Cell Reports Medicine, Volume *5*

Supplemental information

Quantifiable TCR repertoire changes in prediagnostic blood specimens among patients with high-grade ovarian cancer

Xuexin Yu, Mingyao Pan, Jianfeng Ye, Cassandra A. Hathaway, Shelley S. Tworoger, Jayanthi Lea, and Bo Li

Figure S1

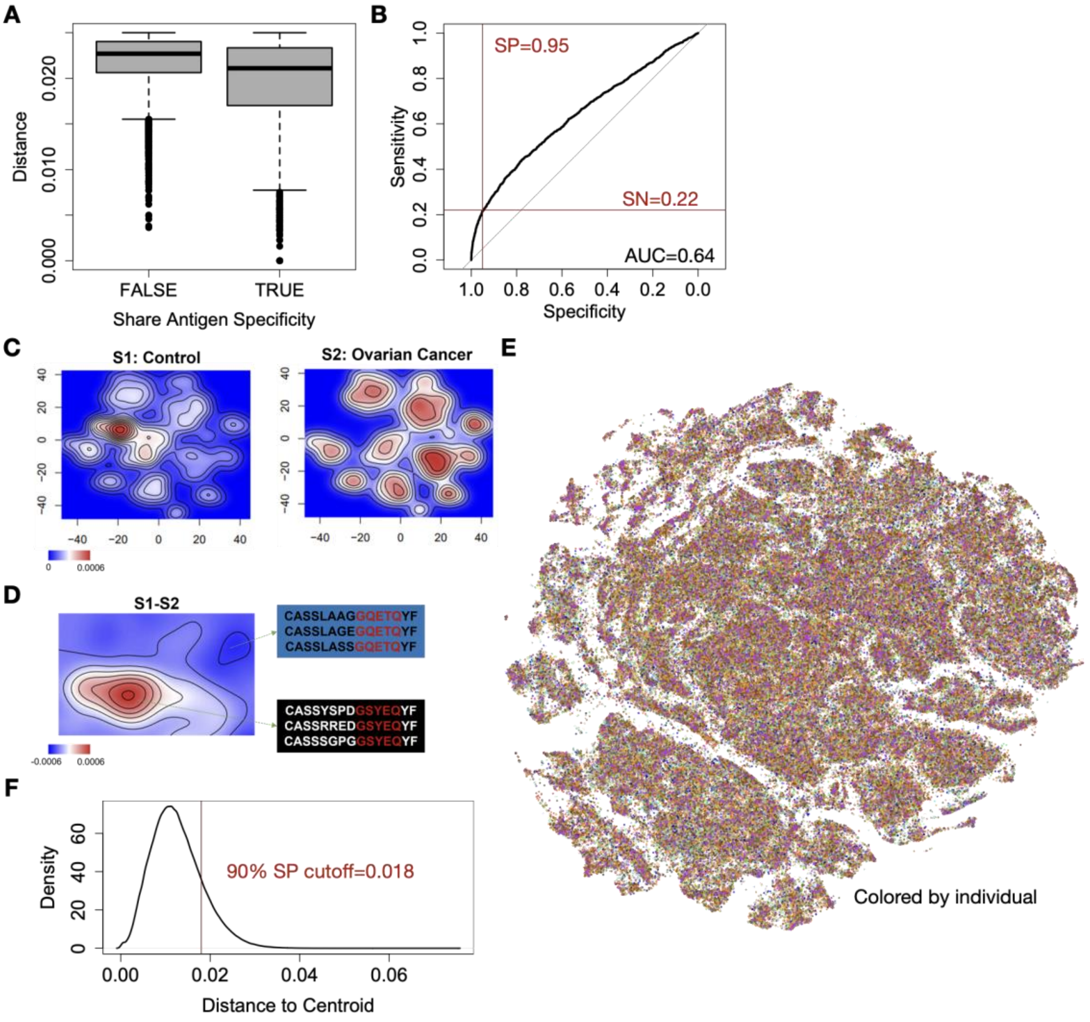

Figure S1. Benchmark of trimer-based embedding and repertoire functional units. Related to Figure 1.

(A) Boxplot showing the distributions of Euclidean distances between a pair of TCRs with known antigen specificities in the benchmark dataset.

(B) Prediction accuracy of Euclidean distance on if or not the pair of TCRs sharing specificity by ROC curve.

(C) 2-D density plot showing TCR distributions in a healthy control and an ovarian cancer patient.

(D) Density difference from (C) showing the enriched or depleted regions in the TCR embedding space. Selected TCR motifs were associated with these regions.

(E) Same t-SNE plot as in Figure 1G, except colored by different individuals.

(F) Distribution of Euclidean distance between any TCR to its assigned k-means cluster centroid. 90% specificity cutoff was determined by the ROC curve in b).

Figure S2

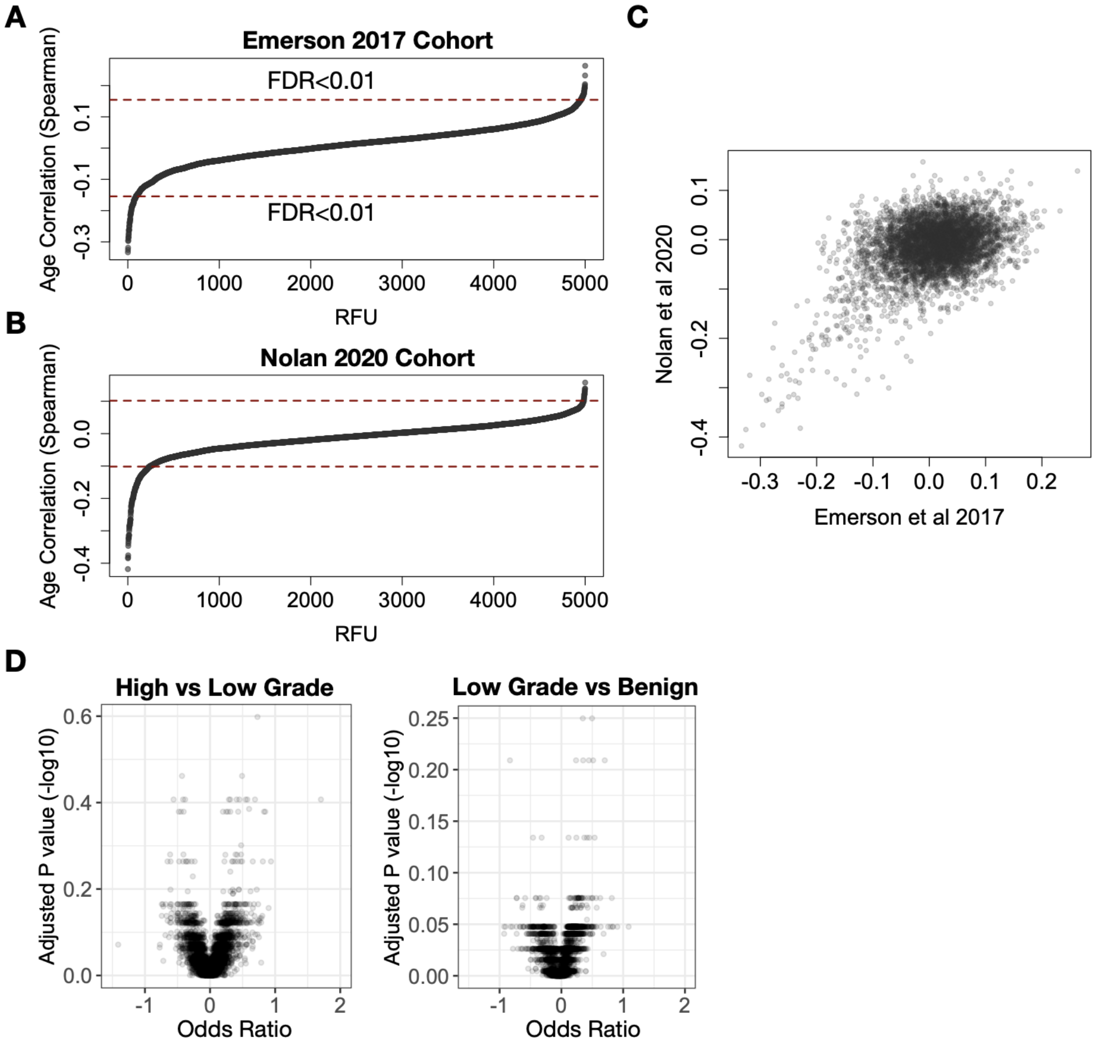

Figure S2. Additional analysis related to ovarian cancer samples collected in the discovery cohort. Related to Figure 2.

(A-B) Ordered Spearman’s correlations of age and each of the 5,000 RFUs with dashed red lines marking FDR<0.01.

(C) Scatter plot showing the relationships of age associations for each RFU between the two large healthy donor cohorts in (A-B).

(D) Volcano plots showing the output of the same analysis as in Figure 2f for the other disease categories in the discovery cohort.

Figure S3

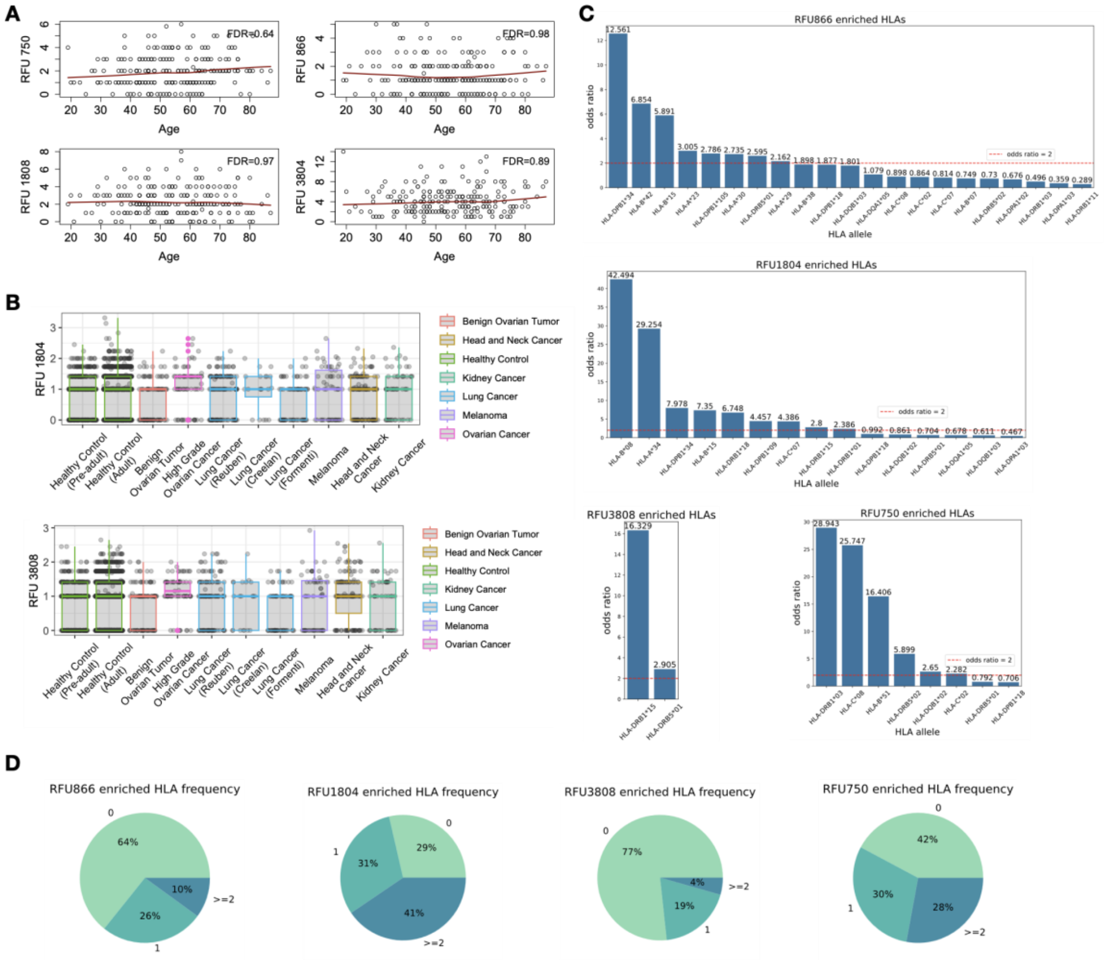

Figure S3. Distribution of selected RFUs across age or in the TCR repertoires of multiple cancers and HLA allele association. Related to Figure 3.

(A) Age association of selected RFUs in Figure 3A. Loess smooth curve was shown as red line in each panel. Spearman's correlation test was used to evaluate statistical significance and FDR was adjusted using the Benjamini-Hochberg method across all 5,000 RFUs.

(B) Same analysis as in Figure 3b showing the distributions of the upregulated RFUs in the top list.

(C) Odds ratio from Fisher's exact test showing the enrichment to individual HLA alleles for each of the 4 selected RFUs. Red horizontal line marks OR=2, which is the cutoff for calling an enrichment. All tests passed FDR=0.05.

(D) Pie chart showing the percentage of individuals in the training cohort carrying 0, 1 or ⩾2 HLA alleles associated with each of the 4 RFUs.

Figure S4

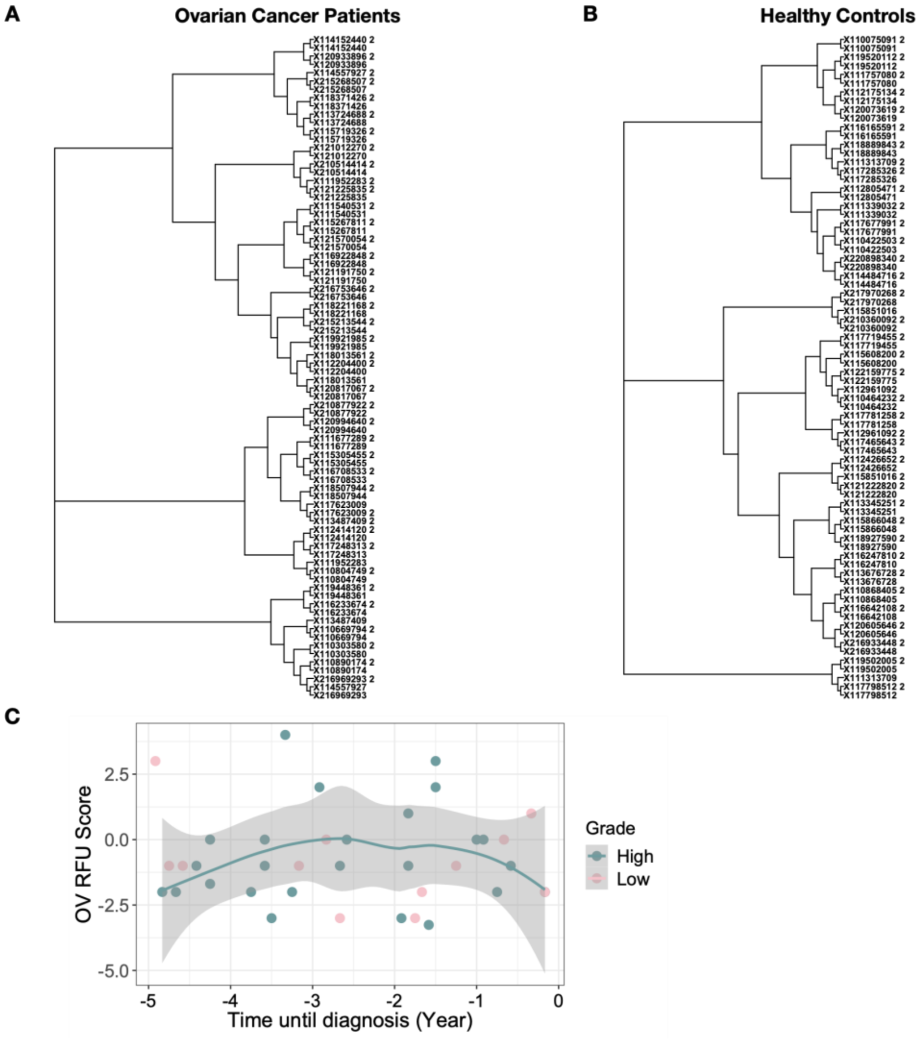

Figure S4. Neighbor-joining tree of TCR repertoire samples from NHS cohort and the prediagnostic curve of OV RFU scores using additional validation samples. Related to Figure 4.

(A-B) Distance matrix was calculated as squared root of 1- Spearman's correlation for the patient **(A)** and donor **(B)** samples separately. Neighbor joining trees were generated using the distance matrices. The second timepoints were marked with '2' at the end of the label.

(C) Same analysis was performed on the OV RFU scores of the second NHS cohort as described in Figure 4D. Smooth curve was generated using high grade samples only. Permutation analysis was performed to evaluate the statistical significance of this curve matching the shape of Figure 4D. 10,000 permutations of OV RFU scores were performed and the number of times when the Loess smooth curve satisfying the following 3 criteria were counted: (1) single peak between -3.5 to -2.5 years prior to diagnosis; (2) only increase before the peak; (3) only decrease after the peak. A total of 381 curves met this standard, with empirical p value = 0.038.