRDD1 and TRDD2. Coding regions are in bold characters. RS heptamers and nonamers, separated by 12 or 23 bases, are underlined. **(C)**: Sequences of the modified signal joints. The amplified signal joints were sequenced with the 5'DD2p oligonucleotide. Identification of TRDD1 and TRDD2 RS was based on the published sequences. The sequences of TRDD1 and TRDD2 RS heptamers are indicated on the first line in bold characters; the first flanking TRDD1 and TRDD2 bases are indicated in italic characters. The nucleotides listed under the "N" column represent putative non-templated nucleotides. The recurrent CA dinucleotide flanking the TRDD1 RS is in bold, underlined characters. The number 2 in the last column indicates that these particular sequences were found twice in a sample from one mouse. Identical sequences found in

different samples were considered as distinct events. The last two sequences represent non-standard rearrangement products generated through the recombination of the 5'TR DD1 and 3'TR DD2 RS (Kanari et al. 2003). In these junctions, the sequence of the *TRDD1* and *TRDD2* genes are underlined. In **(D),** the sequence of two putatively un-modified ApaL1-S signal joints is shown. The additional "CA" dinucleotide adjacent to the TRDD1 RS was also found.

Fig 3: LM-PCR analysis of the 3'TRDD1 RS end from murine thymocytes.

(A) Total thymocyte DNA from two CD3ε-deficient and one RAG-2-deficient mice was amplified as described in Materials and Methods after ligation of a staggered double-stranded linker. The bottom panel shows the amplification of the *p53* gene for the three DNA samples as a DNA input control. The product amplified from RAG-$2^{-/-}$ thymocyte DNA was considered to result from non-specific amplification (see text for details) and its nature was not explored further. **(B)** Total thymocyte DNA from a CD3ε-deficient animal reacted with the DR19/DR20 linker in the presence (T4 Lig +) or in the absence (T4 Lig -) of T4 DNA ligase was amplified with the 3'DD1 and DR20 oligonucleotide primers. DNA input control is shown on the bottom panel for the samples incubated with (1) and without (2) ligase. In **(A)** and **(B)**, one half of the resulting PCR product was digested with ApaL1. After separation on an agarose gel and blotting, un-digested (-) and digested (+) PCR products were revealed by hybridization with the 3'DD1p radio-labelled probe. **(C)** Sequence of the 3' RS23 TRDD1 signal end. The germline sequence of the TRDD1 gene (in bold characters) and its 3'flanking region element is shown on the top "GL" line. The published sequence of the heptamer is underlined. The arrowhead indicates the position of the RAG-induced break. The "SE" line shows the sequence of the 3'TRDD1 signal end as determined following sequencing of the LM-PCR products. The extra "CA" dinucleotide is in bold characters.

Fig 4: Comparison of the 3'TRDD1 RS elements.

The published (Chien et al. 1987) and new (this study) 3'TRDD1 RS23 sequences are aligned with a RS23 consensus sequence **(A)**, a RS sequence assembled from a consensus heptamer, a consensus nonamer and the 23bp spacer with the most mutual information **(B)**, and one another **(C)**. Both sequences are from Cowell et al, 2004. The *TRDD1* gene and the different elements composing the RS are separated by hyphens. DD1: *TRDD1* gene; 7: heptamer; 23: 23 bp spacer; 9: nonamer. In **(A)** and **(B)**, identical bases are replaced by semi columns; in **(C)**, identical bases are in bold capitals and differing bases in lowercase characters.

Fig 5: Alignment of the human 3'TRDD2 and murine 3'TRDD1 RS elements

The human 3'TRDD2 and murine 3'TRDD1 RSs are aligned using either the published **(A)** or revised **(B)** sequence for the murine element. Identical bases are in bold capitals and differing bases are in lowercase characters. The different elements composing the RSs are separated by hyphens. 7: heptamer; 23: 23 bp spacer; 9: nonamer.

Table 1: Oligonucleotides used in this study

3'DD1	5'-GACAATACAGACCAAATATACAGC-3'
3'DD1p	5'-GCGCTGTAGGGAAATATGTAAGTTT-3'
5'DD2	5'-TGGCTTGACATGCAGAAAACACCTG-3'
5'DD2p	5'-GACACGTGATACAAAGCCCAGGGAA-3'
DR20	5'-GCCTATGTACTACCCGGGAATTCGTG-3'
DR19	5'-CACGAATTCCC-3'
X6.5	5'-ACAGCGTGGTGGTACCTTAT-3'
X7	5'-CACATGTACTTGTAGTGGATGG-3

Table 2: RS associated with TRAJ gene segments that have TG as the most 5' nucleotides of the coding region as defined by IMGT. The second RS in each row indicates the RS for which the heptamer includes the coding region TG as its first two nucleotides.

TRAJ gene	RS sequences	RIC score IMGT RS	RIC score shifted RS
>M64239|TRAJ11*01	catttttgtggagaggtttgctgctgtg ttttttgtggagaggtttgctgctgtgtg	-24.60	-36.34
>M64239|TRAJ17*01	tgtttttgcttggcttcagatcactgtg ttttttgcttggcttcagatcactgtgtg	-20.72	-53.66
>M64239|TRAJ2*01	ggcttctgtaaaggtgtcacctgcagtg cttctgtaaaggtgtcacctgcagtgtg	-27.07	-59.52
>M38103|TRAJ2*02	ggcttctgtaaaggtgtcacctgcagtg cttctgtaaaggtgtcacctgcagtgtg	-27.07	-59.52
>M64239|TRAJ21*01	gctttctgtaatggtgctaaccattgtg tttctgtaatggtgctaaccattgtgtg	-23.60	-61.65
>M64239|TRAJ23*01	tgtttttgacagggtatgtaacacagtg ttttttgacagggtatgtaacacagtgtg	-21.74	-53.12
>M64239|TRAJ24*01	ccattttgtagacgtgtttgtcacagtg attttgtagacgtgtttgtcacagtgtg	-28.30	-42.01
>M64239|TRAJ30*01	cgttttgggtatggtcccaatcacagtg ttttgggtatggtcccaatcacagtgtg	-21.20	-61.81
>X02858|TRAJ32*01	agttattgtaaggctctgcagggctgtg ttattgtaaggctctgcagggctgtgtg	-22.82	-49.57
>M64239|TRAJ33*01	tgtttttgttaaggttttgtgtctgtg **ttttttgttaaggttttgtgtctgtgtg**	-22.46	-45.63
>M64239|TRAJ39*01	ggtttttgctgagctggagatcactgtg ttttttgctgagctggagatcactgtgtg	-19.20	-42.07
>M64239|TRAJ4*01	aattcttgtaaagcacctttctactgtg ttcttgtaaagcacctttctactgtgtg	-30.19	-53.24
>X02859|TRAJ45*01	ggtttatgtcaaggcttgcctcagggtg tttatgtcaaggcttgcctcagggtgtg	-19.42	-60.55
>M64239|TRAJ47*01	agttttgtcacaggagtttgaggctgtg ttttgtcacaggagtttgaggctgtgtg	-26.15	-40.71
>M64239|TRAJ56*01	agtttttgtagagtcccgtgtcattgtg ttttttgtagagtcccgtgtcattgtgtg	-23.89	-49.61
>M64239|TRAJ57*01	agtatttgtaaggcagtgtgtgggtgtg tatttgtaaggcagtgtgtgggtgtgtg	-26.48	-47.20
>M64239|TRAJ58*01	agtttttgcaaagcccttcagtgcagtg ttttttgcaaagcccttcagtgcagtgtg	-24.44	-44.54

Fig 1

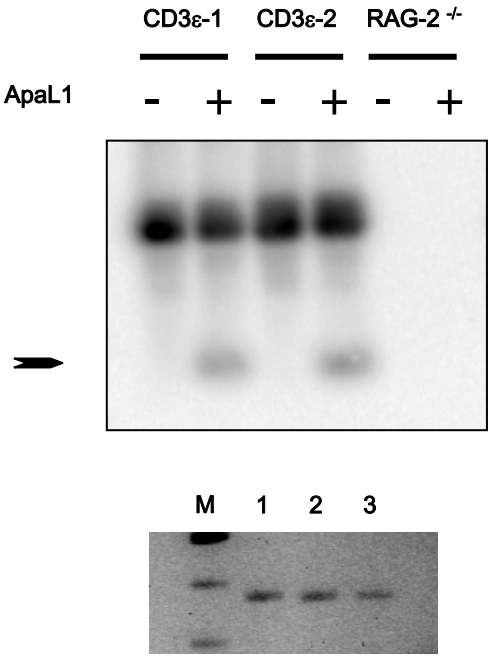

Fig2 :

A : Germline sequences

```
                                    DD1        7            23                    9
TRDD1                        ..//GTGGCATATCACACAGGTTGAAGTATATTAAACCTCTGTTCAGAAACACT

                    9         12        7        DD2
TRDD2          GGTTTTTGCAAAGCTCTGTAGCACCGTGATCGGAGGGATACGAG//..
```

B: Signal and coding ends following RAG-mediated cleavage

```
                          ----------CE       SE-----------------------------------
TRDD1                  ..//GTGGCATATCA       CACAGGTTGAAGTATATTAAACCTCTGTTCAGAAACAC

                  --------------------------SE        CE-------------
TRDD2          GGTTTTTGCAAAGCTCTGTAGCACCGTG         ATCGGAGGGATACGAG//..
```

C: Sequences of the ApaL1-R TR DD1/DD2 signal joints

DD2		**N**		**DD1**	
CACCGTG	*ATCGGA*		*ATATCA*	**CACAGGT**	
CACCGTG		GTGGG**CA**		CACAGGT	
CACCGTG		GGG**CA**		CACAGGT	2
CACCGTG		GGT**CA**		CACAGGT	
CACCGTG		GAG**CA**		CACAGGT	
CACCGTG		GAG**CA**		CACAGGT	
CACCGTG		CCC**CA**		CACAGGT	2
CACCGTG		CAC**CA**		CACAGGT	
CACCGTG		TGT**CA**		CACAGGT	
CACCGTG		TTC**CA**		CACAGGT	
CACCGTG		GG**CA**		CACAGGT	2
CACCGTG		CC**CA**		CACAGGT	
CACCGTG		AC**CA**		CACAGGT	
CACCGTG		G**CA**		CACAGGT	2
CACCGTG		C**CA**		CACAGGT	
CACCGTG		C**CA**		CACAGGT	
CACCGTG		T**CA**		CACAGGT	
CACCGTG		T**CA**		CACAGGT	
CACCGTG		GG**CA**		CACAGGT	
CACCGTG	ATCGGAGGGATACGAGGGGGG**CA**			CACAGGT	
CACCGTG	ATCGGAGGGATACCTGTGGCATAT**CA**			CACAGGT	

D: Sequences of the ApaL1-S TR DD1/DD2 signal joints

DD2		**N**		**DD1**	
CACCGTG	*ATGCGA*		*ATATCA*	**CACAGGT**	
CACCGTG		**CA**		CACAGGT	2

Fig 3

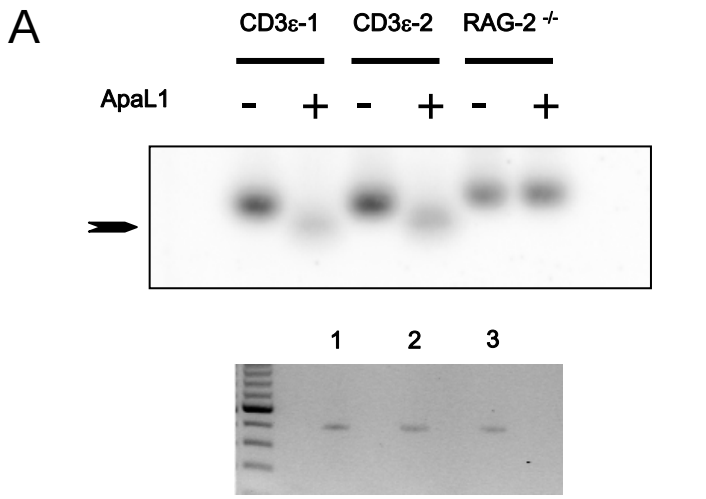

Fig 4

A

```
                        DD1       7           23                    9
RS23 Cons           GTGGCATATCA-CACAGTG-TTGGAACCACATCGGGAGCCTGT-ACAAAAACC
3'DD1 (Chien)       GTGGCATATCA-:::::GT-:GAAGTAT:TTAAACCTCTG:TC-:G:::C::T
3'DD1 (this study)  GTGGCATAT  -::::CA:-G:T:::GT:T::TAAACCT::::-T::G::::A
```

B:

```
                        DD1       7           23                    9
RS23 MI             GTGGCATATCA-CACAGTG-TTGCAACCACATCCTGAGTGTGT-ACAAAAACC
3'DD1 (Chien)       GTGGCATATCA-:::::GT-:GAAGTAT:TTAAACCTC:::TC-:G:::C::T
3'DD1 (this study)  GTGGCATAT  -::::CA:-G:TG::GT:T::TAAACC:C:::-T::G::::A
```

C:

```
                        DD1       7           23                    9
3'DD1 (Chien)       GTGGCATATCA-CACAggt-tgaagtaTATtaaAcctCTgTtc-agAaAcACt
3'DD1 (this study)  GTGGCATAT-  CACAcag-gttgaagTATattAaacCTcTgt-tcAgAaACa
```

Fig 5

A:

```
                 7                  23                 9
h3'TRDD2     CACAcag-gttgGagTgcattAagcCTGTgt-gcAgAaACa
m3'TRDD1     CACAggt-tgaaGtaTattaaAcctCTGTtc-agAaAcACt
```

B:

```
                 7                  23                 9
h3'TRDD2     CACACAG-GTTGgAGTgcATTAAgCCTtTGT-cCAGAAACA
m3'TRDD1     CACACAG-GTTGaAGTatATTAAaCCTcTGT-tCAGAAACA
```