Supplemental Material and Methods

Spectratype analysis

TCR β chains were amplified from cDNA by 26 individual PCRs. One TCR Cβ-primer and 26 TCR Vβ specific primers were used.[1] Subsequently, run-off reactions employing fluorescently labeled Jβ primers were analyzed on an ABI 3730 DNA sequencer (Applied Biosystems, Darmstadt, Germany) as described.[1, 2]

The complexity of Vβ chains was evaluated by counting the number of peaks and was graded on a score of 0-8.[3] A normal CDR3 variability is represented by 8 (to 10) distinct peaks and assigned a score of 8. In the absence of peaks a score of 0 was assigned. Overall TCR complexity was estimated by summing up all scores for each Vβ family. Scoring was performed double-blinded.

Analysis of T cell diversity by the TCR-Profiler

For unambiguous characterization of each individual TCR β-chain configuration, the TCR-Profiler performed the following steps. First, raw sequencing data were pre-processed using quality values (q-values) to trim the 3' end of TCR β-chain sequences**.** Phred-like quality values provided by the Roche/454 sequencer software were used as a measure for individual base-call reliability. Trimming was done by computing the average q-value over a specified number of bases starting from the 3' end. When the average q-value was below a given threshold (q = 30), the regarded number of bases were discarded (trimmed) and the next bases were considered iteratively.

Trimming stopped when the average q-value exceeded the threshold. We considered the remaining sequence to be of high quality, because the accuracy of the sequence has been shown to be higher for sequenced bases nearer to the 5’ end of the amplicon (with respect to the sequencing primer).[4]

Rearranged germline TRBV, TRBD and TRBJ genes were then identified by performing a Smith-Waterman local alignment[5] against each human TCR β-chain germline gene of the IMGT/GENE-DB reference directory.[6] Analysis quality was improved by using the sequencer-specific q-values in each alignment step. We incorporated these q-values into the Smith-Waterman local alignment algorithm by calculating a reliability score $r = 1 – (1/10^{(q/10)})$ for each base, which is then included as an additional weight to the substitution, insertion and deletion scores in the dynamic programming recurrence of the Smith-Waterman algorithm.

Finally CDR3 regions were delimited using specific amino acid sequence motifs flanking the junctional region at the 3' end of the TRBV gene and at the 5' end of the TRBJ gene. The amino acid sequence at the CDR3 3' end comprises the motif [W/F]GXG (IUPAC code, where X denotes any of the 20 common amino acids) that is conserved in all TCR β-chains. The CDR3 5' end discriminating amino acid sequence motif varies dependent on the rearranged germline TRBV gene. We identified the CDR3 5' end using the IMGT/GENE-DB reference directory sequence set. To examine the CDR3 length polymorphism in a patient sample, for each sequence the length of the CDR3 was identified as previously described.[7, 8] Automated screening for in-frame stop codons and out-of-frame transcripts was done to exclude non-productive CDR3 sequences.

Analysis of T cell diversity by the TCR-Profiler

The diversity measurement was based on the overall TCR repertoire and the complexity of all TCR β-chains that were predicted to be in-frame and without premature stop codons in the junction sites. More precisely, the detected CDR3 length polymorphism, the different combinations of TRBV and TRBJ families, as well as a nonparametric measure for characterizing the CDR3 sequence heterogeneity were evaluated. We tested three different scores regarding their individual ability to represent TCR diversity.

Similar to the evaluation of the TCR repertoire by immunospectratyping, our Length Complexity Score (*CS_L*) was calculated as the sum of the numbers of different CDR3 lengths L_v observed in each of the TRBV families *(v)*. The maximum value of L_v per TRBV family contributing to CS_L was limited to the value of 8.

$$CS_L = \sum_{v=1}^{m} min(L_v, 8)$$

The different combinations of rearranged TRBV and TRBJ genes were taken into account in our Combination Complexity Score (*CS_C*). The score calculates the sum of the numbers of different TRBJ gene families that were rearranged with each of the TRBV gene families.

$$CS_C = \sum_{v=1}^{m} \sum_{j=1}^{n} I_{vj}$$

where m is the number of different TRBV genes, n is the number of different TRBJ genes, and I_{vj} is the indicator function that expresses if a specific TRBV – TRBJ combination was observed:

$$I_{vj} = \begin{cases} 0 & = \text{combination not observed} \\ 1 & = \text{combination observed} \end{cases}$$

Each observed gene combination contributes a value of 1 to the complexity score.

Finally, we addressed diversity on the basis of individual CDR3 amino acid sequences in our next-generation-sequencing-spectratyping complexity score (CS_{NGS}). To this end we calculated the distribution of the occurrence frequencies of all individual CDR3 sequences of a sample. The upper tail of this distribution is of particular interest, because it represents CDR3 sequences occurring in elevated amounts due to a distortion of diversity. Percentiles were used as a nonparametric measure for description of this upper tail of the CDR3 distribution via the $\mathrm{CS}_{\mathrm{NGS}}$ score. To construct the score we used the logit of probability (ln(p/(1-p)) ≈ ln(p) because the largest value for low diversity is $p \ll 1$, and expected $\ln(p) < 0$. The CS_{NGS} score is constructed in a way, that lower values indicate less diversity; therefore ln(p) was multiplied by -1. From the resulting distribution, the lower tail was considered and low percentiles Q_i of distributions of -ln(p) were calculated, whereby CS_{NGS} presents the 5th percentile.

CDR3 coverage and error estimation

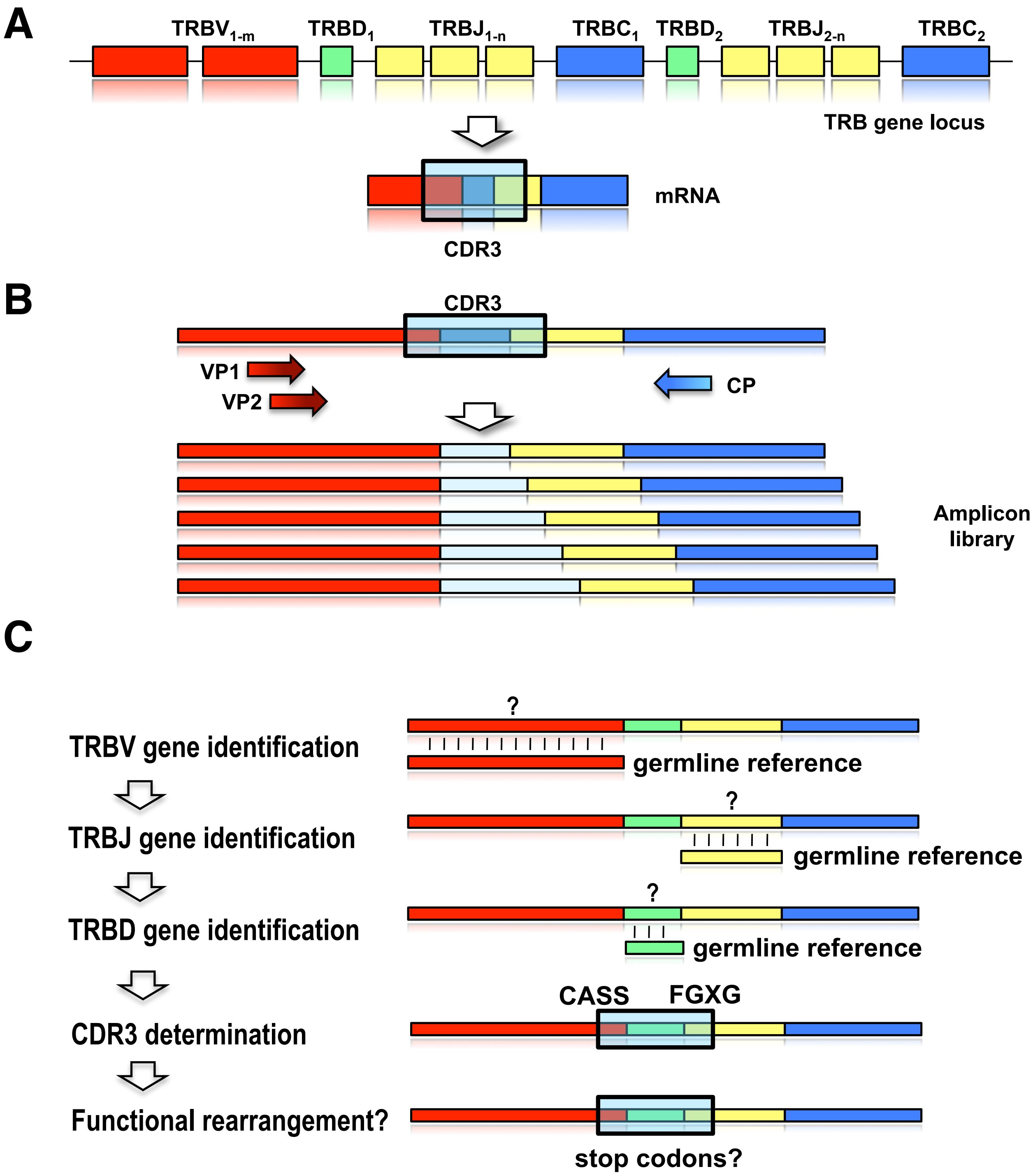
The mean coverage of the CDR3 region was 22.822 per patient sample. On average 13.434 input sequences per sample were identified as unique β-chain nucleotide sequences. 52% of these sequences were predicted to be productively rearranged. Non-productive rearrangements were predominantly caused by out-of-frame mutations and mutations altering the defining CDR3 boundary motifs CASS and FGXG (on average 65% and 27% of all unfunctional rearrangements, respectively). The frequency of premature stop codons due to the insertion of non-template nucleotides was low in all cohorts (3%).

Assessment of CDR3 diversity by NGS-S is not critically compromised by sequence errors caused during the workflow. For a CDR3 region of average length (36 nucleotides) the error rate of the employed high fidelity polymerase (2.8×10^{-6}) leads to an incorporation of one false base in 0.4% of all sequences after 40 cycles of PCR. For 454 pyrosequencing a mean error rate of 1.07% has been obtained.[22] With an average of 6989 productively rearranged CDR3 sequences analyzed per patient sample 28 sequences with 1 error due to PCR and 75 sequences with 1 error due to pyrosequencing can be expected.

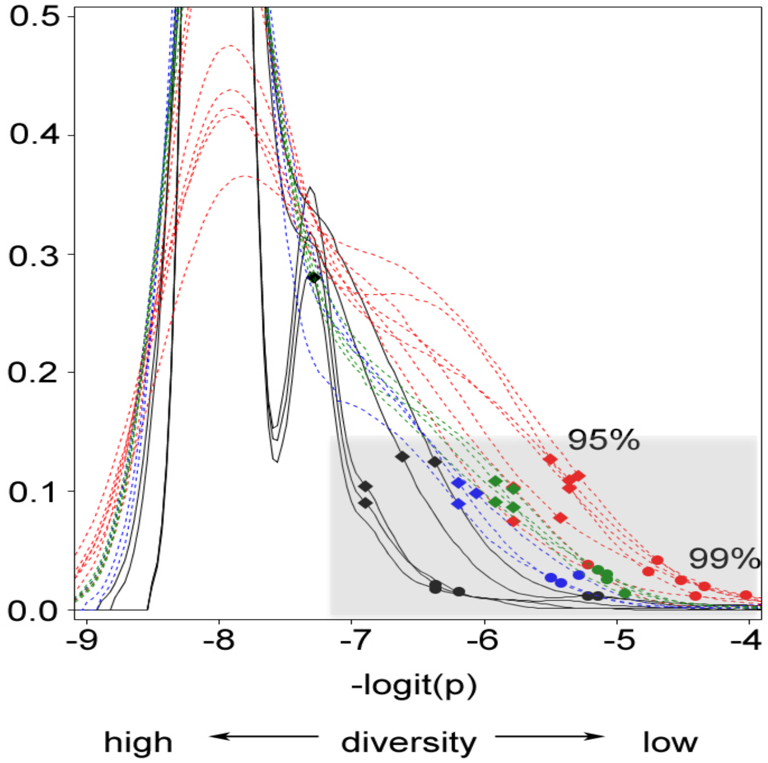
The impact of sequence errors introduced during PCR amplification and sequencing on TCR diversity determination was analyzed by processing a cDNA clone (Origene, Rockville, MD, USA) encoding TRBV5-4 (accession number BC028083.1) using the same workflow. The resulting phred quality score assigned to each sequenced base in the CDR3 region was ≥35.

References:

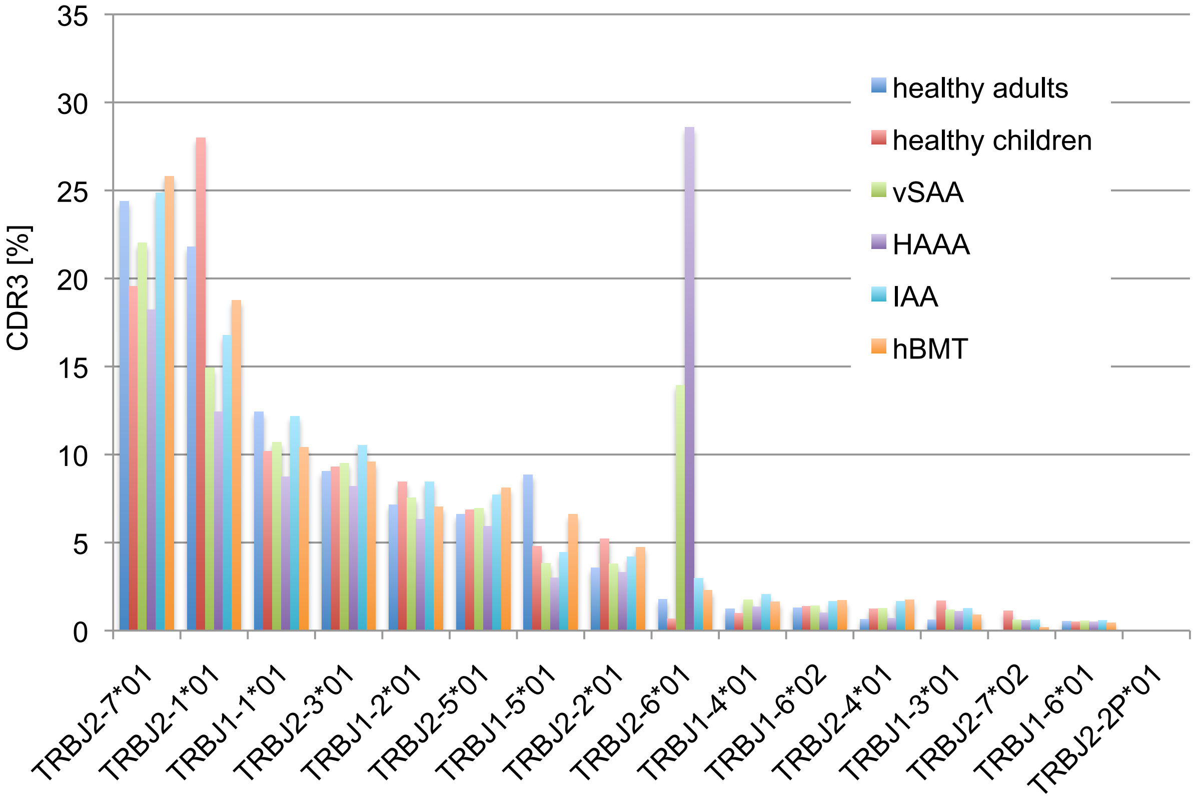
1. Monteiro J, Hingorani R, Peroglizzi R, Apatoff B, Gregersen PK. Oligoclonality of CD8+ T cells in multiple sclerosis. Autoimmunity. 1996;23(2):127-38.
2. Seitz S, Schneider CK, Malotka J, Nong X, Engel AG, Wekerle H, et al. Reconstitution of paired T cell receptor alpha- and beta-chains from microdissected single cells of human inflammatory tissues. Proc Natl Acad Sci U S A. 2006;103(32):12057-62.
3. Schuster FR, Hubner B, Führer M, Eckermann O, Gombert M, Dornmair K, et al. Highly skewed T-cell receptor V-beta chain repertoire in the bone marrow is associated with response to immunosuppressive drug therapy in children with very severe aplastic anemia. Blood Cancer Journal. 2011;1, e8; doi:10.1038/bcj.2011.6; published online 4 March 2011
4. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin JF. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics. 2011;12:245.
5. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195-7.
6. Giudicelli V, Chaume D, Lefranc MP. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. Nucleic Acids Res. 2005;33(Database issue):D256-61.
7. Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. 2009;19(10):1817-24.
8. Yousfi Monod M, Giudicelli V, Chaume D, Lefranc MP. IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. Bioinformatics. 2004;20 Suppl 1(i379-85.



Supplemental Figure 1: Next-generation-sequencing of the T-cell receptor β chain. (A) The TCR β-chain gene locus (TRB). In T lymphocytes variable (V), diversity (D) and joining (J) genes of the TRB gene locus are somatically rearranged to encode the V region. The junction site of V, D and J segments encodes the CDR3 region. V and C region of the β-chain are joined by posttranscriptional splicing to form the translated mRNA. **(B)** Generation of TCR β-chain amplicon libraries for next-generation-sequencing. RNA was extracted from bone marrow derived CD8+ T-cells and transcribed into cDNA. CDR3 amplicon libraries were generated by PCR employing two mixtures of degenerated wobble primers (VP1, VP2) designed to amplify most of the known TRBV genes and a universal reverse primer (CP) that anneals in the constant region. The mean overall coverage of the CDR3 region achieved was 22.822 amplicons per patient. **(C)** The TCR analysis workflow. The software automatedly identified rearranged germline TRBV, TRBJ and TRBD genes were identified by Smith-Waterman local alignment against each human TCR β-chain germline gene of the IMGT/GENE-DB reference directory. CDR3 regions were delimited using conserved flanking amino acid sequence motifs (CASS and FGXG, respectively). Screening for in-frame stop codons was done to identify non-productive TCR β-chain transcripts.



Supplemental Figure 2: Probability and robustness of complete separation between samples derived from healthy controls and patients with different T cell pathologies.
The plot shows distributions of probabilities of occurrences of individual CDR3 chains as yielded by kernel-density approach. The x-axis corresponds to probabilities (log scaled axis). The CS_{NGS} refers to negatives of x-axis values. The plot presents distributions of 19 data sets: 5 data sets of healthy children (black lines), 7 data sets of children after hBMT (red lines), 3 data sets of children with HAAA (blue lines) and 4 data sets of children with IAA (green lines).
–logit(p) values are high for patients with low TCR diversity as indicated. The grey rectangle underlies the region of percentiles where a complete separation between cohorts can be demonstrated. The result for the CS_{NGS} score is very robust when the tail of the distribution is considered (ln(p)> 90th percentile CS_{NGS}<10) and the chosen percentile is not too small (CS_{NGS}>1). CS_{NGS}>10 may reflect parts of the distribution which are not related to differences in diversity, and CS_{NGS}<1 may be strongly influenced by individual CDR3 sequences which hinders generalization. Thus, in our analyses we used CS_{NGS}=5.



Supplemental Figure 3: Frequency of TRBJ gene usage in the analyzed cohorts. Most frequently rearranged TRBJ genes and allels in the three analyzed cohorts (vSAA, hBMT and healthy control) are shown on the abscissa in descending order. The ordinate shows the number of reads in percent of all analyzed productively rearranged CDR3 sequences found in the respective cohort. The vSAA-subcohorts HAAA and IAA are also indicated.