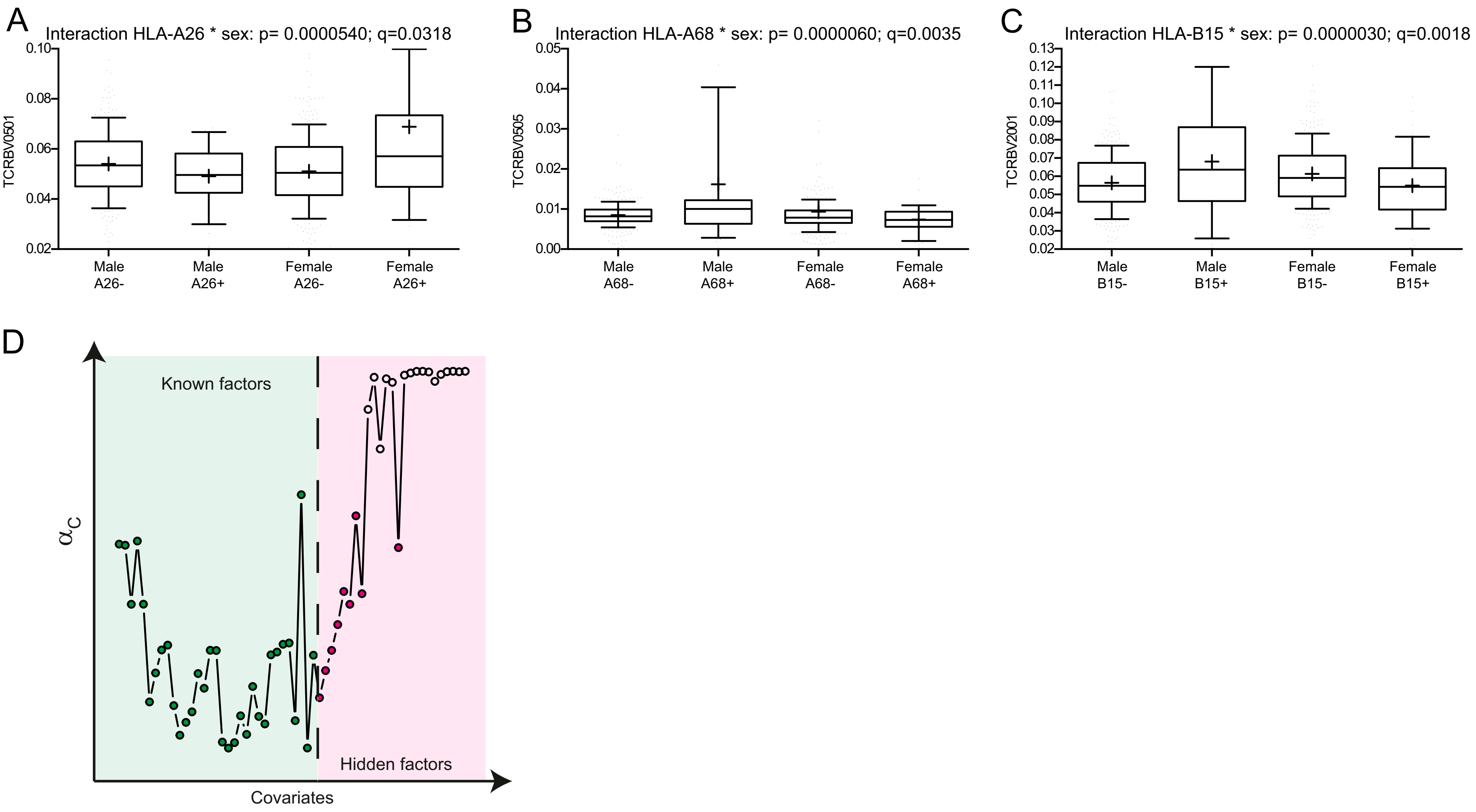
# Supporting Information

## Schneider-Hohendorf et al. 10.1073/pnas.1716146115

### SI Materials and Methods

Sequencing source material was either gDNA (cohorts 1 and 3–5) (1–5) or cDNA (cohort 2) (6), isolated from PBMC (cohort 1), CD4 T cells (cohort 2), CD8 T cells (cohorts 2 and 3), CD4, and CD8 naïve and memory T cells (cohorts 4 and 5). In brief, a multiplex PCR system was applied to amplify functional TCRBV CDR3 sequences from 31 nonpseudogenic, nonsegregating TCRBV gene segments in 19 gene segment families, both D genes and the 13 functional J segments (7, 8). This approach generated an 87-base pair fragment capable of identifying the VDJ region spanning each unique CDR3β (9). Amplicons were sequenced using the Illumina HiSeq platform. Using a baseline developed from a suite of synthetic templates, primer concentrations and computational corrections were used to correct for the primer bias common to multiplex PCR reactions. Raw sequence data were filtered based on the TCRβ V, D, and J gene definitions provided by the IMGT database (www.imgt.org) and binned using a modified nearest-neighbor algorithm to merge closely related sequences and remove both PCR and sequencing errors (10). Immunosequencing data from all presented cohorts is openly available for analysis and download using the ImmuneAccess database (https://clients.adaptivebiotech.com/immuneaccess). Data used for Fig. 1, such as TCRBV usage, HLA-A/B background, as well as all parameters and covariates are additionally listed in Dataset S2. BLOSUM scores for CDR1 and -2 similarities are listed in Dataset S5. TCRBV usage for all samples used in Figs. 2 and 3 are listed in Dataset S6 for CD4 and CD8 T cells and Dataset S7 for naïve and memory CD4 and CD8 T cell subsets.

1. Emerson R, et al. (2013) Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. *J Immunol Methods* 391:14–21.
2. Kanakry CG, et al. (2016) Origin and evolution of the T cell repertoire after post-transplantation cyclophosphamide. *JCI Insight* 1:e86252.
3. Dandekar S, et al. (2016) Shared HLA class I and II alleles and clonally restricted public and private brain-infiltrating αβ T cells in a cohort of Rasmussen encephalitis surgery patients. *Front Immunol* 7:608.
4. Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49:659–665.
5. Savola P, et al. (2017) Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis. *Nat Commun* 8:15869.
6. Schneider-Hohendorf T, et al. (2016) CD8(+) T-cell pathogenicity in Rasmussen encephalitis elucidated by large-scale T-cell receptor sequencing. *Nat Commun* 7:11153.
7. Dean J, et al. (2015) Annotation of pseudogenic gene segments by massively parallel sequencing of rearranged lymphocyte receptor loci. *Genome Med* 7:123.
8. Lefranc MP, et al. (1998) IMGT, the International ImMunoGeneTics database. *Nucleic Acids Res* 26:297–303.
9. Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114:4099–4107.
10. Carlson CS, et al. (2013) Using synthetic templates to design an unbiased multiplex PCR assay. *Nat Commun* 4:2680.



E

| Comparisons from Fig. 1A | Healthy<br>n=597 | With ethnicity as covariate<br>n=393 | With age as covariate<br>n=599 | With hidden factors<br>n=673 | Only determined HLA<br>n=602 |
|---|---|---|---|---|---|
| TCRBV0501 (HLA-A26*sex) | 0.003815 | 0.014163 | 0.000227 | 0.000538 | 0.000054 |
| TCRBV0505 (HLA-A68*sex) | 0.000002 | 0.000012 | 0.000002 | 0.000022 | 0.000006 |
| TCRBV2001 (HLA-B15*sex) | 0.002346 | 0.004051 | 0.000003 | 0.003060 | 0.000006 |

**Fig. S1.** Sensitivity analysis of comparisons from Fig. 1. (*A*–*C*) Data plotted for the Bonferroni-passed significant HLA/TCRBV combinations. Whiskers are at 10:90 percentiles, mean is shown as a +, p indicates uncorrected *P* values from Fig. 1, and q indicates Bonferroni-corrected *P* values. (*D*) The PEER package was used to infer hidden factors potentially influencing the analysis. AlphaC ($\alpha_C$) is a measurement for the importance of covariates according to the PEER algorithm with lower numbers indicating higher importance. For a sensitivity analysis of the Bonferroni-passed combinations from Fig. 1, nine hidden factors (filled magenta circles) in the range of known factors with regard to $\alpha_C$ ($\alpha_C < 100$) were used as covariates in the multivariate linear model in addition to the known factors (filled green circles). Dataset S2 lists the raw values of the 25 hidden factors. (*E*) Bonferroni-passed combinations from *A*–*C* were challenged by only assessing healthy individuals, by including ethnicity as covariate, by including age as a covariate, by including relevant hidden factors derived from PEER (1, 2) as covariates, or by only assessing individuals with determined HLA status (not inferred HLA status) (3). Shown are uncorrected *P* values.

1. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLOS Comput Biol* 6:e1000770.
2. Stegle O, Parts L, Piipari M, Winn J, Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 7:500–507.
3. Emerson RO, et al. (2017) Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 49:659–665.

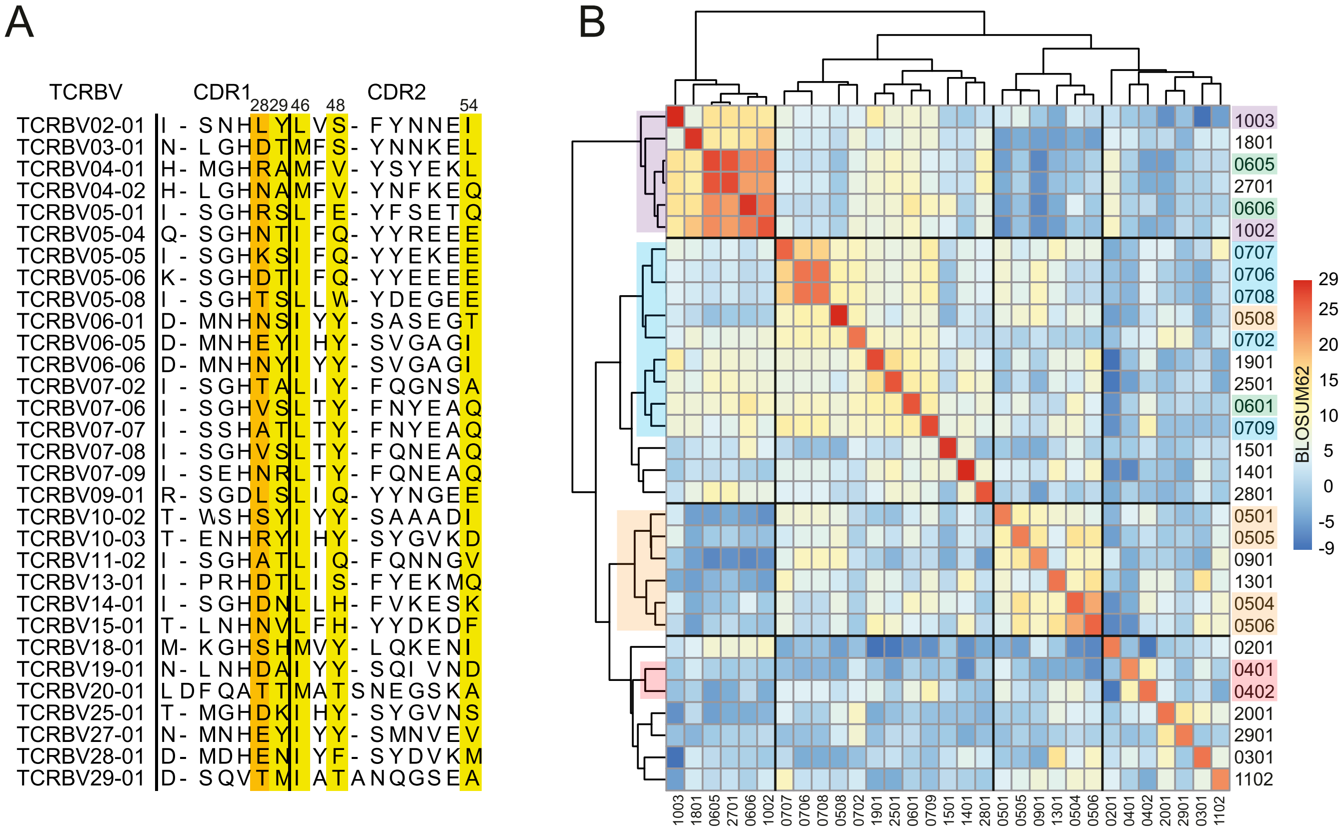


**Fig. S2.** Amino acid sequences and MHC I-binding positions of TCRBV CDR1 and CDR2. (*A*) Sequences of the complementarity determining regions (CDR) 1 and 2 and their MHC-binding amino acids (CDR1: positions 28 and 29 and CDR2: positions 46, 48, and 54). Anchor amino acids are highlighted in yellow; amino acids involved in the binding between CDR1 and MHC without evolutionary conservation are highlighted in orange (1). (*B*) BLOSUM62 scores of the CDR1 and CDR2 binding amino acids clustered according to Ward (2). The large TCRBV subgroups/families are highlighted in red (TCRBV04), orange (TCRBV05), green (TCRBV06), blue (TCRBV07), or purple (TCRBV10).

1. Marrack P, Scott-Browne JP, Dai S, Gapin L, Kappler JW (2008) Evolutionarily conserved amino acids that control TCR-MHC interaction. *Annu Rev Immunol* 26:171–203.
2. Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58:236–244.

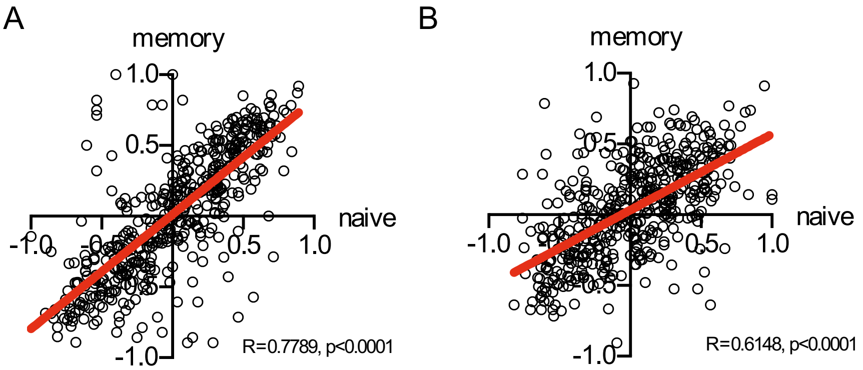


**Fig. S3.** Correlation of CD4 and CD8 naïve and memory TCRBV rhos. Correlation of TCRBV rhos from CD4 (*A*) and CD8 (*B*) naïve (*x* axis) and memory (*y* axis) T cells. Spearman's correlation coefficient (R) and respective *P* value are shown. Red lines illustrate linear regressions.

**Dataset S1. Details of the assessed cohorts**

Datasets S1

**Dataset S2. Raw data for Fig. 1**

Datasets S2

**Dataset S3. *P* values for Fig. 1**

Datasets S3

**Dataset S4. Effect sizes for Fig. 1**

Datasets S4

**Dataset S5. BLOSUM62 scores and Spearman's rhos for Figs. 2 and 3**

Datasets S5

**Dataset S6. Raw data for Fig. 2**

Datasets S6

**Dataset S7. Raw data for Fig. 3**

Datasets S7