Pan-Cancer Single-Cell Profiling Uncovers the Biological Characteristics of Cancer-Testis Genes

Chunyang Fu[1], Ke Liu[1,*]

Affiliations:

1. [1]Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China

[*]To whom correspondence should be addressed:

keliu.iluke@email.sdu.edu.cn

Abstract

Cancer-testis genes (CTGs), characterized by their restricted expression in testicular and neoplastic tissues, have emerged as pivotal modulators of tumor progression and immunogenicity. Here, leveraging single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq) data from 828 tumor samples spanning 13 cancer types, we systematically identified 407 CTGs based on their expression level in malignant cells. Compared to previously reported CTG sets, our curated genes exhibited broader expression among tumor samples and were predominantly enriched on autosomes. These CTGs displayed pronounced co-expression patterns, with a subset consistently activated in proliferating malignant cells, whereas non–cell cycle CTGs were highly correlated with epigenetic regulatory pathways. Functional analyses pinpointed chromatin-modifying genes, including *ARID4B* and *YEATS2*, as putative upstream regulators, whose knockdown resulted in widespread CTG downregulation. Moreover, we demonstrated that high CTG-count constitutes a robust and scalable feature for malignant cell identification across diverse omics data, including scRNA-seq, single-

cell ATAC sequencing, and spatial transcriptomics. By quantifying transcriptomic similarity to a high CTG-count reference population, we developed a streamlined classification framework that enables rapid and accurate annotation of malignant cells in tumor sc/snRNA-seq datasets across diverse cancer types. Clinically, CTG expression levels were associated with patient prognosis and immune microenvironment composition. In hepatocellular carcinoma, elevated CTG expression defined a stem-like, immune-suppressive malignant cell state enriched for regulatory T cells, linked to advanced disease and poor survival. Collectively, our study refines the understanding of CTGs in cancer biology and provides a valuable resource to facilitate future research and therapeutic targeting.

Introduction

Cancer-testis genes (CTGs) comprise a distinct group of genes characterized by high expression in the testis, aberrant activation in various cancer cells, but minimal or absent expression in most normal tissues[1,2]. Due to the immune-privileged status of the testis, many CTG-encoded proteins, termed cancer-testis antigens (CTAs), are immunogenic and represent attractive candidates for cancer vaccine development[3,4]. Beyond their immunogenicity, emerging evidence indicates that CTGs contribute to diverse intrinsic oncogenic programs, including malignant proliferation[5,6], metastatic dissemination[7,8], and therapeutic resistance[9,10]. In addition, CTG expression in malignant cells can reshape the tumor immune microenvironment, thereby supporting tumor progression through extrinsic mechanisms. For example, CTGs such as *KDM5B*[11] and *ITPRIPL1*[12] have been shown to regulate T cell infiltration and activation, ultimately enabling immune evasion.

A comprehensive catalog of CTGs, detailing their expression patterns across cancer types, would substantially advance functional annotation and target selection.

CTdatabase[13], which compiles over 200 CTGs identified in previous studies, has provided a foundation for CTG research. However, the gene expression profiles of certain CTGs have been validated only in a limited number of samples. Recently, the use of large-scale transcriptomic data from the Cancer Genome Atlas (TCGA)[14] and Genotype-Tissue Expression (GTEx)[15] project has led to the identification of hundreds of previously unreported CTGs, significantly expanding the CTG repository[16,17]. Nevertheless, bulk RNA sequencing (bulk RNA-seq) measures the average gene expression across heterogeneous cellular populations, potentially confounding CTG detection due to stromal and immune cell contributions within the tumor microenvironment.

The advent of single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq) technologies has enabled the precise quantification of transcriptomic profiles within individual malignant cells, revealing dynamic interactions between malignant cells and the tumor microenvironment. Several publicly available sc/snRNA-seq resources, such as 3CA[18], TISCH2[19], and HTAN[20], cover hundreds of cancer studies and provide detailed cell annotation information, offering unprecedented granularity for CTG discovery. In this study, by analyzing malignant cell–specific expression profiles derived from curated sc/snRNA-seq datasets, we systematically identified 407 CTGs across 13 cancer types and comprehensively depict their biological characteristics (**Fig 1A**). These CTGs exhibited broader expression across tumor samples than previously reported CTGs and were significantly associated with patient survival and immune microenvironment composition in multiple cancer types. Furthermore, we introduce CTG-count, a novel metric defined as the number of unique CTGs detected per cell, and demonstrate its utility in distinguishing malignant from normal cells. Building upon this metric, we developed and validated an efficient computational pipeline for malignant cell identification across diverse tumor types.

Results

Definition of cancer-testis genes using single-cell transcriptomic data

We designed a pipeline to systematically identify CTGs for various cancer types (**Fig 1B**). We first compiled a CTG candidate gene list comprising 1,294 genes from two sources: 242 previously characterized CTGs from the CTdatabase and 1,212 testis-enriched protein-coding genes from the Human Protein Atlas[21] (**Table 1**; **Fig S1E**). Most candidates exhibited maximal expression in germ cells among 81 normal cell types (**Fig S1F**). Next, we curated 62 publicly available 10x Genomics sc/snRNA-seq datasets (discovery cohort), comprising 828 tumor samples (each with ⩾10 malignant cells) across 13 cancer types (**Table 2**; **Fig S1A-D**): glioblastoma (GBM); triple-negative and non-triple-negative breast cancer (TNBC and non-TNBC); colorectal cancer (CRC); head and neck squamous cell carcinoma (HNSCC); clear cell renal cell carcinoma (ccRCC); hepatocellular carcinoma (HCC); lung adenocarcinoma (LUAD); neuroblastoma (NB); high-grade serous ovarian cancer (HGSOC); pancreatic ductal adenocarcinoma (PDAC); prostate cancer (PRAD); and melanoma (SKCM). For each cancer type, at least 16 tumor samples were included.

Given a CTG candidate gene *g* and a cancer type *C*, we defined *g* as a CTG of *C* if its expression was detected in >1% of malignant cells in at least 50% of samples from *C*. Applying this criterion, we identified 407 unique CTGs in total (**Table 4**; **Figure 1C and S2A**), with a median of 173 CTGs per cancer (range: 77–246). Among these, SKCM and ccRCC exhibited the highest and lowest number of CTGs, respectively. The number of CTGs identified per cancer type showed no significant association with either sample size or the number of total expressed genes (**Figure S2B**), suggesting that CTG abundance reflects intrinsic tumor biology rather than technical confounders and may serve as a supplementary axis for tumor classification. Notably, 36.7% of CTGs

113 were shared across more than six cancer types (**Figure 1C**), indicating potential
114 convergence in their regulatory mechanisms across cancer types.
115
116 We validated CTG activation in malignant cells using an assembled Smart-seq2 cohort
117 (validation cohort 1; 172 tumor samples across nine cancer types; **Table 2**; **Fig S1A-**
118 **D**). In most cancer types, over 65% of CTGs exhibited median expression levels
119 exceeding 1 Transcripts Per Million (TPM) (**Figure 1D**). Furthermore, a strong
120 correlation was observed between the sample-level median CTG expression
121 percentages in the 10x Genomics and Smart-seq2 platforms (**Figure S2C**), indicating
122 that, despite limited cohort sizes in certain cancer types, the expression levels estimated
123 for most CTGs were robust and reproducible.
124
125 Given that CTGs are typically restricted to express in testis and cancerous tissues, we
126 investigated whether the identified CTGs could effectively discriminate tumor from
127 normal samples using TCGA and GTEx bulk RNA-seq datasets. Although our CTG
128 definition considered only expression in malignant cells and disregarded potential
129 expression in normal cells, overall CTG expression still accurately distinguished
130 tumors from normal tissues across most cancer types (median AUC = 0.91; **Figure 1E**)
131 in most cancer type; in addition, we observed a significant positive correlation between
132 overall CTG expression and tumor purity (median Spearman's $\rho = 0.42$; **Figure 1F**).
133 By contrast, ccRCC displayed both minimal tumor-normal expression difference and
134 weak purity correlation, potentially attributable to the smaller number of CTGs
135 identified in this tumor type, consistent with its previously reported CTG-poor
136 phenotype[22]. It is worth noting that the two types of analyses were also performed in
137 TCGA cancer types not included in the discovery cohort, such as CESC, where
138 consistent trends were observed, supporting the universality and reliability of our CTG
139 set. We further compared our CTG set against the initial candidate genes and previously
140 defined CTGs[13,16,17]. While exhibiting comparable tumor-normal discriminative power
141 to established CTG sets, our set markedly outperformed the original candidate genes

142 (**Figure 1E**). Moreover, it demonstrated significantly higher tumor purity correlation
143 than prior CTG sets (**Figure 1F**).
144
145 In summary, by leveraging large-scale single-cell transcriptomic data, we identified 407
146 CTGs with highly restricted expression to testis and cancer tissues. These CTGs were
147 reproducibly detected across platforms, generalized to diverse tumor types, and hold
148 promise as biomarkers for cancer diagnostics.
149
150

151 **X-CTGs exhibit pronounced sample-restricted expression patterns**

152 In CTG research, CTGs are classified into X-chromosome CTGs (X-CTGs) and
153 autosomal CTGs according to chromosomal location[1]. Interest in X-CTGs dates back
154 to the discovery of MAGEA1 at Xq28, the first characterized cancer-testis antigen[22].
155 Among the initial 1,294 CTG candidates, 14.5% mapped to X chromosome; however,
156 in the final CTG list, the proportion declined to 6.1%, and multiple canonical X-CTGs
157 were not detected by our pipeline, such as *CTAG1B* (also known as NY-ESO-1B), and
158 MAGEA family members (e.g., *MAGEA2*, *MAGEA4*). Fisher's exact test demonstrated
159 that the CTGs we identified were predominantly enriched on the autosome (OR = 3.55;
160 **Figure 2A**).
161
162 This unexpected exclusion of X-CTGs prompted us to further investigate their
163 expression profiles in malignant cells. Although some X-CTGs were undetectable at
164 single-cell resolution, notably, other X-CTGs—particularly MAGEA family genes—
165 showed high expression in a minority of tumor samples, with negligible or absent
166 expression in the remainder (**Figure 2B and S3A**). This sample-restricted expression
167 pattern contrasts starkly with the broadly expressed CTGs in our list.
168
169 To quantify sample specificity, we calculated the tau[22] index for each CTG within each
170 cancer. This index ranges from 0 to 1, with values closer to 1 indicating a higher

propensity for expression in a particular sample. Remarkably, we observed that X-CTGs exhibited significantly higher tau values than autosomal CTGs across most cancers in the discovery cohort (median tau > 0.95 in most cases; **Figure 2C**). An exception was HCC, which displayed the lowest median tau for X-CTGs—slightly below that of autosomal CTGs, though not statistically significant. Analysis of TCGA and TARGET bulk RNA-seq data revealed elevated expression of multiple X-CTGs in HCC compared to other cancers (**Figure S3B**), suggesting an aberrant activation of X-CTGs in HCC.

Independent validation using RNA-seq data from CCLE[23] further corroborated the heightened sample specificity of X-CTGs (**Figure S3C**), indicating that this pattern is not an artifact of scRNA-seq data sparsity. Moreover, CTGs excluded by our pipeline—whether X-linked or autosomal—consistently displayed more restricted expression patterns than those retained. These findings collectively underscore that our identified CTGs represent a subset with broad tumor expression profiles. Accordingly, all subsequent analyses were confined to the 407 CTGs.

Cancer-type specificity of CTGs

We performed Uniform Manifold Approximation and Projection (UMAP) analysis[24] on TCGA tumor samples using the identified CTGs and found that the samples clustered distinctly according to cellular origin (**Figure 3A**). Samples derived from cognate tissues—e.g., brain (GBM, LGG), skin (SKCM, UVM), and kidney (KIRP, KIRC)—exhibited spatial proximity within the embeddings. Of note, BRCA samples were further divided into two distinct clusters corresponding to TNBC and non-TNBC subtypes. Conversely, squamous cancers, including CESC, LUSC, ESCA, BLCA and HNSC, occupied adjacent regions in UMAP space, corroborating prior evidence of conserved molecular signatures across squamous cell carcinomas arising from

199 divergent anatomical sites[25,26]. These results highlighted substantial heterogeneity in
200 CTG expression across tumor lineages.
201
202 Using a tau index >0.9 as a threshold for tumor specificity, we identified 144 CTGs
203 with markedly confined expression to individual cancer types (**Table 4**; **Figure 3B**).
204 This specificity was further validated using bulk RNA-seq data from TCGA and
205 TARGET[27] (**Figure S4A**). For instance, *ROPN1* showed predominant expression in
206 TNBC and SKCM; *IL13RA2* was selectively enriched in GBM; and *PRAME* was
207 confined to SKCM. All three genes have been proposed as promising candidates for
208 adoptive T-cell therapy (ACT) in their respective malignancies[28–30]. Robust tumor
209 expression coupled with negligible expression in normal tissues constitutes the
210 foundation for ACT target identification, ensuring selective tumor cytotoxicity while
211 minimizing off-tumor toxicity[31–33]. Interestingly, we noticed that CTGs with higher
212 cancer specificity exhibited lower expression in non-testis normal tissues (**Figure 3C**).
213 This inverse relationship persisted when we extended this analysis to all 33 TCGA
214 cancer types (**Figure S4B**). Given the immune-privileged status of the testis, the
215 expression pattern of these cancer-specific CTGs render them favorable candidates for
216 ACT.
217
218

219 **Epigenetically Driven Co-Expression of CTGs in Malignant Cells**

220 We observed that individual CTGs exhibited significantly higher correlation
221 coefficients with other CTGs than with non-CTGs (**Figure S5A**), suggesting
222 coordinated transcriptional regulation. To investigate this phenomenon further, we
223 employed CS-CORE[34] to reconstruct co-expression networks of CTGs in malignant
224 cells. This analysis revealed multiple strongly correlated CTG pairs across cancer types,
225 such as *IL13RA2*/*LCA5L* in GBM, and *ROPN1*/*IGSF11* in SKCM; while *KDM5B*
226 (CT31), a lysine demethylase, was broadly co-expressed with numerous CTGs across
227 tumor types (**Figure 4A and S6A**). Remarkably, a CTG cluster enriched in canonical

228 cell cycle-associated genes (e.g., *ATAD2* and *CENPF*) emerged across cancer types
229 (**Figure 4A and S6A**), indicating a functional coupling between neoplastic proliferation
230 and CTG transcriptional activation. To determine whether these proliferative CTGs
231 were consistently activated across different cancers, we stratified tumor cells into
232 cycling and non-cycling groups based on scores from two cell cycle gene sets[34]. This
233 stratification identified 50 CTGs upregulated in cycling malignant cells, most of which
234 exhibited highly concordant expression patterns across diverse cancers (**Table 4**;
235 **Figure S6B**). Among CTGs not been recorded in the two gene sets, some CTGs have
236 already been proven to take an important role in the proliferation of tumor cells, such
237 as *CDCA5*[9] and *OIP5*[35].
238
239 Given the extensive co-expression observed among CTGs, we hypothesized that their
240 upstream transcriptional regulators would exhibit broad positive correlations with
241 multiple CTGs. To investigate pathways linked to CTG co-expression, we conducted
242 pathway enrichment analysis on genes strongly correlated with multiple CTGs, using
243 pseudo-bulk transcriptomic profiles derived from the malignant cells in the discovery
244 cohort. In addition to cell cycle-related genes, we found a surprising enrichment of
245 genes involved in cilium assembly (**Figure 4B**), echoing our earlier observation that
246 numerous germ cell–specific genes are also expressed in ciliated cells (**Figure 1F**).
247 Importantly, we observed significant enrichment of genes involved in epigenetic
248 processes, including "chromatin modifying enzymes" and "chromatin organization",
249 spanning multiple cancer types (**Figure 4B**). When we grouped CTGs into "Cell Cyle-
250 CTGs" and "Non-Cell Cycle CTGs", distinct enrichment patterns emerged: the former
251 exclusively enriched for pathways involved in mitotic processes, whereas the latter
252 showed enrichment in epigenetic pathways (**Figure S7A-B**). Similar findings were
253 obtained in TCAG and TARGET bulk RNA-seq data (**Figure S7C-E**), underscoring the
254 critical influence of epigenetic mechanisms on CTG regulation, particularly for non-
255 cell cycle-associated CTGs.
256

To pinpoint potential epigenetic regulators of CTG activation, genes enriched in the "chromatin modifying enzymes" pathway were extracted and prioritized them based on their frequency across cancer types (**Figure 4C**). Notably, histones were identified as the predominant targets for most of these genes according to annotations from the EpiFactors database[36], highlighting the significance of histone modifications in CTG activation. Top-ranked epigenetic regulators from both sc/snRNA-seq and bulk RNA-seq datasets revealed six shared candidates: *ARID4B*, *KDM3B*, *PBRM1*, *SETD2*, *SUZ12*, and *YEATS2* (**Figure 6C**). To validate the functional role of these regulators, we analyzed differentially expressed genes in *YEATS2* knockdown versus control H1299 cells (a human lung large cell carcinoma cell line)[37]. Consistent with our hypothesis, the majority of CTGs were significantly downregulated in *YEATS2* knockdown cells (**Figure 4D**), thereby confirming the role of *YEATS2* in activating CTG expression. Similarly, knockdown of *ARID4B* in PC3 cells[38] (a prostate cancer cell line) also resulted in CTG downregulation (**Figure S7F**). Given the high mutation frequency of *PBRM1* in ccRCC samples[39] and its positive correlation with CTG expression in our analysis, we examined whether the CTG-poor phenotype observed in ccRCC is associated with *PBRM1* inactivation. Differential expression analysis between *PBRM1* non-silent mutated and wild-type ccRCC tumors showed that *PBRM1* was downregulated in mutated samples (**Figure S7G**), with CTGs identified in ccRCC significantly enriched in the mutated group while the remaining candidate genes were enriched in the wild-type group (**Figure 4E**), suggesting a potential regulatory role for PBRM1 in modulating CTG expression.

Rapid annotation of malignant cells in tumor sc/sn-RNAseq data using CTGs

The identification of malignant cells in single-cell RNA sequencing experiments is pivotal for elucidating tumor biology and developing targeted therapeutic strategies[40,41]. Building on our previous demonstration that CTGs distinguish tumor from normal

samples in bulk RNA-seq datasets, we evaluated their capacity to discriminate malignant and non-malignant cells at single-cell resolution.

Our analysis revealed that while CTG-count (the number of unique CTGs detected per cell) varied across normal cell types, malignant cells consistently exhibited significantly elevated CTG-counts compared to normal counterparts across multiple cancer types (**Figure 5A and S8A**). Since CTG-counts derived from the full 407 CTGs correlated strongly with those based on CTGs identified in corresponding individual cancers (**Figure S8B**), we employed the complete CTG set for all CTG-count associated analyses to ensure consistency.

To validate the generalizability of these findings, we assessed CTG-count differences in 21 independent sc/snRNA-seq datasets (validation cohort 2; **Table 2**; **Figure S1A-D**) encompassing 17 cancer types, 12 of which were not even included in the discovery cohort. Consistent with our initial findings, malignant cells maintained significantly higher CTG-counts (**Figure 5B, S8A and S9A**). To discern whether this elevation was attributable to malignancy-associated upregulation rather than lineage-intrinsic expression patterns, we defined normal reference cells for each cancer type based on its cellular origin (e.g., epithelial cells for carcinoma, glial cells for GBM). In the majority of datasets, malignant cells displayed significantly higher CTG-counts than their corresponding reference populations (**Figure S9B**), thereby substantiating high CTG-count as a malignancy-specific feature. In addition, analysis of a single-nucleus ATAC-seq (snATAC-seq) dataset[42] confirmed that malignant cells displayed consistently higher CTG-counts than normal cells across cancer types (**Figure S9C**), suggesting the utility of CTG-count as a robust and scalable metric for malignant cell identification across multi-omics data.

We next evaluated the predictive performance of CTG-count for classifying malignant cells within individual samples. Notably, in the discovery cohort, 84.6% (11/13) of cancer types achieved a median ROC-AUC ≥ 0.75 (**Figure 5C**), although ccRCC

exhibited the lowest performance (median ROC-AUC = 0.55), consistent with its CTG-deficient characteristic. The predictive performance of CTG-count remained robust across sequencing platforms and library complexities, with optimal classification observed on the Smart-seq2 platform (**Figure S10D-E**). Importantly, CTG-count outperformed both aggregate CTG expression and gene-count (total number of genes detected per cell) in discriminating malignant from normal cells in both discovery and validation cohorts (**Figure S10A**).

To further enhance the discriminatory power of CTG-count in identifying malignant cells, we next developed a classification framework termed SingleR-CTG (**Figure 5D**). In this workflow, cells within a given cancer sc/snRNA-seq dataset were ranked by CTG-count, with the top 5% designated as malignant-like and the bottom 5% as normal-like, respectively. Following pseudo-bulk aggregation, a two-column reference expression matrix was generated, with one column representing malignant-like cells and the other representing normal-like cells. SingleR[43] was then employed to compute the Spearman correlation of each cell with the malignant-like reference, generating a malignant score.

In a representative TNBC sample from the discovery cohort, this score substantially improved the separation between malignant and normal cells relative to raw CTG-count (**Figure 5E**). Benchmarking against established malignant cell identification tools, our method demonstrated comparable classification accuracy to inferCNV[44] and slightly outperformed ikarus[45] and scCancer2[46] in both discovery and validation cohorts (discovery cohort: ROC-AUC: 0.95, PR-AUC: 0.91; validation cohort 1: ROC-AUC: 0.99, PR-AUC: 0.96; validation cohort 2: ROC-AUC: 0.95, PR-AUC: 0.91; **Figure 5F**). When the analysis was limited to tumor and normal reference cells only, similar results were obtained (**Figure S10F**). Notably, while inferCNV and our approach yielded highly similar annotations, our method offered substantial computational advantages (**Figure 5G**). For instance, in datasets with ~5,000 cells, inferCNV required ~11 minutes for completion, whereas the computation of malignant score via SingleR took

only 1.3 seconds. This computational advantage increased with dataset size, making our method particularly suitable for large-scale malignancy classification.

CTG-count as a surrogate of tumor purity in spatial transcriptomics data

Spatial transcriptomics (ST) technologies enable the interrogation of tumor subclonal architecture and spatial interactions between cancer cells and the tumor microenvironment[47,48]. Efficient and generalizable identification of malignant spots is a prerequisite for exploiting expanding public ST datasets.

To investigate the ability of CTG-count in malignant spot identification, we first analyzed 12 oral squamous cell carcinoma (OSCC) samples with expert pathological annotations[49]. As anticipated, CTG-counts were significantly higher in malignant versus non-malignant spots in all cases (**Figure S11A**). When we used CTG count to predict OSCC spots, the median ROC-AUC reached 0.74 (**Figure 6A**). Similarly, in an HCC dataset with definitive cell annotations[50], CTG-counts were markedly elevated in tumor cells relative to all other cell types, including normal hepatocytes (**Figure 6B**), reinforcing the malignancy-specific nature of CTG-count.

To comprehensively assess the robustness of CTG-counts across diverse cancer contexts, we subsequently collected 245 ST slides encompassing 15 cancer types (**Table 2**). Given the spots typically contain multiple cells (ranging from 1 to 10), the ESTIMATE algorithm[51] was used to infer tumor purity for each spot. Notably, CTG-count demonstrated a strong positive correlation with estimated tumor purity (Spearman's $\rho > 0.5$) in 10 of 15 cancer types (**Figure 6C-D and S11C**), confirming its utility as a purity surrogate. As observed in single-cell data, ccRCC displayed weaker correlations, and CTG-count showed superior correlation with tumor purity than either aggregate CTG expression or total gene count (**Figure 6D and S11B**). Recognizing that aneuploidy is a hallmark of malignancy and a widely accepted indicator of malignant

374 cell identity, we further assessed the relationship between CTG-count and copy number
375 variation (CNV). Interestingly, our results revealed that CTG-counts were positively
376 correlated with CNV scores across cancer types (**Figure S11D**), and were significantly
377 higher in CNV-high clusters compared to CNV-low counterparts (**Figure 6E**).
378 Collectively, these findings suggest that CTG-count serves as a robust proxy for tumor
379 purity in ST data and provides a practical index for the rapid annotation of malignant
380 spots.
381
382

383 **Elevated CTG expression defines an immune-evasive and stem-like subtype in**
384 **HCC**

385 To evaluate the prognostic relevance of CTGs, we systematically analyzed associations
386 between individual CTG expression levels and patient overall survival across cancer
387 types (**Table 5**). Among the 407 CTGs assessed, 129 (31.2%) demonstrated significant
388 associations with survival in at least one cancer type. Notably, context-dependent
389 survival associations were observed for several CTGs. For example, elevated *SPATA24*
390 expression correlated with poorer survival in HCC (HR = 1.43, p = 0.00098) and NB
391 (HR = 1.73, p = 0.00023), while being associated with improved outcomes in PDAC
392 (HR = 0.60, p = 0.00236) and ccRCC (HR = 0.67, p = 0.00004). Of particular note,
393 CTGs were significantly enriched for survival-associated genes compared to non-CTGs
394 in GBM, HCC, LUAD, and PDAC (**Figure 7A**). A similar trend was observed in
395 ccRCC, though not reaching statistical significance, reinforcing the potential of CTGs
396 as prognostic biomarker candidates. In addition, elevated overall CTG expression was
397 significantly associated with adverse overall survival in HCC (HR = 5.10, p = 0.00017;
398 **Figure 7B**), suggesting that co-expression of survival-associated CTGs may contribute
399 to aggressive disease phenotypes in HCC.
400

401 Given the established role of CTGs in modulating tumor immunogenicity[11,12], we
402 subsequently investigated the relationship between CTG expression level in malignant

403 cells and immune cell infiltration utilizing sc/snRNA-seq data from the discovery
404 cohort (**Figure 7C**). To minimize confounding from immune remodeling in metastatic
405 lesions, this analysis was restricted to primary tumors. In half of the tumor types
406 evaluated, CTG expression in malignant cells correlated significantly with the
407 abundance of various immune cell subsets, most frequently demonstrating a negative
408 correlation with $CD8^+$ T cells and a positive correlation with M2 macrophages. Notably,
409 these dual associations co-occurred in GBM, TNBC, HNSCC, and PDAC. In contrast,
410 HCC exhibited a distinct pattern, with elevated CTG expression strongly associated
411 with increased infiltration of regulatory T cells (Tregs) (**Figure 7C and S12A**). These
412 findings support a consistent link between CTG-high malignant states and
413 immunosuppressive microenvironments, suggesting that CTG-high phenotypes may
414 promote immune evasion through engagement of specific immunoregulatory
415 populations. Consistent with this, elevated CTG expression was associated with
416 downregulation of immune activation pathways (e.g., interferon gamma response,
417 inflammatory response) in malignant cells across multiple cancer types (**Figure 7C**).
418
419 The dual associations in HCC linking elevated CTG levels to adverse survival outcomes
420 and enhanced Treg infiltration prompted a deeper investigation of the cellular and
421 molecular features of CTG-high HCC malignant cells. In an HCC scRNA-seq dataset[52]
422 (n = 13) from the discovery cohort, differential expression analysis comparing CTG-
423 high versus CTG-low malignant cells revealed significant upregulation of cell cycle
424 regulators (e.g., *BIRC5*, *TOP2A*) and stemness-associated genes (e.g., *EPCAM*, *MDK*),
425 alongside downregulation of gene involved in antigen presentation (e.g., *CD74*, *HLA-*
426 *DRA*), cholesterol biosynthesis (e.g., *INSIG*), cytochrome P450 enzymes (e.g.,
427 *CYP27A1*, *CYP2A6*), and apolipoproteins (e.g., *APOB*, *APOC1*) (**Figure 7D**). Hallmark
428 pathway analysis corroborated these observations, showing increased activity of
429 proliferation-related programs (e.g., E2F and MYC targets) and repression of immune
430 response (e.g., interferon gamma response, complement cascade) and metabolic
431 pathways (e.g., cholesterol homeostasis, bile acid metabolism) (**Figure 7E and S12C**).

Considering the liver's central role in metabolism, the profound suppression of hepatic metabolic programs in CTG-high cells suggests a dedifferentiated, less hepatocyte-like state. Consistently, CytoTRACE2[53] analysis indicated higher stemness in CTG-high malignant cells (**Figure 7E**).

These findings were independently validated in a separate HCC scRNA-seq dataset[54] (n = 18), where gene expression and pathway profiles were highly concordant (**Figure 7D and S12B–C**). Consistent with the discovery dataset, Treg infiltration was markedly elevated in CTG-high tumors (**Figure 7F**). Complementary analysis of TCGA-LIHC bulk RNA-seq data using CIBERSORT deconvolution[55] further confirmed increased Treg abundance in CTG-high tumors (**Figure S12D**). Moreover, CTG expression was higher in advanced-stage (stage II/III) compared to early-stage (stage I) HCC (**Figure 7G**), and strongly correlated with the stemness indices (**Figure 7H and S12E**). Taken together, these findings suggesting that elevated CTG[56] expression characterizes an immune-evasive, stem-like subtype of HCC, which is associated with tumor progression and unfavorable clinical outcomes.

Discussion

CTGs have garnered considerable attention in cancer research due to their restricted expression in testis and crucial roles in tumorigenesis and tumor progression[1–4]. In this study, we refined a catalogue of 407 CTGs across 13 prevalent cancer types by leveraging single-cell and single-nucleus transcriptomic data. Diverging from previous identification strategies[16,17], our stringent threshold (expression in ≥50% of samples per cancer type) specifically enriches for CTGs with broad expression across tumor samples, rather than those highly expressed in only a small subset of cases. Strikingly, these CTGs predominantly localized to autosomes, whereas X-chromosome CTGs exhibited pronounced sample-restricted expression. Although X-CTGs have received

substantial attention and served as the basis for multiple vaccine strategies[3,4], outcomes from several clinical trials have been unsatisfactory[57,58]. Consequently, prioritizing autosomal CTGs with broad expression may enhance the therapeutic applicability of CTG-based strategies and benefit a larger proportion of patients.

Although many CTGs are shared across cancer types, we observed pronounced cancer-type specificity in both the number of CTGs identified per cancer and their expression patterns. Notably, CTGs with greater cancer specificity exhibited markedly lower expression in non-testicular tissues. A major limitation of ATC therapies is the risk of on-target, off-tumor toxicity due to the expression of target genes in normal cells. In this context, cancer-specific CTGs, defined by minimal expression in non-testicular tissues and robust expression in tumors, represent attractive ACT targets, offering the potential for selective tumor eradication while minimizing adverse effects. Indeed, several ACT therapies targeting cancer-specific CTGs have demonstrated encouraging outcomes in clinical trials[28–30], underscoring their translational potential.

Beyond their immunotherapeutic potential, our findings establish CTGs as versatile biomarkers. CTG expression robustly distinguished tumor from normal tissues and showed strong correlations with tumor purity, highlighting their potential utility in cancer detection. Furthermore, CTG expression patterns effectively discriminated cancer types according to their cellular origins. Even within breast cancers, CTG expression differed between TNBC and non-TNBC subtypes. Notably, a previous study[59] reported that selective activation of two CTGs in basal-like breast cancers was prognostic and associated with enhanced tumor aggressiveness. CTGs were also significantly enriched for genes associated with patient survival relative to non-CTGs. Pan-cancer analysis revealed that elevated CTG expression in malignant cells was broadly associated with immunosuppressive tumor microenvironments. In HCC, malignant cells with high CTG expression exhibited transcriptional hallmarks of stemness and enhanced proliferative activity, coupled with suppression of hepatic metabolic functions and immune signaling pathways. These tumors also showed

490 increased infiltration of Tregs and were associated with poorer clinical outcomes and
491 more aggressive disease progression.
492
493 Previous studies have reported CTG co-expression in several cancer types[60–62].
494 Extending these findings, we uncovered extensive CTG co-expression networks
495 orchestrated by distinct biological programs, particularly those involving the cell cycle
496 and chromatin-modifying enzymes. Cell-cycle CTGs were broadly activated in
497 proliferating malignant cells across cancer types and formed conserved co-expression
498 modules. Functional relevance of these genes has been demonstrated; for example,
499 knockdown of *CDCA4* in esophageal cancer[9] and *OIP5* in GBM[35] significantly
500 impaired tumor cell proliferation. Whether these cell cycle–associated CTGs share
501 conserved oncogenic functions, and whether their inhibition could serve as a
502 generalized anti-tumor strategy across malignancies warrant further investigation. In
503 parallel, and in contrast to earlier reports that emphasized DNA methylation as the
504 primary regulator of CTG expression[16,63–66], our results highlight histone modifications
505 as important contributors to CTG activation. Specifically, knockdown of the epigenetic
506 regulators *YEATS2* and *ARID4B* in distinct cancer cell lines led to significant
507 downregulation of CTG expression, supporting their regulatory roles. Moreover,
508 variation in histone-modifying landscapes may account for inter-cancer differences in
509 CTG expression frequency. For instance, the CTG-poor phenotype observed in ccRCC
510 was significantly associated with mutations in the chromatin remodeler *PBRM1*,
511 suggesting a mechanistic link between epigenetic dysregulation and impaired CTG
512 activation in specific tumor contexts.
513
514 A key innovation of this study is the identification and validation that malignant cells
515 consistently exhibit higher CTG-counts than normal cell types across more than 20
516 cancer types, as revealed by scRNA-seq, snATAC-seq, and ST data. While prior studies
517 have reported that tumor cells generally display elevated gene-counts[67–69] (**Figure**
518 **S10B**), our findings demonstrate that elevated CTG-counts in malignant cells are not

merely a byproduct of global transcriptional increases. Indeed, CTG-counts exhibited significantly greater discriminatory power for malignant cell identification compared to random-gene counts, calculated from size- and expression-matched non-CTG control gene sets (**Figure S10A**). Furthermore, even after adjusting for transcriptional activity, CTG-counts remained significantly elevated in malignant cells compared to normal cells, whereas not for adjusted random-gene counts (**Figure S10C**), further highlight the obligate participation of CTGs in multiple hallmarks of cancer[70].

Accurate discrimination of malignant cells from normal cell populations remains a significant challenge in single-cell cancer analyses. Numerous tools have been developed to address this challenge, including aneuploidy-based methods (e.g., inferCNV, CopyKAT[71]) and marker-based approaches (e.g., ikarus[56], scCancer2[46]). Despite these advancements, the potential of CTGs for identifying malignant cells remains unexplored. Motivated by the consistent higher CTG-counts in malignant cells, we developed Single-CTG, a pipeline that quantifies the transcriptomic similarity of individual cells to CTG-counts high malignant cells for robust malignant cell classification. Across multiple cancer types and cohorts, Single-CTG demonstrated performance comparable to the widely used inferCNV, while achieving substantially faster runtimes, particularly for large datasets. This method provides a scalable and efficient tool for large-scale malignant cell identification and enables rapid exploratory analyses without requiring the selection of reference normal cells, a prerequisite for inferCNV.

In summary, we present a refined CTG catalog and delineate its multifaceted biological properties. As multi-omics cancer datasets continue to expand, our CTG compendium can be further updated to capture dynamic expression programs across tumor development and progression. Previous studies have suggested that, in addition to the testis, genes normally restricted to other immune-privileged tissues, such as the placenta[72], also contribute to cancer progression. Extending similar methodologies to

548 investigate these cancer-associated genes could provide a more comprehensive
549 understanding of cancer biology and uncover new avenues for therapeutic exploration.
550
551

552 # Methods

553 **Cancer sc/snRNA-seq datasets collection and preprocessing**

554 A total of 96 well-annotated cancer sc/snRNA-seq datasets were collected (**Table 2**),
555 primarily sourced from the 3CA, HTAN, TISCH2, and Single Cell Portal
556 (https://singlecell.broadinstitute.org/single_cell). For dataset processing, the following
557 criteria were applied: (1) datasets encompassing distinct cancer subtypes or generated
558 by different sequencing platforms were stratified into separate entries; (2) only tumor
559 samples containing ⩾10 malignant cells and adjacent normal samples were retained;
560 (3) specimens obtained from the same patient at multiple timepoints (e.g., pre- and post-
561 treatment) or anatomical sites (e.g., tumor core and edge) were treated as independent
562 samples when adequate metadata were available.
563

564 The **discovery cohort** was constituted from 62 publicly available 10x Genomics
565 sc/snRNA-seq datasets, encompassing 828 tumor samples spanning 13 cancer types:
566 GBM, TNBC, non-TNBC, CRC, HNSCC, ccRCC, HCC, LUAD, NB, HGOSC, PDAC,
567 PRAD, and SKCM. This cohort was utilized for the majority of analyses within this
568 study, including CTG identification, examination of CTG co-expression patterns, CTG-
569 count associated analyses, and investigation of the relationship between CTG
570 expression levels in malignant cells and immune cell infiltration. **Validation cohort 1**
571 was assembled from 13 Smart-seq2 datasets, containing 172 tumor samples across 9
572 cancer types: GBM, TNBC, non-TNBC, CRC, HNSCC, LUAD, HGOSC, PRAD, and
573 SKCM. This validation cohort was specifically employed to confirm CTG expression
574 patterns in malignant cells and to perform CTG-count associated analyses. **Validation**
575 **cohort 2** was composed of 21 sc/snRNA-seq datasets generated using diverse

sequencing platforms (including 10x Genomics, Smart-seq2, inDrop, Drop-seq, Microwell array, and Seq-Well), representing 17 distinct cancer types: GBM, PDAC, PRAD, HGSOC, LUAD, lung squamous cell carcinoma, lung large cell carcinoma, lung pleomorphic carcinoma, SCLC, papillary RCC, Wilms tumor, chromophobe RCC, nasopharyngeal carcinoma, endometrial cancer, intrahepatic cholangiocarcinoma, cutaneous squamous cell carcinoma, and cutaneous basal cell carcinoma. Validation cohort 2 was exclusively utilized for CTG-count associated analyses.

Normalization, feature selection, dimensionality reduction and visualization

Given that all the datasets have already gone through strict quality control in their corresponding original analysis, no additional quality control procedure is performed in this analysis.

All the procedures were conducted using the Seurat[73] R package (version 4.4.0). For each dataset, the raw UMI count or TPM expression matrix were normalized using NormalizeData function with LogNormalize method to reduce the influence of different count depth per cell. Later, top 2,000 highly variable genes (HVGs) were identified using vst method in FindVariableFeatures function. Subsequently, the expression matrix was scaled using ScaleData function and principal component analysis (PCA) was performed for dimensionality reduction using RunPCA function. Lastly, a two-dimensional embeddings was created with RunUMAP function on the top 30 principal components (PCs) for visualization.

When the analysis restricted to normal cells, after PCA analysis, harmony[74] (version 1.2.0) was run with default parameters to remove potential batch effects between different samples within a dataset. Then UMAP analysis was performed using the top 30 batch-corrected PCs.

606 **Cell identity validation and immune subtype annotation**

607 To harmonize the original cell annotations provided in each dataset, all cells were
608 classified into eight major categories: **Normal Epithelial**, **Stromal**, **Immune**, **Brain**
609 **Glial**, **Skin Melanocyte**, **Malignant**, **Other** (cell types not falling into the preceding
610 categories, e.g., adipocytes and myocytes), and **Unknown** (e.g., low-quality cells or
611 cells lacking clear identity).

612

613 To validate the identity of the five major normal cell types (Normal Epithelial, Stromal,
614 Immune, Brain Glial, and Skin Melanocyte), we performed two complementary
615 analyses for each dataset. First, we assessed whether each cluster primarily comprised
616 cells of the same annotated type on the UMAP (restricted to normal cells, following
617 batch correction). Second, differentially expression analysis using FindAllMarkers
618 function in Seurat package was preformed to obtain the top 20 up-regulated for each
619 cell type, then enrichment analysis was performed against CellMarker[75] gene sets using
620 compareCluster function in the clusterProfiler[76] R package (version 4.6.2) to check if
621 the corresponding gene sets were up-regulated in each cell types.

622

623 Malignant cell annotations were validated using two orthogonal criteria: (1) Malignant
624 cells often exhibit patient-specific transcriptional profiles due to high inter-patient
625 heterogeneity; therefore, malignant cells from distinct patients were expected to form
626 separate clusters in the UMAP (all cells, without batch correction), whereas normal
627 cells of the same type should cluster together. (2) Malignant cells typically display
628 extensive copy number alterations. Thus, their inferred CNV scores (see "Inference of
629 copy number variations in sc/snRNA-seq datasets" section) were expected to be
630 significantly higher than those of normal cells.

631

632 Immune cell subtype annotations were refined through automated classification using
633 the SingleR package (v2.0.0). Initial annotation was performed against the reference
634 constructed from Blueprint and ENCODE data to confirm immune lineage identity. One

635 dataset ("Data_Kidney_Obradovic2021") was excluded from downstream immune
636 infiltration analysis, since only 17.5% of its annotated immune cells mapped to immune
637 references. High-resolution subtype annotation was then conducted using the LM22[55]
638 signature matrix as the reference.
639
640
641 **Candidate genes definition and CTGs identification**
642 A total of 276 cancer–testis (CT) antigens were initially retrieved from CTDatabase
643 (http://www.cta.lncc.br/). Each antigen was queried against the HGNC database
644 (https://www.genenames.org/), resulting in 242 genes with approved HGNC symbols
645 and corresponding Ensembl ids. In parallel, testis-enriched protein-coding genes were
646 extracted from the Human Protein Atlas (HPA; https://www.proteinatlas.org/). Genes
647 were selected if their expression level in testis or a tissue group (2–5 tissues including
648 testis) was at least fourfold higher than in any other tissue. A total of 1,212 genes with
649 approved HGNC symbols satisfied this criterion. The union of these two sources
650 yielded 1,294 candidate genes, among which 160 were common to both datasets (**Table**
651 **1**).
652
653 Candidate gene expression in malignant cells was then assessed using 10x Genomics
654 sc/snRNA-seq data from the discovery cohort. For each cancer type, only those
655 candidate genes shared across all relevant datasets were retained. A gene was
656 considered expressed in a given cell if its log1p-normalized expression value exceeded
657 zero. For each tumor sample, the expression level of a candidate gene was quantified
658 as the percentage of malignant cells in which the gene was expressed. If the expression
659 percentage of a candidate genes is bigger than 1% in at least half samples for a given
660 cancer type, this gene was defined as CTGs for that cancer type.
661
662
663 **Inference of copy number variations in sc/snRNA-seq datasets**

Large-scale copy number variations (CNVs) in individual cells were inferred on the sample level using the infercnv R package (version 1.17.0). For each sample, the raw count matrix was used as input, and reference cells were defined as the "Stromal" or "Immune" population with the highest number of cells. When only normalized expression matrices were available, platform-specific rescaling was implemented: UMI-based platforms (i.e., 10x Genomics, inDrop, Drop-seq, Microwell-seq, and Seq-Well) were converted to transcripts per 10,000 (TP10k = TPM/100), while full-length platforms (SmartSeq-2) were converted to transcripts per 100,000 (TP100k = TPM/10). The resultant matrices were subsequently utilized as input. In samples where only malignant cell annotations were available, all non-malignant cells were designated as reference cells. The minimum average read count threshold among reference cells was set to 0.1 for UMI-based datasets and 1.0 for full-length datasets, in accordance with the recommendations of inferCNV.

Cell-level CNV quantification was achieved by calculating a CNV score for each cell, defined as the mean of the squared gene-level CNV values. Gene-level CNV values were derived by subtracting 1 from the residual expression values. This metric reflects the deviation of gene expression from the inferred diploid baseline state and serves as a proxy for the magnitude of large-scale chromosomal alterations within individual cells.

Construction of pseudo-bulk transcriptome profiles of malignant cells

For each tumor sample i in the discovery cohort, expression values from all malignant cells were aggregated by applying Seurat's AggregateExpression function to the exponentiated log1p-normalized counts, thereby producing a pseudo-bulk profile per sample. For each cancer type c, after intersecting genes across datasets, the aggregated counts were converted to counts per million (CPM), generating a pseudo-bulk

692 expression matrix (genes × samples) specific to *c*. An identical procedure was applied
693 to malignant cells within the validation cohort 1, except that aggregated counts were
694 normalized to TPM.
695
696
697 **Bulk RNA-seq datasets collection and related analyses**
698 Multiple TCGA, TARGET, and GTEx bulk RNA-seq datasets were obtained from
699 UCSC Xena (https://xenabrowser.net/datapages/), and dataset selection was determined
700 by specific analytical objectives as detailed in **Table 3**. CCLE bulk RNA-seq data were
701 retrieved from the DepMap portal (https://depmap.org/portal/).
702
703 For each sample, the expression of CTG sets and candidates genes was quantified using
704 single-sample Gene Set Enrichment Analysis (ssGSEA)[77], implemented via the gsva
705 function from the GSVA[78] package (version 1.46.0). Only protein-coding genes within
706 the expression matrices were included in ssGSEA score calculation. Differential
707 expression of CTG sets between tumor (TCGA) and normal (TCGA and GTEx)
708 samples was evaluated using ROC-AUC values, computed with the auc function from
709 the pROC[79] package (version 1.18.5). Consensus measurement of purity estimations
710 (CPE)[80] were adopted to represent tumor purity for TCGA samples, and Spearman's
711 rank correlation coefficients were computed to assess the associations between CTG
712 set expression and tumor purity within each cancer type.
713
714 CTG-based UMAP analysis was performed on batch-normalized TCGA expression
715 data. Specifically, $\log_2$-transformed expression values for CTGs were first centered and
716 scaled via the ScaleData function in Seurat, followed by principal component analysis
717 via RunPCA. Two-dimensional UMAP projections were then derived from the top 30
718 PCs using the RunUMAP function. Expression levels of individual CTGs in
719 non-testicular tissues were summarized as the median transcripts per million (TPM)
720 across all GTEx samples per tissue.

721

722

723 **Sample specificity analysis of CTG expression**

724 For a given gene g in cancer type C, let x_i denotes its expression in sample i (i = 1, 2, ...,
725 N), where N is the total number of samples in C. The tau (τ) index for gene g was
726 calculated as:

727 $$\tau = \frac{\sum_{i=1}^{N}(1-x_i/x_{\max})}{N-1}$$

728 where x_{max} representing the maximum expression observed across all samples. In the
729 discovery cohort, each x_i was measured as the percentage of malignant cells in sample
730 i exhibiting detectable expression of gene g. In CCLE dataset, x_i was taken as the $\log_2$-
731 transformed TPM value for gene g, and only solid tumor types represented by more
732 than 30 cell lines were included.

733

734 To facilitate comparative analysis of sample specificity, CTGs were stratified into four
735 categories: (1) "Autosomal CTGs (Identified)", comprising CTGs located on
736 autosomes that were detected by our pipeline; (2) "Autosomal CTGs (Unidentified)",
737 comprising autosomal CTGs annotated in CTdatabase but not recovered by our pipeline;
738 (3) "X-CTGs (Identified)", comprising CTGs located on the X chromosome that were
739 detected by our pipeline; and (4) "X-CTGs (Unidentified)", comprising X-CTGs
740 annotated in CTdatabase but not recovered by our pipeline.

741

742

743 **Cancer specificity analysis of CTG expression**

744 Tau index was similarly employed to quantify cancer-specific expression patterns.
745 While identical in computation to the sample specificity analysis, x_i here corresponded
746 to the median expression percentage of g across all tumor samples for cancer type i,
747 with N representing the total number of cancer types and x_{max} representing the
748 maximum median expression observed across cancer types. In TCGA data, x_i was
749 defined as the median TPM for g across all tumors in cancer type i. A CTG was

classified as cancer-specific ($\tau > 0.9$) for the cancer type exhibiting maximal median expression, conditional upon its prior identification as a CTG in that cancer type.

Inference of CTG co-expression

The CS-CORE R package (version 0.0.0.9000) was employed to infer co-expression patterns of CTGs with each dataset individually. As this method requires cell-level sequencing depth information, only 10x Genomics datasets containing raw UMI counts were included. For each cancer type, only CTGs identified in that type were used for co-expression analysis. Within each dataset, the CS-CORE estimate was used to quantify the co-expression strength between CTG pairs. CTG pairs with Benjamini–Hochberg adjusted p-values ≥ 0.01 were considered not significant, and their CS-CORE estimates were set to zero. To ensure directional consistency, the CS-CORE estimates for any given CTG pair across datasets within a cancer type should be either both ≥ 0 or both ≤ 0. To summarize results across datasets for a given cancer type, the median CS-CORE estimate for each CTG pair was calculated and used as its representative co-expression score.

Co-expression networks were visualized using the ggraph R package (version 2.1.0). Only CTG pairs with a median CS-CORE estimate greater than 0.25 were retained. CTG clusters within each network were identified using the cluster_louvain function with default parameters from the igraph package (version 1.5.1).

Classification of cycling malignant cells and identification of cell cycle-associated CTGs

For each dataset, AUCell[81] (v1.20.2) was employed to quantify the activity of G2/M and S-phase gene sets with malignant cells. Let x and y denote the AUCell scores for

the G2/M and S phase gene sets, respectively. As the majority of malignant cells were non-cycling, a malignant cell i was classified as cycling if

$$x_i > \text{median}(x) + 5 \times \text{MAD}(x) \quad \text{or} \quad y_i > \text{median}(y) + 5 \times \text{MAD}(y),$$

where MAD refers to median absolute deviation.

For each dataset d within cancer type c, differential expression analysis was conducted between cycling and non-cycling malignant cells, focusing exclusively on CTGs identified in cancer type c. The Wilcoxon rank-sum test was performed using the FindMarkers function in Seurat, with pseudocount.use = 1e-9 to account for the generally low expression levels of CTGs. Multiple hypothesis testing was corrected using the false discovery rate (FDR) method. A CTG was deemed significantly upregulated in dataset d if it met all of the following criteria: FDR < 0.01, average $\log_2$ fold-change > 1, and a difference in the fraction of expressing cells > 0.1. Within each cancer type c, CTGs up-regulated in ≥ 50% of datasets were designated as cell cycle–associated; for cancer types with fewer than three datasets, the threshold was adjusted to ≥ 2 datasets.

Enrichment Analysis of Genes Correlated with Multiple CTGs

For each cancer type c, CTGs identified in c were designated as set M, while non-CTGs (genes not classified as CTGs in any malignancy) were defined as set N. To systematically identify non-CTGs with potential multi-CTG interactions, Spearman's rank correlation coefficients (ρ) were computed between each non-CTG $n_i \in N$ and all CTGs $m_j \in M$ using sample level expression profiles. Strong positive correlations were stringently filtered using dual criteria ($\rho > 0.5$ and Benjamini-Hochberg adjusted $p < 0.01$). Non-CTGs were then rank-ordered based on the count of significantly positively correlated CTGs, retaining top-ranked non-CTGs per cancer type for downstream analysis. This process generated a gene set list where each element represented the top non-CTGs correlated with multiple CTGs for a cancer type.

Functional enrichment analysis was performed using the compareCluster function in the clusterProfiler package (v4.6.2) with the REACTOME[82] pathway database. Specifically, for pseudo-bulk samples aggregated from malignant cells in sc/snRNA-seq datasets, top 2000 non-CTGs were used; for TCGA/TARGET bulk RNA-seq datasets, top 500 non-CTGs were used.

Differential gene expression analysis and gene set enrichment analysis (GSEA)

DESeq2[83] (v1.38.3) was used to identify differentially expressed genes (DEGs) between *PBRM1* non-silent mutated and wild-type ccRCC tumors. The DEGs between *YEATS2* knockdown and control H1299 cells was obtained from GSE90781. The DEGs between *ARID4B* knockdown and control PC3 cells was obtain from GSE116668. DEGs were defines with absolute log2(Fold change) ≥ 0.25 and FDR ≤ 0.05. GSEA for the DEGs were conducted using clusterProfiler package. CTG set we defined and the CTG candidate genes were used for GSEA.

Calculation of CTG-count, gene-count, random gene-count, and CTG expression in sc/snRNA-seq datasets

For each cell *i* in sc/snRNA-seq datasets, CTG-count was defined as the number of CTGs with detectable expression (counts > 0). Gene-count was similarly defined as the total number of expressed genes (counts > 0) per cell. To derive random gene-count, all genes were stratified into 24 expression bins based on their mean expression levels across all cells. For each CTG, a non-CTG gene with matched expression levels was randomly selected from the same bin, resulting in a background gene set that recapitulated the global expression distribution of the CTG set. This procedure was repeated 100 times to generate 100 independent gene sets. For each cell *i*, the number of genes with non-zero expression was computed for each random gene set, and the median value across the 100 iterations was used as random gene-count for that cell. CTG expression score for each cell was computed using the AddModuleScore function

in Seurat. All analyses were performed at the sample level. When comparing differences in these variables between malignant and non-malignant cells at the sample level, only samples containing at least 10 malignant and 10 non-malignant cells were included.

To extend the CTG-count metric to ST data, the same definition was applied to each spatial spot, where CTG-count was the number of CTGs with detectable expression (counts > 0). In snATAC-seq data, gene activity was first inferred by aggregating fragment counts within gene bodies and promoter regions using the GeneActivity function implemented in the Signac[83] (v1.10.0) package. CTG-count was then defined as the number of CTGs with non-zero gene activity per cell.

Benchmarking of malignant cell identification methods

To evaluate the performance of our method in identifying malignant cells, we compared it against three established approaches: inferCNV, ikarus, and scCancer2. CNV scores were derived from the output of the inferCNV package, following the procedure detailed in the "Inference of copy number variations in sc/snRNA-seq datasets" section. The ikarus score, reflecting the predicted malignancy probability of each cell, was obtained by applying the default tumor/normal marker gene sets and pretrained classifier included in the ikarus pipeline. For scCancer2, we employed the author-provided XGBoost classifier on a scaled expression matrix composed of the top 5,000 highly variable genes, and the predicted probability of malignancy for each cell was recorded as the scCancer2 score. All benchmarking analyses were conducted at the sample level to minimize the influence of inter-sample batch effects.

ST datasets collection and processing

A total of 245 tumor sections spanning 15 cancer types were curated from publicly available repositories, including GEO, SCAR[84], HTAN, Mendeley Data, and CROST[85] (see **Table 2** for detailed metadata). For each section, the raw count matrix was

normalized and scaled using the SCTransform function. Dimensionality reduction was then performed using principal component analysis (PCA) via the RunPCA function. A shared nearest-neighbor (SNN) graph was constructed based on the top 30 principal components using FindNeighbors, followed by clustering with the original Louvain algorithm (FindClusters) at a resolution of 1.5. All above analyses were conducted using the Seurat R package (v4.4.0).

ESTIMATE scores were computed for each spatial spot using the ESTIMATE R package (v1.0.13), and tumor purity was inferred using a previously validated formula[51]:

$$\textit{Tumor purity} = \cos\left(0.6049872018 + 0.0001467884 \times \textit{ESTIMATE score}\right)$$

Inference of copy number variations in ST data

Copy number variation inference for each ST section was performed as described by Xun *et al*[86]. To define appropriate reference (i.e., normal) spots, immune scores were calculated for each spot using the AddModuleScore function in Seurat, based on canonical immune cell markers (*PTPRC*, *CD2*, *CD3D*, *CD3E*, *CD3G*, *CD5*, *CD7*, *CD79A*, *MS4A1*, *CD19*). For each section, the cluster with the highest average immune score, as determined from the prior clustering results, was considered the candidate reference cluster. To ensure immune enrichment of the selected reference, a Wilcoxon rank-sum test was performed to compare immune scores of the candidate cluster against all other spots. Only when the immune score was significantly higher (ROC-AUC > 0.6 and FDR < 0.01) was the cluster accepted as the reference for CNV analysis.

Subsequently, infercnv package was applied at the sample level using the raw count matrix as input, with the following parameters: cutoff = 0.1, analysis_mode = "subclusters", cluster_by_groups = FALSE, denoise = TRUE, HMM = TRUE, HMM_type = "i6", BayesMaxPNormal = 0, and tumor_subcluster_partition_method = "random_trees". The random tree partitioning approach divides observation spots into eight CNV-defined subclusters, enhancing discrimination between malignant and non-

malignant regions. Given a spot s and a gene $i \in \{1, 2, ..., N\}$, where x_i denotes the inferred CNV state of gene i, the CNV score of spot s was computed as:

$$CNV\ score(s) = \sum_{i=1}^{N} |x_i - 3|$$

All resulting CNV clusters, including the normal reference cluster, were ranked by their average CNV scores. The two clusters with the highest average scores were designated as CNV-high, whereas the two with the lowest were defined as CNV-low.

Survival analysis

To systematically identify genes associated with patient survival, univariate Cox proportional hazards regression was performed for each gene within each cancer type using the coxph function from the survival package (v3.5.7). For a given gene i in cancer type c, the model was specified as:

Overall Survival (months) ~ *Expression of gene i*.

Benjamini–Hochberg false discovery rate (FDR) correction was applied within each cancer type to control for multiple hypothesis testing, and genes with $FDR < 0.05$ were defined as survival-associated genes. The enrichment of survival-associated genes was then assessed between two predefined gene sets: (1) CTGs identified in cancer type c (set M), and (2) non-CTGs that were not classified as CTGs in any cancer type (set N, as defined in the "Enrichment Analysis of Genes Correlated with Multiple CTGs" section). Proportional differences in enrichment between sets M and N were assessed using Fisher's exact test.

The association between the overall expression level of CTGs and survival in cancer type c was assessed using the Cox model:

Overall Survival (months) ~ *ssGSEA score of set M*.

Additionally, patients were stratified into "High" and "Low" groups based on whether their ssGSEA scores were above or below the median. Kaplan–Meier (KM) survival curves for the two groups were generated using the survfit function, and differences

between two survival curves were tested using the log-rank test via the survdiff function in the survival package.

Characterization of CTG-high malignant hepatocytes in HCC

To investigate the cellular and molecular features of CTG-high malignant cells in HCC, one discovery dataset (GEO: GSE149614) and an independent validation dataset (CNGBdb: CNP0000650) were analyzed. In the validation dataset, cell-type annotations provided by the original authors were evaluated using SingleR against the Blueprint and ENCODE reference. Due to insufficient granularity of T cell subclassification in the original metadata, re-annotation was performed using the Monaco Immune Data reference. The identification of Tregs was subsequently validated based on the expression of canonical markers *CD4*, *IL2RA*, and *FOXP3*.

In both datasets, the expression levels of the HCC CTG signature and the 50 hallmark gene sets in malignant cells were quantified using the AddModuleScore function in Seurat. Malignant cells were stratified into three groups based on CTG expression levels: “High” (top 25%), “Middle”, and “Low” (bottom 25%). Differential expression between “High” and “Low” groups was assessed using the Wilcoxon rank-sum test via Seurat’s FindMarkers function, and genes with $|avg_log_2FC| > 1$, adjusted $p < 0.01$ and min.pct > 0.01 were considered significant. Differences in hallmark[87] pathways were assessed using two-sided Wilcoxon tests with FDR correction and effect sizes were estimated by ROC-AUC values. Developmental potential of malignant cells was inferred using the cytotrace2 function from the CytoTRACE2 package (v1.0.0).

Statistical analysis

If not specified, the two-sided Wilcoxon rank-sum test was used to was performed to analyze the differences between the two groups. Spearman’s correlation test was used to assess the correlations between two variables. A two-tailed *P*-value of 0.05 was

considered statistically significant. R 4.2.3 was used for the entire data processing, statistical analysis, and plotting procedures.

Data Availability

All datasets utilized in this study are publicly available. The corresponding repositories and accession numbers are provided in Tables 2 and 3.

Code Availability

R scripts used to analyze data and generate figures are available upon request to the corresponding author.

Acknowledgments

The research is supported by National Natural Science Foundation of China (Fund 32370715), Science Fund Program for Distinguished Oversea Young Scholars of Shandong Province (2023HWYQ-015), and the Cheeloo Young Scholar Program of Shandong University. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

982 ## Authors' contribution

983 C.F. and K.L. conceived and designed the study. C.F. performed formal analysis. All
984 authors contributed to writing, reviewing, and editing the manuscript and approved the
985 manuscript.
986
987
988

989 ## Competing Interests

990 The researchers declare no competing interests.
991
992
993

994 ## References

995 1. Gibbs, Z. A. & Whitehurst, A. W. Emerging Contributions of Cancer/Testis Antigens to
996 Neoplastic Behaviors. *Trends in Cancer* **4**, 701–712 (2018).
997 2. Gordeeva, O. Cancer-testis antigens: Unique cancer stem cell biomarkers and targets for
998 cancer therapy. *Seminars in Cancer Biology* **53**, 75–89 (2018).
999 3. Fan, C. *et al.* Cancer/testis antigens: from serology to mRNA cancer vaccine. *Seminars in*
1000 *Cancer Biology* **76**, 218–231 (2021).
1001 4. Yang, P., Meng, M. & Zhou, Q. Oncogenic cancer/testis antigens are a hallmarker of
1002 cancer and a sensible target for cancer immunotherapy. *Biochimica et Biophysica Acta (BBA)*
1003 *- Reviews on Cancer* **1876**, 188558 (2021).
1004 5. Deng, Q. *et al.* RNA Interference Against Cancer/Testis Genes Identifies Dual Specificity
1005 Phosphatase 21 as a Potential Therapeutic Target in Human Hepatocellular Carcinoma.
1006 *Hepatology* **59**, 518–530 (2014).
1007 6. Wang, C. *et al.* A cancer-testis non-coding RNA LIN28B-AS1 activates driver gene
1008 LIN28B by interacting with IGF2BP1 in lung adenocarcinoma. *Oncogene* **38**, 1611–1624 (2019).

1009 7. Xie, Z. *et al.* USP4-mediated CENPF deubiquitylation regulated tumor metastasis in
1010 colorectal cancer. *Cell Death Dis* **16**, 81 (2025).
1011 8. Xia, A. *et al.* The cancer-testis lncRNA LINC01977 promotes HCC progression by
1012 interacting with RBM39 to prevent Notch2 ubiquitination. *Cell Death Discov.* **9**, 169 (2023).
1013 9. Xu, J. *et al.* Systematic cancer-testis gene expression analysis identified CDCA5 as a
1014 potential therapeutic target in esophageal squamous cell carcinoma. *eBioMedicine* **46**, 54–65
1015 (2019).
1016 10. AlHossiny, M. *et al.* Ly6E/K Signaling to TGFβ Promotes Breast Cancer Progression,
1017 Immune Escape, and Drug Resistance. *Cancer Research* **76**, 3376–3386 (2016).
1018 11. Zhang, S.-M. *et al.* KDM5B promotes immune evasion by recruiting SETDB1 to silence
1019 retroelements. *Nature* **598**, 682–687 (2021).
1020 12. Deng, S. *et al.* ITPRIPL1 binds CD3ε to impede T cell activation and enable tumor immune
1021 evasion. *Cell* **187**, 2305-2323.e33 (2024).
1022 13. Almeida, L. G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data
1023 on cancer-testis antigens. *Nucleic Acids Research* **37**, D816–D819 (2009).
1024 14. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer
1025 analysis project. *Nat Genet* **45**, 1113–1120 (2013).
1026 15. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–
1027 585 (2013).
1028 16. Wang, C. *et al.* Systematic identification of genes with a cancer-testis expression pattern
1029 in 19 cancer types. *Nat Commun* **7**, 10499 (2016).
1030 17. Carter, J. A. *et al.* Identification of pan-cancer/testis genes and validation of therapeutic
1031 targeting in triple-negative breast cancer: Lin28a-based and Siglece-based vaccination
1032 induces antitumor immunity and inhibits metastasis. *J Immunother Cancer* **11**, e007935 (2023).
1033 18. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand
1034 tumours. *Nature* **618**, 598–606 (2023).
1035 19. Han, Y. *et al.* TISCH2: expanded datasets and new tools for single-cell transcriptome
1036 analyses of the tumor microenvironment. *Nucleic Acids Research* **51**, D1425–D1431 (2023).
1037 20. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions
1038 across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).

1039 21. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
1040 22. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level
1041 relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
1042 23. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of
1043 anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
1044 24. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and
1045 Projection for Dimension Reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426
1046 (2020).
1047 25. Campbell, J. D. *et al.* Genomic, Pathway Network, and Immunologic Features
1048 Distinguishing Squamous Carcinomas. *Cell Reports* **23**, 194-212.e6 (2018).
1049 26. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of
1050 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
1051 27. Pugh, T. J. *et al.* The genetic landscape of high-risk neuroblastoma. *Nat Genet* **45**, 279–
1052 284 (2013).
1053 28. Kortleve, D. *et al.* TCR-Engineered T Cells Directed against Ropporin-1 Constitute a Safe
1054 and Effective Treatment for Triple-Negative Breast Cancer. *Cancer Discovery* **14**, 2450–2470
1055 (2024).
1056 29. Brown, C. E. *et al.* Locoregional delivery of IL-13Rα2-targeting CAR-T cells in recurrent
1057 high-grade glioma: a phase 1 trial. *Nat Med* **30**, 1001–1012 (2024).
1058 30. Wermke, M. *et al.* Abstract PR018: IMA203 TCR-T targeting PRAME demonstrates potent
1059 anti-tumor activity in patients with different types of metastatic solid tumors. *Molecular*
1060 *Cancer Therapeutics* **22**, PR018–PR018 (2023).
1061 31. Olson, D. J. & Odunsi, K. Adoptive Cell Therapy for Nonhematologic Solid Tumors. *JCO*
1062 **41**, 3397–3407 (2023).
1063 32. Yang, J., Chen, Y., Jing, Y., Green, M. R. & Han, L. Advancing CAR T cell therapy through
1064 the use of multidimensional omics data. *Nat Rev Clin Oncol* **20**, 211–228 (2023).
1065 33. Jing, Y. *et al.* Expression of chimeric antigen receptor therapy targets detected by single-
1066 cell sequencing of normal cells may contribute to off-tumor toxicity. *Cancer Cell* **39**, 1558–
1067 1559 (2021).

1068 34. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-
1069 cell RNA-seq.
1070 35. He, J. *et al.* Cancer-testis specific gene OIP5: a downstream gene of E2F1 that promotes
1071 tumorigenesis and metastasis in glioblastoma by stabilizing E2F1 signaling. *Neuro-Oncology*
1072 **20**, 1173–1184 (2018).
1073 36. Marakulina, D. *et al.* EpiFactors 2022: expansion and enhancement of a curated database
1074 of human epigenetic factors and complexes. *Nucleic Acids Research* **51**, D564–D570 (2023).
1075 37. Mi, W. *et al.* YEATS2 links histone acetylation to tumorigenesis of non-small cell lung
1076 cancer. *Nat Commun* **8**, 1088 (2017).
1077 38. Wu, R.-C. *et al.* Identification of the PTEN-ARID4B-PI3K pathway reveals the dependency
1078 on ARID4B by PTEN-deficient prostate cancer. *Nat Commun* **10**, 4332 (2019).
1079 39. Teh, B. T., Farber, L. J. & Furge, K. Molecular Characterization of Renal Cell Carcinoma. in
1080 *Renal Cell Carcinoma* (eds Figlin, R. A., Rathmell, W. K. & Rini, B. I.) 91–111 (Springer US,
1081 Boston, MA, 2012). doi:10.1007/978-1-4614-2400-0_5.
1082 40. Yan, P. *et al.* Midkine as a driver of age-related changes and increase in mammary
1083 tumorigenesis. *Cancer Cell* **42**, 1936-1954.e9 (2024).
1084 41. Yang, C. *et al.* Targeting the immune privilege of tumor-initiating cells to enhance cancer
1085 immunotherapy. *Cancer Cell* **42**, 2064-2081.e19 (2024).
1086 42. Terekhanova, N. V. *et al.* Epigenetic regulation during cancer transitions across 11 tumour
1087 types. *Nature* **623**, 432–441 (2023).
1088 43. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a
1089 transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).
1090 44. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV.
1091 45. Dohmen, J. *et al.* Identifying tumor cells at the single-cell level using machine learning.
1092 *Genome Biol* **23**, 123 (2022).
1093 46. Chen, Z. *et al.* scCancer2: data-driven in-depth annotations of the tumor
1094 microenvironment at single-level resolution.
1095 47. Mo, C.-K. *et al.* Tumour evolution and microenvironment interactions in 2D and 3D space.
1096 *Nature* **634**, 1178–1186 (2024).

1097 48. Gulati, G. S., D'Silva, J. P., Liu, Y., Wang, L. & Newman, A. M. Profiling cell identity and
1098 tissue architecture with single-cell and spatial transcriptomics. *Nat Rev Mol Cell Biol* **26**, 11–
1099 31 (2025).
1100 49. Arora, R. *et al.* Spatial transcriptomics reveals distinct and conserved tumor core and
1101 edge architectures that predict survival and targeted therapy response. *Nat Commun* **14**,
1102 5029 (2023).
1103 50. Liu, Y. *et al.* Identification of a tumour immune barrier in the HCC microenvironment that
1104 determines the efficacy of immunotherapy. *Journal of Hepatology* **78**, 770–782 (2023).
1105 51. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from
1106 expression data. *Nat Commun* **4**, 2612 (2013).
1107 52. Lu, Y. *et al.* A single-cell atlas of the multicellular ecosystem of primary and metastatic
1108 hepatocellular carcinoma. *Nat Commun* **13**, 4594 (2022).
1109 53. Kang, M. *et al.* Mapping single-cell developmental potential in health and disease with
1110 interpretable deep learning. Preprint at https://doi.org/10.1101/2024.03.19.585637 (2024).
1111 54. Sun, Y. *et al.* Single-cell landscape of the ecosystem in early-relapse hepatocellular
1112 carcinoma. *Cell* **184**, 404-421.e16 (2021).
1113 55. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
1114 *Nat Methods* **12**, 453–457 (2015).
1115 56. Malta, T. M. *et al.* Machine Learning Identifies Stemness Features Associated with
1116 Oncogenic Dedifferentiation. *Cell* **173**, 338-354.e15 (2018).
1117 57. Vansteenkiste, J. *et al.* Adjuvant MAGE-A3 Immunotherapy in Resected Non–Small-Cell
1118 Lung Cancer: Phase II Randomized Study Results. *JCO* **31**, 2396–2403 (2013).
1119 58. Vansteenkiste, J. F. *et al.* Efficacy of the MAGE-A3 cancer immunotherapeutic as adjuvant
1120 therapy in patients with resected MAGE-A3-positive non-small-cell lung cancer (MAGRIT): a
1121 randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet Oncology* **17**, 822–
1122 835 (2016).
1123 59. Laisné, M. *et al.* A novel bioinformatic approach reveals cooperation between
1124 Cancer/Testis genes in basal-like breast tumors. *Oncogene* **43**, 1369–1385 (2024).

1125 60. Hong, J. A. *et al.* Reciprocal Binding of CTCF and BORIS to the NY-ESO-1 Promoter
1126 Coincides with Derepression of this Cancer-Testis Gene in Lung Cancer Cells. *Cancer Research*
1127 **65**, 7763–7774 (2005).
1128 61. Gure, A. O. *et al.* Cancer-Testis Genes Are Coordinately Expressed and Are Markers of
1129 Poor Outcome in Non–Small Cell Lung Cancer. *Clinical Cancer Research* **11**, 8055–8062 (2005).
1130 62. Cheng, W. *et al.* Growth differentiation factor 1-induced tumour plasticity provides a
1131 therapeutic window for immunotherapy in hepatocellular carcinoma. *Nat Commun* **12**, 7142
1132 (2021).
1133 63. Simpson, A. J. G., Caballero, O. L., Jungbluth, A., Chen, Y.-T. & Old, L. J. Cancer/testis
1134 antigens, gametogenesis and cancer. *Nat Rev Cancer* **5**, 615–625 (2005).
1135 64. Link, P. A. *et al.* Distinct Roles for Histone Methyltransferases G9a and GLP in Cancer
1136 Germ-Line Antigen Gene Regulation in Human Cancer Cells and Murine Embryonic Stem
1137 Cells. *Molecular Cancer Research* **7**, 851–862 (2009).
1138 65. Karpf, A. R., Bai, S., James, S. R., Mohler, J. L. & Wilson, E. M. Increased Expression of
1139 Androgen Receptor Coregulator MAGE-11 in Prostate Cancer by DNA Hypomethylation and
1140 Cyclic AMP. *Molecular Cancer Research* **7**, 523–535 (2009).
1141 66. Van Tongelen, A., Loriot, A. & De Smet, C. Oncogenic roles of DNA hypomethylation
1142 through the activation of cancer-germline genes. *Cancer Letters* **396**, 130–137 (2017).
1143 67. Bhuva, D. D. *et al.* Library size confounds biology in spatial transcriptomics data. *Genome*
1144 *Biol* **25**, 99 (2024).
1145 68. Lambrechts, D. *et al.* Phenotype molding of stromal cells in the lung tumor
1146 microenvironment. *Nat Med* **24**, 1277–1289 (2018).
1147 69. Moorman, A. *et al.* Progressive plasticity during colorectal cancer metastasis. *Nature* **637**,
1148 947–954 (2025).
1149 70. Maxfield, K. E. *et al.* Comprehensive functional characterization of cancer–testis antigens
1150 defines obligate participation in multiple hallmarks of cancer. *Nat Commun* **6**, 8840 (2015).
1151 71. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from
1152 single-cell transcriptomes. *Nat Biotechnol* **39**, 599–608 (2021).
1153 72. Yu, J. *et al.* Progestogen-driven B7-H4 contributes to onco-fetal immune tolerance. *Cell*
1154 **187**, 4713-4732.e19 (2024).

1155 73. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29
1156 (2021).
1157 74. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with
1158 Harmony. *Nat Methods* **16**, 1289–1296 (2019).
1159 75. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in
1160 human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research* **51**, D870–
1161 D876 (2023).
1162 76. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
1163 *The Innovation* **2**, 100141 (2021).
1164 77. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven
1165 cancers require TBK1. *Nature* **462**, 108–112 (2009).
1166 78. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray
1167 and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).
1168 79. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare
1169 ROC curves. *BMC Bioinformatics* **12**, 77 (2011).
1170 80. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat*
1171 *Commun* **6**, 8971 (2015).
1172 81. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat*
1173 *Methods* **14**, 1083–1086 (2017).
1174 82. Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*
1175 **52**, D672–D678 (2024).
1176 83. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state
1177 analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
1178 84. Deng, Y. *et al.* SCAR: Single-cell and Spatially-resolved Cancer Resources. *Nucleic Acids*
1179 *Research* **52**, D1407–D1417 (2024).
1180 85. Wang, G. *et al.* CROST: a comprehensive repository of spatial transcriptomics. *Nucleic*
1181 *Acids Research* **52**, D882–D890 (2024).
1182 86. Xun, Z. *et al.* Reconstruction of the tumor spatial microenvironment along the malignant-
1183 boundary-nonmalignant axis. *Nat Commun* **14**, 933 (2023).

87. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems* **1**, 417–425 (2015).