

Lineage tracking reveals dynamic relationships of T cells in colorectal cancer

Lei Zhang^{1,6}, Xin Yu^{2,6}, Liangtao Zheng^{1,6}, Yuanyuan Zhang^{3,6}, Yansen Li⁴, Qiao Fang¹, Ranran Gao³, Boxi Kang³, Qiming Zhang³, Julie Y. Huang², Hiroyasu Konno², Xinyi Guo³, Yingjiang Ye⁴, Songyuan Gao⁵, Shan Wang⁴, Xueda Hu³, Xianwen Ren³, Zhanlong Shen^{4*}, Wenjun Ouyang^{2*} & Zemin Zhang^{1,3*}

T cells are key elements of cancer immunotherapy¹ but certain fundamental properties, such as the development and migration of T cells within tumours, remain unknown. The enormous T cell receptor (TCR) repertoire, which is required for the recognition of foreign and self-antigens², could serve as lineage tags to track these T cells in tumours³. Here we obtained transcriptomes of 11,138 single T cells from 12 patients with colorectal cancer, and developed single T cell analysis by RNA sequencing and TCR tracking (STARTRAC) indices to quantitatively analyse the dynamic relationships among 20 identified T cell subsets with distinct functions and clonalities. Although both CD8⁺ effector and ‘exhausted’ T cells exhibited high clonal expansion, they were independently connected with tumour-resident CD8⁺ effector memory cells, implicating a TCR-based fate decision. Of the CD4⁺ T cells, most tumour-infiltrating T regulatory (T_{reg}) cells showed clonal exclusivity, whereas certain T_{reg} cell clones were developmentally linked to several T helper (T_H) cell clones. Notably, we identified two IFNG⁺ T_H1-like cell clusters in tumours that were associated with distinct IFNγ-regulating transcription factors—the GZMK⁺ effector memory T cells, which were associated with EOMES and RUNX3, and CXCL13⁺BHLHE40⁺ T_H1-like cell clusters, which were associated with BHLHE40. Only CXCL13⁺BHLHE40⁺ T_H1-like cells were preferentially enriched in patients with microsatellite-instable tumours, and this might explain their favourable responses to immune-checkpoint blockade. Furthermore, IGFLR1 was highly expressed in both CXCL13⁺BHLHE40⁺ T_H1-like cells and CD8⁺ exhausted T cells and possessed co-stimulatory functions. Our integrated STARTRAC analyses provide a powerful approach to dissect the T cell properties in colorectal cancer comprehensively, and could provide insights into the dynamic relationships of T cells in other cancers.

Tumour-infiltrating lymphocytes are highly heterogeneous with respect to their cell-type compositions, gene expression profiles and functional properties, which might contribute to diverse responses to cancer immunotherapies¹. Recent clinical trials have demonstrated that patients with colorectal cancer (CRC) who display microsatellite instability (MSI) but not microsatellite-stable (MSS) phenotypes respond to the immune-checkpoint blockade of PD-1⁴, but the underlying mechanisms are not fully understood^{5–7}. Several single-cell RNA sequencing (RNA-seq) studies have revealed diverse subsets and functions of T cells in various cancer types^{8–10}. Here we developed an integrated approach, STARTRAC, to track further the dynamic relationships among T cell subsets identified inside colorectal carcinoma, adjacent normal mucosa and peripheral blood, based on both single-cell transcriptome and TCR α- and β-chain sequences as lineage-specific markers^{2,11} (Extended Data Fig. 1a, b). STARTRAC incorporated several unique indices, including STARTRAC distribution (dist), expansion (expa), migration (migr) and transition (tran), to quantitatively describe tissue distribution, clonal expansion, migration and developmental transition or

differentiation, respectively, which are essential for anti-tumour immunity by T cells (Extended Data Fig. 1b, Methods).

We obtained transcriptome data for 11,138 single T cells from 12 patients with CRC, including 4 MSI and 8 MSS patients (Extended Data Figs. 1a, 2a, Supplementary Table 1). Genomic alterations of these tumours were consistent with the characteristics of CRC from The Cancer Genome Atlas (TCGA)¹² (Extended Data Fig. 2b, Supplementary Table 2). CD8⁺ T cells, CD4⁺CD25^{-int} T_H cells and CD4⁺CD25^{hi} T_{reg} cells were assessed by multi-colour immunohistochemistry (IHC) (Extended Data Fig. 1c) and collected by fluorescence-activated cell sorting (FACS) before deep single-cell RNA-seq analyses (Extended Data Fig. 1a, d, Methods). Overall, we obtained an average of 1.25 million uniquely mapped read pairs (Extended Data Fig. 3a, b, Supplementary Table 3). After a series of quality control filtering, 10,805 cells remained—of which 91.4% had at least one pair of full-length productive α and β chains (Extended Data Fig. 3c, d, Supplementary Table 4). There were 7,274 clonotypes, each of which had unique productive α–β chain pairs and out of which 870 were represented by two or more cells that resulted in 3,474 clonal T cells.

A total of 8 CD8⁺ and 12 CD4⁺ T cell clusters were identified, each exhibiting a distinct distribution of clonotypes and clonal T cells (Fig. 1a, Extended Data Fig. 3e). The stability of clusters was supported by different clustering methods, down-sampling analysis (Extended Data Fig. 3f, g) and distinct signature genes (Extended Data Fig. 4a–c, Supplementary Table 5). In addition to typical CD8⁺ and CD4⁺ T cell clusters including naïve (T_N), central memory (T_{CM}) and effector memory (T_{EM}) T cells, recently activated effector memory or effector T cells (TEMRA/T_{EFF}, designated T_{EMRA} hereafter), mucosal-associated invariant T (MAIT) cells, blood-T_{reg} cells, tumour-T_{reg} cells, and dysfunctional or ‘exhausted’ CD8⁺ T (T_{EX}) cells, we also identified two IFNG⁺ T_H1-like cell clusters, with the CD4_C07-GZMK cluster expressing several markers of T_{EM} cells (designated as CD4⁺ TEM cells) and the CD4_C09-CXCL13 cluster showing higher expression of CXCL13 and BHLHE40 (designated as CXCL13⁺BHLHE40⁺ T_H1-like cells; Extended Data Fig. 5a). In contrast to the hepatocellular carcinoma (HCC) and non-small-cell lung carcinoma (NSCLC) datasets^{9,10}, we found additional T cell subsets, including T_H17 (CD4_C08-IL23R), follicular T helper cells (CD4_C06-CXCR5), follicular T regulatory cells (CD4_C11-IL10), and two additional subsets of CD8⁺ T cells (CD8_C05-CD6 and CD8_C06-CD160). The latter two highly expressed CD69 and ITGAE, known markers of tissue-resident memory T (T_{RM}) cells¹³ (Extended Data Fig. 5b). Whereas CD8_C05-CD6 probably represented lamina propria T_{RM} cells¹³, CD8_C06-CD160 was characterized as intraepithelial lymphocytes (IELs) based on the highly expressed natural killer cell markers¹⁴.

Our STARTRAC-dist index revealed distinct patterns of tissue distribution of different T cells (Fig. 1b, Extended Data Fig. 5c, d). Within the CD8⁺ subtypes, T_N, T_{CM} and T_{EMRA} cells were predominantly enriched

¹Beijing Advanced Innovation Centre for Genomics, Peking-Tsinghua Centre for Life Sciences, Peking University, Beijing, China. ²Department of Inflammation and Oncology, Discovery Research, Amgen, South San Francisco, CA, USA. ³BIOPIIC and School of Life Sciences, Peking University, Beijing, China. ⁴Department of Gastroenterological Surgery, Peking University People's Hospital, Beijing, China. ⁵Department of Pathology, Peking University People's Hospital, Beijing, China. ⁶These authors contributed equally: Lei Zhang, Xin Yu, Liangtao Zheng, Yuanyuan Zhang.
*e-mail: shenlong1977@163.com; wouyang@amgen.com; zemin@pku.edu.cn

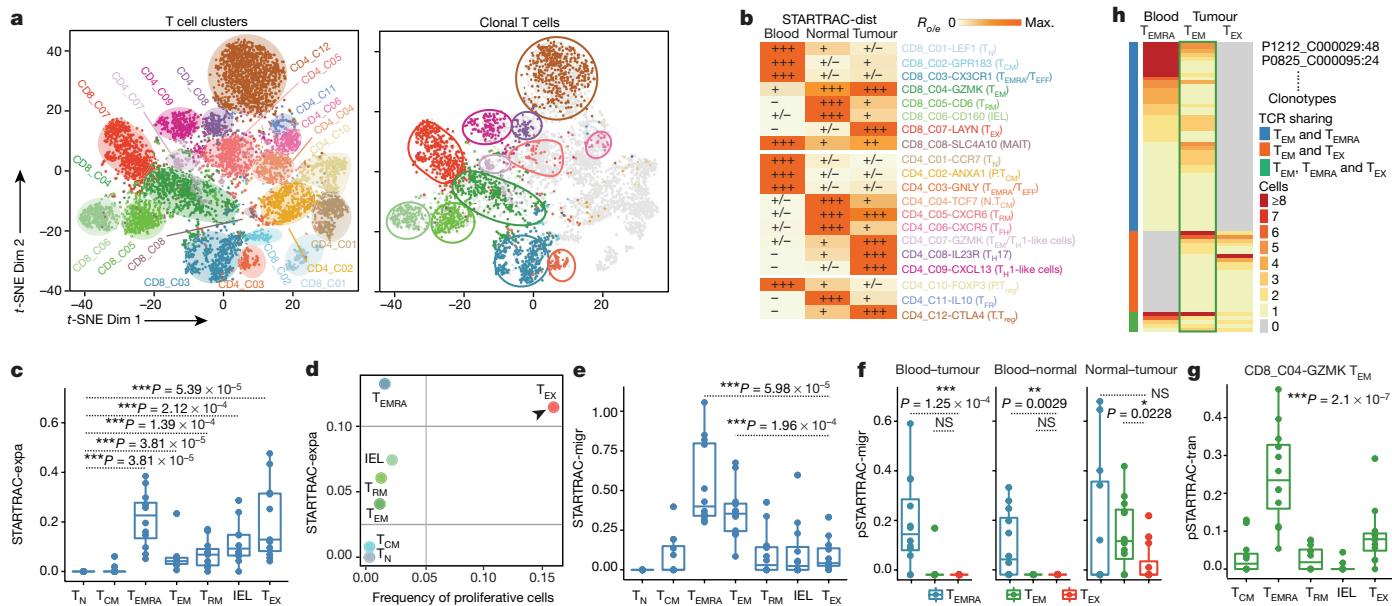


Fig. 1 | Properties of CD8⁺ T cell clonal expansion, migration and developmental transition. **a**, Left, *t*-distributed stochastic neighbour embedding (*t*-SNE) plot of 8,530 T cells from 12 patients with CRC showing 20 major clusters (8 for 3,628 CD8⁺ and 12 for 4,902 CD4⁺ T cells; functional interpretations in Extended Data Fig. 5d). Right, highlighted clonal T cells ($n = 3,065$). Each dot denotes an individual T cell; colour denotes cluster origin. **b**, Tissue preference of each cluster estimated by the STARTRAC-dist index. +++, $R_{o/e} > 1$; ++, $0.8 < R_{o/e} \leq 1$; +, $0.2 \leq R_{o/e} \leq 0.8$; +/-, $0 < R_{o/e} < 0.2$; -, $R_{o/e} = 0$; in which $R_{o/e}$ denotes the ratio of observed to expected cell number. N, normal tissue; P, peripheral blood; T, tumour. **c**, Clonal expansion levels of CD8⁺ clusters quantified by STARTRAC-expa for each patient ($n = 12$). **d**, Frequencies of proliferative CD8⁺ T cells (x axis) versus the STARTRAC-expa index

in blood (Fig. 1b). T_{EX} cells were specifically enriched in tumours, whereas two subsets of T_{RM} cells were predominantly found in normal mucosa. Likewise, among CD4⁺ subtypes, naive and effector-like cells were enriched in blood. Follicular T helper cells were enriched in normal mucosa, whereas two IFNG⁺ T_{H1}-cell-like subsets and T_{H17} cells were enriched in tumours. Three FOXP3⁺ T_{reg} cell clusters, CD4_{C10}-FOXP3, CD4_{C11}-IL10 and CD4_{C12}-CTLA4, were enriched in blood, normal mucosa and tumours, respectively.

Focusing on CD8⁺ T cells, the STARTRAC-expa index revealed the CD8⁺ T_{EX} and T_{EMRA} cells as the clusters with the highest degree of clonal expansion, followed by IELs (Fig. 1c). T_{EX} cells in CRC contained the highest percentage of proliferative cells (Fig. 1d), and enriched with MKI67hi cells and proliferation-related pathways, as confirmed by IHC (Extended Data Fig. 6a–c), although these high-proliferative cells resembled the low-proliferative cells with respect to the expression of key genes, including TBX21, EOMES and PDCD1¹⁵ (Extended Data Fig. 6d, e). Consistently, although 79.26% of high-proliferative cells were clonal, most of their clonotypes (40 out of 46) were shared with low-proliferative cells (Extended Data Fig. 6f); this suggests that proliferative status was not a distinguishing feature, as was observed between progenitor and terminally differentiated exhausted states from chronic infections¹⁵. Furthermore, although ex vivo reactivation experiments have demonstrated that these CD8⁺ T_{EX} cells produce less effector cytokines¹⁶, our analyses revealed that these cells expressed higher levels of effector molecules—such as IFNG, GZMB, GZMH and PRF1—than other CD8⁺ subsets (Extended Data Fig. 6g); this indicates that T_{EX} cells may not have completely lost their anti-tumour effector potential in vivo. Several transcription factors were preferentially expressed in the CD8⁺ T_{EX} cell subset. Although PRDM1 and BATF were the only previously known factors¹⁷, RBPJ, TOX and BHLHE40 were functionally uncharacterized in T_{EX} cells (Extended Data Fig. 6g).

(y axis). **e**, Migration potentials of CD8⁺ T cell clusters quantified by overall STARTRAC-migr indices for each patient ($n = 12$). **f**, Comparison of migration potentials of CD8⁺ T_{EMRA}, T_{EM} and T_{EX} cells by pairwise STARTRAC-migr (pSTARTRAC-migr) indices for each patient ($n = 12$). **g**, Developmental transition of CD8⁺ T_{EM} cells with other CD8⁺ cells quantified by pairwise STARTRAC-tran indices for each patient ($n = 12$). *** $P < 0.001$, Kruskal–Wallis test. **h**, The distribution of clonal clonotypes in indicated CD8⁺ subsets. Tumour T_{EM} cells showing mutually exclusive TCR sharing with blood T_{EMRA} and tumour T_{EX} cells (Extended Data Fig. 8e). NS, not significant. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-sided Wilcoxon test (**c**, **e** and **f**). For box plots in all figures, centre lines denote median values; whiskers denote $1.5 \times$ the interquartile range; coloured dots denote outliers.

The STARTRAC-migr analysis of CD8⁺ clusters revealed that T_{EMRA} cells were associated with the highest mobility, followed by T_{EM} cells, and both clusters had higher mobility than T_{EX} cells (Fig. 1e, Extended Data Fig. 7a). Furthermore, pairwise STARTRAC-migr analyses revealed a high degree of TCR sharing of T_{EMRA} cells among blood, normal mucosa and tumours, whereas T_{EX} exhibited tumour exclusivity (Fig. 1f). Accordingly, these clusters expressed different sets of genes related to migration¹⁸, including chemokine receptors, integrin and other trafficking-related molecules such as S1P receptors (Extended Data Fig. 7b). For example, T_{EMRA} cells highly expressed S1PR1, S1PR5 and ITGB7, which supports their capability to circulate in the periphery and home to normal mucosa and tumours.

Although T_{EMRA} cells also expressed many effector molecules such as PRF1, GZMB and GZMH, they did not express T_{EX} cell markers such as PDCD1 and HAVCR2 (also known as TIM3) (Extended Data Fig. 6g). Notably, STARTRAC-tran analysis indicated that both T_{EMRA} and T_{EX} cell clusters were highly associated with T_{EM} cells (Fig. 1g, Extended Data Fig. 8a). T_{EX} cells were only linked to T_{EM} cells, but both T_{EMRA} and T_{EM} cells were associated with T_{CM} cells (Extended Data Fig. 8b). Furthermore, T_{EM} cells were also moderately connected to normal-enriched T_{RM} cells. Accordingly, Monocle trajectory analysis of these CD8⁺ cell clusters also corroborated a developmental trajectory from T_{EM} cells to either T_{EMRA} or T_{EX} cells (Extended Data Fig. 8c).

The transition from T_{EM} to T_{EX} cells may predominantly occur in tumours based on their tissue distribution patterns (Fig. 1b). Although only 19.35% (24 out of 124) of T_{EMRA} cell-expanded clonotypes had clonal cells located inside tumours, 44.35% (55 out of 124) were linked to tumour-infiltrating T_{EM} cells, whereas only 5.65% (7 out of 124) were associated with blood T_{EM} cells (Extended Data Fig. 8d, Supplementary Table 6); this supports the developmental connection between T_{EMRA} cells and tumour T_{EM} cells. Notably, tumour T_{EM} cell clones linked to

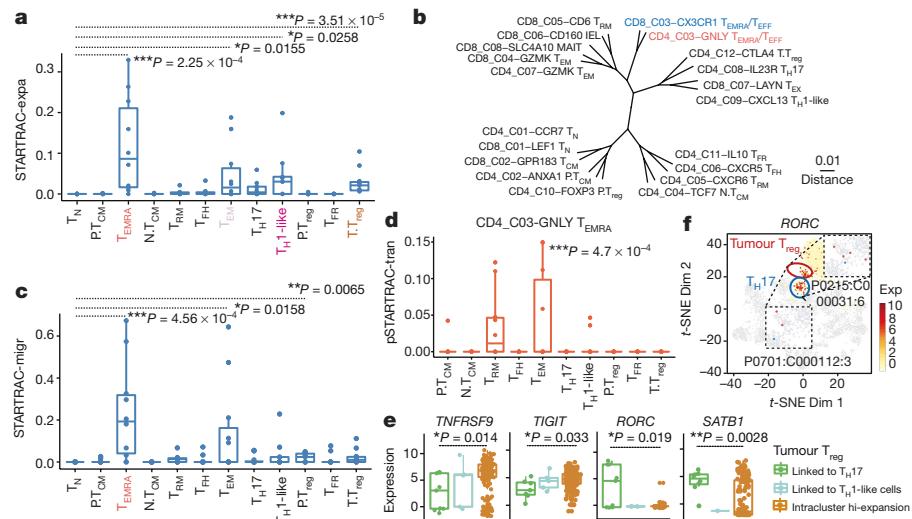


Fig. 2 | Properties of CD4⁺ T cell clonal expansion, migration and developmental transition. **a**, Clonal expansion levels of CD4⁺ T cell clusters quantified by STARTRAC-expa indices for each patient ($n = 11$). **b**, Similarities of the signature gene expressions of T cell clusters. Distance = (1 – Pearson correlation coefficient)/2. **c**, Migration potentials of CD4⁺ T cell clusters quantified by overall STARTRAC-migr indices for each patient ($n = 11$). **d**, Developmental transition of CD4⁺ T_{EMRA} cells with other CD4⁺ cells quantified by pSTARTRAC-tran indices for each patient ($n = 10$). *** $P < 0.001$, Kruskal–Wallis test. **e**, Normalized

blood T_{EMRA} cells were mutually exclusive with those linked to T_{EX} cells (Fig. 1h, Extended Data Fig. 8e). This pattern was also confirmed in individual patients (Extended Data Fig. 8f). Thus, TCR clonotypes may have a role in determining the developmental trajectories between T_{EM} and T_{EX} cells and between T_{EM} and T_{EMRA} cells.

In contrast to CD8⁺ cells, CD4⁺ cells exhibited lower clonal expansion overall. Among all CD4⁺ clusters, CD4_C03-GNLY exhibited the

expression of a series of genes in three tumour T_{reg} cell subpopulations that shared TCRs with T_H17 cells ($n = 9$ cells), CXCL13⁺BHLHE40⁺ T_H1-like cells ($n = 5$) and exclusive to their own respective group ($n = 228$).

* $P < 0.05$, ** $P < 0.01$, two-sided Student's *t*-test. **f**, Representative clonotypes of tumour T_{reg} cells ($n = 1,320$) with high expression of RORC (coloured dots in ellipse) and shared TCRs with T_H17 cells ($n = 244$). Red denotes T_{reg} cells; blue denotes T_H17 cells. Exp, centred normalized expression. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-sided Wilcoxon test (a, c).

highest clonal expansion (Fig. 2a). This cluster contained signature genes that include NKG7, PRF1, GNLY and GZMH—a signature that is profoundly similar to that of CD8⁺ T_{EMRA} cells (Fig. 2b, Extended Data Fig. 9a); CD4_C03-GNLY cells were therefore designated as CD4⁺ T_{EMRA} cells. Notably, the CD4_C03-GNLY cluster showed migration properties that were comparably high to those of CD8⁺ T_{EMRA} cells (Fig. 2c), and it was linked to the CD4⁺ T_{EM} population (Fig. 2d).

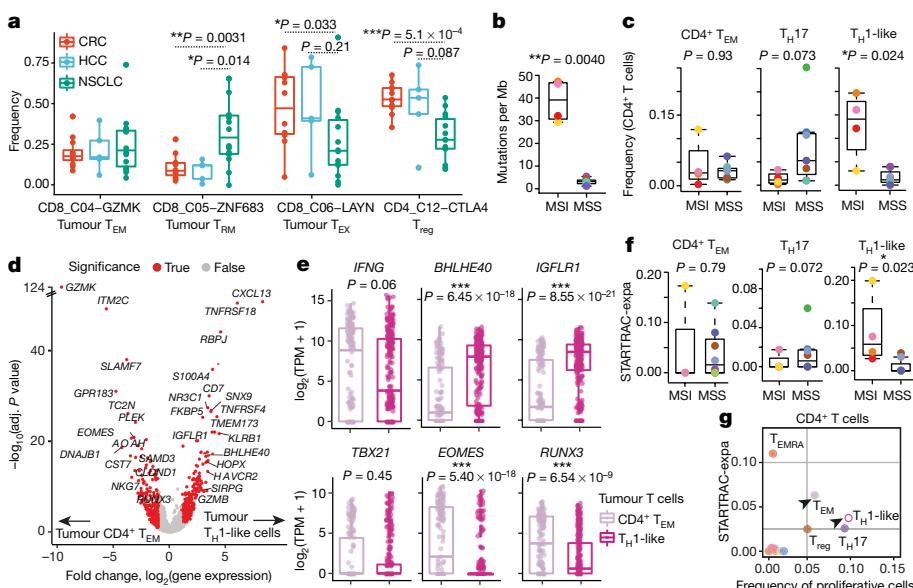


Fig. 3 | Clonal T_H1-like T cells are enriched in MSI tumours. **a**, Comparison of the proportions of different CD8⁺ and CD4⁺ T cell clusters in tumours from patients with CRC ($n = 12$), HCC⁹ ($n = 5$), and NSCLC¹⁰ ($n = 14$) after re-clustering of the combined dataset (Extended Data Fig. 10). Each dot denotes an individual patient. **b**, Box plot showing mutation load of patients with MSI ($n = 4$) and MSS ($n = 8$) CRC. **c**, Percentages of tumour-enriched T_H cells in the overall CD4⁺ T cells from patients with MSI ($n = 4$) and MSS ($n = 7$) CRC. **d**, Volcano plot showing differentially expressed genes between CXCL13⁺BHLHE40⁺ T_H1-like cells

($n = 315$) and CD4⁺ T_{EM} cells ($n = 161$) in tumours. $P < 0.01$, Benjamini–Hochberg adjusted two-sided unpaired limma-modulated *t*-test; fold change ≥ 2 . **e**, Gene expression comparison between CXCL13⁺BHLHE40⁺ T_H1-like cells ($n = 315$) and CD4⁺ T_{EM} cells ($n = 161$). **f**, STARTRAC-expa indices for tumour-enriched T_H cells in patients with MSI ($n = 4$) and MSS ($n = 7$) CRC. **g**, Frequencies of proliferative cells versus STARTRAC-expa index for CD4⁺ cells. Each dot in **b**, **c** and **f** represents an individual patient. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-sided Wilcoxon test (a–c, e, f). TPM, transcripts per million.

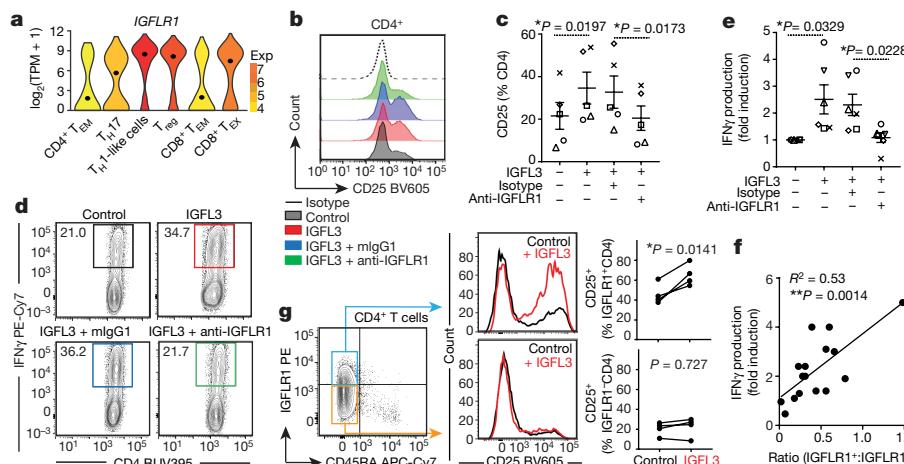


Fig. 4 | IGFLR1 functions as a co-stimulatory receptor in T cells.

a, Violin plots showing *IGFLR1* expression in tumour-enriched T cell clusters, including CD4⁺ T_{EM} cells ($n = 185$), T_{H17} cells ($n = 244$), T_{H1}-like cells ($n = 319$), T_{reg} cells ($n = 1,320$), CD8⁺ T_{EM} cells ($n = 773$) and CD8⁺ T_{EX} cells ($n = 860$). Black dots denote mean values; widths denote cell densities. b, Representative histograms of CD25 expression in activated and indicated CD4⁺ T cells on day 2 ($n = 5$ donors, $n = 3$ independent experiments). mIgG1 denotes mouse anti-IgG1 control antibody. c, Quantification of CD25 expression in b. Symbols are individual donors. Data are mean \pm s.e.m. ($n = 5$; c, e). d, Flow cytometry

of intracellular IFN γ expression of activated CD4⁺ T cells on day 2 ($n = 6$ donors, $n = 3$ independent experiments). e, Fold changes in IFN γ supernatant from d measured by ELISA relative to the control group. f, Correlation of IFN γ induction by IGFL3 and the ratio of IGFLR1⁺ to IGFLR1⁻ CD4⁺ T_{MEM} cells ($n = 16$ donors; F-test). g, Activated CD4⁺ T_{MEM} cells were gated as IGFLR1⁺ and IGFLR1⁻ cells by FACS (left). Representative histograms (middle) and percentages (right) of CD25 expression levels in IGFLR1⁺ and IGFLR1⁻ CD4⁺ T_{MEM} cells ($n = 4$ donors). Two-sided paired Student's *t*-test (c, e and g).

The potential tumour-killing activities of the CD4_C03-GNLY cluster should therefore be further explored.

Tumour-infiltrating T_{reg} cells were among the highly expanded populations (Fig. 2a). Most (88%) clonal CRC-infiltrating T_{reg} cells contained TCR clonotypes exclusive to themselves, indicating their potential for recognizing tumour-associated antigens and local expansion characteristics^{9,10} (Extended Data Fig. 9b). A few tumour T_{reg} cells also shared TCRs with T_{reg} cells from blood (CD4_C10-FOXP3) or normal mucosa (CD4_C11-IL10). Several other tumour T_{reg} cells shared TCRs with T_H cells in tumours (Extended Data Fig. 9b, Supplementary Table 7), indicating that they were induced T_{reg} (iT_{reg}) cells¹⁹. STARTRAC-tran analysis suggested that these potential iT_{reg} cells were developmentally linked to T_{H17} and CXCL13⁺BHLHE40⁺ T_{H1}-like cells (Extended Data Fig. 9c). iT_{reg} cells that share TCRs with T_{H17} cells fell into a sub-cluster of a tumour T_{reg} cell group with higher expression of ROR γ (Fig. 2e, f). The co-expression of ROR γ and FOXP3 at the protein level was confirmed by IHC (Extended Data Fig. 9d). In addition, SATB1²¹ was selectively expressed in T_{reg} cells linked to T_{H17} cells, whereas BACH2²² was preferentially expressed in T_{reg} cells linked to CXCL13⁺BHLHE40⁺ T_{H1}-like cells (Fig. 2e, Supplementary Table 8). By contrast, those T_{reg} cells with high intra-cluster expansion had relatively high expression of TNFRSF9 and TIGIT (Fig. 2e), suggesting that at least some of these cells might belong to natural T_{reg} cells²³. The roles of these different subsets of T_{reg} cells needs further investigation.

When comparing T cell populations across cancer types^{9,10} (Extended Data Fig. 10a, b), we found that although the composition and abundance of blood-derived T cells were highly similar, T cell patterns were distinct in both tumours and adjacent normal tissues (Extended Data Fig. 10c-f). Notably, CRC and HCC tumours exhibited a higher abundance of CD8⁺ T_{EX} and CD4⁺ T_{reg} cells, whereas NSCLC tumours exhibited enrichment of tumour T_{RM} cells with low expression of PDCD1 and CTLA4 but high expression of ZNF683 (Fig. 3a). Likewise, the IELs were specifically present in normal mucosa and tumours in patients with CRC (Extended Data Fig. 10e, f).

Next, by focusing on the differences between heavily mutated MSI tumours (Fig. 3b) and MSS tumours, we found that MSI tumours exhibited abundant CXCL13⁺BHLHE40⁺ T_{H1}-like cells, whereas MSS tumours were moderately enriched with T_{H17} cells (Fig. 3c). Accordingly, the CXCL13⁺BHLHE40⁺ T_{H1}-like cell signal was

increased in the TCGA MSI-high patients with CRC, whereas the T_{H17} signal was enriched in the MSS cohort (Extended Data Fig. 10g). Although increased overall IFNG⁺ T_{H1} cells has been suggested in MSI CRC tumours^{6,7}, of the two IFNG⁺ T_{H1}-like cell subsets we identified, only the CXCL13⁺BHLHE40⁺ T_{H1}-like cell cluster and not another GZMK⁺IFNG⁺ T_{EM} cell cluster was enriched in MSI tumours. Notably, although TBX21 showed similar expression in both subsets, other IFN γ -regulating transcription factors EOMES and RUNX3²⁴ were preferentially expressed in GZMK⁺ T_{EM} cells, whereas BHLHE40 was selectively expressed in CXCL13⁺BHLHE40⁺ T_{H1}-like cells (Fig. 3d, e, Supplementary Table 9); this suggests distinctive transcriptional control for these two IFNG⁺ subsets. Indeed, BHLHE40 not only regulates IFN γ but also represses IL-10 production^{25,26}. Furthermore, STARTRAC analyses revealed that CXCL13⁺BHLHE40⁺ T_{H1}-like cells were clonally expanded and enriched in MSI patients, and proliferative in tumours (Fig. 3f, g). Moreover, these cells exhibited developmental connection with GZMK⁺ T_{EM} cells, indicating the potential inter-conversion of these two IFNG⁺ subsets (Extended Data Fig. 9c). Given the involvement of T_{H1}-like cells in response to anti-CTLA4 therapy in melanoma²⁷, we speculate that the enrichment of CXCL13⁺BHLHE40⁺IFNG⁺ T_{H1}-like cells in MSI patients might contribute to their favourable response to immunotherapies.

These CXCL13⁺BHLHE40⁺ T_{H1}-like cells in tumours showed high expression of CXCR3, HAVCR2, PDCD1, ICOS and TIGIT (Extended Data Fig. 11a, Supplementary Table 10). Notably, a less-characterized gene, *IGFLR1*, was also upregulated in these cells. *IGFLR1* belongs to the TNFR superfamily²⁸ and has three potential ligands, with IGFL1 and IGFL3 showing high-affinity interactions. *IGFLR1* was also upregulated in tumour T_{EX} cells (Extended Data Fig. 11b, Supplementary Table 11) and T_{reg} cells (Fig. 4a). In vitro, suboptimal T cell activation was sufficient to upregulate *IGFLR1* in memory CD4⁺ T cells (Extended Data Fig. 11c-e). Importantly, IGFL3 enhanced CD25 induction and IFN γ secretion in CD4⁺ T cells under this condition (Fig. 4b-e). The degree of IFN γ induction in total CD4⁺ T cells was correlated with the induction of *IGFLR1* in memory T cells (Fig. 4f). In addition, the IGFL3-augmented CD25 expression was observed only in IGFLR1⁺ memory T cells, but not in the IGFLR1⁻ cells in the same cultures (Fig. 4g). Furthermore, the synergistic effect by IGFL3 could be specifically blocked by an anti-IGFLR1 antibody (Fig. 4c-e).

To study the role of IGFLR1 in exhausted CD8⁺ T cells, we adopted a protocol for low-degree stimulation of chronic stimulation CD8⁺ T (T_{CS}) cells (see Methods) to induce certain features of in vivo exhausted T cells (Extended Data Fig. 11f–h). These CD8⁺ T_{CS} cells exhibited higher IGFLR1 expression than activated conventional CD8⁺ effector T (T_{conv}) cells (Extended Data Fig. 11i, j). The addition of IGFL3 synergized the TCR-induced HAVCR2 expression²⁹ in these cells, which could be blocked by an anti-IGFLR1 antibody (Extended Data Fig. 11k, l). Altogether, these data suggest that IGFLR1 could synergize with TCR signalling and serve as a co-stimulatory molecule.

In summary, STARTRAC analyses unveiled various functional, migratory and developmental connections among different T cell subsets in CRC. Our data and previous findings in mouse models³⁰ revealed TCR-dependent trajectories for T_{EMRA} and T_{EX} cells from tumour-resident CD8⁺ T_{EM} cells, suggesting therapeutic strategies to promote the conversion from T_{EX} to T_{EMRA} cells. The enrichment of CXCL13⁺BHLHE40⁺IFNG⁺ Th1-like cells in MSI patients not only provides a rationale for the high response rate to checkpoint blockade in these patients, but also solicits therapeutic focus on these cells. The compendium dataset of differentially expressed genes such as *IGFLR1*, available through an interactive portal at <http://crc.cancer-pku.cn>, can be used as a resource for further T cell exploration and the identification of regulatory mechanisms and therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0694-x>.

Received: 26 February 2018; Accepted: 19 October 2018;

Published online 29 October 2018.

- Chen, D. S. & Mellman, I. Elements of cancer immunity and the cancer-immune set point. *Nature* **541**, 321–330 (2017).
- Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
- Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
- Le, D. T. et al. PD-1 Blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
- Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
- Llosa, N. J. et al. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer Discov.* **5**, 43–51 (2015).
- Mlecnik, B. et al. Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability. *Immunity* **44**, 698–711 (2016).
- Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
- Zheng, C. et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. *Cell* **169**, 1342–1356.e16 (2017).
- Guo, X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. *Nat. Med.* **24**, 978–985 (2018).
- Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
- The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
- Schenkel, J. M. & Masopust, D. Tissue-resident memory T cells. *Immunity* **41**, 886–897 (2014).
- Cheroutre, H., Lambolez, F. & Mucida, D. The light and dark sides of intestinal intraepithelial lymphocytes. *Nat. Rev. Immunol.* **11**, 445–456 (2011).

- Paley, M. A. et al. Progenitor and terminal subsets of CD8⁺ T cells cooperate to contain chronic viral infection. *Science* **338**, 1220–1225 (2012).
- Thommen, D. S. et al. A transcriptionally and functionally distinct PD-1⁺ CD8⁺ T cell pool with predictive potential in non-small-cell lung cancer treated with PD-1 blockade. *Nat. Med.* **24**, 994–1004 (2018).
- Chihara, N. et al. Induction and transcriptional regulation of the co-inhibitory gene module in T cells. *Nature* **558**, 454–459 (2018).
- Griffith, J. W., Sokol, C. L. & Luster, A. D. Chemokines and chemokine receptors: positioning cells for host defense and immunity. *Annu. Rev. Immunol.* **32**, 659–702 (2014).
- Adeegbe, D. O. & Nishikawa, H. Natural and induced T regulatory cells in cancer. *Front. Immunol.* **4**, 190 (2013).
- Sefik, E. et al. Individual intestinal symbionts induce a distinct population of ROR γ ⁺ regulatory T cells. *Science* **349**, 993–997 (2015).
- Beyer, M. et al. Repression of the genome organizer SATB1 in regulatory T cells is required for suppressive function and inhibition of effector differentiation. *Nat. Immunol.* **12**, 898–907 (2011).
- Roychoudhuri, R. et al. BACH2 represses effector programs to stabilize T_{reg}-mediated immune homeostasis. *Nature* **498**, 506–510 (2013).
- Zhang, Y. et al. Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells. *Blood* **122**, 2823–2836 (2013).
- Djuretic, I. M. et al. Transcription factors T-bet and Runx3 cooperate to activate *Ifng* and silence *Iif4* in T helper type 1 cells. *Nat. Immunol.* **8**, 145–153 (2007).
- Yu, F. et al. The transcription factor Blhhe40 is a switch of inflammatory versus antiinflammatory Th1 cell fate determination. *J. Exp. Med.* **215**, 1813–1821 (2018).
- Huynh, J. P. et al. Blhhe40 is an essential repressor of IL-10 during *Mycobacterium tuberculosis* infection. *J. Exp. Med.* **215**, 1823–1838 (2018).
- Wei, S. C. et al. Distinct cellular mechanisms underlie anti-CTLA-4 and anti-PD-1 checkpoint blockade. *Cell* **170**, 1120–1133.e17 (2017).
- Lobito, A. A. et al. Murine insulin growth factor-like (IGFL) and human IGFL proteins are induced in inflammatory skin conditions and bind to a novel tumor necrosis factor receptor family member, IGFLR1. *J. Biol. Chem.* **286**, 18969–18981 (2011).
- Singer, M. et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. *Cell* **166**, 1500–1511.e9 (2016).
- Schiettlinger, A. et al. Tumor-specific T cell dysfunction is a dynamic antigen-driven differentiation program initiated early during tumorigenesis. *Immunity* **45**, 389–401 (2016).

Acknowledgements We thank C. X. Ye for sample preparation and F. Wang, X. Zhang and J. S. Li for assistance with FACS. We thank the Computing Platform of the CLS (Peking University). This project was supported by Beijing Advanced Innovation Centre for Genomics at Peking University, Key Technologies R&D Program (2016YFC0900100), National Natural Science Foundation of China (81573022, 31530036, 91742203 and 81672375) and Amgen Corporation (USA). L.Z. was supported by the Postdoctoral Foundation of CLS.

Reviewer information *Nature* thanks N. Haining, M. Suva and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.Z., W.O. and L.Z. designed experiments. L.Z., X.Y., Q.F., R.G., O.Z., J.Y.H., H.K. and X.G. performed the experiments. L.Z., L.T.Z., Y.Z., X.R. and X.H. analysed sequencing data. Y.L., J.Y., S.W., Y.G. and Z.S. collected clinical samples. B.K. constructed the website. L.Z., W.O. and Z.Z. wrote the manuscript with input from all authors.

Competing interests W.O., X.Y., H.K. and J.Y.H. are employees of Amgen Inc.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0694-x>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0694-x>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to Z.S., W.O. or Z.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Human specimens. Twelve patients with CRC, including eight women and four men, were enrolled and pathologically diagnosed with colorectal adenocarcinoma at Peking University People's Hospital. Written informed consent was provided by all patients. This study was approved by the Research and Ethical Committee of Peking University People's Hospital and complied with all relevant ethical regulations. Fresh tumour and adjacent normal tissue samples (at least 2 cm from matched tumour tissues) were surgically resected from the above-described patients. Patients P0701, P0909, P1212, P1228, P0215, P0411, P0413, P0825, P0123 and P0309 had peripheral blood and paired tumour and adjacent normal tissues obtained, whereas patients P1012 and P1207 had only fresh tumour tissue and matched peripheral blood. Their ages ranged from 35 to 82 with a median of 67. None of them was treated with chemotherapy or radiation before tumour resection. The stages of these patients were classified according to the guidance of AJCC version 8. Among these patients, one was diagnosed at stage I, five at stage II, five at stage III, and one at stage IV. Among the four MSI-high (MSI-H) patients, three had positive lymph nodes (P0123, P0413 and P0909), and two had poorly differentiated disease (P0825 and P0909). Although we did not purposely exclude the stage IV patient, none of the MSI-H patients had distal metastasis, as evidenced by the enhanced computerized tomography (CT) results for abdomen, chest and pelvic areas before surgery. The available clinical characteristics are summarized in Supplementary Table 1. For the IGFLR1 study, **human peripheral blood mononuclear cells (PBMCs)** were obtained from 16 healthy donors after informed consent and authorization by the Amgen Research Blood Donor Program.

Single-cell collection. Tumours and adjacent normal tissues were cut into approximately 1-mm³ pieces in the RPMI-1640 medium (Invitrogen) with 10% fetal bovine serum (FBS; ScienCell), and enzymatically digested with MACS Tumour Dissociation Kit (Miltenyi Biotec) for 30 min on a rotor at 37°C, according to the manufacturer's instruction. The dissociated cells were subsequently passed through a 40-μm cell-strainer (BD) and centrifuged at 400g for 10 min. After the supernatant was removed, the pelleted cells were suspended in red blood cell lysis buffer (Solarbio) and incubated on ice for 2 min to lyse red blood cells. After washing twice with PBS (Invitrogen), the cell pellets were re-suspended in sorting buffer (PBS supplemented with 1% FBS).

PBMCs were isolated using HISTOPAQUE-1077 (Sigma-Aldrich) solution as previously described⁹. In brief, 3 ml of fresh peripheral blood was collected before surgery in EDTA anticoagulant tubes and subsequently layered onto HISTOPAQUE-1077. After centrifugation, lymphocyte cells remained at the plasma-HISTOPAQUE-1077 interface and were carefully transferred to a new tube and washed twice with PBS. Red blood cells were removed via the same procedure described above. These lymphocytes were re-suspended in sorting buffer.

Single-cell sorting, reverse transcription, amplification and sequencing. Single-cell suspensions were stained with antibodies against CD3, CD4, CD8 and CD25 (anti-human CD3, UCHT1; anti-human CD4, OKT4; anti-human CD8, OKT8; anti-human CD25, BC96; eBioscience) for FACS sorting, performed on a BD Aria III instrument. Single cells of different subtypes including cytotoxic T (T_c) cells, T_H cells and T_{reg} cells were enriched by gating 7AAD⁻CD3CD8⁺, 7AAD⁻CD3⁻CD4⁺CD25^{-/int} and 7AAD⁻CD3⁻CD4⁺CD25^{hi} T cells, respectively, and sorted into 96-well plates (Axygen) chilled to 4°C, prepared with lysis buffer with 1 μl 10 mM dNTP mix (Invitrogen), 1 μl 10 μM Oligo dT primer, 1.9 μl 1% Triton X-100 (Sigma), and 0.1 μl 40 U μl⁻¹ RNase Inhibitor (Takara).

The single-cell lysates were sealed and stored frozen at -80°C immediately. Single-cell transcriptome amplifications were performed according to the Smart-Seq2 protocol^{9,31}. The External RNA Controls Consortium (ERCC; Ambion; 1:4,000,000) was added into each well as the exogenous spike-in control before the reverse transcription. The amplified cDNA products were purified with 1 × Agencourt XP DNA beads (Beckman). A procedure of quality control was performed following the first round of purification, which included the detection of CD3D by qPCR (forward primer, 5'-TCATTGCCACTCTGCTCC-3'; reverse primer, 5'-GGTCACCTGTTCCGAGCC-3') and fragment analysis by analyser (AATI). For those single-cell samples with high quality after quality control (cycle threshold <30), the DNA products were further purified with 0.5 × Agencourt XP DNA beads, and the concentration of each sample was quantified by Qubit HsDNA kits (Invitrogen). Multiplex (384-plex) libraries were constructed and amplified using the TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme Biotech). The libraries were then purified with Agencourt XP DNA beads and pooled for quality assessment by fragment analyser. For all the 12 patients, purified libraries were analysed by an Illumina Hiseq 4000 sequencer with 150-bp pair-end reads. For patient P1207, only CD8⁺ T cells were collected, thus this patient was not included when analysing CD4⁺ T cells.

Primary human T cell isolation and in vitro activation. PBMCs from healthy donors were isolated by density gravity centrifugation (Ficoll-Paque PREMIUM; GE Healthcare). T cell subsets were isolated from PBMCs with appropriate magnetic beads following the manufacturer's protocol (StemCell Technologies). Isolated T cells were cultured in RPMI-1640 supplemented with 10% heat-inactivated FBS, 100 U ml⁻¹ penicillin and streptomycin and 2-mercaptoethanol (all from Gibco). Human T cells were activated with anti-CD3 (UCHT1, 2 μg ml⁻¹, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml⁻¹, BD Biosciences). When indicated, 100 ng ml⁻¹ recombinant human IGFL3 (Adipogen) was added in culture medium alone or with 20 μg ml⁻¹ of anti-IGFLR1 blocking antibody (clone 905338, R&D Systems). The same amount of mouse IgG1 antibody (R&D Systems) was used as a control. Cells were collected and stained with indicated monoclonal antibodies (anti-human CD4, OKT4; anti-human CD8, RPA-T8; anti-human CD45RA, HI100; anti-human CCR7, G043H7; anti-human HAVCR2, F38-2E2; anti-human CD25, M-A251; anti-human IFNγ, B27; anti-human IGFLR1, 905338; Biolegend, BD Biosciences, or R&D Systems). Flow cytometry data were acquired with an LSR-II analyser and analysed using FlowJo software (Tree Star). Cytokine concentrations were measured in cell culture supernatants 40 h after stimulation with enzyme-linked immunosorbent assay (ELISA) kits specific for human IFNγ (BD Bioscience). All data were from three independent experiments with more than four donors.

In vitro CD8⁺ T_{Cs} cells. The CD8⁺ T_{Cs} cells were generated using an in vitro chronic low-degree stimulation protocol as described in previous studies^{32,33}. Purified human CD8⁺ T cells at 1 × 10⁶–2 × 10⁶ cells ml⁻¹ were subjected to anti-CD3 (UCHT1, 0.2 μg ml⁻¹, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml⁻¹, BD Biosciences) stimulation for 3–4 days followed by at least an additional two rounds of re-stimulation every 3–4 days with anti-CD3 (UCHT1, 1 μg ml⁻¹, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml⁻¹, BD Biosciences) to generate CD8⁺ T_{Cs} cells. CD8⁺ T_{Cs} cells were then subjected to stimulation with or without human IGFL3 as described above. Cells were stained with indicated monoclonal antibodies (anti-human CD8, RPA-T8; anti-human HAVCR2, F38-2E2; anti-human PD-1, EH12.2H7; anti-human CD39, eBioA1; anti-LAG3, 305223H; anti-IGFLR1, 905338; Thermal Fisher Scientific, BD Biosciences, or R&D Systems).

T cell stimulation. Human memory CD4⁺ T cells were isolated from healthy donor PBMCs to a purity of >94% with memory CD4⁺ T cell isolation kit following manufacturer's protocol (Miltenyi Biotec). Isolated T cells were cultured in RPMI-1640 supplemented with 10% heat-inactivated FBS, 100 U ml⁻¹ penicillin and streptomycin and 2-mercaptoethanol (all from Gibco). Human T cells were activated with anti-CD3 (UCHT1, 2 μg ml⁻¹, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml⁻¹, BD Biosciences) for 2 days. T cells were then rested in culture medium for 2 days followed by 16 h starvation in RPMI-1640 plus 0.5% FBS. Next, T cells were collected and subjected to stimulation. For TCR stimulation, T cells were incubated on ice for 30 min with 1 μg ml⁻¹ anti-CD3 (OKT3, Thermo Fisher Scientific) and 1 μg ml⁻¹ anti-CD28 followed by a 15-min incubation with 5 μg ml⁻¹ anti-mouse IgG (Thermo Fisher Scientific). Cells were activated by incubation in a 37°C water bath for 25 min. For IGFL3 stimulation, T cells were incubated on ice for 30 min with 100 ng ml⁻¹ recombinant human IGFL3 (Adipogen). Cells were then activated by incubation in a 37°C water bath for 25 min.

Cytokine production detection. Human memory CD4⁺ T cells were isolated as described above. Cells were activated with anti-CD3 (UCHT1, 2 μg ml⁻¹, BD Biosciences) and anti-CD28 (CD28.2, 2 μg ml⁻¹, BD Biosciences) or anti-CD3 and anti-CD28 plus IGFL3 (100 ng ml⁻¹). Cells were seeded in flat-bottom 96-well plates with 10⁵ cells in 100 μl culture medium per 96 wells. Cytokine concentrations were measured in cell culture supernatants 40 h after stimulation with ELISA kits specific for human IFNγ (BD Bioscience).

Bulk DNA and RNA isolation and sequencing. Genomic DNA of peripheral blood and tissue samples of patients with CRC were extracted using the QIAamp DNA Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of DNA were quantified using the Qubit HsDNA Kits (Invitrogen) and the qualities of DNA were evaluated with agarose gel electrophoresis. Exon libraries were constructed using the SureSelectXT Human All Exon V5 capture library (Agilent). Samples were sequenced on the Illumina Hiseq 4000 sequencer with 150-bp paired-end reads. For bulk RNA analysis, small fragments of tumour tissues and adjacent normal tissues were first stored in RNAlater RNA stabilization reagent (QIAGEN) after surgical resection and kept on ice to avoid RNA degradation. RNA of tumour and adjacent normal tissue samples were extracted using the RNeasy Mini Kit (QIAGEN) according to the manufacturer's specification. The concentrations of RNA were quantified using the NanoDrop instrument (Thermo) and the qualities of RNA were evaluated with fragment analyser (AATI). Libraries were constructed using NEBNext Poly (A) mRNA Magnetic Isolation Module kit (NEB) and NEBNext Ultra RNA Library Prep Kit for Illumina Paired-end Multiplexed Sequencing Library (NEB). Samples were sequenced on the Illumina Hiseq 4000 sequencer with 150-bp paired-end reads.

Microsatellite instability testing. DNA purified from tumour tissues using QIAamp DNA Mini Kit (QIAGEN) was subjected to multiplex fluorescent PCR-based assay (Promega) by amplifying seven loci including five mononucleotide repeats (NR21, BAT26, BAT25, NR24 and Mono27) and two pentanucleotide repeats (PentaC and PentaD) and was compared with DNA extracted from matched adjacent normal tissues. Multiplex PCR products were analysed by ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). Patients were defined as MSI-H status by the presence of two or more mononucleotide loci showing instability. MSS was defined as no loci of instability. Among 12 patients in this study, 4 of them were MSI-H (P0413, P0825, P0123 and P0909), and the other 8 were MSS (P0215, P0411, P0701, P1012, P1207, P1212, P1228 and P0309).

Immunohistochemistry. The specimens were collected from Peking University People's Hospital within 30 min of the tumour resection and fixed in 10% formalin for 48 h. Dehydration and embedding in paraffin was performed as the following routine methods⁹. These paraffin blocks were cut into 5-μm sections and adhered to a glass slide. Then, the paraffin sections were placed in the 70 °C paraffin oven for 1 h before being deparaffinised in xylene and then rehydrated in 100%, 90% and 70% alcohol successively. The antigens were retrieved by the Epitope Retrieval Solution (Leica Biosystems), and the sections were incubated with ready-to-use primary antibodies (mouse anti-human MLH1, clone ES05; mouse anti-human MSH2, clone 25D12; mouse anti-human MSH6, clone PU29; mouse anti-human PMS2, clone M0R4G, all from Leica Biosystems) on the BOND system (Leica Biosystems) according to the manufacturer's protocol.

Multi-colour immunohistochemistry. The specimens were collected and prepared for the formalin-fixed paraffin-embedded tissues sections as previously mentioned⁹. The confirmation of ROR γ^+ T_{reg} cells was analysed using Opal 7-Colour Manual IHC Kit (PerkinElmer, NEL811001KT) according to the manufacturer's protocol, as previously described¹⁰. In brief, antigen was retrieved by AR9 buffer (pH 6.0, PerkinElmer) and boiled in the oven for 15 min. After a pre-incubation with blocking buffer at room temperature for 10 min, the sections were incubated at room temperature for 1 h with rabbit anti-human CD3 (Abcam, clone SP7, 1:100), rabbit anti-human ROR γ (Abcam, 1:50), and mouse anti-human FOXP3 (Abcam, clone mAbcam22510, 1:100). A secondary horseradish peroxidase-conjugated antibody (PerkinElmer) were added and incubated at room temperature for 10 min. Signal amplification was performed using TSA working solution diluted at 1:100 in 1 × amplification diluent (PerkinElmer) and incubated at room temperature for 10 min. The other validations by multi-colour IHC were performed using the same protocols with different primary antibodies as follows. The primary antibodies and IHC metrics used in the validation of T_C, T_H and T_{reg} cells were rabbit anti-human CD3 (Abcam, clone SP7, 1:400), rabbit anti-human CD4 (Abcam, clone EPR6855, 1:400), mouse anti-human CD8 (Abcam, clone 144B, 1:500) and mouse anti-human FOXP3 (Abcam, clone mAbcam22510, 1:500). The primary antibodies and IHC metrics used in the validation of proliferative CD8 $^+$ T_{EX} cells were: rabbit anti-human TIM-3 (also known as HAVCR2) (Cell Signaling, clone D5D5R, 1:100), mouse anti-human Ki67 (Abcam, B126.1, 1:200), mouse anti-human PD-1 (Abcam, clone NAT105, 1:200) and mouse anti-CD8 (Abcam, clone 144B, 1:200). The multispectral imaging was collected by Mantra Quantitative Pathology Workstation (PerkinElmer, CLS140089) at 20× magnification and analysed by InForm Advanced Image Analysis Software (PerkinElmer) version 2.3. For each patient, a total of 8–15 high-power fields were taken based on their tumour sizes.

Quality control and preprocessing of single-cell RNA-seq data. Low-quality read pairs of single-cell RNA sequencing (scRNA-seq) data were filtered out if at least one end of the read pair met one of the following criteria: (1) 'N' bases account for $\geq 10\%$ of the read length; (2) bases with quality <5 account for $\geq 50\%$ of the read length; and (3) the read contains adaptor sequence. The filtered read pairs were processed using HTSeqGenie pipeline (R package version 4.8) to obtain the gene expression table. Specially, read pairs were then mapped to human ribosomal RNA sequences (download from RFam database) and the read pairs with both ends unmapped were kept for downstream analysis. Read pairs passing this filter for rRNA were aligned to human reference sequence (hg19) using GSAP³⁴, with parameters '-novelsplice 1 -n 10 -i 1 -M 2'. To calculate the expression levels of genes, the gene model file 'knownGene.txt' (30 June 2013 version), downloaded from UCSC, was used. The R function findOverlaps was used to count the number of uniquely mapped read pairs located in each gene and the count table tabulated as genes by cells was used for downstream analysis. The transcripts per million (TPM) table was derived from the count table and the TPM value was calculated by

$$\frac{10^6 \times C_{ij}/\text{length of gene } i}{\sum_i C_{ij}/\text{length of gene } i}$$

in which C_{ij} was the count value of gene i in cell j .

Low-quality cells were filtered if the library size or the number of expressed genes (counts larger than 0) was smaller than predefined thresholds. Both thresholds

were defined as the medians of all cells minus $3 \times$ the median absolute deviation. Furthermore, if the proportion of mitochondrial gene counts was larger than 10%, these cells were discarded. Only cells with the average TPM of CD3D, CD3E and CD3G larger than 10 were kept for subsequent analysis. We further identified CD4 $^+$, CD8 $^+$, CD4 $^-$ CD8 $^-$ (double negative) and CD4 $^+$ CD8 $^+$ (double positive) T cells based on the gene expression data. Given the average TPM of CD8A and CD8B, one cell was considered as CD8 positive or negative if the value was larger than 30 or less than 3, respectively; given the TPM of CD4, one cell was considered as CD4 positive or negative if the value was larger than 30 or less than 3, respectively. Hence, the cells can be in silico classified as CD4 $^+$ CD8 $^-$, CD4 $^-$ CD8 $^+$, CD4 $^+$ CD8 $^+$, CD4 $^-$ CD8 $^-$ and other cells that cannot be clearly defined. A total of 52 cells were filtered out owing to the inconsistent classifications based on transcriptome data and FACS.

After discarding genes with average counts of fewer than or equal to 1, the count table of the cells passing the above filtering was normalized using a pooling strategy implemented in the R function computeSumFactors³⁵. With this strategy, size factors for individual cells were deconvoluted from size factors of pools, the sizes of which ranged from 20, 40, 60, 80 to 100. To avoid violating the assumption that most genes were not differentially expressed, hierarchical clustering based on Spearman's rank correlation was performed first, then normalization was performed in each cluster separately. The size factor of each cluster was further re-scaled to enable comparison between clusters. The normalized data were in log₂ space. To remove the possible effects of different donors on expression, the normalized table was further centred by patient. Thus, in the centred expression table, the mean values of the cells for each patient were zero. A total of 12,548 genes and 10,805 cells were retained in the final expression table. If not explicitly stated, 'normalized read count' or 'normalized expression' in this study refers to the normalized and centred count data for simplicity.

Analysis pipelines of bulk exome sequencing and RNA-seq data. The bulk exome sequencing data were cleaned following the same procedure for the scRNA-seq data processing. The cleaned read pairs were then processed according to the BWA-PICARD/GATK-strelka pipeline. In brief, the cleaned read pairs were aligned to human genome reference version b37 (downloaded from <ftp://ftp.broadinstitute.org/>) by the BWA-MEM algorithm³⁶. The alignments were then sorted and de-duplicated by PICARD (Broad Institute). GATK³⁷ was used to realign multiple reads around putative INDEL by Smith-Waterman alignment algorithm and re-calibrate base quality. The analysis-ready bam files were input into the GATK UnifiedGenotyper module to call SNP/INDEL and into strelka³⁸ to call somatic SNV/INDEL and into ADTEx³⁹ (version 1.0.4) to call somatic copy number alterations. The mutations were annotated with annovar⁴⁰.

TCR assembly. TraCeR was used to deduce the TCR sequences of each cell³. The outputs of TraCeR include the assembled nucleotide sequences for both α and β chains, the coding potential of the nucleotide sequences (that is, productive or not), the translated amino acid sequence, the CDR3 sequences and the estimated TPM value of α or β chains. Only cells with TPM values larger than 10 for the α chain and larger than 15 for the β chain were kept.

For cells with two or more α or β chains assembled, the α - β pair that was productive and of the highest expression level was defined as the dominant α - β pair in the corresponding cell. If two cells had identical dominant α - β pairs, the dominant α - β pair were identified as clonal TCRs. To integrate with the gene expression data, the TCR-based analysis was performed only for cells that passed the aforementioned quality control pipeline (total 10,805). Thus, 9,878 cells with TCR information were used in the integrative analysis (Supplementary Table 4).

If one cell had an α chain composed of V segment TRAV1-2 and one of the following J segments (TRAJ33, TRAJ20 and TRAJ12), the cell was classified as a MAIT cell⁴¹. If the α chain of one cell was rearranged by V segment TRAV10 and J segment TRAJ18, the cell was classified as an invariant natural killer T cell⁴². In the 9,878 cells with at least one pair of productive α and β chains, only 3 cells were identified as invariant natural killer T cells, and 102 cells were identified as MAIT cells, including 71 CD8 $^+$ CD4 $^-$ T cells classified in silico.

Unsupervised clustering analysis of CRC scRNA-seq dataset. The expression tables of CD8 $^+$ CD4 $^-$ T cells and CD8 $^-$ CD4 $^+$ T cells as defined by the aforementioned in silico classification but excluding MAIT cells and invariant natural killer T cells, were fed into an iteratively unsupervised clustering pipeline separately. Specifically, given an expression table, the top n genes with the largest variance were selected, and then the expression data of the n genes were analysed by single-cell consensus clustering (SC3)⁴³. n was tested from 500, 1,000, 1,500, 2,000, 2,500 and 3,000. In SC3, the distance matrices were calculated based on Spearman correlation and then transformed by calculating the eigenvectors of the graph Laplacian. Then, the k-means algorithm was applied to the first d eigenvectors multiple times, in which d was chosen as between 4% and 7% of the total number of input cells. Finally, hierarchical clustering with complete agglomeration was performed on the SC3 consensus matrix and k clusters were inferred. The SC3 parameter k , which was used in the k-means and hierarchical clustering, was tried from 2 to 10.

For each SC3 run, the silhouette values were calculated, the consensus matrix was plotted and cluster-specific genes were identified. Such information was used to determine the optimal k and n . Once the stable clusters were determined, the above procedure was iteratively applied to each of these clusters to reveal the sub-clusters. The *in silico* classified CD8 $^+$ CD4 $^-$ MAIT cells had distinct gene expression patterns compared with other CD8 $^+$ CD4 $^+$ T cells, and were defined as cluster 'CD8_C08-SLC4A10'.

When the clustering results were obtained, one-way ANOVA implemented by R function 'aov' was performed to identify the differentially expressed genes among the clusters. R function TukeyHSD was used to identify which cluster pairs showed a significant difference. A gene was defined as being significantly differentially expressed based on the following criteria: (1) adjusted P value (Benjamini-Hochberg method) of F -test of less than 0.05; (2) the absolute difference of any one significant cluster pair (P value of Tukey's honest significant difference method less than 0.01) larger than 1. The significantly differentially expressed genes were categorized in the cluster that showed the highest expression (Supplementary Table 5).

The *t*-SNE method implemented in R package Rtsne was used for clustering visualization. To visualize the cell density on the *t*-SNE plot, kernel density estimation was performed using R function 'kde' (ks package), and the contour lines encompassing the top 10%, 20%, ... 90% cells with highest densities were shown. A total of 8,530 T cells, including 3,628 CD8 $^+$ CD4 $^-$ and 4,902 CD8 $^-$ CD4 $^+$ T cells with clustering definitions, were used in the *t*-SNE projection. Other cells such as CD8 $^+$ CD4 $^+$ and CD8 $^-$ CD4 $^-$ T cells were not included in this visualization.

To validate the clustering results from the SC3 pipeline, we also performed clustering analysis using two additional pipelines, Seurat⁴⁴ and sscClust¹⁰. Raw read-count tables were provided to the Seurat pipeline. For each cell, the counts were normalized by the total counts then multiplied by a scale factor of 100,000 before transforming to the log₂ scale. To identify highly variable genes, the relationship between mean expression and dispersion was fitted with log(VMR) (variance to mean ratio) as dispersion function. The mean expression cut-offs were set at 0.0125 and 8 for low and high limit, respectively, and dispersion cut-off was set at 0.5 for low limit. The donor covariate effect was removed by regression, and the resulting data were used to perform PCA. The top 15 principal components were kept and clusters were identified by the SNN algorithm. Resolution parameters of 0.7 and 1.0 were set for CD8 $^+$ T cell data and CD4 $^+$ T cell data, respectively.

The sscClust method is a two-round clustering pipeline that uses both PCA and *t*-SNE for dimension reduction and uses a density-based clustering method. The normalized expression data used in SC3 analysis were also used in the sscClust analysis. In the first round, the top 1,500 genes with the highest standard deviation were used for PCA, and top principal components were used for *t*-SNE, implemented in R package Rtsne. The R package densityClust⁴⁵ was used for density peak identification and cluster assignment. Subsequently, differentially expressed genes were identified by analysis of variance, and the top 1,000 genes were used in the second round of PCA/*t*-SNE/densityClust analysis. In both rounds, top 15 principal components were used. The rho and delta parameters of densityClust, which denote the density of each cell and the minimum distance to other cells with density larger than the cell in consideration, were chosen based on the rho-delta decision plot. Specifically, for the CD8 $^+$ T cell data, rho/delta of 40/5 and 30/3 were used for the first and second round analysis, respectively; for the CD4 $^+$ T cell data, 50/4 and 50/4 were used for first and second round analysis, respectively.

Down-sampling analysis of CRC T cell scRNA-seq dataset. To evaluate the effect of cell numbers on clustering results, we iteratively repeated the clustering analysis after down-sampling the CRC T cell data to 100, 200, 500, 1,000, 1,500, 2,000 and 3,000 cells. For each down-sampling number, 10 replicates were performed. Each down-sampled dataset was used for clustering analysis by sscClust (for speed considerations and similarity to the SC3 results), the resulting cluster labels were compared with our benchmark labels, as obtained from the whole dataset analysis, using the normalized mutual information (NMI) index⁴⁶. A higher NMI index means more accurate cluster assignment in the down-sampled dataset. The sscClust pipeline was run with largely the same procedure described in the previous section but with different (rho and delta) parameters. The same (rho and delta) parameters were used in both rounds of clustering. The rho parameter was set at 1.5, 3, 7.5, 15, 20, 20 and 20 for cell numbers ranging from 100 to 3,000, respectively. For the delta parameter, values of 2 to 7 were tested and the value that gave the highest NMI value was chosen as the optimal parameter.

Analysis of combined CRC, HCC and NSCLC T cell scRNA-seq datasets. The expression data of cells passing the quality control filters in the three studies for HCC, NSCLC and CRC T cells (GEO accessions GSE98638, GSE99254 and GSE108989) were fetched, and then re-processed using the same aforementioned pipeline for *in silico* classification and normalization. The *in silico*-classified CD8 $^+$ CD4 $^-$ and CD8 $^-$ CD4 $^+$ T cells, excluding invariant natural killer T cells, were used in combined clustering analysis. The large combined dataset demanded the use of the sscClust method for its high computational efficiency. For CD8 $^+$ T cell data, rho and delta parameters (100, 5) and (100, 3) were used for the first and

second round analysis respectively, and for CD4 $^+$ T cell data, (120, 2) and (130, 2) were used for the first and second round analysis, respectively.

Gene set enrichment analysis. Pre-ranked analysis module in GSEA⁴⁷ was used for gene set enrichment analysis. The gene sets we used were from the database MSigDB⁴⁸. In single cluster enrichment analysis, the normalized and centred expression data were transformed to z -scores. For each cluster, the z -scores across cells were averaged per gene. Each cluster had an average expression profile. The z -score profile was used as input for GSEA.

Identification of proliferative cells. The average expression of known proliferation-related genes was defined as the proliferation score. These proliferation genes include ZWINT, E2F1, FEN1, FOXM1, H2AFZ, HMGB2, MCM2, MCM3, MCM4, MCM5, MCM6, MKI67, MYBL2, PCNA, PLK1, CCND1, AURKA, BUB1, TOP2A, TYMS, DEK, CCNB1 and CCNE1⁴⁹. Proliferative cells were identified using an outlier detection procedure implemented in the R package extremevalues. Specifically, as most of the proliferation scores came from low-proliferation cells, a normal distribution was fitted using cells with proliferation scores between 10% and 90% quantiles. Cells with a proliferation score larger than a threshold were classified as proliferative cells, and this threshold value was optimally set by the getOutliersI function of the extremevalues package.

Trajectory analysis. To characterize the potential process of T cell functional changes and determine the potential lineage differentiation between diverse T cells, we applied the Monocle (version 2) algorithm⁵⁰ with the top 700 signature genes of CD8 $^+$ T cells excluding MAIT cells, based on the rank of F statistic generated by ANOVA (Supplementary Table 5). Cells were ordered through the inferred pseudotime to indicate their differentiation progress. The Monocle function relative2abs was used to convert TPM measurement into mRNAs per cell (RPC), and then the CellData Set object was created with the parameter 'expressionFamily = negbinomial'. Then the CD8 $^+$ T cell differentiation trajectory was inferred after dimension reduction and cell ordering with the default parameters of R package Monocle.

TCGA data analysis. The TCGA colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) data were used to confirm the differences of the T cell subtype compositions between patients with MSI-H ($n = 62$) and MSS ($n = 286$). None of the patients from the TCGA COAD and READ had any previous record of immunotherapy treatment. The gene expression data and clinical data were downloaded from UCSC Xena (<http://xena.ucsc.edu/>). We calculated the average expression of known marker genes of T_H17 cells (*IL17A*, *IL17F*, *IL23R*, *CCR6*, *RORC* and *CD4*) and T_H1-like cells (*CXCL13*, *HAVCR2*, *IFNG*, *CXCR3*, *BHLHE40* and *CD4*) after z -score normalization with log-transformed expression profiles. P values from the Wilcoxon test were used to determine the statistical significance in R.

Definition of STARTRAC indices for tissue distribution, clonal expansion, tissue migration and state transition. We present STARTRAC as a framework, defined by four indices, to analyse different aspects of T cells based on paired single-cell transcriptomes and TCR sequences. The first index, STARTRAC-dist, uses the ratio of observed over expected cell numbers in tissues to measure the enrichment of T cell clusters across different tissues. Given a contingency table of T cell clusters by tissues, we first apply chi-squared test to evaluate whether the distribution of T cell clusters across tissues significantly deviates from random expectations. We then calculate the STARTRAC-dist index for each combination of T cell clusters and tissues according the following formula:

$$I_{\text{dist}}^{\text{STARTRAC}} = R_{o/e} = \frac{\text{observed}}{\text{expected}}$$

in which $R_{o/e}$ is the ratio of observed cell number over the expected cell number of a given combination of T cell cluster and tissue. The expected cell number for each combination of T cell clusters and tissues are obtained from the chi-squared test. Different from the chi-squared values, which are defined as $\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$ can only indicate the divergence of observations from random expectations, $I_{\text{dist}}^{\text{STARTRAC}}$ defined by $R_{o/e}$ can indicate whether cells of a certain T cell cluster are enriched or depleted in a specific tissue. For example, if $R_{o/e} > 1$, it suggests that cells of the given T cell cluster are more frequently observed than random expectations in the specific tissue, that is, enriched. If $R_{o/e} < 1$, it suggests that cells of the given T cell cluster are observed with less frequency than random expectations in the specific tissue, that is, depleted. By calculating the STARTRAC-dist indices via $R_{o/e}$, we can quantify the tissue preference of T cell clusters efficiently.

The other three STARTRAC indices, STARTRAC-expa, STARTRAC-migr and STARTRAC-tran, are designed to measure the degree of clonal expansion, tissue migration, and state transition of T cell clusters upon TCR tracking, respectively. The MAIT cells (CD8_C08-SLC4A10) were not included in these types of analyses because they have distinct TCRs. For STARTRAC-expa, which uses the standard TCR clonality measurement⁵¹ but is specifically applied to different T cell clusters in our analyses, we first adopt the normalized Shannon entropy to calculate the evenness of the TCR repertoire of the given T cell cluster and then define the

STARTRAC-expa index as 1 – evenness. Mathematically, the STARTRAC-expa index of a specific cluster with N clonotypes is defined by the following formula:

$$I_{\text{expa}}^{\text{STARTRAC}} = 1 - \text{evenness} = 1 - \frac{-\sum_{i=1}^N p_i \log_2 p_i}{\log_2 N}$$

in which p_i is the cell frequency of clonotype i in the cluster, and a clonotype is defined by identical, full-length, paired α and β TCR chains. Although the definition of STARTRAC-expa is mathematically identical to the clonality scores frequently used in bulk TCR repertoire sequencing studies⁵¹, two distinctions should be noted. First, STARTRAC-expa is defined for T cell clusters while the traditional TCR clonality is defined for specific specimens. A T cell cluster in STARTRAC framework can consist of T cells from several tissues and patients, but the specimens subject to bulk TCR repertoire sequencing are typically from a unique tissue and patient. Second, STARTRAC-expa uses a more stringent clonotype definition. For traditional bulk TCR sequencing studies, clonotypes are generally defined based on identical CDR3 (the complementarity determining region 3) sequences of TCR α or β chains, owing to technological limitations. However, STARTRAC-expa is defined using the strictest clonotype definition, which requires that both the full-length α and β chains of TCRs are identical at the nucleotide level. Thus, although STARTRAC-expa has an identical mathematical formula to that of the traditional TCR clonality definition, they have distinct biological meanings. STARTRAC-expa ranges from 0 to 1, with 0 indicating no clonal expansion for each clonotype while 1 indicating that the cluster is composed of only one clonally expanded clonotype. If a cluster is composed of multiple clonotypes and each clonotype is subject to distinct extent of clonal expansion, STARTRAC-expa will be between 0 and 1, with high STARTRAC-expa indicating high clonality.

Even if T cells with identical TCR clonotypes are present in different tissues or in different development states, logically they could likely derive from a single naive T cell, clonally expanded initially at one location and migrated across tissues, or have undergone state transitions. Based on this principle, we define STARTRAC-migr and STARTRAC-tran to evaluate the extent of tissue migration and state transition of each clonotype, respectively. For each clonotype, given its distribution across tissues (peripheral blood, adjacent normal mucosa and tumour), we define its STARTRAC-migr index I_{migr}^t as:

$$I_{\text{migr}}^t = -\sum_{j=1}^J p_j^t \log_2 p_j^t$$

in which p_j^t is the ratio of the number of cells with TCR clonotype t in tissue j to the total number of cells with TCR clonotype t and $\sum_{j=1}^J p_j^t = 1$. For two T cell clusters with similar clonal expansion and clonal size, the one with clonal cells broadly distributed in various tissues would probably be more mobile. Similarly, the STARTRAC-tran index I_{tran}^t can be defined as:

$$I_{\text{tran}}^t = -\sum_{k=1}^K p_k^t \log_2 p_k^t$$

in which p_k^t is the ratio of the number of cells with TCR clonotype t in cluster k to the total number of cells with TCR clonotype t , $\sum_{k=1}^K p_k^t = 1$, and K is the total number of cell clusters. Although both definitions use Shannon entropy for calculation, they are distinct from the measurement of TCR clonality in bulk TCR repertoire sequencing. As described above, the traditional TCR clonality is defined at the sample level; however, STARTRAC-migr and STARTRAC-tran are defined primarily at the clonotype level. Given one clonotype, the evenness or diversity of its TCR repertoire will be zero because all the cells have identical TCRs, while the STARTRAC-migr and STARTRAC-tran indices will be non-trivial because cells of the same clonotype can migrate across tissues or change their transcriptional states. Thus, the inputs of the formulas of STARTRAC-migr and STARTRAC-tran are also different from the traditional TCR clonality measurement and STARTRAC-expa. The input of STARTRAC-migr is the observed cell frequency across tissues of a certain clonotype, while the input of STARTRAC-tran is the observed cell frequency across cell clusters of a certain clonotype. By contrast, the input of STARTRAC-expa is the observed cell frequency across clonotypes of a certain cell cluster, and the input for the traditional TCR clonality measure is the observed sequence frequency across a TCR repertoire of a given sample. For the calculation of STARTRAC-migr, to exclude the possible influence of different extent of expansion or local proliferation of T cells, we also calculate the proliferation-normalized STARTRAC-migr index, which normalizes the number of expanded cells of clonotypes in each tissue as 1. As expected, we found a similar trend of T cell migration potentials for both CD8⁺ and CD4⁺ T cells evaluated by this proliferation-normalized STARTRAC-migr as those calculated by STARTRAC-migr (data not shown). To make our calculation consistent, we used STARTRAC-migr for subsequent analyses.

After the extent of tissue migration of each clonotype is quantified by STARTRAC-migr, given a cluster with total T clonotypes, the STARTRAC-migr index at the cluster level $I_{\text{migr}}^{\text{STARTRAC}}$ can be defined as the weighted average of all TCR clonotype migration indices contained in the cluster:

$$I_{\text{migr}}^{\text{STARTRAC}} = \sum_{t=1}^T p_{\text{cls}}^t I_{\text{migr}}^t$$

in which p_{cls}^t is the ratio of the number of cells with clonotype t in cluster cls to the total number of cells in cluster cls.

Similarly, when the extent of state transition of each clonotype is quantified by STARTRAC-tran, given a cluster with total T clonotypes, the STARTRAC-tran index at the cluster level can be defined as the weighted average of all TCR clonotypes state transition indices contained in the cluster:

$$I_{\text{tran}}^{\text{STARTRAC}} = \sum_{t=1}^T p_{\text{cls}}^t I_{\text{tran}}^t$$

in which p_{cls}^t is the ratio of the number of cells with clonotype t in cluster cls to the total number of cells in cluster cls.

Of note, both STARTRAC-migr and STARTRAC-tran are defined at two different levels (clonotypes and clusters), with the clonotype-level definitions describing the extent of migration and state transition of a given clonotype, and the cluster-level definitions depicting the summarization of such properties of all clonotypes within a cluster.

Besides the overall evaluation of the extents of migration and state transitions by STARTRAC-migr and STARTRAC-tran, we also define pairwise STARTRAC-migr (pSTARTRAC-migr) and STARTRAC-tran (pSTARTRAC-tran) indices for precise quantification. For example, given a clonotype t and two tissue types (for example, blood and tumour), the pSTARTRAC-migr index $p I_{\text{migr}}^t$ is calculated by the following formula:

$$p I_{\text{migr}}^t = -\sum_{j=1}^2 p_j^t \log_2 p_j^t$$

in which p_j^t is the ratio of the number of cells with TCR clonotype t in tissue j to the total number of cells with TCR clonotype t in tissues 1 and 2 (that is, blood and tumour), and $\sum_{j=1}^2 p_j^t = 1$. In other words, pSTARTRAC-migr uses the same formula as STARTRAC-migr but limits the number of tissues to two and the frequencies of cells between two specified tissues are re-calculated. Likewise, given a clonotype t and two T cell clusters (for example, T_{EM} and T_{EX} cells), the pSTARTRAC-tran index $p I_{\text{tran}}^t$ is calculated by the following formula:

$$p I_{\text{tran}}^t = -\sum_{k=1}^2 p_k^t \log_2 p_k^t$$

in which p_k^t is the ratio of the number of cells with TCR clonotype t in cluster k to the total number of cells with TCR clonotype t in clusters 1 and 2 (that is, T_{EM} and T_{EX} cells), and $\sum_{k=1}^2 p_k^t = 1$. Thus, pSTARTRAC-tran uses the same formula as STARTRAC-tran but limits the number of clusters to two, and the frequencies of cells between the two specified clusters are re-calculated. Once pairwise STARTRAC-migr and STARTRAC-tran for clonotypes are obtained, the corresponding indices for clusters are calculated via weighted average according to their clonotype compositions. As all STARTRAC-migr and STARTRAC-tran indices are defined by Shannon entropy, high values indicate high migration and state transition, respectively.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

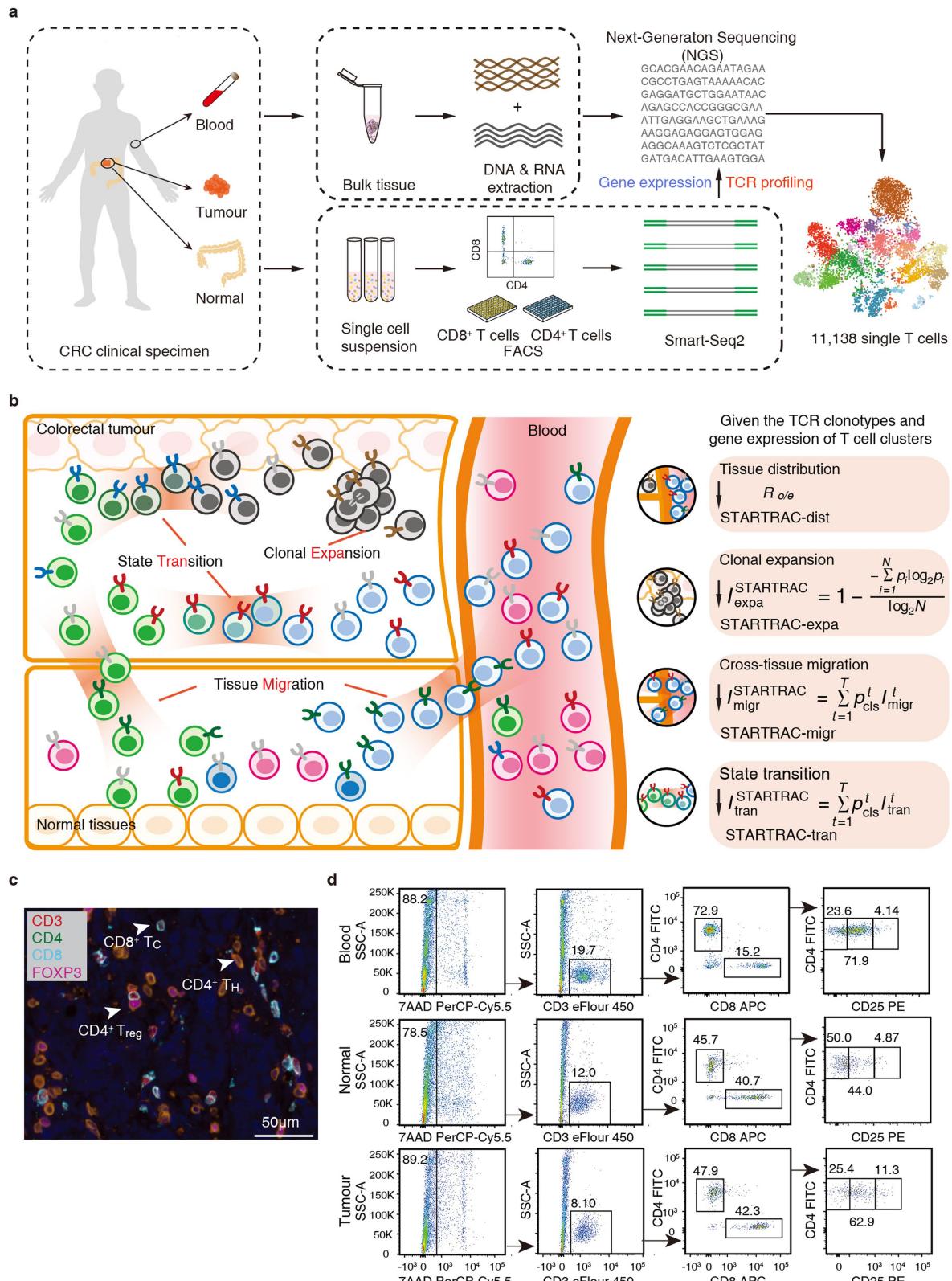
Code availability. The open source code is available at GitHub. Code for ssc-Clust clustering is available on GitHub (<https://github.com/Japrin/sscClust>). Code for STARTRAC analysis is available on GitHub (<https://github.com/Japrin/STARTRAC>).

Data availability

The data that support the findings of this study are available from the corresponding author upon request. Sequencing data are available at EGA (accession number EGAS00001002791), and processed gene expression data can be obtained from Gene Expression Omnibus (GEO) (accession number GSE108989).

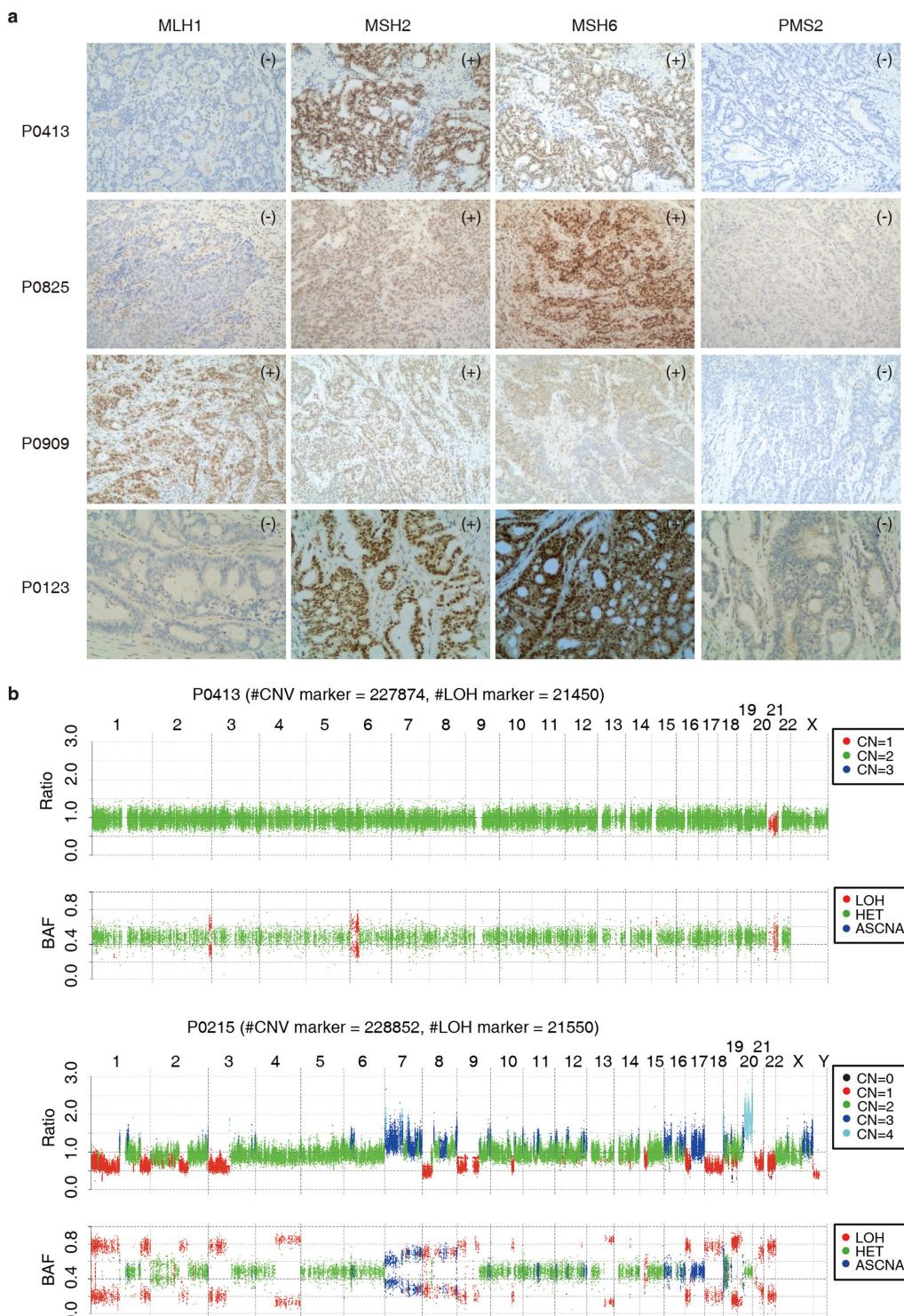
31. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
32. Emtage, P. C. et al. Second-generation anti-carcinoembryonic antigen designer T cells resist activation-induced cell death, proliferate on tumor contact, secrete cytokines, and exhibit superior antitumor activity *in vivo*: a preclinical evaluation. *Clin. Cancer Res.* **14**, 8112–8122 (2008).

33. Chodisetti, S. B. et al. Triggering through Toll-like receptor 2 limits chronically stimulated T-helper type 1 cells from undergoing exhaustion. *J. Infect. Dis.* **211**, 486–496 (2015).
34. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
35. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
36. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
37. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
38. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
39. Amarasinghe, K. C. et al. Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15**, 732 (2014).
40. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
41. van Wilgenburg, B. et al. MAIT cells are activated during human viral infections. *Nat. Commun.* **7**, 11653 (2016).
42. Godfrey, D. I., Stankovic, S. & Baxter, A. G. Raising the NKT cell family. *Nat. Immunol.* **11**, 197–206 (2010).
43. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
44. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
45. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
46. Strehl, A. & Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002).
47. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
48. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
49. Whitfield, M. L., George, L. K., Grant, G. D. & Perou, C. M. Common markers of proliferation. *Nat. Rev. Cancer* **6**, 99–106 (2006).
50. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).
51. Kirsch, I., Vignali, M. & Robins, H. T-cell receptor profiling in cancer. *Mol. Oncol.* **9**, 2063–2070 (2015).



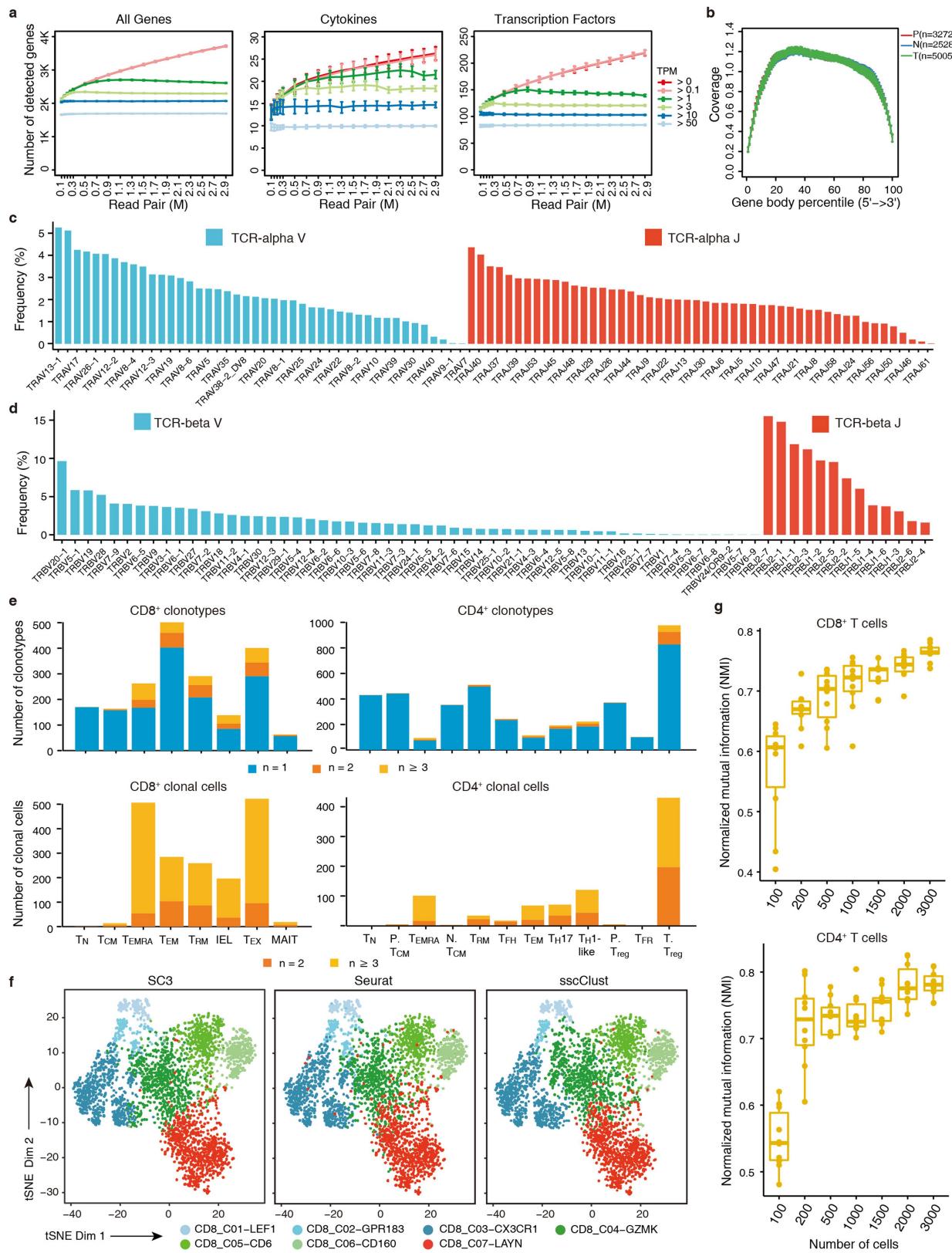
Extended Data Fig. 1 | Study design and tracking T cell dynamics of patients with CRC by STARTRAC. **a**, The experimental flowchart of this study. **b**, A cartoon illustrating four indices defined by STARTRAC to characterize T cell dynamics. STARTRAC-dist, tissue preference of a cluster estimated by ratios of observed cell numbers to random expectations ($R_{o/e}$); STARTRAC-expa, degree of clonal expansion of a cluster defined as ‘1 – evenness’, with evenness as the normalized Shannon entropy of its TCR distribution; STARTRAC-migr, migratory potential of a cluster estimated by the average entropy of its clonotypes across tissues; STARTRAC-tran, potentials of developmental transitions of a cluster,

estimated by the average entropy of its clonotypes across two different functional clusters. The detailed definitions of STARTRAC indices are in Methods. **c**, Opal multi-colour IHC staining with anti-CD3, -CD4, -CD8 and -FOXP3 antibodies to validate the existence of T cells in CRC tumours (exemplified by patient P0215). Original magnification, $\times 20$. T_c, CD8⁺ cytotoxic T cells; T_H, CD4⁺ T helper cells. **d**, Gating strategy for single T cell sorting in this study (exemplified by patient P0215). T_c, T_H and T_{reg} cells were enriched by sorting 7AAD⁻CD3⁺CD8⁺, 7AAD⁻CD3⁺CD4⁺CD25^{-/int} and 7AAD⁻CD3⁺CD4⁺CD25^{hi} T cells, respectively.



Extended Data Fig. 2 | Pathological and genomic characteristics of CRC tumours in the study. **a**, Deficiency of mismatch repair proteins including MLH1, MSH2, MSH6 and PMS2 in all MSI patients (P0413, P0825, P0909 and P0123) measured by IHC ($n = 12$ patients). +, proficiency; -, deficiency. Original magnification, $\times 200$. **b**, Profiles of DNA copy numbers of two representative patients (MSI patient, top; MSS patient,

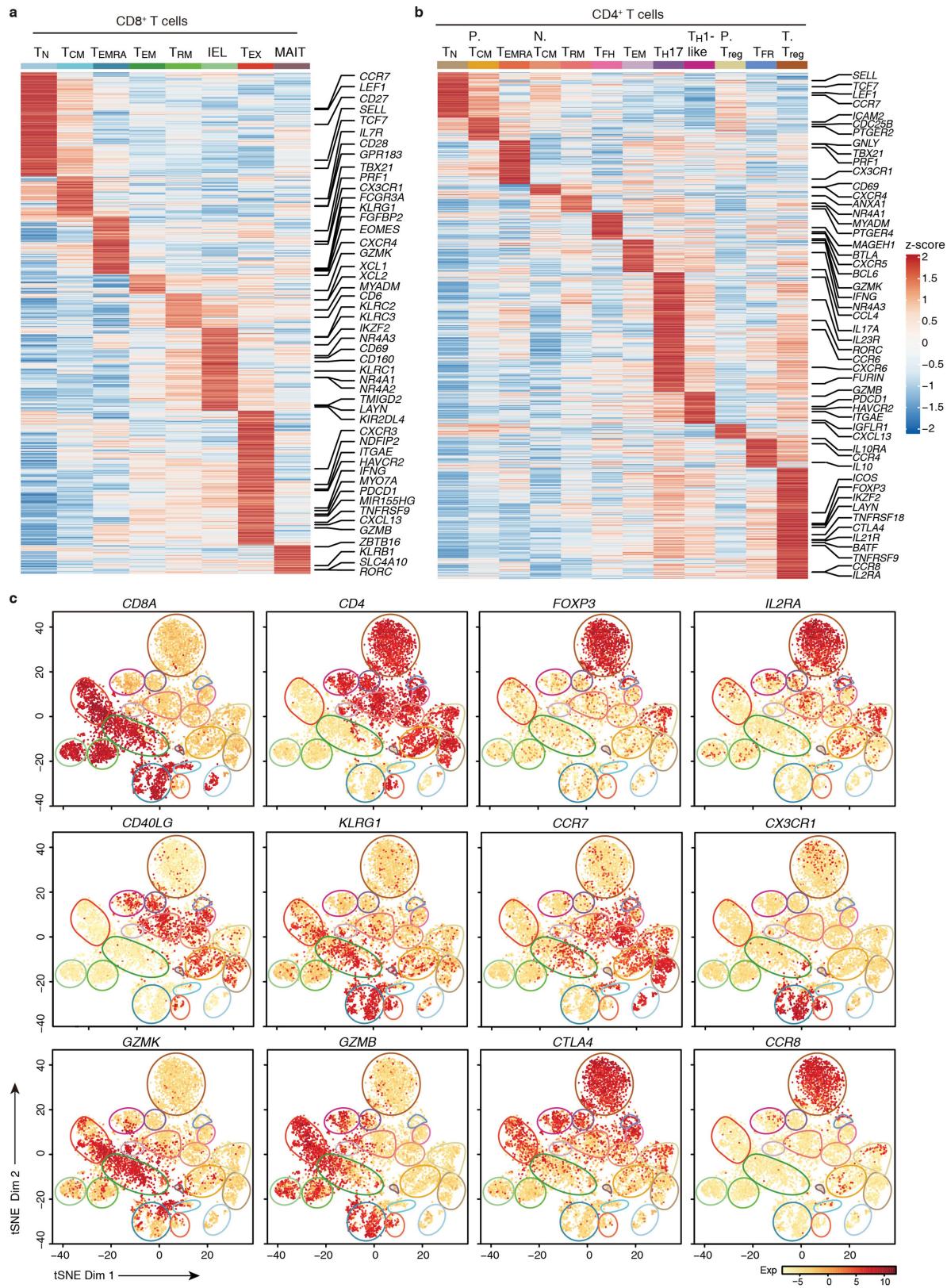
bottom). The copy number information was obtained by ADTex and depicted in bin count plots across chromosomes. The read count ratios ('1' in y axis means baseline copy number) and B allele frequencies (BAF) are shown. Various coloured dots in the ratio graph represent different copy number status of each segment. ASCNA, allele-specific copy number alteration; HET, heterozygous; LOH, loss of heterozygosity.



Extended Data Fig. 3 | See next page for caption.

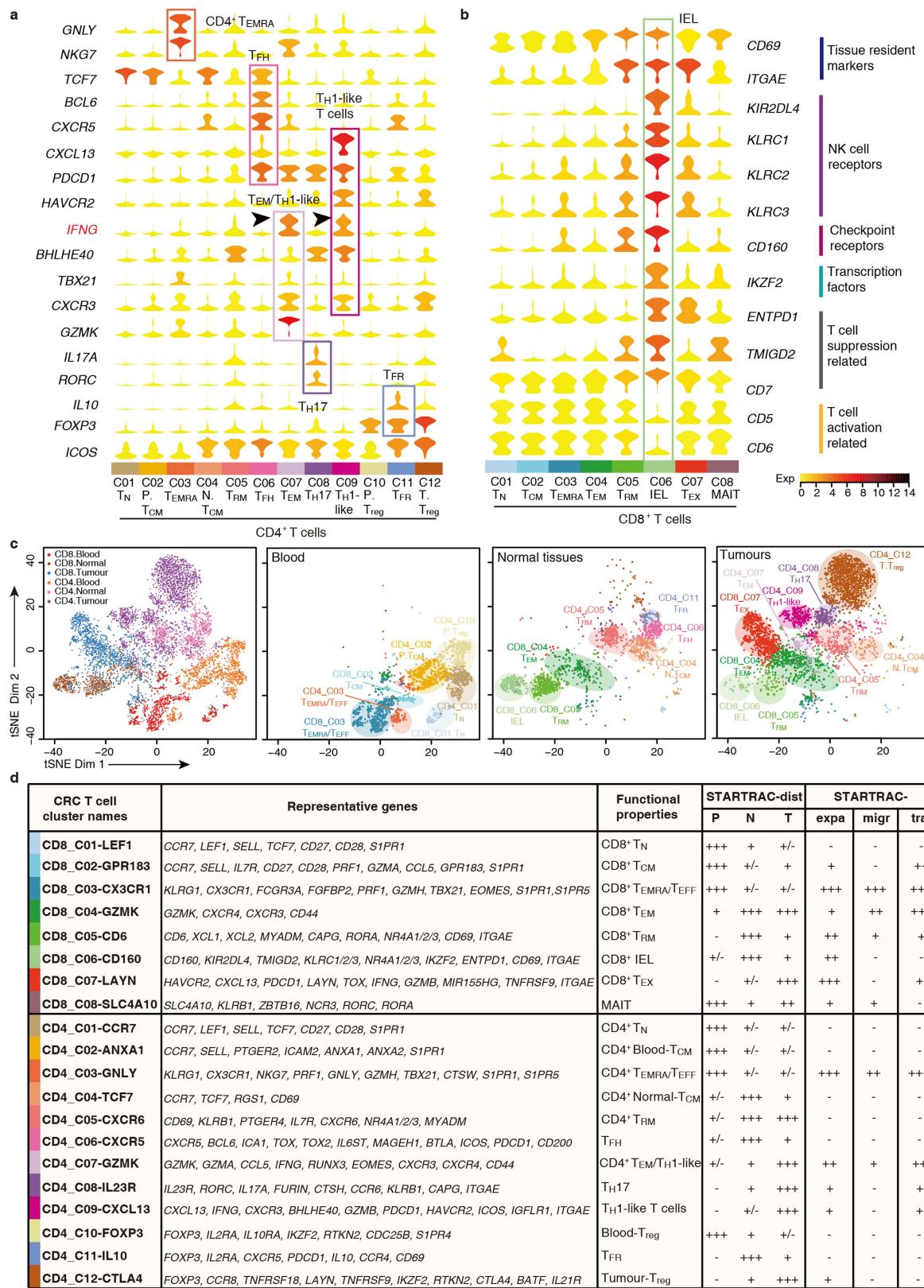
Extended Data Fig. 3 | Basic information of the single T cell RNA-seq data. **a**, Saturation curves of the number of detected genes against sequencing depth (exemplified by cell NTC-53 from patient P1228). Each point on the curve is derived from calculations based on the random selection of a fraction of raw reads from each sample, representing the average of 100 replicate sub-samplings. Error bars denote s.d. Each line with a different colour shows how fast a gene can reach detection saturation at different expression levels, represented by a particular TPM value. **b**, Unbiased coverage of gene body from 5' to 3' between blood, tumours and adjacent normal tissues. **c**, Frequencies of the V and J segments of the TCR α chains. **d**, Frequencies of the V and J segments of the TCR β chains. **e**, Bar plots showing the number of clonotypes and clonal cells in each CD8 $^+$ and CD4 $^+$ T cell cluster. The clonotypes are categorized as unique ($n = 1$) and clonal ($n = 2$ and $n \geq 3$) based on their

cell numbers. Clonal cells are defined as those clonotypes containing at least two cells. **f**, t-SNE projection of 3,557 CD8 $^+$ T cells (CD8_C01-LEF1, $n = 174$; CD8_C02-GPR183, $n = 169$; CD8_C03-CX3CR1, $n = 743$; CD8_C04-GZMK, $n = 773$; CD8_C05-CD6, $n = 487$; CD8_C06-CD160, $n = 351$; CD8_C07-LAYN, $n = 860$) based on different clustering methods including SC3, Seurat and sscClust. Each point represents one single cell coloured by cluster label. **g**, Box plots showing the down-sampling analysis of clustering performed on CD8 $^+$ and CD4 $^+$ T cell dataset. Each dot represents an individual clustering of a given number of T cells. The down-sampling and clustering were performed iteratively for each cell number ($n = 10$ times). Each down-sampled clustering was compared to the clustering performed on the entire dataset, using the NMI index. Higher NMI values indicate more accurate cluster assignment.



Extended Data Fig. 4 | Expression levels of signature genes in each T cell cluster. **a**, Gene expression heat map of 8 CD8⁺ T cell ($n = 3,628$) clusters. Rows represent signature genes and columns represent different clusters. **b**, Gene expression heat map of 12 CD4⁺ T cell clusters ($n = 4,902$). **c**, t-SNE plot of expression levels of selected genes in different clusters indicated by the coloured oval corresponding to Fig. 1a. Number of cells contained in each cluster: CD8_C01-LEF1, $n = 174$; CD8_C02-GPR183, $n = 169$; CD8_C03-CX3CR1, $n = 743$; CD8_C04-GZMK, $n = 773$;

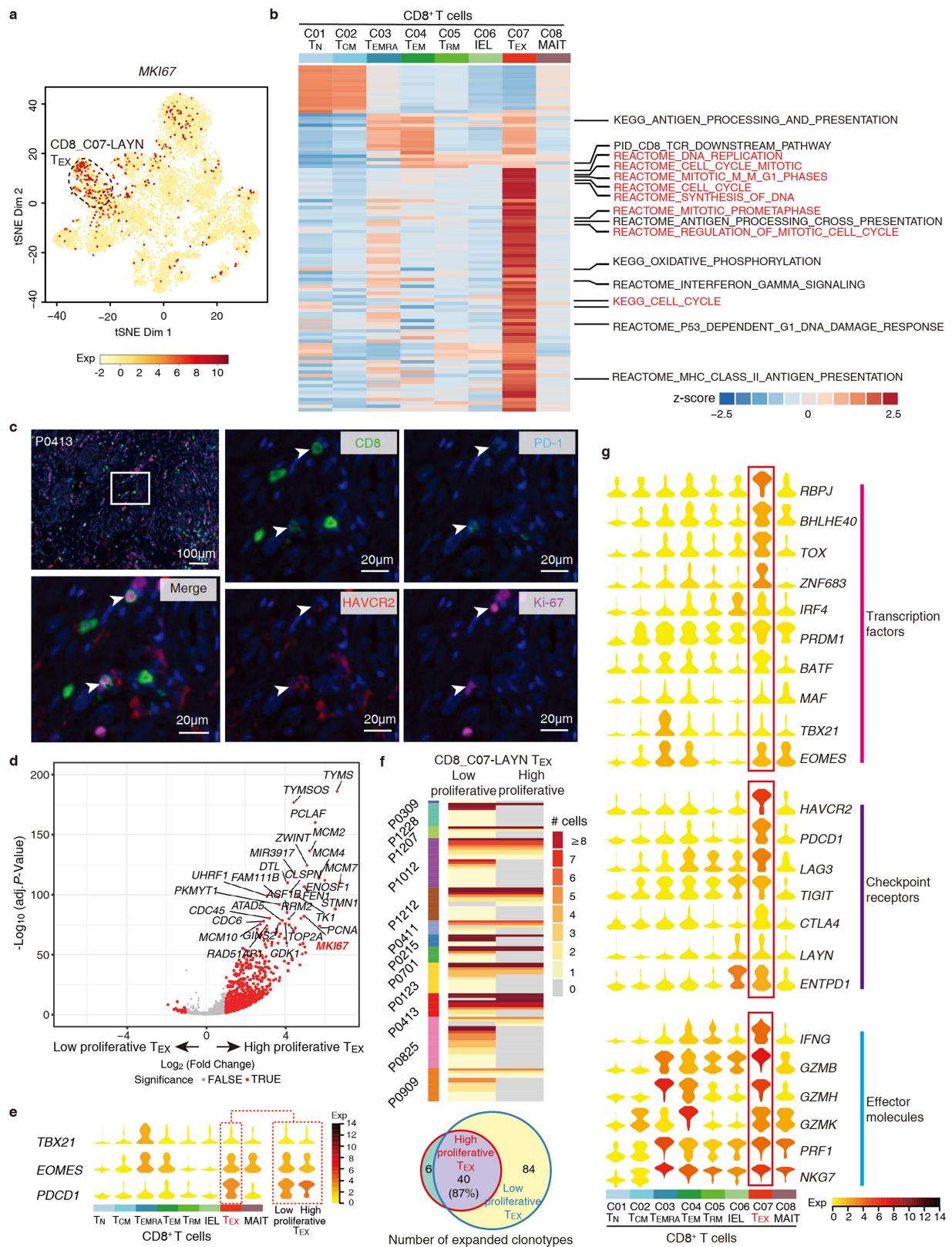
CD8_C05-CD6, $n = 487$; CD8_C06-CD160, $n = 351$; CD8_C07-LAYN, $n = 860$; CD8_C08-SLC4A40, $n = 71$; CD4_C01-CCR7, $n = 462$; CD4_C02-ANXA1, $n = 472$; CD4_C03-GNLY, $n = 190$; CD4_C04-TCF7, $n = 388$; CD4_C05-CXCR6, $n = 568$; CD4_C06-CXCR5, $n = 262$; CD4_C07-GZMK, $n = 185$; CD4_C08-IL23R, $n = 244$; CD4_C09-CXCL13, $n = 319$; CD4_C10-FOXP3, $n = 389$; CD4_C11-IL10, $n = 103$; CD4_C12-CTLA4, $n = 1,320$.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Summary of functional properties of various T cell clusters. **a**, Functional subsets of CD4⁺ T cell ($n = 4,902$) clusters defined by a set of known marker genes. Number of cells contained in each CD4⁺ cluster: T_N, $n = 462$; P.T_{CM}, $n = 472$; T_{EMRA}, $n = 190$; N.T_{CM}, $n = 388$; T_{RM}, $n = 568$; follicular T helper (T_{FH}), $n = 262$; T_{EM}, $n = 185$; T_{H17}, $n = 244$; T_{H1}-like cells, $n = 319$; P.T_{reg}, $n = 389$; N.T_{reg}, $n = 103$; T.T_{reg}, $n = 1,320$. N, normal tissue; P, peripheral blood; T, tumour. **b**, Characteristics of the CD8⁺ IEL T cells as defined by the expression properties of a panel of functionally relevant genes in CD8⁺ T cells ($n = 3,628$). Number of cells contained in each CD8⁺ cluster: T_N, $n = 174$; T_{CM}, $n = 169$; T_{EMRA}, $n = 743$; T_{EM}, $n = 773$; T_{RM}, $n = 487$; IEL, $n = 351$; T_{EX}, $n = 860$; MAIT, $n = 71$. For violin plots in **a** and **b**, colours denote average expression levels; widths denote cell densities. **c**, t-SNE plot

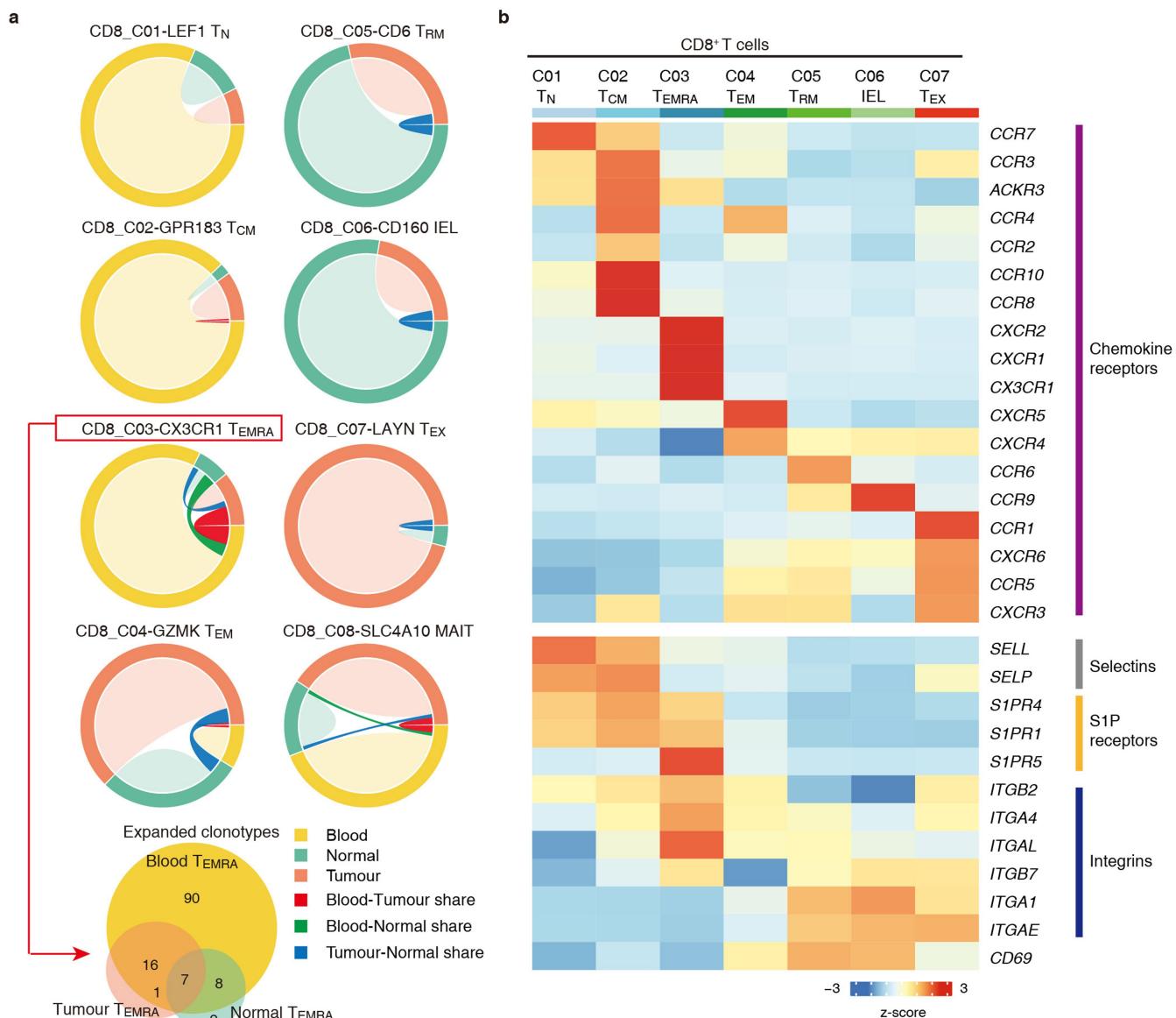
showing the presence of different T cell clusters in peripheral blood ($n = 2,449$; CD8⁺ T cells, $n = 1,021$; CD4⁺ T cells, $n = 1,428$), adjacent normal tissues ($n = 1,962$; CD8⁺ T cells, $n = 961$; CD4⁺ T cells, $n = 1,001$) and tumours ($n = 4,119$; CD8⁺ T cells, $n = 1,646$; CD4⁺ T cells, $n = 2,473$). **d**, Overview of T cell cluster characteristics. STARTRAC-dist: +++ indicates $R_{o/e} > 1$; ++, $0.8 < R_{o/e} \leq 1$; +, $0.2 < R_{o/e} \leq 0.8$; +/-, $0 < R_{o/e} \leq 0.2$; -, $R_{o/e} = 0$. STARTRAC-expa: +++ indicates $I_{\text{expa}}^{\text{STARTRAC}} > 0.10$; ++, $0.06 < I_{\text{expa}}^{\text{STARTRAC}} \leq 0.10$; +, $0.005 < I_{\text{expa}}^{\text{STARTRAC}} \leq 0.05$; -, $I_{\text{expa}}^{\text{STARTRAC}} \leq 0.005$. STARTRAC-migr: +++ indicates $I_{\text{migr}}^{\text{STARTRAC}} > 0.50$; ++, $0.21 < I_{\text{migr}}^{\text{STARTRAC}} \leq 0.50$; +, $0.1 < I_{\text{migr}}^{\text{STARTRAC}} \leq 0.20$; -, $I_{\text{migr}}^{\text{STARTRAC}} \leq 0.1$. STARTRAC-tran, +++ indicates $I_{\text{tran}}^{\text{STARTRAC}} > 0.20$; ++, $0.10 < I_{\text{tran}}^{\text{STARTRAC}} \leq 0.20$; +, $0.05 < I_{\text{tran}}^{\text{STARTRAC}} \leq 0.10$; -, $I_{\text{tran}}^{\text{STARTRAC}} \leq 0.05$.



Extended Data Fig. 6 | See next page for caption.

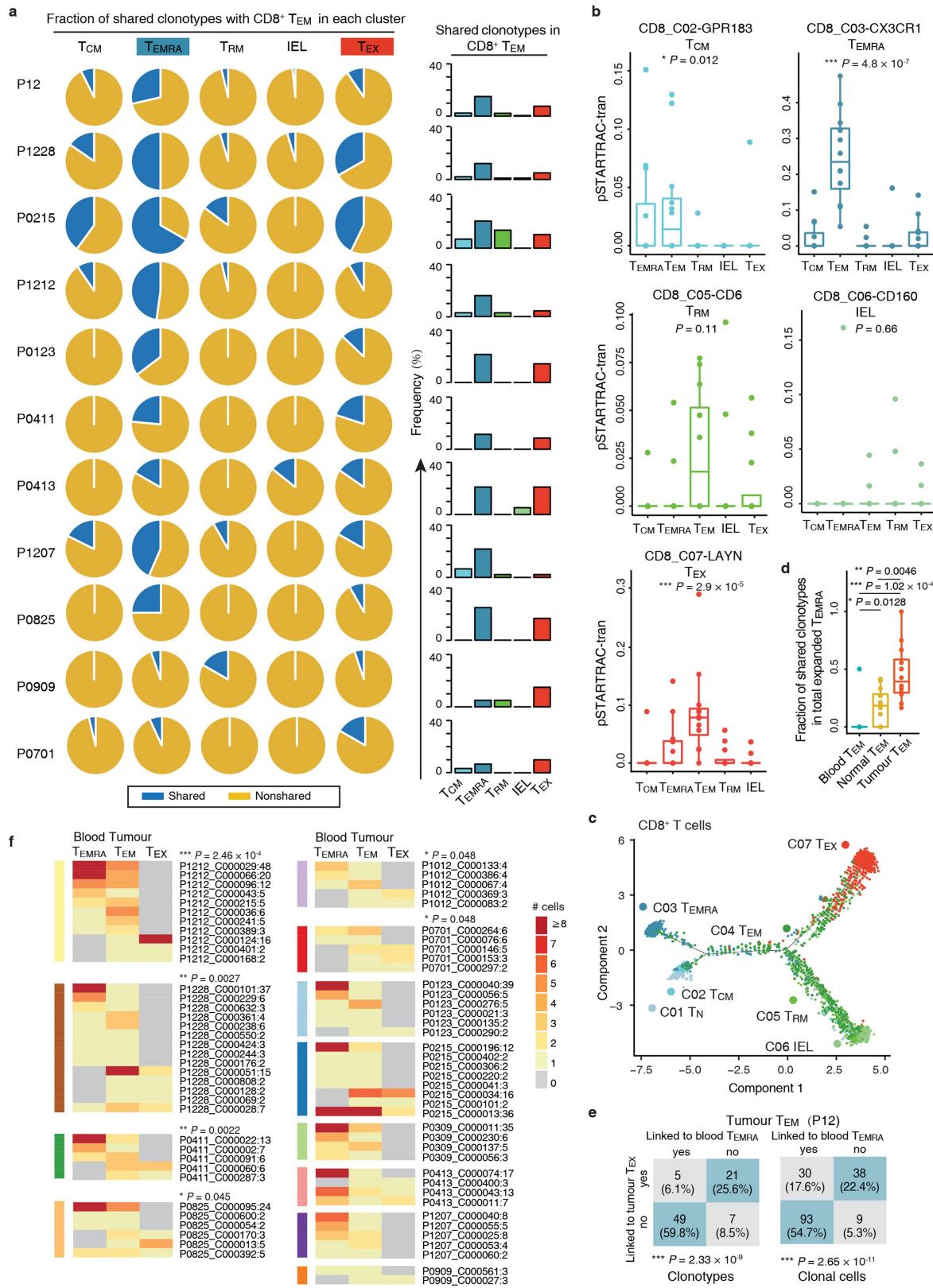
Extended Data Fig. 6 | CD8⁺ T_{EX} cells are characterized by high proliferation property and production of effector molecules. **a**, A subpopulation of CD8⁺ T_{EX} shows high expression of *MKI67* among 8,530 T cells. **b**, Gene set enrichment analysis (GSEA) showing the enrichment of proliferation-related pathways in CD8⁺ T_{EX} cells ($n = 3,628$; false discovery rate < 0.01 ; labelled in red). **c**, Representative example of a CRC tumour stained by multi-coloured IHC showing co-expression of Ki67, CD8, PD-1 and HAVCR2 in CD8⁺ T_{EX} cells (exemplified by P0413; $n = 2$ patients). Original magnification, $\times 20$. **d**, Volcano plot showing the differentially expressed genes between high-proliferative ($n = 140$) and low-proliferative ($n = 720$) T_{EX} cells. Most of the highly expressed genes in high-proliferative T_{EX} cells are related to cell proliferation. Adjusted

$P < 0.01$; fold change ≥ 2 ; two-sided unpaired limma-moderated *t*-test; Benjamini–Hochberg adjusted P value **e**, Violin plot showing the expression of *TBX21*, *EOMES* and *PDCD1* in each CD8⁺ T cell ($n = 3,628$) cluster and the low-proliferative ($n = 720$) or high-proliferative ($n = 140$) T_{EX} cell subsets. **f**, Most of the clonotypes of high-proliferative T_{EX} cells were also found in low-proliferative T_{EX} cells (top). Each row represents an individual clonotype from one patient. Venn diagram showing overlapped clonal clonotypes (≥ 2 cells) of high- and low-proliferative T_{EX} cells (bottom). **g**, Characteristics of CD8⁺ T_{EX} cells ($n = 3,628$) as defined by the gene expression of a series of transcription factors, checkpoint receptors, and effector molecules. For violin plots in **e** and **g**, colours denote average expression levels; widths denote cell densities.



Extended Data Fig. 7 | Distinct migration capabilities of different CD8⁺ T cell clusters. **a**, Top, chord diagrams showing the distribution of clonotypes in blood, normal mucosa and tumours for different CD8⁺ T cell clusters. CD8⁺ T_{EMRA} cells show remarkable TCR sharing among different tissues. The shadows coloured in transparent yellow, green and orange represent blood, normal and tumour-specific clonotypes, respectively. The bridges coloured in dark green, dark red and dark blue represent clonotypes shared by blood–normal, blood–tumour and normal–tumour, respectively. Bottom, Venn diagram showing the distribution of expanded clonotypes in blood, tumour and normal T_{EMRA} cells. Number of cells contained in each cluster: T_N, n = 174; T_{CM}, n = 169; T_{EMRA}, n = 743; T_{EM}, n = 773; T_{RM}, n = 487; IEL, n = 351; T_{EX}, n = 860.

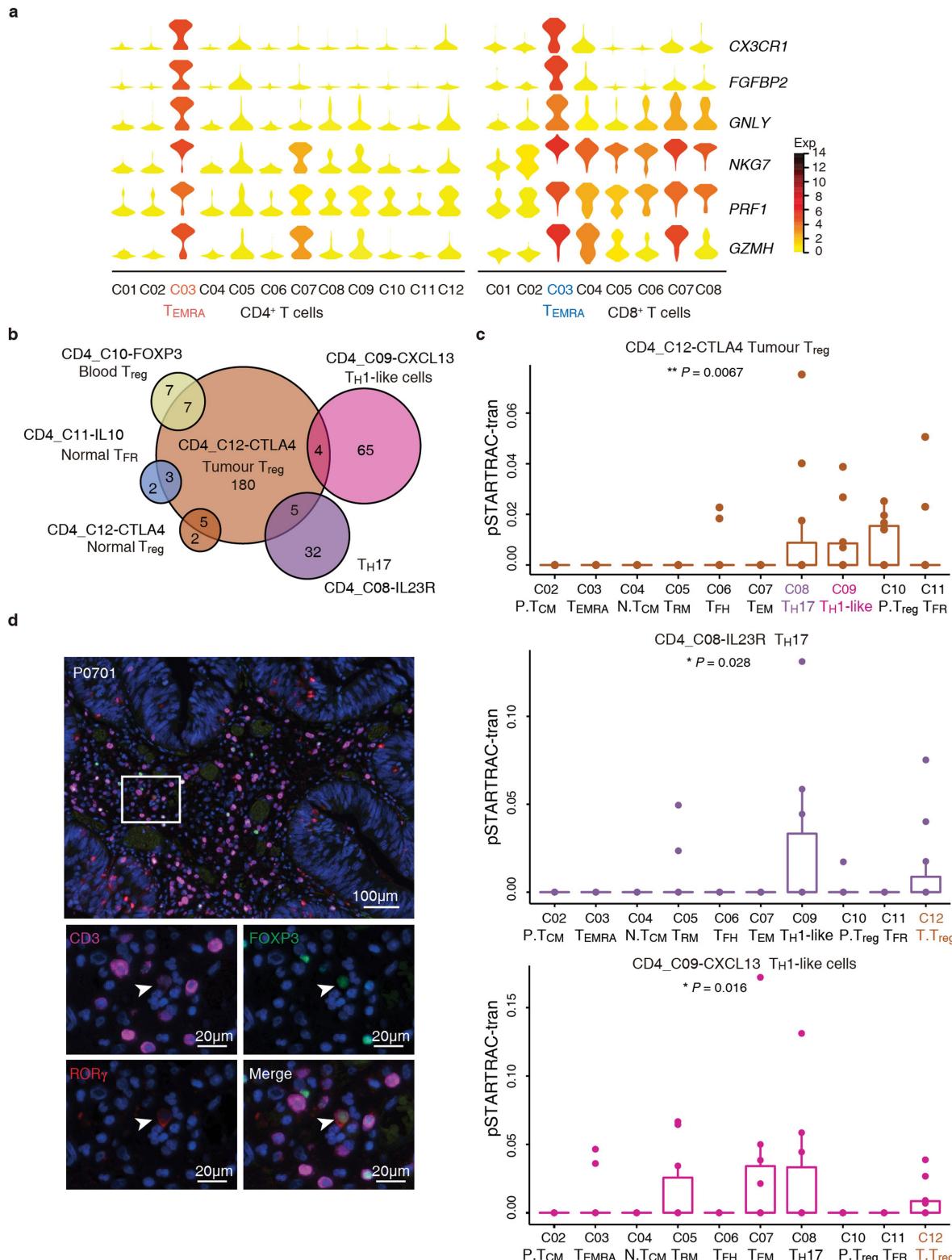
blue represent clonotypes shared by blood–normal, blood–tumour and normal–tumour, respectively. Bottom, Venn diagram showing the distribution of expanded clonotypes in blood, tumour and normal T_{EMRA} cells. Number of cells contained in each cluster: T_N, n = 174; T_{CM}, n = 169; T_{EMRA}, n = 743; T_{EM}, n = 773; T_{RM}, n = 487; IEL, n = 351; T_{EX}, n = 860.



Extended Data Fig. 8 | See next page for caption.

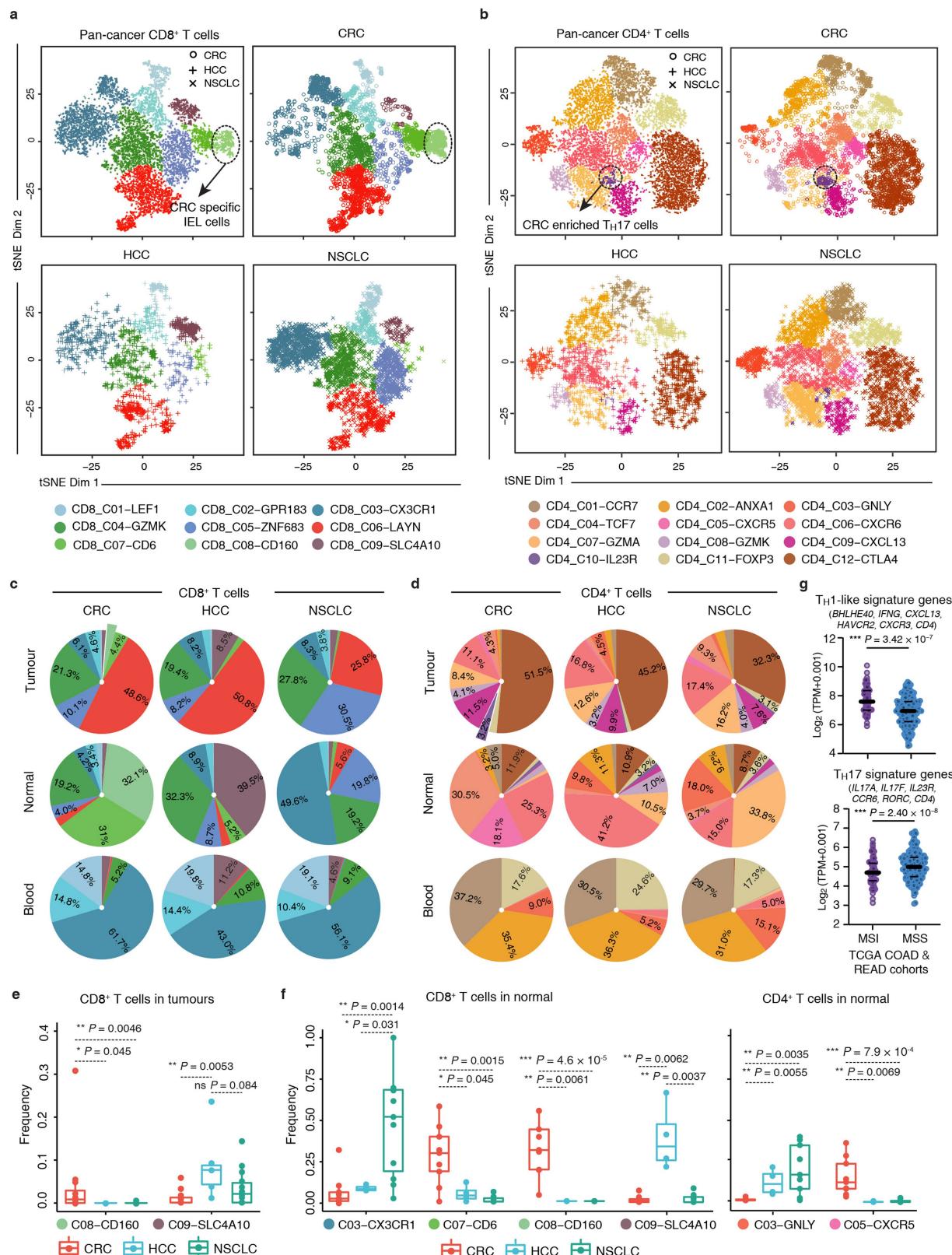
Extended Data Fig. 8 | TCR sharing and state transitions of CD8⁺ T cell clusters implicated by STARTRAC-tran indices. **a**, Pie charts showing the fraction of shared clonotypes with CD8⁺ T_{EM} cells within the other indicated clusters (left). P12 represents merged data of 12 patients with CRC. Bar plots showing the fraction of shared clonotypes of CD8⁺ T_{EM} with other clusters within the CD8⁺ T_{EM}. **b**, pSTARTRAC-tran indices of CD8⁺ T_{CM}, T_{EMRA}, T_{RM}, IEL and T_{EX} cells for each patient (depicted by dots). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, Kruskal-Wallis test. **c**, Potential developmental trajectory of CD8⁺ T cells ($n = 3,557$, excluding MAIT cells) inferred by Monocle2 based on gene expressions. **d**, Frequency of shared clonotypes in CD8⁺ T_{EMRA} cells with various T_{EM} cell subsets in each patient ($n = 12$). **e**, Statistical analysis of tumour T_{EM}

shared TCRs with blood T_{EMRA} and tumour T_{EX} cells based on the number of clonotypes and clonal cells (related to Fig. 1h). *** $P < 0.001$, two-sided Fisher's exact test. **f**, Clonotypes of tumour T_{EM} cells crossing different clusters showing mutually exclusive TCR sharing of tumour T_{EM} cells with blood T_{EMRA} and tumour T_{EX} cells. Each row represents an individual clonotype from one patient. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, two-sided Fisher's exact test (based on the number of clonal cells in each patient). Number of clonal cells analysed in each patient: P1212, $n = 30$; P1228, $n = 27$; P0411, $n = 11$; P0825, $n = 10$; P1012, $n = 7$; P0701, $n = 9$; P0123, $n = 9$; P0215, $n = 17$; P0309, $n = 9$; P0413, $n = 2$; P1207, $n = 7$; P0909, $n = 2$.



Extended Data Fig. 9 | Characterization of CD4⁺ TEMRA and tumour T_{reg} cells by STARTTRAC analysis. **a**, Violin plots showing normalized expression of cytotoxic related molecules in 12 CD4⁺ ($n = 4,902$ cells) and 8 CD8⁺ ($n = 3,628$) T cell clusters. Colours denote mean values; width denotes cell densities. **b**, Venn diagram highlighting common clonotypes ($n_{cell} \geq 2$) shared between tumour T_{reg} and other CD4⁺ T cell clusters.

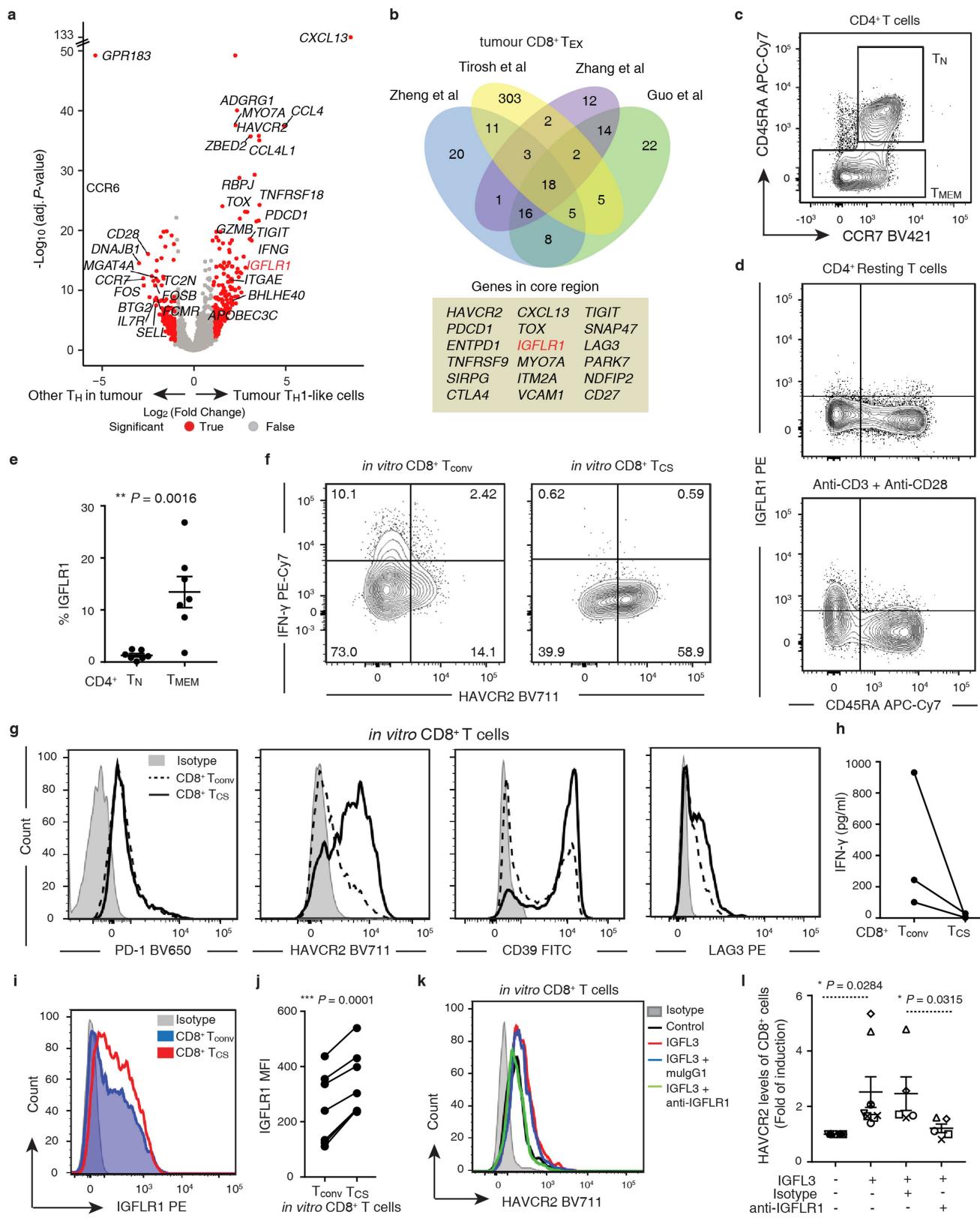
c, Developmental transition of tumour T_{reg} cells, T_{H17} cells and T_{H1}-like cells with other CD4⁺ cells quantified by pSTARTTRAC-tran indices for each patient ($n = 11$). **d**, Representative example of a CRC tumour stained by IHC, with white arrow showing co-expression of CD3, FOXP3 and ROR γ ($n = 2$ patients). Original magnification, $\times 20$.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Comparative analysis of T cells from different cancer indications based on integrated analyses. **a**, *t*-SNE plot of 8,874 single CD8⁺ T cells from this CRC study ($n = 3,632$), and previous HCC⁹ ($n = 1,467$) and NSCLC¹⁰ ($n = 3,775$) studies. Nine CD8⁺ clusters were generated by sscClust based on the integrated dataset. The CRC-specific IEL cells (CD8_C06-CD160) are highlighted. **b**, *t*-SNE plot of 12,635 single CD4⁺ T cells from this CRC study ($n = 4,929$), and previous HCC⁹ ($n = 2,472$) and NSCLC¹⁰ ($n = 5,234$) studies. The CRC-enriched T_H17 cells (CD4_C10-IL23R) are highlighted. Each dot represents one single cell coloured by clusters and shaped by tumour types in **a** and **b**. **c**, Composition of different CD8⁺ T cells in each tumour type by different tissue origins. CD8⁺ T cell clusters with frequencies below 3% are not labelled. **d**, Composition of different CD4⁺ T cells in each tumour type

by different tissue origins. CD4⁺ T cell clusters with frequencies below 3% are not labelled. **e**, Comparison of the fractions of CD8⁺ IEL (CD8_C08-CD160) and MAIT (CD8_C09-SLC4A10) cells in tumours from patients with CRC ($n = 12$), HCC ($n = 5$) and NSCLC ($n = 14$). **f**, Comparison of the fractions of different CD8⁺ T cells and CD4⁺ T cells in control tissues from patients with CRC ($n = 12$), HCC ($n = 5$) and NSCLC ($n = 14$). **g**, Validation of the enrichments of CXCL13⁺BHLHE40⁺ T_H1-like cells in patients with MSI-H CRC ($n = 62$) and T_H17 cells in patients with MSS CRC ($n = 286$) in the TCGA COAD and READ cohorts by comparison of the indicated signature gene expression. Centre lines denote the median, top and bottom lines denote the 25th and 75th percentiles. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; two-sided Wilcoxon test (**e-g**).



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | IGFLR1 expression in activated CD4⁺ T cells and exhausted CD8⁺ T cells. **a**, Volcano plot showing differentially expressed genes between tumour CXCL13⁺BHLHE40⁺ T_{H1}-like T cells ($n = 203$) and other T_H cells in tumours ($n = 723$; Supplementary Table 10). Adjusted $P < 0.01$ (two-sided unpaired limma-moderated t -test; Benjamini–Hochberg adjusted P value) and fold change ≥ 2 . **b**, Venn diagram showing the overlap of tumour CD8⁺ exhaustion-related genes identified in this study ($n = 68$, Supplementary Table 11) with those from previous melanoma⁸ ($n = 349$), HCC⁹ ($n = 82$) and NSCLC¹⁰ ($n = 90$) studies. The detailed overlaps of CD8⁺ exhaustion-related genes in different cancer types are in Supplementary Table 11. $P < 2.2 \times 10^{-16}$, hypergeometric test. **c**, CD4⁺ naive (T_N) and memory (T_{MEM}) T cells were gated as CD45RA⁺CCR7⁺ and CD45RA⁻CCR7^{+/-} cells by FACS. **d**, FACS plots of IGFLR1 expression in activated CD4⁺ T cells ($n = 6$ donors, $n = 3$ independent experiments). **e**, Quantification of IGFLR1 expression levels from **d** as a percentage of IGFLR1⁺ T_N or T_{MEM} CD4⁺ subsets under suboptimal activation conditions ($n = 7$). Each symbol represents a donor with mean \pm s.e.m. shown (**e**, **l**). **f**, Representative

FACS plots for HAVCR2 and IFN γ expression levels in CD8⁺ T_{conv} (activated by anti-CD3 plus anti-CD28) and T_{CS} cells (in vitro chronically stimulated exhausted CD8⁺ T cells from corresponding individuals). Numbers in quadrants indicate the percentage of positive cells ($n = 5$ donors, $n = 2$ independent experiments). **g**, Representative histograms of PD-1, HAVCR2 ($n = 8$ donors, $n = 3$ independent experiments), CD39 and LAG3 ($n = 4$ donors, $n = 2$ independent experiments) expression levels in CD8⁺ T_{conv} and T_{CS} cells. **h**, Quantification of IFN γ levels produced by CD8⁺ T_{conv} and T_{EX} cells from **g** of three donors. **i**, Representative histogram of IGFLR1 expression levels in CD8⁺ T_{conv} and T_{CS} cells. **j**, Expression levels of IGFLR1 in activated CD8⁺ T_{conv} and T_{CS} cells determined by FACS (MFI, mean fluorescent intensity; $n = 6$ donors, $n = 4$ independent experiments). **k**, Representative histograms of HAVCR2 expression in T_{CS} cells subjected to re-stimulation with anti-CD3 alone (control) or together with recombinant human IGFL3 as well as indicated antibodies for 2 days ($n = 5$ donors, $n = 3$ independent experiments). **l**, Quantification of HAVCR2 levels from **k**. Two-sided paired Student's t -test (**e**, **j** and **l**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
 - Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
 - Give P values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- Clearly defined error bars
 - State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

No special or proprietary software was used.

Data analysis

The following software was used in this study:

FlowJo v10, InForm Advanced Image Analysis v2.3, GSNAp (version 2014-10-22), TraCeR (version 2015-10-21), GSEA (version 2.2.4), bwa (version 0.7.17), samtools (version 0.1.19), GATK (version 3.8-1-0), picard (version 2.18.9), strelka (version 1.0.14), ADTEx (version 1.0.4), annovar (version 2018-04-16).
R (version 3.5.0) and the additional packages: HTSeqGenie_4.8.0, scran_1.8.2, SC3_1.7.2, Seurat_2.3.2, monocle_2.8.0, Rtsne v0.13, densityClust v0.3, ggplot2_2.2.1, ggpubr_0.1.7, VennDiagram_1.6.20, ks_1.11.2, limma_3.36.2, circlize_0.4.4, beeswarm_0.2.3, Vennerable_3.1.0.9, ape_5.1 and ComplexHeatmap_1.18.1.
Code for sscClust clustering is available on GitHub (<http://github.com/Japrin/sscClust>). Code for STARTRAC analysis is available on GitHub (<https://github.com/Japrin/STARTRAC>). Other ad hoc scripts for analysing data are available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing raw data have been uploaded to the EGA database under the accession number EGAS00001002791, and processed gene expression data can be obtained from GEO dataset under accession number GSE108989. Such data will be publicly released upon acceptance of this manuscript. In addition, we also developed an interactive web server (<http://crc.cancer-pku.cn>) for analysing, visualizing and downloading the single cell data for individual or multiple user-input genes.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. The sample size were chosen according to sample availability and to achieve a sufficient cell number after sorting for subsequent processing, based on previous experiments within the lab (Zheng et al, Cell, 2017; Guo et al, Nat Med, 2018). Statistical differences provide the rationale for sufficiency of the sample sizes.
Data exclusions	For sequencing data, we excluded low-quality cells according to the criteria we established in our previous single cell studies (Zheng et al, Cell, 2017; Guo et al, Nat Med, 2018), if abnormalities exist in (1) cell library size; (2) the number of expressed genes; (3) the proportion of mitochondrial gene counts; (4) gene expression value of CD3/CD4/CD8. After these filtering, 10,805 of total 11,138 cells remained. The details of such cut-off lines could be checked in the section of Methods.
Replication	All replications were successful, and the detailed information was provided in corresponding figure legends.
Randomization	Not applicable for this study as no treatment strategies are compared.
Blinding	Not applicable since there was no specific grouping.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

FACS antibodies

The antibodies used for FACS sorting of single T cells were anti-human CD3 (UCHT1, eFluor450, eBioscienc 48-0038-42), anti-human CD4 (OKT4, FITC, eBioscience, 11-0048-42), anti-human CD8a (OKT8, APC, eBioscience 17-0086-42), anti-human CD25 (BC96, PE, eBioscience 12-0259-42).

The antibodies used for primary human T cell isolation and in vitro activation were anti-human CD3 (UCHT1) and anti-CD28 (CD28.2), anti-human CD4 (OKT4); anti-human CD8 (RPA-T8), anti-human CD45RA (HI100), anti-human CCR7 (G043H7), anti-

human HAVCR2 (F38-2E2), anti-human CD25 (M-A251), anti-human IFN- gamma (B27), anti-human IGFLR1 (905338), mouse IgG1 and were from Biolegend, BD Biosciences, or R&D Systems.

The antibodies used for *in vitro* chronic stimulated CD8+ T cells were anti-human CD8 (RPA-T8), anti-human HAVCR2/Tim-3 (F38-2E2), anti-human PD1 (EH12.2H7), anti-human CD39 (eBioA1), anti-LAG3 (305223H), anti-IGFLR1 (905338) and were from Thermal Fisher Scientific, BD Biosciences, or R&D Systems.

IHC antibodies

The antibodies used for validation of Tc, TH and Treg cells by multi-colour IHC were rabbit anti-human CD3 (SP7, Abcam ab16669, 1/400), rabbit anti-human CD4 (EPR6855, Abcam ab133616, 1/400), mouse anti-human CD8 (144B, Abcam ab14147, 1/500), mouse anti-human FOXP3 (mAbcam22510, Abcam, ab22510, 1/500).

The antibodies used for validation of ROR gamma+ Treg cells by multi-colour IHC were rabbit anti-human CD3 (SP7, Abcam ab16669, 1/100), rabbit anti-human ROR gamma (Abcam ab219496, 1/50), and mouse anti-human FOXP3 (mAbcam22510 Abcam ab22510, 1/100).

The antibodies used for validation of proliferative CD8+ Tex cells by multi-colour IHC were rabbit anti-human HAVCR2/TIM-3 (D5D5R, Cell Signaling 45208, 1/100), mouse anti-human PD1 (NAT105, Abcam ab52587, 1/200), mouse anti-human CD8 (144B, Abcam ab17147, 1/200), and mouse anti-human Ki67 (B126.1, Abcam ab8191, 1/200).

The antibodies used for dMMR detection were mouse anti-human MLH1 (ES05, Leica Biosystems PA0610), mouse anti-human MSH2 (25D12, Leica Biosystems PA0048), mouse anti-human MSH6 (PU29, Leica Biosystems PA0597) and mouse anti-human PMS2 (MOR4G, Leica Biosystems).

Validation

All the antibodies used in this study were commercial antibodies and were only used for applications, with validation procedures described on the following sites of the manufacturers:

<https://www.thermofisher.com>; <https://www.bdbsciences.com>; <https://www.biologel.com>; <https://www.rndsystems.com>; <https://www.abcam.com>; <https://www.leicabiosystems.com>; <https://www.cellsignal.com>.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Characteristics of CRC patients which freshly resected tumours, adjacent normal tissues, and peripheral blood were collected: P0701, female, 68, rectum, adenocarcinoma, I, MSS; P0909, male, 45, colon, adenocarcinoma, III B, MSI; P1212, female, 42, colon, adenocarcinoma, II, MSS; P1228, female, 77, colon, adenocarcinoma, II, MSS; P0215, male, 75, colon, adenocarcinoma, IV, MSS; P0411, male, 75, rectum, adenocarcinoma, II B, MSS; P0413, female, 82, colon, adenocarcinoma, III B, MSI; P0825, female, 83, colon, adenocarcinoma, II B, MSI; P0123, female, 65, colon, adenocarcinoma, III B, MSI; P0309, male, 55, rectum, adenocarcinoma, III C, MSS.
Characteristics of CRC patients which freshly resected tumours and peripheral blood were collected: P1012, female, 35, colon, adenocarcinoma, III C, MSS; P1207, female, 66, colon, adenocarcinoma, II, MSS.

Recruitment

Twelve patients who were pathologically diagnosed with CRC were enrolled in this study. None of them were treated with chemotherapy or radiation prior to tumour resection. Detailed information can be found in the section of "Human specimens" in Methods and Supplementary Table 1. There are no biases on the selection of patients.

Flow Cytometry

Plots

Confirm that:

- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Information provided in Methods section

Instrument

BD Aria III and BD LSR-II analyzer

Software

FlowJo v10

Cell population abundance

The abundance of the relevant cell populations, determined by testing the sorted cells again by FACS, reached >99%.

Gating strategy

Information available on Extended Data Fig. 1, Extended Data Fig. 11 and Methods sections.

- Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.