Major Research Project Thesis

By: Gabe van den Hoeven

Studentnumber: 8254567

Date: 5-11-2024

Research group: Theoretical Biology and Bioinformatics

Main supervisor: Rob de Boer

Secondary examiner: Can Kesmir

Layman summary

The immune system is built in such a way that it can differentiate the body's own cells from foreign cells, to keep out foreign bacteria and viruses. Most immune cells recognise foreign material with arm-like components on their cell surface called "receptors". To be able to respond to a large diversity in bacteria and viruses, an incredibly diverse set of specific receptors is made. This diversity is due to the highly variable binding regions of the receptors. These variable parts are made up of 3 segments of DNA that are joined together when the cell is born. The parts where these segments meet can be altered slightly: small parts can be broken off, and/or extra pieces can be added. This allows for at least 100000000 different receptors.

Prior research has found that 10% of receptors that occurred repeatedly in different individuals, lacked the middle of the 3 segments. Cells lacking this middle segment also occurred more often in embryo's, which led them to suspect that these receptors are made in early development. At that early time in development, the protein "TdT", which adds parts in between the segments is not active yet. They hypothesised that without these additions, the middle segment can be completely deleted. In this research project we tested this hypothesis. The value of this theoretical research is fundamental for understanding how these receptors of immune cells are being made, and why some lack the middle segment. Understanding the properties of these exceptional cells may later contribute to the development of immunotherapies.

To test the hypothesis, we analysed data of receptors sequences from immune cells in mice. One group of mice was genetically modified to not produce the protein TdT, the other group was used as a control to compare with. Using a computer algorithm, we determined the length of the middle segment in each receptor sequence in the data. As expected, the two groups differed in the extra parts that were added in between the segments. But we did not find a significant difference in the length of the middle segment between the two groups. We then verified that the results from the prior study (that proposed the hypothesis), could be confirmed in our dataset. We indeed found that 13% of the receptor sequences that were present in all 10 control mice, lacked the middle segment. This means that we have to discard the original hypothesis, and that, during early development, there is another process responsible for making receptors that lack the middle segment.

N-nucleotide additions by Terminal deoxynucleotidyl Transferase do not protect against deletion of the D segment during VDJ-recombination in T-cell receptors

T-cell receptor (TCR) diversity is fundamental to the immune system's ability to recognise foreign antigens. TCRs are put together in a semi-stochastic process called V(D)J-recombination, where the α-chain consists of a variable (V) and junction (J) gene segment, and the β-chain consists of a V, diversity (D), and J gene segment. During V(D)J-recombination one segment of each of these genes are recombined, with deletions by exonucleases and non-template nucleotide additions by a protein called Terminal deoxynucleotidyl Transferase (TdT) occurring at the junctions of these segments. Prior research has shown that some abundant β-chain sequences lack the D segment. We test their hypothesis that the absence of TdT may cause the deletion of the D segment. By comparing sequences of TdT knock-out and wild-type mice, we find that abundant β-chain sequences often have no D segment, but see no significant increase in the TdT knock-out group, suggesting that TdT does not protect against the deletion of the D segment. Additionally, our analyses revealed that V and J gene segment usage differs significantly between TdT knock-out and wild-type sequences, and that almost 60% of abundant wild-type sequences used either the TRBV1 or TRBV16 gene segment.

Introduction

T lymphocytes, or T-cells, are a crucial part of the adaptive immune system. When activated, they start proliferating to clear pathogens. By binding their T-cell receptors (TCRs) to peptides presented by the major histocompatibility complex they differentiate between self and foreign cells. To accommodate a large diversity of specific T-cell receptors, these receptors are configured semi-stochastically through a process called V(D)J-recombination. In this process, small segments of three genes are combined: variable (V), and junction (J) for the α-chain (TRA), and V, diversity (D), and J for the β-chain (TRB) into the complementary determining region 3 (CDR3) of the TCR. Double stranded breaks are introduced at the ends of these gene segments and subsequently, segments are put together in a so-called non-homologous end joining (NHEJ) reaction. V(D)J-recombination allows for an estimated repertoire diversity of at least 10^{20} (Mora & Walczak, 2016) of which around 10^{8} is actually attained in the TCR repertoire (Qi et al., 2014).

During V(D)J-recombination, the α-chain and β-chain of the TCR are subjected to what is called the 12/23 rule. This refers to the spacers of 12 or 23 base pairs (bp) in the residual signal sequences (RSSs) that flank the V, D, and J gene segments, and guide the recombination process. The RSSs play a significant role in determining which gene segments are used. Epigenetic regulation of the chromatin structure by histone modification of the nucleosome allows an RSS to be made accessible, or inaccessible, for binding to the recombination activating gene (RAG) protein complex, which binds to the RSSs to start the recombination (Feeny, 2009; Gopalakrishnan et al., 2013; Krangel, 2003, 2015; Sleckman et al., 2000; Stanhope-Baker et al., 1996). Two gene segments can only be combined if the spacers between their RSSs have different lengths. This is due to the 3-dimensional conformation RAG takes when it binds to the DNA to splice the DNA at the ends of the gene segments (Figure 1). The 5’ side of the D and J segments have a nonamer-12 bp-heptamer RSS, while the 3’ side of the V and D segments have a heptamer-23 bp-nonamer RSS (Ma et al., 2016). This typically binds the D to J and V to D. It is also possible to have VJ recombination’s in the β-chain without a D segment, but this is rarely seen in humans or mice, and called the beyond 12/23 restriction (Bassing et al., 2000; Tillman et al., 2003).

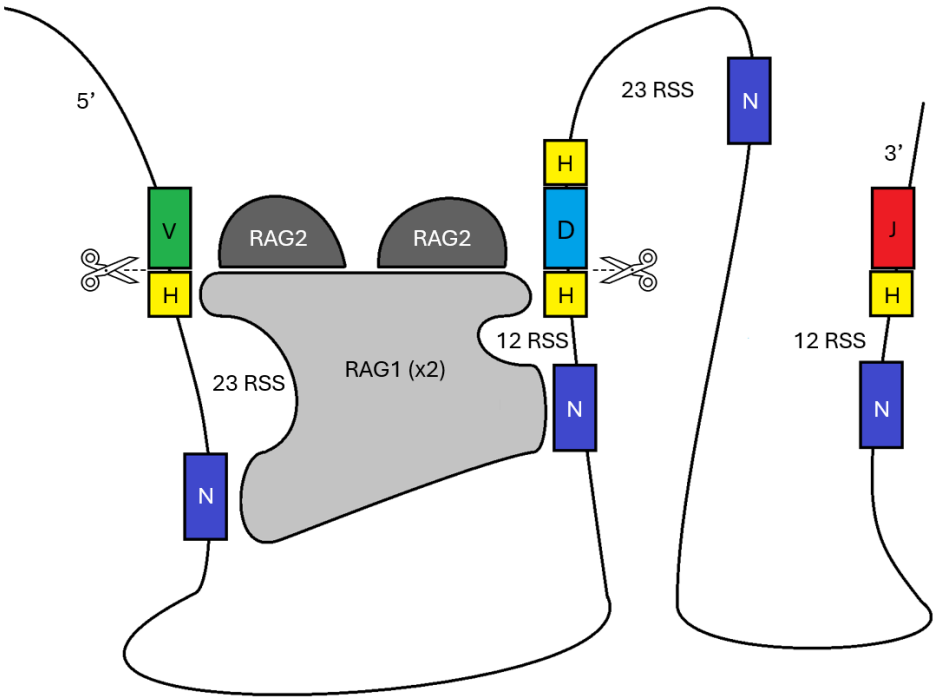

Figure 1: Visualisation of the RAG protein complex positioning near a V and D gene segment to initiate V(D)J-recombination. The RAG protein complex consists of 2 *RAG1* and 2 *RAG2* proteins. It is guided into position by the residual signal sequences (RSSs) that directly flank the gene segments. An RSS consists of a conserved heptamer (H) and nonamer (N) separated by a 12 or 23 bp spacer. The conformation of RAG restricts which gene segments can be recombined. RAG introduces a double strand break in between the gene segments and the heptamer of their RSS, triggering a NHEJ reaction.

After RAG cleaves the DNA in between the heptamers of the RSSs and the gene segments, the two strands at the end of each gene segment form a hairpin (stem-loop) structure while the ends of the RSS form a blunt end (Helmink & Sleckman, 2012). First, the D and J gene segments are recombined after which the V gene segment is recombined with the DJ sequence. NHEJ enzymes, including the protein Artemis, are recruited by DNA protein kinase to form a protein complex that phosphorylates and activates Artemis. Artemis then opens the hairpins by nicking the strands with single strand cuts for each gene segment, so that the 3' ends can be altered before the segments are joined (Figure 2) (Ma et al., 2005). If both strands are nicked, this results in the deletion of a few nucleotides. Alternatively, Artemis can nick only a single strand. Then, the covalent bond of the hairpin allows the nucleotides of the nicked strand to join the 3' side of the other strand, resulting in a short added palindromic sequence (Figure 2, left). The reverse complementary nucleotides that make up this palindromic sequence are called palindromic (P) nucleotides (Srivastava & Robins, 2012). Various exonucleases may delete nucleotides, while terminal deoxynucleotidyl transferase (TdT) may add non-templated (N) nucleotides to the 3' ends. After the pairing of the strands, the gaps in the sequence are filled with complementary nucleotides and ligated to form the new CDR3 sequence.

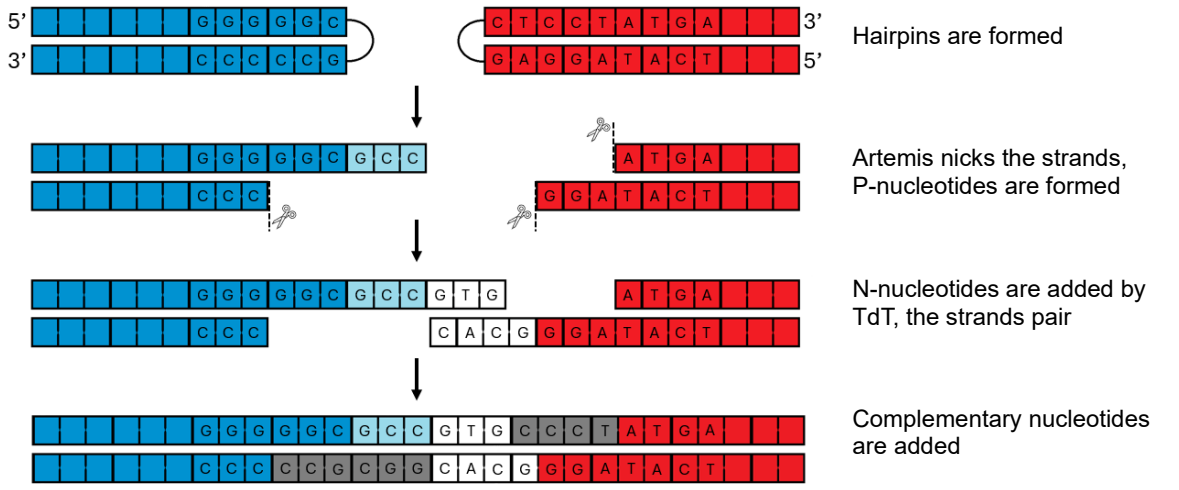

Figure 2: Visualisation of the NHEJ reaction between the CDR3 sequences of a D gene segment on the left in blue, and a J gene segment on the right in red. After RAG cleaves the DNA, a hairpin loop is formed between the strands at the ends of both gene segments. Here, the D segment is only nicked on one strand by Artemis creating a palindromic (P) sequence of 3 nt in light blue on the other strand. At the same time, 2 nt are deleted from the J segment by two single strand cuts. TdT adds non-templated (N) nucleotides to the 3' of each segment in white. After the pairing of the strands, the gaps are filled with complementary nucleotides in grey and the strands are ligated to form the new CDR3 sequence.

Even though TRB sequences usually contain a D segment, prior research provided evidence that TRB sequences that are relatively abundant in the repertoire are enriched in sequences without a D segment. These sequences persist over time and were primarily found at higher levels in neonate samples, suggesting they may be arising from an early stage in development when TdT is downregulated (De Greef & De Boer, 2021). De Greef and De Boer argue that due to this downregulation, the lack of N-additions may lead the exonucleases to delete parts, or the entirety of the D segment during the V(D)J-recombination process. However, Murugan et al. (2012) showed that probabilities of deletions and insertions during V(D)J-recombination are independent processes, which suggests deletions happen before the N-additions by TdT. Since this is only a suggestion, we here analyse characteristics of high incidence TRB CDR3 sequences to test the hypothesis put forward by De Greef & De Boer (2021). We analyse repertoire data from Textor et al. (2023), on TRB CDR3 sequences from $CD4^+$ T-cells of TdT knock-out (TdTKO) mice (n=13) and wild-type (WT) mice (n=10). We extend the algorithm developed by De Greef & De Boer (2021) to allow us to infer the length of the D segment from the TRB sequence. We confirm that TRB sequences without a D segment are relatively common in sequences with a high incidence but reject the hypothesis because we fail to observe a protective effect of TdT.

Results

Previous work found that ~10% of the abundant TRB sequences have no D segment (De Greef & De Boer, 2021). These sequences were primarily found in neonate samples. The combination of these sequences being widespread and more so present in samples of early development, led them to hypothesize that the downregulation of TdT in early development is the cause of these shorter sequences. TdT can insert nucleotides in between the gene segments in the CDR3 sequence, while exonucleases can delete nucleotides. A downregulation of TdT could therefore cause shorter TRB sequences which are lacking a D segment due to deletions. Therefore, we searched for differences between TRB CDR3 sequences of TdTKO mice and WT mice.

TdTKO TRB sequences are significantly less likely to have N-additions than those of WT mice and reflect TRB sequences of embryonic mice.

We investigated the difference in percentages of sequences with insertions between the two groups to validate the data. The level of confidence was represented by the number of reads for any given sequence. Thresholds were set for 1, 2, 3, 5, 10, 15, 20, 25 and 50 reads, and for each of these thresholds, all unique sequences with more reads than the threshold were considered. Using the method described in De Greef & De Boer (2021) we found that as much as 45% of the sequences with more than five reads still had insertions in TdTKO mice (Figure s1). This led us to suspect that the method was overestimating the number of N-additions by assuming all nucleotides that did not match exactly with the V, inferred D, or J segment sequences were N-additions. We therefore extended the algorithm of De Greef & De Boer by taking P-additions into account and used that from here on. Figure 3a shows that TdTKO sequences are far less likely to have insertions than WT sequences. Interestingly, the percentages of sequences with more than 20 and 50 reads that still had insertions were 16% and 11%, which was more than we had expected. We then analysed nucleotide sequences generated using OLGA (Sethna et al., 2019), of which we knew the exact V, D, and J sequences used in the CDR3 sequence, to validate our method. We found that almost all sequences without N-additions also had no insertions when analysed with the extended method, leading us to conclude that the insertions found in the data are most likely sequencing errors and or PCR artifacts.

To test whether these TdTKO sequences are representative of sequences from mice in early development, we also compared our findings to those found in another study, comparing TRB sequences from mice of different ages (Sethna et al., 2017). Figure 3b shows the distribution of the number of N-additions in unique sequences for TdTKO and WT, and is directly comparable to the results found by Sethna et al (2017), comparing the number of N-additions inferred from out-of-frame thymic TRB sequences across a set of ages. The similarities are striking, both the D42 post birth and WT group have ~10% of sequences without insertions, peak at 3 nt with 20-25%, and reduce back to ~10% at 5 nt. Moreover, the embryonic E17 and TdTKO groups also match almost exactly, having no insertion in 90% and 75% of all unique sequences respectively. Together, this indicates that sequences from TdTKO mice do in fact have significant reductions in TdT activity, and that the TdTKO sequences accurately reflect TdT activity of embryonic sequences.

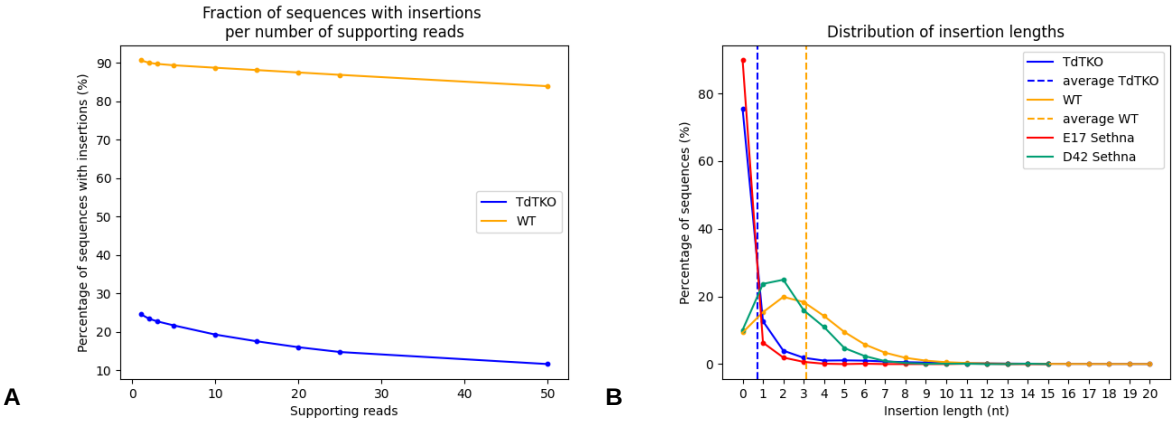

Figure 3: (**A**) Percentage of unique TRB CDR3 sequences containing insertions, or N-additions, per number of supporting reads. It shows that sequences in TdTKO mice have insertions significantly less frequently than sequences of WT mice. (**B**) Distribution of insertion lengths, that is the number of nucleotides in between the V and J segments that did not match the D segment and could not be identified as P-additions. The data for the E17 and D42 graphs were inferred from a figure in another study. For these graphs the y-axis represents the probability of the insertion length. TdTKO average: 0.72, WT average: 3.14.

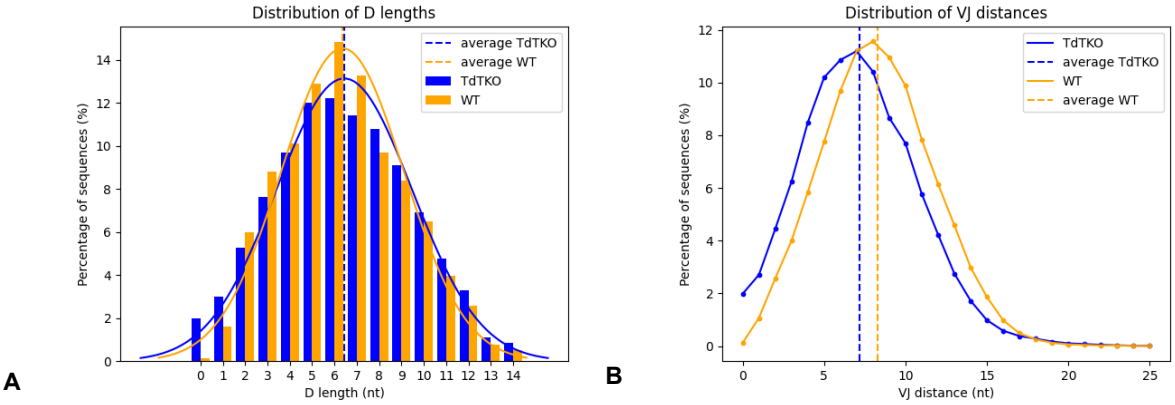

Figure 4: (**A**) Distribution of the inferred D lengths. It shows the percentages of the TRB CDR3 sequences with a particular inferred length of their D segment. The bars of both TdTKO and WT each sum up to 100%. In the sequences with no D segment there is a clear overrepresentation of the TdTKO sequences. However, only 2% of the TdTKO sequences have an inferred D length of 0 nt, while the majority has a D length of 5-8 nt. TdTKO average: 6.43, WT average: 6.36. (**B**) Distribution of VJ distances, that is the number of nucleotides in between the V and J gene segments that do not match the germline sequence of either segment. This includes a potential D segment as well as N-additions and or P-additions. TdTKO average: 7.16, WT average: 8.30.

The length of the D segment hardly differs between TdTKO mice and WT mice.

The analysis of the inferred length of the D segments in the unique TRB sequences revealed that despite sequences without a D segment being overrepresented in TdTKO mice, almost all sequences still had an inferred D segment. This was 98.02% and 99.88% for TdTKO and WT respectively. Moreover, the mean D length of TdTKO sequences was longer (6.57 nt) than that of the WT sequences (6.37 nt). The method used to infer the D segment in the

sequences allows for a single nucleotide to be identified as such. Therefore, we repeated this analysis requiring D segments of at least three nucleotides. Assuming that many TdTKO sequences would have a D segment of just a few nucleotides, we expected to see a significant decrease in the percentage of sequences with an inferred D segment. Surprisingly, we found that 89.75% of TdTKO sequences and 92.29% of WT sequences had an inferred D segment of three or more nucleotides, with a mean of 7.02 and 6.75 nt respectively. Figure 4a shows a distribution of the inferred D lengths for both TdTKO and WT mice, each with their respective normal distribution. The overrepresentation of sequences having a D segment length of 0 or 1 nt in TdTKO mice, made it seem as though the TdTKO distribution was slightly skewed. However, skewedness and kurtosis analyses revealed that neither distribution was significantly skewed and though the TdTKO distribution was slightly more platykurtic (-0.52 and -0.45 for TdTKO and WT respectively), this was insignificant.

Additionally, we compared the VJ distance between TdTKO and WT, that is the number of nucleotides between the last and first nucleotide of the V and J gene segments respectively, which includes the D segment and any N- or P-additions (Figure 4b). Interestingly, the distributions are very similar with TdTKO sequences being only 1 nt shorter on average. Considering that on average TdTKO has almost three nt fewer insertions (Figure 3b), this suggests that the D segment might in fact be a bit longer in TdTKO mice, although this is not revealed by the data. We also looked at differences in deletions of V and J segments in TdTKO and WT sequences as a proxy for D segment deletions (Figure s2). We found significant differences between the groups; WT mice had more deletions in the V segment, but TdTKO mice had more deletions in the J segment. Therefore, we could not conclude either group had more deletions than the other. All together, these results indicate that sequences from TdTKO mice do not have a significantly shorter D segment compared to sequences of WT mice.

Abundant TRB sequences frequently lack a D segment.

As we were unable to find compelling evidence that supported our hypothesis, we aimed to verify the findings of De Greef & De Boer (2021), in this dataset as well. To do so, we stratified the data for incidence (Figure 5). Here, incidence is the number of mice in which a sequence was present. We considered sequences identical when they had the same CDR3 nucleotide sequence and V and J gene segments. We define Low incidence as the group of sequences in one mouse, and High incidence as the group of sequences in the highest two incidence groups.

Figure 5a shows the percentages of sequences with no inferred D segment per incidence level for TdTKO and WT mice. We found a positive correlation between the incidence and the percentage of sequences without an inferred D segment. Moreover, 13% of the High incidence sequences in WT mice do not have an inferred D segment, which agrees with the findings of De Greef & De Boer (2021). Interestingly, only 7% of sequences present in all TdTKO mice had no inferred D segment. Figure 5b includes inferred D segments of 1 and 2 nt and has a similar correlation. Again, the results for WT mice are in accordance with the findings of De Greef & De Boer. However, one would expect that especially in sequences with a high incidence, the representation of the shorter D segment sequences would be more apparent in the TdTKO sequences if absence of TdT allowed for shorter D segments.

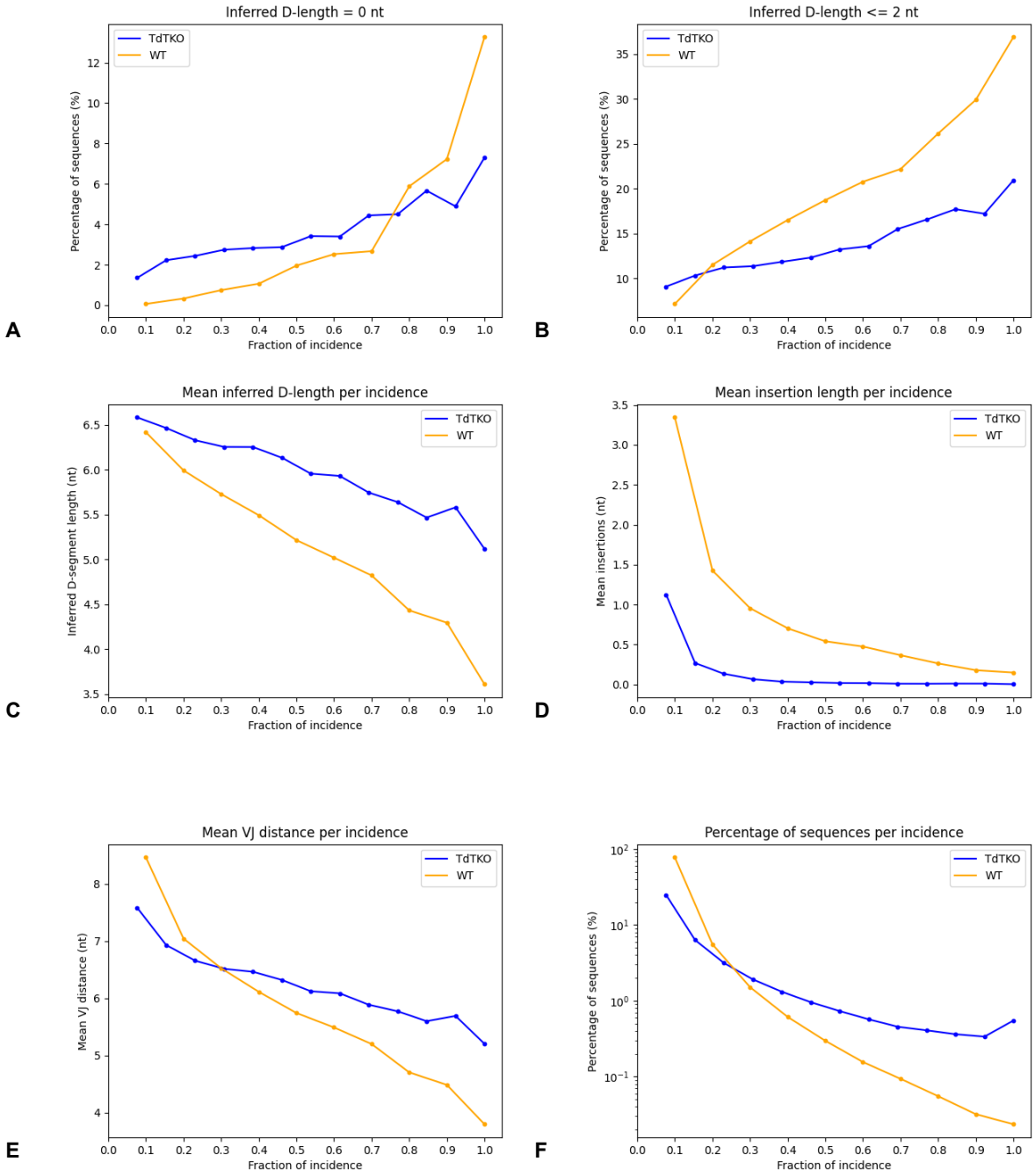

Figure 5: Fraction of incidence plots, which is the number of mice a given TRB CDR3 sequence is present in, in TdTKO and WT mice. The number of TdTKO and WT mice were n=13 and n=10 respectively. (**A**) The percentage of sequences per incidence that have no inferred D segment. 13% of WT sequences present in all mice have no inferred D segment which agrees with the findings of De Greef & De Boer (2021). (**B**) The percentage of sequences per incidence that have an inferred D length of less than or equal to 2 nucleotides. (**C**) The mean inferred D-length per incidence. Inferred D-length of WT sequences tend to be shorter overall, but especially at high incidence. (**D**) The mean insertion length per incidence. There is a strong negative correlation between insertion length and incidence. Because insertions are unlikely, sequences with many insertions are rare and will have a low incidence. (**E**) The mean VJ distance per incidence. At low incidence WT mice have longer VJ distances, likely due to the added insertions. At high incidence, the inferred D segment of WT tends to be shorter than that of TdTKO, influencing this plot. (**F**) Percentage of the number of sequences per incidence. The vertical axis is log-scaled, the total number of unique sequences for TdTKO and WT are 247100 and 1380014 respectively. Specifics for the number of sequences per incidence can be found in Table s1.

To add to this, Figure 5c shows the mean inferred D length as a function of the incidence. As expected, we see a negative correlation between the incidence and mean inferred D length, yet the mean D length in TdTKO sequences remains higher than that of WT sequences across all levels of incidence. Figure 5d displays the mean insertion length as a function of the incidence. Here, we observe that TdTKO sequences that are present in more than one mouse generally have no insertions, and WT sequences tend to have fewer insertions the higher the incidence. Figure 5e shows the mean VJ distance per incidence and can be interpreted as the combined effect of the data shown in figures 5c and 5d. At Low incidence, the insertions by TdT cause the VJ distance of WT sequences to be longer compared to TdTKO sequences. As the incidence gets higher, the mean VJ distance seems to be dominated by the mean D length, following the trend in Figure 5c. Finally, Figure 5f shows the percentage of sequences per incidence on a log scale. Most of the sequences can be found at Low incidence, and as incidence increases, we see a steady decrease in the number of sequences with only 889 sequences left in WT mice at High incidence. Specifics can be found in Table s1.

Though we were able to validate the findings of De Greef & De Boer in the WT sequences, the results for the TdTKO sequences were unexpected, with exception of Figure 5d. We had expected to see that the length of the D segment in TdTKO mice would be comparable to that of WT mice, if not shorter. Instead, these figures suggest that the WT sequences are shorter than the TdTKO sequences at High incidence. We suspect that the generation probability of TdTKO sequences might be disproportionately higher than that of WT sequences, which could cause sequences to have a higher incidence, however this is difficult to compare as OLGA has no option for calculating generation probabilities in the absence of TdT. Therefore, we are unable to accurately estimate generation probabilities for TdTKO mice. All together, these results provide further evidence that the D segment in TdTKO sequences is not significantly shorter than those in WT sequences, and we should therefore consider discarding the hypothesis.

Generation probabilities of WT mice are higher in High incidence as well as D segment lacking sequences.

We suspected that generation probabilities might influence the incidence, therefore we calculated the generation probabilities for WT sequences using OLGA (Sethna et al., 2019). As mentioned above, we were unable to accurately estimate TdTKO generation probabilities using OLGA. However, Sethna et al. (2017) estimated generation probabilities of sequences from embryonic E17 mice, and found these generation probabilities to be higher than those of D42 post birth mice. Because TdT activity from the embryonic sequences reflects that of TdTKO mice, this data can be used as a proxy. This supports the idea that TdTKO sequences have a higher generation probability than WT sequences and that this could have increased their incidence.

Figure 6 shows the generation probabilities of WT sequences per incidence and per length of inferred D segment. In Figure 6a we see a moderate positive correlation between incidence and generation probability. A spearman rank correlation test showed this to be 0.318. A positive correlation was expected, as sequences with a high generation probability would be more likely to be made by multiple mice. Figure 6b shows that there is a substantially higher generation probability for sequences without an inferred D segment. This was quite unexpected and may be an inaccurate representation of the true generation probabilities by OLGA for sequences without D segment (Thierry Mora, personal communication).

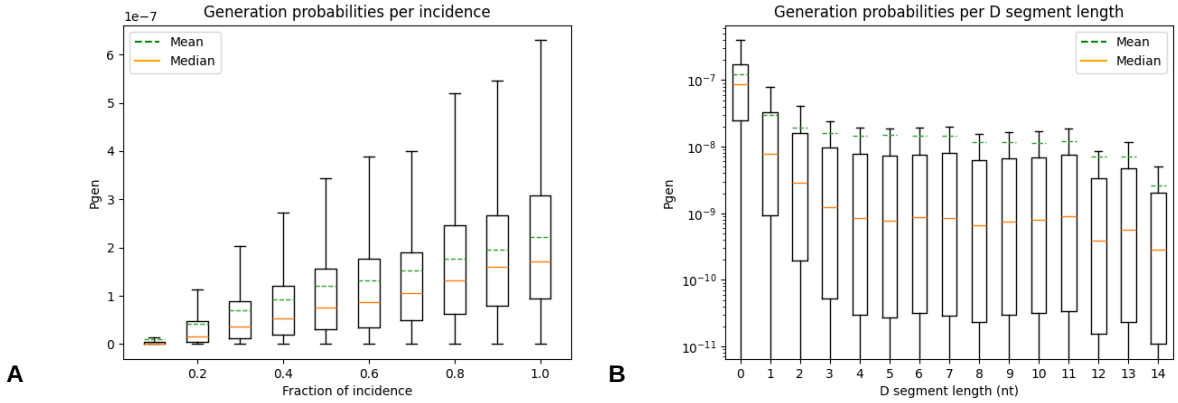

Figure 6: Boxplots showing the generation probabilities of WT sequences as a function of (**A**) incidence and (**B**) D segment length. We observe a moderate correlation between generation probability and incidence. Sequences lacking a D segment appear to have a substantially higher generation probability. The vertical axis in Figure **B** is log-scaled. The number of WT sequences with a specific inferred D segment can be found in Table s2.

Analysis of VJ segment usage reveals that TRBV1 and TRBV16 make up almost 60% of V usage in High incidence sequences of WT mice.

Finally, we investigated the V and J gene segment usage of TdTKO and WT sequences to further identify any possible underlying patterns in the data. Figure 7 shows a comparison in V and J gene segment usage between TdTKO and WT mice, as well as V and J usage of High and Low incidence sequences in WT mice. Because RTCR can only identify functional VJ gene segments, only those segments are shown in this analysis.

Usage of gene segments for V (Figure 7a) and J (Figure 7b) was calculated per mouse, indicated by the blue (TdTKO) and orange (WT) dots, as well as the corresponding 95% confidence intervals and means. Notably, the total usage of the gene segments is significantly different between TdTKO and WT mice for over half of both V and J segments. This suggests that TdT influences the segment usage during V(D)J-recombination. Furthermore, we looked at differences in VJ usage between High and Low incidence sequences in WT mice. We found a staggering difference in usage for TRBV1 and TRBV16, making up almost 60% of V segment usage in High incidence sequences combined (Figure 7c). Further analysis of these two segments indicates that segment usage is positively correlated with incidence (Figure s3). For the J segment usage, the biggest differences were in TRBJ1-1 and TRBJ2-1 (Figure 7d). Here, the usage of TRBJ1-1 was almost doubled in High incidence sequences, while TRBJ2-1 was halved, compared to Low incidence sequences. We also compared V and J gene segment usage of WT sequences with and without inferred D segment (Figure s4). The J gene segment usage suggested that the first cluster of J gene segments was used more often in sequences without an inferred D segment. Taken together, these results indicate that gene segment usage changes at some point during development.

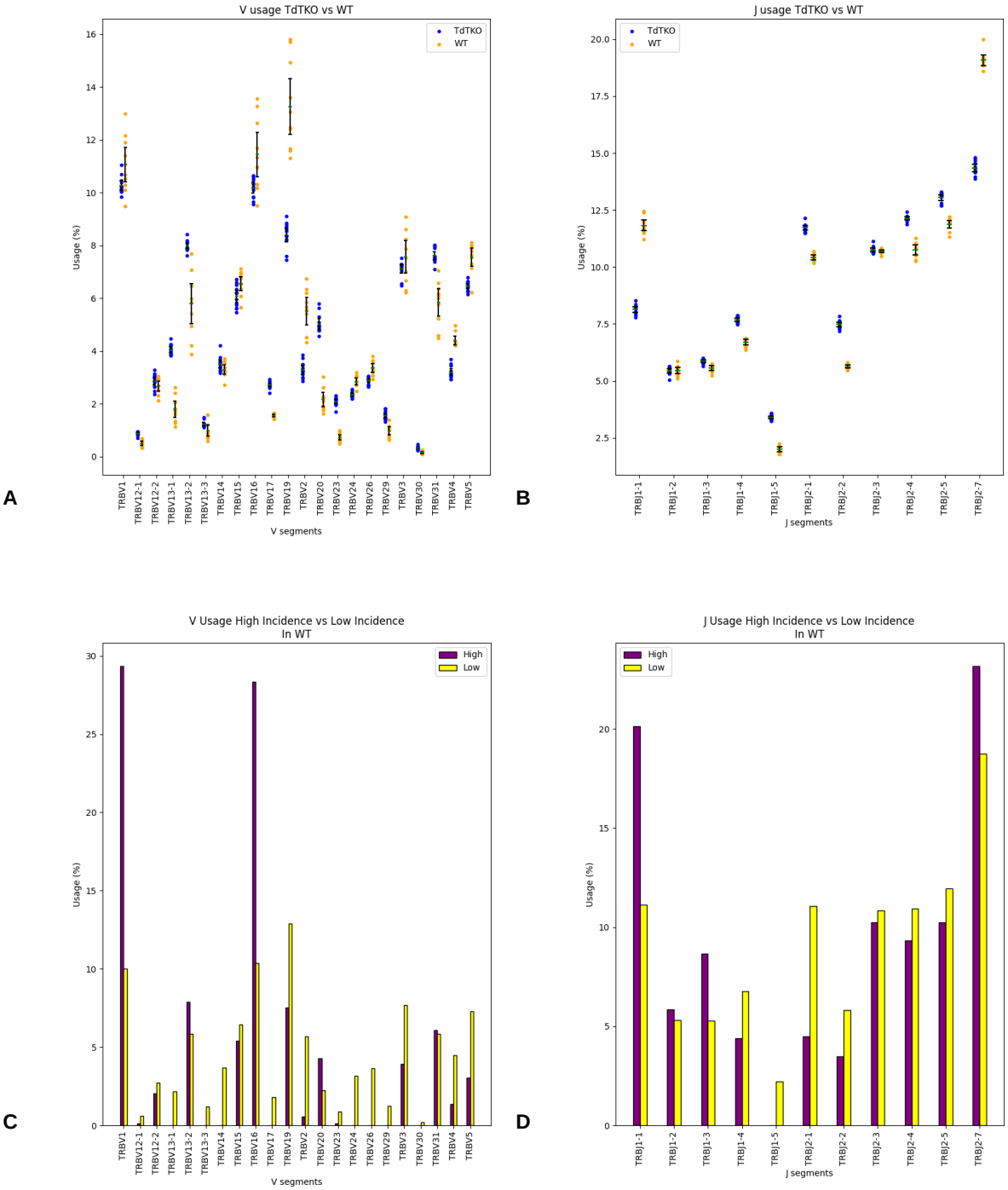

Figure 7: Usage of functional (**A**) V and (**B**) J gene segments in TdTKO and WT sequences. Each of the dots represents one mouse, 95% confidence ranges are shown in black with the green lines representing the mean usage of that segment of all mice in their respective group. There are clear differences in the usage of some segments between the two groups. Only functional V and J gene segments are shown in these plots, because the segments were identified using RTCR, which only considers the functional VDJ gene segments. (**C**) Usage of V segments in High and Low incidence WT sequences. While Low incidence V usage is spread out, High incidence V usage is dominated by TRBV1 and TRBV16. (**D**) Usage of J segments in High and Low incidence WT sequences. Low incidence J usage is comparable to the general WT J usage as it makes up the majority of all WT sequences. High incidence J usage is substantially higher in TRBJ1-1 compared to Low incidence.

Discussion

We analysed TRB sequencing data from $CD4^+$ T-cell repertoires of TdTKO and WT mice to test the hypothesis that the absence of TdT may lead to TRB sequences that lack the D segment due to deletions. We found that TdTKO sequences accurately reflect TdT activity of embryonic sequences but could not find any evidence that supports the hypothesis, and therefore we discard the hypothesis. We then confirmed that 13% of high incidence sequences found in all WT mice lack a D segment. These observations suggest that there is a different process than TdT mediated N-additions and deletions by exonucleases that is causing these abundant and short sequences that lack the D segment.

The incidence data showed that, on average, high incidence WT sequences lack a D segment more frequently than TdTKO sequences of the same incidence, and that their average D segment length is also shorter than that of TdTKO sequences. This may be explained by higher generation probabilities of TdTKO sequences. We could not directly estimate generation probabilities of TdTKO sequences because OLGA has no option for calculating generation probabilities for sequences without N-additions. However, Sethna et al. (2017) estimated generation probabilities for embryonic sequences, whose TdT activity reflects that of TdTKO sequences allowing us to use them as a proxy. Sethna et al. found a significant difference in repertoire diversity that was almost entirely due to the change in TdT activity, with embryonic sequences having a higher generation probability than post birth sequences. The difference in generation probability could have caused TdTKO sequences, which would otherwise have a lower generation probability due to insertions, to now be more frequent. Moreover, we show that generation probability and incidence is correlated in WT sequences. Altogether, this could cause the TdTKO data to be shifted in the plots, hence why it seems as though TdTKO sequences have longer D segments and lack a D segment less often. The high estimated generation probabilities of sequences without inferred D segment may also be inaccurate. Generation probabilities estimated by OLGA are most heavily influenced by TdT mediated N-additions, with less insertions being more likely. Due to the method used to infer D segment usage, sequences without an inferred D segment by definition have no insertions. This may have caused OLGA to estimate a disproportionately high generation probability for sequences without an inferred D segment, as sequences without TdT additions are relatively likely.

The analysis of the V and J gene segments used in the sequences showed that the usage of segments differed significantly between TdTKO and WT sequences. We can only speculate on why this is, but this could indicate that N-additions by TdT influence the tolerance induction during thymic selection of sequences having certain gene segments. The selection for the least self-reactive sequences would then explain the difference in VJ segment usage, as different V and J segments would be more self-reactive in sequences with N-additions than in sequences without N-additions. The number of deleted nucleotides in V and J segments shown in Figure s2 could also play a role in this, as we also observe significant differences in them between TdTKO and WT sequences.

We did not find evidence suggesting that the sequences that lack a D segment do so due to deletions. An alternative hypothesis is that in those sequences the D segment was skipped. Figure s4 shows the V and J segment usage for sequences that lack a D segment compared to those that do not. The J segment usage seems to suggest that sequences that lack a D segment are more often recombined with a J segment from the TRBJ1 cluster. We speculate that sequences with a J segment from the TRBJ1 cluster are more likely to skip the D segment.

What exactly is causing the sequence characteristics we observe at high incidence remains unclear. If these sequences are indeed created in early development, we should see more of these characteristics in embryonic sequences, but so far consistent evidence is still lacking. We saw an increase in the number of High incidence sequences in the TdTKO group that was likely due to the higher generation probability of having no N-additions and they did not show the same sequence characteristics as the High incidence sequences from WT mice. However, given the correlation of generation probability and incidence in WT sequences, we can assume that the High incidence sequences usually have a high generation probability. The simplest explanation would be that this high generation probability is the sole reason for the High incidence sequences being so widespread, and that any correlations we observe are due to coincidence. On the other hand, a more appealing hypothesis would be that the high incidence sequences are short and give rise to potentially more cross-reactive TCRs, with the short length making them have a higher generation probability. We can only speculate as to when and why these sequences are made, but further research with a sizable dataset of embryonic and wild-type TRB sequences from multiple mice could potentially provide an answer as to whether the High incidence sequences we observe are indeed produced in early development or not.

Materials & methods

TdTKO and WT sequence data

We acquired TRB CDR3 sequence data from Textor et al. (2023) of $CD4^+$ T-cells from TdT knock-out (n=13) and wild-type (n=10) C57BL/6A mice. These datasets were pre-processed and contained for each sequence an identifier for the mouse, the phenotype ($CD4^+$ or other), the read count, the Phred score, the nucleotide and amino acid sequences, and the V and J gene segments that were used in the sequence together with their stop and start positions in the sequence respectively. The V and J segments were identified using RTCR with default settings. The data files can be found in the GEO database under the accession code GSE221703. Specific sample identifiers can be found in Table 1.

Table 1: Sample identifiers for the processed data files that were used.

Phenotype	Sample identifiers
TdTKO	GSM6893351
	GSM6893356
	GSM6893359
	GSM6893360
	GSM6893362
WT	GSM6893365
	GSM6893366
	GSM6893367
	GSM6893369
	GSM6893370

Inferring the D segment of CDR3 nucleotide sequences

The processing of these data files was done using in-house python scripts that will be available on GitHub (https://github.com/GabevandenHoeven/TdTKO_mice/tree/master). We filtered the data for $CD4^+$ sequences that had more than 1 read and were shorter than 64 nt. We extended the method by De Greef & De Boer (2021) for inferring the D segment length. For each sequence, the sequence was matched to the reference CDR3 sequence of the V and J segments identified by RTCR until there was no exact match. The remaining sequence was subsequently searched for the longest exact match to either of the D gene segments. In the extended method every possibility of D segment length was tested where the sequences left and right of the potential D segment were searched for P-nucleotides. The sum of the length of the potential D segment and the maximum number of P-nucleotides for that match is then taken as a score, of which the highest is considered to be the best match. In cases where there were multiple matches with the highest score, the match with the longest D segment was chosen as the best match. Any nucleotides that could not be identified as a part of the D segment or P-nucleotides were considered to be N-additions by TdT. This method minimises the number of insertions and prefers longer D segments as to cause minimal bias.

Statistical calculations, plotting of results and generated sequences

All processing of the sequence data for calculations was done using in-house Python scripts. The SciPy and NumPy packages were used in some statistical calculations and Matplotlib pyplot was used to plot the figures. Generation probabilities were calculated using OLGA (Sethna et al., 2019). In silico sequences, that were generated to validate the extended method of inferring D segment length, were also made using OLGA. Generated sequences without N-additions required an adaption in the source code of OLGA in the file “sequence_generation.py”. The edited version of this file, which included a copy of the function that creates the sequence but left out the N-additions, is also available on GitHub (https://github.com/GabevandenHoeven/TdTKO_mice/tree/master).

References

Bassing, C. H., Alt, F. W., Hughes, M. M., D'Auteuil, M., Wehrly, T. D., Woodman, B. B., Gärtner, F., White, J. M., Davidson, L., & Sleckman, B. P. (2000). Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule. *Nature*, *405*(6786), 583–586. https://doi.org/10.1038/35014635

De Greef, P. C., & De Boer, R. J. (2021). TCRβ rearrangements without a D segment are common, abundant, and public. *Proceedings of the National Academy of Sciences*, *118*(39), e2104367118. https://doi.org/10.1073/pnas.2104367118

Feeney, A. J. (2009). Genetic and Epigenetic Control of V Gene Rearrangement Frequency. In P. Ferrier (Ed.), *V(D)J Recombination* (Vol. 650, pp. 73–81). Springer New York. https://doi.org/10.1007/978-1-4419-0296-2_6

Gopalakrishnan, S., Majumder, K., Predeus, A., Huang, Y., Koues, O. I., Verma-Gaur, J., Loguercio, S., Su, A. I., Feeney, A. J., Artyomov, M. N., & Oltz, E. M. (2013). Unifying model for molecular determinants of the preselection Vβ repertoire. *Proceedings of the National Academy of Sciences*, *110*(34). https://doi.org/10.1073/pnas.1304048110

Helmink, B. A., & Sleckman, B. P. (2012). The Response to and Repair of RAG-Mediated DNA Double-Strand Breaks. *Annual Review of Immunology*, *30*(1), 175–202. https://doi.org/10.1146/annurev-immunol-030409-101320

Krangel, M. S. (2003). Gene segment selection in V(D)J recombination: Accessibility and beyond. *Nature Immunology*, *4*(7), 624–630. https://doi.org/10.1038/ni0703-624

Krangel, M. S. (2015). Beyond Hypothesis: Direct Evidence That V(D)J Recombination Is Regulated by the Accessibility of Chromatin Substrates. *The Journal of Immunology*, *195*(11), 5103–5105. https://doi.org/10.4049/jimmunol.1502150

Ma, L., Yang, L., Bin Shi, He, X., Peng, A., Li, Y., Zhang, T., Sun, S., Ma, R., & Yao, X. (2016). Analyzing the CDR3 Repertoire with respect to TCR—Beta Chain V-D-J and V-J Rearrangements in Peripheral T Cells using HTS. *Scientific Reports*, *6*(1), 29544. https://doi.org/10.1038/srep29544

Ma, Y., Schwarz, K., & Lieber, M. R. (2005). The Artemis:DNA-PKcs endonuclease cleaves DNA loops, flaps, and gaps. *DNA Repair*, *4*(7), 845–851. https://doi.org/10.1016/j.dnarep.2005.04.013

Mora, T., & Walczak, A. (2016). *Quantifying lymphocyte receptor diversity*. https://doi.org/10.48550/ARXIV.1604.00487

Murugan, A., Mora, T., Walczak, A. M., & Callan, C. G. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, *109*(40), 16161–16166. https://doi.org/10.1073/pnas.1212755109

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., & Goronzy, J. J. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, *111*(36), 13139–13144. https://doi.org/10.1073/pnas.1409155111

Sethna, Z., Elhanati, Y., Callan, C. G., Walczak, A. M., & Mora, T. (2019). OLGA: Fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, *35*(17), 2974–2981. https://doi.org/10.1093/bioinformatics/btz035

Sethna, Z., Elhanati, Y., Dudgeon, C. R., Callan, C. G., Levine, A. J., Mora, T., & Walczak, A. M. (2017). Insights into immune system development and function from mouse T-cell repertoires. *Proceedings of the National Academy of Sciences*, *114*(9), 2253–2258. https://doi.org/10.1073/pnas.1700241114

Sleckman, B. P., Bassing, C. H., Hughes, M. M., Okada, A., D'Auteuil, M., Wehrly, T. D., Woodman, B. B., Davidson, L., Chen, J., & Alt, F. W. (2000). Mechanisms that direct ordered assembly of T cell receptor β locus V, D, and J gene segments. *Proceedings of the National Academy of Sciences*, *97*(14), 7975–7980. https://doi.org/10.1073/pnas.130190597

Srivastava, S. K., & Robins, H. S. (2012). Palindromic Nucleotide Analysis in Human T Cell Receptor Rearrangements. *PLoS ONE*, *7*(12), e52250. https://doi.org/10.1371/journal.pone.0052250

Textor, J., Buytenhuijs, F., Rogers, D., Gauthier, È. M., Sultan, S., Wortel, I. M. N., Kalies, K., Fähnrich, A., Pagel, R., Melichar, H. J., Westermann, J., & Mandl, J. N. (2023). Machine learning analysis of the T cell receptor repertoire identifies sequence features of self-reactivity. *Cell Systems*, *14*(12), 1059-1073.e5. https://doi.org/10.1016/j.cels.2023.11.004

Tillman, R. E., Wooley, A. L., Khor, B., Wehrly, T. D., Little, C. A., & Sleckman, B. P. (2003). Cutting Edge: Targeting of Vβ to Dβ Rearrangement by RSSs Can Be Mediated by the V(D)J Recombinase in the Absence of Additional Lymphoid-Specific Factors. *The Journal of Immunology*, *170*(1), 5–9. https://doi.org/10.4049/jimmunol.170.1.5

Supplementals

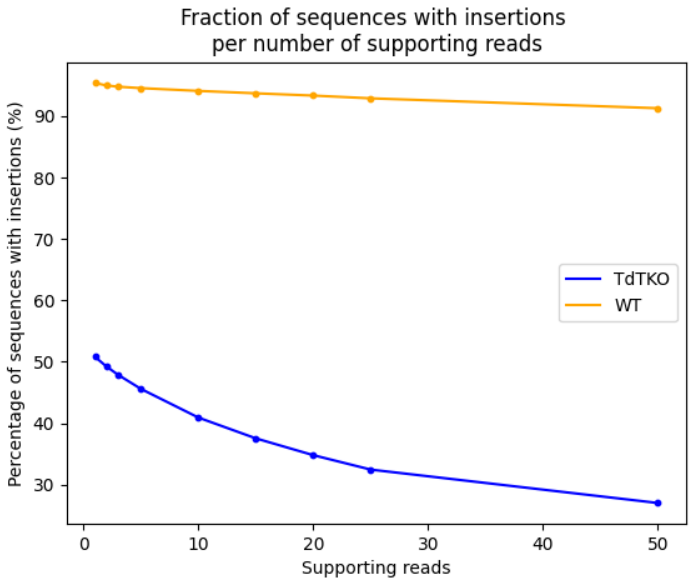

Figure s1: Percentage of unique TRB CDR3 sequences containing insertions, or N-additions, per threshold of number of supporting reads using the method of inferring D segment length described by De Greef & De Boer (2021). This method is overestimating the number of insertions.

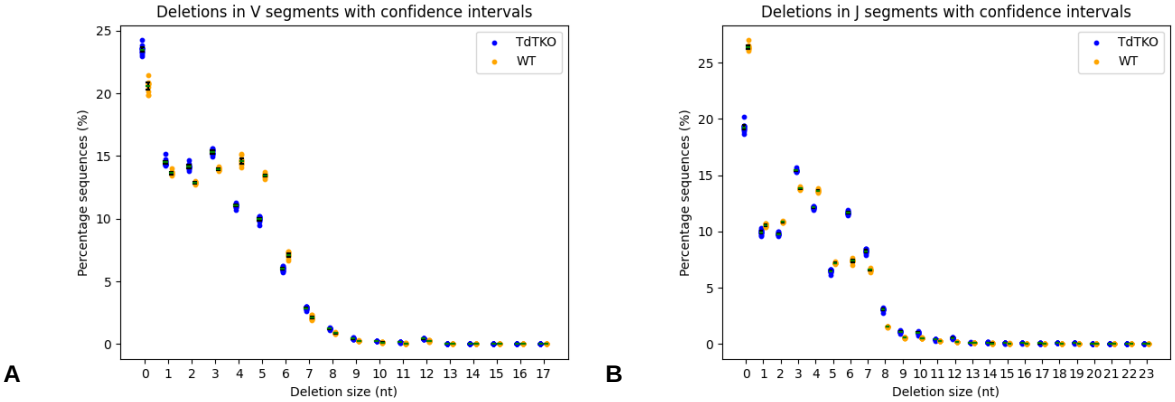

Figure s2: Percentage of sequences that have a specific number of deletions in the (**A**) V gene segment and (**B**) J gene segment for TdTKO and WT mice. There are significant differences between the two groups for <7 deleted nt in V gene segments, and <10 deleted nt in J gene segments, but from these plots it cannot be concluded that TdTKO sequences or WT sequences have more deletions than the other.

Table s1: Number of sequences per incidence for TdTKO and WT.

Number of mice	TdTKO sequences	WT sequences
1	146,484	1,246,356
2	37,303	89,089
3	18,547	24,230
4	11,275	9,814
5	7,751	4,758
6	5,642	2,496
7	4,306	1,498
8	3,358	884
9	2,681	512
10	2,401	377
11	2,136	NA
12	1,984	NA
13	3,232	NA
Total	247,100	1,380,014

Table s2: Number of WT sequences per inferred D segment length.

Length of inferred D segment (nt)	Number of sequences
0	1,648
1	22,317
2	82,498
3	121,565
4	139,589
5	177,987
6	204,443
7	183,219
8	133,874
9	115,422
10	89,358
11	54,543
12	35,746
13	10,798
14	7,007

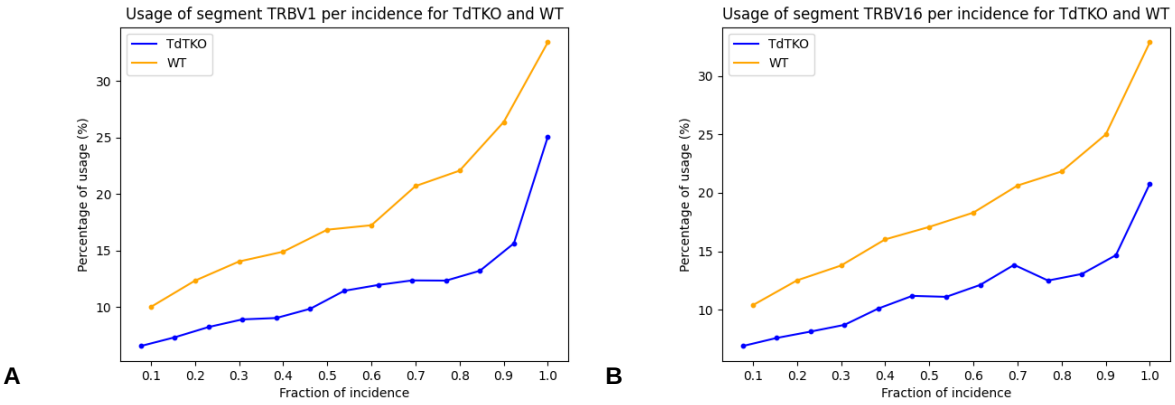

Figure s3: Percentage of gene segment usage per incidence for (**A**) TRBV1 and (**B**) TRBV16. We observe a positive correlation between segment usage and incidence for both TRBV1 and TRBV16.

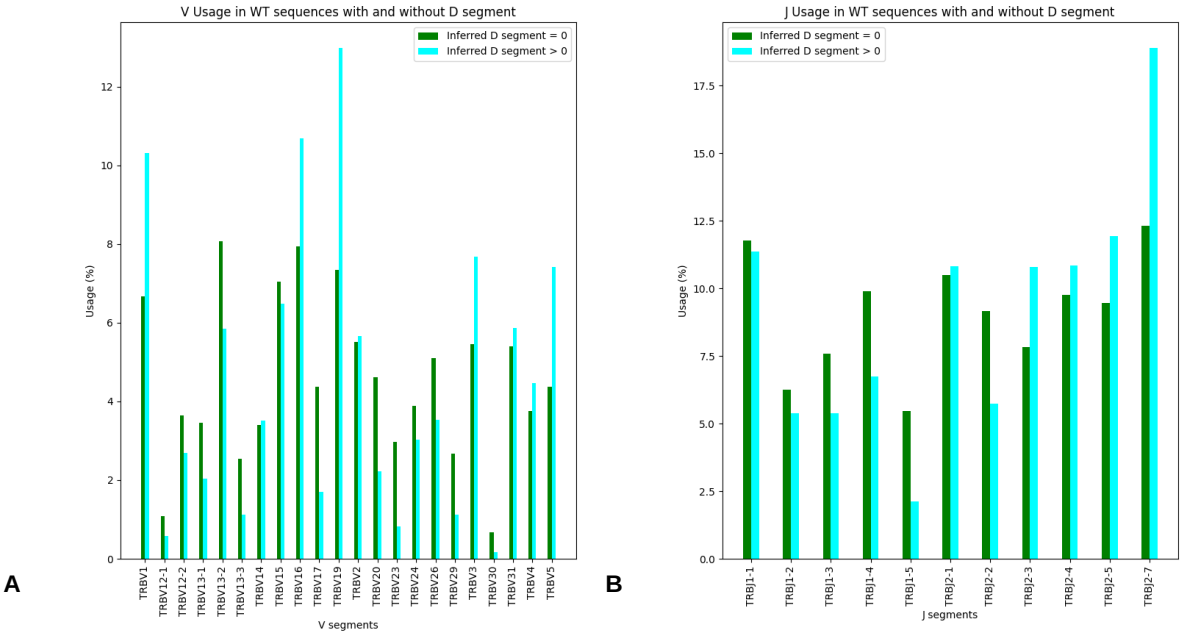

Figure s4: (**A**) V and (**B**) J gene segment usage in WT sequences with and without an inferred D segment. The number of unique sequences without an inferred D segment was 1,648. The number of unique sequences with an inferred D segment was 1,378,366. Figure **B** suggests that the first cluster of J segments (TRBJ1-1 to TRBJ1-5) is used more often in sequences without an inferred D segment.

Do T_{regs} have shorter TRB CDR3 sequences than naïve T-cells and what are their respective generation probabilities?

In this research project we tried to emulate TRB CDR3 sequences produced in early development using TdTKO mice and compared them to sequences from WT mice. We found that there was a difference in VJ segment usage between these two groups and speculated that the lack of n-additions by TdT might drive sequences towards a different segment usage due to the selection for non-self-reactivity. TCRs that have hydrophilic, or “sticky” amino acids are more likely to be self-reactive, because they bind more easily to a peptide. Data from prior research has indicated that TCRs with sticky amino acids are enriched in T_{reg} cells (Lagattuta et al., 2022), and a different study found that the TRA CDR3 amino acid sequences of TCRs with a fixed β-chain from CD4 T_{reg} cells had a higher generation probability than those of conventional CD4 T-cells (De Greef et al., 2024). We wondered if TRB CDR3 sequences from CD4 T_{reg} cells also have higher generation probabilities than conventional naïve CD4 T-cells, and if there are differences in CDR3 sequence length.

To address these questions, we compared TRB CDR3 sequence length and generation probabilities of CD4 T_{reg} and CD4 naïve T-cells from humans (Gomez-Tourino et al., 2018). The data was retrieved from the ImmuneACCESS database using the immunoSEQ Analyzer. For both the T_{reg} and naïve sequences, only the eight healthy control samples HD1-HD8 were used in this analysis. Only in-frame CDR3 sequences were used. The generation probabilities were calculated using OLGA (Sethna et al., 2019) using the nucleotide sequence and default settings with no masks. The python scripts that were used to process the files and analyse the data can be found on GitHub (https://github.com/GabevandenHoeven/TdTKO_mice/tree/master).

Figure S2-1 shows the amino acid and nucleotide generation probabilities as a frequency distribution and per incidence. In Figure S2-1A and S2-1B, the distributions of the log generation probabilities of T_{reg} and naïve cells are almost identical for both the amino acid and the nucleotide sequences. Naturally, the generation probabilities for amino acid are higher than those of the nucleotide sequence because of codon degeneracy, by having multiple codons for the same amino acid, different nucleotide sequences translate into the same amino acid sequence. Remarkably, the generation probabilities for amino acid sequences and nucleotide sequences differ substantially between T_{reg} and naïve T-cells when stratified for incidence (Figures S2-1C and S2-1D). Here, the amino acid generation probabilities of T_{reg} cells tend to be higher than those of the naïve cells, with the difference increasing with incidence. Conversely, the nucleotide generation probabilities of T_{regs} tend to be lower than those of naïve cells.

This result is quite puzzling. We can only speculate as for why this is, but an explanation could be that T_{regs} have more substitutions or insertions and deletions into codons that would produce the same amino acid sequences when compared to naïve T-cells. In this way, the nucleotide sequences of T_{regs} would get a lower generation probability while that of the corresponding amino acid sequence would stay the same. We also compared the nucleotide sequence CDR3 length of T_{reg} and naïve T-cells (Figure S2-2). We found that, in general, there is no difference in CDR3 length between T_{reg} and naïve T-cells.

This analysis raises many new questions. For example, is there a difference in CDR3 length when stratifying for incidence and which amino acids are preferentially used in the naïve and T_{reg} repertoire. These questions, fall outside the scope of this project and could be answered in a more in-depth analysis.

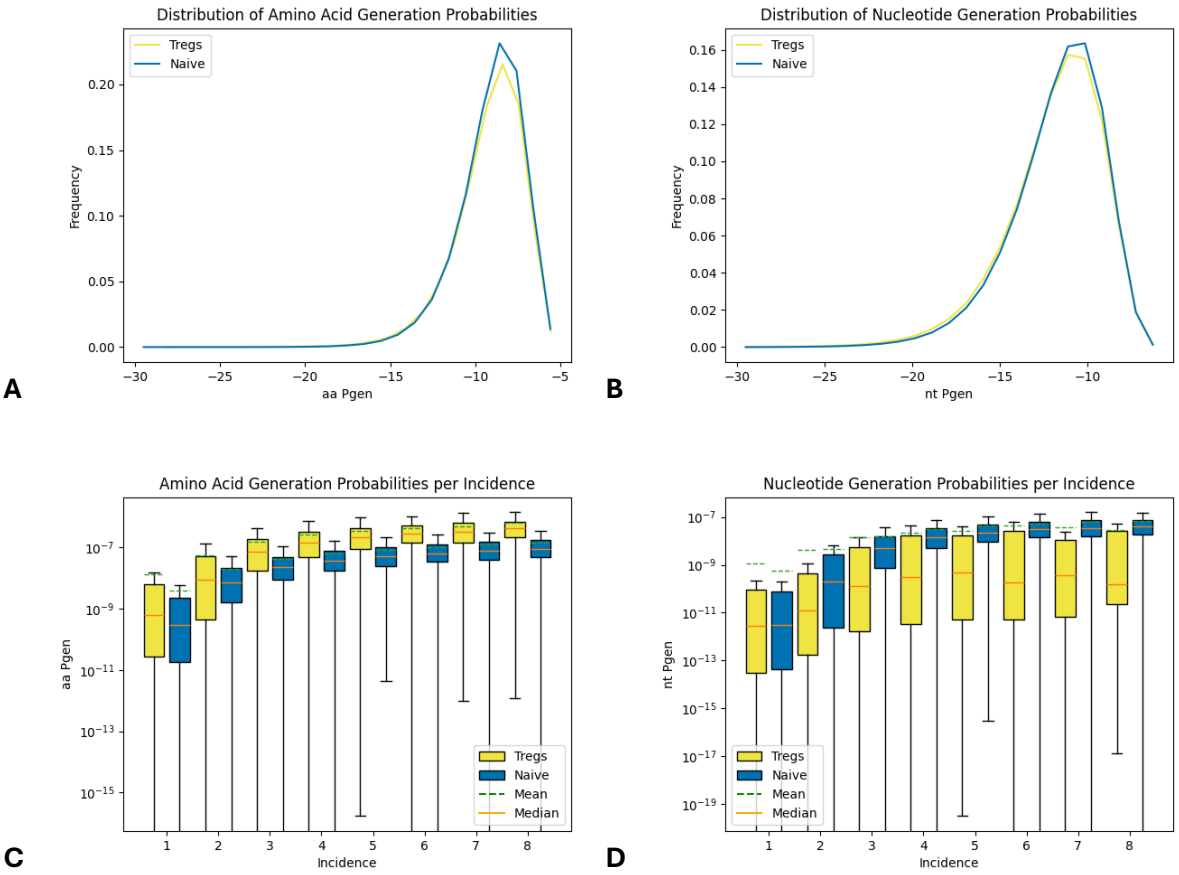

Figure S2-1: Generation probabilities of amino acid and nucleotide sequences from T_{reg} and conventional naïve $CD4^+$ T-cells as a frequency distribution and per incidence. Here, incidence is the number of samples a sequence was found in. We only use unique sequences, and the sequences were considered identical when their nucleotide sequence was the same. The frequencies in the generation probability distributions add up to 1 for both T_{reg} and naïve.

Table S2-1: Fraction of sequences at each level of incidence.

Incidence	1	2	3	4	5	6	7	8
Treg	0,897	0,078	0,013	0,005	0,002	0,001	0,0008	0,0005
Naïve	0,882	0,075	0,020	0,009	0,005	0,003	0,002	0,001

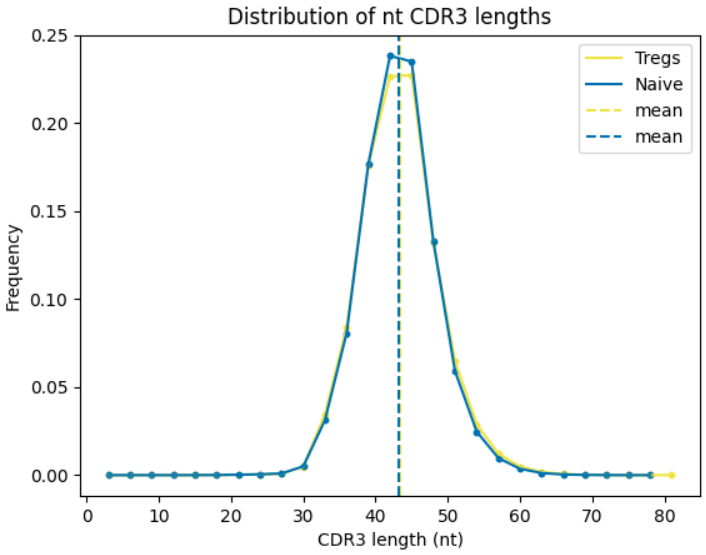

Figure S2-2: Distribution of TRB CDR3 nucleotide sequence lengths of CD4 T_{reg} and CD4 naïve T-cells. There are no significant differences.

SI references

De Greef, P. C., Njeru, S. N., Benz, C., Fillatreau, S., Malissen, B., Agenès, F., De Boer, R. J., & Kirberg, J. (2024). The TCR assigns naive T cells to a preferred lymph node. *Science Advances*, *10*(30). https://doi.org/10.1126/sciadv.adl0796

Gomez-Tourino, I., Kamra, Y., Lorenc, A., & Peakman, M. (2018). T-cell receptor β chains show abnormal shortening, repertoire diversity and sharing in type 1 diabetes [Dataset]. In *immuneACCESS*. https://doi.org/10.21417/b7c88s

Lagattuta, K. A., Kang, J. B., Nathan, A., Pauken, K. E., Jonsson, A. H., Rao, D. A., Sharpe, A. H., Ishigaki, K., & Raychaudhuri, S. (2022). Repertoire analyses reveal T cell antigen receptor sequence features that influence T cell fate. *Nature Immunology*, *23*(3), 446–457. https://doi.org/10.1038/s41590-022-01129-x

Sethna, Z., Elhanati, Y., Callan, C. G., Walczak, A. M., & Mora, T. (2019). OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics*, *35*(17), 2974–2981. https://doi.org/10.1093/bioinformatics/btz035