



### **Supplementary Information for**

A set point in the selection of the  $\alpha\beta$ TCR T cell repertoire imposed by pre-TCR signaling strength

Elena R. Bovolenta, Eva M. García-Cuesta, Lydia Horndler Julia Ponomarenko, Wolfgang W. Schamel, Mario Mellado, Mario Castro, David Abia, Hisse M. van Santen

Hisse M. van Santen  
Email: [hvansanten@cbm.csic.es](mailto:hvansanten@cbm.csic.es)

#### **This PDF file includes:**

Supplementary texts 1 and 2  
Figures S1 to S5  
Dataset S1 legends  
SI References

#### **Other supplementary materials for this manuscript include the following:**

Datasets underlying TCR $\beta$  repertoire analysis are deposited at the NLM/NCBI BioProject Sequence Read Archive and can be accessed via the link:  
<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA609042>

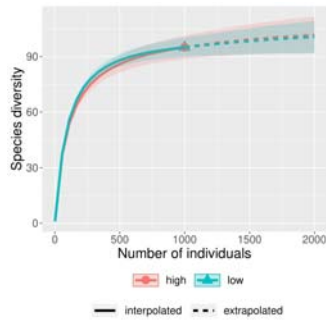
## Supplementary Information Text

### Supplementary Text 1: Comparing diversities and sample sizes

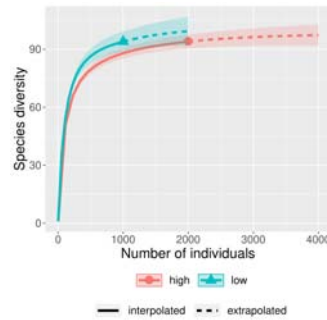
In Fig.ST1 we compare different *synthetic* samples in which we have the same underlying diversity ( $N_1 = N_2$ ) and different sample sizes ( $M_1, M_2$ ). Overall if the underlying diversity is the same (Fig.ST1A ) the curves overlap and if not ( $N_1 \neq N_2$ ), they are distinctly different regardless the sample size (Fig.ST1B).

**(A)** Same diversity provides same curves despite differences in the sample size

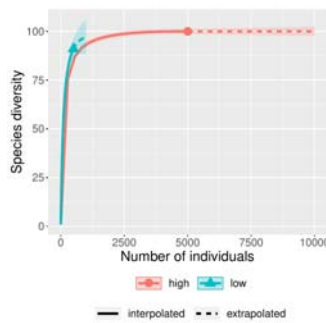
$N_1 = 100 \ N_2 = 100 \ M_1 = 1000 \ M_2 = 1000$



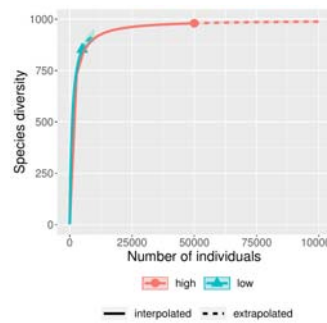
$N_1 = 100 \ N_2 = 100 \ M_1 = 1000 \ M_2 = 2000$



$N_1 = 100 \ N_2 = 100 \ M_1 = 500 \ M_2 = 5000$

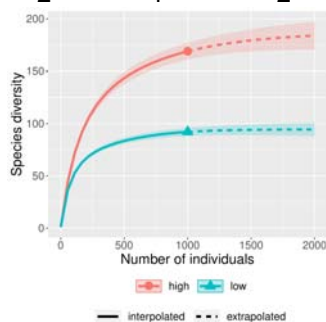


$N_1 = 1000 \ N_2 = 1000 \ M_1 = 5000 \ M_2 = 50000$

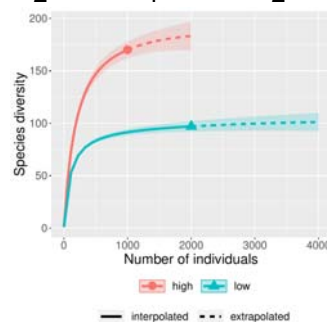


**(B)** Different diversities provide different curves despite the sample size

$N_1 = 100 \ N_2 = 200 \ M_1 = 1000 \ M_2 = 1000$

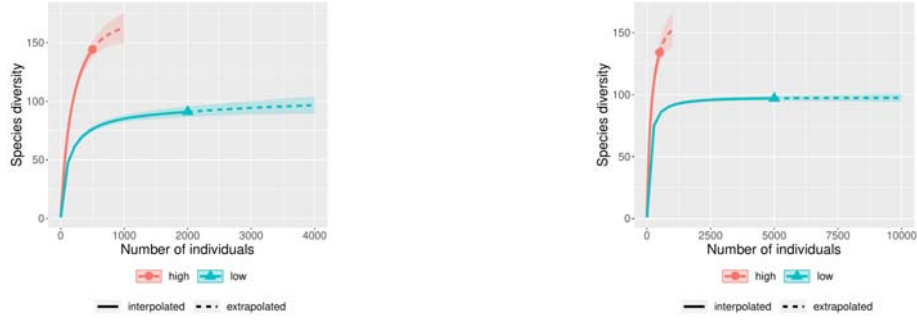


$N_1 = 100 \ N_2 = 200 \ M_1 = 2000 \ M_2 = 1000$



$N_1 = 100 \ N_2 = 200 \ M_1 = 2000 \ M_2 = 500$

$N_1 = 100 \ N_2 = 200 \ M_1 = 5000 \ M_2 = 500$



**Fig ST1. (A)** We compare different *synthetic* samples in which we have the same underlying diversity ( $N_1 = N_2$ ) and different samples sizes ( $M_1, M_2$ ). Overall, note how the diversity curves capture the diversity even in extreme cases where  $M_2$  is an order of magnitude larger than  $M_1$ . **(B)** Different diversities provide non-overlapping diversity curves. In both cases, the extrapolated diversity cannot be attributed to the sample size.

### Supplementary Text 2: A stochastic model of DN cell population

Consider that the number of DN cells are measured (sampled) at a certain time. Initially, there are  $n_0$  cells in the thymus and new cells migrate from the bone marrow at a rate  $v$ . There, they can proliferate at a rate  $\lambda$ , die at a rate  $\mu$  and transform into DP at a rate  $\delta$ . Using the theory of continuous-time branching processes (57), it can be shown that the generating function of the distribution is.

$$P(z, t) = \frac{(\lambda - \delta - \mu)^{v/\lambda} (-\delta - \mu - (z - 1)(\delta + \mu)e^{-t(\delta - \lambda + \mu)} + \lambda z)^{n_0}}{(-\delta - \mu - \lambda(z - 1)e^{-t(\delta - \lambda + \mu)} + \lambda z)^{n_0 + \frac{v}{\lambda}}}$$

From this distribution, the probability of having exactly  $N$  DP cells in the thymus would be

$$p_N = \frac{d^N P(z)}{dz^N}$$

Also, the mean or the variance can be computed using the *cumulant* generating function,  $K(\theta)$ , related to  $P(z)$

$$K(\theta) = \log P(e^\theta) \Rightarrow \begin{cases} \text{Mean: } m &= K'(0) \\ \text{Variance: } \sigma^2 &= K''(0) \end{cases}$$

In this case,

$$m(t) = \frac{e^{-t(\delta - \lambda + \mu)}}{\delta - \lambda + \mu} (n_0(\delta - \lambda + \mu) + v(-e^{t(\delta - \lambda + \mu)} 1))$$

$$\sigma^2(t) = \frac{e^{-2t(\delta - \lambda + \mu)} (-e^{t(\delta - \lambda + \mu)} 1)}{(\delta - \lambda + \mu)^2} (-\lambda v + n_0(\delta - \lambda + \mu)(\delta + \lambda + \mu) + v(\delta + \mu)e^{t(\delta - \lambda + \mu)})$$

Two important consequences can be extracted from these equations: (i) If the experiment is performed at different times might provide different mean values of the number of cells. Note that the mean and the variance have an exponential sensitivity with time; (ii) even if the experiment is performed at a *steady-state* where there is a balance between immigration, proliferation, death and selection into DP, the steady-state coefficient of variation, will be

$$CV \equiv \lim_{t \rightarrow \infty} \frac{\sigma(t)}{m(t)} = \frac{\sqrt{\frac{v(\delta + \mu)}{(\delta - \lambda + \mu)^2}}}{\frac{v}{\delta - \lambda + \mu}} = \sqrt{\frac{\delta + \mu}{v}},$$

so the magnitude of the (relative) fluctuations will be relevant ( $CV \gg 0$ ) unless there is no death and selection ( $\mu = \delta = 0$ ) or the immigration rate is infinitely large. In conclusion, even if the

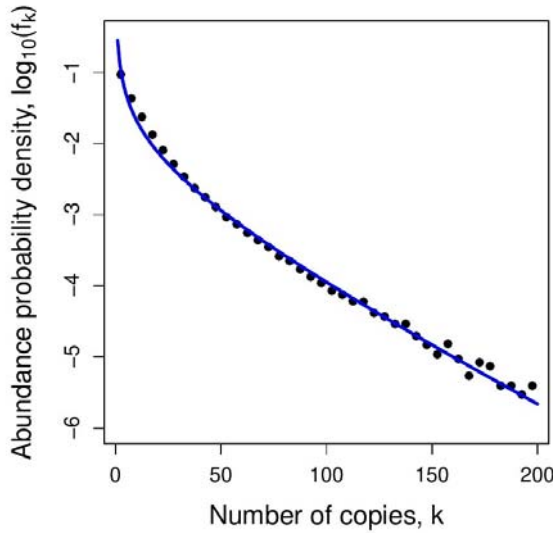
experiments are performed at exactly the same mice age, the intrinsic stochasticity of proliferation, death or migration, the total number of sampled cells is highly variable. In summary, one cannot simply attribute the measured diversity to the experimental sample size.

### Birth-death on top of the diversity distribution

In the previous section, we described the total population of cells, without any mention to the underlying diversity,  $D_0$ . Now, we consider different populations of the recently selected DP cells (namely, they are produced at a rate  $\delta$  from the previous stage). Figure ST2 shows the normalized empirical distribution of abundances for wild-type cells using the experimental data described in the main text, defined as

$$\text{Abundance probability density} \equiv f_k = \frac{N_k}{N}, \text{ with } N = \sum_{k=1}^{D_0} N_k,$$

where  $D_0$  is the 0-diversity (also known as *richness*).



**Fig. ST2.** Empirical probability density of abundances,  $P_k$  (circles). The blue solid line is a fit to Fisher's log-series distribution for  $f_k$  with parameter  $\pi = 0.967$ . For the sake of clarity, we have omitted in the logarithmic plot frequencies below  $2.5 \times 10^{-6}$ .

The blue line in Fig.ST2 shows a fit to Fisher's log-series distribution (58) given by,

$$f_k = \left( \sum_{k=1}^{D_0} \frac{\pi^k}{k} \right)^{-1} \frac{\pi^k}{k}, \quad k = 1, 2, \dots, D_0 \quad \pi \in (0,1).$$

The fit is not perfect, but we use this distribution (commonly observed in studies of diversity in Ecology) for the sake of the argument. For large diversities,

$$\sum_k \frac{\pi^k}{k} \simeq -\log(1 - \pi) \Rightarrow \text{so } f_k \simeq \frac{-1}{\log(1 - \pi)} \frac{\pi^k}{k}.$$

To estimate the role of death and proliferation of DP cells, we model each clone as an independent birth-death process, following

$$P(z, t) = \frac{(\lambda - \delta - \mu)^{\nu/\lambda} (-\delta - \mu - (z-1)(\delta + \mu)e^{-t(\delta-\lambda+\mu)} + \lambda z)^{n_0}}{(-\delta - \mu - \lambda(z-1)e^{-t(\delta-\lambda+\mu)} + \lambda z)^{n_0 + \frac{\nu}{\lambda}}}$$

where we have  $n_0 = 1$  for the  $N_1$  cells with abundance equal to unity,  $n_0 = 2$  for the  $N_2$  cells with two copies, ... We assume that the death and proliferation rates are identical for all the cells and, as we are describing *labeled* clones, there is not any immigration term. Note that the diversity is decreased by one unit if all the copies of a cell go extinct. In particular, for  $\delta = \nu = 0$ , we find that (the subindex  $d$  stand for DP)

$$p_0(t) = \left( \frac{\mu_d - \mu_d e^{-t(\lambda_d - \mu_d)}}{\lambda_d - \mu_d e^{-t(\lambda_d - \mu_d)}} \right)^{n_0} \Rightarrow \lim_{t \rightarrow \infty} p_0(t) = \begin{cases} 1 & \text{if } \mu_d \leq \lambda_d \\ \left( \frac{\mu_d}{\lambda_d} \right)^{n_0} & \text{if } \mu_d < \lambda_d \end{cases}$$

As we observe a huge population of cells, we can assume that the death rate,  $\mu_d$  is smaller than the proliferation rate (namely,  $\lim_{t \rightarrow \infty} p_0(t) \ll 1$ ).

Note that —as cells do not interact with each other— we have  $N_1$  different cells with just one clone (abundance= 1),  $N_2$  with two clones, and so on. Thus, we can define the mean decrease in the diversity as

$$\Delta D_0 = \text{Number of cells} \times (\text{Fraction of clones with just 1 cell} \times \text{Probability of 1 cell dies} + \text{Fraction of clones with two cells} \times \text{Probability of 2 cells die} + \dots)$$

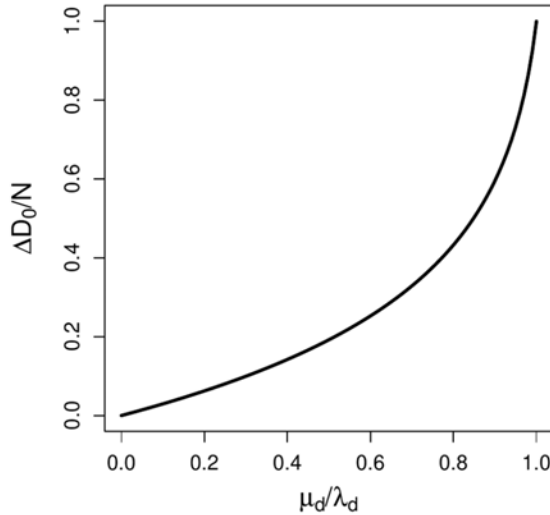
Mathematically,

$$\Delta D_0 = N \left( f_1 \left( \frac{\mu_d}{\lambda_d} \right)^1 + f_2 \left( \frac{\mu_d}{\lambda_d} \right)^2 + \dots \right)$$

Using Fisher's log series distribution defined above, we find that the mean diversity decrease would be

$$\Delta D_0 = - \frac{N}{\log(1 - \pi)} \sum_{k=1}^{D_0} \frac{\pi^k}{k} \left( \frac{\mu_d}{\lambda_d} \right)^k \simeq N \frac{\log(1 - \pi \mu_d / \lambda_d)}{\log(1 - \pi)}.$$

So we can conclude that, in the absence of death,  $\Delta D_0 = 0$ , as expected; namely, all the *species* are conserved. In general, Fig. ST3 shows the relative change in diversity,  $\Delta D_0 / N$  with the ratio  $\mu_d / \lambda_d$ , also known as the unit extinction probability. The fact that, empirically, the diversity drops for L19A mice, might be caused by an increased death rate and a reduced proliferation rate. So, the collected data do not allow to identify the dominant mechanism.

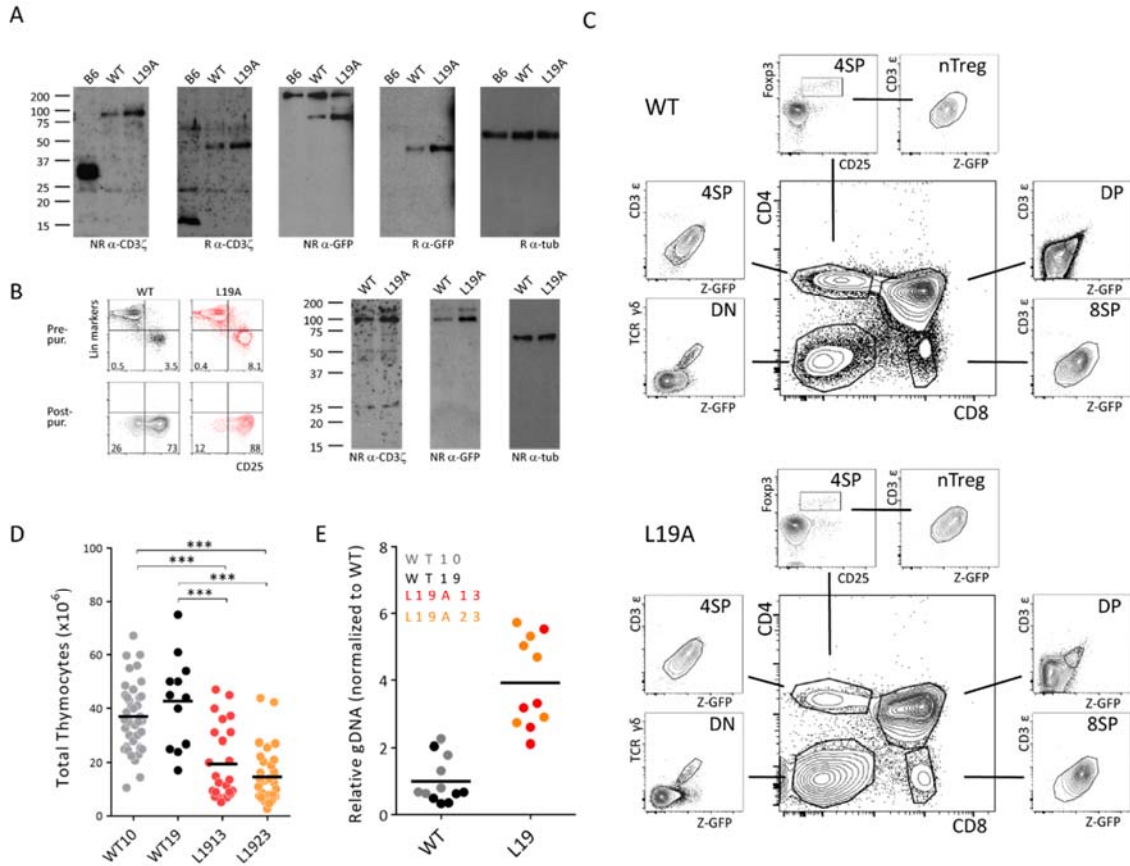


**Figure ST3.** Relative decrease in diversity,  $D_0$ , as a function of the extinction probability,  $\mu_d / \lambda_d$ . In the absence of death, the diversity does not change. When extinction is guaranteed ( $\mu_d > \lambda_d$ ), the diversity reduces to 0 in the long term. The fact that, empirically, the diversity drops for L19A mice, might be caused indistinctly by an increased death rate and a reduced proliferation rate.

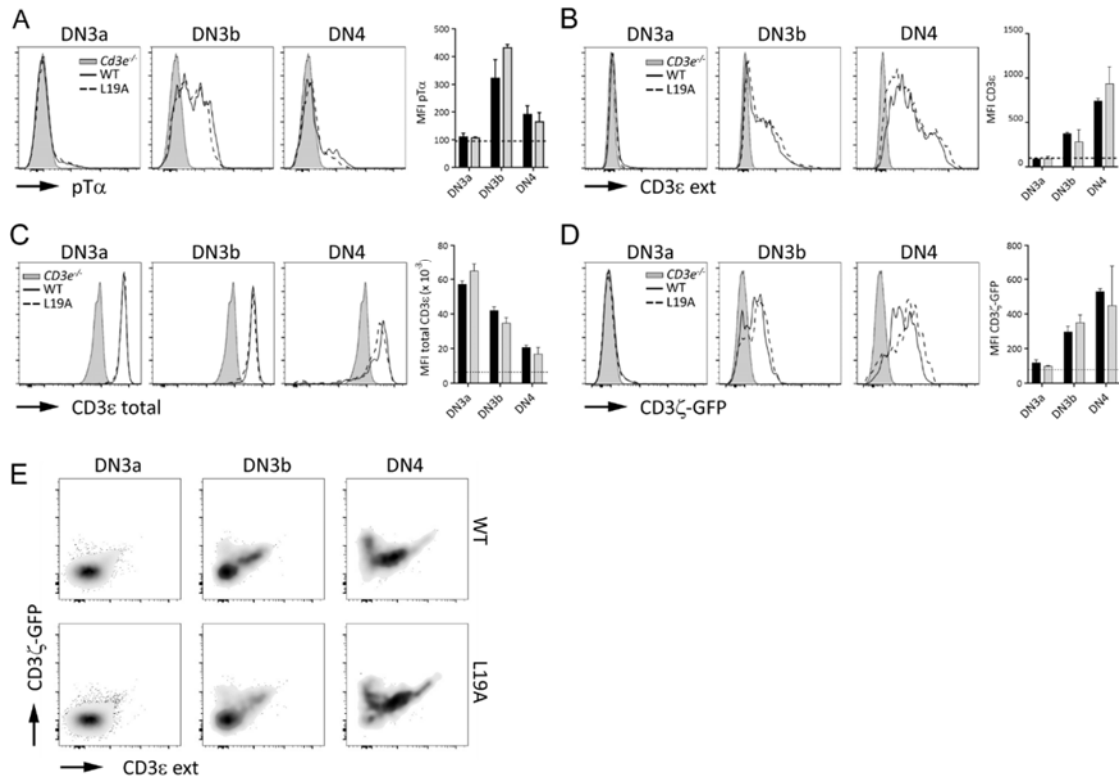
Finally, we can compute the *relative sensitivities*,  $\sigma_{\mu_d}, \sigma_{\lambda_d}$  of the diversity change with death and proliferation,

$$\begin{cases} \sigma_{\mu_d} \equiv \frac{\mu_d}{\Delta D_0} \frac{\partial \Delta D_0}{\partial \mu_d} \\ \sigma_{\lambda_d} \equiv \frac{\lambda_d}{\Delta D_0} \frac{\partial \Delta D_0}{\partial \lambda_d} \end{cases}$$

Note that, as  $\lambda_d$  and  $\mu_d$  enter as their ratio in  $\Delta D_0$ , so they have the same relative sensitivity but with opposite signs,  $\sigma_{\mu_d} = -\sigma_{\lambda_d}$ .

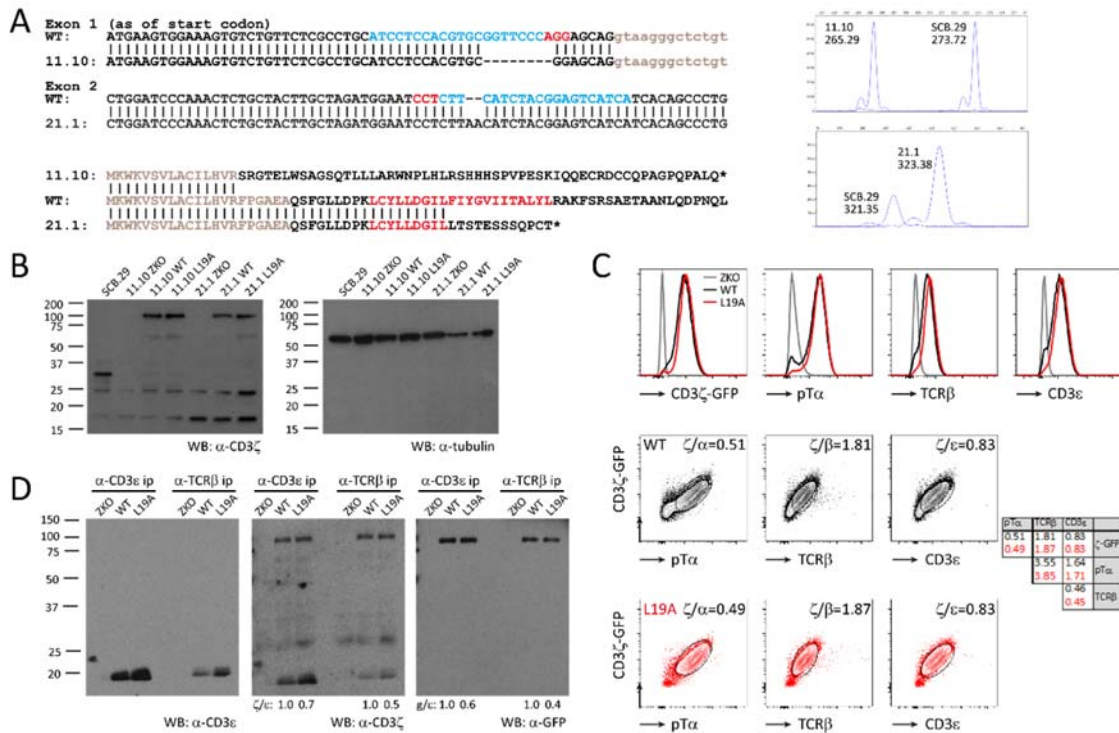


**Supplementary Figure 1. Expression of transgenic WT and L19A CD3 $\zeta$ -GFP chains under control of a human CD2 promoter/enhancer cassette.** (A and B) Immunoblot analysis of total lysates, prepared in 0.3% Brij96 detergent supplemented with a protease inhibitor cocktail (Roche 11697498001), from C57BL/6J, WT and L19A whole thymus (A) and WT and L19A DN thymocytes purified via negative selection with biotinylated CD4 $\alpha$ -, CD8 $\alpha$ -, CD19 $\alpha$ -, TCR $\gamma\delta$ -, NK1.1- and CD44-specific antibodies and biotin-binder Dynabeads (ThermoFisher Scientific 11047) (B). Purification of the DN subset was checked by incubating antibody-labeled pre- and post-purification samples with fluorescently-labeled streptavidin and anti-CD25 and analysis by flow cytometry. Lysates were separated in parallel via 10% reducing or non-reducing SDS-PAGE, dry-blotted to PVDF membranes. Membranes were incubated in parallel with antibodies against CD3 $\zeta$  (448 rabbit serum) or GFP (Roche 11814460001), followed by incubation with a tubulin-specific antibody (Sigma T5168). Bound antibodies were detected with HRP-coupled anti-rabbit or -mouse antibodies [ThermoFisher 31460 and 32430] and ECL (BioRad 1705061), followed by exposure to medical X-ray films (AGFA). (C) Expression pattern of the transgenes. Thymocyte suspensions obtained from 6 weeks old WT (upper panels) and L19A (lower panels) mice were stained with fluorescently-labeled antibodies specific for the indicated proteins and analyzed by flow cytometry. (D) Quantification of the total number of thymocytes in each mouse line. Each dot shows the total number of thymocytes in an individual mouse. Mean number of cells in each line is indicated by the black vertical bar. P-values were calculated using an unpaired two-tailed Student's t test with 95% CI (\*\*p<0.01, \*\*\*p<0.001). (E) Relative copy numbers of the transgenes. Genomic DNA isolated from tail snips from 5-6 mice of each transgenic lines was used as a template for a qPCR with oligonucleotides spanning the CD3 $\zeta$ -GFP junction: FW: CCT TTA CCA GGG TCT CAG CAC TG; RV: CCT CAG GGT CAG CTT GCC GTA G For normalization, a qPCR amplifying the genomic sequence of the Foxo1 gene was used: FW: CTT TGC CCC AGA TGC CTA T; RV: GGT TCA ATC CTC CGT AAC TTG A. Expression levels were normalized to the mean of the combined WT samples.



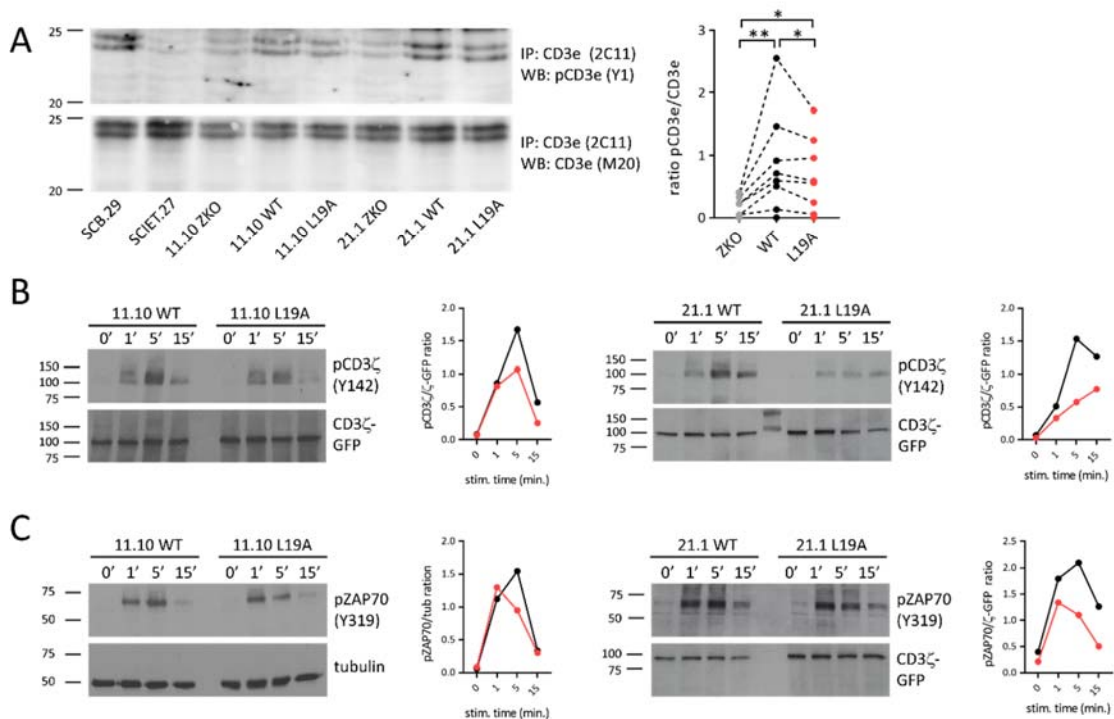
**Supplementary Figure 2. Expression of the pre-TCR at the surface of DN thymocytes.** Expression levels of the (A) pTα chain and (B) CD3εγ and CD3εδ heterodimers at the cell surface of DN3a (Lin-CD44-CD25<sup>hi</sup>), DN3b (Lin-CD44-CD25<sup>lo</sup>) and DN4 (Lin-CD44-CD25<sup>-</sup>) thymocytes from WT and L19A mice transgenic. (C) Total CD3ε expression, detected by staining permeabilized cells with the 2C11 mAb and (D) expression of CD3ζ-GFP fusion proteins (detected via the GFP moiety) on indicated populations defined as in (A) and (B). Background levels of staining were determined using Cd3e<sup>-/-</sup> thymocytes and the staining profile obtained with these cells was repeated in all corresponding panels as background reference. Graphs present the mean  $\pm$  SEM of the expression level calculated for 1 out of 4 experiments; dashed horizontal line indicates staining level obtained in Cd3e<sup>-/-</sup> thymocytes. (E) Correlation between cell surface labeling for the CD3ε chain and WT and L19A GFP-CD3ζ chains at the surface of DN3 and DN4 thymocytes. P-values were calculated using an unpaired two-tailed Student's t test with 95% CI.



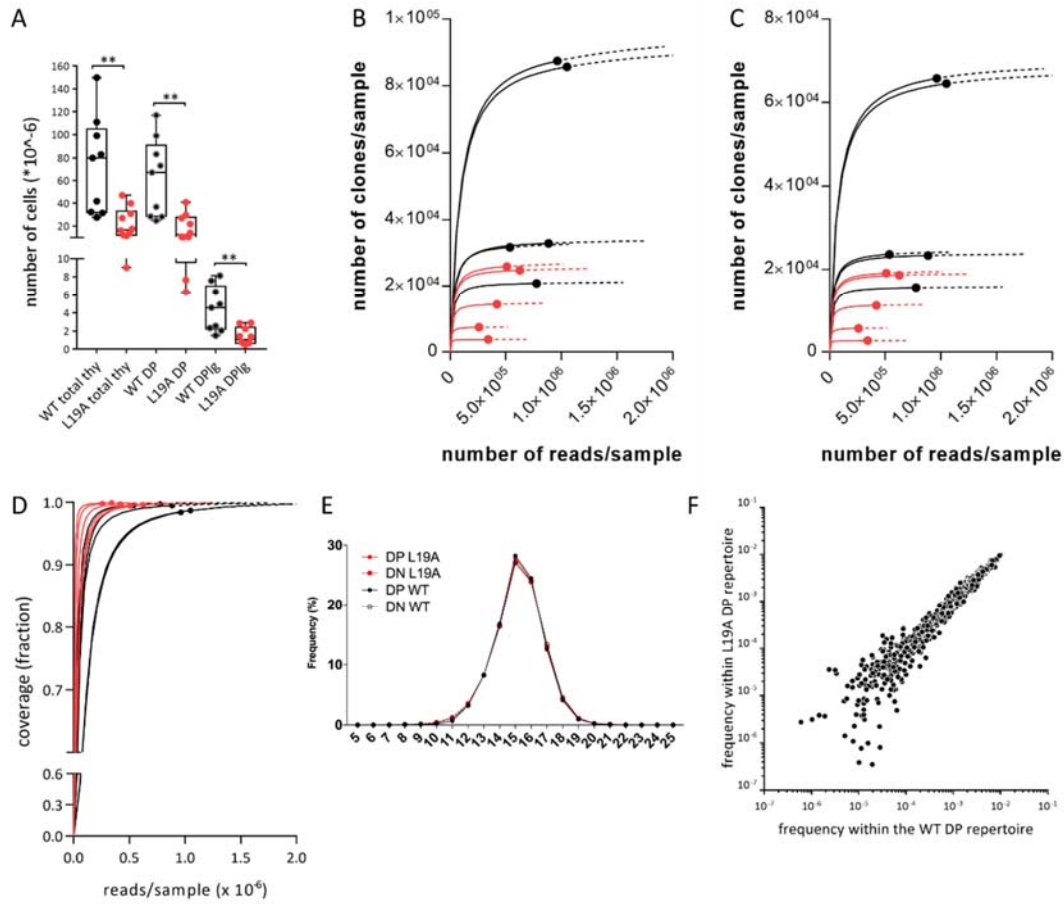


**Supplementary Figure 3. Assembly of WT and L19A pre-TCR in pre-T cell lines.** (A) CRISPR-Cas9 mediated targeting of exon 1 and exon 2 of the *Cd3z* gene in the pre-T cell line SCB.29. Forward or reverse oligos encoding sgRNAs specific for exon 1 and exon 2 of *Cd3z* (top panel, blue), upstream of PAM sequences (top panel, red), were cloned in the PX458 vector (Addgene #48138) (59) and transfected into SCB.29 cells. Two days later GFP<sup>+</sup> cells were single cell-sorted into 96 well plates. Individual clones were tested for cell surface expression of the pre-TCR and 2 pre-TCR-negative clones (11.10 and 21.1) were analyzed by CRISPR-STAT (60) (panels on right, overlays of elution profiles for original cell line and targeted clones indicating the size in bps of the amplified fragments). CRISPR-STAT and Sanger sequencing revealed an 8 bp deletion in clone 11.10 and a 2 bp insertion in clone 21.1. These indels give rise to proteins (lower panel) that in the 11.10 clone has homology with CD3ζ up to the 15<sup>th</sup> aa of the signal sequence (grey) and a 56 aa stretch with no homology to proteins in the databases. The 2 bp insertion in clone 21.1 causes a frameshift after the 9<sup>th</sup> aa of the transmembrane domain (red), resulting in a 12 aa extension with no homology to proteins in the databases. (B) Sequential immunoblotting with anti-CD3ζ and -tubulin antibodies of whole lysates of the SCB.29 cell line, *Cd3z*-deficient (ZKO) lines 11.10 and 21.1 and their derivatives reconstituted with lentiviral vectors encoding the WT or L19A CD3ζ-GFP fusion proteins (8). (C) Flow cytometric analysis of 11.10 ZKO, WT and L19A cell lines upon concomitant cell surface labeling with fluorescently labeled pTα-, TCRβ- and CD3ε-specific antibodies. Top panel: overlays of signals for each individual antibody and the CD3ζ-GFP protein. Middle and lower panels show bivariate plots of CD3ζ-GFP vs each of the antibodies for WT and L19A cells, respectively. Ratios of the MFIs of each combination is shown in the top right corner of each plot. Ratios between MFIs of the antibody-derived signals did neither show differences between WT and L19A cells (table on the right; black text WT, red text L19A). These findings were confirmed with the 21.1 ZKO, WT and L19A cell lines. (D) Sequential immunoblot analysis of anti-CD3ε and -TCRβ immunoprecipitates (ip) of 11.10 ZKO, WT and L19A cells with the indicated antibodies. Densitometry of the exposures showed a relative enrichment for the CD3ε protein vs the CD3ζ-GFP protein (normalized to  $\zeta/\epsilon$  and  $gfp/\epsilon$  ratio of WT cells). The lack of antibodies that recognize TCRβ or pTα in immunoblots did not allow us to distinguish between enrichment for CD3ε vs loss of the CD3ζ-GFP during the experimental procedure. Reduced or negligible amounts of CD3ε were recovered in anti-CD3ε and -TCRβ IPs from the

ZKO cell line as compared to the L19A and WT cell lines in various experiments, suggesting that in absence of CD3 $\zeta$  and/or assembly with TCR $\alpha\beta$  this protein is prone to degradation.



**Supplementary Figure 4. Pre-TCR mediated signaling in pre-T cell lines.** (A) Steady state phosphorylation of CD3ε. Immunoprecipitates with the CD3ε-specific antibody 2C11 from the SCB.29 cell line, its TCRβ-negative parental line SC1ET (20) and the two panels of ZKO, WT and L19A lines were separated by SDS-PAGE and immunoblotted with the phospho-εY1 antiserum (61) followed by blotting with the pan-CD3ε antibody M20 (Santa Cruz sc-1127). The graph on the right combines quantifications for 5 comparisons of 11.10 ZKO, WT and L19A cells and 4 comparisons of 21.1 ZKO, WT and L19A cells. P-values were calculated using a paired two-tailed Student's t test with 95% CI (\* p<0.05; \*\* p<0.01). (B) Stimulation-induced phosphorylation of Y142 of CD3ζ in 11.10 and 21.1 WT and L19A cell lines. Cell lines were stimulated for the indicated times with 5 μg/ml soluble 2C11 antibody followed by lysis in 0.3% Brij96 lysis mix supplemented with protease and phosphatase inhibition cocktails (Roche 11697498001 and 04906837001). Whole lysates were separated by NR SDS PAGE and immunoblotted with a CD3ζ-pY142-specific antibody (Sigma SAB4200334), followed by blotting with anti-tubulin or anti-GFP antibodies. Densitometric quantification, using tubulin or GFP signals as normalizers, is shown in the graphs to the right of the blots. (C) Lysates as prepared in B were immunoblotted with an antibody against pY319 of ZAP70 (Cell Signaling 27011). Quantification of signals, normalized by the corresponding signal with the anti-GFP antibody are shown to the right.



**Supplementary Figure 5. Properties of the TCR $\beta$  repertoire in DN and early DP thymocytes of WT and L19A mice.** (A) Comparison of compartment size of total thymocyte population (total thy), DP compartment (DP) and the CD69-FSChi DP thymocytes (DPig) in WT and L19A mice. Data were compiled from three independent experiments with a total of 9 WT and 10 L19A mice. P-values were calculated using an unpaired t test with Welch's correction (\*\* p < 0.01). (B) Number of clones per sample estimated with the 1D for WT (black) and L19A (red) early DP thymocytes. Solid lines: Rarefaction (interpolation) curve; dashed lines: extrapolation curves; symbols: observed diversity. (C) Number of clones per sample estimated with the 2D for WT (black) and L19A (red) early DP thymocytes. Solid lines: Rarefaction (interpolation) curve; dashed lines: extrapolation curves; symbols: observed diversity. (D) The coverage of the sampling process, that estimates the fraction of the real number of CD69-FSChi early DP clones actually detected in the experiment for OD. Solid lines: rarefaction (interpolated) coverage; Dashed lines: extrapolated coverage; Symbols: computed coverage using the actual experimental data. It can be seen that the coverage is almost 1 in all cases (the symbols are close to the asymptotic 100% coverage), meaning that the sampling has covered the real underlying repertoire almost completely. (E) Length distribution of the CDR3 region of WT and L19A DN3 and early DP thymocytes. (F) Scatter plot of frequencies of VDJ combinations encountered in the repertoires of WT and L19A early DP thymocytes. The coordinates of each dot correspond to the mean of frequencies of each VDJ combination in WT and L19A early DP thymocytes. None of the 831 VDJ combinations encountered in these early DP repertoires showed significant differences in representation between both repertoires as determined by using multiple t-tests corrected with the two-stage step-up method of Benjamini, Krieger and Yekutieli. Twenty-two out of the 831 VDJ combinations were not present in either the WT or L19A DP repertoire and have not been included in the logarithmic representation of the data.

**Dataset 1: Resampling analysis of TCR $\beta$  CDR3 repertoire in WT and L19A DN3 and CD69<sup>FSC<sup>hi</sup></sup> DP thymocytes.** Each individual repertoire was randomly sampled 200 times without replacement for 32000 reads (corresponding to the number of clones in the smallest repertoire obtained) and the number of times that each clone was selected was recorded. The number of times that each sequence (column A) was selected in the 200 resamplings is recorded in column B through Q and p-values of two-tailed, unpaired T-tests and adjusted p- values according to the method of Benjamini & Hochberg are shown for the comparisons of WT vs L19A DP repertoires (columns R and U), combined WT and L19A DN repertoires vs WT DP repertoires (columns S and V) and combined WT and L19A DN repertoires vs L19A DP repertoires (columns T and W).

## SI References

57. Linda JS Allen. *An introduction to stochastic processes with applications to biology*. CRC Press, 2010. <https://doi.org/10.1201/b12537>
58. Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology* **12**, 42–58 (1943)
59. F. A. Ran, *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308 (2013).
60. B. Carrington, G. K. Varshney, S. M. Burgess, R. Sood, CRISPR-STAT: an easy and reliable PCR-based method to evaluate target-specific sgRNA activity. *Nucleic Acids Res* **43**, e157 (2015).
61. E. P. Dopfer, *et al.*, Analysis of novel phospho-ITAM specific antibodies in a S2 reconstitution system for TCR–CD3 signalling. *Immunology Letters* **130**, 43–50 (2010).