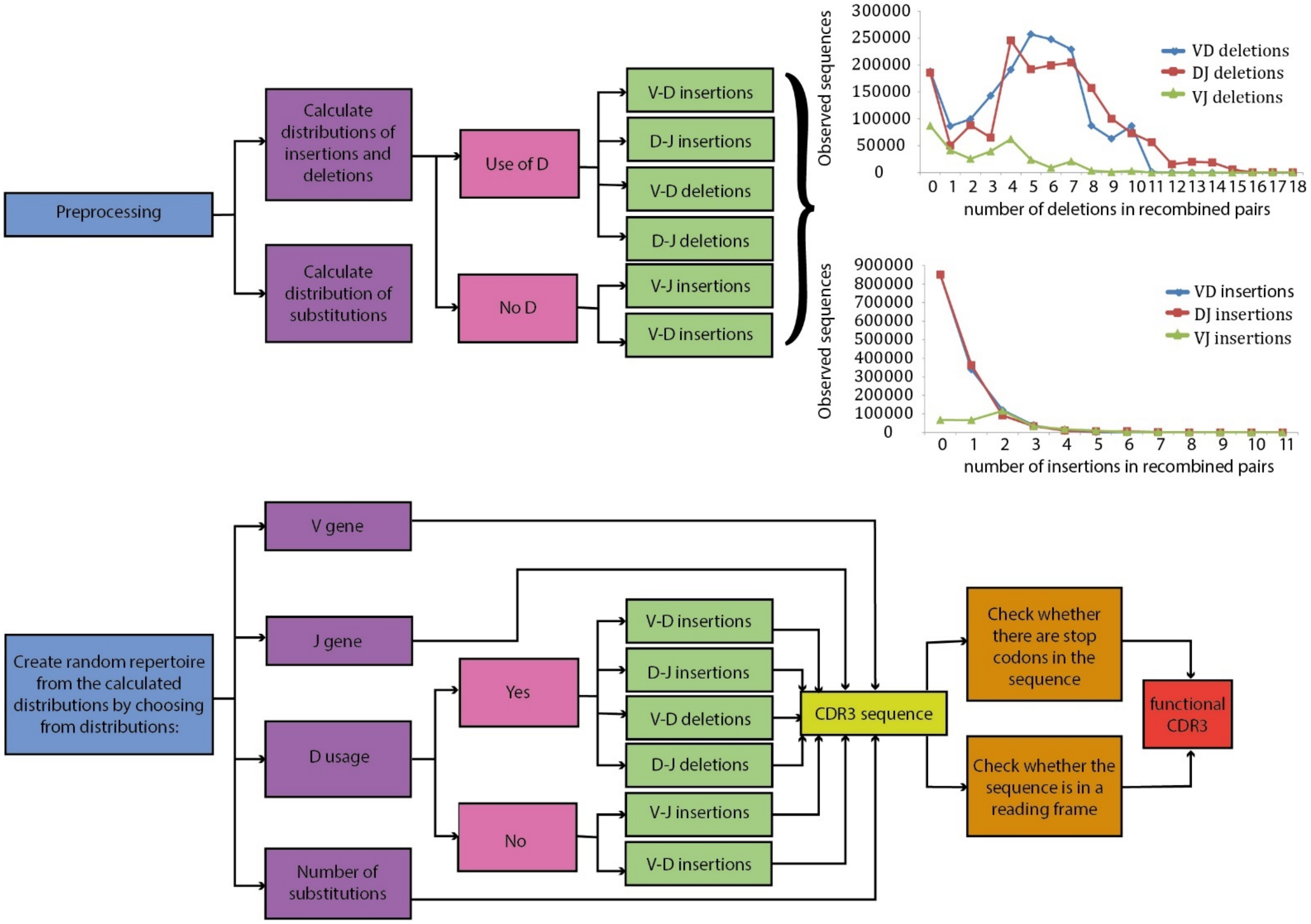
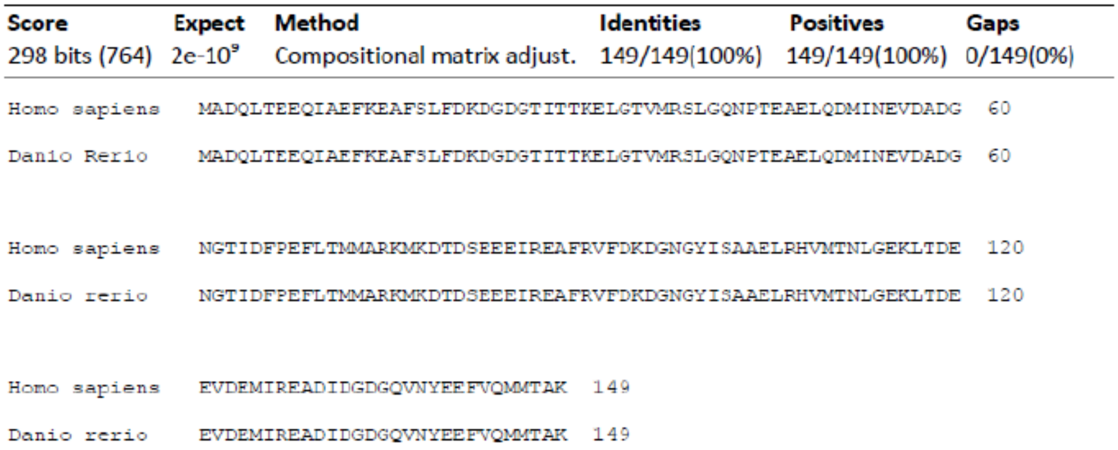

Preprocessing
Calculate distributions of insertions and deletions
Calculate distribution of substitutions
Use of D
No D
V-D insertions
D-J insertions
V-D deletions
D-J deletions
V-J insertions
V-D insertions
Observed sequences
300000
250000
200000
150000
100000
50000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
number of deletions in recombined pairs
VD deletions
DJ deletions
VJ deletions
Observed sequences
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
0 1 2 3 4 5 6 7 8 9 10 11
number of insertions in recombined pairs
VD insertions
DJ insertions
VJ insertions
Create random repertoire from the calculated distributions by choosing from distributions:
V gene
J gene
D usage
Number of substitutions
Yes
No
V-D insertions
D-J insertions
V-D deletions
D-J deletions
V-J insertions
V-D insertions
CDR3 sequence
Check whether there are stop codons in the sequence
Check whether the sequence is in a reading frame
functional CDR3

Figure S1

A



Score	Expect	Method	Identities	Positives	Gaps
298 bits (764)	2e-10^{9}	Compositional matrix adjust.	149/149(100%)	149/149(100%)	0/149(0%)

```
Homo sapiens    MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADG  60

Danio Rerio     MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADG  60


Homo sapiens    NGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDE  120

Danio rerio     NGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDE  120


Homo sapiens    EVDEMIREADIDGDGQVNYEEFVQMMTAK  149

Danio rerio     EVDEMIREADIDGDGQVNYEEFVQMMTAK  149
```

B

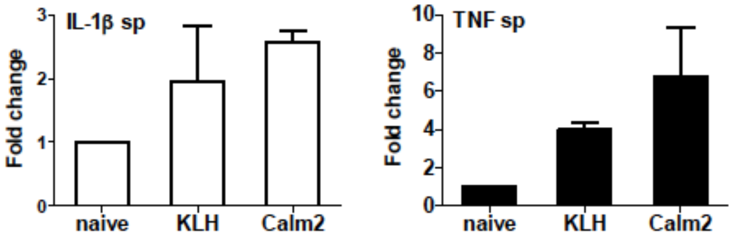


Figure S2

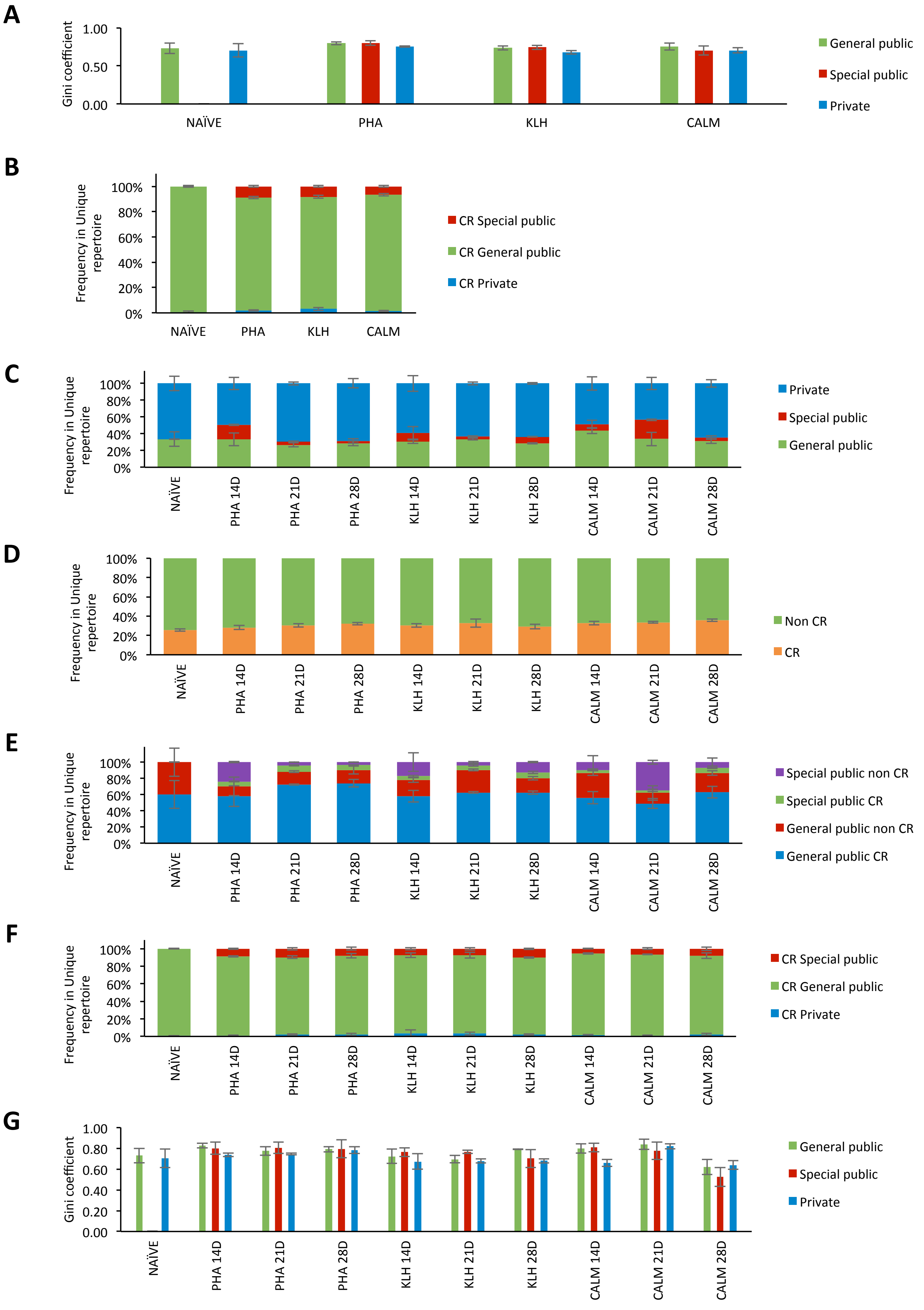

A
Gini coefficient
1.00
0.50
0.00
NAÏVE
PHA
KLH
CALM
General public
Special public
Private
B
Frequency in Unique repertoire
100%
80%
60%
40%
20%
0%
NAÏVE
PHA
KLH
CALM
CR Special public
CR General public
CR Private
C
Frequency in Unique repertoire
100%
80%
60%
40%
20%
0%
NAÏVE
PHA 14D
PHA 21D
PHA 28D
KLH 14D
KLH 21D
KLH 28D
CALM 14D
CALM 21D
CALM 28D
Private
Special public
General public
D
Frequency in Unique repertoire
100%
80%
60%
40%
20%
0%
NAÏVE
PHA 14D
PHA 21D
PHA 28D
KLH 14D
KLH 21D
KLH 28D
CALM 14D
CALM 21D
CALM 28D
Non CR
CR
E
Frequency in Unique repertoire
100%
80%
60%
40%
20%
0%
NAÏVE
PHA 14D
PHA 21D
PHA 28D
KLH 14D
KLH 21D
KLH 28D
CALM 14D
CALM 21D
CALM 28D
Special public non CR
Special public CR
General public non CR
General public CR
F
Frequency in Unique repertoire
100%
80%
60%
40%
20%
0%
NAÏVE
PHA 14D
PHA 21D
PHA 28D
KLH 14D
KLH 21D
KLH 28D
CALM 14D
CALM 21D
CALM 28D
CR Special public
CR General public
CR Private
G
Gini coefficient
1.00
0.80
0.60
0.40
0.20
0.00
NAÏVE
PHA 14D
PHA 21D
PHA 28D
KLH 14D
KLH 21D
KLH 28D
CALM 14D
CALM 21D
CALM 28D
General public
Special public
Private

Figure S3

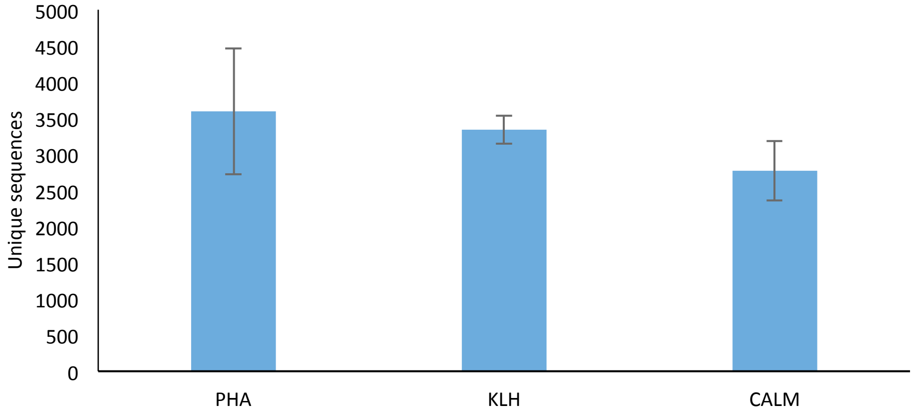

5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Unique sequences
PHA
KLH
CALM

Figure S4

Supplemental Information

Supplemental Figure Legends

Figure S1. **Simulation of V,D,J recombination in zebrafish, related to Figure 1.** Different stages during the simulation that produces random TCRβ1 repertoire sequences. The first step in the simulation is the pre-processing of sequencing data to obtain the distribution of the different events (deletions, insertions, substitutions) composing the variations leading to a full TCRb1 zebrafish repertoire. The second stage in the simulation is the production of the sequences that compose this synthetic simulated repertoire using the data obtain in the first stage.

Figure S2. **Zebrafish immunization, related to Figure 3. (A)** Alignment of human and zebrafish calmodulin 2 (Calm2) protein sequences. **(B)** Cytokine expression measured by qPCR in WKM+Spleens of male zebrafish immunized with KLH or Calm2 in IFA supplemented with LPS and CpG, one week after the booster injection. Mean ± S.E.M, n=3, each n constitutes of pooled organs from 2-3 fish.

Figure S3. **Time course analysis of the response of the TCRβ1 repertoire to stimulation, related to Figure 5. (A)** Gini coefficient analysis. **(B)** Fraction of general public, special public and private T cell clones within clones generated by convergent recombination in the unique repertoire. **(C)** Private, general public and special public fractions of the TCRb1 repertoire after immunization during different time points. **(D)** Convergent recombined clone fractions of the TCRb1 repertoire at different timepoints after treatment. **(E)** Fraction of the public repertoire within convergent recombined and non-convergent recombined clones of the TCRb1 repertoire at different timepoints after treatment. **(F)** Fraction of general public, special public and private clones within convergent recombined clones in the unique TCRb1 repertoire during different time points. **(G)** Gini coefficient analysis at different time points after treatment.

Figure S4: Response of the TCRα repertoire to stimulation, related to Figure 6. TCRa unique sequences in the different treatment groups.

Table S1: Primers used, related to the Experimental Procedures.

	Sequence 5'-3'	Purpose
Oligo(dT)12-18 primer	Invitrogen	RT-PCR
Smarter II A Oligo	AAGCAGTGGTATCAACGCAGAGTACXXXXX	RT-PCR
5' PCR primer IIA	SMARTer™Pico PCR cDNA Synthesis kit	Library amplification
Cβ1 primer	CATTTAGAATCTTTACGGATGGTTCACTCTTGGGA	Library amplification
α primer	CACTTGGTAAATATTCGGCTTCACTTCAGT	Library amplification
PE1 FCB ILL 1_2 V2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	Barcode addition
PE2 CB1	CAAGCAGAAGACGGCATACGAGATCATTTAGAATCTTTACGGATGGTTCACTCTTGGGA	Barcode addition
PE1 ILL BC 5PIIA 2_2 1	CACGACGCTCTTCCGATCTNNCGTGATAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 2	CACGACGCTCTTCCGATCTNNNACATCGAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 3	CACGACGCTCTTCCGATCTNNNNGCCTAAAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 4	CACGACGCTCTTCCGATCTNNTGGTCAAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 5	CACGACGCTCTTCCGATCTNNNCACTGTAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 6	CACGACGCTCTTCCGATCTNNNNATTGGCAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 7	CACGACGCTCTTCCGATCTNNGATCTGAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 8	CACGACGCTCTTCCGATCTNNNTCAAGTAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 9	CACGACGCTCTTCCGATCTNNNNCTGATCAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 10	CACGACGCTCTTCCGATCTNNAAGCTAAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 11	CACGACGCTCTTCCGATCTNNNGTAGCCAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 12	CACGACGCTCTTCCGATCTNNNNTACAAGAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 13	CACGACGCTCTTCCGATCTNNTTGACTAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 14	CACGACGCTCTTCCGATCTNNNGGAACTAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 15	CACGACGCTCTTCCGATCTNNNNTGACATAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 16	CACGACGCTCTTCCGATCTNNGGACGGAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 17	CACGACGCTCTTCCGATCTNNNCTCTACAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 18	CACGACGCTCTTCCGATCTNNNNGCGGACAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 19	CACGACGCTCTTCCGATCTNNTTTCACAAGCAGTGGTATCAACGCAGAGT	Barcode addition
PE1 ILL BC 5PIIA 2_2 20	CACGACGCTCTTCCGATCTNNNNGGCCACAAGCAGTGGTATCAACGCAGAGT	Barcode addition

Table S2. The 100 most abundant TCRα and TCRβ clones and their frequencies. Related to Figure 7.

TCR α		TCR β	
AA SEQ	**Frequency**	**AA seq**	**Frequency**
ALNSQFKIF	0.0313	AAYGQISGSYPAY	0.0124
ALNYGNKIT	0.0288	AARDRNYGDPAY	0.0079
VTDSGGWKVI	0.0177	AKQDRGRSEAH	0.0067
ALQHTGSLGKII	0.0143	AKQDGNNYNPAY	0.0067
ALQPAGYKKII	0.0139	AARERVGSSQAF	0.0064
ALQPVGLEKII	0.0122	ASRDNNRPAY	0.0064
ALQPNYNKIT	0.0107	AVSAAGNYQAY	0.0063
ALNNNYKII	0.0098	AARTATTSPAY	0.0062
ALDAARKII	0.0094	AASTGNNVNPAY	0.0061
ALRTDSGGWKVI	0.0093	AASAGQYDPAY	0.0060
ALVPGTGLQKVI	0.0086	AASTGNNYDPAY	0.0054
ALVKTGGYGKII	0.0085	AAGTGYRPAY	0.0038
ALVTDTGGRKVI	0.0080	AASRTGVNRPAY	0.0037
ALHPGFNKLM	0.0080	AAYRQGVGSSQAF	0.0036
ALNTGYKII	0.0075	AASNTQAY	0.0036
ALEPSGVKLI	0.0056	AVSAARFNPAY	0.0032
ALTSNWKVI	0.0052	AANRVMGSEAH	0.0032
ALDAGGGRKII	0.0051	AAYYRTGASQAF	0.0030
ALNQAGFQKLI	0.0051	AAYDRSGGGAQAF	0.0029
ALTSGVKLI	0.0049	AAYYRDANPAI	0.0029
ALQPTGNYKII	0.0049	AVSTGDSASPAV	0.0028
ALQPQSGSWKIH	0.0049	AARGESQPAY	0.0028
ALETSGDFAY	0.0048	AARTGNRDPAY	0.0027
ALDGSGRKII	0.0047	AASLTSGYPAY	0.0026
ALQYLGNKIV	0.0042	AASHNRPAY	0.0025
ALNDGTWKLH	0.0041	AAFYTGTSGYPAY	0.0025
ALMTNNRKIV	0.0040	AARHNANPAI	0.0025
ALVTAGYKLI	0.0040	AACYNYQAY	0.0024
ALQPAGNKII	0.0039	AAGQSGLQAY	0.0024
ALQPYGNNKIT	0.0039	AGHSGSYQAY	0.0023
ALQTTSVKIV	0.0038	AARELGSGGGAQAF	0.0022
ALQTTGGGGYKII	0.0038	AASKGHMGSEAH	0.0022
ALRPNNQKLI	0.0038	ASSIDTGASQAF	0.0022
ALTQSGFKFI	0.0037	AARTVYQDDPAY	0.0021
ALALNANKII	0.0036	AVRTGSANPAI	0.0021
ALRTSSQWKLM	0.0035	AAYTNNRPAY	0.0021
AMENFNKIT	0.0035	AAYNTGASQAF	0.0020
ALNRGGVDKLI	0.0035	AARDSFGSSQAF	0.0020
ALVPYGNKIT	0.0035	AKQMRQNYQAY	0.0020
ALQPNNQKLI	0.0035	ASSIGRNNRPAY	0.0020

ALDANTNKMI	0.0033	AASIQQSQPAY	0.0019
ALTQNAFKLI	0.0033	AAYYRNNANPAI	0.0019
ALQPNNYKII	0.0033	AARDRQYDPAY	0.0019
ALENYGNKII	0.0033	AALDINTANPAI	0.0018
ALATSVKIV	0.0032	AAYYNFNPAY	0.0018
ALQNYNKIT	0.0032	AAYSTSGYPAY	0.0017
AMEPDATRKII	0.0032	AAMTISGLQAY	0.0017
ALVTGTGVNKVI	0.0031	AASGTNYQAY	0.0017
ALNTGGLNKLI	0.0030	AASGQGYPAY	0.0017
ALQPNDGFKLF	0.0030	AARQNFNPAY	0.0017
ALNNNIKIV	0.0030	AVSTNQYDPAY	0.0017
ALVTQSGFKFI	0.0030	AAFPGTQTQPAY	0.0016
ALQPTGMASKIL	0.0030	AARGLGSEAH	0.0016
ALVPTGSLGKII	0.0028	AAYYIGYRPAY	0.0016
ALDTGGYGKII	0.0028	AVRDNSGTYPAY	0.0016
ALTSSQWKLM	0.0028	AASSGSYPAY	0.0016
ALQDANTNKMI	0.0028	AAYGGGAQAF	0.0016
ALRLTDTGGRKVI	0.0028	AACTGTGNPAF	0.0016
ALNVNKIT	0.0028	AARDINYGDPAY	0.0016
ALQNNNIKIV	0.0027	AVRDRNYGDPAY	0.0016
ALTFGATKII	0.0027	SVSQQGTGNPAF	0.0016
ALQPTAGYKLI	0.0027	AALDRTQPAY	0.0015
VTNNQKLI	0.0026	AASRDNQYDPAY	0.0015
ALNTGGYGKII	0.0026	AVSDRSTTSPAY	0.0015
ALVPGTGVNKVI	0.0026	AAYDNYQAY	0.0015
AMVPSGSGLYKVI	0.0026	AAYYYSTSGYPAY	0.0015
AMETQTGLQKIL	0.0025	AARQGKSQPAY	0.0015
ALVNDAYKIY	0.0025	AAWTNSGYPAY	0.0015
AMAQTGLQKIL	0.0025	AAYRENRPAY	0.0015
ALQNTGYKMV	0.0025	AASINSGGGAQAF	0.0015
ALQPSSYGGKLI	0.0025	AVRTGFGSSQAF	0.0014
ALDGSGLKII	0.0024	AARQGNTQAY	0.0014
ALRPAGYKLI	0.0024	AAYYQSQPAY	0.0014
AMVKTGGYGKII	0.0024	AAYGGFGSSQAF	0.0014
ALTDSGGWKVI	0.0024	AARQDNYDPAY	0.0014
ALRNQAGFQKLI	0.0024	AARTGNYGDPAY	0.0014
ALTTSGGIKII	0.0024	AARQGSSQAF	0.0014
ALVPGSGLKII	0.0024	AVSAGGYQAY	0.0014
ALMTTGVKII	0.0024	AAFYGNTQAY	0.0013
ALVPDAARKII	0.0023	AARDNYDPAY	0.0013
ALVPTGGYGKII	0.0023	AAYYQGNYQAY	0.0013
ALDGAARKIF	0.0023	AASLGAGYPAY	0.0013
ALVTTSVKIV	0.0023	AARTGYDPAY	0.0013
ALNYGNNKIT	0.0023	AKQDTGSGAGAQAF	0.0013
ALVPNNRKIV	0.0023	AASNMGSEAH	0.0013

ALKPTGGGYKLI	0.0022	AARAGGMGSEAH	0.0013
ALRPDSQFKIF	0.0022	AASDTQAY	0.0013
ALNYGSGNYKLI	0.0022	AAYNNFNPAY	0.0013
ALQPDSGGWKVI	0.0022	AASLGNYQAY	0.0013
ALEPTGGLNKLI	0.0022	AARELGTTSPAY	0.0013
ALQDGSGRKII	0.0021	AARESQPAY	0.0013
ALQPTSGGIKII	0.0021	AAYYDNNRPAY	0.0013
ALDGSGTKII	0.0021	AAYYSTANPAI	0.0013
ALEPTNNLKIV	0.0021	AAYTGDYNPAY	0.0013
ALRPQSGFKFI	0.0021	AAREGGTQPAY	0.0013
ALQTGSGNWKII	0.0021	AASYAQYDPAY	0.0013
ALQPNAGGLSKLM	0.0021	AAYSGTGYGQAY	0.0013
ALVPNAQKII	0.0021	AVSADYQAY	0.0012
ALNAGQKLI	0.0020	AARQNNYDPAY	0.0012
ALTYTGAQKLI	0.0020	AASYTSGYPAY	0.0012

Experimental Procedures

Fish maintenance. 1 year old male zebrafish (AB strain) were maintained in a 28-30°C system with a 14/10 hrs light/dark cycle. All experiments were carried out in accordance with guidelines by the Institutional Animal Care and Use Committee of Harvard Medical School.

Immunization. Fish were anaesthetized using 0.02% Tricaine methanesulfonate (Sigma-Aldrich) and immunized intra-peritoneally (i.p.) with a 10µl emulsion containing 1:1 Incomplete Freund's Adjuvant (IFA, Difco Laboratories) and 90% PBS (Invitrogen), 0.25µg lipopolysaccharide (ultrapure LPS, Invivogen), 0.7µg CpG Oligonucleotide ODN 1826 (Invivogen) and 2 µg of either PHA (Sigma-Aldrich), KLH (Sigma-Aldrich) or CALM (Creative BioMart, NY, USA). Two weeks later the fish were boosted with PHA, KLH or CALM in 1:1 IFA: 90% PBS.

We measured cytokine transcripts in spleens, whole kidney marrow and intestines from immunized fish, 7 days after boosting. Total RNA was purified with TRIzol® Reagent. The reverse transcription reaction was prepared from 800ng total RNA using the High capacity cDNA reverse transcription kit (Applied Biosystems). The qRT-PCR was performed on a VIIA™7 Real-time PCR System (Applied Biosytems) using a SybrGreen-based reaction from Takara (SYBR®*Premix Ex Taq*™ II, Takara Bio Inc.). The amplification program was 95°C 30sec; 95°C 5 sec, 60°C 30sec for 40 cycles; and the melting curve 95°C 15 sec, 60°C 1min for 1 cycle. The primers were designed to span over exon-exon borders using the Primer3Plus program (http://www.bioinformatics.nl/cgibin/primer3plus/primer3plus.cgi). Relative mRNA expression was calculated using the 2^-ΔCt formula with β-actin as the reference gene.

mRNA isolation, reverse transcription and 5' RACE. Total RNA was extracted from whole fish homogenate using Trizol®reagent (Life Technologies) and mini spin columns (Qiagen). After phase separation using Trizol and chloroform (Sigma-Aldrich), the upper phase was mixed with equal amounts of 70% ethanol (Sigma-Aldrich) and placed on top of a mini spin column (Qiagen). The rest of the procedure was performed according to the manufacturer's protocol (Qiagen) including the on-column DNAse digestion (RNase-free DNase set from Qiagen). mRNA was purified using µMACS™ mRNA isolation kits from Miltenyi Biotech and 160ng was used for reverse transcription using SuperScript™ II Reverse Transcriptase (Invitrogen) 0.5µg Oligo(dT)12-18 primer (Invitrogen) and 12pmol of SMARter II A Oligonucleotide (Clontech). The cDNA was double purified using NucleoSpin® columns (Clontech) and Agencourt® AMPure®XP beads (Beckman Coulter).

Library amplification and barcode addition. cDNA from each of fish (70ng) was used for TCRβ/α chain library amplification using Advantage® cDNA Polymerase mix (Clontech) with 10pmol of the 5´PCR primer IIA from the SMARTer™ Pico PCR cDNA Synthesis kit (Clontech) and the constant region primer **(Table S1)**. SybrGreen (Invitrogen) diluted in TE buffer was added to the reaction at 0.4X end dilution. As amplification references, the fluorescent standards from the Real-time Library Amplification kit (KAPA Biosystems) were used. The library was amplified using a VIIA™7 Real-time PCR System (Applied Biosystems) and the amplification program was: 95°C 5min; 95°C 30 sec, 72°C 2min15sec for 5 cycles; 95°C 30 sec, 68°C 2min 15 sec, n cycles (number of cycles needed for the PCR to reach an amplification level between fluorescent standard 1 and 3). The library was gel-purified using a double-comb 2% agarose E-gel® (Invitrogen) and barcodes were added to 150ng of library DNA using the same reaction as for the library amplification and the primers listed in **Table S1**. The program used was: 95°C 5min; 98°C 20 sec, 54°C 30 sec, 68°C 2min15sec for 2 cycles; 98°C 20 sec, 68°C 2min 15 sec for 10 cycles. The PCR products were purified using Agencourt AMPure XP beads and quantified a Library Quantification kit (KAPA Biosystems).

Verification of TCR library amplification. To verify the specificity of the C-region primer, the PCR library product amplified with the oligo dT primer and the C-region primer was sequenced. First, the PCR product was ligated into a TOPO-XL vector (Invitrogen). Chemically competent OneShot®TOP10 *Echerichia coli* (Invitrogen) were transformed with the library construct and 5 different colonies were picked and sequenced at the Dana-Farber/Harvard Cancer Center DNA Resource Core using the T7 primer.

Annotation and quality control. TCRβ and TCRa annotation was performed by using NCBI BLAST+ to identify the V and J germline genes of a TCR read, and then the CDR3 was determined by finding the conserved cysteine at the 5' end of the CDR3 and the conserved Phenylalanine at the 3' end of the CDR3.

The numbers of nucleotide additions was determined by taking the length of the CDR3 nucleotides and subtracting the number of nucleotides encoded by the V, J and D genes. A germline index was calculated by dividing the number of nucleotides in the CDR3 encoded by V, J and D genes by the length of the CDR3 producing a value between 0.0 and 1.0.

To identify potential sequencing errors, TCRβ species represented by a single sequencing read (that is, having a count of 1) were discarded. Additionally, the coverage was calculated for each sequenced sample. The coverage was defined as: total count of annotated reads / cell count used for library generation. TCRβ1 species represented by fewer reads than half of the coverage were then discarded, so that if a TCRβ1 species is represented by 4 reads and the coverage is 10x, that TCRβ1 species will be discarded.

PCR amplification error was addressed by identifying species in which the same V and J genes were used and the CDR3 was of the same length, but varied by only one nucleotide. The count of the species pair was then probed and if species A was found to be less than 5% of pecies B, then species A would be discarded as the likely product of PCR amplification error.

Simulation. To simulate V(D)J recombination and production of CDR3 sequences we used two distinct approaches – data based and random based. In both cases, the simulation follows the following mechanisms: V (D) combination, VD (J) combination, VD nucleotide deletion, DJ nucleotide deletion, nucleotide substitutions **(Suppl. Fig. 1)**. In one approach, we calculated these parameters out of available sequencing by iterating over sequenced recombined regions and obtaining V,J,D, deletion, insertion and substitution. During simulation, once V and D segments were chosen (in zebrafish, the choice for D stands between the single D segment being used or not used), parameters for deletion/insertion were selected according to the distribution of deletion segments obtained from sequencing; see **Suppl. Fig 1** for the actual values obtained for deletions, insertions, and the obtained V(D)J segment lengths used in CDR3. The random based version of the simulation was using random choices (within observed limits) for nucleotide combinations instead of measured values.

Probability of re-usage of sequences. To calculate the probability of a sequence in an immunized fish, based on that sequence being a shared on a non-shared sequence we followed the following procedure. For a group ***M*** including all CDR3 sequences in immunized fish, and a group ***N*** including all CDR3 sequences sequenced in naïve fish, we tagged a sequence ***s***

$$s \in N$$

according to whether or not it was shared within the naïve group. A sequence was tagged **public** if it was shared between at least two fish

$$s_{public}\colon s \in N_i, s \in N_j;\ i \neq j$$

And define $S = \{s_{public}\}$

The probability of observing a CDR3 sequence *z* in an immunized fish, based on its inclusion as a public sequence was therefore:

$$p(z \in M | z \in S)$$

Then, given the following table:

z	$z \notin M$	$z \in M$
$z \notin S, z \in N$	$n_{0,0}$	$n_{0,1}$
$z \in S$	$n_{1,0}$	$n_{1,1}$

Where $n_{i,j}$ is the number of sequences in each sub category, the probability becomes:

$$p(z \in M | z \in S) = \frac{n_{1,1}}{n_{1,0} + n_{1,1}}$$

This value was easily obtained from available sequence affiliations and was calculated for different groups in Figure 4.

In a similar manner, to obtain the probability of observing a sequence in the immunized group, based on the fact it was not a shared sequence in naïve fish, we calculated the probability:

$$p(z \in M | z \notin S) = \frac{n_{0,1}}{n_{0,0} + n_{0,1}}$$

We then used bootstrapping to obtain significance values, re-sampling with repetitions from the population of naïve sequences to obtain sequence pools and comparing with the immunized repertoire. This process was repeated 1000 times to produce distributions for $p(z \in M | z \in S)$ and for $p(z \in M | z \notin S)$.

V(D)J combinations. For each sequence, V D and J were identified according to the above section. As the ability to produce sequences without the use of the D segment is unique to zebrafish, we further divided the plots to the VJ usage with or without D segment. The Unique repertoire is defined as the subset of the total repertoire, without repetitions. Multiple usages of the same VJ pairs may appear even in the Unique repertoire, as multiple CDR3 sequences may originate from the same V(D)J segments.

Nucleotide sequence sharing. To compute CDR3 nucleotide or aa sequence sharing we first found all the unique sequences within a group and then we created a Venn diagram using "VENNY" tool representing the number of shared nucleotide sequences in all the different group combinations.

Convergent recombination. For each nucleotide sequence we found its amino-acid sequence, the groups sharing this nucleotide sequence and the groups sharing the amino-acid sequence. We counted the number of nucleotides passed from one combination of the groups to another - represent convergence. We visualized the results using Circos software package.

Calculating the upper and lower limit of TCR species. To estimate the potential number of zebrafish TCRab species we used our measurements of total TCRa and TCRb repertoires. Since the TCRab is composed of TCRa and TCRb chains, these limits are based on the possible combinations of the two chains. Calculation of the upper limit is a straightforward calculation of all possible combinations of TCRa and TCRb chains. That is, if we denote the upper limit as N_{upper} with N_{alpha} as the total number of unique TCRa sequences and N_{beta} as the total number of unique TCRb sequences. Then

$$N_{upper} = N_{alpha} \times N_{beta}$$

To calculate the lower limit of zebrafish TCRab species we used the number of copies detected of each TCRa and TCRb chain sequence. For example, we expect the most highly abundant TCRb chain to co-exist with the most highly abundant TCRa chain. Thus, to estimate the lower possible number of zebrafish TCRab species we matched the rankings of the copy numbers of TCRa and TCRb chains. That is, in the matrix M in which rows are the unique set of TCRa chains and of which columns are the unique set of TCRb chains, we tagged with 1 identical ranking from both sets.

The pseudo code for this is therefore:

```
$if\ \left(rank\left(alpha(i)\right) == rank\left(beta(j)\right)\right)$
$then$
    $m(i,j) = 1$
```

Since TCRa chains and TCRb chains do not have the same number of unique clones, and since this estimate is in essence an upper limit to the lower estimate, we spanned the vector of ranking of the alpha sequences to provide a complete overlap (an identical number of elements) to the beta vector. The pseudo code thus becomes

```
$if\ \left(spanned\_rank\left(alpha(i)\right) == rank\left(beta(j)\right)\right)$
$then$
$m(i,j) = 1$
```

Finally, we provide an additional layer of flexibility to the estimate, but relaxing the exact match to allow for nearest neighbor match on the matrix, that is:

```
$if\ \left(spanned\_rank\left(alpha(i \pm 1)\right) == rank\left(beta(j \pm 1)\right)\right)$
$then$
$m(i,j) = 1$
```

Which is the final (pseudo) code used for matrix assignment. The lower limit of zebrafish TCRab combinations was then calculated as a sum over the values of the matrix.

These calculations generated the matrix visualized in **Fig. 7**. We subsampled this hypothetical repertoire to represent the roughly 200,000 T cells in each fish. We sampled without repetition from the distribution of TCRab combinations described above. The mode of sampling has been determined so as to make sure the composition of the single fish repertoire maintains the rules of the distribution dictated by the possible combination of TCRa and TCRb sequences. To study the stability of these results we repeated the procedure 1000 times. The obtained result over 1000 repetitions has been $(1.7 \pm 0.002) \times 10^5$ and is therefore extremely stable.