Highly diverse TCR α chain repertoire of pre-immune CD8⁺ T cells reveals new insights in gene recombination

Raphael Genolet¹, Brian J Stevenson^{1,2},
Laurent Farinelli³, Magne Østerås³
and Immanuel F Luescher^{1,*}

¹Ludwig Centre for Cancer Research, University of Lausanne, Epalinges, Switzerland, ²Vital-IT Group, Swiss Institute of Bioinformatics, Lausanne, Switzerland and ³Fasteris SA, Plan-les-Ouates, Switzerland

Although the T-cell receptor $\alpha\delta$ (TCR $\alpha\delta$) locus harbours large libraries of variable (TRAV) and junctional (TRAJ) gene segments, according to previous studies the TCR α chain repertoire is of limited diversity due to restrictions imposed by sequential coordinate TRAV-TRAJ recombinations. By sequencing tens of millions of TCR α chain transcripts from naive mouse CD8⁺ T cells, we observed a hugely diverse repertoire, comprising nearly all possible TRAV-TRAJ combinations. Our findings are not compatible with sequential coordinate gene recombination, but rather with a model in which contraction and DNA looping in the TCR $\alpha\delta$ locus provide equal access to TRAV and TRAJ gene segments, similarly to that demonstrated for IgH gene recombination. Generation of the observed highly diverse TCR α chain repertoire necessitates deletion of failed attempts by thymic-positive selection and is essential for the formation of highly diverse TCR $\alpha\beta$ repertoires, capable of providing good protective immunity.

The EMBO Journal (2012) 31, 1666–1678. doi:10.1038/emboj.2012.48; Published online 28 February 2012

Subject Categories: immunology

Keywords: CD8; illumina; repertoire; sequencing; TCR

Introduction

CD4⁺ and CD8⁺ T cells express T-cell receptor $\alpha\beta$ (TCR $\alpha\beta$) consisting of disulphide linked α and β chains that belong to the Ig superfamily and recognize with their complementary determining regions (CDRs) major histocompatibility complex (MHC) molecules displaying antigen-derived peptides (Starr *et al*, 2003; Rudolph *et al*, 2006). CDR1 and CDR2 are encoded by TCR V α (TRAV) and V β (TRBV) gene segments and primarily interact with MHC helices, whereas the highly diverse CDR3 sequences interact mainly with MHC bound peptides. The efficacy of T-cell immunity depends on the diversity of the pre-immune TCR repertoire. TCR genes are formed by recombination of V, (D), J and C gene segments during thymic development (Bassing *et al*, 2002; Krangel,

2009). The TCR β genes are formed in CD4, CD8 double-negative (DN) thymocytes by sequential TRBD, TRBJ and TRBV recombination obeying allelic exclusion. Rearranged genes need to produce a TCR β chain capable of associating with the pTCR α chain, resulting in an autosomal signal that prevents further recombination and promotes developmental progression to CD4 CD8 double-positive (DP) thymocytes (von Boehmer, 2005). In these cells, TRAV and TRAJ gene segments rearrange on both alleles in multiple rounds, until a TCR capable of recognizing an MHC-peptide ligand on the thymic epithelium and of mediating positive selection is formed (Starr *et al*, 2003). After lineage commitment, thymocytes undergo negative selection, during which cells expressing self-reactive TCR are deleted (Starr *et al*, 2003). The surviving CD4 or CD8 single-positive (SP) thymocytes leave the thymus to populate the peripheral lymphoid system and serve as progenitors for T-cell immunity (Starr *et al*, 2003; Jenkins *et al*, 2010).

V(D)J recombination is mediated by the lymphocyte-specific recombinase activating genes (RAG) 1 and 2 (Yannoutsos *et al*, 2001; Gellert, 2002; Schatz and Ji, 2011). RAG-1 and RAG-2 have been shown to have distinct patterns of DNA binding. While RAG-1 binds to accessible recombination signal sequences (RSSs) that flank V, (D) and J segments, RAG-2 preferentially binds DNA at sites containing histone H3K4 trimethylated at lysine 4 (H3K4me3) via its PHD domain (Matthews *et al*, 2007; Ji *et al*, 2010a,b). Recombination centres, comprising RAG-1 and RAG-2, are formed in activated TRAJ clusters that can capture TRAV gene segments with accessible RSS for recombination. The RAG proteins mediate double-stranded DNA breaks (DSB) and orchestrate VDJ recombination. DSBs are subsequently repaired via the non-homologous end-joining pathway (Bassing *et al*, 2002; Krangel, 2009). The variability in the position of cleavage of the hairpin loop by Artemis and the random nucleotide addition by terminal deoxynucleotidyl transferase (TdT) generates highly variable junctional, CDR3 α sequences (Cabaniols *et al*, 2001; Gellert, 2002). The diversity of the pre-immune TCR repertoire largely determines the ability of an organism to eliminate infections and transformed cells (Starr *et al*, 2003; Jenkins *et al*, 2010). While the repertoire of TCR β chains is remarkably diverse ($3\text{--}4 \times 10^6$ distinct sequences) (Robins *et al*, 2009), the diversity of the pre-immune TCR α chain repertoire, based on limited sequence analysis, has been estimated to be considerably smaller ($0.79\text{--}1.18 \times 10^4$ species) (Cabaniols *et al*, 2001; Pasqual *et al*, 2002). This would imply that the diversity of the TCR repertoire is largely determined by the TCR β chain, which does not agree with available sequence data (Wang *et al*, 2010; Dash *et al*, 2011).

The C57BL/6J (B6) mouse TCR $\alpha\delta$ locus spans ~ 1.7 megabases (Mb), is largely occupied by 132 TRAV gene segments, and consists of original TRAV, most of which have undergone

*Corresponding author. Ludwig Centre for Cancer Research, University of Lausanne, Chemin des Boveresses 155, 1066 Epalinges, Switzerland. Tel.: +41 21 692 5988; Fax: +41 21 692 5995; E-mail: immanuel.luescher@unil.ch

Received: 3 November 2011; accepted: 2 February 2012; published online: 28 February 2012

serial duplication to produce the N and duplicate TRAV subloci (see Supplementary Figure S1 and <http://www.imgt.org/IMGTrepertoire/LocusGenes/>; Church *et al*, 2009). At the 3' end of the locus is a cluster of 60 TRAJ (covering about 64 kb) followed by one TRAC gene segment. The TCR δ locus is unique in that it contains TCR δ genes upstream of the TRAJ region that can rearrange in DN thymocytes upon activation of the TCR δ enhancer (E δ) (Bassing *et al*, 2003; Krangel, 2009). Unless a functional TCR $\gamma\delta$ is formed and commits thymocytes to the TCR $\gamma\delta$ T-cell lineage, TRAV-TRAJ gene recombination starts in DP thymocytes upon activation of the TCR α enhancer (E α), located downstream of the TRAC. This enhancer lies over 500 kb upstream and activates the T early activation (TEA) and proximal TRAJ promoters, resulting in transcription, histone modifications and chromatin remodelling in proximal TRAV and TRAJ regions, restricting primary TRAV-TRAJ rearrangements to 3' TRAV and 5' TRAJ segments (Villey *et al*, 1996; Mauvieux *et al*, 2001; Hawwari *et al*, 2005; Krangel, 2009).

Numerous studies pertain to the view that secondary TRAV-TRAJ recombination occurs in a similar focussed manner, namely that after deletion of a large segment of intervening DNA comprising TEA and TCR δ genes, promoters of proximal TRAV and TRAJ segments become activated, again favouring recombination of proximal TRAV and TRAJ segments (Roth *et al*, 1991; Huang and Kanagawa, 2001; Pasqual *et al*, 2002; Fuschiotti *et al*, 2007). Such progressive opening of the TRAV and TRAJ loci, rendering gradually more distal TRAV and TRAJ gene segments amenable to recombination, has been proposed to generate a diverse repertoire by allowing extensive use of the TRAV and TRAJ segment libraries. This view is consistent with the observations that proximal TRAV and TRAJ regions exhibit a 5'-decontracted, 3'-contracted conformation, increased transcription, histone modifications and accessibility of RSS (Shih and Krangel, 2010; Ji *et al*, 2010b; Shih *et al*, 2011). However, this model has been challenged based on TCR α chain sequencing of hundreds of T-cell clones that showed allelic coincidence for TRAJ, but not TRAV segments and no restriction on TRAV-TRAJ gene usage, which is implied by coordinated, sequential recombination (Davodeau *et al*, 2001).

A different mechanism of V(D)J recombination has been elucidated for the mouse immunoglobulin heavy chain (IgH) locus, which spans 3 Mb and comprises from 5' to 3' end about 100 functional V, 12 D, 4 J and 8 C genes that rearrange in developing B cells, first with DJ and then with DJ-V recombinations. Quantitative 3D fluorescence *in-situ* hybridization (3D FISH) studies have demonstrated that V gene recombinations involve V locus contraction, chromatin remodelling and dynamic DNA loop formation, providing V genes similar proximity to DJ segments (Jhunjhunwala *et al*, 2008, 2009). Whether a similar mechanism operates for TRAV-TRAJ recombination is controversial. According to one 3D FISH study the TCR δ locus in DP thymocytes is contracted (Skok *et al*, 2007), which is at variance with another study according to which the 5' portion of the locus is decontracted (Shih and Krangel, 2010).

To comprehensively assess the diversity of the pre-immune mouse TCR α chain repertoire, we performed paired-end Illumina deep sequencing of RNA derived from CD8⁺ T cells of B6 mice. Analysis of tens of millions of TCR α chain sequences indicated that (i) the TCR α chain repertoire

comprises nearly all (94%) possible TRAV-TRAJ recombinations; (ii) the frequencies of TRAV-TRAJ recombination depend on the TRAV and to a lesser degree on TRAJ gene segment usage and (iii) the CDR3 α diversity correlated with TRAV-TRAJ recombination frequencies. Our results are not compatible with sequential coordinate TRAV-TRAJ rearrangements, but rather with the view that secondary TRAV-TRAJ recombination relies on contraction and DNA looping of the TRAV locus providing TRAV gene segments similar access to recombination centres in the TRAJ locus.

Results

TCR α chain sequencing on naive B6 CD8⁺ T cells

To assess the TCR α repertoire of naive CD8⁺ T cells from B6 mice ($n=5$), their RNA was subjected to two rounds of paired-end Illumina sequencing, each yielding >33 million sequences. Of these, ~15 million (46%) allowed unambiguous and ~13 million (40%) ambiguous TCR α sequence assignment, leaving ~14% unidentified reads (Figure 1A). The high proportion of ambiguous sequences, for which alignments with two or more different genomic TRAV sequences had the same score, is explained by high sequence identity between the TRAV subfamily members, originating from serial duplication of the original TRAV locus, giving rise to the duplicate and N loci (Supplementary Figure S1A). Ambiguous sequence assignment was more frequent for TRAV families comprising numerous members, often having high sequence similarity, than for families containing one or only a few members with distinct sequences. Thus, in order to obtain an integral and unbiased view of the repertoire, we distributed the ambiguous sequence counts to TRAV gene segments according to the ratio of the unambiguous reads assigned to each TRAV for the whole sequencing sample.

The B6 TCR α locus contains 132 TRAV and 60 TRAJ genes, which theoretically allow 7920 possible combinations or 4142 if the 23 TRAV and 22 TRAJ gene segments classified as pseudogenes are omitted (Supplementary Figure S1; <http://www.imgt.org>). Inspection of our sequences indicated that pseudogene sequences were indeed absent, except for five TRAJ (TRAJ47, TRAJ44, TRAJ26, TRAJ7 and TRAJ4) and four TRAV segments (TRAV5-4, TRAV9N-1, TRAV11N and TRAV12N-3), which were found in nearly 12% of all recombinations. With the exception of TRAV11N, which has a stop codon in its sequence, the data suggest that these ought to be reconsidered as functional gene segments (Figure 1B). The other pseudogenes were observed in only 0.2% of all sequences. For simplicity, we present data only for genes referred to as functional in the IMGT database; however, similar results were obtained when analysing all TCR α chain sequences (data not shown).

To assess the impact of possible sequencing errors on TCR α chain repertoire assessment, we calculated the difference in expression frequency for each TRAV-TRAJ recombination between the two sequenced samples (technical replicates, Seq. S1 and Seq. 2). There was less than a two-fold difference for the vast majority of recombinations (Figure 1C). Only ~2.5% of the recombinants exceeded two-fold, and since these were expressed at low levels, they would have minimal impact on the global distribution of the TRAV-TRAJ recombinants (Figure 1C and D). Since the two technical replicates were similar, the results shown in the main figures were

Figure 1 Comparison of two TCR α chain deep sequences. (A) RNA from naive C57BL/6 CD8⁺ T cell was subjected to two paired-end Illumina sequencings. In Seq. 1 33,403,339 and in Seq. 2 32,672,980 total sequences were obtained, of which the indicated fractions and absolute numbers allowed unambiguous, ambiguous (between two TRAV segments of different subloci) or no TCR α chain assignment. (B) For the unambiguous sequences Seq. 1 and Seq. 2, the percentages of TRAV-TRAJ gene segment recombinants are summarized (see Supplementary Figure S1). (C) The fold differences of TRAV-TRAJ recombination frequencies (z axis) of Seq. 1 and Seq. 2 are plotted against TRAV (x axis) and TRAJ gene segments (y axis) listed according to their chromosomal locations. White spheres represent recombinants for which the difference was <2-fold and red spheres difference ≥ 2 -fold. The sizes of the spheres represent the frequencies of the recombinants; for x and y axis numbering, see Supplementary Figure S1. (D) The sum of the white (blue bars) and red spheres (yellow bars) shown in percent of total recombinants. (E) Percentages of in-frame (blue bars) and out-of-frame sequences (orange bars) found in Seq. 1 data.

based on Seq. 1 data. As described below, very similar results were obtained when Seq. 2 data were used. It is noteworthy that on average 16.9% of the TCR α chain sequences was out-of-frame (Figure 1E). However, there was considerable bias in the percentage of out-of-frame sequences across all TRAV-TRAJ recombinants, arguing that the formation of out-of-frame TCR α sequences is not random.

Primary TRAV-TRAJ rearrangements follow coordinate gene recombination

Because intervening DNA is deleted in secondary TCR α chain rearrangements, TRAV-TRAJ recombinants containing the proximal TRAV21 (1 in Figures) and TRAJ58 (1 in Figures) gene segments, respectively, originate from primary rearrangements. The frequency of recombinants containing TRAV21 (1) exhibited clear preferential rearrangements with proximal TRAJ segments (Figure 2A). TRAJ58 (1) also preferentially recombined with the most proximal TRAV segments; however, it also recombined with upstream TRAV segments, including distal ones, albeit with lower frequencies

(Figure 2B). When considering all TRAV-TRAJ recombinants, preferential recombination of proximal TRAJ and TRAV gene segments was limited to the three proximal TRAV segments, which are under the control of the E α enhancer (Figure 2C; Bassing *et al*, 2003; Hawwari and Krangel, 2005). Strikingly, however, for the vast majority of recombinants the frequency of distribution was not consistent with coordinated, sequential TRAV and TRAJ gene segment recombination (Figure 2C).

The frequency of TRAV-TRAJ recombination correlates with TRAV and to a lesser degree with TRAJ segment usage

Examination of all TCR α chain sequences indicated that 94% of all possible TRAV-TRAJ recombinants were expressed in naive CD8⁺ T cells (Figure 3A). The same findings were obtained when using data from Seq. 2 (Supplementary Figure S2A). A fraction of the unseen recombinants is most likely accounted for by selection of cells in other lineages, for example, DP thymocytes expressing TRAV11D (92) may be preferentially selected as NKT cells, the TCR of which

Figure 2 TRAV and TRAJ gene segment usage in primary rearrangements. (A, B) The distribution of TRAJ gene segments associated with TRAV21 (1) (A) or of TRAV segments associated with TRAJ58 (1) (B) is represented in percent for all TRAV21 (1) and TRAJ58 (1) containing recombinants. (C) The frequencies of TRAV-TRAJ gene segment recombinants are represented by a colour code, as defined in the bar, in a 2D graph, in which functional TRAV and TRAJ gene segments are plotted on the x and y axis, respectively. For x and y axis numbering, see Supplementary Figure S1.

exclusively express this TRAV (MacDonald and Mycko, 2007). To gain an overview of TRAV-TRAJ recombination, we plotted their frequency against the TRAV and TRAJ gene segments listed in order of their chromosomal location. This 3D landscape revealed a striking, apparently random dependence of the recombination frequency on TRAV segment usage (Figure 3B). Essentially, the same results were obtained when Seq. 2 data were used (Supplementary Figure S2B). This was also observed when the TRAV segments were clustered according to their TRAV-TRAJ recombination profile, using hierarchical clustering based on Manhattan distances (Figure 3C). If one assumes that TRAV-TRAJ rearrangements occur in a coordinated manner, proximal TRAV segments should recombine preferentially with proximal TRAJ segments, and distal TRAV segments with distal TRAJ gene segments. Therefore, TRAJ profiles of proximal TRAV segments should be similar, but different compared with those of distal TRAV segments. Comparison of the hierarchical ranking of TRAV-TRAJ combinations with their ranking in chromosomal order did not reveal any correlation, indicating that there is no obvious link between the position of TRAV segments and their recombination with TRAJ segments (Figure 3D). Similar results were obtained when TRAJ gene segments were ranked by hierarchical clustering (data not shown). These results strongly argue against sequential, coordinated recombination on a locus-wide level.

During evolution, the TCR α locus has undergone two duplication events involving most of its genes, resulting in three sets of highly homologous TRAV segments, referred to as the original, N and duplicate loci, separated by several kilobases on the chromosome (Supplementary Figure S1; Church *et al*, 2009). To appreciate the impact of TRAV gene segment usage on TRAV-TRAJ recombination, we compared the recombination frequencies between the three subloci. 3D graphs for each of the TRAV subloci revealed strikingly different patterns of TRAV-TRAJ recombination (Figures 4A–C). In addition, they revealed sublocus-specific differences in TRAJ segment usage; for example, more distal TRAJ segments were preferentially associated with the N locus, whereas more proximal TRAJ segments were more commonly paired with the duplicate TRAV locus. To better appreciate the differences in TRAV-TRAJ recombinant frequencies between the original and the N locus, we calculated for each the fold changes and plotted these as spheres against TRAJ and TRAV gene segments. As represented by rows of blue spheres, recombinants of TRAV segments of the 3' portion of the N locus were ≥ 2 -fold more frequent than those of the original locus (Figure 4D). The inverse was true for 5' TRAV segments, that is, their recombinants were ≥ 2 -fold more frequent in the original sublocus, as shown by rows of yellow spheres. As indicated by the size of the spheres, the former recombinants (blue spheres) were on

Figure 3 TRAV-TRAJ gene segment recombinants: an overview. (A) The observed TRAV-TRAJ gene segment recombinants are represented as orange squares in a 2D graph, in which the functional TRAV and the TRAJ gene segments are listed on the x and y axis, respectively. About 6% of the possible recombinants were not observed (black squares). (B) The frequencies (% of total) of TRAV-TRAJ gene recombinants (z axis) are plotted against the TRAV (x axis) and TRAJ gene segments (y axis), listed according to their chromosomal locations. For x and y axis numbering, see Supplementary Figure S1. The frequencies are additionally indicated by a colour code, as defined in the bar. (C) 3D plot as in (B), but with the TRAJ gene segments ranked in a hierarchical clustering (y axis), using a simple link algorithm based on the Manhattan distance method (XLstat). (D) Scatter plot in which TRAV-TRAJ recombinants are plotted against the TRAJ gene segments listed in hierarchical ranking (y axis) and in order of their chromosomal location (x axis), respectively. The TCR α locus is located on chromosome 14 (14C2) at 19.7 cM.

average more numerous than the latter (yellow spheres). Remarkably different situations were observed when the original and the duplicate loci (Figure 4E) or the N and the duplicate loci (Figure 4F) were compared likewise. For example, TRAV-TRAJ recombinant frequencies were high for 3' TRAV segments of the N locus (Figure 4B) but not the duplicate locus (Figure 4C), and therefore, their fold differences were large (Figure 4F). In nearly all cases the TRAV-TRAJ recombination frequencies (Figure 4A–C) and hence the fold differences between the subloci (Figure 4D–F) depended on TRAV usage. The results were remarkably different for each sublocus, indicating that TRAV-TRAJ recombination frequencies are governed by TRAV usage in a non-sublocus-specific manner. Because TRAV segments of the three subloci are highly homologous, this argues that TRAV-TRAJ recombination depends not on TRAV sequences, but on TRAV accessibility determined by factors beyond the subloci.

We then examined the dependence of TRAV-TRAJ recombination on TRAV usage locus-wide. A nearly linear relationship (Spearman coefficient R^2 0.567) was observed when the TRAV segments were ranked in order of increasing expression

frequency and plotted against the frequency of their TRAV-TRAJ recombinants (Figure 5A). Similar analysis of the TRAJ segments showed a more random correlation (Spearman coefficient R^2 0.133) between TRAV-TRAJ recombination frequency and TRAJ usage (Figure 5B). To find out whether there exists preferential TRAJ usage within the TRAV, we calculated the TRAJ recombination frequency for each TRAV as percentage of all recombinants and plotted these against TRAJ gene segments, ranked in order of their expression frequency. This revealed an improved correlation (Spearman coefficient R^2 0.38), indicating that recombination also depends on TRAJ usage, albeit clearly less strikingly than on TRAV usage (Figure 5C). Essentially, the same findings were obtained when Seq. 2 data were used (Supplementary Figure S2C–E).

To appreciate the distribution of TRAV-TRAJ recombination locus-wide, we plotted their frequency against the TRAV segments ordered by their real chromosomal distances, spanning nearly 1.5 Mb (Supplementary Figure S3A). A striking correlation between frequency and TRAV usage was observed, but not between the chromosomal position of TRAV segments

Figure 4 Comparison of TRAV-TRAJ recombinant frequencies observed in the original, N and duplicate loci. (A–C) The frequencies of TRAV-TRAJ gene segment recombinants observed in the original (A), N (B) and duplicate (C) loci are represented as Figure 2B for their common TRAV gene segments. The colours indicate the recombination frequencies, as indicated in the bar. (D–F) The fold differences of TRAV-TRAJ recombinants between the original and N (D), the original and duplicate (E) or the N and duplicate loci (F) were calculated and represented as fold changes on the z axis in 3D plots, in which TRAV and TRAJ functional gene segments are listed on the x and y axis, respectively (as in A). For x and y axis numbering, see Supplementary Figure S1. The size of the spheres represents the number of TRAV-TRAJ recombinants and their colours designate the differences, namely: in (D) blue spheres indicate ≥ 2 fold higher frequencies of recombinants in the N than in the original locus and orange spheres indicate ≥ 2 fold higher frequencies in the original than in the N locus; in (E) blue spheres represent ≥ 2 fold higher frequencies in the duplicate than in the original locus and orange spheres ≥ 2 fold higher frequencies in the original than in the duplicate locus; in (F) blue spheres indicate ≥ 2 fold higher frequencies in the duplicate than in the N locus and orange spheres ≥ 2 fold higher frequencies in the N than in the duplicate locus. In all, graphs black spheres indicate no changes; yellow, red and green spheres recombinants that are present only in the original, N and duplicate locus, respectively.

and their recombination frequency. The TRAV segments exhibiting high recombination frequencies ($\geq 2\%$) were mostly located in the N and duplicate loci, whereas for most TRAV segments of the original and duplicate loci the recombination frequencies were low ($<0.5\%$). Similar analysis of the TRAJ segments showed that most had recombination frequencies in the range of 1–4%, with the notable exception of TRAJ30 (21) and TRAJ22 (26), which were found in over 8% of all recombinants (Supplementary Figure S3B). The TRAJ usage for given TRAV in TRAV-TRAJ recombination generally increased with their recombination frequency; for example, it was remarkably diverse for TRAV16N (43) and V12D-2 (84), which exhibited the most frequent recombination (Figure 5D). However, this correlation was not very tight, since the TRAJ distribution was tighter for some (e.g., TRAV13N-4 (42) and TRAV13D-1 (89)) and looser for others (e.g., TRAV17 (3), 19 (2) and 21 (1)), irrespective of the recombination frequency. Similar analysis performed for TRAV usage for given TRAJ segment in TRAV-TRAJ recombination showed an overall tighter correlation between the TRAV distribution and TRAV-TRAJ recombination frequency (Figure 5E).

Taken collectively, our data indicate that (i) TRAV-TRAJ recombination frequencies strongly depend on the TRAV and to a lesser degree on the TRAJ usage; (ii) the highest frequencies were observed for TRAV segments of the N locus and a few TRAV segments of the duplicate locus and (iii) there was no obvious link between the recombination frequency and the location of TRAV segments neither in the sublocus nor in the locus.

Analysis of the junctional diversity

The diversity of the TCR α chain repertoire relied mainly on the hypervariable CDR3 α region. Global junctional diversity can be represented by CDR3 α length distribution (spectratyping) (Cabaniols *et al*, 2001), which showed a length distribution in the range of 30–54 bp (10–18 amino acids) with a maximum at 42 bp (14 amino acids) (Figure 6A). To assess the distribution of junctional diversity, we plotted the frequency of the different CDR3 α sequences against the TRAV and TRAJ segments listed in order of their chromosomal localization (Figure 6B). This 3D landscape resembled that of the TRAV-TRAJ recombination frequencies (Figure 3B), suggesting a link between recombination frequency and the

Figure 5 TRAV gene segment usage correlates with TRAV-TRAJ recombinant frequencies. (A, B) The TRAV (A) and TRAJ (B) gene segments are listed on the x axis in order of ascending expression frequency and plotted against the number of their TRAV-TRAJ recombinants (counts; y axis). (C) The percentages of TRAV-TRAJ recombinants were calculated for each TRAV segment (y axis) and plotted against the TRAJ segments (x axis) ranked as in (B). The median values are represented as coloured dots and the Spearman coefficient R^2 as numerical inserts. (D) For each TRAV segment, listed on the x axis in order of their chromosomal location, the percentages of their TRAV-TRAJ recombinants were calculated for each TRAJ and plotted on the y axis as 25/75 (box), 50 (inner lines) and 10–90 percentiles (whisker). For colour coding and TRAV segment numbering, see Supplementary Figure S1. (E) Analogous analysis was performed for the TRAV segment distribution in TRAV-TRAJ recombinants (y axis) and represented likewise for TRAJ gene segments.

diversity of CDR3 α sequences. Indeed, plotting the number of different CDR3 α sequences against the number of TRAV-TRAJ recombinants showed an overall good linear relationship between TRAV-TRAJ recombination frequency and CDR3 α diversity with a Spearman correlation coefficient of 0.807 (Figure 6C). However, this graph also revealed deviations from a linear correlation, since some recombinants exhibited high frequencies and low CDR3 α sequence diversity and a few others the inverse. Essentially, the same results were obtained when using data from Seq. 2 (Supplementary Figure S4).

For more detailed analysis, we plotted the frequency of unique CDR3 α sequences against the TRAV segments ordered by their real chromosomal distances (Supplementary Figure S5A). The distribution of CDR3 α sequences over the TRAV locus was reminiscent of that for the TRAV-TRAJ recombinant frequencies (Supplementary Figure S3A), with a few notable exceptions; for example, for TRAV14D-3 (68) and TRAV12D-2 (84), the frequencies of CDR3 α sequences were higher than the corresponding TRAV-TRAJ recombinants. Similar findings were obtained when comparing the distribution of CDR3 α sequences or TRAV-TRAJ recombinants plotted against the real chromosomal distances of TRAJ genes (Supplementary Figures S3B and S5B). Again, while these distributions overall were similar, some

deviations were evident, for example, for TRAJ52 (5) and TRAJ30 (21).

For more stringent comparison, we calculated for each TRAV gene segment the ratio between the number of unique CDR3 α sequences and the frequency of TRAV-TRAJ recombinants. This revealed considerable, TRAV sublocus-specific variations, for TRAV segments (Figure 7A). For the majority of TRAV segments, the ratio was larger than one, that is, there were more CDR3 α sequences than TRAV-TRAJ recombinants. For a few TRAV gene segments the inverse was true, most prominently so for TRAV of the N sublocus. Conversely, for most of the TRAV gene segments of the original sublocus the average median values were around four and the 10–90 percentiles often above ten. The average median values for the unique, original, N and duplicate TRAV subloci were 1.44, 2.19, 1.05 and 1.46, respectively. These TRAV sublocus-specific variations in CDR3 α sequences were also observed when plotting for each TRAV segment the frequencies of the CDR3 α sequences against those of the TRAV-TRAJ recombinants (Figure 7B). In this representation, for the points below the linear regression, the frequency of CDR3 α sequences was lower than the frequency of TRAV-TRAJ recombinants; most of these points were red, that is, corresponded to TRAV segments of the N sublocus. Conversely, most of the points

Figure 6 Junctional (CDR3 α) diversity: on overview. **(A)** The numbers of unique CDR3 α sequences (counts, y axis) are plotted against the CDR3 α lengths (in bp; x axis). **(B)** The frequencies of unique CDR3 α sequences were calculated for each TRAV-TRAJ recombinant (z axis) and plotted against the TRAV (x axis) and TRAJ gene segment (y axis), listed in order of their chromosomal localization (Supplementary Figure S1). The frequencies of CDR3 α sequences are represented by a colour code, as indicated by the bar. **(C)** The numbers of unique CDR3 α sequences (y axis) are plotted against the number of TRAV-TRAJ recombinants (x axis). The blue line represents a linear correlation between the number of TRAV-TRAJ recombinants and unique CDR3 α sequences and the numerical insert the corresponding Spearman coefficient R^2 .

above the line, for which the frequency of CDR3 α sequences was higher than the frequency of TRAV-TRAJ recombinants, were yellow, that is, corresponded to TRAV segments of the original sublocus. Analogous analysis performed for the TRAJ genes showed an overall better correlation between junctional diversity and TRAV-TRAJ recombination frequency (Figure 7C and D), arguing that the generation of junctional diversity depends more on TRAV than on TRAJ segment usage.

To further analyse the impact of TRAV segment usage on CDR3 α diversity, we calculated the number of unique and shared CDR3 α sequences for TRAV segments belonging to the original, N, duplicate loci or the unique 3' and 5' terminal regions, and represented these in a Venn diagram. This revealed that the distribution of CDR3 α sequences was remarkably TRAV specific, namely that most of the CDR3 α sequences were unique for a given TRAV sublocus (Supplementary Figure S6A). The recombinants of TRAV segments from the original locus constituted 34.7% of all CDR3 α sequences, but only 21.3% of all TRAV-TRAJ recombinants (Supplementary Figure S6B). Conversely, TRAV gene segments from the N locus constituted merely 25% of unique CDR3 α sequences, but 36.7% of all TRAV-TRAJ recombinants. Taken collectively, these results indicate that the junctional diversity correlated overall with the frequency of

TRAV-TRAJ recombination, but that there were deviations that depended on TRAV, much less than TRAJ gene segment usage, arguing that nucleotide insertions and deletions during the recombination depend to some extent on the genomic organization of the TRAV locus.

To obtain an estimate of the total TCR α chain repertoire diversity of naive CD8⁺ T cells, we used an approach frequently applied to the unseen species problem (Ionita-Laza *et al*, 2009; Robins *et al*, 2009). The estimation was based on the distribution of observed frequencies of each TCR sequence to determine whether new (as yet unobserved) sequences are present in the population. The more the frequency distribution is right skewed, that is, there are proportionally more single sequences (singletons) in the observed population, the more likely it is that new sequences will be discovered upon resampling, leading to a higher estimated total population. By using the Chao 1984, ChaoLee 1992 and jackknife functions provided in the SPECIES R package (Wang, 2011), with the default parameters recommended by the authors, we calculated $(2.36 \pm 0.23) \times 10^6$ distinct TCR α chain sequences. However, this value is not reliable because these functions are sensitive to the number of observed singletons, some of which result from Illumina sequencing errors (Nakamura *et al*, 2011; Nguyen *et al*, 2011). Estimation of the sequencing error rate

Figure 7 Relation between CDR3 α diversity and TRAV and TRAJ gene segment usage. **(A)** The ratio (x axis) between the number of unique CDR3 α sequences (% of total) and the frequency of TRAV-TRAJ recombinants (% of total) was calculated for each TRAV and plotted against the TRAV segments, listed in order of their chromosomal location as 25/75 (box), 50 (inner lines) and 10–90 percentiles (whisker). The colours indicate the TRAV subloci (see Supplementary Figure S1). **(B)** The frequency of TRAV-TRAJ recombinants for given TRAV gene segments (x axis, in % of total) are plotted against the frequency of unique CDR3 α sequences (y axis, in % of total). The colours of the dots represent the different TRAV segment loci and the blue line the linear regression of all data points. **(C, D)** Analogous analyses were performed for TRAJ gene segments.

would require further independent sequencing experiments and technical replicates. However, even if one assumes that all singletons were sequencing errors the number of remaining unique TCR α sequences is $(1.75 \pm 0.04) \times 10^5$, which is considerably higher than the murine TCR α chain repertoire previously estimated to be $0.79\text{--}1.18 \times 10^4$, due to rearrangement restriction (Cabaniols *et al*, 2001; Pasqual *et al*, 2002).

Discussion

To obtain a comprehensive view of the pre-immune mouse TCR α chain repertoire, we performed deep sequencing of RNA derived from B6 peripheral CD8 $^+$ T cells. The observed TCR α chain repertoire comprised nearly all (94%) possible TRAV-TRAJ recombinants. This high diversity was attributed to nearly complete usage of the TRAJ and TRAV libraries (Figures 2C, 3A and B). Only a few TRAJ genes were infrequently used, namely the most 3' and 5' end ones (Figures 2C and 3B; Supplementary Figure S3B). The paucity of the former is probably explained by the fact that they are preferentially recombined in primary rearrangements and deleted in the course of subsequent rearrangements (Huang and Kanagawa, 2001; Pasqual *et al*, 2002; Fuschiotti *et al*, 2007; Krangel, 2009). There were four TRAV gene segments distributed across the TRAV locus that failed to recombine

with several TRAJ genes (TRAV6.3 (38), 6N-5 (32), 16D (70) and 11D (92)) (Figure 3A). These holes in the repertoire are accounted for, at least in part, by thymic selection in other lineages, for example, NKT T cells, CD4 $^+$ T cells or IEL (MacDonald and Mycko, 2007; Jenkins *et al*, 2010).

The observed high diversity of the pre-immune TCR α chain repertoire, especially the complete use of TRAJ and TRAV libraries, is not consistent with the widely accepted view that TRAV-TRAJ recombinations proceed in a coordinated, sequential manner from proximal to progressively more distal TRAV and TRAJ segments (Roth *et al*, 1991; Huang and Kanagawa, 2001; Pasqual *et al*, 2002; Fuschiotti *et al*, 2007; Jouvin-Marche *et al*, 2009; Krangel, 2009). As a consequence of sequential coordinate recombinations, TRAV and TRAJ library usages are restricted, which substantially reduces the number of realizable recombinations. Our comprehensive TCR α chain sequencing clearly shows that recombination of TRAV and TRAJ gene segments is not restricted by their genomic location, for example, distal TRAV can efficiently rearrange with proximal TRAJ segments and *vice versa*. Explanations for the diverging conclusions between our and previous studies include: (i) In previous studies typically only a few TRAV segments were examined; often these comprised distal unique TRAV (e.g., TRAV2 (108) or TRAV6–1 (105)) and proximal unique TRAV segments

(TRAV21 (1), TRAV 19 (2) and TRAV17 (3)). Because the former TRAV preferentially recombine with relatively distal TRAJ, and the latter with proximal TRAJ segments (Figures 2C and 3B), it seems hazardous to derive from such limited analysis general conclusions for locus-wide rearrangements, as these can vary greatly for different TRAV (Figures 2–4; Supplementary Figure S3A). (ii) In older studies, often TRAV-specific primers were used that are specific not for one TRAV gene segment, but for a group of segments that may have considerably different chromosomal locations, which compromises the validity of the conclusions reached. In fact, it was discovered only in 2009 that the mouse TRAV locus contains two duplications and not one, as had been accepted previously (Supplementary Figure S1; Church *et al*, 2009). (iii) In many previous studies, TRAV-TRAJ recombinant analysis was focussed on TRAJ usages for a few TRAV segments. Such data may not provide conclusive support for general coordinate and sequential TRAV-TRAJ gene segment recombination, because the TRAJ locus contains regulatory elements that favour progression of secondary rearrangements from its 5' to 3' end (Hawwari *et al*, 2005; Abarrategui and Krangel, 2006; Hawwari and Krangel, 2007; Krangel, 2009).

There is compelling evidence that primary recombinations are focussed on proximal TRAV and TRAJ gene segments (Villey *et al*, 1996; Huang and Kanagawa, 2001; Mauvieux *et al*, 2001; Yannoutsos *et al*, 2001; Guo *et al*, 2002; Hawwari *et al*, 2005; Abarrategui and Krangel, 2006; Hawwari and Krangel, 2007; Krangel, 2009). On CD71⁺ DP thymocytes, the E δ enhancer activates TEA and promoters of proximal TRAJ and TRAV segments, resulting in transcription, histone modifications, chromatin remodelling, which favour coordinate rearrangements of proximal TRAJ and TRAV genes. Consistent with this, we observed that primary recombinations, that is, those containing TRAV21 (1) recombined with 5' TRAJ genes with high preference (Figure 2A). Quite the same was observed for the adjacent TRAV19 (2) and 17 (3), which are also under the control of the E α enhancer (Figure 2C). These recombinants were the only ones exhibiting a pattern consistent with coordinated sequential gene rearrangement and even this not strictly, as proximal TRAV could recombine with very distal TRAJ segments (Figures 2 and 3). Importantly, when considering all transcripts these rearrangements constituted only a scant fraction.

Because the frequency of TRAV-TRAJ recombination critically depended on TRAV segment usage, our data argue that this reflects TRAV accessibility to recombination centres (Figures 4 and 5; Supplementary Figure S3). We proposed that overcoming the vast distances separating distal TRAV segments from recombination centres relies on large-scale chromatin contraction and dynamic DNA rosette formations, similarly to that shown for IgH gene rearrangements (Jhunjhunwala *et al*, 2008, 2009). In pre-pro-B cells, chromatin exhibits dynamic formation of DNA rosettes, separated by linkers of variable size. Subsequently in pro-B cells, further chromatin contraction in the IgH locus takes place, enabling recombination between V_H and D_HJ_H segments separated by large genomic distance with high frequency, thus allowing the formation of highly diverse IgH repertoires. Proteins involved in this mechanism include Pax5, E2A and CTCF/cohesion (Fuxa *et al*, 2004; Degner *et al*, 2011; Ebert *et al*, 2011). CTCF and cohesion have been shown to be

involved in cooperatively mediating long distance chromatin loop formation and regulating transcription (Degner *et al*, 2009, 2011; Degner-Leisso and Feeney, 2010; Seitan *et al*, 2011). The view that secondary TRAV-TRAJ rearrangements rely on analogous DNA looping of the TCR α locus is consistent with the observations that on DP thymocytes the TCR α locus is contracted, as assessed by 3D FISH experiments (Skok *et al*, 2007). Moreover, Rad21, a component of the cohesion complex, has been shown to be required for secondary TRAV-TRAJ recombination, transcription and H3K histone trimethylation, the formation of recombination centres and long distance DNA looping (Seitan *et al*, 2011). Cohesion has been shown to be dispersed over the entire TRAV locus (Seitan *et al*, 2011), which is likely also the case for CTCF binding sites, as has been shown for the V_H locus and because cohesion binds to CTCF (Degner-Leisso and Feeney, 2010; Ohlsson *et al*, 2010; Degner *et al*, 2011; Seitan *et al*, 2011).

Although recombination frequencies varied mainly with TRAV segment usage, there were also clear, albeit less striking variations with TRAJ usage, suggesting a second level of recombination regulation at the level of TRAJ segments (Figure 5; Supplementary Figure S3B). The most 5' TRAJ segments (TRAJ 50–58 (6–1)) exhibited low recombination frequencies (Supplementary Figure S3B), which is consistent with their preferential usage in primary recombination, and deletion during subsequent secondary rearrangements (Hawwari and Krangel, 2007; Krangel, 2009). The most frequent recombinations were observed for TRAJ segments in the 3' half portion of the TRAJ locus, namely TRAJ 30 (21) and TRAJ 22 (26) making up nearly 8% of all recombinations (Figure 3B; Supplementary Figure S3B). These TRAJ segments become amenable to recombination after deletion of the TEA and 5' TRAJ promoters (Hawwari *et al*, 2005; Abarrategui and Krangel, 2006; Hawwari and Krangel, 2007), supporting the view that extensive and wide ranging TRAV and TRAJ recombination occurs after primary rearrangements. Interestingly, the usage of some TRAJ segments was linked to TRAV segment usage. For example, while TRAV gene segments in the middle of the duplicate locus efficiently recombined with all TRAJ, those at the 5' end of the original locus recombined poorly with 5' TRAJ (Figures 2C and 3B). These differences may be explained by recombination centres that are formed in different locations in the TRAJ locus or in different topologies, providing different recombination probabilities for different arrays of TRAV segments. Cohesin may be involved in this, because in Rad21^{−/−} thymocytes the TRAJ locus exhibited a dramatic reduction in H3K4me3 sites and binding of RAG-1 and -2 (Seitan *et al*, 2011).

Based on these observations, we propose an alternative model for secondary TRAV-TRAJ recombination that operates on CD71-small DP thymocytes after deletion of the long intervening DNA and that was not recognized in previous studies, many of which used RAG KO systems, in which this deletion does not take place. After deletion of TEA and promoters of 5' TRAJ segments, downstream TRAJ genes become activated (Hawwari *et al*, 2005). This renders RSS of TRAJ genes accessible for RAG-1 binding and promotes trimethylation of H3K4 (H3K4me3) that are binding sites for RAG-2 (Matthews *et al*, 2007), thus promoting the formation of recombination centres in the TRAJ locus; which in turn mediate recombination of TRAJ with accessible TRAV

gene segments (Ji *et al*, 2010a,b; Schatz and Ji, 2011). This model is supported by several findings: (i) TRAV genes can recombine with virtually every TRAJ segment, irrespective of their chromosomal location, which is not possible with sequential coordinate TRAV-TRAJ recombination (Figures 2 and 3). (ii) The recombination frequency varied considerably depending on the TRAV segment location (Figures 2C and 3B; Supplementary Figure S3A) in a pattern reminiscent of the distribution of cohesion binding sites in the TRAV locus (Seitan *et al*, 2011). (iii) The overall good correlation between the recombination frequency and junctional sequence diversity strongly suggests that the accessibility of TRAV segments to recombination centres determines the probability of recombination and hence the generation of CDR3 α diversity (Figures 6C and 7; Supplementary Figures S3 and S5).

It is noteworthy, however, that there were deviations between recombination frequency and CDR3 α sequence diversity (Figures 6C and 7; Supplementary Figure S6), arguing that the mechanisms mediating nucleotide insertions and deletions in the course of non-homologous end joining are not stochastic, but depend on the genomic location of TRAV segments and the genomic organization of the TCR $\alpha\delta$ locus (Bassing *et al*, 2002; Gellert, 2002). Interestingly, recombination of the proximal unique TRAV gene segments exhibited remarkably low CDR3 α sequence diversity (Figures 2C, 3B, 6B and 7A; Supplementary Figure S6). Because these TRAV segments are mainly found in primary rearrangements, it seems that these yield lower junctional diversity than secondary recombination.

A key argument for coordinate sequential gene recombination is that it prevents premature exhaustion of TRAV and TRAJ libraries (Jouvin-Marche *et al*, 2009; Krangel, 2009). According to our model, it is indeed possible that distal TRAV segments recombine in secondary rearrangements, resulting in deletion of intervening TRAV gene segments and hence limitation of the number of possible rearrangements. This is alleviated by the fact that recombination takes place on both alleles and by thymic-positive selection. Positive selection is an extraordinarily wasteful process, in which 95–97% of DP thymocytes, produced continuously in copious numbers, are deleted. DP thymocytes have a lifespan of about 3.5 days, in which they attempt several rounds of TCR α chain recombination. Positive selection imposes severe criteria on TCR α gene recombination, namely that it needs to generate functional TCR/CD3 complexes, capable of recognizing an endogenous MHC-peptide ligand on the thymic epithelium (Huesmann *et al*, 1991; Kisielow and von Boehmer, 1995; Yannoutsos *et al*, 2001; Starr *et al*, 2003; Jenkins *et al*, 2010). This necessitated the evolution of TRAV and TRAJ loci and recombination mechanisms capable of generating highly diverse TCR α chain repertoires. Our model allows this, however at a price, namely the high frequency of apoptosis of DP thymocytes, that is, of failed attempts to generate selecting TCR. By contrast, sequential coordinated recombination implies restricted TRAV and TRAJ gene segment usage and hence lower TCR α chain diversity (Pasqual *et al*, 2002; Jouvin-Marche *et al*, 2009; Krangel, 2009), which is expected to compromise the generation of selecting TCR and to put the organism at risk of failing to mount protective immunity.

Our view that the TCR α chain repertoire is at least as vast as the TCR β chain repertoires is supported by high-throughput TCR sequencing on human T cells (Wang *et al*, 2010) and

a study in which TCR α and β chains were sequenced on single CD8 $^{+}$ T cells (Dash *et al*, 2011). The latter study further indicated that the *de-facto* TCR α diversity is significantly increased by 10% or more of the cells having two in-frame TCR α transcripts. It should be noted that our TCR α chain repertoire diversity calculations underestimate the diversity generated on DP thymocytes due to deletions of specificities by negative thymic selection, unsuccessful pairing with the preformed TCR β chain and selection of thymocytes to other lineages.

In conclusion, we performed comprehensive TCR α chain repertoire analysis on naive CD8 $^{+}$ T cells and observed a much higher diversity than previous studies have indicated. Nearly all possible TRAV-TRAJ recombinations were observed, each of which exhibited variable, often high junctional diversity. Our results are incompatible with the widely accepted view that TRAV-TRAJ recombination follows sequential coordinated rearrangements of proximal to gradually more distal TRAV and TRAJ gene segments. Instead they strongly argue that secondary TRAV-TRAJ rearrangements rely on TCR α locus contraction and DNA looping that provides TRAV segments opportunities to recombine with TRAJ segments, irrespective of their genomic location. This model bears close similarities to VDJ gene recombination at the IgH locus and explains the observed nearly complete usage of TRAJ and TRAV libraries and generation of pre-immune TCR α chain repertoires orders of magnitude more diverse than previously reported. We argue that the generation of highly diverse TCR α chain repertoires is necessary for the formation of broad TCR $\alpha\beta$ repertoires, which in turn are required for providing broad protective immunity.

Materials and methods

Mice

Seven-week-old C57BL/6 (B6) mice were purchased from Harlan laboratory (Horst, The Netherlands).

RNA preparation

CD8 $\alpha\beta^{+}$ T cells (80×10^6) were purified from the spleen of five mice by negative selection, according to the manufacturer instruction (Mouse CD8 $^{+}$ T-cell enrichment kit; StemCell Technologies, Grenoble, France). The total RNA was then extracted using Trizol purification.

In-vitro transcription

We took the advantage of the linear amplification of the *in-vitro* transcription (IVT) to amplify the TCR (Polacek *et al*, 2003). In all, 1 μ g of total RNA was used for TCR amplification, using specific primer in the constant region fused with the T7 promoter. The IVT was performed according to the instructions of the MessageAmp II aRNA Amplification Kit (Ambion, Rotkreuz, Switzerland).

Illumina sequencing

A specific ssDNA adapter (adenyl-AGATCGGAAGAGCGGTTCAGddC) was ligated to the 3' end of the RNA fragments using truncated T4 RNA ligase. The first strand cDNA was then synthesized using a primer (CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT) complementary to the ssDNA adapter. The library was prepared by adding the Illumina adapters sequences by PCR amplification using a primer specific to the 5' constant region (AATGATACGGCGACCACCGAGATCTACACAGCAGGTTCTGGGTCTG) and the standard Illumina PCR primer. A quality control analysis was done by cloning a fraction of the library in a TOPO vector and sequencing a few clones. The paired-end 100 bp high-throughput sequencing on the Illumina 'Hi-Seq 2000' was performed following Illumina's standard protocol.

Illumina sequencing read analysis

Sequences of the different reads were identified by an all-by-all sequence alignment to published TRAJ http://www.imgt.org/IMGTrepertoire/index.php?section=LocusGenes&repertoire=genetable&species=Mus_musculus&group=TRAJ or TRAV http://www.imgt.org/textes/IMGTrepertoire/LocusGenes/index.php?repertoire=genetable&species=Mus_musculus&group=TRAV sequences, respectively, using align0 (Myers and Miller, 1988) on the command line with the following parameters: $-f -24 -g -1$ (penalties for gap opening and gap extension). TRAJ and TRAV assignment was determined by the best global alignment match score to the reference sequences. All TRAJ sequences could be determined unambiguously. When two or more TRAV shared the best match score, sequence counts were attributed to the matched (ambiguous) TRAV according to the ratio of the unambiguous reads assigned to each TRAV for the whole sequencing sample. Because the nucleotide sequences of TRAV12N-2 (56), TRAV14N-2 (53) and TRAV14N-3 (41) are identical to those of TRAV12D-2 (84), TRAV14D-2 (80) and TRAV14D-3 (68), respectively, the counts were distributed equally to each pair. The align0 output was processed using *ad hoc* Perl scripts to: (i) identify the TRAJ and TRAV sequence within each read in order to determine the composition of each TRAV-TRAJ recombination; (ii) classify the TRAV-TRAJ-TRAC recombination as being in-frame or out-of-frame; (iii) determine the sequence and the length of each unique CDR3 region; (iv) determine the amino-acid sequence of each unique CDR3 region; (iv) determine the abundance (sequence count) of the CDR3 region identified for each TRAV-TRAJ recombination event.

References

- Abarrategui I, Krangel MS (2006) Regulation of T cell receptor-alpha gene recombination by transcription. *Nat Immunol* **7**: 1109–1115
- Bassing CH, Swat W, Alt FW (2002) The mechanism and regulation of chromosomal V(D)J recombination. *Cell* **109**(Suppl): S45–S55
- Bassing CH, Tillman RE, Woodman BB, Canty D, Monroe RJ, Sleckman BP, Alt FW (2003) T cell receptor (TCR) alpha/delta locus enhancer identity and position are critical for the assembly of TCR delta and alpha variable region genes. *Proc Natl Acad Sci USA* **100**: 2598–2603
- Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM (2001) Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med* **194**: 1385–1390
- Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, Bult CJ, Agarwala R, Cherry JL, DiCuccio M, Hlavina W, Kapustin Y, Meric P, Maglott D, Birtle Z, Marques AC, Graves T, Zhou S, Teague B, Potamocis K *et al* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biol* **7**: e1000112
- Dash P, McClaren JL, Oguin III TH, Rothwell W, Todd B, Morris MY, Becksfort J, Reynolds C, Brown SA, Doherty PC, Thomas PG (2011) Paired analysis of TCRalpha and TCRbeta chains at the single-cell level in mice. *J Clin Invest* **121**: 288–295
- Davodeau F, Difilippantonio M, Roldan E, Malissen M, Casanova JL, Couedel C, Morcet JF, Merckenschlager M, Nussenzweig A, Bonneville M, Malissen B (2001) The tight interallelic positional coincidence that distinguishes T-cell receptor Jalpha usage does not result from homologous chromosomal pairing during ValphaJalpha rearrangement. *EMBO J* **20**: 4717–4729
- Degner SC, Verma-Gaur J, Wong TP, Bossen C, Iverson GM, Torkamani A, Vettermann C, Lin YC, Ju Z, Schulz D, Murre CS, Birshtein BK, Schork NJ, Schlissel MS, Riblet R, Murre C, Feeney AJ (2011) CCCTC-binding factor (CTCF) and cohesin influence the genomic architecture of the Igh locus and antisense transcription in pro-B cells. *Proc Natl Acad Sci USA* **108**: 9566–9571
- Degner SC, Wong TP, Jankevicius G, Feeney AJ (2009) Cutting edge: developmental stage-specific recruitment of cohesin to CTCF sites throughout immunoglobulin loci during B lymphocyte development. *J Immunol* **182**: 44–48
- Degner-Leisso SC, Feeney AJ (2010) Epigenetic and 3-dimensional regulation of V(D)J rearrangement of immunoglobulin genes. *Semin Immunol* **22**: 346–352
- Ebert A, McManus S, Tagoh H, Medvedovic J, Salvagiotto G, Novatchkova M, Tamir I, Sommer A, Jaritz M, Busslinger M (2011) The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* **34**: 175–187
- Fuschiotti P, Pasqual N, Hierle V, Borel E, London J, Marche PN, Jouvin-Marche E (2007) Analysis of the TCR alpha-chain rearrangement profile in human T lymphocytes. *Mol Immunol* **44**: 3380–3388
- Fuxa M, Skok J, Souabni A, Salvagiotto G, Roldan E, Busslinger M (2004) Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes Dev* **18**: 411–422
- Gellert M (2002) V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem* **71**: 101–132
- Guo J, Hawwari A, Li H, Sun Z, Mahanta SK, Littman DR, Krangel MS, He YW (2002) Regulation of the TCRalpha repertoire by the survival window of CD4 (+) CD8 (+) thymocytes. *Nat Immunol* **3**: 469–476
- Hawwari A, Bock C, Krangel MS (2005) Regulation of T cell receptor alpha gene assembly by a complex hierarchy of germline Jalpha promoters. *Nat Immunol* **6**: 481–489
- Hawwari A, Krangel MS (2005) Regulation of TCR delta and alpha repertoires by local and long-distance control of variable gene segment chromatin structure. *J Exp Med* **202**: 467–472
- Hawwari A, Krangel MS (2007) Role for rearranged variable gene segments in directing secondary T cell receptor alpha recombination. *Proc Natl Acad Sci USA* **104**: 903–907
- Huang C, Kanagawa O (2001) Ordered and coordinated rearrangement of the TCR alpha locus: role of secondary rearrangement in thymic selection. *J Immunol* **166**: 2597–2601
- Huesmann M, Scott B, Kisielow P, von Boehmer H (1991) Kinetics and efficacy of positive selection in the thymus of normal and T cell receptor transgenic mice. *Cell* **66**: 533–540
- Ionita-Laza I, Lange C, M Laird N (2009) Estimating the number of unseen variants in the human genome. *Proc Natl Acad Sci USA* **106**: 5008–5013
- Jenkins MK, Chu HH, McLachlan JB, Moon JJ (2010) On the composition of the preimmune repertoire of T cells specific for Peptide-major histocompatibility complex ligands. *Annu Rev Immunol* **28**: 275–294
- Jhunjunwala S, van Zelm MC, Peak MM, Cutchin S, Riblet R, van Dongen JJ, Grosveld FG, Knoch TA, Murre C (2008) The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**: 265–279
- Jhunjunwala S, van Zelm MC, Peak MM, Murre C (2009) Chromatin architecture and the generation of antigen receptor diversity. *Cell* **138**: 435–448

The size of the CDR3 α was calculated relative to the conserved C residue of the germinal TRAV and the conserved FG residues of the germinal TRAJ segments. The deep sequencing data can be found in the Gene Expression Omnibus (GEO) public database (<http://www.ncbi.nlm.nih.gov/geo/>) under the accession number GSE35626.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (<http://www.embojournal.org>).

Acknowledgements

This work was supported by grant 310030_12533/1 of the Swiss National Science Foundation (SNF). We are grateful to Drs J Leignadier and C Iseli for ideas and discussions.

Author contributions: RG designed and performed experiments, analysed the data and wrote the manuscript. BJS performed the bioinformatics analysis of the data and corrected the manuscript. LF and MO contributed to the development of the sequencing method and performed Illumina sequencing. IFL initiated the study, supervised the work and wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

- Ji Y, Little AJ, Banerjee JK, Hao B, Oltz EM, Krangel MS, Schatz DG (2010a) Promoters, enhancers, and transcription target RAG1 binding during V(D)J recombination. *J Exp Med* **207**: 2809–2816
- Ji Y, Resch W, Corbett E, Yamane A, Casellas R, Schatz DG (2010b) The *in vivo* pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell* **141**: 419–431
- Jouvin-Marche E, Fuschioti P, Marche PN (2009) Dynamic aspects of TCRalpha gene recombination: qualitative and quantitative assessments of the TCRalpha chain repertoire in man and mouse. *Adv Exp Med Biol* **650**: 82–92
- Kisielow P, von Boehmer H (1995) Development and selection of T cells: facts and puzzles. *Adv Immunol* **58**: 87–209
- Krangel MS (2009) Mechanics of T cell receptor gene rearrangement. *Curr Opin Immunol* **21**: 133–139
- MacDonald HR, Mycko MP (2007) Development and selection of Valpha 14i NKT cells. *Curr Top Microbiol Immunol* **314**: 195–212
- Matthews AG, Kuo AJ, Ramon-Maiques S, Han S, Champagne KS, Ivanov D, Gallardo M, Carney D, Cheung P, Ciccone DN, Walter KL, Utz PJ, Shi Y, Kutateladze TG, Yang W, Gozani O, Oettinger MA (2007) RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature* **450**: 1106–1110
- Mauvieux L, Villey I, de Villartay JP (2001) T early alpha (TEA) regulates initial TCRVAJA rearrangements and leads to TCRJA coincidence. *Eur J Immunol* **31**: 2080–2086
- Myers EW, Miller W (1988) Optimal alignments in linear space. *Comput Appl Biosci* **4**: 11–17
- Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**: e90
- Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL (2011) Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics* **12**: 106
- Ohlsson R, Bartkuhn M, Renkawitz R (2010) CTCF shapes chromatin by multiple mechanisms: the impact of 20 years of CTCF research on understanding the workings of chromatin. *Chromosoma* **119**: 351–360
- Pasqual N, Gallagher M, Aude-Garcia C, Loiodice M, Thuderoz F, Demongeot J, Ceredig R, Marche PN, Jouvin-Marche E (2002) Quantitative and qualitative changes in V-J alpha rearrangements during mouse thymocytes differentiation: implication for a limited T cell receptor alpha chain repertoire. *J Exp Med* **196**: 1163–1173
- Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert Jr CJ, Davies PF (2003) Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA. *Physiol Genomics* **13**: 147–156
- Robins HS, Campregher PV, Srivastava SK, Wachter A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* **114**: 4099–4107
- Roth ME, Holman PO, Kranz DM (1991) Nonrandom use of J alpha gene segments. Influence of V alpha and J alpha gene location. *J Immunol* **147**: 1075–1081
- Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. *Annu Rev Immunol* **24**: 419–466
- Schatz DG, Ji Y (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* **11**: 251–263
- Seitan VC, Hao B, Tachibana-Konwalski K, Lavagnoli T, Mira-Bontenbal H, Brown KE, Teng G, Carroll T, Terry A, Horan K, Marks H, Adams DJ, Schatz DG, Aragon L, Fisher AG, Krangel MS, Nasmyth K, Merkenschlager M (2011) A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature* **476**: 467–471
- Shih HY, Hao B, Krangel MS (2011) Orchestrating T-cell receptor alpha gene assembly through changes in chromatin structure and organization. *Immunol Res* **49**: 192–201
- Shih HY, Krangel MS (2010) Distinct contracted conformations of the Tcr α /Tcr δ locus during Tcr α and Tcr δ recombination. *J Exp Med* **207**: 1835–1841
- Skok JA, Gislser R, Novatchkova M, Farmer D, de Laat W, Busslinger M (2007) Reversible contraction by looping of the Tcr α and Tcr β loci in rearranging thymocytes. *Nat Immunol* **8**: 378–387
- Starr TK, Jameson SC, Hogquist KA (2003) Positive and negative selection of T cells. *Annu Rev Immunol* **21**: 139–176
- Villey I, Caillol D, Selz F, Ferrier P, de Villartay JP (1996) Defect in rearrangement of the most 5' TCR-J alpha following targeted deletion of T early alpha (TEA): implications for TCR alpha locus accessibility. *Immunity* **5**: 331–342
- von Boehmer H (2005) Unique features of the pre-T-cell receptor alpha-chain: not just a surrogate. *Nat Rev Immunol* **5**: 571–577
- Wang C, Sanders CM, Yang Q, Schroeder Jr HW, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson Jr JR, Davis RW, Han J (2010) High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci USA* **107**: 1518–1523
- Wang JP (2011) SPECIES: an R package for species richness estimation. *J Stat Softw* **40**: 1–15
- Yannoutsos N, Wilson P, Yu W, Chen HT, Nussenzweig A, Petrie H, Nussenzweig MC (2001) The role of recombination activating gene (RAG) reinduction in thymocyte development *in vivo*. *J Exp Med* **194**: 471–480