**Supplementary Figure Legends:**

**Figure S1. Performance evaluation for iSMART, TCRdist and GLIPH. a**) Fraction of >3 TCR clusters called by TCRdist at different distance cutoffs. A dramatic increase is observed at 12, which is selected as the default cutoff for TCR grouping. **b**) Cluster purity and count distributions across different cluster sizes. **c**) Purity distributions for large TCR clusters with more than 3 sequences.

**Figure S2.** **Cluster antigen-specificity analysis for iSMART, TCRdist and GLIPH.** **a**) Heatmap representation for cross-antigen classification errors for iSMART. Each entry in the off-diagonal matrix is an integer representing the number of CDR3 groups showing co-clustering of the two antigens. Diagonal entries recorded the number of clusters with at least 2 CDR3s assigned to the corresponding antigen. **b**) Clustering specificity as measured by purity shown for each of the 15 antigens.

**Figure S3. Summary of iSMART identified CDR3 clusters. a**) Histogram of CDR3 cluster size distribution. Median cluster size is 2, with range 2 to 248. There are 670 clusters with more than 3 CDR3s. **b**) Distribution of the counts for clustered CDR3s carried by each individual in the analysis. The median count is 1, with range 1 to 119.

**Figure S4. Clustered CDR3s as an indicator for activated T cells.** Volcano plot for genes positively correlated with number of clustered CDR3s. Median values across different cancer types for each gene were calculated for both p value and partial Spearman’s correlation with tumor purity correction. Top genes (p≤10-15) were zoomed in for visualization. Genes related to negative regulation for Treg cells and selected T cell activation markers were highlighted with dark red color.

**Figure S5. Differentially expressed gene analysis for the tissue resident memory subpopulations. a**) Volcano plot for Trm1 group against all the other T cells in sample BC10. Wilcoxon rank sum test was applied to evaluate the statistical significance, and p values were corrected by Bejamini-Hochberg method. We labeled the genes with mean fold change greater than 3 or smaller than -3, and FDR≤0.01. Established markers for tissue-resident memory T cells were highlighted with colors: red for up-regulation and green for down-regulation. **b**) Volcano plot for Trm2 cells against Trm1, with both groups defined in Figure 3c. Same analysis was performed to estimate p values and fold change. Genes with FDR≤10-6 were displayed.

**Figure S6. Additional tSNE plot visualization for selected markers.** Previously reported Trm markers including CD69 (general T cell activation and memory differentiation), CXCR6, Blimp-1 and CD103 (ITGAE) were visualized by tSNE plots. Expression patterns for two putative T cell exhaustion markers, PD-1 and TIM3 were also presented. Figure legend was in log scale.

**Figure S7. Mitochondrial gene fraction in different T cell subpopulations.** Mitochondrial gene expression levels were measured by UMI counts, with gene symbols starting with MT-, as defined in the Seurat package. The fraction was defined as the total UMI counts for all the mitochondrial genes over the total UMI count of the sample. Three T cell subpopulations of interest were displayed, in addition to the overall population. A horizontal line at 10% was displayed on the plot as the cutoff for dying cells.

**Figure S8. Pseudotime trajectory plots for individual clonotypes in sample BC10.** βCDR3 sequences were displayed as figure titles and each point on the plot represent a cell. Red circles label cells with the corresponding CDR3 sequence. The numbers in the figure titles are the number of cells in the corresponding clonotype. Two evolutionary patterns were observed. Pattern 1 is from Tpre to Trm1, including clonotypes CASRPPAGELAFF, CASSLWGDTQYF, CASSLSGSPKGEQYF, CASRTSGASTDTQYF and CASRTSGDFSYEQYF. Pattern 2 is from Trm1 to Trm2, including clonotype CSARDGNTEAFF.

**Figure S9. Single cell trajectory analysis for breast cancer sample BC11. a**) Same trajectory analysis in BC11, with exception that for the selected markers of the cell clusters, bold characters indicate FDR<0.05, where normal font otherwise. **b**) Boxplot for representative metabolic genes, with statistical significance evaluated by Wilcoxon rank sum test. **c**) Trajectory plots with individual clonotype overlaid by red circles. The number after each CDR3 sequence is the number of cells in the corresponding clonotype. Two evolutionary patterns were observed. Pattern 1 is from Tpre to Trm1, including clonotypes CASTDREGRYEQYF, CASSPDGKETQYF, CASSRDGQGNTIYF, CASSYSKVVLYGYTF and CASSPPSGSLGETQYF. Pattern 2 is from Trm1 to Trm2, including clonotypes CASSGTSGSYNEQFF, CASSLAPVSNYGYTF, CACSSGRYTGELFF and CQPVRDRGIYNEQFF.

**Figure S10. *HSFX1* expression across cancer types and normal tissues.** **a**) TPM values for HSFX1 expression were displayed in boxplots across 32 cancer types. Adjacent normal samples with sufficient sample size (n≥20) were also included in the plot. **b**) TPM values for HSFX1 expression displayed in boxplot across 53 normal tissue types as reported by the GTEx data portal. Box colors distinguish major tissue types: yellow for brain, pink for ovary, light gray for prostate, dark gray for testis and olive green for spleen.

**Figure S11. Generation of a potentially immunogenic peptide by *HSFX1* expression.** Violin plot showing significant difference in *HSFX1* expression between clustered and non-clustered individuals from all TCGA cancers. 9 individuals with solved HLA genotypes have positive *HSFX1* expression.

**Figure S12. Immunospot assays for *HSFX1*-derived peptide using naïve C57BL/6J mice and immunohistochemistry staining of *HSFX1* in normal human tissues.** **a**) Representative results showed IFNγ secreting cells from indicated groups. Column texts labeled the 3 treatment groups, where the T cells were collected from mouse models with humanized HLA\*A02:01 (left) or mouse H-2Kb alleles (right). Row texts labeled simulants used in the ELISPOT assay. Experiments performed using naïve immunocompetent C57BL/6J mice (n=4, all female for each group). PMA+ionomycin: standard positive control. **b**) IHC results of *HSFX1* on a normal tissue array. All slides were viewed and interpreted by board-certified surgical pathologist. Different resolutions were selected to visualize the structures of the corresponding tissues. Placenta is the only tissue with positive staining. Stromal staining between colonic glands (Colonic Mucosa) potentially results from binding to non-specific immunoglobulin Fc ligands expressed by B lymphocytes or macrophages, a known artifact for colonic tissues. The colonic epithelial cells are strictly negative.

**Figure S13. TSSK2 expression across human cancers.** Boxplot for *TSSK2* expression across different tumor and normal tissues.

**Supplementary Table Legends:**

**Table S1.** Gene-Ontology enrichment analysis was performed for the top 500 genes and the top 10 pathways were displayed. Highlighted pathways were related to immune cell activation.

**Table S2.** NetMHC prediction for a HSFX1 derived 9-mer peptide, with strong binding affinities to 3 common HLA alleles.

**Table S3.**  HLA genotype information table for the 9 individuals with HSFX1 expression shown in Figure S11. All 9 samples were colon or endometrial cancers, bearing at least one matched common HLA binder(s) as shown in Table S2 (highlighted in yellow). Expression levels in transcript per million (TPM) were also displayed in the table.

**Table S4.** NetMHC predicted binding peptide information for TSSK2.

**Table S5.** HLA genotypes for samples with TSSK2 expression.

**Supplementary Dataset 1:**

**Supplementary file -15\_antigens.** Information for the 15 selected antigens for methodology performance evaluation.

**Supplementary file -TCRcount\_correlation.** Summary of association analysis between gene expression levels and clustered CDR3 count in each individual.

**Supplementary file -New\_Group\_genes.** Summary of differential gene expression analysis between the single cells from new defined group and others, with fold change and FDR estimations.

**Supplementary file -Trm2vsTrm1\_DETable.** The differential expressed genes of Trm2vsTrm1.

**Supplementary file -CDR3\_Cluster\_Gene\_Correlation:** Summary of differential gene expression analysis between each CDR3 cluster and other TCGA tumors, with fold change and FDR estimations.

**Supplementary file -TCGAtcrClustered\_iSMART.** 4.6K iSMART clustered TCRs from TCGA tumor RNA-seq data.

**Supplementary Dataset 2.** TCR clusters obtained from the top 10000 most abundant clonotypes of 666 HCMV cohort.