Title: Overexpression of transposable elements underlies immune overdrive and poor clinical outcome in cancer patients

Authors: Xiaoqiang Zhu[1], Hu Fang[1], Kornelia Gladysz[1], Jayne A. Barbour[1], Jason W. H. Wong*[1,2]

Affiliations:

[1]School of Biomedical Sciences, The University of Hong Kong, Pokfulam, Hong Kong SAR

[2] Centre for PanorOmic Sciences, The University of Hong Kong, Pokfulam, Hong Kong SAR

*Corresponding authors: Jason W. H. Wong, E-mail: jwhwong@hku.hk.

One Sentence Summary: Transposable element expression is an independent predictor of immune infiltration and poor clinical outcome in cancer patients.

Abstract:

Increased immune infiltration in tumor tissue is usually associated with improved clinical outcome, but excessive infiltration can lead to worst prognosis. The factors underlying such immune overdrive phenotype remains unknown. Here, we investigate the contribution of transposable element (TE) expression to immune response and clinical outcome in cancer. Using colorectal cancer as a model, we develop a TE expression score, showing that highest scores are predicative of immune overdrive and poor outcome independent of microsatellite instability and tumor mutation burden. TE expression scores from cell lines treated with DNA methyltransferase inhibitors show that TEs are directly responsible for driving excess immune infiltration. A pan-cancer survey of TE expression further identify a subset of kidney cancer patients with similar immune overdrive phenotype and adverse prognosis. Together, our

findings reveal that in cancer, TE expression underlies the immune overdrive phenotype and is an independent predictor of immune infiltration and prognosis.

Introduction:

Transposable elements (TEs) are a major component of the human genome. They are divided into class 1 retrotransposons and class 2 DNA transposons which are further subclassified into subclasses, superfamilies and over 1,000 subfamilies *(1)*. As much as 45% of the human genome is comprised of TEs, with many copies of near identical TE sequences from each subfamily located throughout the genome *(2-4)*. In normal somatic tissue, TEs are mainly epigenetically silenced *(5, 6)*, however, in cancer, TEs can become reactivated due to DNA hypomethylation *(7)*, resulting in the transcription of retrotransposons into RNA or direct transposition of DNA transposons. One potential consequence of the reactivation of TEs is to stimulate the immune system via viral mimicry *(8, 9)*. For instance, upon the treatment of DNA methyltransferase (DNMT) inhibitors, one type of TE, the human endogenous retrovirus (hERV) was shown to be reactivated and was accompanied by the up-regulation of viral defense pathways in ovarian *(10)* and colorectal *(8)* cancer cells. Several hERVs such as LTR21B, MER57F, HERVL74-int, were shown to be positively associated with immune infiltration (e.g. CD8 + T cells expression) in multiple cancer types *(11)*. Recently, it has been shown that hERVs can serve as tumor antigen signals and some specific hERVs are of prognostic value and predictive of response of immunotherapy *(9).* For instance, ERV3-2 was found to be correlated with immune checkpoint activation across 11 cancer types *(12).* Additionally, hERV 4700 derived epitope may function as a target by which anti-PD1 could trigger anti-tumor immunity in kidney renal clear cell carcinoma (KIRC) *(13).* Apart from hERVs, other classes of TEs could also be the potential source of immunogenic peptides from tumor cells such as long

interspersed nuclear elements (LINE), short interspersed nuclear elements (SINE) and SINE-VNTE-Alu (SVA) *(11).* Together, these results have demonstrated the critical roles of TEs in anti-tumor immunity. Nonetheless, how TE reactivation impacts cancer progression and clinical outcome still remains unclear.

Patients whose cancer have higher immune cell infiltration tend have better prognosis. For instance, Immunoscore has been developed based on the density of CD3+ and cytotoxic CD8+ T cells in the tumor and the invasive margin in colorectal cancer (CRC) *(14),* and has been shown to have prognostic value superior to American Joint Committee on Cancer (AJCC) stage classification *(15).* Intriguingly, a recent study identified a high risk subgroup of CRC patients with high tumor immune infiltration as indicated by high CD8A and CD274 gene expression *(16).* Termed "immune overdrive" signature, these patients' cancer included both microsatellite instability (MSI) and stable (MSS) status, increased TGF-β activation and overexpression of immune response and checkpoint genes. While whether such patients are likely to benefit from immune checkpoint inhibitors therapy remain to be evaluated, the underlying factor behind this phenotype and whether a similar immune overdrive phenotype exists in other cancer types remains unknown.

Given the recent evidence for the role of TEs in triggering cancer immune response, in this study, we set about investigating whether TE expression may be the underlying mechanism of the immune overdrive phenotype observed in CRC. To do this, we quantified the expression of over 1,000 subfamilies of TEs at the RNA level in The Cancer Genome Atlas (TCGA) CRC cohort. We identified a nine TE expression signature that classified CRC patients into four groups with distinctive prognosis. The group with the highest TE score was characterised by markers of immune overdrive and had the poorest prognosis, while the upper intermediate TE score group is characterised by moderately elevated immune marker expression had the

best prognosis – mirroring the risk profile reported by Fakih et al *(16).* Significantly, our TE score was predicative of prognosis independent of MSI status and tumour mutation burden (TMB). To demonstrate that TE expression is directly linked to the immune overdrive phenotype, we compared immune response and gene expression pathways in high TE score samples using data from cell lines treated with DNA demethylating agents. Finally, a pan-cancer analysis of our TE expression signature uncovered a similar immune overdrive subgroup in high risk KIRC patients, validating the prognostic significance of TE expression signature across cancer types.

Results:

Identification of TEs associated with survival and immune activation in CRC

To establish whether TEs are associated with CRC prognosis and immune activation, we first quantified the TE expression landscape in CRC by applying the recently developed "REdiscoverTE" pipeline *(11)* on RNA sequencing data from TCGA (Fig. S1A). In brief, the pipeline quantifies the number of reads mapping to each TE subfamily without uniquely identifying unique instances in the genome. Our down-stream analysis was focused on 1,072 TE subfamilies which were classified into six classes including LTR (long terminal repeats), DNA, LINE, SINE, Satellite and Retroposon (Fig. S1B). The expression pattern of these six classes is shown in Fig. 1A, indicating that Retroposon and SINE had higher expression following by LINE and LTR while Satellite and DNA had lowest expression *(11)*.

To comprehensively identify prognostic TEs, we performed survival analysis on each TE in terms of four survival endpoints, respectively, including overall survival (OS), DSS (disease specific survival), DFI (disease-free interval) and PFI (progression-free interval) *(17).* There were 372 candidate TEs that had survival difference for at least one endpoints (Fig. 1B, Table. S1) with seven significant in all the four endpoints (Fig. S1C-G). Using a permutation test (see

Methods), we estimated the false discover rate (FDR) to be 1.35% (Fig. S1H). Interestingly, almost all of the hazard ratio of the candidate TEs was greater than one indicating that higher TE expression generally contributed to worse survival (Fig. 1C). Similar analysis was performed at family and class level, respectively. Retroposon showed significant differences in terms of three endpoints (Fig. S1I and J). Retroposon includes six subfamilies named SVA-A to SVA-F. Further multivariable Cox regression analysis for each of these Retroposons suggested five of them could be independent predictors of survival for at least one endpoint except for SVA-B (Fig. S1K-M).

Next, to identify immune associated TEs, we evaluated 29 immune indices including 17 immunologically relevant gene sets from ImmPort *(18)*, one overall immune infiltration, termed ImmuneScore calculated by ESTIMATE *(19)* and 11 immune associated genes (Fig. S1N). Gene set variation analysis (GSVA) was performed to estimate expression scores of the 17 immune gene sets. Further correlation analysis between individual TEs and the immune indices indicated that 14 out of the 1,072 TEs had significant positive correlation with at least one immune indice (Spearman's correlation ≥ 0.4, $p < 0.0001$, Fig. 1D and E). The FDR the significant immune-TE associations was estimated to be 0.7% by permutation test (Fig. S1O).

To integrate the results of the prognostic and immune TE associations, we overlapped the 372 survival relevant TEs and the 14 immune positively correlated TEs. By doing so, nine TEs were identified and used for further exploration (Fig. 1F).

Generation of CRC subtypes based on TE score and clinical outcome

Based on the nine TEs identified above, we first explored the relationship between their expression, finding that most of them were positively correlated with each other (Fig. S2A). As such, to generate a combined TE score, we averaged the normalized TE expression (log counts per million mapped reads, logCPM) across the CRC samples. We applied the kaps

algorithm *(20)* to the normalized TE score and identified 4 clusters based on OS (see Methods, Fig. S2B-G). Based on these clusters, samples were classified into four risk groups termed TE clusters from cluster 1 to cluster 4 with increasing TE score (Fig. 2A). Cluster 4 accounted for 9% (n = 51) of the total cohort while cluster 3 for 8% (n = 47), cluster 2 for 19% (n = 113) and cluster 1 for 64% (n = 379). We found that there were significant differences among these four clusters in terms of some molecular features (Fig. 2B-E, Fig. S2H, Table. S2). Specifically, cluster 4 showed higher fraction of MSI samples (33%) while cluster 1 had lowest fraction (11%) (chi-squared P = 0.0001, Fig. 2B). Cluster 4 also had more samples with CpG island methylator phenotype (CIMP) (30%) (chi-squared P = 0.0228, Fig. 2C). We also observed some overlaps between TE clusters and other two molecular subtypes including consensus molecular subtype (CMS) *(21)* and immune subtypes *(22)*. CMS consists of four subtypes characterized by MSI, CIMP high and immune infiltration for CMS1, epithelial, WNT and MYC signals activation for CMS2, metabolic dysregulation for CMS3, mesenchymal and TGF-β signal activation for CMS4. Predominantly, high fraction of samples in cluster 4 belong to CMS1 subtype (35%) while half of the cluster 1 were CMS2 and 40% of the cluster 2 belong to CMS4 (chi-squared P < 0.0001, Fig. 2D). Moreover, as for immune subtypes derived from pan-cancer analysis, cluster 4 showed higher proportion of IFN-gamma dominant (39%) and inflammatory subtype (10%) while cluster 1 displayed highest proportion of wound healing subtype (85%) followed by cluster 2 (69%), cluster 3 (58%) and cluster 4 (51%) (chi-squared P < 0.0001, Fig. 2E). Lastly, as expected, there was significant survival difference among these four clusters (log-rank P = 0.0035 for OS, P = 0.011 for DSS, P = 0.12 for DFI and P = 0.022 for PFI, Fig. 2F-I). Generally, cluster 4 showed worse survival while cluster 3 showed the most favorable outcomes (cluster 4 versus 3 with HR = 2.36, 95%CI = 0.96-5.80, P = 0.05 for OS, HR = 3.99, 95%CI =1.12-14.16, P = 0.02 for DSS and HR = 3.05, 95%CI = 1.34-6.94, P = 0.005 for

PFI). Notably, cluster 4 had a very poor survival rate after relapse while cluster 3 had superior survival rate after relapse.

TE score is a prognostic and immune infiltration predictor independent of MSI and tumor mutation burden

In CRC, it is well established that patients with tumors that are MSI or high TMB generally have better prognosis. As TE cluster 4 are slightly enriched with MSI tumors compared with other groups (33% versus 19%, 16% and 11% in clusters 3, 2 and 1 respectively), we sought to determine whether TE score is an independent predictor of prognosis and immune infiltration. To do so, we performed multivariable Cox regression on the TE clusters adjusted by clinical features including MSI status. Cluster 4 remained an independent prognostic variable for all the three endpoints with HR against cluster 3 of 3.98 (95%CI: 1.09-14.57, P = 0.037) for OS, 9.52 (95%CI: 1.18-76.54, P = 0.034) for DSS and 2.79 (95%CI:1.07-7.31, P = 0.036) for PFI (Fig. 3A-C). Notably, MSI was not significant for any of the endpoints. Further analysis was performed on MSI and MSS samples separately. It was found that the four clusters were well separated especially in MSI samples (Fig. 3D and E). We also explored the correlation between TE cluster and TMB, defined as non-silent mutation count per megabase (Mb) *(23)*. TMB was poorly correlated with TE score (r = 0.13, Fig. 3F) and there was no difference in terms of TMB among TE clusters (Fig. 3G).

To test if TE cluster could independently predict immune infiltration, we used CD8A, a maker of CD8+ T cells and a T cell-inflamed gene expression profile (GEP) *(24)*. Firstly, we observed strong correlations between TE score with CD8A (r = 0.43, Fig. 3H) and GEP (r = 0.51, Fig. 3I), respectively. A multinomial logistic regression analysis incorporating clinical parameters and MSI or TMB was then performed to predict CD8A and GEP. Cluster 4 was a significant predictor for CD8A and GEP and notably displayed the highest odds ratio (OR) of

6.3 for CD8A (Fig. 3J) and 5.4 for GEP (Fig. 3K). Furthermore, TE score showed much higher OR than TMB with 2.3 versus 1.3 for CD8A (Fig. 3L) and 3.0 versus 1.4 for GEP (Fig. 3M). Our results demonstrate that TE clusters and TE score were independent predictors of immune infiltration independent of MSI and TMB.

Immune overdrive is associated with TE score and expression

Given that the TE signature expression was associated with immune activation, we further explored the differences of tumor immune microenvironment (TME) among TE clusters. We first compared 28 cell fractions including 26 immune cells and two stroma cells (see Table. S3) based on GSVA and found that cluster 4 displayed higher fractions of most of these cells especially for T cells, macrophages and dendritic cells, while cluster 1 showed a lack of immune infiltration (Fig. 4A). Similar results were obtained for 10 cell fractions based on the MCPcounter method (Fig. S3A). Further, cluster 4 displayed the highest gene signatures of immune infiltration followed by cluster 3, 2 and lowest for cluster 1 (Fig. 4B). These signatures included T cell, lymphocyte, leukocyte infiltration signatures, hot tumor signature and tumor associated macrophage (TAM) ratio. Cluster 4 also displayed highest expression of T helper 1 (Th-1) immune response and regulatory genes as well as MCH II and MHC I molecules (Fig. 4B, Fig. S3B and C). Cluster 4 also had higher IFN-γ response rate but exhibited T cell exhaustion (Fig. 4B). Most of the TCR/BCR receptor indices were also highest in cluster 4 and lowest in cluster 1 (Fig. 4C).

Next, we investigated association of the TE clusters with genetic changes and did not found distinctive differences among the clusters except for neo-antigens (Fig. 4D). Specifically, we investigated associations with hotspot mutations by focusing on CRC specific drivers *(25, 26)*. To test if any drivers contributed to TE reactivation, we compared the differences in the frequency of hotspot mutations between cluster 4 and other clusters. Eleven cancer genes

had sufficient mutated samples in each group for further statistical evaluation (see Methods, Fig. S3D). Only TP53 (29% in cluster 4 versus 17% in other three clusters) and BRAF (14% in cluster 4 versus 2.4% in other three clusters) displayed significance (chi-square P = 0.0001 and P = 0.0378 for TP53 and BRAF respectively). As for the copy number changes showing in Fig. S3E, there was no distinctive differences among TE clusters.

To elucidate the molecular phenotype associated with the TE clusters, we further analyzed dysregulated pathways among TE clusters. We found that cluster 4 and 2 displayed stronger immune evasion-associated signatures including TGF-β response, extracellular matrix (ECM) gene expression, VEGF target, Epithelial–mesenchymal transition (EMT) and innate anti-PD1 resistance (IPRES) signatures (Fig. 4E, Fig. S3F). By analysis of 50 MSigDB hallmark gene sets *(27)* and 39 gene program and canonical drug targetable pathways *(28)*, we found 32 out of 50 hallmark gene sets were dysregulated among TE clusters and most of them were upregulated in cluster 4 (Fig. 4F). Similarly, 24 out of 39 drug targetable pathways were significantly different among TE clusters indicating the potential therapy targets for individual clusters such as ALK and PI3K pathways in cluster 4, anti-apoptosis and epidermal growth factor (EGF) signals in cluster 3, plasma membrane signal in cluster 2 and MYC pathway in cluster 1 (Fig. S3G).

Finally, we compared the expression of two markers including CD8A and CD274 which were used to identify immune overdrive by Fakih et al *(16)*. Our results demonstrated that our cluster 4 also displayed higher expression of these two markers (Fig. S3H and I), implying the comparability of immune overdrive identified by TE cluster and these two markers. More importantly, based on the classification of risk groups identified by Fakih et al *(16)*, we found that risk group (IV*) characterized by immune overdrive signature also displayed highest TE score, followed by risk group III* and I/II (Fig. S3J). Together, these results above reflect that

the immune overdrive phenotype of cluster 4 is characterized by TE reactivation, highest immune infiltration, poorest survival and immune evasion activity (e.g. TGF-β signal), and higher expression of immune response and checkpoint genes.

Activation of innate immune response in CRC with high TE score recapitulates TE reactivation by DNMT inhibitors

DNMT inhibitors such as 5-aza-2'-deoxycytidine (5-aza) have been shown to result in TE reactivation through induction of genome-wide DNA demethylation *(2, 5)*. This in turn results in the activation of a range of innate immune response pathways in cell lines *(8, 10)*. To show that TE expression is also directly responsible for triggering the activation of immune response in the CRC patient samples, we compared immune pathway activation in 5-aza treated cells and the CRC TE clusters. Generally, activity of hallmark gene sets was consistent in these three 5-aza treated cell line datasets (Fig. 4F). Interestingly, we observed large overlaps of the significant up-regulated pathways in treated groups of these three data sets and CRC TE cluster 4 (n=21, Fig. 4G, Table. S4). Specifically, these pathways were mainly associated with immune response such as complement, inflammatory response and IFN γ and α response signals. Besides, several oncogenic pathways were also up-regulated including P53 pathway, TGF-β, EMT, apoptosis pathways. Furthermore, by quantifying the TE expression in GSE80137 derived from RNA sequencing, we found that TE score were highest in the 5-aza treated groups compared to control group (Fig. 4H).

As the 5-aza treated samples are all derived from cell lines, to further confirm that TE reactivation signal is mainly derived from tumor cells in bulk CRC tissue rather than the TME, we investigated TE expression in a high-depth single cell RNA sequencing (scRNA-seq) breast cancer dataset *(29)*. Clustering analysis based on the global TE expression profiles after dimensionality reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE) displayed

that tumor and non-tumor cells formed separated clusters (Fig. S3K). Importantly, the proportion of reads mapping to TEs was highest in tumor cells (Kruskal-Wallis test P = 0.0019, Fig. S3L). This is consistent with previous studies that compared TE expression profiles across healthy human tissues, finding that samples from whole blood displayed relatively lower expression of TE compared with other solid tissues, suggesting that hematopoietic cells (e.g. T and B cells) is contributed less to TE expression *(30, 31).* Together, these results suggest that TE reactivation is likely to be the underlying mechanism of the immune overdrive phenotype of cluster 4 and TE expression signal is mainly derived from tumor cells.

TEs trigger intracellular immune response by viral mimicry

Increasing evidence indicated that TE reactivation (e.g. hERVs) could stimulate immune system via intracellular antiviral responses *(8, 10)*. To elucidate a comprehensive pathway regulation behind TE reactivation in CRC, we applied weighted correlation network analysis (WGCNA) *(32)* to find module genes that were associated with TE score. WGCNA is a popular systems biology approach aimed to not only build gene networks but also detect gene modules associated with phenotypic trait (see Methods). Our results suggested that two of the module genes were strongly positively correlated with TE score (r= 0.5 for greenyellow module, r = 0.46 for brown module, Fig. 5A, Fig. S4A-G, Table. S5). There were 39 and 389 genes in greenyellow and brown module, respectively. To determine the function of these modules, we performed GO and KEGG enrichment analysis using these genes (Fig. 5B and C). Genes in the greenyellow module were mainly enriched in pathways associated with innate immune pathways such as defense to virus, type I IFN response, dsRNA sensing, NOD-like, RIG-I-like, MDA-5 and TLRs signals. Genes in brown module were more involved in adaptive immune response such as differentiation, migration and activation of immune cells, IFN γ response, JAK-STAT and NF-kB signals.

We then further compared the expression of some critical genes that might be involved in the response to TE reactivation. As expected, most of these genes were highly expressed in cluster 4 followed by cluster 3 and 2 and lowest in cluster 1 (Fig. 5D). Specifically, these included RIG-I-like and interferon-stimulated genes (ISGs), OASL/2/3, IFNs secretion and production process which have proofed to be associated with TE reactivation. For instance, upon TE reactivation, the secretion and production of IFNs were increased in which IFN α and β indicated type I IFN response. It have been proven that type I IFN response is indicative of the upregulation of intracellular antiviral pathways which are generally induced by dsRNA sensing signals simulated by TE reactivation *(8, 10).* Similarly, some literature evidenced TE suppressors such as APOBECs, ADAR, NOD2, MOV10, MOV10L1, CTCFL, etc were also upregulated in cluster 4. Specifically, SVA-C and F were two Retroposons that are part of the nine TE signature. It has been shown that CTCFL, a germline-specific transcription factor, functions as suppressor of SVA expression by directly binding to and regulate SVA repeats *(33).* Together, these results show that our TE score recapitulate the immune response known to be induced by both endogenous transposable element and exogenous viruses.

Pan-cancer analysis identified immune overdrive phenotype in KIRC

Finally, to confirm if the TE induced immune overdrive phenotype is also present in other cancer types, we examined TE expression in an additional 23 cancer types. We first compared the nine TE score among cancer types. Several cancer types showed higher TE score such as KIRC, Diffuse Large B-cell Lymphoma (DLBC) and Head and Neck squamous cell carcinoma (HNSC) while CRC and Adrenocortical carcinoma (ACC) were generally lower (Fig. S5A, Table. S6). By performing univariable Cox regression analysis on TE score in each cancer type, we found that apart from CRC, another cancer type, KIRC also had significantly increased HR > 1 (HR =1.78, 95%CI = 1.43 - 2.22, P value < 0.0001, Fig. 6A, Table. S7). Several cancers also had

significant HR but less than one including Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), HNSC, Liver hepatocellular carcinoma (LIHC) and Ovarian serous cystadenocarcinoma (OV) (Table. S7). We then further correlated TE score with immune infiltration based on GEP. Overall, TE score correlated well with GEP across the cancer types (r = 0.45, Fig. S5B). Individually, the best correlations were observed in SKCM, HNSCC and CESC followed by CRC and PRAD (Fig. 6B).

As some KIRC appear to also exhibit an immune overdrive phenotype, we carried out further analyses as we had done for CRC. Eight of the nine TEs had similar trends of expression across the samples except for Trigger12A (Fig. 6C). As with CRC, using the kaps algorithm, we identified four KIRC clusters with differing prognostic outcomes (Fig. 6D-F) and enrichment of molecular subtypes (Fig. S5C). Notably, cluster 4 had worse OS (cluster 4 versus 3 with HR = 2.21, 95%CI = 1.29-3.77, P = 0.003, Fig. 6D), DSS (cluster 4 versus 3 with HR = 2.77, 95%CI = 1.38-5.56, P = 0.0027, Fig. 6E), and poor survival rate after relapse (cluster 4 versus 3 with HR = 2.56, 95%CI = 1.38-4.76, P = 0.0021, Fig. 6F). Cluster 2 also had relative short OS while cluster 3 had favorable survival rate after relapse. After adjusted by clinical features, cluster 4 remained significant for all three endpoints and cluster 2 were significant for OS and PFI (Fig. S5D-F).

Consistently, cluster 4 in KIRC also displayed highest immune infiltration and immune evasion phenotypes (Fig. S5G-J). Similar with CRC, there were no distinctive differences for genetic changes among these TE clusters in KIRC (Fig. S5K). Finally, most of the genes involved in the response of TE reactivation had higher expression in cluster 4 showing similar immune regulation patterns with CRC (Fig. S5L-R).

Discussion:

Molecular subtyping based on genomic and transcriptomic data has facilitated improved understanding of molecular features in cancers and has guided targeted strategies in cancer treatment *(34-36)*. For instance, MSI is a critical subtype in CRC which has been associated with high immune infiltration (e.g. $CD8^+$ T cells) *(37-40)* and lower risk of relapse *(41, 42).* Generally, cancer samples with higher immune infiltration had better survival which has been observed in cancers including CRC *(14, 22).* However, an immune overdrive phenotype is also observed in CRC characterized by high immune infiltration but poorest survival *(16)*. Here, we link TE expression to the immune overdrive phenotype in CRC and proposed that reactivation and overexpression of TE might be the potential reason of this phenotype. The immune overdrive phenotype can be reproduced directly using our TE signature. TE cluster 4 is characterized as the immune overdrive phenotype with the highest TE score, poorest survival but also highest immune infiltration. Importantly, a similar immune overdrive subgroup was also present in KIRC which to our knowledge has not previously been reported. Our findings suggest that immune overdrive is mainly contributed by TE reactivation, highlighting the importance of TE reactivation on immune infiltration in cancers.

Nearly 50% of the human genome consists of TEs which are critical for sustaining genomic stability, chromosomal structure and transcriptional regulation *(43, 44)*. Generally, TEs are strictly regulated from early embryonic development and in human differentiated somatic cells mainly through epigenetic mechanisms such as DNA methylation and histone modification *(45, 46)*. TEs are highly mutagenic and normally regarded as harmful because their activation is conflicting to the fitness of the human host *(47)*. Compelling evidence indicates the critical roles of reactivated TEs in cancer development and progression resulting from the loss of TEs suppression *(48-50)*. Generally, epigenetic regulation, especially DNA methylation and histone modification, are the best known mechanisms of TE silencing. Indeed,

studies have demonstrated that epigenetic alterations could lead to carcinogenesis in which TE reactivation might be a potential secondary cause *(51)*. The global loss of methylation can lead to TE reactivation and is often accompanied by the hypermethylation of tumor suppressor genes in cancers *(52)*. For instance, the reactivation of LINE1 caused by DNA hypomethylation has been observed in several cancer types including CRC *(53)*, LIHC *(54)*, and BRCA *(55)*. It has been shown that 5-aza treatment could stimulate innate immune response accompanied by TE reactivation including hERVs and other class of TEs *(8, 10, 11)*. Our analysis based on cell line data also confirmed immune response was stimulated after 5aza treatment. More importantly, three cell lines of GBM derived RNA sequencing obtained significant higher TE score after treatment. Up-regulated hallmark pathways largely overlapped between 5-aza treated cells and TE cluster 4, most of these were immune associated signals as well as oncogenic pathways. These results suggested that the phenotype of TE cluster could be reflected by 5-aza treatment, implying TE reactivation to be the likely cause of immune overdrive. Our TE cluster 4 with highest TE score is comprised of a higher fraction of samples with CIMP, a phenotype characterized by hypermethylation of promoter CpG island sites, but importantly is also associated with genome-wide global hypomethylation *(56)* which may contribute to the reactivation of TEs. Currently, few tools are available to specifically investigate epigenetic regulation on TEs as the repetitive nature of TEs makes it difficult to assign reads derived from next generation sequencing technology to individual TE copies, termed multi-mapping problems *(57)*. Bisulfite sequencing (BS-Seq) and methyl-DNA immunoprecipitation sequencing (MeDIP-Seq) are two well-established methods to measure DNA methylation at high throughput level *(58)*. However, mapping these short reads from bisulfite converted genomic DNA to TEs is still challenging. Although most older TEs with accumulated sufficient nucleotide diversity can be uniquely identified, the insertions of

younger TEs are often indistinguishable from corresponding source elements by using short-read sequencing *(59)*. Given that most of the quantification of DNA methylation in TCGA cohorts were based on the Infinium human methylation 450K or 27K array platform, this limits the proper analysis on the correlation between DNA methylation alterations and TE reactivation in this study. Nevertheless, by using TCGA Illumina 450K array data, Kong et al found that 431 out of 1,007 TEs displayed inverse correlation between TE expression and methylation in at least one out of 10 cancer types while only 13 TEs were significant across multiple cancer types *(11)*. One recent pre-published paper has demonstrated it is possible to accurately assess methylation of TEs by using long-read nanopore sequencing *(59)*. This new technique would facilitate our investigation of TE biology in the future. Moreover, heterochromatin formation has also been proven to regulate TE silencing by cooperating with DNA methylation and small RNAs *(7).* TE-associated nucleosomes are generally methylated at the histone 3 lysine (H3K9) representing signals. To date, chromatin immunoprecipitation sequencing (ChIP-seq) is the best known to determine the chromatin landscape of TEs by mapping reads from ChIP-seq data to consensus sequences of TEs *(60)*. Since no ChIP-seq data is available for TCGA data, we could not compare TE associated chromatin pattern among our TE clusters at tissue level.

Previous studies have drawn the landscape of TE expression across human tissues and indicated that TE expression is much higher in solid tissues compared with in whole blood *(30, 31).* This is in line with our findings based on high depth scRNA-seq dataset of breast cancer that cancer cells are the main contributor of TE expression. Thus, TE expression might reflect cancer-cell intrinsic characteristics *(3)*. Besides, our further analysis revealed that TE score was generally correlated with not only the proportion of reads mapping to TEs but also TE expression at class level in both bulk CRC and KIRC, and scRNA-seq datasets (Fig. S6A-D). This

indicates that TE score might reflect global TE expression. Our TE cluster identified immune overdrive phenotype in CRC and KIRC characterised by highest immune infiltration but worst survival. As for survival difference, TE cluster 4 and 2 had shorter OS than cluster 3 and 1. Cluster 4 had an extreme poor survival rate after relapse, while cluster 3 had superior survival rate after relapse. The independent predictive value of cluster 4 was confirmed after adjusted by clinical factors including MSI status, suggesting the clinical significance of TE cluster as biomarker for prognosis. TE clusters were also separated well in MSI samples, implying that MSI tumours might be further classified based on TE score (Fig. 3D). Although it has been known that higher TMB normally leads immune activation, no significant difference of TMB was observed amongst TE cluster. Importantly, we found two driver genes, TP53 and BRAF, with enriched hotspot mutations in cluster 4 compared with other three clusters. Indeed, studies have reported that p53 can function to restrain TEs and TP53 mutations may potentially cause reactivation of TEs *(61, 62)*. Hence, the enrichment of TP53 mutations in cluster 4 is consistent with these studies. As for the enrichment of BRAF in cluster 4, the possible reason might be the higher fraction of MSI samples in cluster 4. For CRC, although the fraction of MSI status across TE clusters was different, no difference was observed between cluster 4 and 3. These results implied that TE expression is likely mechanism behind immune overdrive rather than TMB or MSI. Importantly, immunosuppressive phenotype was observed in cluster 4 and 2 marked by high expression score of TGF-β response, ECM genes, VEGFA and EMT. Indeed, recent studies have demonstrated that increased TGF-β response signal is the primary mechanism of immune evasion and could lead to worse survival *(63)*. Moreover, a pan-cancer investigation demonstrated that up-regulation of 30 ECM genes was involved with poor prognosis *(64)*. These results to some extent explain why cluster 4 and 2 had worst overall survival.

Innate immune system is essential for pathogen recognition and initiation of protective immune response through the recognition of pathogen associated molecular patterns (PAMPs) by its pattern recognition receptors (PRRs) *(65)*. Nucleic acids including RNA and DNA are critical PAMPs especially for viruses. We found that, upon TE reactivation, some important PRRs of innate immune signals were up-regulated. These included TLRs, RIG-I like receptors, NOD-like receptors, MDA5 (IFIH1), APOBECs, etc. TLRs including TLR3, 7 and 8 are endosome RNA sensors while other PPRs including RIG-I, MDA5, NOD-like receptors belong to cytoplasm RNA sensors *(66-68)*. Indeed, TE-derived RNAs are very prevalent and can form dsRNA in the nucleus. Annealing of these hybirds is relaxed by adenosine (A)-to-inosine (I) editing through ADAR or cytidine (C)-to-uridine (U) deamination editing through APOBEC3s *(3, 69)*. However, the unedited hybirds are prone to bind with RNA sensors and further stimulate downstream immune response by viral mimicry represented by increased IFN response with higher expression of IFN-stimulated genes (ISGs) and IFN regulatory factors such as IRF3 and IRF7. To date, it remains unknown whether individual classes of TEs are prone to activate different PRRs and whether this would cause distinct downstream biological response *(70)*. In addition to triggering of innate immune activation, some other evidence also support that TE reactivation can stimulate immune response through other mechanisms. For instance, some TEs, such as LTR can function as prompters or enhancers of ISGs *(71)*. Moreover, TEs, such as hERVs, have also been shown to provide a source of antigens in KIRC *(13)*.

It has been suggested that CRC with MSI could benefit from immune checkpoint blockade (ICB) therapy. Moreover, epigenetic therapy has been proven to increase tumor immunogenicity and modulate the response to immunotherapy *(8, 72)*. Thus, there are several potential strategies for the treatment of patients amongst TE clusters. MSI tumors in cluster 4 are prone to relapse, therefore, patients in this cluster might benefit from combined

ICB and chemotherapy. Currently, some clinical trials such as ATOMIC, are ongoing with the aim to investigate the efficiency of ICB for MSI-H CRC. Our findings indicate that patients with highest TE score might have higher risk of recurrence and benefit from ICB. Furthermore, our results found that activation of TGF-β, ALK, PI3K pathways were enriched in cluster 4. Therefore, these specific pathways might be used as targets for therapy by combining with ICB. Finally, epigenetic therapy combining with ICB might be suitable for patients in cluster 1 and 2 with relative lower expression TE. More studies and clinical trials will be needed to confirm these strategies.

There are some limitations of our study. This is a retrospective study linking TEs with immune overdrive using TCGA CRC cohort without independent validation. To our knowledge there are currently no other suitable publicly available CRC RNA-seq data with large sample size that can be used to interrogate TE expression. However, since immune overdrive in CRC has been previously reported and we have identified a similar phenotype in KIRC, this means that immune overdrive and its association with TE reactivation is not a cohort specific phenomenon. Future large cohort of CRC or KIRC with comprehensive RNA sequencing and survival information is needed to validate our findings. Although we did not observe immune overdrive phenotype in other cancer types, it does not mean similar phenotype does not exist in those cancers types. Some other TEs may specifically contribute to immune overdrive in other cancer types rather than our nine TEs. Thus, more studies are needed to identify other potential associations between immune infiltration and TE expression in other cancers.

In conclusion, our results suggest that immune overdrive phenotype is not only characterized by high immune infiltration and poor survival, but also overexpression of TEs. Importantly our findings suggest that TE reactivation is the potential cause of immune overdrive in CRC and KRIC. Moreover, patients with relative lower TE expression might be

suitable for strategies combining ICB with epigenetic therapy while patients in cluster 4 with highest TE expression may need more comprehensive treatment and clinical monitoring.

Materials and Methods

TE expression quantification using REdiscoverTE pipeline

We used the REdiscoverTE pipeline to quantify TE expression based on RNA sequencing data as described by Kong et al *(11)*. Briefly, REdiscoverTE uses Salmon (Version, 0.8.2) to perform quantification adjusted by GC content bias and sequence specific bias. The reference transcriptome include RNA transcript sequences from GENCODE release 26 basic *(73)*, RepeatMasker element *(74)* and GENCODE RE-containing introns. TE and gene transcripts were separately quantified and the output of read counts were used for further normalization. TEs were aggregated into TE subfamily, family and class defined by the human Repeatmasker for hg38. TEs were also classified based on their individual genomic locations with respect to genes including exon, intron and intergenic.

TCGA RNA sequencing fasta files were obtained from NCI Genomic Commons (https://www.cancer.gov/tcga). Reads were trimmed firstly using bbduk.sh pipeline *(75)* and then quantified by REdiscoverTE. For normalization, the read counts were aggregated to gene level for gene transcripts and subfamilies for TE, respectively. The two matrixes including gene and TE expression were combined into one as input of edgeR *(76)*. The normalization was conducted using "RLE" algorithm and log2CPM was obtained with prior counts of 5. Then a total of 1,204 TEs were used for further analysis. Additionally, the proportion of reads mapping to TEs was calculated as the total counts mapped to TE divided by the total counts mapped to TE and genes.

Indeed, Kong et al has provided the normalized TE expression in pan cancer level. However, not all the samples were included for some cancer types. Since our analysis started from CRC, we applied REdiscoverTE pipeline only to the CRC cohort. For other cancer types, processed data from Kong et al were used .

Search for TEs associated with survival

To select out the potential TEs associated with survival, we included four survival endpoints for survival analysis in CRC including OS, DSS, DFI and PFI. We performed Univariable Cox regression analysis for individual TEs in each of the four endpoints. The median expression of each TE was used as the cutoff point to separate samples into high or low expression group. A TE was considered as significant if the log-rank p-value was less than 0.05. To estimate the FDR of the candidate TEs, we shuffled samples and performed survival analysis for 100 times. TEs that were significant in at least one endpoint were used for further analysis.

Search for TEs associated with immune activation

To estimate the correlation of TEs with immune activity, we included a total of 29 immune activation indices for analysis consisted of (i) ImmuneScore, which was measured by ESTIMATE algorithm *(19)*; (ii) 17 immunologically relevant signatures reflecting immune pathways derived from ImmPort *(18)*; (iii) mRNA expression of 11 markers representing immune activation or checkpoint pathway up-regulation including *CD8A*, *CD86*, *CD80*, *CTLA4*, *PDCD1LG2*, *CD274*, *PDCD1*, *LAG3*, *TNFRSF14*, *BTLA* and *HAVCR2*. We performed GSVA to obtain the 17 gene signatures score(see method below). We calculate Spearman's correlation coefficients between individual TEs and immune sets. To estimate the FDR of the candidate TEs, we shuffled samples and performed Spearman's correlation analysis 100 times. TEs with correlation coefficients ≥ 0.4 and p-value < 0.0001 were regarded as immunogenic TEs. TEs significantly correlated with at least one immune sets were used for further analysis.

Generation of TE clusters

Using the nine survival and immune associated TEs, we defined subgroups with distinctive OS differences based on kaps algorithm *(20)*. Briefly, this algorithm estimates multi-way split points simultaneously and finds the optimal set of cut off points for a prognostic variable (e.g. OS). A multi-way partition is identified based on subgroup pairwise test and the best number of subgroups is defined by a permutation test. Based on the output, Xk and X1 indicate the overall and worst-pair test statistics. The "adj.Pr (|X1|)" represents the Bonferroni corrected permuted p-value which can be used to select the optimal K. In this study, we input OS as prognostic variable and found that the worst pairs of comparisons were significant with significance level $\alpha = 0.05$ when k = 2 (adjusted P = 0.001) and 4 (adjusted P = 0.04). To comprehensively compare the survival differences among samples, we chose k = 4 and identified four subgroups for downstream analysis.

Estimation of gene signature expression score

We included 70 gene signatures from previous publications associated with cell types in tumour microenvironment (n = 28), immune infiltration (n = 14), immune evasion (n = 4) and IPRES *(n = 24)*. See Table. S3 for details of each signature. The R package GSVA was applied to calculate most of these gene signature expression scores. GSVA is a non-parametric and unsupervised method that can be used to evaluate gene set enrichment based on gene expression profiles derived from microarrays or RNA-seq data *(77)*. It can evaluate the given pathway activity variation by transforming the gene by sample matrix into a gene set by sample matrix. Therefore, it can assess pathway enrichment for individual cases. Importantly, GSVA also provide a method called "ssgsea", which can compute a gene set enrichment score per sample as the normalized difference in empirical cumulative distribution functions of gene expression ranks inside and outside a given gene set. Single sample GSEA (ssGSEA) was

firstly described by Berbie et al *(78)*. To validate the estimation of cell fractions from GSVA, we additionally used Microenvironment Cell Population (MCP)-counter algorithm to estimate the fraction of 10 cell type including endothelial cells, fibroblasts and another eight immune cells populations. This method can robustly quantify the abundance of these cell types based on transcriptomic data for each sample *(79)*. As for the IPRES signature which consisted of 24 gene sets, we firstly calculated the expression score of each gene sets and averaged them as the IPRES score. In terms of GEP, we referred to the previous approach *(24)* based on 18 inflammatory genes including *CCL5, CD27, CD274, CD276, CD8A, CMKLR1, CXCL9, CXCR6, HLA-DQA1, HLA-DRB1, HLA-E, IDO1, LAG3, NKG7, PDCD1LG2, PSMB10, STAT1* and *TIGIT*. The GEP score was estimated by a weighted sum of normalized expression values of these 18 genes adjusted by the expression of 11 housekeeping genes.

Estimation of pathway regulation

To compare the pathway regulation, we included pathways from two resources including hall markers (n= 50) *(27)* and gene program and canonical targetable pathways (n = 39) *(28)*. We calculated the 50 hallmark pathways score using GSVA. As for the 39 targetable pathways, we directly downloaded processed gene expression signature score data from previous TCGA pan-cancer analysis *(28)*. The Kruskal-Wallis test was used to test the difference of these pathways amongst TE clusters.

Somatic mutation and copy number variation analysis

To explore the genetic changes, we performed somatic mutation and copy number variation (CNV) analysis. Specifically, we focused on hotspot mutations of a total of 95 CRC specific driver genes *(25, 26)*. The mutation profile for these 95 genes were downloaded from cBioPortal *(80).* Only 11 driver genes that displayed at least 5 hotspot mutations in either cluster 4 or the other clusters combined were retained for further analysis. Fisher's exact test

was performed to compare the differences among groups. In terms of CNV, Copy Number GISTIC2 level 4 data was download from Broad GDAC Firehose (https://gdac.broadinstitute.org/). Segment file was used as input for GISTIC 2.0 from GenePattern *(81)*. The derived GISTIC score was used for visualization.

Cell line datasets treated with DNA methylation inhibitor

We explored the expression profiles three cell line datasets which were treated with 5-aza. (i) GSE5816, consisted of 11 lung cancer cells, 1 breast cancer and 1 CRC cell line *(82)*. Cells were treated with DMSO as the control group, while low and high groups were treated with 0.1 and 1 uM 5-aza, respectively. Cells were collected after 6 days. (ii) GSE80137, consisted of three GMB cell lines. Cells were treated with 1 uM 5-aza and collected after 3 days *(83)*. (iii) GSE22250, consisted of 7 breast cancer cell lines, treated with 5-aza of 1 uM and collected after 4 days *(84).* The processed transcript data for these datasets were directly accessed under individual GEO accession identifiers and used for analysis. The batch effect was removed based on the type of cell lines in GSE80137, GSE22250 and GSE5816, respectively using the Combat method in the sva R package.

Weighted correlation network analysis (WGCNA)

A WGCNA network was generated for CRC data using the top 5,000 genes with the highest median absolute deviation *(32)*. This approach firstly calculates Pearson's correlation coefficients for all the genes to get the correlation matrix of these genes. A soft threshold of β value can be determined based on a scale-free topology criterion (a β of 10 was chosen in this study). The topological overlap metric (TOM) and dissTOM = 1-TOM are obtained from the resulting adjacent matrix. Then hierarchical clustering is performed based on the blockwiseModules and dynamic tree cutting functions to obtain a cluster dendrogram representing the gene co-expression modules in which the genes are densely interconnected.

Individual gene modules are marked using different colors while grey module indicates these genes are not assigned into any modules. To link the association of modules to TE expression, a module-trait association was performed between gene module and the phenotype (TE score). The gene significance (GS) indicates the absolute value of the association between the expression profile and phenotype while the module membership (MM) represents the correlation between the expression profile and each module. Finally, to explore the biological function behind TE expression, module genes from brown and greenyellow modules were used for GO and KEGG pathway enrichment analyses as these two modules correlated well with TE score.

Single cell RNA sequencing data processing

Raw data was downloaded under the SRA accession number "SRP066982" which consists of 563 fastq files of single cells derived from breast cancer. Only 515 of these cells with high quality and annotation of cell types, as indicated by the original paper was used for further analysis *(29)*. The REdiscoverTE pipeline was applied to these data to quantify TE expression as with the CRC cohort.

Statistics analysis

All statistical analyses were carried out using the program R. Enumeration data were examined by Chi-square test or Fisher's exact test. The comparisons among multiple groups were performed by nonparametric Kruskal-Wallis test. Survival analysis was evaluated by the Kaplan–Meier survival curve and the Log-rank test. Differences were considered significant with a value of $P < 0.05$ unless otherwise stated.

Supplementary Materials

Fig. S1. Screening of candidate TEs.

Fig. S2. Clinical and molecular comparison among TE clusters.

Fig. S3. Comparison among TE cluster in terms of immune overdrive.

Fig. S4. Construction of co-expression module using WGCNA.

Fig. S5. Pan cancer analysis of TE score.

Fig. S6. Correlation between TE score and global TE expression.

Table. S1. Results of screening TEs associated with survival.

Table. S2. summarized clinical data of CRC.

Table. S3. Gene signatures used in this study.

Table. S4. Overlapped significant pathways among 5-aza treated datasets and TE cluster.

Table. S5. Two module gene list derived from WGCNA.

Table. S6. TE score of pan cancer.

Table. S7. Univariable Cox regression analysis of TE score across 24 cancer types.

References and Notes:

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome biology.* 2018;19(1):199.
2. Gorbunova V, Boeke JD, Helfand SL, and Sedivy JM. Human Genomics. Sleeping dogs of the genome. *Science (New York, NY).* 2014;346(6214):1187-8.
3. Burns KH. Our Conflict with Transposable Elements and Its Implications for Human Disease. *Annual review of pathology.* 2020;15:51-70.
4. Deniz Ö, Frost JM, and Branco MR. Regulation of transposable elements by DNA modifications. *Nature reviews Genetics.* 2019;20(7):417-31.
5. Anwar SL, Wulaningsih W, and Lehmann U. Transposable Elements in Human Cancer: Causes and Consequences of Deregulation. *International journal of molecular sciences.* 2017;18(5).
6. He J, Fu X, Zhang M, He F, Li W, Abdul MM, et al. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nature communications.* 2019;10(1):34.
7. Slotkin RK, and Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nature reviews Genetics.* 2007;8(4):272-85.
8. Roulois D, Loo Yau H, Singhania R, Wang Y, Danesh A, Shen SY, et al. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell.* 2015;162(5):961-73.

9. Hegde PS, and Chen DS. Top 10 Challenges in Cancer Immunotherapy. *Immunity.* 2020;52(1):17-35.
10. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell.* 2015;162(5):974-86.
11. Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nature communications.* 2019;10(1):5228.
12. Panda A, de Cubas AA, Stein M, Riedlinger G, Kra J, Mayer T, et al. Endogenous retrovirus expression is associated with response to immune checkpoint blockade in clear cell renal cell carcinoma. *JCI insight.* 2018;3(16).
13. Smith CC, Beckermann KE, Bortone DS, De Cubas AA, Bixby LM, Lee SJ, et al. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *The Journal of clinical investigation.* 2018;128(11):4804-20.
14. Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet (London, England).* 2018;391(10135):2128-39.
15. Galon J, Mlecnik B, Bindea G, Angell HK, Berger A, Lagorce C, et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *The Journal of pathology.* 2014;232(2):199-209.
16. Fakih M, Ouyang C, Wang C, Tu TY, Gozo MC, Cho M, et al. Immune overdrive signature in colorectal tumor subset predicts poor clinical outcome. *The Journal of clinical investigation.* 2019;129(10):4464-76.
17. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell.* 2018;173(2):400-16.e11.
18. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific data.* 2018;5:180015.
19. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications.* 2013;4:2612.
20. Eo S-H, Kang HJ, Hong S-M, and Cho H. K-adaptive partitioning for survival data, with an application to cancer staging. *arXiv preprint arXiv:13064615.* 2013.
21. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine.* 2015;21(11):1350-6.
22. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Yang T-HO, et al. The immune landscape of cancer. *Immunity.* 2018;48(4):812-30. e14 %@ 1074-7613.
23. Garcia-Garijo A, Fajardo CA, and Gros A. Determinants for Neoantigen Identification. *Frontiers in immunology.* 2019;10:1392.
24. Cristescu R, Mogg R, Ayers M, Albright A, Murphy E, Yearley J, et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science (New York, NY).* 2018;362(6411).
25. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods.* 2013;10(11):1081-2 %@ 548-7105.
26. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology.* 2016;34(2):155-63.

27. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, and Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems.* 2015;1(6):417-25.
28. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell.* 2014;158(4):929-44 %@ 0092-8674.
29. Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications.* 2017;8:15081.
30. Bogu GK, Reverter F, Marti-Renom MA, Snyder MP, and Guigó R. Atlas of transcriptionally active transposable elements in human adult tissues. *bioRxiv.* 2019:714212.
31. Larouche J-D, Trofimov A, Hesnard L, Ehx G, Zhao Q, Vincent K, et al. Widespread and tissue-specific expression of endogenous retroelements in human somatic tissues. *Genome Medicine.* 2020;12:1-16.
32. Langfelder P, and Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics.* 2008;9(1):559 %@ 1471-2105.
33. Pugacheva EM, Teplyakov E, Wu Q, Li J, Chen C, Meng C, et al. The cancer-associated CTCFL/BORIS protein targets multiple classes of genomic repeats, with a distinct binding and functional preference for humanoid-specific SVA transposable elements. *Epigenetics & chromatin.* 2016;9(1):35.
34. Roszik J, and Subbiah V. Mining Public Databases for Precision Oncology. *Trends in cancer.* 2018;4(7):463-5.
35. Yu Y. Molecular classification and precision therapy of cancer: immune checkpoint inhibitors. *Frontiers of medicine.* 2018;12(2):229-35.
36. Malone ER, Oliva M, Sabatini PJB, Stockley TL, and Siu LL. Molecular profiling for precision cancer therapies. *Genome Med.* 2020;12(1):8.
37. Boland CR, and Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology.* 2010;138(6):2073-87.e3.
38. Gryfe R, Kim H, Hsieh ET, Aronson MD, Holowaty EJ, Bull SB, et al. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *The New England journal of medicine.* 2000;342(2):69-77.
39. Popat S, Hubner R, and Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2005;23(3):609-18.
40. Maby P, Galon J, and Latouche JB. Frameshift mutations, neoantigens and tumor-specific CD8(+) T cells in microsatellite unstable colorectal cancers. *Oncoimmunology.* 2016;5(5):e1115943.
41. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *The New England journal of medicine.* 2003;349(3):247-57.
42. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, Hamilton SR, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2010;28(20):3219-26.
43. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860-921.

44. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, 3rd, et al. Landscape of somatic retrotransposition in human cancers. *Science (New York, NY).* 2012;337(6097):967-71.
45. Rebollo R, Romanish MT, and Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annual review of genetics.* 2012;46:21-42.
46. Gerdes P, Richardson SR, Mager DL, and Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. *Genome biology.* 2016;17:100.
47. Labrador M, and Corces VG. Transposable element-host interactions: regulation of insertion and excision. *Annual review of genetics.* 1997;31:381-404.
48. Criscione SW, Zhang Y, Thompson W, Sedivy JM, and Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC genomics.* 2014;15:583.
49. Wolff EM, Byun HM, Han HF, Sharma S, Nichols PW, Siegmund KD, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS genetics.* 2010;6(4):e1000917.
50. Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, et al. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *International journal of cancer.* 2009;124(1):81-7.
51. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature genetics.* 2013;45(7):836-41.
52. Esteller M. Epigenetics in cancer. *The New England journal of medicine.* 2008;358(11):1148-59.
53. Swets M, Zaalberg A, Boot A, van Wezel T, Frouws MA, Bastiaannet E, et al. Tumor LINE-1 Methylation Level in Association with Survival of Patients with Stage II Colon Cancer. *International journal of molecular sciences.* 2016;18(1).
54. Harada K, Baba Y, Ishimoto T, Chikamoto A, Kosumi K, Hayashi H, et al. LINE-1 methylation level and patient prognosis in a database of 208 hepatocellular carcinomas. *Annals of surgical oncology.* 2015;22(4):1280-7.
55. Terry MB, Delgado-Cruzata L, Vin-Raviv N, Wu HC, and Santella RM. DNA methylation in white blood cells: association with risk factors in epidemiologic studies. *Epigenetics.* 2011;6(7):828-37.
56. Issa JP. CpG island methylator phenotype in cancer. *Nature reviews Cancer.* 2004;4(12):988-93.
57. Lerat E, Casacuberta J, Chaparro C, and Vieira C. On the Importance to Acknowledge Transposable Elements in Epigenomic Analyses. *Genes.* 2019;10(4).
58. Lea AJ, Vockley CM, Johnston RA, Del Carpio CA, Barreiro LB, Reddy TE, et al. Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife.* 2018;7.
59. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Cheetham SW, et al. Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *bioRxiv.* 2020.
60. Day DS, Luquette LJ, Park PJ, and Kharchenko PV. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome biology.* 2010;11(6):R69.
61. Tiwari B, Jones AE, and Abrams JM. Transposons, p53 and Genome Security. *Trends in genetics : TIG.* 2018;34(11):846-55.

62. Wylie A, Jones AE, and Abrams JM. p53 in the game of transposons. *BioEssays : news and reviews in molecular, cellular and developmental biology.* 2016;38(11):1111-6.
63. Tauriello DVF, Palomo-Ponce S, Stork D, Berenguer-Llergo A, Badia-Ramentol J, Iglesias M, et al. TGFβ drives immune evasion in genetically reconstituted colon cancer metastasis. *Nature.* 2018;554(7693):538-43.
64. Chakravarthy A, Khan L, Bensler NP, Bose P, and De Carvalho DD. TGF-β-associated extracellular matrix genes link cancer-associated fibroblasts to immune evasion and immunotherapy failure. *Nature communications.* 2018;9(1):4692.
65. Chen N, Xia P, Li S, Zhang T, Wang TT, and Zhu J. RNA sensors of the innate immune system and their detection of pathogens. *IUBMB life.* 2017;69(5):297-304.
66. Hayashi F, Means TK, and Luster AD. Toll-like receptors stimulate human neutrophil function. *Blood.* 2003;102(7):2660-9.
67. Schlee M. Master sensors of pathogenic RNA - RIG-I like receptors. *Immunobiology.* 2013;218(11):1322-35.
68. Kell AM, and Gale M, Jr. RIG-I in RNA virus recognition. *Virology.* 2015;479-480:110-21.
69. Schumann GG. APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition. *Biochemical Society transactions.* 2007;35(Pt 3):637-42.
70. Jones PA, Ohtani H, Chakravarthy A, and De Carvalho DD. Epigenetic therapy in immune-oncology. *Nature reviews Cancer.* 2019;19(3):151-61.
71. Chuong EB, Elde NC, and Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (New York, NY).* 2016;351(6277):1083-7.
72. Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nature genetics.* 2017;49(7):1052-60.
73. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research.* 2012;22(9):1760-74.
74. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics.* 2004;Chapter 4:Unit 4.10.
75. Bushnell B. BBDuk: Adapter. *Quality Trimming and Filtering httpssourceforgenetprojectsbbmap.*
76. Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England).* 2010;26(1):139-40.
77. Hänzelmann S, Castelo R, and Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics.* 2013;14(1):7 %@ 1471-2105.
78. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature.* 2009;462(7269):108-12 %@ 1476-4687.
79. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology.* 2016;17(1):218.
80. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling.* 2013;6(269):pl1.
81. Reich M. Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. *GenePattern.*2:500-1.

82. Shames DS, Girard L, Gao B, Sato M, Lewis CM, Shivapurkar N, et al. A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. *PLoS medicine.* 2006;3(12).
83. Shraibman B, Kadosh DM, Barnea E, and Admon A. Human leukocyte antigen (HLA) peptides derived from tumor antigens induced by inhibition of DNA methylation for development of drug-facilitated immunotherapy. *Molecular & Cellular Proteomics.* 2016;15(9):3058-70 %@ 1535-9476.
84. Dedeurwaerder S, Desmedt C, Calonne E, Singhal SK, Haibe-Kains B, Defrance M, et al. DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO molecular medicine.* 2011;3(12):726-41 %@ 1757-4684.

Acknowledgments: Funding: This work is support by seed funding to JWHW from The University of Hong Kong. **Author contributions:** XZ and JWHW conceived this study. HF, KG, and JAB helped to collect the public data. XZ and JWHW conducted statistical analysis. HF, KG, and JAB helped to interpret the results. XZ and JWHW wrote the manuscript. Data used in this study is derived from TCGA database (https://cancergenome.nih.gov). Part of the data was analyzed on Cancer Genomic Cloud (CGC, http://www.cancergenomicscloud.org/). The CGC team helped to build custom application program for the analysis on CGC with pilot funds provided by Seven Bridges Genomics. **Competing interests:** The authors have declared that no conflict of interest exists. **Data availability**: TCGA CRC RNA sequencing data were directly analysed on Cancer Genomics Cloud using the custom pipeline [http://www.cancergenomicscloud.org/]. The processed TE expression for pan-cancer was downloaded from [http://research-pub.gene.com/REdiscoverTEpaper/]. CRC Copy Number GISTIC2 level 4 data was downloaded from Broad GDAC Firehose [http://gdac.broadinstitute.org/runs/analyses_2016_01_28/data/COADREAD/20160128/]. The clinical data, gene program pathways were obtained from [https://xenabrowser.net/]. The TCR/BCR index scores and genetic changes were downloaded from the Supplementary table of the pan-cancer immune landscape paper *(22)*. Three cell line datasets were

downloaded from [https://www.ncbi.nlm.nih.gov/geo/] under the accession number of (i) GSE5816, (ii) GSE80137, and (iii) GSE22250.

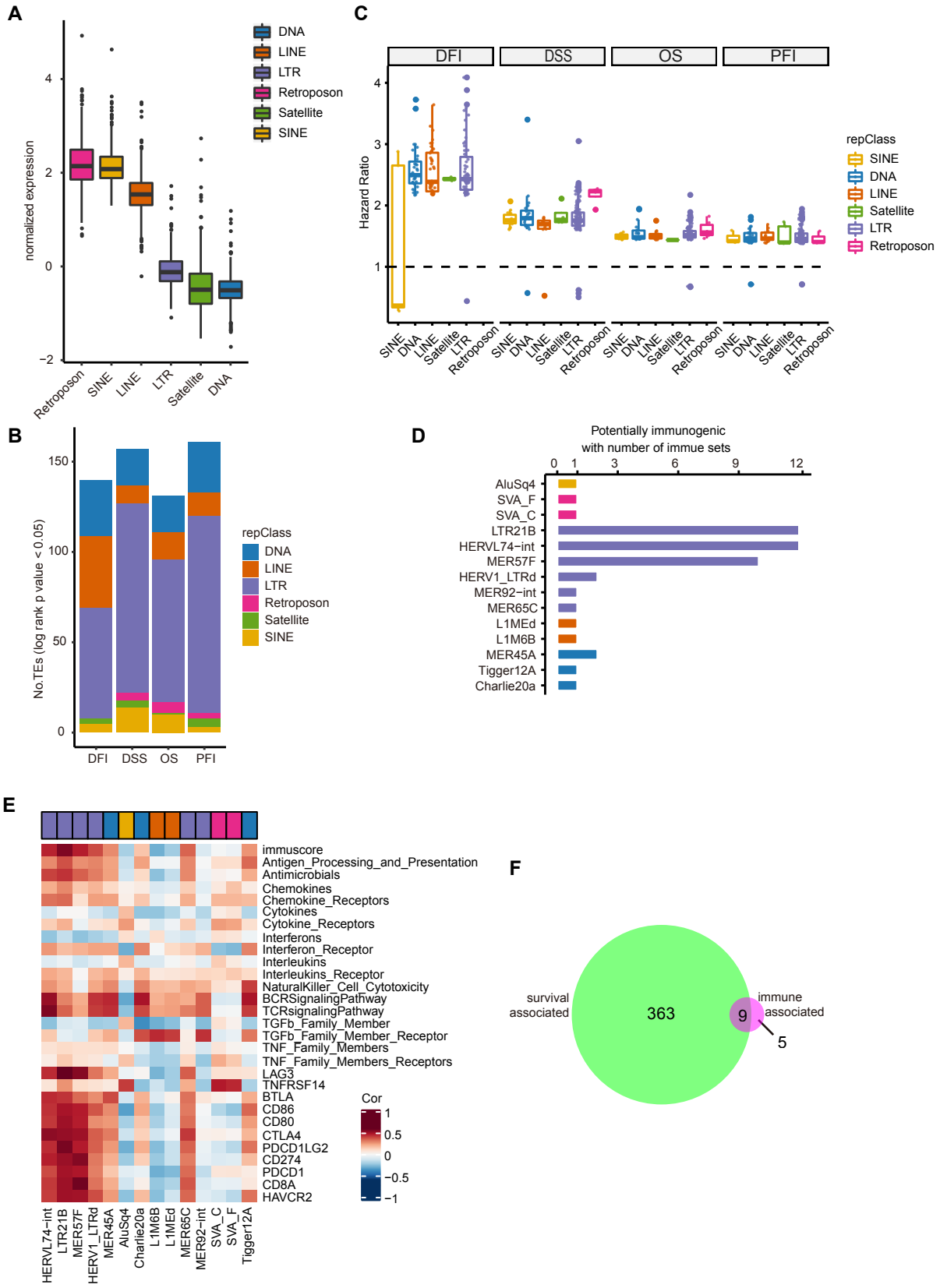


Figure 1. Identification of TEs associated with survival and immune sets in CRC. (A) Normalized TE expression pattern at TE class level including Retroposon, SINE, LINE, LTR, Satellite and DNA. **(B)** Stacked plot showing the number of subfamily TEs with significant log rank p value ($p < 0.05$) for each of the four endpoints including DFI, DSS, OS and PFI. TEs were annotated at class level. **(C)** Distribution of hazard ratios of significant TEs from (B) for each of the four endpoints. TEs were annotated at class level. **(D)** Number of immune sets significantly correlated with each TE (Cor ≥ 0.4, $p < 0.0001$). **(E)** Spearman's correlation between candidate TE expression (n=14) and 29 immune sets. Heatmap colors indicate the correlation coefficient. **(F)** Venn diagram showing 9 TEs overlapped between candidate prognostic TEs and immunogenic TEs.

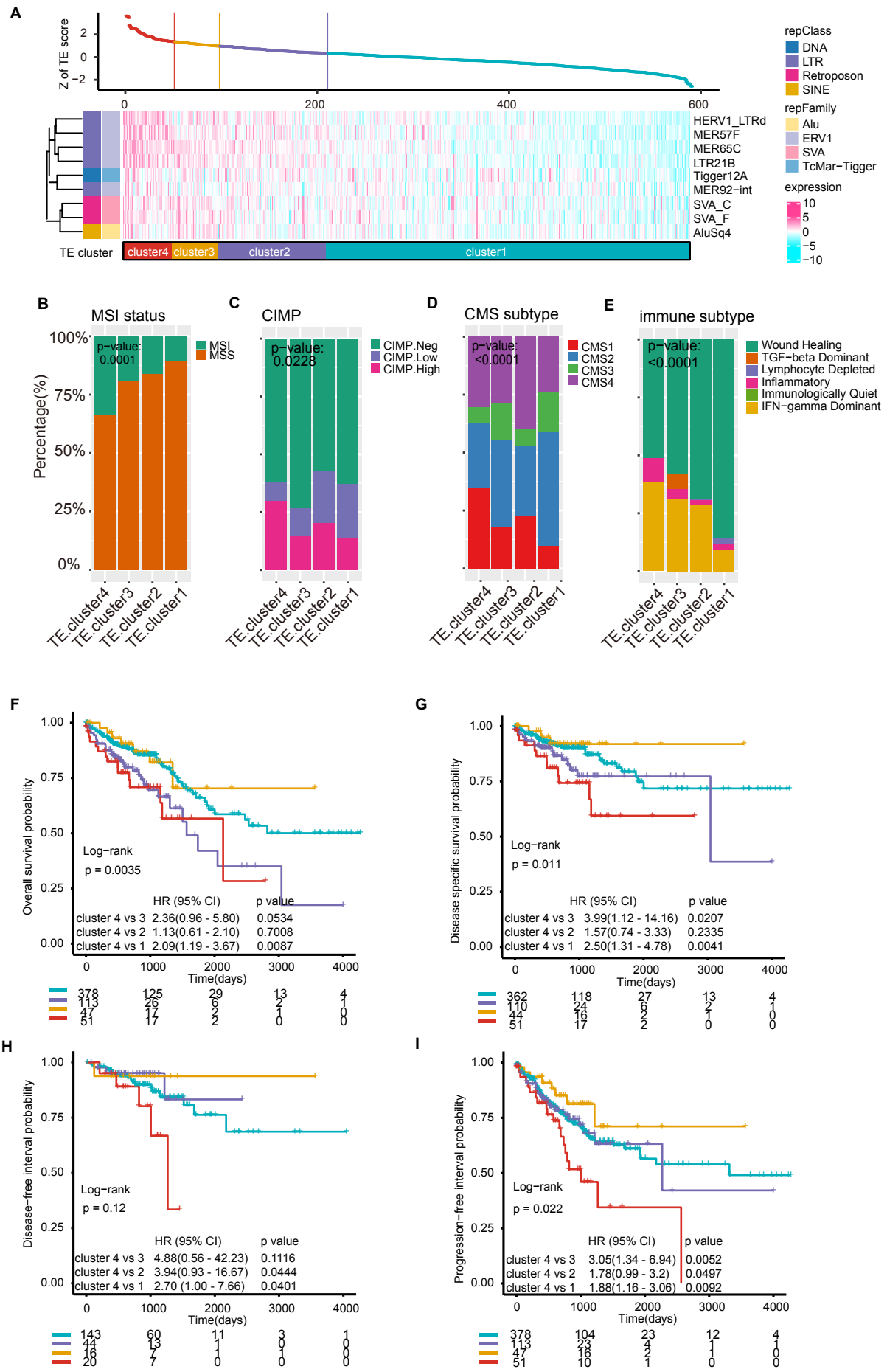


Figure 2. Generation of TE score-based CRC clusters and comparison of molecular association. (A) Top scatter plot showed the TE score in decreasing order from left to right. Bottom heatmap displayed the expression profiles of the 9 TEs across four TE clusters as ordered by the TE score. Each TE was annotated at family and class level, respectively. **(B-E)** Stacked plots showing the fractions of molecular features across four TE clusters including MSI status **(B)**, CIMP **(C)**, CMS subtypes **(D)** and immune subtypes **(E)**. **(F-I)** Prognostic value of four TE clusters with Kaplan-Meier survival analysis for OS (n = 589) **(F)**, DSS (n = 567) **(G)**, DFI (n = 223) **(H)** and PFI (n = 589) **(I)**. The hazard ratios (HR) and 95% confidence intervals (CIs) for pairwise comparisons in univariable analyses (log-rank test) are displayed in each Kaplan-Meier plot. Numbers below the x-axes represent the number of patients at risk at the selected time points. The tick marks on the Kaplan-Meier curves indicated the censored patients.

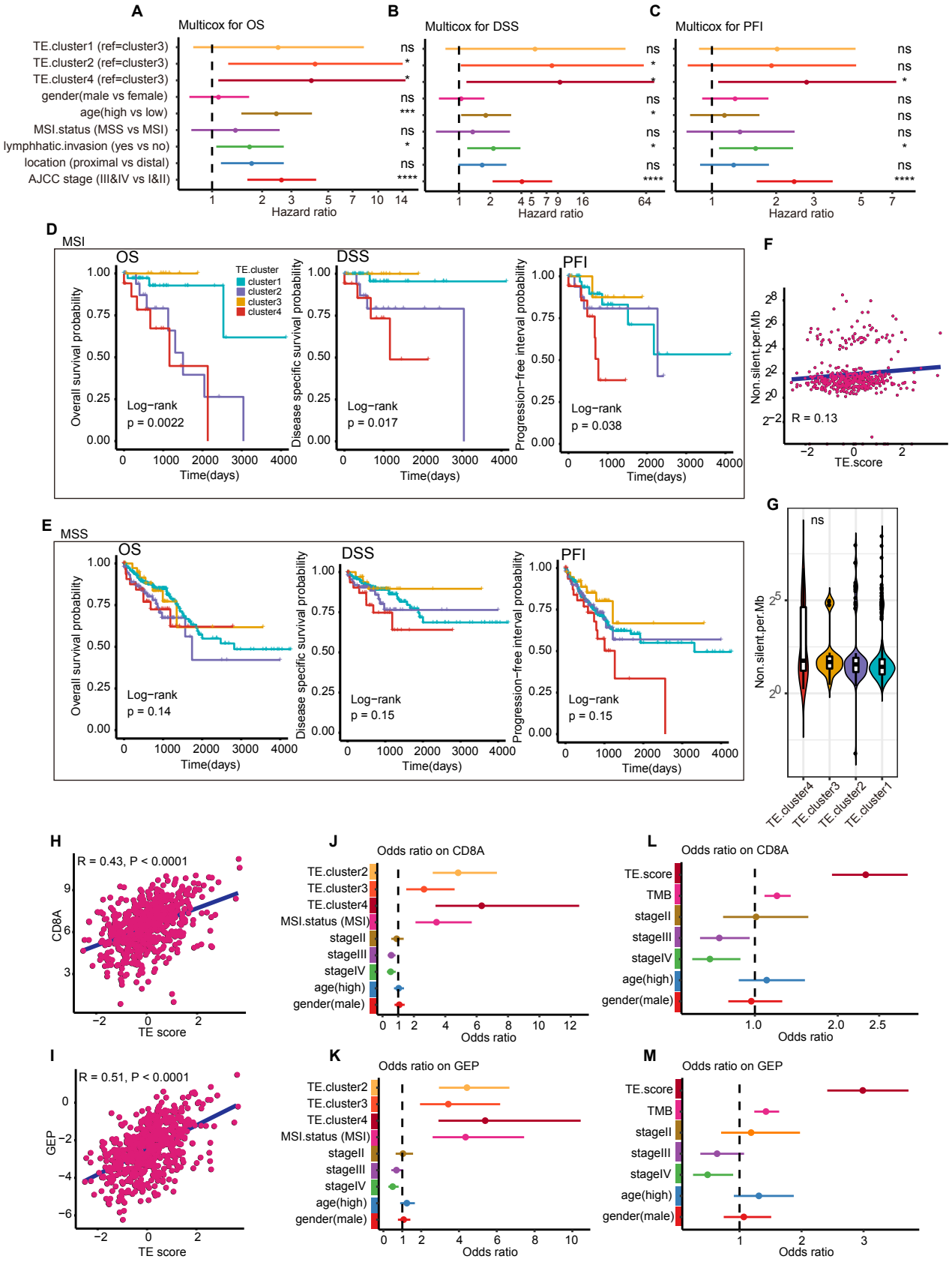


Figure 3. Prognostic value of TE cluster and immune infiltration prediction. (A-C) Forest plots showing multivariable Cox regression analysis of TE cluster adjusted by clinical features for OS **(A)**, DSS **(B)** and PFI **(C)**. All variables were set as categorial variable. Samples with age < 65 was set as age low group and ≥ 65 for high group. Solid dots represent the HR of death and open-ended horizontal lines represent the 95 % confidence intervals (CIs). All p-values were calculated using Cox proportional hazards analysis (ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001). **(D-E)** Prognostic value of four TE clusters with Kaplan-Meier survival analysis in two subgroups separated by MSI status (MSI in **D**, MSS in **E**) for three endpoints, respectively. DFI was excluded because of non-comparable sample size among TE clusters. P-value was calculated using log-rank test. Numbers below the x-axes represent the number of patients at risk at the selected time points. The tick marks on the Kaplan-Meier curves indicate the censored patients. **(F)** Spearman's correlation between normalized TE score and non-silent mutation per Mb. **(G)** Violin plot comparing non-silent mutation per Mb among TE clusters (n.s.: p > 0.05). **(H-I)** Spearman's correlation between normalized TE score and CD8A expression **(H)** and GEP **(I)**, respectively. **(J-K)** Forest plots showing the odds ratio indicating immune infiltration determined by CD8A expression **(J)** and GEP **(K)** using multinomial logistic regression analysis adjusted by MSI status. **(L-M)** Forest plots showing the odds ratio indicating immune infiltration determined by CD8A expression **(L)** and GEP **(M)** using multinomial logistic regression analysis adjusted by TMB. Solid dots represent the adjusted OR and open-ended horizontal lines represent the 95 % confidence intervals (CIs). OR to the right of dashed line (where OR = 1) indicates higher odds of immune infiltration while OR to the left of the dashed line indicates lower odds of immune infiltration.

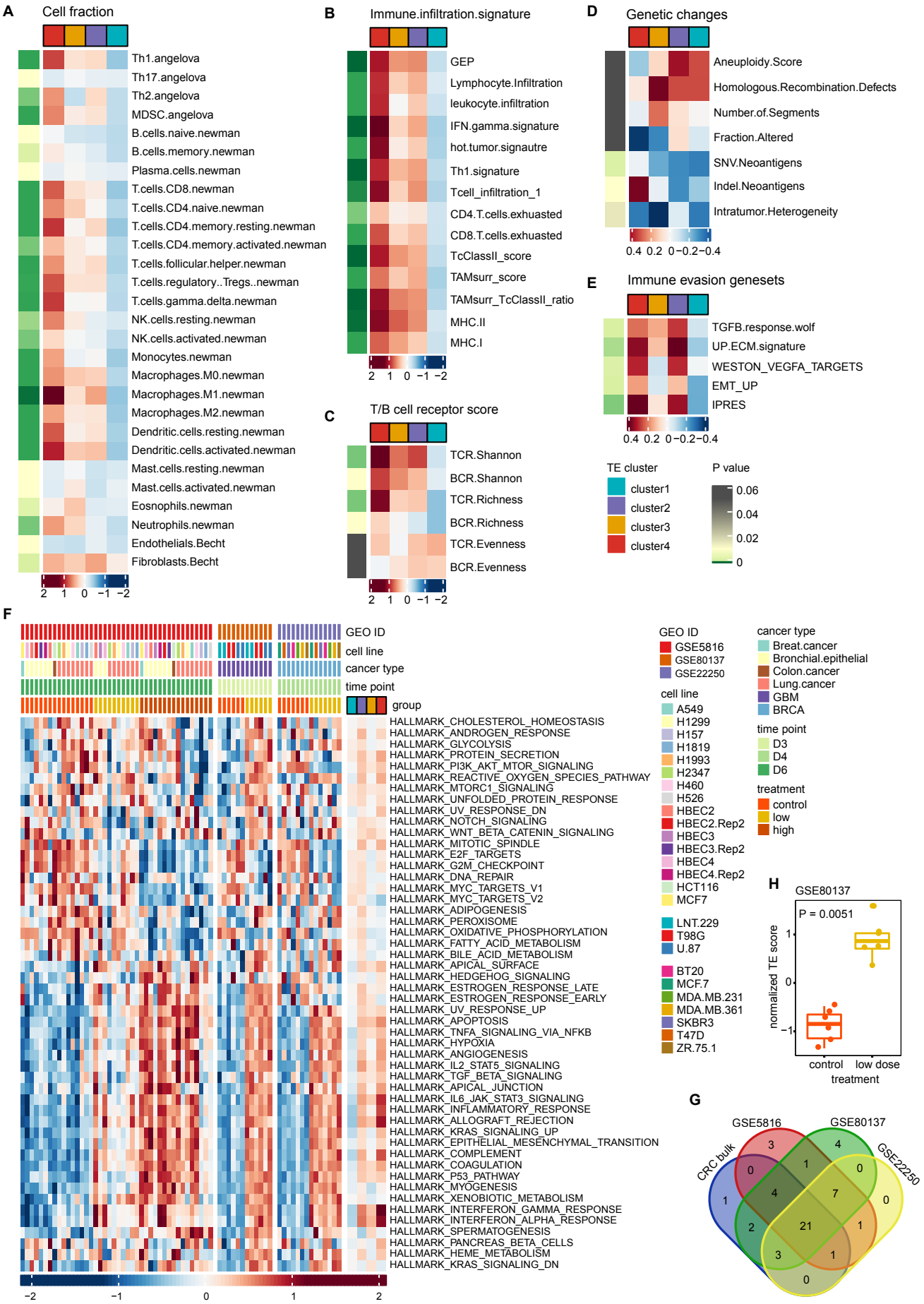


Figure 4. Exploration of immune overdrive phenotype. (A) Gene set variation analysis showing fraction of 28 cell types. **(B)** Gene set variation analysis showing immune infiltration signatures. **(C)** TCR/BCR indexes comparison among TE clusters. **(D)** Genetic changes comparison among TE clusters. **(E)** Gene set variation analysis showing immune evasion signatures. P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown. **(F)** Heatmap showing 50 hallmark gene sets score based on gene set variation analysis in three 5-aza treated cell line datasets across multiple cell lines and CRC TE clusters. **(G)** Venn diagram showing the overlapped significant pathways among cell line datasets and bulk CRC. **(H)** Comparison of TE score between treated and control groups in GSE80137.

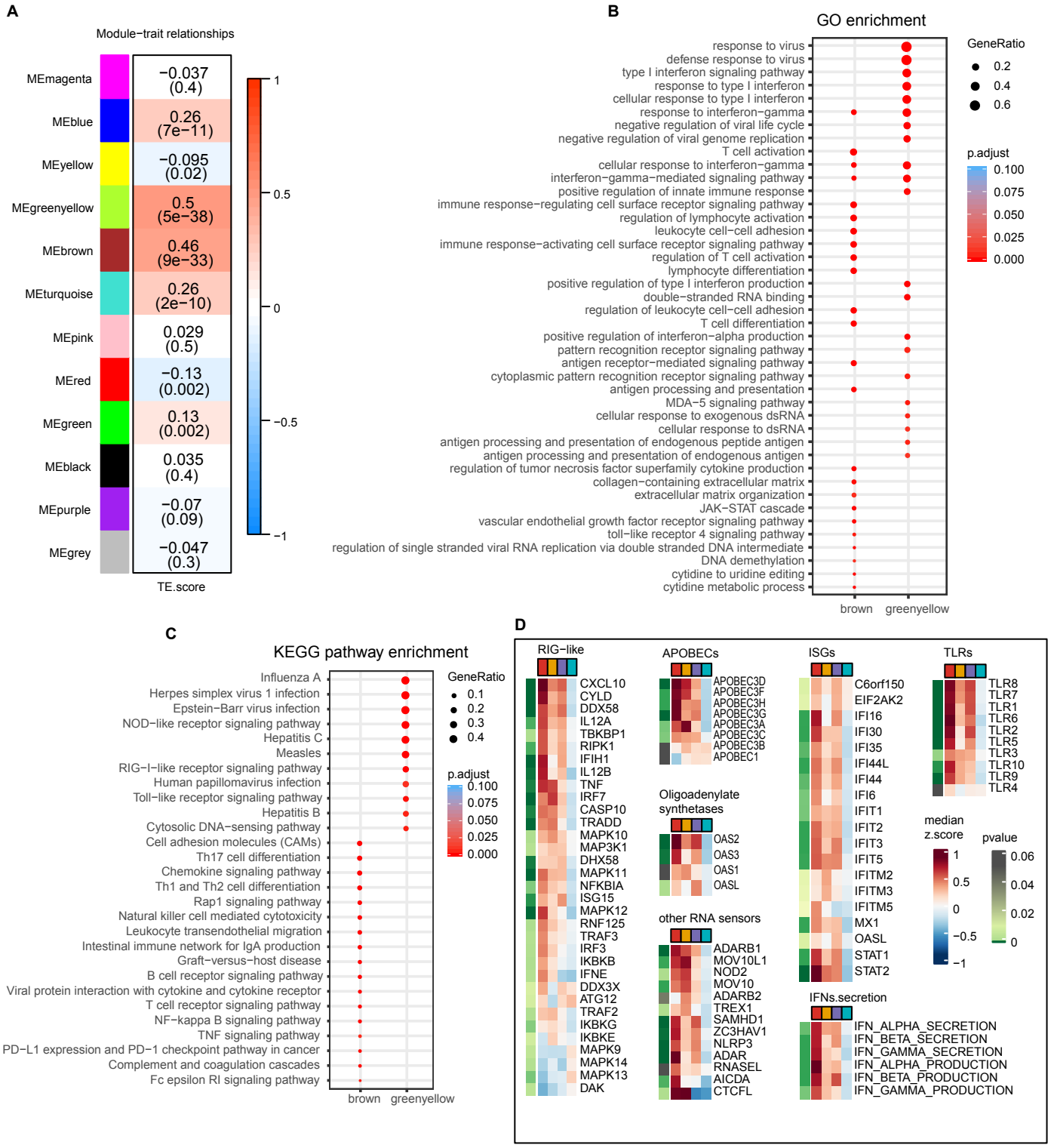


Figure 5. Weighted correlation network analysis (WGCNA) based on TE score. (A) WGCNA consensus network modules correlated with TE score. Each row corresponds to a module, column to the TE score, respectively. Each cell contains the corresponding correlation coefficient and p-value. Individual gene modules were marked using different colors. **(B)** GO enrichment analysis of the genes in the brown and greenyellow module, respectively. **(C)** KEGG pathway enrichment analysis of the genes in the brown and greenyellow module, respectively. The size of the circle indicates the ratio of the genes mapped to each pathway. **(D)** Representative expression of genes or signatures involved in immune response and RNA sensor signals including RIG-I-like pathways, APOBECs, Oligoadenylate synthetases, RNA sensors, interferon-stimulated genes (ISGs), interferon secretion process and Toll-like receptors (TLRs). P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown.

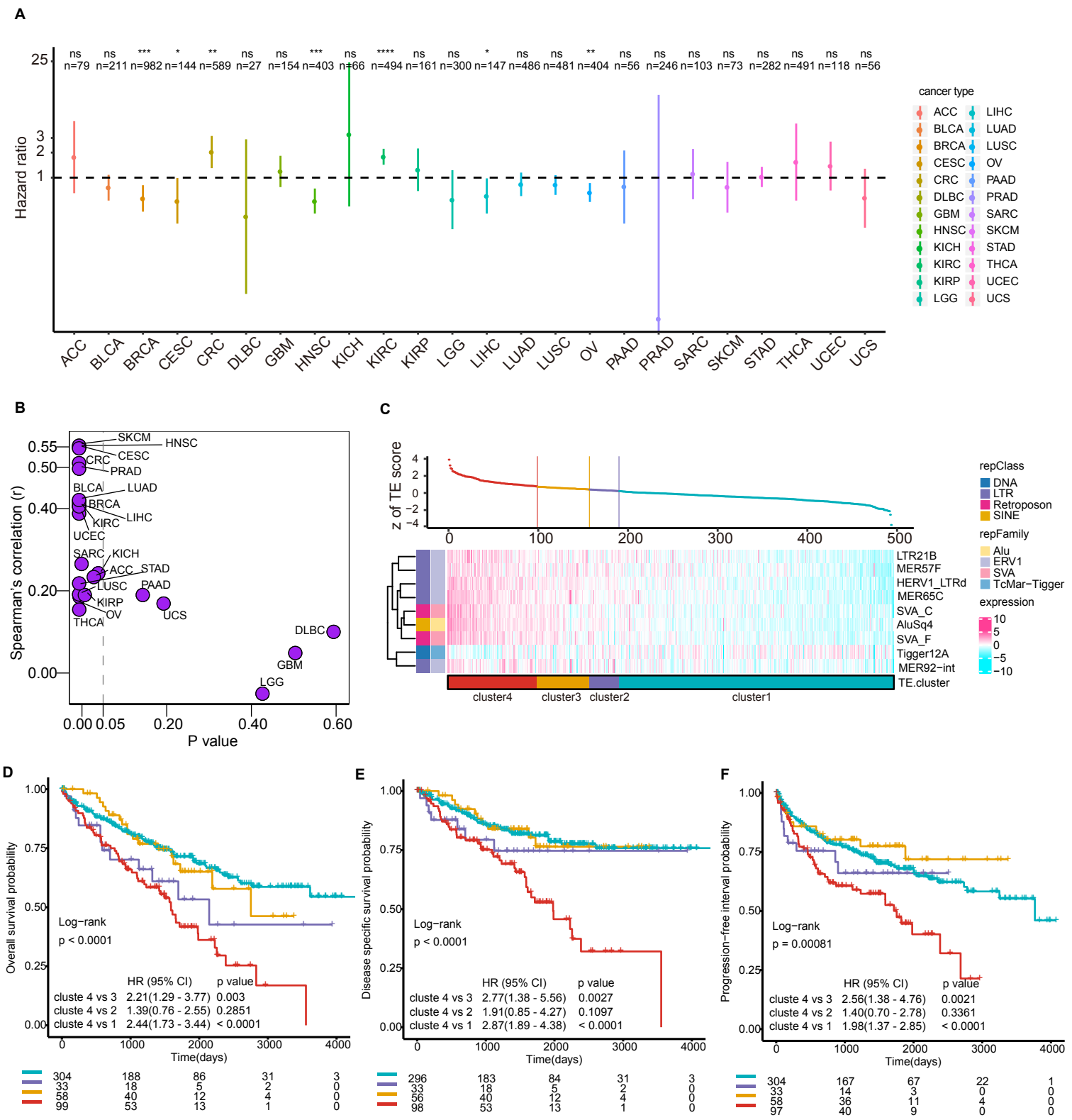
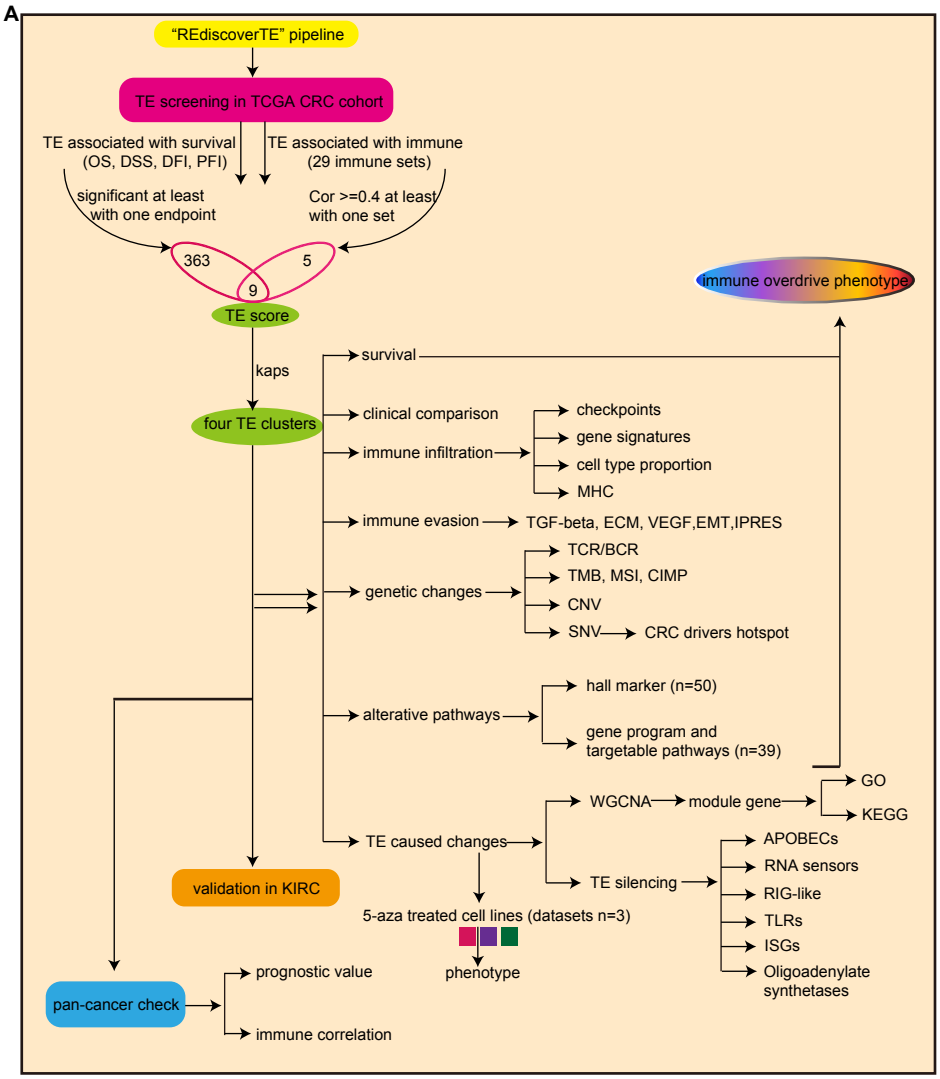
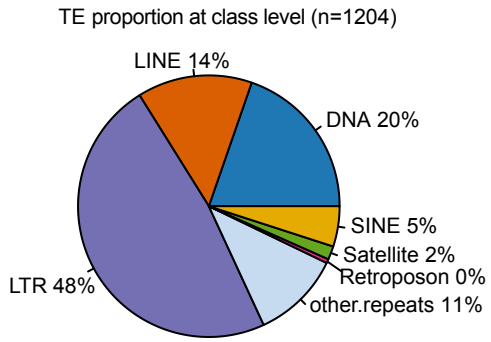
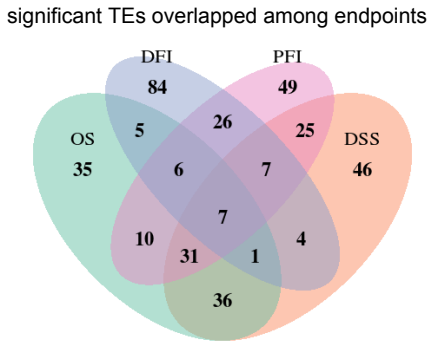
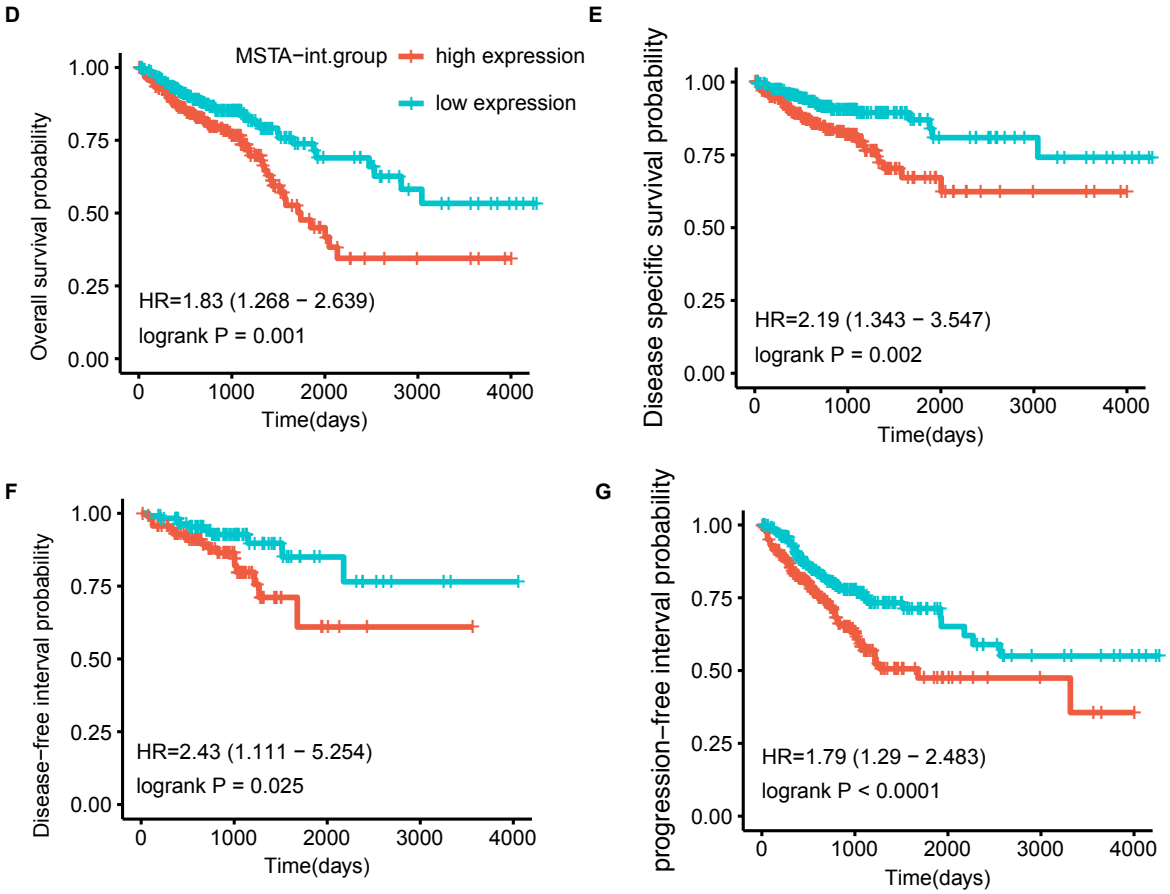


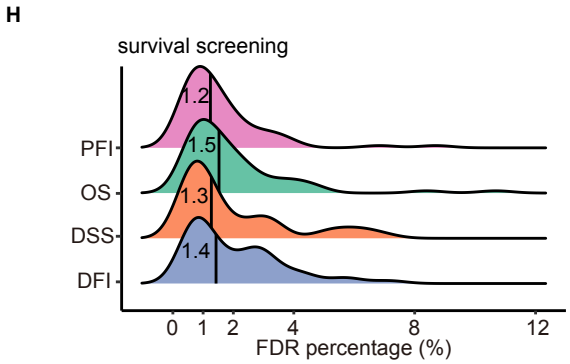
Figure 6. Pan-cancer analysis of TE score and identification overdrive phenotype in KIRC. (A) Forest plot showing the univariable Cox regression analysis of OS on TE score across 24 cancer types. Solid dots represent the HR of death and open-ended horizontal lines represent the 95 % CIs. All p-values were calculated using Cox proportional hazards analysis (ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001). **(B)** Spearman's correlation between TE score and GEP across 24 cancer types. X-axis indicates p-value and y-axis indicates correlation coefficient. **(C)** Top scatter plot showed the distribution TE score in decreasing order from left to right in KIRC. Bottom heatmap displayed the expression profiles of 9 TEs across four TE clusters as ordered by the TE score. Each TE was annotated at family and class level, respectively. **(D-F)** Prognostic value of four TE clusters with Kaplan-Meier survival analysis in KIRC for OS (n = 494) **(D)**, DSS (n = 483) **(E)** and PFI (n = 492) **(F)**. The hazard ratios (HR) and 95% CIs for pairwise comparisons in univariate analyses (log-rank test) are displayed in each Kaplan-Meier plot. Numbers below the x-axes represent the number of patients at risk at the selected time points. The tick marks on the Kaplan-Meier curves indicate the censored patients.


A
"REdiscoverTE" pipeline
TE screening in TCGA CRC cohort
TE associated with survival (OS, DSS, DFI, PFI)
TE associated with immune (29 immune sets)
significant at least with one endpoint
Cor >=0.4 at least with one set
363
5
9
TE score
immune overdrive phenotype
kaps
four TE clusters
survival
clinical comparison
checkpoints
gene signatures
immune infiltration
cell type proportion
MHC
immune evasion
TGF-beta, ECM, VEGF,EMT,IPRES
TCR/BCR
TMB, MSI, CIMP
genetic changes
CNV
SNV
CRC drivers hotspot
hall marker (n=50)
alterative pathways
gene program and targetable pathways (n=39)
GO
WGCNA
module gene
KEGG
TE caused changes
APOBECs
RNA sensors
TE silencing
RIG-like
validation in KIRC
TLRs
5-aza treated cell lines (datasets n=3)
ISGs
phenotype
Oligoadenylate synthetases
prognostic value
pan-cancer check
immune correlation

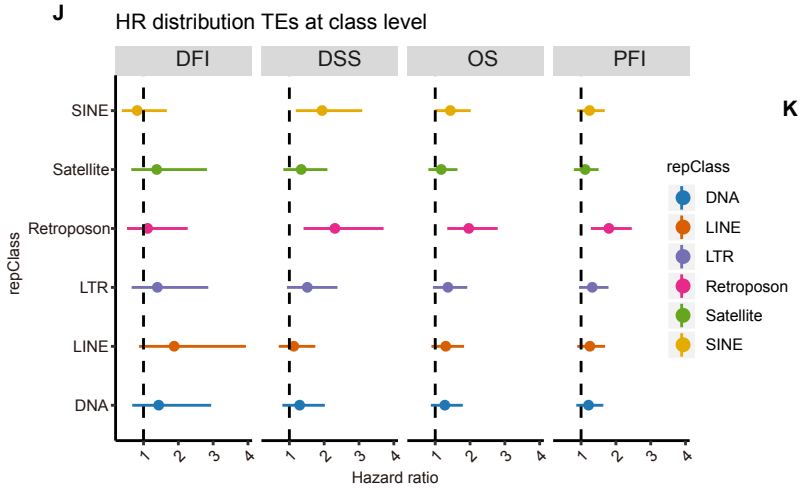

B
TE proportion at class level (n=1204)
LINE 14%
DNA 20%
SINE 5%
Satellite 2%
Retroposon 0%
other.repeats 11%
LTR 48%

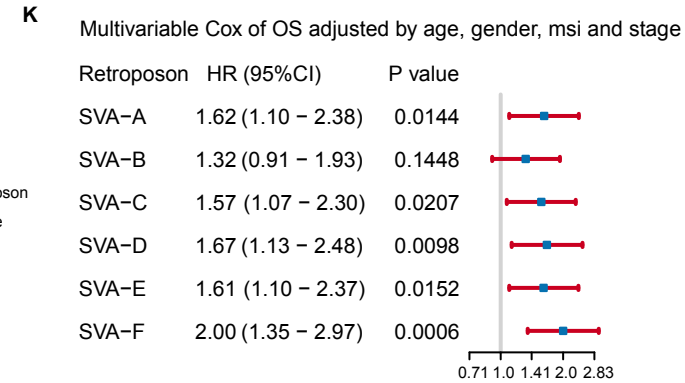

C
significant TEs overlapped among endpoints
DFI
PFI
OS
DSS
84
49
35
46
5
26
25
6
7
7
10
31
1
4
36

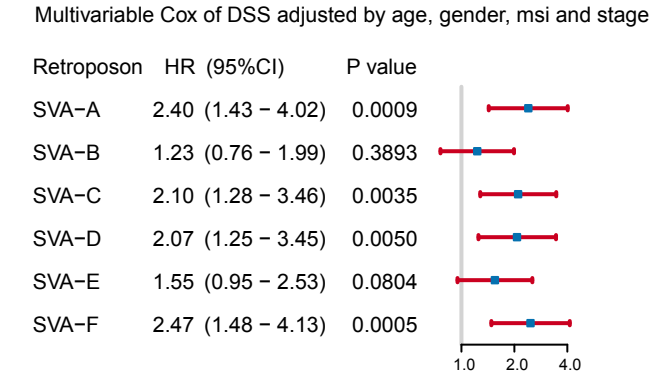

D
MSTA−int.group
high expression
low expression
Overall survival probability
HR=1.83 (1.268 − 2.639)
logrank P = 0.001
Time(days)
E
Disease specific survival probability
HR=2.19 (1.343 − 3.547)
logrank P = 0.002
Time(days)
F
Disease−free interval probability
HR=2.43 (1.111 − 5.254)
logrank P = 0.025
Time(days)
G
progression−free interval probability
HR=1.79 (1.29 − 2.483)
logrank P < 0.0001
Time(days)

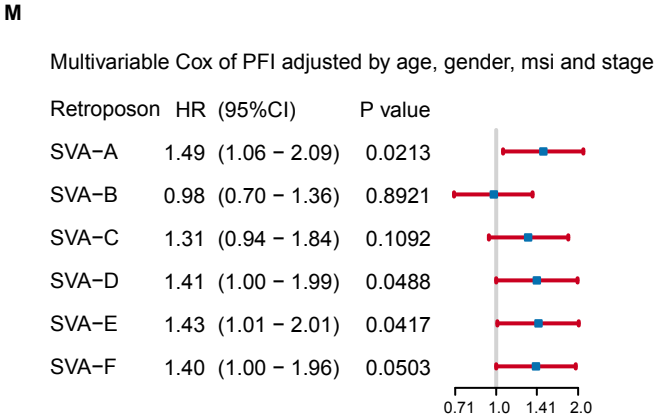

H
survival screening
PFI
OS
DSS
DFI
1.2
1.5
1.3
1.4
0 1 2 4 8 12
FDR percentage (%)

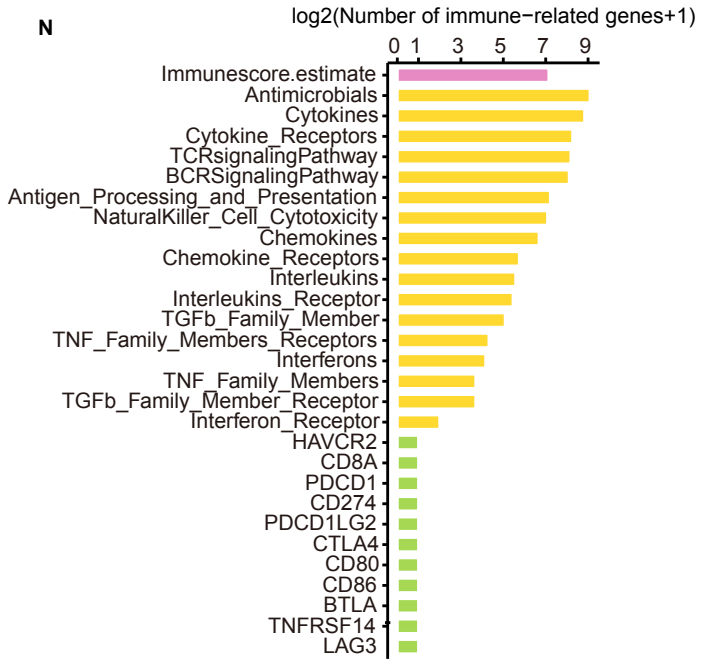
I
HR distribution of significant TEs at family level
DFI
DSS
OS
PFI
repFamily
tRNA
telo
TcMar−Tc2
TcMar
TcMar
SVA
MIR
L2
hAT−Blackjack
hAT
Gypsy
ERVL−MaLR
ERVL
ERVK
ERV1
centr
Alu
Hazard ratio

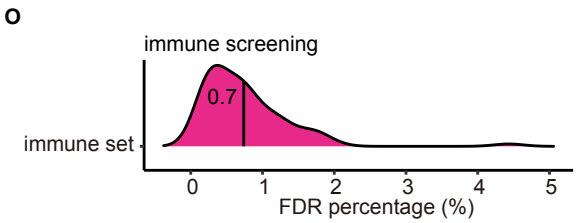

J
HR distribution TEs at class level
DFI
DSS
OS
PFI
repClass
SINE
Satellite
Retroposon
LTR
LINE
DNA
Hazard ratio
repClass
DNA
LINE
LTR
Retroposon
Satellite
SINE


K
Multivariable Cox of OS adjusted by age, gender, msi and stage
Retroposon HR (95%CI) P value
SVA−A 1.62 (1.10 − 2.38) 0.0144
SVA−B 1.32 (0.91 − 1.93) 0.1448
SVA−C 1.57 (1.07 − 2.30) 0.0207
SVA−D 1.67 (1.13 − 2.48) 0.0098
SVA−E 1.61 (1.10 − 2.37) 0.0152
SVA−F 2.00 (1.35 − 2.97) 0.0006
0.71 1.0 1.41 2.0 2.83

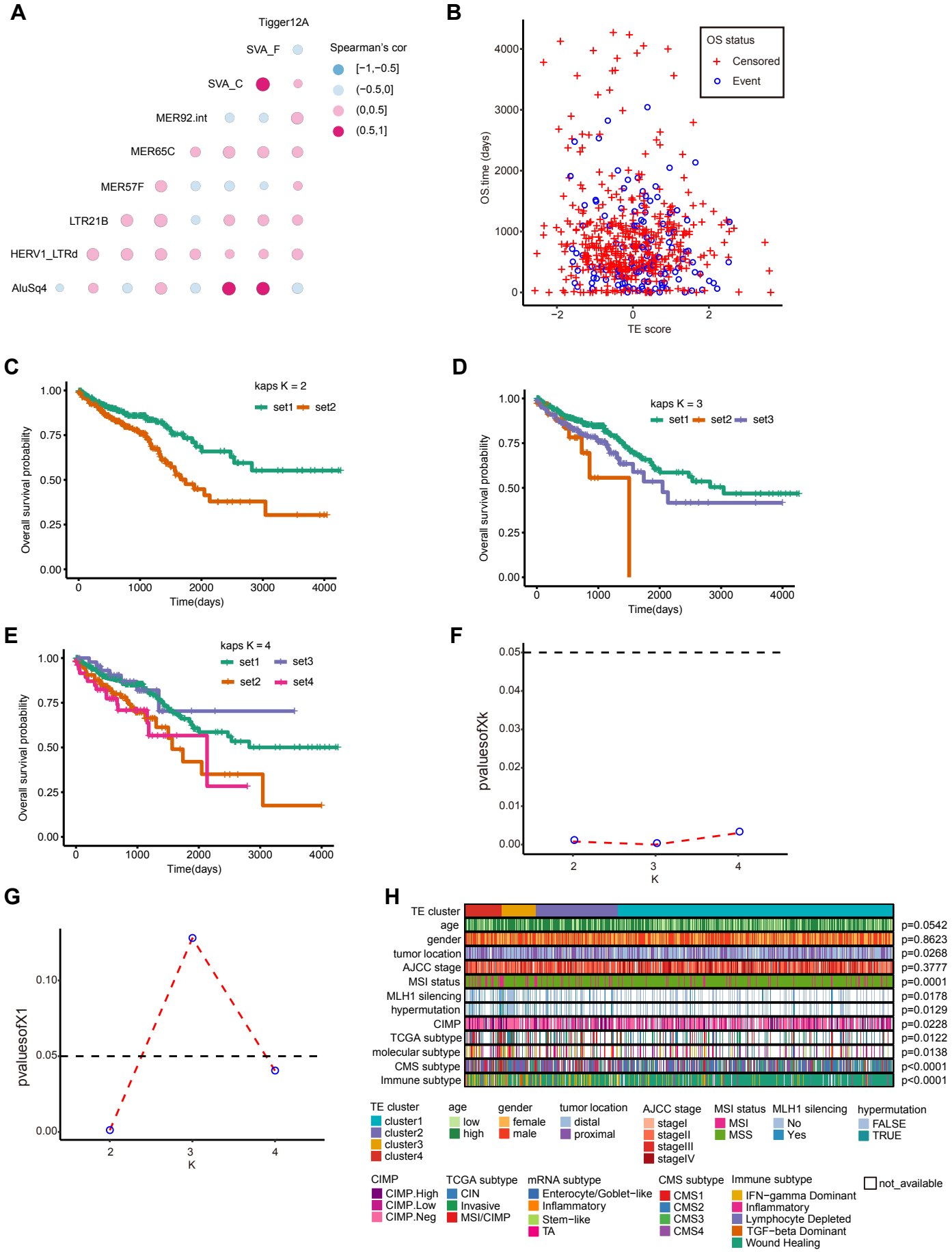

L
Multivariable Cox of DSS adjusted by age, gender, msi and stage
Retroposon HR (95%CI) P value
SVA−A 2.40 (1.43 − 4.02) 0.0009
SVA−B 1.23 (0.76 − 1.99) 0.3893
SVA−C 2.10 (1.28 − 3.46) 0.0035
SVA−D 2.07 (1.25 − 3.45) 0.0050
SVA−E 1.55 (0.95 − 2.53) 0.0804
SVA−F 2.47 (1.48 − 4.13) 0.0005
1.0 2.0 4.0


M
Multivariable Cox of PFI adjusted by age, gender, msi and stage
Retroposon HR (95%CI) P value
SVA−A 1.49 (1.06 − 2.09) 0.0213
SVA−B 0.98 (0.70 − 1.36) 0.8921
SVA−C 1.31 (0.94 − 1.84) 0.1092
SVA−D 1.41 (1.00 − 1.99) 0.0488
SVA−E 1.43 (1.01 − 2.01) 0.0417
SVA−F 1.40 (1.00 − 1.96) 0.0503
0.71 1.0 1.41 2.0

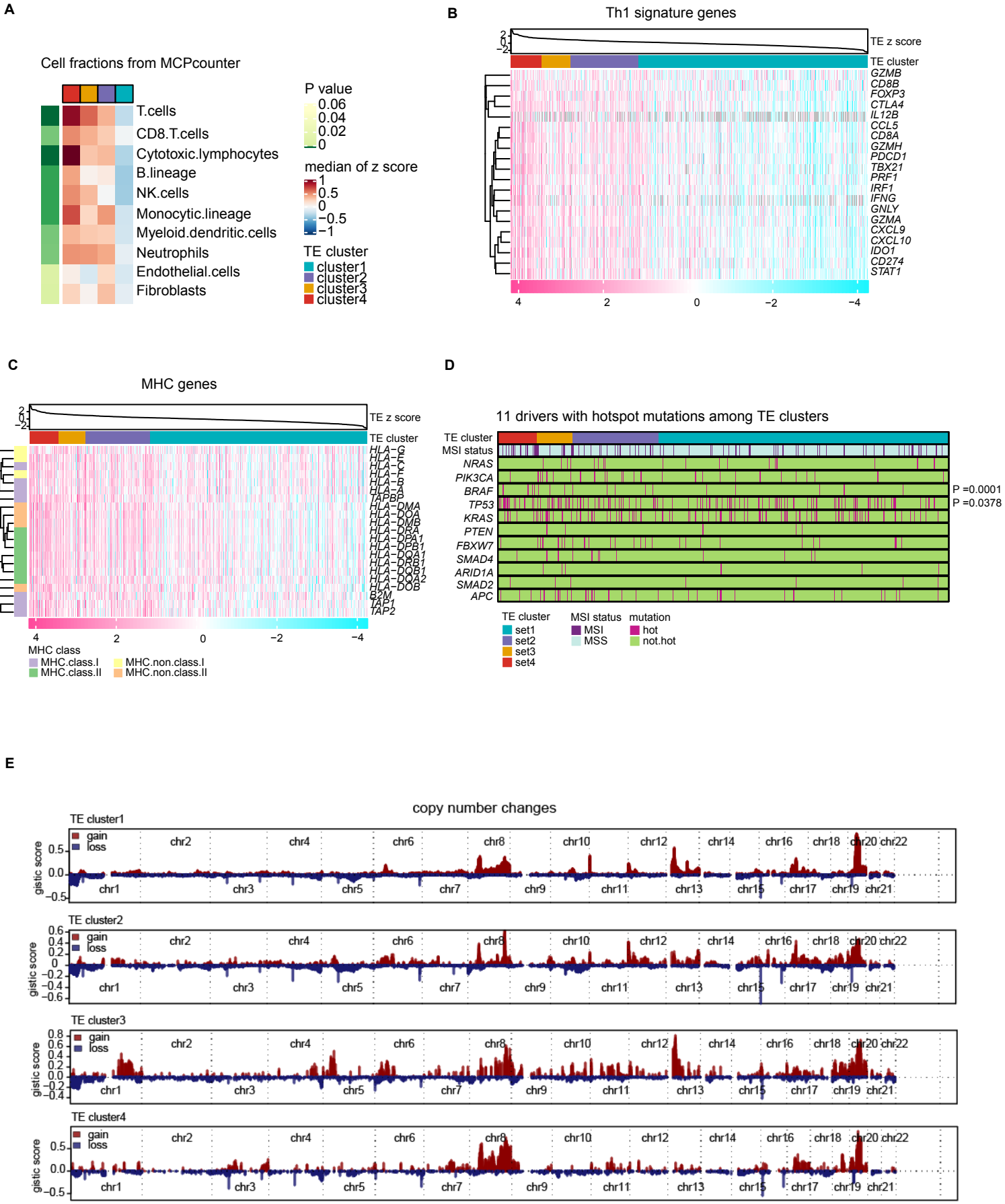

N
log2(Number of immune−related genes+1)
0 1 3 5 7 9
Immunescore.estimate
Antimicrobials
Cytokines
Cytokine_Receptors
TCRsignalingPathway
BCRSignalingPathway
Antigen_Processing_and_Presentation
NaturalKiller_Cell_Cytotoxicity
Chemokines
Chemokine_Receptors
Interleukins
Interleukins_Receptor
TGFb_Family_Member
TNF_Family_Members_Receptors
Interferons
TNF_Family_Members
TGFb_Family_Member_Receptor
Interferon_Receptor
HAVCR2
CD8A
PDCD1
CD274
PDCD1LG2
CTLA4
CD80
CD86
BTLA
TNFRSF14
LAG3

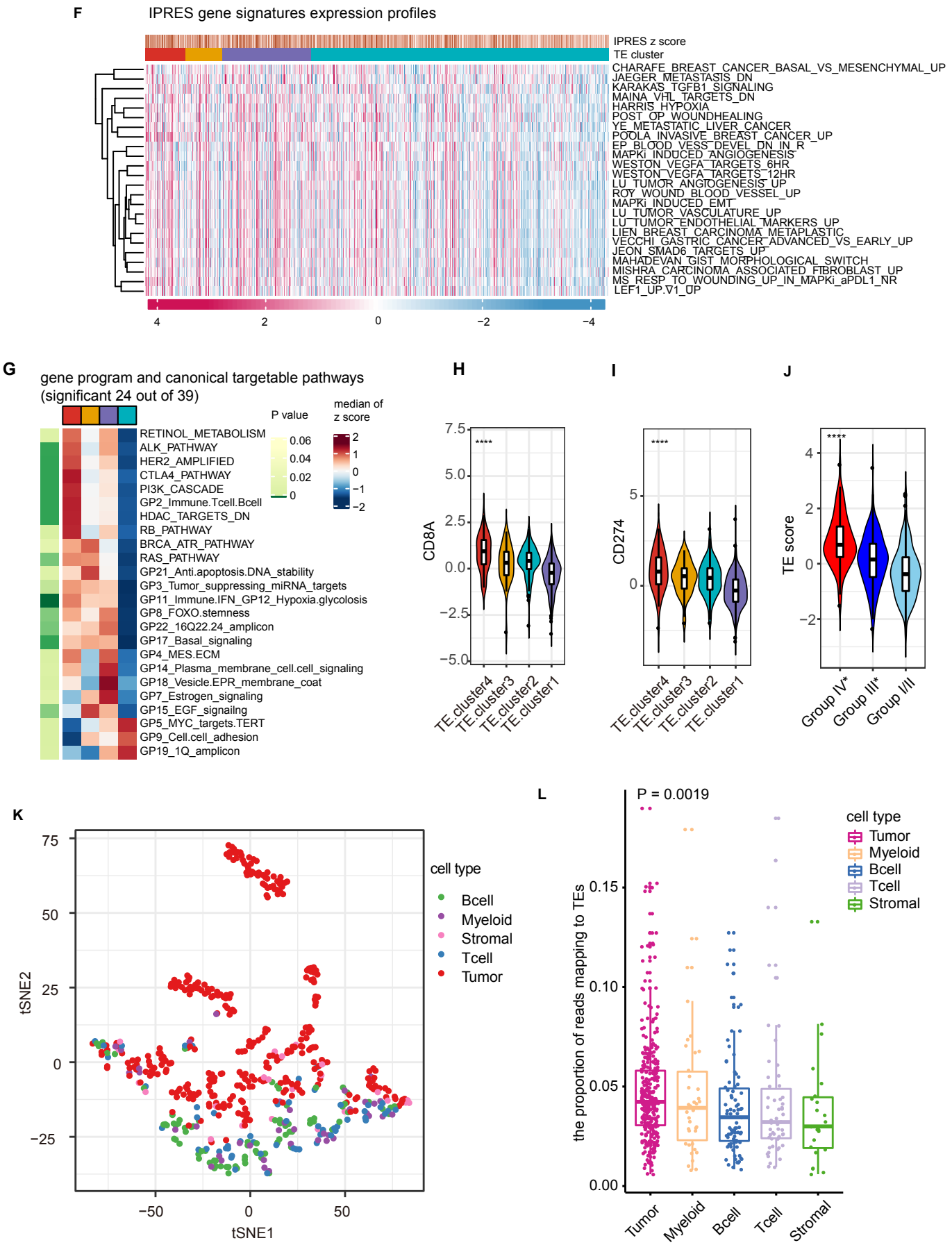

O
immune screening
immune set
0.7
0 1 2 3 4 5
FDR percentage (%)

Supplementary Figure S1. Screening of candidate TEs. (A) Schematic workflow of this study. DSS, disease free survival; OS, overall survival; DFI, disease free interval; PFI, progression free interval; ECM, extracellular matrix; VEGF, vascular endothelial growth factor; EMT, epithelial-mesenchymal transition; IPRES, innate anti-PD1 resistance signature; TMB, tumor mutation burden; CIMP, CpG island methylator phenotype; CNV, copy number variation; SNV, single-nucleotide variant; TLR, Toll-like receptors; ISGs, interferon-stimulated genes. **(B)** Pie chart showing the fraction of 1,204 TEs at class level. **(C)** Venn diagram showing the overlaps of significant candidate TEs associated with four endpoints including OS, DFI, PFI, DSS. **(D-G)** Prognostic value of one representative TE (MSTA-int) with Kaplan-Meier survival analysis for OS **(D),** DSS **(E)**, DFI **(F)** and PFI **(G)**. The hazard ratios (HR) and 95% CIs for pairwise comparisons in univariable analyses (log-rank test) are displayed in each Kaplan-Meier plot. **(H)** Density ridgeline plot showing FDR of univariable Cox regression analysis for each endpoint. Vertical line indicates the median value. **(I-J)** Forest plots showing univariable Cox regression analysis of TEs for four endpoints at family **(I)** and class **(J)** level, respectively. Solid dots represent the HR of death and open-ended horizontal lines represent the 95 % CIs. For TEs at family level, only those families significant with at least one endpoint were shown in (I). Six main TEs at class level were shown. **(K-M)** Forest plots showing multivariable Cox regression analysis of six Retroposon at three endpoints including OS **(K)**, DSS **(L)** and PFI **(M)**. For each Retroposon at each endpoint, four clinical features were included for multivariable Cox regression analysis including age, gender, MSI status and AJCC stage. Solid dots represent the HR of death and open-ended horizontal lines represent the 95% CIs. **(N)** Bar plot showing the number of genes in 29 immune sets. **(O)** Density ridgeline plot showing FDR of Spearman's correlation between TEs and immune sets. Vertical line indicates the median value.

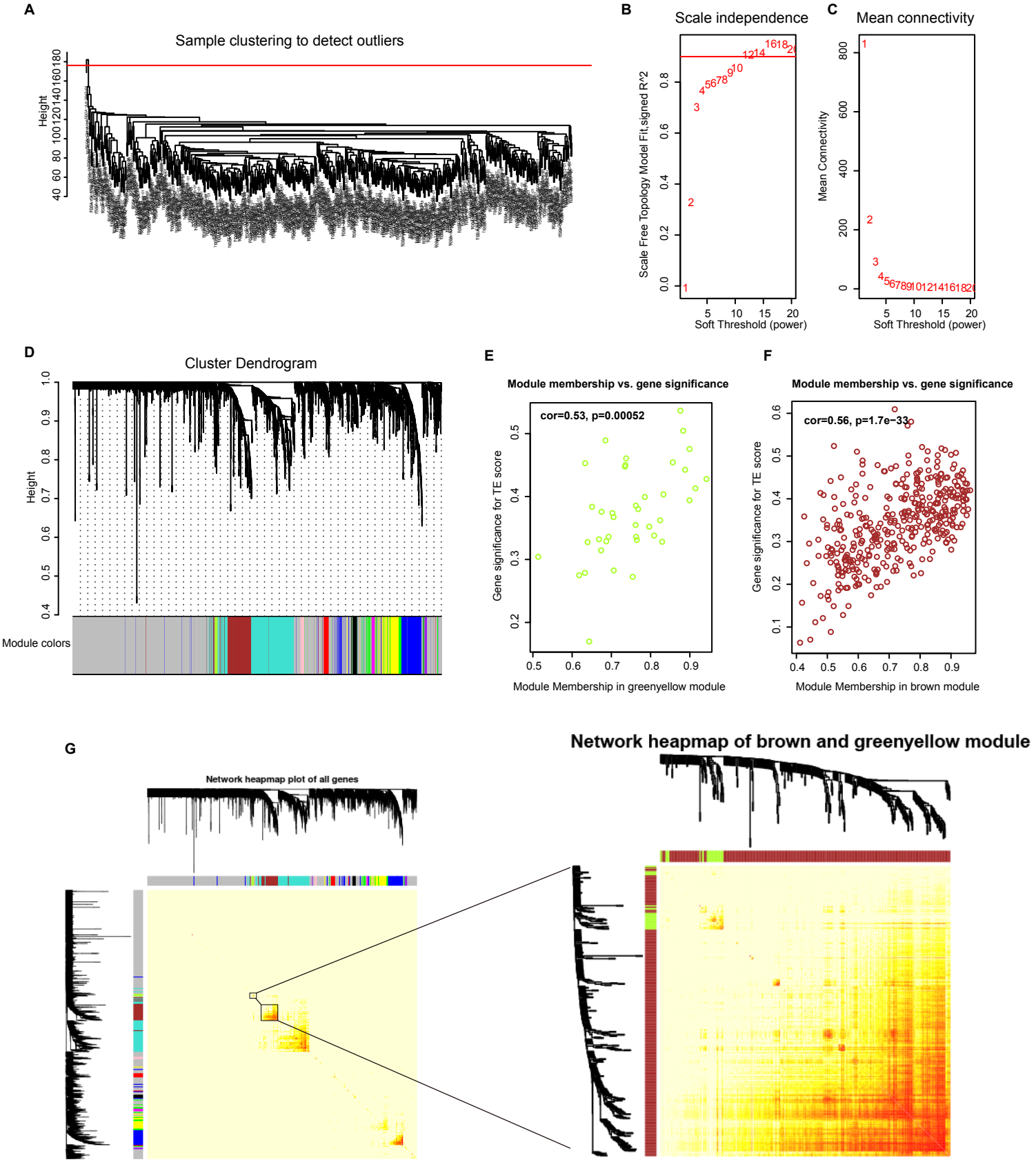

A
Tigger12A
SVA_F
SVA_C
MER92.int
MER65C
MER57F
LTR21B
HERV1_LTRd
AluSq4
Spearman's cor
[−1,−0.5]
(−0.5,0]
(0,0.5]
(0.5,1]
B
OS status
Censored
Event
OS.time (days)
TE score
C
kaps K = 2
set1
set2
Overall survival probability
Time(days)
D
kaps K = 3
set1
set2
set3
Overall survival probability
Time(days)
E
kaps K = 4
set1
set2
set3
set4
Overall survival probability
Time(days)
F
pvaluesofXk
K
G
pvaluesofX1
K
H
TE cluster
age
gender
tumor location
AJCC stage
MSI status
MLH1 silencing
hypermutation
CIMP
TCGA subtype
molecular subtype
CMS subtype
Immune subtype
p=0.0542
p=0.8623
p=0.0268
p=0.3777
p=0.0001
p=0.0178
p=0.0129
p=0.0228
p=0.0122
p=0.0138
p<0.0001
p<0.0001
TE cluster
cluster1
cluster2
cluster3
cluster4
age
low
high
gender
female
male
tumor location
distal
proximal
AJCC stage
stageI
stageII
stageIII
stageIV
MSI status
MSI
MSS
MLH1 silencing
No
Yes
hypermutation
FALSE
TRUE
CIMP
CIMP.High
CIMP.Low
CIMP.Neg
TCGA subtype
CIN
Invasive
MSI/CIMP
mRNA subtype
Enterocyte/Goblet−like
Inflammatory
Stem−like
TA
CMS subtype
CMS1
CMS2
CMS3
CMS4
Immune subtype
IFN−gamma Dominant
Inflammatory
Lymphocyte Depleted
TGF−beta Dominant
Wound Healing
not_available

Supplementary Figure S2. Clinical and molecular comparison among TE clusters. (A) Correlation matrix showing Spearman's correlation coefficient among 9 TEs with each other. **(B)** Scatter plot of survival times (OS) against the prognostic factor (TE score). **(C-E)** Kaplan-Meier survival curves of the selected groups for K = 2 **(C)**, K = 3 **(D)** and K = 4 **(E)**. (F) Plot of the overall p-values against K with significance level α = 0.05. **(G)** Plot of the worst-pair p-values against K with significance level α = 0.05. **(H)** Heatmap showing the distribution of clinical and molecular features among four TE clusters. Each row represents one feature, column to each sample. P-value was calculated using chi-square test.

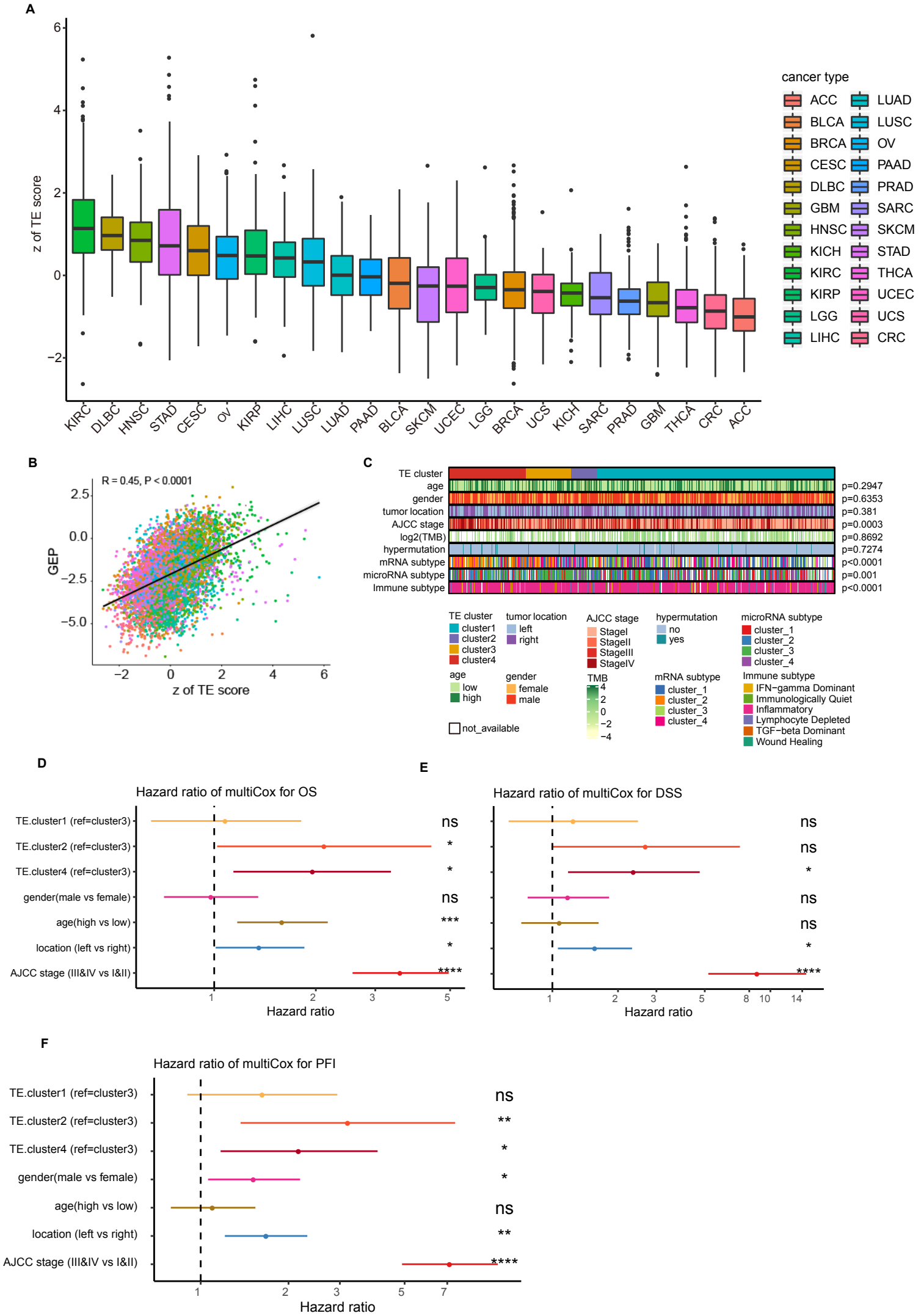

A
Cell fractions from MCPcounter
T.cells
CD8.T.cells
Cytotoxic.lymphocytes
B.lineage
NK.cells
Monocytic.lineage
Myeloid.dendritic.cells
Neutrophils
Endothelial.cells
Fibroblasts
P value
0.06
0.04
0.02
0
median of z score
1
0.5
0
−0.5
−1
TE cluster
cluster1
cluster2
cluster3
cluster4
B
Th1 signature genes
TE z score
TE cluster
GZMB
CD8B
FOXP3
CTLA4
IL12B
CCL5
CD8A
GZMH
PDCD1
TBX21
PRF1
IRF1
IFNG
GNLY
GZMA
CXCL9
CXCL10
IDO1
CD274
STAT1
C
MHC genes
TE z score
TE cluster
HLA−G
HLA−E
HLA−C
HLA−F
HLA−B
HLA−A
TAPBP
HLA−DMA
HLA−DOA
HLA−DMB
HLA−DRA
HLA−DPA1
HLA−DPB1
HLA−DQA1
HLA−DRB1
HLA−DQB1
HLA−DQA2
HLA−DOB
B2M
TAP1
TAP2
MHC class
MHC.class.I
MHC.class.II
MHC.non.class.I
MHC.non.class.II
D
11 drivers with hotspot mutations among TE clusters
TE cluster
MSI status
NRAS
PIK3CA
BRAF
TP53
KRAS
PTEN
FBXW7
SMAD4
ARID1A
SMAD2
APC
P =0.0001
P =0.0378
TE cluster
set1
set2
set3
set4
MSI status
MSI
MSS
mutation
hot
not.hot
E
copy number changes
TE cluster1
TE cluster2
TE cluster3
TE cluster4
gain
loss
gistic score

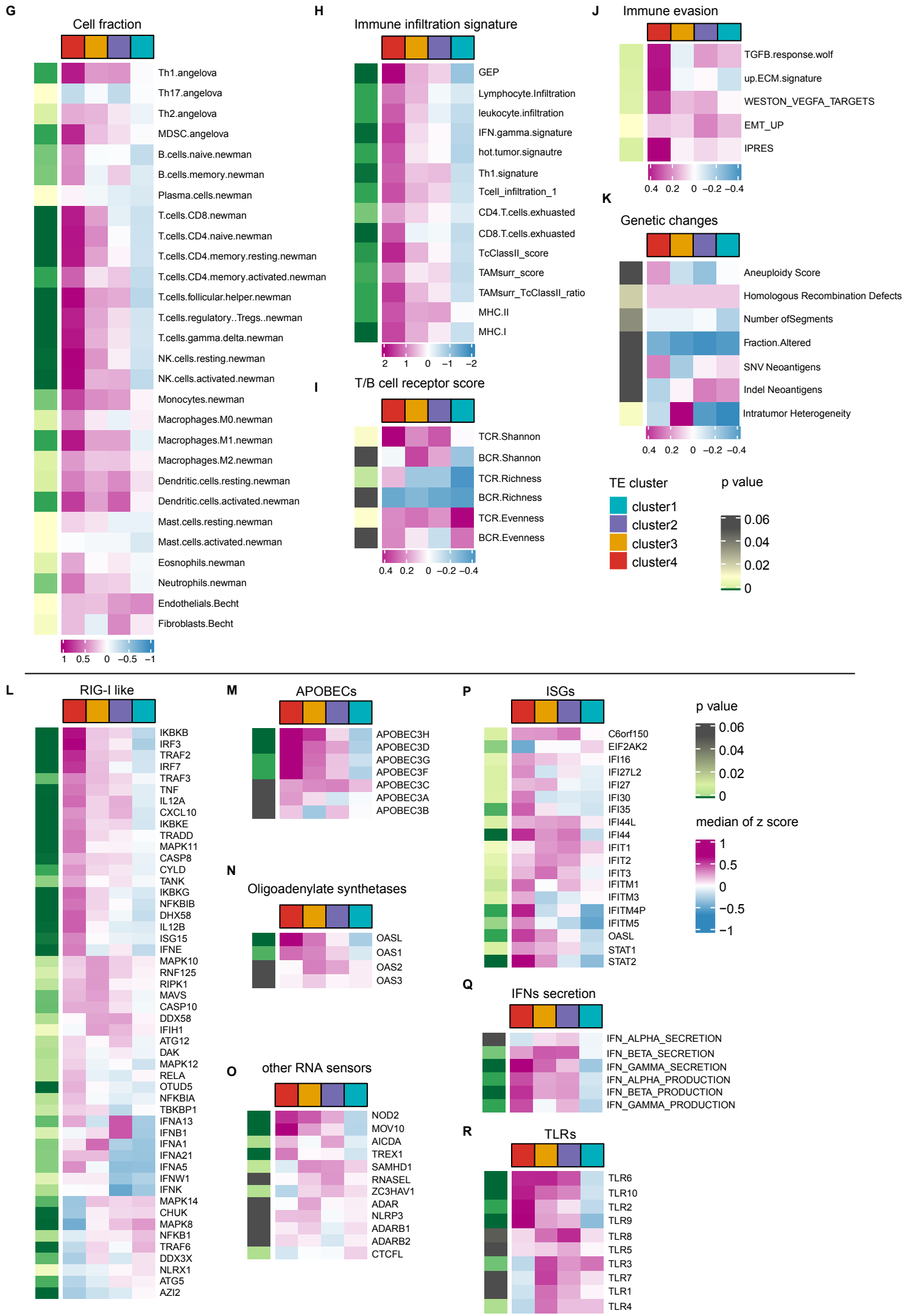

F
IPRES gene signatures expression profiles
IPRES z score
TE cluster
CHARAFE_BREAST_CANCER_BASAL_VS_MESENCHYMAL_UP
JAEGER_METASTASIS_DN
KARAKAS_TGFB1_SIGNALING
MAINA_VHL_TARGETS_DN
HARRIS_HYPOXIA
POST_OP_WOUNDHEALING
YE_METASTATIC_LIVER_CANCER
POOLA_INVASIVE_BREAST_CANCER_UP
EP_BLOOD_VESS_DEVEL_DN_IN_R
MAPKi_INDUCED_ANGIOGENESIS
WESTON_VEGFA_TARGETS_6HR
WESTON_VEGFA_TARGETS_12HR
LU_TUMOR_ANGIOGENESIS_UP
ROY_WOUND_BLOOD_VESSEL_UP
MAPKi_INDUCED_EMT
LU_TUMOR_VASCULATURE_UP
LU_TUMOR_ENDOTHELIAL_MARKERS_UP
LIEN_BREAST_CARCINOMA_METAPLASTIC
VECCHI_GASTRIC_CANCER_ADVANCED_VS_EARLY_UP
JEON_SMAD6_TARGETS_UP
MAHADEVAN_GIST_MORPHOLOGICAL_SWITCH
MISHRA_CARCINOMA_ASSOCIATED_FIBROBLAST_UP
MS_RESP_TO_WOUNDING_UP_IN_MAPKi_aPDL1_NR
LEF1_UP.V1_UP
4
2
0
−2
−4
G
gene program and canonical targetable pathways
(significant 24 out of 39)
P value
0.06
0.04
0.02
0
median of
z score
2
1
0
−1
−2
RETINOL_METABOLISM
ALK_PATHWAY
HER2_AMPLIFIED
CTLA4_PATHWAY
PI3K_CASCADE
GP2_Immune.Tcell.Bcell
HDAC_TARGETS_DN
RB_PATHWAY
BRCA_ATR_PATHWAY
RAS_PATHWAY
GP21_Anti.apoptosis.DNA_stability
GP3_Tumor_suppressing_miRNA_targets
GP11_Immune.IFN_GP12_Hypoxia.glycolosis
GP8_FOXO.stemness
GP22_16Q22.24_amplicon
GP17_Basal_signaling
GP4_MES.ECM
GP14_Plasma_membrane_cell.cell_signaling
GP18_Vesicle.EPR_membrane_coat
GP7_Estrogen_signaling
GP15_EGF_signailng
GP5_MYC_targets.TERT
GP9_Cell.cell_adhesion
GP19_1Q_amplicon
H
CD8A
****
TE.cluster4
TE.cluster3
TE.cluster2
TE.cluster1
I
CD274
****
TE.cluster4
TE.cluster3
TE.cluster2
TE.cluster1
J
TE score
****
Group IV*
Group III*
Group I/II
K
tSNE2
tSNE1
cell type
Bcell
Myeloid
Stromal
Tcell
Tumor
L
P = 0.0019
the proportion of reads mapping to TEs
cell type
Tumor
Myeloid
Bcell
Tcell
Stromal

Supplementary Figure S3. Comparison among TE cluster in terms of immune overdrive. (A) Cell fraction of 10 cell types estimated using MCPCounter algorithm. P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown. **(B)** Heatmap showing the expression profiles of Th-1 signatures compromised of 20 genes. Samples in each column was ordered by TE score with decreasing order from left to right. **(C)** Heatmap showing the expression profiles of MHC genes. Samples in each column was ordered by TE score with decreasing order from left to right. **(D)** Heatmap showing hotspot mutation profiles of 11 CRC drivers. Each row indicates one gene, each column indicates one sample. P-value was calculated using chi-square test by comparing between cluster 4 and three clusters combined. **(E)** CNV plot showing the GISTIC score among four TE cluster. **(F)** Heatmap showing the expression profiles of IPRES signatures compromised of 24 pathways. Each row indicates one pathway, each column indicates one sample. Sample in each column was ordered by TE score with decreasing order from left to right. **(G)** Gene set variation analysis of 39 gene program and canonical targetable pathways. 24 significant pathways were shown. P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown. **(H)** Violin plot showing the comparison of CD8A expression amongst TE clusters. **(I)** Violin plot showing the comparison of CD274 expression amongst TE clusters. **(J)** Violin plot showing the difference of TE score among risk groups identified by Fakih et al (13) (****: $p <= 0.0001$). **(K)** t-SNE plot of TE expression profiles at subfamily level for 515 single cells. **(L)** Boxplot showing the proportion of reads mapping to TEs among cell types. P-value for each variable was calculated using Kruskal-Wallis test.

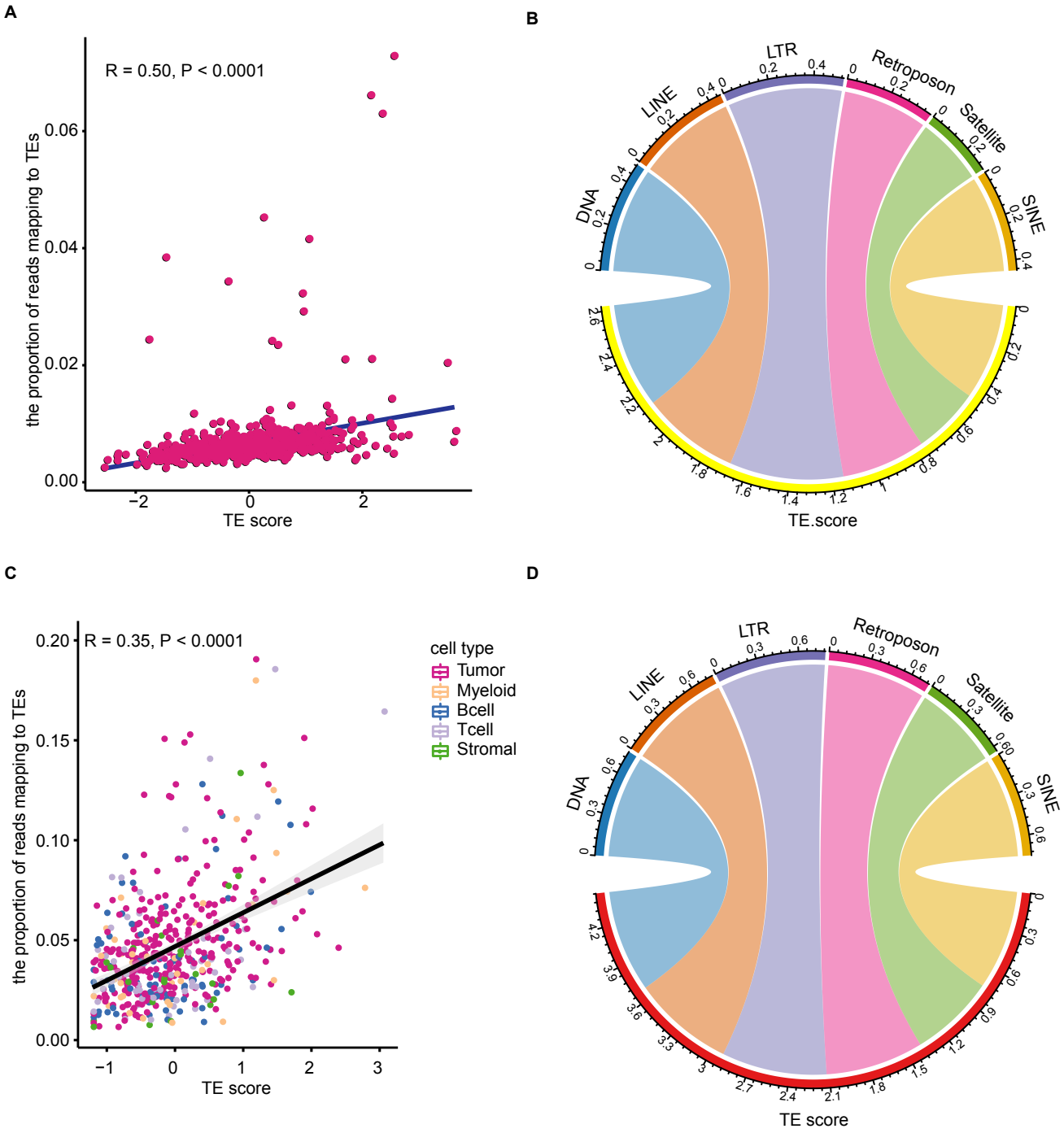

A
Sample clustering to detect outliers
Height
B
Scale independence
Scale Free Topology Model Fit,signed R^2
Soft Threshold (power)
C
Mean connectivity
Mean Connectivity
Soft Threshold (power)
D
Cluster Dendrogram
Height
Module colors
E
Module membership vs. gene significance
cor=0.53, p=0.00052
Gene significance for TE score
Module Membership in greenyellow module
F
Module membership vs. gene significance
cor=0.56, p=1.7e−33
Gene significance for TE score
Module Membership in brown module
G
Network heapmap plot of all genes
Network heapmap of brown and greenyellow module

Supplementary Figure S4. Construction of co-expression module using WGCNA. (A) Clustering dendrogram of CRC samples. One sample (TCGA−AA−3947−01) was considered as outlier and was removed in downstream analysis. **(B-C)** Soft-thresholding power selection in WGCNA. Analysis of the scale-free fit index for individual soft-thresholding powers. Analysis of the mean connectivity for individual soft-thresholding powers. The power = 10 was chosen which is the lowest power for the curve that the scale-free topology fit index flat upon reaching a high value above 0.9 with a moderate mean connectivity. **(D)** Clustering dendrograms of genes included with dissimilarity based on topological overlap, together with assigned module colors. A total of 12 modules were identified and assigned into different colors. **(E-F)** Scatter plots of gene significance for TE score versus module membership in greenyellow **(E)** and brown **(F)** module, respectively. **(G)** Heatmap showing the topological overlap in WGCNA. Each row and column represents a gene, light color indicates low topological overlap, and progressively darker red indicates higher topological overlap. Darker squares along the diagonal represent modules. The gene dendrogram and module assignment are shown along the left and top. Heatmap on the right panel zooms into the brown and greenyellow modules.


A
z of TE score
cancer type
ACC
BLCA
BRCA
CESC
DLBC
GBM
HNSC
KICH
KIRC
KIRP
LGG
LIHC
LUAD
LUSC
OV
PAAD
PRAD
SARC
SKCM
STAD
THCA
UCEC
UCS
CRC
B
R = 0.45, P < 0.0001
GEP
z of TE score
C
TE cluster
age
gender
tumor location
AJCC stage
log2(TMB)
hypermutation
mRNA subtype
microRNA subtype
Immune subtype
p=0.2947
p=0.6353
p=0.381
p=0.0003
p=0.8692
p=0.7274
p<0.0001
p=0.001
p<0.0001
not_available
D
Hazard ratio of multiCox for OS
TE.cluster1 (ref=cluster3)
TE.cluster2 (ref=cluster3)
TE.cluster4 (ref=cluster3)
gender(male vs female)
age(high vs low)
location (left vs right)
AJCC stage (III&IV vs I&II)
Hazard ratio
E
Hazard ratio of multiCox for DSS
Hazard ratio
F
Hazard ratio of multiCox for PFI
Hazard ratio


G
Cell fraction
Th1.angelova
Th17.angelova
Th2.angelova
MDSC.angelova
B.cells.naive.newman
B.cells.memory.newman
Plasma.cells.newman
T.cells.CD8.newman
T.cells.CD4.naive.newman
T.cells.CD4.memory.resting.newman
T.cells.CD4.memory.activated.newman
T.cells.follicular.helper.newman
T.cells.regulatory..Tregs..newman
T.cells.gamma.delta.newman
NK.cells.resting.newman
NK.cells.activated.newman
Monocytes.newman
Macrophages.M0.newman
Macrophages.M1.newman
Macrophages.M2.newman
Dendritic.cells.resting.newman
Dendritic.cells.activated.newman
Mast.cells.resting.newman
Mast.cells.activated.newman
Eosnophils.newman
Neutrophils.newman
Endothelials.Becht
Fibroblasts.Becht
1 0.5 0 −0.5 −1
H
Immune infiltration signature
GEP
Lymphocyte.Infiltration
leukocyte.infiltration
IFN.gamma.signature
hot.tumor.signautre
Th1.signature
Tcell_infiltration_1
CD4.T.cells.exhausted
CD8.T.cells.exhausted
TcClassII_score
TAMsurr_score
TAMsurr_TcClassII_ratio
MHC.II
MHC.I
2 1 0 −1 −2
I
T/B cell receptor score
TCR.Shannon
BCR.Shannon
TCR.Richness
BCR.Richness
TCR.Evenness
BCR.Evenness
0.4 0.2 0 −0.2 −0.4
J
Immune evasion
TGFB.response.wolf
up.ECM.signature
WESTON_VEGFA_TARGETS
EMT_UP
IPRES
0.4 0.2 0 −0.2 −0.4
K
Genetic changes
Aneuploidy Score
Homologous Recombination Defects
Number ofSegments
Fraction.Altered
SNV Neoantigens
Indel Neoantigens
Intratumor Heterogeneity
0.4 0.2 0 −0.2 −0.4
TE cluster
cluster1
cluster2
cluster3
cluster4
p value
0.06
0.04
0.02
0
L
RIG-I like
IKBKB
IRF3
TRAF2
IRF7
TRAF3
TNF
IL12A
CXCL10
IKBKE
TRADD
MAPK11
CASP8
CYLD
TANK
IKBKG
NFKBIB
DHX58
IL12B
ISG15
IFNE
MAPK10
RNF125
RIPK1
MAVS
CASP10
DDX58
IFIH1
ATG12
DAK
MAPK12
RELA
OTUD5
NFKBIA
TBKBP1
IFNA13
IFNB1
IFNA1
IFNA21
IFNA5
IFNW1
IFNK
MAPK14
CHUK
MAPK8
NFKB1
TRAF6
DDX3X
NLRX1
ATG5
AZI2
M
APOBECs
APOBEC3H
APOBEC3D
APOBEC3G
APOBEC3F
APOBEC3C
APOBEC3A
APOBEC3B
N
Oligoadenylate synthetases
OASL
OAS1
OAS2
OAS3
O
other RNA sensors
NOD2
MOV10
AICDA
TREX1
SAMHD1
RNASEL
ZC3HAV1
ADAR
NLRP3
ADARB1
ADARB2
CTCFL
P
ISGs
C6orf150
EIF2AK2
IFI16
IFI27L2
IFI27
IFI30
IFI35
IFI44L
IFI44
IFIT1
IFIT2
IFIT3
IFITM1
IFITM3
IFITM4P
IFITM5
OASL
STAT1
STAT2
p value
0.06
0.04
0.02
0
median of z score
1
0.5
0
−0.5
−1
Q
IFNs secretion
IFN_ALPHA_SECRETION
IFN_BETA_SECRETION
IFN_GAMMA_SECRETION
IFN_ALPHA_PRODUCTION
IFN_BETA_PRODUCTION
IFN_GAMMA_PRODUCTION
R
TLRs
TLR6
TLR10
TLR2
TLR9
TLR8
TLR5
TLR3
TLR7
TLR1
TLR4

Supplementary Figure S5. Pan cancer analysis of TE score. (A) Comparison of TE score across 24 cancer types. **(B)** Spearman's correlation between TE score and GEP in pooled cancer samples (n = 6,554). **(C)** Heatmap showing the comparison of clinical and molecular features among four TE clusters in KIRC. Each row represents one feature, while each column represents one sample. P-value was calculated using the chi-square test. **(D-F)** Forest plots showing multivariable Cox regression analysis of TE cluster adjusted by clinical features for OS **(D)**, DSS **(E)** and PFI **(F)** in KIRC. All variable was set as categorial variable. Samples with age < 65 was set as age low group and ≥ 65 for high group. Solid dots represent the HR of death and open-ended horizontal lines represent the 95 % CIs. All P-values were calculated using Cox proportional hazards analysis (ns: p > 0.05, *: p <= 0.05, **: p <= 0.01, ***: p <= 0.001, ****: p <= 0.0001). **(G)** Gene set variation analysis showing fraction of 28 cell types in KIRC. **(H)** Gene set variation analysis showing immune infiltration signatures in KIRC. **(I)** TCR/BCR indexes comparison among TE clusters in KIRC. **(J)** Genetic changes comparison among TE clusters in KIRC. **(K)** Gene set variation analysis showing immune evasion signatures in KIRC. P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown. **(L-R)** Representative expression of genes or signatures involved in immune response and RNA sensor signals in KIRC including RIG-I-like pathways **(L)**, APOBECs **(M)**, Oligoadenylate synthetases **(N)**, RNA sensors **(O)**, interferon-stimulated genes (ISGs) **(P)**, interferon secretion process **(Q)** and Toll-like receptors (TLRs) **(R)**. P-value for each variable was calculated using Kruskal-Wallis test. For each variable, the median of normalized value in each cluster was shown.


A
R = 0.50, P < 0.0001
the proportion of reads mapping to TEs
TE score
B
LTR
Retroposon
Satellite
SINE
LINE
DNA
TE.score
C
R = 0.35, P < 0.0001
the proportion of reads mapping to TEs
TE score
cell type
Tumor
Myeloid
Bcell
Tcell
Stromal
D
LTR
Retroposon
Satellite
SINE
LINE
DNA
TE score

Supplementary Figure S6. Correlation between TE score and global TE expression. (A) Spearman's correlation between TE score and the proportion of reads mapping to TEs in CRC. **(B)** Circle plot showing Spearman's correlation between TE score and TE expression at six main class level (DNA, LINE, LTR, SINE, Retroposon and Satellite) in CRC. **(C)** Spearman's correlation between TE score and the proportion of reads mapping to TEs in scRNA-seq data of breast cancer. **(D)** Circle plot showing Spearman's correlation between TE score and TE expression at six main class level (DNA, LINE, LTR, SINE, Retroposon and Satellite) in KIRC.