communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-09154-w

Multi-omic analysis of hepatocellular carcinoma reveals aberrant *cis*-regulatory changes and dysregulated retrotransposons with prognostic potential

Clooney C. Y. Cheng[1], Ming Fung Cheung[2], Ah Young Lee[2], Qiong Wu[3], Savio Ho-Chit Chow[4], Julie Y. J. Ang[1], Ignacio Riquelme Medina[1], Grace Lo[2], Haoran Wu[3], Weiqin Yang[3], Paul B. S. Lai[5], Kevin Y. Yip[4,6] ✉, Alfred S. Cheng[3] ✉ & Danny Leung[1,2] ✉

Hepatocellular carcinoma (HCC) is the most common form of primary liver cancer and has limited therapeutic options. Epigenetic dysregulation plays a crucial role in hepatocarcinogenesis; however, its impact on *cis*-regulatory elements (CREs) and retrotransposons remains underexplored. Here, we investigated the epigenetic changes underlying the aberrant CRE and retrotransposon activity in HCC. We show that focal DNA hypomethylation of these elements is associated with transcriptional reprogramming. Notably, we uncover a dual regulatory mechanism for *GPC3*, a key diagnostic biomarker and immunotherapeutic target in HCC. This requires the concomitant reactivation of a fetal liver super-enhancer (SE) and DNA hypomethylation of CpG islands. Furthermore, DNA methylation loss drives cryptic activation of retrotransposons, some of which exhibit prognostic potential. Intriguingly, we identify a HERVE-int-derived long non-coding RNA that shows higher expression in patients with more aggressive tumors, poorer disease outcomes, and is correlated with molecular signatures associated with improved response to immunotherapy. Collectively, our findings reveal widespread epigenomic dysregulation of CREs and retrotransposons in HCC, highlighting new potential therapeutic strategies and biomarkers.

Hepatocellular carcinoma (HCC) is the most common form of primary liver cancer and the third leading cause of cancer-related deaths worldwide[1]. The key risk factors for HCC include alcohol abuse, chronic hepatitis B/C virus (HBV/HCV) infections, and non-alcoholic fatty liver disease (NAFLD)[2]. These conditions promote persistent inflammation, liver fibrosis, and cirrhosis that often precede the onset of HCC[2]. With a 5-year survival rate of 18%, prognosis of HCC remains poor due to limited therapeutic options[3]. Sorafenib and lenvatinib have been the standard of care for patients with advanced and unresectable HCC; however, the efficacy of these treatments is modest[4,5]. In recent years, immune checkpoint inhibitors (ICIs), either used individually or in combination with other targeted therapies, have improved patient survival relative to sorafenib[6–8]. Nevertheless, many of these patients do not respond or develop resistance[9]. Combining epigenetic drugs with targeted therapies and/or immunotherapies holds promise for improving patient outcomes[10–14]. The development of these drugs requires a comprehensive understanding of the precise molecular alterations. However, the

[1]Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. [2]Center for Epigenomics Research, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. [3]School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China. [4]Cancer Genome and Epigenetics Program, NCI-Designated Cancer Centre, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [5]Department of Surgery, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China. [6]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China. ✉e-mail: kyip@sbpdiscovery.org; alfredcheng@cuhk.edu.hk; dcyleung@ust.hk

aberrant changes in multiple layers of the epigenome that contribute to transcriptional dysregulation in HCC, including DNA methylation and histone modifications, have not been fully elucidated. Hence, integrative epigenomic profiling is needed to uncover the epigenetic vulnerabilities in this disease, which may be leveraged to maximize patient benefits from existing therapeutic regimens.

Cis-regulatory elements (CREs), including promoters and enhancers, play critical roles in transcriptional reprogramming during tumorigenesis[15]. These sequences are tightly regulated by epigenetic mechanisms. Low DNA methylation, increased chromatin accessibility, and enrichment of specific histone modifications such as histone H3 lysine 27 acetylation (H3K27ac) are features associated with their activation[16–19].These epigenetic mechanisms work in concert with transcription factors (TFs) and coactivators to direct gene expression and facilitate 3D chromatin interactions. Super-enhancers (SEs), which are large enhancer clusters, also show strong enrichment of H3K27ac and coactivators[20,21]. De novo SE acquisition near oncogenes such as *MYC* and *TAL1* can promote pro-tumorigenic signaling[22,23]. Despite the extensive enhancer and SE reprogramming in HCC[24–26], the impact of epigenetic dysregulation on CRE activities and their functional consequences require further investigation.

Comprising ~40% of the human genome, retrotransposons are an abundant reservoir of CREs and contribute to gene regulation through disseminating TF binding sites[27–29]. Although most are epigenetically silenced in normal tissues, some elements can become derepressed in cancer and can have therapeutic implications. For example, global DNA hypomethylation can activate long-terminal repeat (LTR) retrotransposons, including endogenous retroviruses (ERVs), which can function as alternative promoters for oncogenes[30,31]. The resulting chimeric proteins can potentially be targeted as neoantigens to sensitize cancer cells to immunotherapy[30,31]. Furthermore, expression of specific retrotransposon subfamilies can induce anti-tumor immunity and immunotherapy responses[14,32,33]. In HCC, pervasive activation of LTR retrotransposon-driven long non-coding RNAs (lncRNAs), particularly those from the LTR12C subfamily, is associated with HBV etiology and less differentiated tumors[34]. These transcripts can promote DNA double-strand break repair and tumor metastasis[35–37]. Despite the evidence of retrotransposon dysregulation in HCC, the prognostic value of retrotransposon-derived CREs and their clinical relevance to immunotherapy remain largely unexplored.

In this study, we integrated the DNA methylomes, histone modification profiles, and transcriptomes of HCC tumors and matched tumor-adjacent normal tissues to decipher the transcriptional regulatory landscape of HCC. With this multi-omic approach, we defined epigenetically dysregulated CREs and retrotransposons, revealing their functional link to the altered transcriptional programs in HCC. Specifically, we elucidated the dual regulatory mechanism of glypican-3 (*GPC3*), a diagnostic marker and immunotherapeutic target for HCC[38–40]. Furthermore, we showed that patients with high expression of a HERVE-int element-driven lncRNA were associated with significantly poorer outcomes but had molecular signatures associated with better response to immunotherapy.

Results

Widespread DNA methylation and transcriptional dysregulations in HCC

We performed whole genome bisulfite sequencing (WGBS), chromatin immunoprecipitation followed by sequencing (ChIP-seq), and stranded total RNA-sequencing on seven pairs of HBV-associated HCC and matched tumor-adjacent normal (TAN) tissues (Fig. 1a and Supplementary Data 1). All datasets meet the quality standards of the International Human Epigenome Consortium (IHEC) (Supplementary Data 2)[41]. Due to sample availability, some experiments were omitted. To bolster our analyses, we included four publicly available DNA methylomes of TAN tissues[42]. Moreover, we used Avocado, a machine learning algorithm trained on ENCODE data, to impute the missing ChIP-seq datasets[43]. We demonstrated accuracy and reliability of the imputed data by comparing the imputed and observed ChIP-seq signals (Supplementary Fig. 1).

We observed widespread DNA methylation loss at both genic and non-coding regions in HCC (Fig. 1b and Supplementary Fig. 2a, b), of which approximately 78% of the hypomethylated regions (HypoDMRs) overlapped with partially methylated domains (PMDs). These domains were enriched with repressive histone modifications and associated with transcriptional repression (Supplementary Fig. 2c, d)[44–46]. We found heterogeneity in DNA methylation levels at these loci among the HCC samples (Supplementary Fig. 2e). On the other hand, hypermethylated regions (HyperDMRs) in HCC were enriched in the promoters of developmental genes and in genic features such as untranslated regions (UTRs) and exons (Fig. 1b and Supplementary Fig. 2f)[47,48]. Overall, our data are consistent with previous reports of global DNA hypomethylation and focal DNA hypermethylation in HCC[49–52].

RNA-seq revealed upregulation of cell cycle regulators and downregulation of immune- and extracellular matrix-related genes (Fig. 1c and Supplementary Fig. 3a,b,c). Based on transcriptomic signatures, HCC can be categorized into the proliferation and non-proliferation classes. The proliferation class is characterized by the activation of cell proliferation pathways and more aggressive tumors. We found that 25% of the upregulated genes (n = 120) were defined as proliferation class signature genes (Supplementary Fig. 3d)[53,54], most of which (93 out of 120) were associated with cell cycle progression (Supplementary Fig. 3e). Moreover, weighted gene co-expression network analyses (WGCNA) of The Cancer Genome Atlas-Liver Hepatocellular Carcinoma (TCGA-LIHC) cohort showed that some of these dysregulated genes were hub genes within the modules corresponding to HCC or TAN (Supplementary Fig. 3f, g). These results confirm that patients in our cohort have transcriptomic signatures of the proliferation class[50].

Epigenetic reprogramming of CREs drives transcriptional dysregulation in HCC

Next, we examined the epigenetic alterations at the promoters of dysregulated genes (Fig. 1d). Concordant changes of H3K27ac and H3K4me3, but not H3K9me3 or H3K27me3, were detected at their promoters. Notably, approximately half of the upregulated genes (n = 239) exhibited promoter DNA hypomethylation. These genes included *AFP*, a diagnostic marker for HCC, several HCC-related lncRNAs (*LINC01419*, *LINC02163*, and *LINC02835*), and tumor-associated antigens such as the *MAGE* gene family[35–37].

Enhancer and SE rewiring has also been reported in HCC[24–26]. Thus, we further investigated the transcriptional programs linked to these dysregulated elements. We defined 3637 and 2924 typical enhancers that gained and lost activity, respectively (Fig. 2a, Supplementary Fig. 4a, and Supplementary Data 6). Interestingly, while inactivated typical enhancers were broadly shared between most HCC samples, substantial numbers of the activated typical enhancers were restricted to specific patients. This finding aligns with the recently reported HCC enhancer heterogeneity[24]. The activated typical enhancers were associated with developmental gene ontology (GO) terms and were enriched for binding motifs of development-related TFs such as CEBPα/β, NK2-1, and ZIC3 (Fig. 2a, b)[55–57]. In contrast, inactivated typical enhancers were associated with collagen fibril organization and chemotaxis GO terms and were enriched for binding motifs of immune-related TFs, including STAT1 (Fig. 2a, b)[12,13]. We also compared the levels of H3K9me3, H3K27me3, and DNA methylation at the dysregulated typical enhancers in HCC and TAN (Supplementary Fig. 4b). While there was no significant change in H3K27me3, we observed loss of H3K9me3 at the activated typical enhancers.

Next, we defined SEs in HCC and TAN samples and found that SEs in both HCC and TAN were associated with GO terms related to metabolic genes (e.g. *CYP3A4*, *SLC38A4*, and *IGFBP1*) (Fig. 2c)[20,21,58]. This was consistent with the established role of SEs in regulating cell- and tissue-specific processes[20,21,58]. We identified activated and inactivated SEs in HCC (n = 234 and n = 86, respectively) (Fig. 2d, Supplementary Fig. 4c, and Supplementary Data 6). The activated SEs also exhibited heterogeneity across HCC

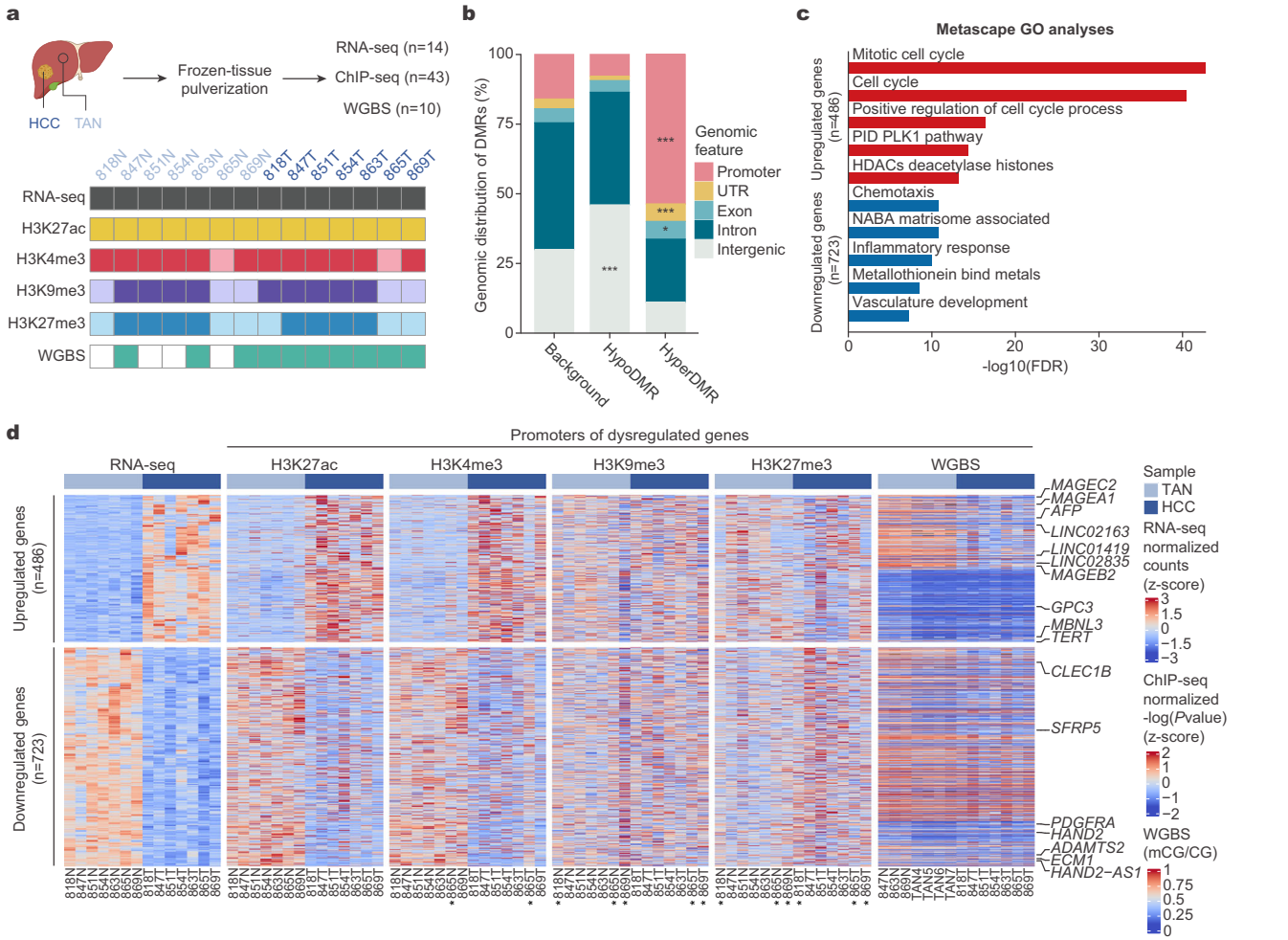


Fig. 1 | Multi-omic profiling reveals widespread DNA methylation alterations and epigenetic changes at promoters of dysregulated genes in HCC. a Schematic of experimental design. Frozen HCC or TAN tissue samples were pulverized and subjected to stranded total RNA-seq, ChIP-seq for four histone modifications (H3K27ac, H3K4me3, H3K9me3, and H3K27me3), and WGBS. Omitted histone marks were imputed using Avocado. Shading of cells indicates experimental (dark) or imputed (light) datasets. **b** Genomic distribution of hypoDMRs and hyperDMRs in HCC. HypoDMR: intergenic ($P = 1.1e{-}15$). $^{*}P < 0.05$ and $^{**}P < 0.001$; HyperDMR: promoter ($P = 2.653e{-}195$), UTR ($P = 3.621e{-}6$) and exon ($P = 0.0404$), one-tailed hypergeometric test corrected by the Benjamini-Hochberg approach. **c** GO analyses for dysregulated genes in HCC. *P* values (one-tailed hypergeometric test) for the top 5 GO terms are shown for both the upregulated and downregulated genes, respectively. **d** Heatmaps showing the expression of dysregulated genes in HCC and the epigenetic signatures of their promoters (± 1 kb of TSS) in HCC and TAN samples. For each heatmap, each row represents a dysregulated gene and each column represents a sample. Selected dysregulated genes are labeled to the right of the WGBS heatmap. Asterisks (*) indicate Avocado-imputed signals.

samples and were associated with metabolism, early development, and extracellular matrix formation GO terms (Fig. 2d). Importantly, SE-associated genes showed concomitant change in expressions (Fig. 2e). For instance, the upregulation of *MBNL3*, which encodes for a splicing factor highly expressed in both fetal livers and in HCC, was associated with the reactivation of an upstream SE (Supplementary Fig. 4d)[59]. Conversely, the downregulation of *ADAMTS2* was coupled with the inactivation of a SE and several typical enhancers located within a PMD in HCC (Fig. 2f). Together, our findings illustrate the epigenetic changes at aberrant CREs and the associated transcriptional networks in HCC.

Aberrant reactivation of SE drives *GPC3* overexpression in HCC

To characterize the functional roles of activated SEs in HCC, we focused on the SE upstream of the glypican-3 (*GPC3*) gene. *GPC3* encodes for a heparan sulfate proteoglycan that regulates cell growth and differentiation of various tissues, including the liver, during embryonic development[60]. While *GPC3* is broadly silenced in normal adult livers, it becomes overexpressed in more than 70% of HCC, which we validated in both our patient samples and the TCGA-LIHC cohort (Supplementary Fig. 5a)[61]. Importantly, *GPC3* is highly expressed in HCC relative to other cancer types (Fig. 3a)[40,61]. These features support *GPC3* as both a diagnostic biomarker and a prognostic indicator. Furthermore, in recent years, GPC3 has emerged as a promising immunotherapeutic target for HCC[62–65].

The *GPC3* SE consisted of seven constituent elements (E1-E7), which exhibited similar open chromatin patterns in fetal hepatoblasts (Fig. 3b). The chromatin accessibility of the SE positively correlated with that of the *GPC3* promoter and its gene expression (Supplementary Fig. 5b). Importantly, this SE was differentially reactivated in HCC but not in other analyzed cancer types (Supplementary Fig. 5c). Furthermore, *GPC3* overexpression was accompanied by significant 3D interactions amongst constituent enhancers as well as between constituent enhancers and the promoter (Fig. 3b)[66,67].

To validate the *cis*-regulatory potentials of E1-E7, we performed luciferase assays to test the enhancer and promoter activities of these elements. We found that most candidates demonstrated strong enhancer but minimal promoter activities (Fig. 3c and Supplementary Fig. 5g). We then used the CRISPRi system to epigenetically silence their activities. This resulted in at least 40% reduction in expression, suggesting that all constituent enhancers contribute to the transcriptional regulation of *GPC3* (Fig. 3d).

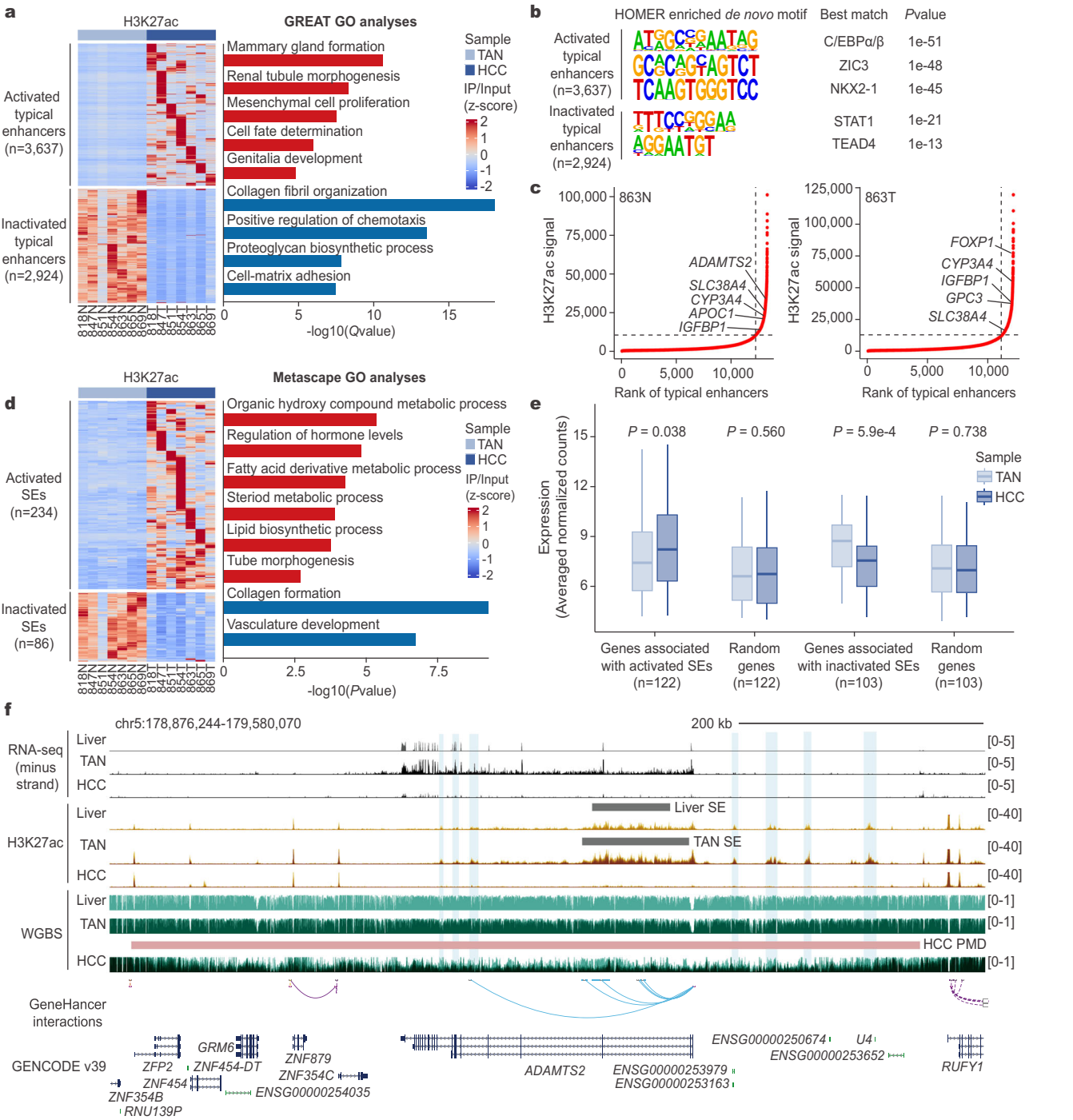
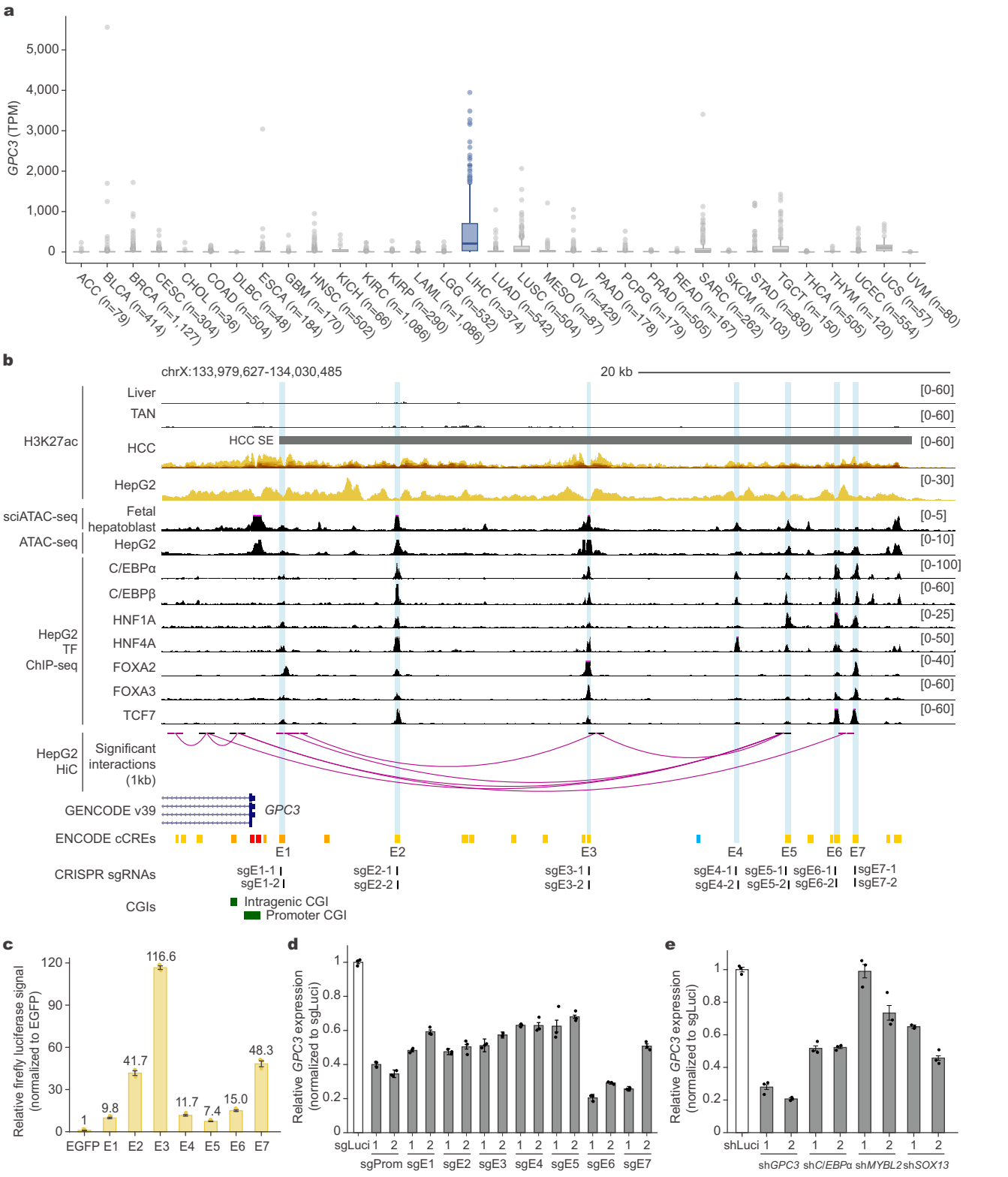


Fig. 2 | Enhancer and SE reprogramming are associated with transcriptional dysregulation in HCC. a Left: heatmap of H3K27ac ChIP-seq signal of dysregulated typical enhancers. Each row represents a typical enhancer and each column represents a sample. Right: Top relevant GO terms associated with the activated and inactivated typical enhancers (red and blue bars, respectively). **b** HOMER motif analyses of the dysregulated typical enhancers. **c** Input-normalized H3K27ac signal distribution of stitched enhancers in HCC and TAN samples from patient 863. Enhancers were ranked in increasing order based on their input-normalized H3K27ac signals by the ROSE algorithm. SEs were defined as enhancers above the inflection point of the curve. Representative genes associated with SEs are labeled. **d** Left: heatmap of H3K27ac ChIP-seq signal of dysregulated SEs. Each row represents a SE and each column represents a sample. Right: Top relevant GO terms associated with the activated and inactivated SE (red and blue bars, respectively). **e** Boxplots comparing the expression of dysregulated SE-associated genes between HCC and TAN. *P* values (two-tailed Wilcoxon test) are shown. The center and bounds of boxes indicate the median and quartile of all data points, respectively. The minima and maxima of whiskers indicate quartile 1–1.5× the interquartile range and quartile 3+1.5 × the interquartile range, respectively. **f** Genome browser screenshot showing an inactivated SE and several adjacent inactivated typical enhancers located within a PMD at the *ADAMTS2* gene locus. The SEs are annotated as gray bars and typical enhancers are highlighted by cyan shadings. The HCC PMD is annotated with a pink bar. Stranded total RNA-seq tracks are displayed as TPM, H3K27ac ChIP-seq tracks are displayed as fold enrichment over input, and WGBS tracks are displayed as mCG/CG. Liver RNA-seq and WGBS tracks are from Roadmap Epigenomics Project and shown as single replicates. Other tracks are shown as composite signals by overlaying tracks across samples. GeneHancer interactions track is from the UCSC genome browser.



Notably, the overexpression of *GPC3* in HCC did not result from copy-number alterations or mutations (Supplementary Fig. 5d–f). Moreover, we observed no strong correlation between the expressions of *GPC3* and its associated lncRNA, *GPC3-AS1*, which has been reported to enhance *GPC3* expression by recruiting coactivators (Supplementary Fig. 5h)[68]. Hence, aberrant reactivation of the SE is a key regulatory mechanism underlying the transcriptional dysregulation of *GPC3* in HCC.

Aberrant reactivation of the *GPC3* SE is associated with loss of DNA methylation and increased binding of key TFs

Next, we asked how epigenetic changes contributed to the reactivation of the *GPC3* SE in HCC. We observed modest change in H3K9me3 and H3K27me3 at most of the constituent enhancers (E3, E4, E5, E6, and E7) (Supplementary Fig. 6a). In contrast, these elements showed extensive loss of DNA methylation (Fig. 4a). Thus, DNA hypomethylation, rather than

Fig. 3 | *GPC3* overexpression coincides with super-enhancer reactivation in HCC. **a** Boxplots showing the expression of *GPC3* across TCGA cancer types. HCC is highlighted in blue while other cancer types are shown in gray. Cancer type abbreviations follow TCGA definitions. The center and bounds of boxes indicate the median and quartile of all data points, respectively. The minima and maxima of whiskers indicate quartile 1–1.5× the interquartile range and quartile 3 + 1.5× the interquartile range, respectively. Pairwise Wilcoxon's tests were performed for LIHC and all other cancer types. All comparisons were statistically significant with *P* values < 0.01. **b** Genome browser screenshot showing the reactivated SE at the *GPC3* locus. The SE is annotated with a gray bar. *GPC3* constituent enhancers (E1-E7) are highlighted by cyan shadings. sciATAC-seq track is shown as fragment counts. H3K27ac ChIP-seq, TF ChIP-seq, and ATAC-seq tracks are displayed as fold enrichment over input. Significant chromatin interactions are shown as arcs. sciATAC-seq for fetal hepatoblasts is from CATlas, ATAC-seq and TF ChIP-seq for HepG2 are from ENCODE, and Hi-C for HepG2 is from 4DN. All tracks are shown as composite signals by overlaying tracks across samples. Genomic positions of sgRNAs targeting individual *GPC3* constituent enhancers are also shown. **c** Luciferase assay assessing the enhancer activities of E1-E7. Data represent the mean ± standard deviation of three technical replicates from one representative experiment (n = 2 independent experiments). **d** RT-qPCR of the relative expression of *GPC3* upon epigenetic silencing of *GPC3* constituent enhancer. Each dot represents a technical replicate and error bars denote standard deviations. sgRNA targeting luciferase (sgLuci) is used as the negative control, whereas two sgRNAs targeting the *GPC3* promoter (sgProm) are used as the positive controls. *GAPDH* serves as the internal control for the calculation of relative *GPC3* expression and the values are normalized to that of sgLuci. **e** RT-qPCR of the relative expression of *GPC3* upon depletion of *C/EBPα*, *MYBL2*, and *SOX13*. Each dot represents a technical replicate and error bars denote standard deviations. shRNA targeting luciferase (shLuci) is used as the negative control, whereas two shRNAs targeting *GPC3* (sh*GPC3*) are used as the positive controls. *GAPDH* serves as the internal control for the calculation of relative *GPC3* expression and the values are normalized to that of shLuci.

depletion of H3K9me3 or H3K27me3, might precede the reactivation of the *GPC3* SE.

In line with the high TF binding density observed at SEs, we detected binding of liver-restricted TFs (*C/EBPα/β*, *HNF1A*, and *HNF4A*), pioneering TFs (*CEBPα/β*, *FOXA2*, and *FOXA3*), and Wnt-signaling TFs (*TCF7*) at individual constituent enhancers (Fig. 3b)[20,21]. To identify the key TFs involved in reactivating the *GPC3* SE, we examined TF ChIP-seq datasets from ENCODE. This analysis revealed 271 TFs that were enriched at one or more constituent enhancers. We then performed co-expression analysis with the TCGA-LIHC datasets, identifying 137 of these TFs whose expression patterns were positively correlated with *GPC3* (*rho* > 0.4 and *P* < 0.05). By overlapping these lists, we identified five candidate TFs: *ARID3A*, *C/EBPα*, *FOXA3*, *MYBL2*, and *SOX13*, most of which are overexpressed in HCC and are implicated in hepatocarcinogenesis (Supplementary Fig. 6b, c)[69–73]. To test their regulatory function on *GPC3*, we performed shRNA knockdown of each candidate (Supplementary Fig. 6d). For *ARID3A* and *FOXA3*, knockdown efficiencies varied substantially between replicates; therefore, we focused on *C/EBPα*, *MYBL2*, and *SOX13*, which showed significant and reproducible knockdown efficiencies (Supplementary Fig. 6d). Interestingly, we found that the knockdown of *C/EBPα* and *SOX13*, but not *MYBL2*, resulted in a pronounced reduction in *GPC3* expression (35–48%) (Fig. 3e). Taken together, these findings demonstrate that *C/EBPα* and *SOX13* contribute to *GPC3* overexpression through activating the *GPC3* SE.

Aberrant reactivation of the *GPC3* SE and DNA hypomethylation of CpG islands are both necessary for the transcriptional dysregulation of *GPC3* in HCC

GPC3 has two CpG islands (CGIs): one proximal to the promoter (promoter CGI) and the other within its first intron (intragenic CGI). These elements are likely to contribute to *GPC3* regulation, as promoter CGI hypermethylation can silence *GPC3* in vitro[74]. We observed DNA hypomethylation at the promoter CGI in the TCGA-LIHC cohort (Fig. 4b). Unlike the promoter CGI, the role of the intragenic CGI in *GPC3* transcriptional regulation has not been reported. Interestingly, we also observed significant DNA hypomethylation at two CpG probes within the intragenic CGI in the TCGA-LIHC cohort (Fig. 4c). Moreover, the methylation levels of all CpG probes within the intragenic CGI were inversely correlated with *GPC3* expression (Fig. 4d). These data suggest that DNA hypomethylation of both promoter and intragenic CGIs are necessary for *GPC3* overexpression in HCC. To test their functionality, we employed the CRISPR-SunTag-DNMT3A1 system to specifically methylate the promoter or intragenic CGIs[31]. Indeed, targeted methylation of either element reduced *GPC3* expression by 45-55% (Supplementary Fig. 7). These results show that loss of DNA methylation at both the promoter and intragenic CGIs are required for the overexpression of *GPC3* in HCC.

Our observations hint at an interplay between the DNA methylation loss at SE and CGIs and *GPC3* transcriptional upregulation (Fig. 4e, f). Firstly, SE reactivation without hypomethylated CGIs resulted in minimal *GPC3* expression (HCC samples 818, 865, 847, and 854). Notably, methylation of either CGI (HCC samples 818 and 865) was sufficient to reduce *GPC3* expression. On the other hand, despite having lowly methylated CGIs, patients without SE reactivation exhibited relatively low *GPC3* expression (HCC samples 851 and 869). Importantly, *GPC3* was most highly expressed with concordant SE reactivation and CGI hypomethylation (HCC sample 863). Thus, we conclude that aberrant reactivation of the SE and DNA hypomethylation of both CGIs are concomitantly required for *GPC3* overexpression in HCC.

Epigenetic dysregulation of retrotransposon-derived CREs in HCC

Retrotransposons are an abundant source of cryptic regulatory elements in cancers[13,30,75–77]. To study dysregulated retrotransposon-derived CREs, we overlapped retrotransposons annotated in RepeatMasker with the dysregulated CREs in HCC. We identified 3,167 and 1,361 dysregulated retrotransposon-derived typical enhancers (RT-typical enhancers) promoters (RT-promoters), respectively (Supplementary Fig. 8a, b and Supplementary Data 7). Consistent with the enhancer/SE heterogeneity in HCC, we observed varying enrichment of H3K27ac at both dysregulated RT-typical enhancers and RT-promoters amongst patients (Fig. 2a, d and Supplementary Fig. 8a, b). Similarly, dysregulated RT-typical enhancers were also associated with metabolism, early development, and extracellular matrix organization GO terms (Fig. 4a and Supplementary Fig. 8c, d). These elements were enriched for binding motifs of development-related TFs, including that of Wnt signaling (*TCF3* and *TCF7L1*) and Hippo pathways (*TEAD1/2/3/4*) (Supplementary Fig. 8e). In contrast, dysregulated RT-promoters were enriched for binding motifs for ubiquitously expressed TFs, including those for *YY1*, *KLF*, and *NF-Y* (Supplementary Fig. 8e).

Next, we examined the class distribution of the dysregulated retrotransposon-derived CREs (Supplementary Fig. 8f). Interestingly, whereas Long Interspersed Nuclear Elements (LINEs) were over-represented in dysregulated RT-typical enhancers, LTRs were enriched in both dysregulated RT-typical enhancers and RT-promoters. For the activated RT-typical enhancers, 42 out of 51 over-represented subfamilies were LTR retrotransposons. These included LTR2B, LTR8B, and LTR10A, which regulate genes critical in placental development (Supplementary Fig. 8g and Supplementary Data 7)[78,79]. Similarly, for the activated RT-promoters, 32 out of 38 over-represented subfamilies were LTRs. These include LTR12C, which is pervasively activated in HBV-associated HCC (Supplementary Fig. 8h and Supplementary Data 7)[34].

DNA hypomethylation leads to cryptic activation of retrotransposon-driven transcripts associated with poorer disease outcomes

We further overlapped de-repressed retrotransposons ($log_2FC > 3.5$ and $P_{adj} < 0.01$, Supplementary Data 7) with activated RT-promoters,

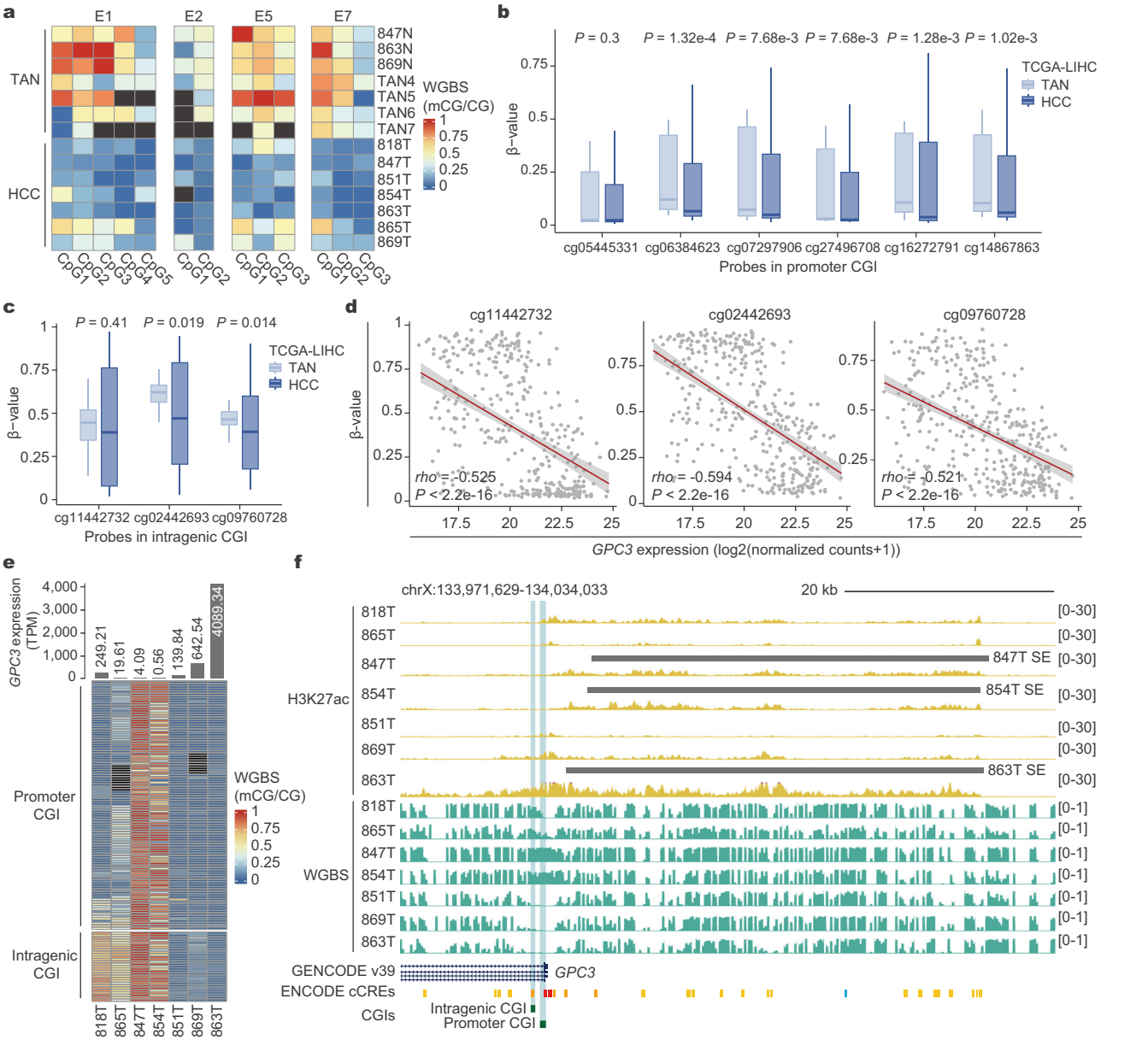


Fig. 4 | SE reactivation and DNA hypomethylation of CGIs work in concert to regulate *GPC3* expression. **a** Heatmaps comparing the DNA methylation of constituent enhancers with at least 2 CpGs (E1, E2, E5, and E7) between HCC and TAN. Each row corresponds to a sample and each column corresponds to a CpG within the constituent enhancer. Dark gray cells indicate CpGs with insufficient read coverage (< 5). Boxplots comparing the DNA methylation of CpGs within the *GPC3* promoter CGI (**b**) and intragenic CGI (**c**) between HCC and TAN in the TCGA-LIHC cohort. *P* values (two-tailed Wilcoxon test) are shown. **d** Scatterplots showing the correlation between the DNA methylation of CpGs within the *GPC3* intragenic CGI and *GPC3* expression in the TCGA-LIHC cohort. Spearman's correlation coefficient (*rho*) and *P* values are indicated at the bottom left of each scatterplot. **e** Heatmaps showing the DNA methylation of the *GPC3* promoter and intragenic CGIs in each HCC sample. The corresponding *GPC3* expression in each patient sample is shown as barplots at the top of the promoter CGI heatmap. Dark gray cells indicate CpGs with insufficient read coverage (< 5). **f** Genome browser screenshot showing the concomitant reactivation of the SE and DNA hypomethylation of the *GPC3* promoter and intragenic CGIs. SEs are annotated with gray bars. *GPC3* promoter and intragenic CGIs are highlighted by cyan shading. H3K27ac ChIP-seq tracks are displayed as fold enrichment over input and WGBS tracks are displayed as mCG/CG.

identifying 24 candidates, of which, 20 were LTR elements (Fig. 5a). Notably, 11 of these elements overlapped with the promoters of annotated lncRNAs (Fig. 5a), consistent with the previous finding that LTRs are enriched in the transcription start sites (TSSs) of lncRNAs[80]. Cryptic activation of these retrotransposons was associated with loss of repressive histone marks and DNA methylation (Fig. 5a).

To prioritize the candidates for downstream analyses and functional experiments, we stratified patients in the TCGA-LIHC cohort based on the expression of each of the 24 candidates. We discovered that the expressions of two elements, belonging to the HERVE (HERVE-int) and THE1D subfamilies, could significantly differentiate patients' survival (Fig. 5b, c and Supplementary Fig. 9a–c). Patients with high HERVE-int expression (HERVE-int-High, top 25%, n = 93) were associated with poorer progression-free survival (Fig. 5c). The expression of HERVE-int also increased across tumor grades (Fig. 5d). Likewise, patients with intermediate to high expression of THE1D (THE1D-Intermediate, 30%<x < 70%, n = 149; THE1D-High, top30%, n = 112) also showed worse progression-free disease outcomes (Supplementary Fig. 9c). It should be noted that different stratifying criteria were applied for HERVE-int and THE1D

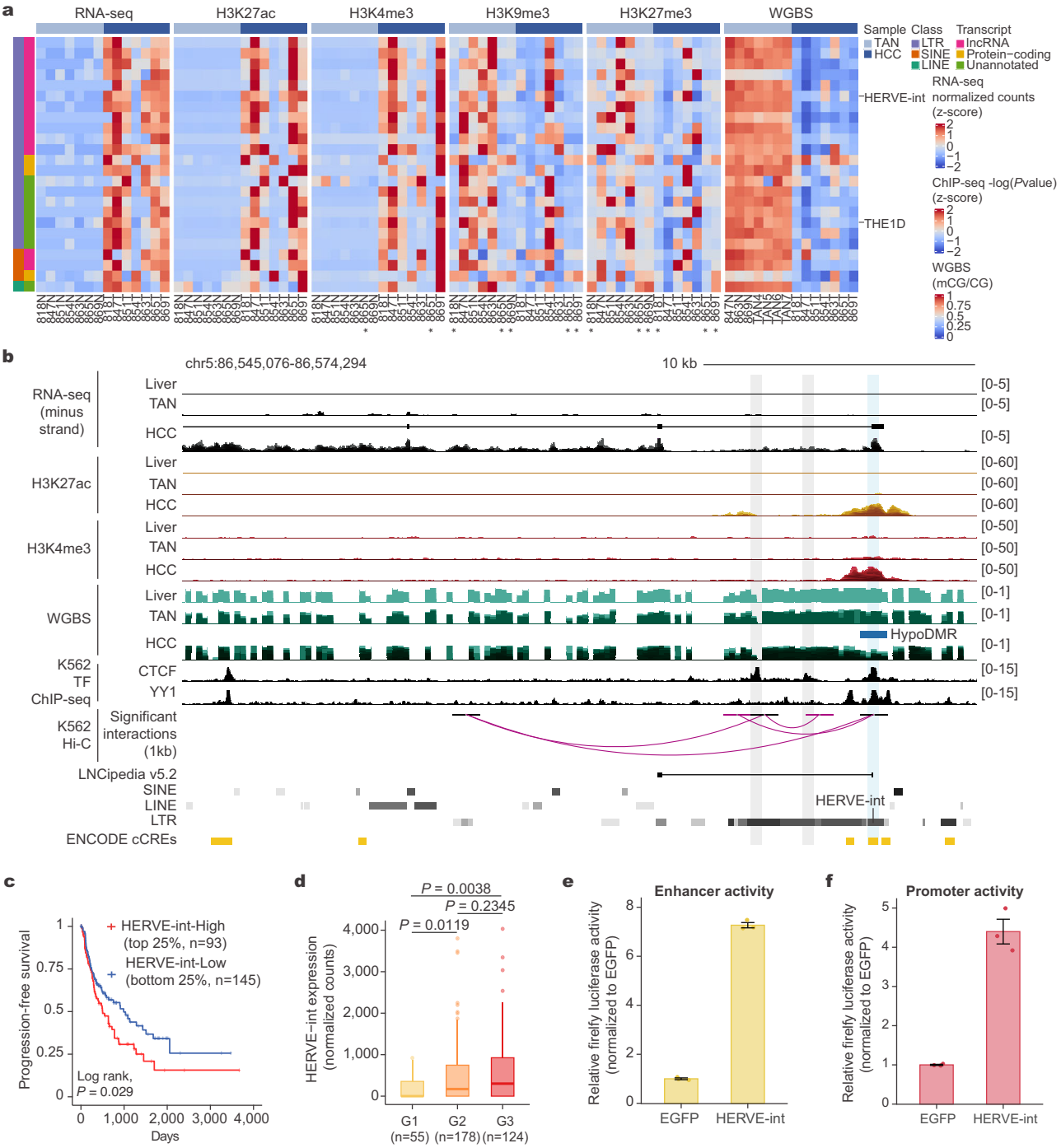


Fig. 5 | Epigenetic dysregulation of retrotransposon-driven transcripts in HCC. **a** Heatmaps showing the expression and epigenetic signatures of 24 activated RT-promoters that gave rise to highly upregulated transcripts in HCC. Each row represents a candidate and each column represents a sample. Class and transcript annotations of the 24 candidates are shown to the left of the RNA-seq heatmap. HERVE-int and THE1D are also labeled to the right of the WGBS heatmap. **b** Genome browser screenshot showing the expression and epigenetic signatures of HERVE-int. StringTie-assembled transcripts are shown above the HCC RNA-seq track. HERVE-int and the neighboring HERVE-int elements in contact are highlighted by cyan and gray shadings, respectively. HypoDMR in HCC is annotated with a blue bar above the HCC WGBS track. Stranded total RNA-seq tracks are displayed as TPM, ChIP-seq tracks are displayed as fold enrichment over input, and WGBS tracks are displayed as mCG/CG. TF ChIP-seq and Hi-C for K562 are from ENCODE and 4DN, respectively. Liver RNA-seq and WGBS tracks are from Roadmap Epigenomics Project and shown as single replicates. Other tracks are shown as composite signals by overlaying tracks across samples. **c** Kaplan-Meier survival plot for HERVE-int expression in the TCGA-LIHC cohort. Based on HERVE-int expression, patients were stratified into the low (bottom 25%) and high (top 25%) groups. *P* value (log-rank test) is shown. **d** Boxplot comparing the expression of HERVE-int across tumor grades in the TCGA-LIHC cohort. *P* values (two-tailed Wilcoxon test) are shown. Luciferase assays assessing the enhancer (**e**) and promoter (**f**) activities of HERVE-int. Data represent the mean ± standard deviation of three technical replicates from one representative experiment (n = 2 independent experiments).

because they have distinct expression distribution in HCC patients (Supplementary Data 8).

The HERVE-int element overlapped with an annotated lncRNA promoter (Fig. 5b), whereas the THE1D element functioned as the alternative promoter for the pseudogene *ENSG00000286540* (Supplementary Fig. 9a, b). Luciferase assays validated that the HERVE-int element functioned as both an enhancer and a promoter, whereas the THE1D element demonstrated predominantly promoter activity (Fig. 5e, f and Supplementary Fig. 9d, e). Focal DNA hypomethylation at HERVE-int and THE1D coincided with the gain of H3K27ac and H3K4me3 (Fig. 5b, Supplementary Fig. 9a, b and Supplementary Fig. 10a, b). To test the mechanistic relationship, we used the CRISPR-SunTag-TET1CD system to achieve targeted DNA demethylation of HERVE-int and THE1D in HCC cell lines that do not express these transcripts (Supplementary Fig. 10c, d). Indeed, targeted DNA demethylation led to significant induction of HERVE-int and THE1D (Supplementary Fig. 10e,f), demonstrating that loss of DNA methylation is sufficient for their derepression in HCC.

In addition to HCC cell lines, HERVE-int is expressed in the K562 cell line. Integrating TF ChIP-seq and Hi-C data from K562, we found that HERVE-int was bound by CTCF and YY1 and had significant chromatin interactions with two other neighboring HERVE-int elements (Fig. 5b). Given that CTCF and YY1 are involved in mediating enhancer-promoter interactions, we hypothesized that loss of DNA methylation at HERVE-int facilitates CTCF and YY1 binding, which leads to aberrant *cis*-regulatory interactions[81–83]. Interestingly, a recent study demonstrated that knockdown of *SETDB1*, the H3K9 methyltransferase, in HCC cells led to the transcriptional derepression of two HERVE-int elements through loss of H3K9me3[63]. These results suggest that depending on the sequence and genomic locations, individual HERVE-int elements are regulated by distinct epigenetic mechanisms.

To determine the functional roles of HERVE-int and THE1D in HCC, we performed CRISPRi experiments. We achieved approximately 80–90% knockdown of both HERVE-int and THE1D expression (Fig. 6a and Supplementary Fig. 10g). RNA-seq revealed that silencing of either element did not broadly affect the transcriptome (Fig. 6b, Supplementary Fig. 10h, and Supplementary Data 9); however, we observed significant dysregulation of genes with established roles in cancers. For instance, in HERVE-int-depleted cells, *ANXA1*, which has anti-inflammatory activities in various cancers, was upregulated[84,85]. Conversely, *DUSP9*, associated with less differentiated HCC tumors, was downregulated (Fig. 6b)[86]. Similarly, in THE1D-depleted cells, *RBP1*, a gene frequently downregulated in HCC, showed increased expression (Supplementary Fig. 10h)[87].

HERVE-int expression associates with poor prognosis and immunotherapy-related molecular features in HCC

We further assessed whether increased expression of HERVE-int or THE1D in HCC correlated with molecular characteristics associated with poorer disease outcomes. A previous TCGA study established the iClust1-3 molecular subclass classifications, each of which is associated with distinct genetic and epigenetic alterations[50]. We found that both HERVE-int-High and THE1D-High patients were enriched in the iClust3 subclass (Supplementary Fig. 10i). In addition to the more pronounced global loss of DNA methylation, patients of the iClust3 subclass show molecular signatures associated with more aggressive tumors, including increased chromosomal instability and expression of *AFP*[50]. Notably, we found that HERVE-int-High, but not THE1D-Intermediate or -High patients, had higher aneuploidy score and *AFP* expression. (Fig. 6c, d). Therefore, we focused our subsequent analyses on HERVE-int.

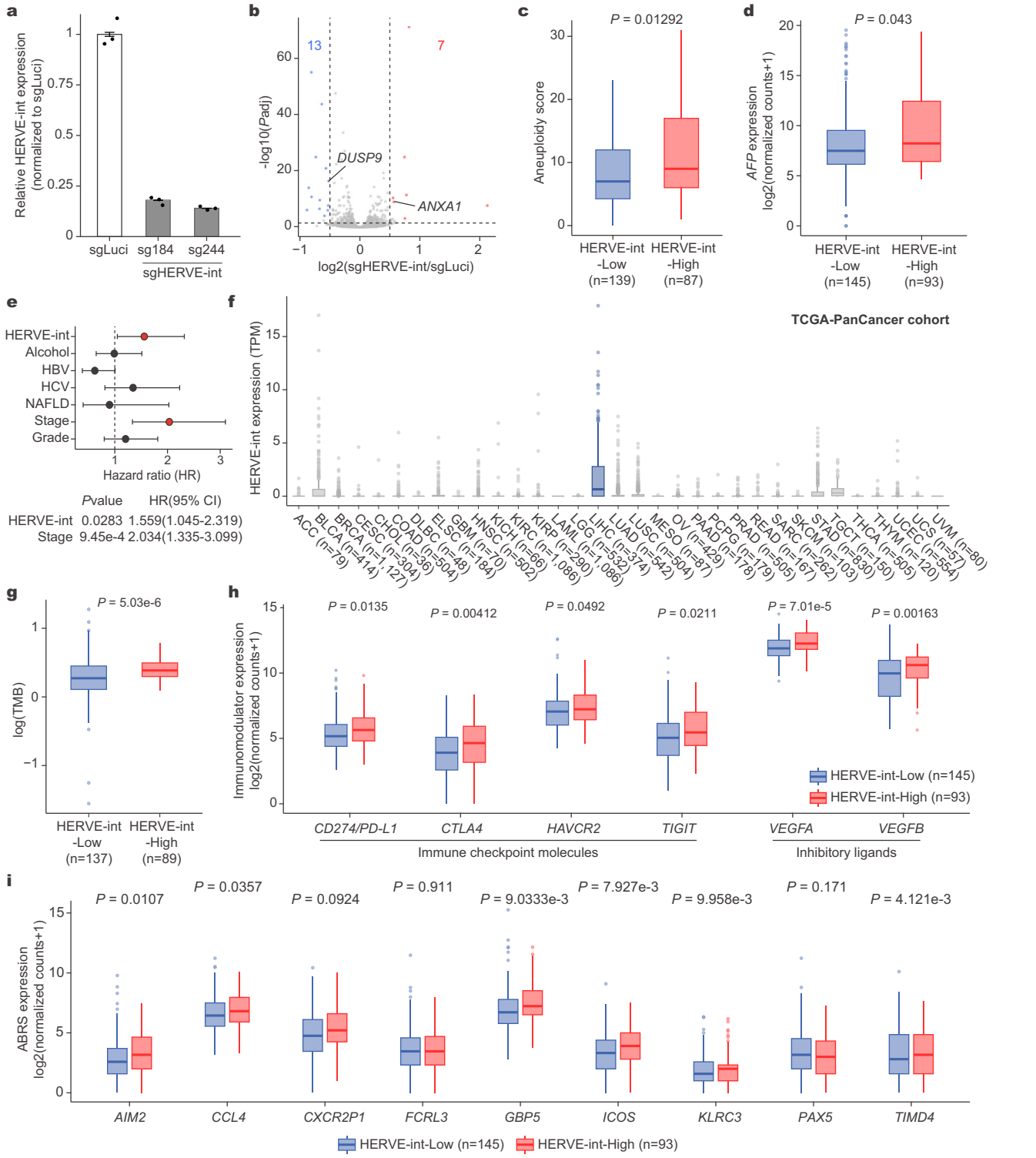
To further examine the interplay between HERVE-int expression and other risk factors on patient survival, we constructed a multivariate Cox regression model by binary categorizing HERVE-int-High and HERVE-int-Low patients based on different clinicopathological conditions, demonstrating that HERVE-int could serve as an independent prognostic factor for HCC (Fig. 6e). Moreover, we discovered that HERVE-int exhibited a highly restricted expression pattern in HCC but not in other cancer types (Fig. 6f).

In parallel with increased aneuploidy, HERVE-int-High patients also had higher tumor mutational burden (TMB) (Fig. 6g). Given the positive association between TMB and patient response to immunotherapy, we asked whether there is also differential expressions of immunomodulators between HERVE-int-Low and HERVE-int-High patients[88]. Indeed, we found that HERVE-int-High patients had higher expressions of key immune checkpoint molecules, including *CD274/PD-L1* and *CTLA-4*. Moreover, these patients showed higher expressions of *VEGF* ligands, which are important angiogenesis mediators that can suppress anti-tumor immune responses (Fig. 6h)[89–91]. In recent years, atezolizumab (anti-PD-L1) and bevacizumab (anti-VEGF) combination therapy have become the new standard of care for patients with unresectable HCC. The atezolizumab +bevacizumab response signature (ABRS) was developed, which comprises of 9 genes that are more highly expressed in patients who respond favorably to the combined therapy[92]. Interestingly, HERVE-int-High patients also showed upregulation of 6 out of the 9 genes (Fig. 6i). While these data suggest that HERVE-int-High patients may be more likely to benefit from ICIs, VEGF inhibitors, and/or their combined therapeutic regimens, more preclinical studies are needed. Collectively, our findings highlight HERVE-int as a potential prognostic biomarker and predictor of patient response to immunotherapy for HCC.

Discussion

Hepatocarcinogenesis is a multi-stage process that involves the accumulation of both genetic and epigenetic aberrations. While the genetic landscape of HCC is well-characterized, the role of epigenetic dysregulation in driving the aberrant activity of CREs and retrotransposons in HCC remains incompletely understood[50–52,93]. In this study, we performed a comprehensive analysis of the transcriptional regulatory landscape of HCC by integrating the genome-wide profiles of DNA methylation, histone modifications, and transcriptomes of HCC patient samples and cell lines. We demonstrated the widespread epigenetic changes at CREs and retrotransposons, which are coupled with transcriptional reprogramming in HCC. Importantly, we uncovered a dual regulatory mechanism for *GPC3*, which is an established HCC-related gene, and demonstrated the clinical and immunotherapeutic implications of specific retrotransposons, including HERVE-int. These findings underscore the value of integrating epigenomic data to discover new therapeutic targets and biomarkers for HCC[93].

A major unmet challenge in HCC immunotherapy is to overcome its highly immunosuppressive tumor microenvironment[9]. Despite being an inflammation-associated cancer, more than 60% of HCC cases exhibit non-inflamed phenotypes, characterized by low expression of immune checkpoint molecules, reduced immune cell infiltration, and blunted inflammatory activities[94]. These molecular features of transcriptional numbness could contribute to poor response and resistance to immunotherapy[95]. Combining epigenetic drugs with immunotherapies holds promise for rejuvenating exhausted immune cells within the HCC tumor microenvironment[96]. The results of our integrative analyses support this combined treatment strategy. Specifically, we elucidated a dual regulatory mechanism for *GPC3*, which is an immunotherapeutic target highly specific to HCC[38–40]. First, we showed the aberrant reactivation of a fetal liver SE drives the transcriptional dysregulation of *GPC3* in HCC. Second, we found that the DNA methylation levels of two CGIs associated with *GPC3*, namely the promoter CGI and intragenic CGI, also modulated the expression of *GPC3*. Both events are required for *GPC3* overexpression in HCC. These results have implications for improving patient response to as well as patient selection for GPC3-targeted immunotherapies. Indeed, while GPC3-derived peptide vaccine, GPC3-targeted antibodies, and GPC3-targeted CAR-T cell therapy are well-tolerated and have demonstrated efficacy in phase I and phase II clinical trials, patients with advanced stage HCC are more refractory to these treatments[62,64,65,97,98]. These patients could potentially benefit from the addition of a DNA demethylating agent such as 5-aza-2'-deoxycytidine, which may upregulate *GPC3*, and in turn, enhance its presentation to immune cells. On the other hand, GPC3-negative HCC patients,



although typically excluded from GPC3-targeted immunotherapies, might also benefit from such combined therapy. Despite the heterogeneity of our patient cohort, it appears that some GPC3-negative HCC patients harbor DNA hypermethylation of the CGIs in the presence of the active SE. Thus, these patients might be sensitized to GPC3-targeted immunotherapies upon DNA demethylation at the CGIs. Moreover, DAC treatment could induce expression of retrotransposons normally silenced by DNA methylation, which could also augment tumor immunogenicity[10–14,99]. However, further studies are needed to evaluate the safety and efficacy of this combined strategy.

Currently, there are no validated predictive biomarkers for patient response to immunotherapy in HCC[100]. While high TMB and immune checkpoint molecule expression, including PD-L1, have been used as predictive biomarkers across different cancer types, they have limited clinical utility when considered in isolation[88,101]. Hence, new predictive biomarkers are needed to identify HCC patients who are most likely to benefit from immunotherapy. Mounting evidence suggests that retrotransposon dysregulation in cancer is associated with active immune responses[14,102]. Intriguingly, increased expression of ERVs, including a full-length HERV4700 element, was significantly associated with improved response to PD-1/PD-

Fig. 6 | HERVE-int dysregulation is associated with specific molecular signatures and more favorable immunotherapy response in HCC. a RT-qPCR of the relative expression of HERVE-int in PLC cells upon CRISPRi. Each dot represents a technical replicate and error bars denote standard deviations. sgRNA targeting luciferase (sgLuci) is used as the negative control. *GAPDH* serves as the internal control for the calculation of relative *GPC3* expression and the values are normalized to that of sgLuci. **b** Volcano plot showing dysregulated genes in HERVE-int-depleted PLC cells. For each gene, the -log10-transformed, two-tailed adjusted *P* values (*P*adj) is plotted against the $\log_2$-transformed fold change. Upregulated (*P*adj < 0.05 and $\log_2$FC > 0.5) and downregulated (*P*adj < 0.05 and $\log_2$FC < -0.5) are labeled in red and blue, respectively. Dashed lines represent the indicated thresholds. Selected dysregulated genes are labeled. Boxplots comparing the aneuploidy score (**c**) and *AFP* expression (**d**) between HERVE-int-Low and HERVE-int-High patients. **c** Only patients with matched RNA-seq and aneuploidy score data were analyzed. **e** Multivariate cox regression analysis for HERVE-int expression and other clinicopathological conditions in the TCGA-LIHC cohort. *P* values (log-rank test) and 95% confidence intervals for hazard ratios of the two independent prognostic indicators of the model are shown. **f** Boxplots comparing HERVE-int expression across all TCGA cancer types. HCC is highlighted in blue while other cancer types are shown in gray. Cancer type abbreviations follow TCGA definitions. **g** Boxplots comparing the TMB between HERVE-int-Low and HERVE-int-High patients. Only patients with matched RNA-seq and mutation data were analyzed. **h** Boxplots comparing the expressions of selected immune checkpoint molecules and inhibitory ligands between HERVE-int-Low and HERVE-int-High patients. **i** Boxplots comparing the expressions of genes comprising the ABRS between HERVE-int-Low and HERVE-int-High patients. For all boxplots in this figure, *P* values (two-tailed Wilcoxon test) are shown. The center and bounds of boxes indicate the median and quartile of all data points, respectively. The minima and maxima of whiskers indicate quartile 1–1.5× the interquartile range and quartile 3+1.5× the interquartile range.

L1 blockade in clear cell renal cell carcinoma[32,33]. Expanding on these findings, we discovered that patients with high expression of a HERVE-int element exhibited molecular signatures of more favorable responses to immunotherapy, including higher TMB and expression of key immune checkpoint molecules and VEGFA/B. Importantly, HERVE-int is more highly expressed in HCC compared to other cancer types, further strengthening its potential as a predictive biomarker for immunotherapy response in HCC. Recently, the combination of atezolizumab (anti-PD-L1) and bevacizumab (anti-VEGF) has become the new standard of care for unresectable HCC[6,7]. Integrating specific retrotransposons, such as HERVE-int, into the existing predictive biomarker panels might improve identification of patients most likely to benefit from combined therapy. Future studies should explore the feasibility of detecting HERVE-int transcripts in circulating cell-free nucleic acids for minimally invasive monitoring. Taken together, these findings offer promising new avenues for therapeutic development and personalized HCC management.

Methods

Patient consent and specimen processing

Seven HBV-associated HCC patients who underwent hepatectomy at the Prince of Wales Hospital (Hong Kong) were enrolled in this study. All ethical regulations relevant to human research participants were followed. Informed consent was obtained from all patients and the study protocol was approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee. Patient clinical metadata is summarized in Supplementary Data 1. Patient specimens were processed immediately after surgery and snap-frozen in liquid nitrogen for storage. Frozen liver tissues were pulverized with mortar and pestle in liquid nitrogen for further analyses.

Cell culture

HepG2 was obtained from ATCC (ATCC HB-8065). Hep3B and PLC were gifts from Prof. Nathalie Wong (Department of Surgery, The Chinese University of Hong Kong). Cell cultures were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum (Gibco), 1% L-glutamine (Gibco), 1% non-essential amino acids (Gibco), 1% sodium pyruvate (Gibco), and 1% penicillin-streptomycin (Gibco) at 37°C in 5% CO_2. Medium was changed every 48 h and cells were passed upon reaching 80% confluency. Cell lines were routinely tested and found to be free of mycoplasma contamination by the MycoStrip™ Test Kit (InvivoGen).

Whole genome bisulfite sequencing

To mitigate amplification bias, four WGBS libraries were prepared per sample. WGBS libraries for the seven HCC tissues were prepared by the Center for PanorOmic Sciences, The University of Hong Kong. Briefly, DNA was sonicated to 200–400 bp fragments. 100 ng of the fragmented DNA was spiked with 1% lambda DNA (Promega) and subjected to bisulfite conversion using EZ DNA Methylation Lighting Kit (Zymo Research Corp., Orange, CA, USA). Libraries were prepared using the EpiGenome™ Methyl-Seq Kit following manufacturer's instructions. Each library was sequenced on a different lane on the Illumina HiSeq1500 platform. WGBS libraries for the three TAN tissues were prepared in-house. Briefly, genomic DNA was extracted using DNeasy® Blood & Tissue Kit (Qiagen) following manufacturer's protocol. 1 μg of genomic DNA spiked with 0.5% lambda DNA (Promega) was sonicated to 100–1500 bp using Covaris® S200 focused-ultrasonicator (Covaris). Adaptor-ligated DNA fragments (200–400 bp) were bisulfite converted using EpiJET Bisulfite Conversion Kit (Thermo Fisher Scientific) and amplified using KAPA HiFi HotStart Uracil +ReadyMix (KAPA) for 7 cycles. Libraries were sequenced on Illumina HiSeq 4000. WGBS quality metrics are listed in Supplementary Data 2.

ChIP-seq

MicroChIP-seq for patient samples were performed as previously described[103]. Briefly, pulverized liver tissues were crosslinked with 1% formaldehyde (Sigma-Aldrich) in 1X PBS for 8 min at room temperature and then quenched with glycine at a final concentration of 125 mM for 5 min at room temperature. After washing with ice-cold 1X PBS twice, pulverized and crosslinked liver tissues were sonicated using Covaris® S200 focused-ultrasonicator (Covaris). Fragmented chromatin was extracted by RIPA dilution buffer. Protein A Dynabeads (Thermo Fisher Scientific) were bound with antibodies of interest (H3K27ac, Active Motif 39133; H3K4me3, Active Motif 39915; H3K9me3, Abcam ab176916 or H3K27me3, Active Motif 39155) and incubated at 4°C for 36 h. Captured chromatin was washed four times with RIPA buffer and eluted at 37°C for 1 h in ChIP elution buffer. Eluted chromatin was reverse crosslinked by incubating with Proteinase K (NEB) at 68°C for 4 h. DNA was purified with QIAquick® PCR Purification Kit (Qiagen). Libraries were prepared using KAPA HyperPrep DNA Kit (Roche) following manufacturer's instruction and sequenced on Illumina NextSeq 500.

Native ChIP for HCC cells was performed as previously described with minor modifications[104]. Briefly, 1×10^7 cells were resuspended in ice-cold douncing buffer and lysed by passing through 25GX 5/8" needle (Terumo) for 50 times. Chromatin was digested with 550 U/mL MNase (Thermo Fisher Scientific) for 15 min at 37°C without shaking. EDTA was added to quench the digestion. Nuclear membrane was disrupted by incubating for 1 h in hypotonic lysis buffer. ChIP was performed by incubating digested chromatin with antibodies that were prebound to Dynabeads™ M-280 Sheep Anti-Mouse IgG (Thermo Fisher Scientific) or Dynabeads™ M-280 Sheep Anti-Rabbit IgG (Thermo Fisher Scientific) in IP buffer overnight at 4°C. Captured chromatin was washed twice with wash buffer, once with final wash buffer, and eluted at 68°C for 2 h in elution buffer together with RNase A (Sigma-Aldrich). Eluted DNA was purified by QIAquick® PCR purification Kit (Qiagen) following manufacturer's instruction and sequenced on Illumina NextSeq 500. ChIP-seq quality metrics are listed in Supplementary Data 2.

RNA-seq and RT-qPCR

For stranded total RNA-seq, total RNA was extracted using RNeasy® Mini Kit (Qiagen). 200 ng (patient samples) or 1 μg (HCC cell lines) of total RNA was subjected to rRNA depletion using Ribo-off rRNA depletion Kit (H/M/R) (Vazyme), followed by library preparation using QIAseq™ Stranded Library Kit (Qiagen) according to manufacturer's protocol. Libraries were sequenced on Illumina HiSeq 4000.

For RT-qPCR, total RNA was extracted using RNeasy® Mini Kit (Qiagen). 1 μg of total RNA was treated with DNaseI (NEB) and purified with Agencourt® RNAClean™ XP (Beckman Coulter). First-strand synthesis was performed using SuperScript® III Reverse Transcriptase (Invitrogen) following manufacturer's protocol. cDNA templates were subjected to 40 cycles of PCR using LightCycler 480 Instrument II (Roche). RT-qPCR primers used are listed in Supplementary Data 3.

Cloning

For luciferase assays, candidates were cloned downstream of the luciferase reporter in pGL3-promoter vector (Promega) to test for enhancer activity, or upstream of the luciferase reporter gene in pGL3-enhancer vector (Promega) to test for promoter activity, respectively. A random EGFP sequence that has no *cis*-regulatory activity was also cloned accordingly and used as negative controls.

For all CRISPR-related experiments, sgRNAs were cloned into pLKO5.sgRNA.EFS.puroR (derived from pLKO5.sgRNA.EFS.GFP by replacing the GFP with puromycin resistance gene; Addgene plasmid # 57822). For shRNA knockdown experiments, shRNAs were cloned into pSIH-H1-puroR (derived from pSIH-H1-copGFP (Systems BioSciences) by replacing copGFP with puromycin resistance gene). For CRISPR-SunTag experiments, pHRdSV40-dCas9-22aaLinker_GCN4_v4-P2A-blastR was constructed by replacing BFP with blasticidin resistance gene.

Luciferase assays

Luciferase assays for *GPC3* constituent enhancers, HERVE-int and THE1D were performed in HepG2 cells. Briefly, the candidates were cloned downstream of the luciferase reporter gene in pGL3-promoter vector (Promega) to test for enhancer activity, or upstream of the luciferase reporter gene in pGL3-enhancer vector (Promega) to test for promoter activity, respectively. A random EGFP sequence that has no *cis*-regulatory activity was also cloned accordingly and used as negative controls. 0.8×10^5 cells were seeded per well in 24-well plate 18 h prior experiment. The cells were co-transfected with pGL3 vector containing the candidate or random EGFP sequence and pGL4.75[hRluc/CMV] vector (Promega) in 1000:1 pmol ratio using Lipofectamine™ 3000 Transfection Reagent (Thermo Fisher Scientific). Luciferase signals were measured 48 h after transfection using Dual-Luciferase® Reporter (DLR™) Assay System as per manufacturer's protocol on a FlexStation3® Multi-Mode Microplate Reader. Enhancer or promoter activity of each candidate was calculated using the firefly-to-renilla signal ratio and normalized to that of the random EGFP sequence. Primers used for cloning are listed in Supplementary Data 3.

shRNA knockdown and CRISPR experiments

shRNA knockdown experiments were performed in HepG2 cells. CRISPRi for HERVE-int was performed in PLC-dCas9 cells. CRISPRi for *GPC3* constituent enhancers and THE1D were performed in HepG2-dCas9 cells. CRISPR-SunTag-DNMT3A1 for *GPC3* intragenic CGI and promoter CGI were performed on HepG2-dCas9-SunTag cells. CRISPR-SunTag-TET1CD for HERVE-int and THE1D were performed on HepG2-dCas9-SunTag cells and Hep3B-dCas9-SunTag cells, respectively. shRNAs were designed using siRNA Wizard Software v3.1 (https://www.invivogen.com/sirnawizard/; InvivoGen) and cloned into pSIH-H1-puroR. sgRNAs were designed using CRISPOR v5.0.1 and cloned into pLKO5.sgRNA.EFS.puroR[105].

Stable cells were generated as previously described with minor modifications[106]. Briefly, HEK293 cells were seeded at 1.2×10^6 cells per well of 6-well plate and co-transfected in equal molar concentrations of lentiviral constructs, pMD2.G (Addgene plasmid # 12259) and pCMVR8.74 (Addgene plasmid # 22036) using Lipofectamine™ 3000 Transfection Reagent (Thermo Fisher Scientific). Cells were incubated at 37°C in 5% CO_2 overnight and media was changed after 18 h. Viral supernatant was harvested at day 2 and 3 post-transfection. Target cells were transduced with the supernatant with 8 μg/mL filter-sterilized polybrene. Transduced cells were selected with 10 μg/mL blasticidin (InvivoGen) for 7 days or 1 μg/mL puromycin (InvivoGen) for 5 days before downstream experiments.

Bioinformatic analyses

WGBS data analyses

Mapping and visualization. Paired-end sequencing reads were aligned to GRCh38/hg38 using Bismark v0.23.0 with default parameters[107]. PCR duplicates were removed by Picard tools v2.23.0 (http://broadinstitute.github.io/picard/). Only CpGs with at least 5 reads were retained for all downstream analyses. Signal tracks with mCG/CG per 200 bp bin were generated for visualization in UCSC genome browser. Circos plot comparing the DNA methylation landscape of HCC and TAN was generated using pyCirclize v1.3.0 (https://github.com/moshi4/pyCirclize) in python 3.12.

Identification of DMRs. DMRs were called with methylKit v1.20.0 per 500 bp non-overlapping window with at least 5 CpGs covered[108]. Significance of %mCG difference in each window was determined by Fisher's exact test with *Q*value adjusted by SLIM method[109]. HyperDMRs in HCC were defined as regions with at least two CpGs showing 30% gain in mCG. Similarly, hypoDMRs in HCC were defined as regions with at least two CpGs showing 30% loss in mCG. HyperDMRs and hypoDMRs in HCC are listed in Supplementary Data 4.

Identification of PMDs. PMDs were defined as previously described[110,111]. Briefly, PMDs were called with methpipe v5.0.1 using default parameters[112]. Only PMDs with sizes no less than 50 kb and do not overlap with gaps in GRCh38/hg38 or the HOX clusters were retained for analyses. For quantification of protein-coding genes or repeats falling within PMDs, only those with at least 80% of length overlapping with PMDs were considered. PMDs in HCC are listed in Supplementary Data 4.

RNA-seq data analyses

Mapping and visualization. Paired-end sequencing reads were aligned to GRCh38/hg38 and GENCODE v39 with STAR v2.7.6 using ENCODE parameters except where –outFilterMultimapNmax was set to 1 to keep only uniquely mapped reads[113]. TPM of genes were quantified with RSEM v1.3.3[114]. TPM signal tracks were generated for visualization in UCSC genome browser, using the bamCoverage program in deepTools v3.5.1 with parameters –normalizeUsing BPM –exactScaling and –filterRNAstrand forward/reverse for minus/plus strand, respectively[115].

Identification of dysregulated genes and retrotransposons. Only transcriptionally active genes (TPM > 1 in ≥ 2 samples) and retrotransposons (RPKM > 1 in ≥3 samples) were used as inputs for differential analyses. Dysregulated genes and retrotransposons were defined using DESeq2 v1.36.0[116]. For genes, the thresholds of $\log_2FC > 1$ and Padj < 0.01 were used for upregulated genes (n = 486), and the thresholds of $\log_2FC < -1$ and Padj < 0.01 were used for downregulated genes (n = 723), respectively. For retrotransposons, the thresholds of $\log_2FC > 3.5$ and Padj < 0.01 were used for upregulated retrotransposons (n = 3005), and the thresholds of $\log_2FC < -3.5$ and Padj < 0.01 were used for downregulated retrotransposons (n = 1152), respectively. Upregulated and downregulated genes are listed in Supplementary Data 5. Upregulated and downregulated retrotransposons are listed in Supplementary Data 7.

Gene ontology and Gene Set Enrichment Analysis. GO analyses for dysregulated genes and genes associated with dysregulated SEs were

performed using Metascape v3.5[117]. GO analyses for dysregulated typical enhancers and dysregulated RT-enhancers were performed using GREAT v4.0.4[118]. GSEA for dysregulated genes was performed using clusterProfiler v4.6.2 with the Hallmark gene sets from MSigDB (Broad Institute)[119].

Retrotransposon-derived transcript assembly. Retrotransposon-derived transcripts were assembled using StringTie2 v2.1.2 with default parameters[120].

ChIP-seq analyses

Avocado imputation. Avocado v0.4.0 was used for model training and signal imputation[43]. keras v2.13.1 library was used in conjunction with CUDA v11.8 and cuDNN v8.8.0.121 to interface with TensorFlow v2.13.0 library for all model training. -log10(P values) signals from each chromosome were binned at 25 bp resolution and arcsinh transformed. Models were trained on a per-chromosome basis with parameters: –batch_size= 40000 –n_celltype_factors=32 –n_assay_factors=256 –n_25bp_factors=25 –n_250bp_factors=40 –n_5kbp_factors=45 –n_layers=2 –n_nodes=2048 and –n_epoch2 = 100. The ADAM optimizer and mean-square error (MSE) loss were employed for fitting the in-house data to the Avocado ENCODE-Core hg38 pre-trained model (https://zenodo.org/records/4774521) using the fit_celltypes and fit functions. The epoch size was adjusted for each chromosome by dividing the number of genomic positions (chromosome length divided by a bin size of 25) by a batch size of 40,000 and rounding to the nearest integer. In the evaluation phase, 10 out of 43 samples were left-out from the training set to assess the per-sample global MSE of the imputed signal (mean of (imputed bin valuesground truth bin values)^2). This evaluation was compared against two baseline measures: (1) the per-sample baseline MSE (mean across all bins (ground truth value of one bin of that sample - mean of all ground truth bin values of that sample)^2), and (2) baseline MSE averaged across all other samples (mean across all bins (ground truth value of one bin of that sample - mean of all ground truth bin values across all other samples)^2).

Mapping and visualization. Single-end sequencing reads were aligned to GRCh38/hg38 using ENCODE ChIP-seq pipeline v2.2.1 (https://github.com/ENCODE-DCC/chip-seq-pipeline2). Signal tracks with fold enrichment over input were generated using MACS2 v2.2.7.1 for visualization in UCSC genome browser[121].

Peak calling and quality controls. Narrow peaks (H3K27ac and H3K4me3) were called using MACS2 v2.2.7.1 with parameters –g 2.7e9 –SPMR –keep-dup all and –q 0.01. High-confidence peaks were retained by further filtering peaks with (1) -log$P$ > 5 and -log$Q$ > 2, (2) input-subtracted RPM > 1, and (3) at least 2-fold enrichment of RPM in ChIP compared to input libraries (RPM_{ChIP}≥ 2xRPMinput) as per ENCODE ChIP-seq pipeline (https://github.com/ENCODE-DCC/chip-seq-pipeline2). Broad peaks (H3K9me3 and H3K27me3) were called using epic2 v0.0.48[122]. For H3K9me3, the parameters –k –cs GRCh38.chrom.sizes –g 1 bin 5000 were used. For H3K27me3, the parameters –k –cs GRCh38.chrom.sizes –g 1 bin 2000 were used. For quality controls on ChIP-seq narrow peak datasets, NSC and RSC were calculated using phantompeakqualtools v1.2.2, and FRiP was calculated by dividing the number of reads falling into peak regions by the total number of mapped reads[123,124].

Identification of CREs and dysregulated CREs. 1 kb flanking regions of H3K27ac and H3K4me3 summit peaks were used to identify putative CREs. Only experimental data are used to define putative CREs. Briefly, active putative enhancers were defined as regions enriched with H3K27ac peaks not overlapping with H3K4me3 peaks or GENCODE v39 TSS regions (1 kb flanking of GENCODE v39 TSS). Active putative promoters were defined as regions overlapped with both H3K27ac and H3K4me3 peaks. For samples that do not have H3K4me3 data, we defined active putative promoters as GENCODE v39 TSS regions that overlap with H3K27ac, and active putative enhancers as H3K27ac peaks that do not overlap with active putative promoters. To generate a non-redundant union set of active putative enhancers or promoters in HCC or TAN, the active putative enhancers or promoters defined in each sample that are within 500 bp were first merged within sample, and once again across samples of the same condition. To generate a master list of active putative enhancers or promoters across all samples, the non-redundant union set of active putative enhancers or promoters in HCC and TAN that are within 500 bp were merged. H3K27ac signals of all active enhancers (n = 62,288) or active promoters (n = 29,467) were subsequently subjected to k-means clustering using ComplexHeatmap v2.14.0124.

Dysregulated typical enhancers in HCC were defined using DESeq2 v1.36.0 with the thresholds of $log_2FC > 2.25$ and Padj < 0.01 for activated typical enhancers (n = 3367), and $log_2FC < -1.5$ and Padj < 0.01 for inactivated typical enhancers (n = 2924), respectively116. Similarly, dysregulated active promoters in HCC were defined using DESeq2 v1.36.0 with the thresholds of $log_2FC > 2$ and Padj < 0.01 for activated promoters (n = 1172) and $log_2FC < -2$ and Padj < 0.01 for inactivated promoters (n = 727), respectively. Dysregulated typical enhancers and promoters in HCC are listed in Supplementary Data 6.

Identification of SEs and dysregulated SEs. SEs were identified using the ROSE algorithm with default parameters[20,21,58]. All SEs across samples (n = 3245) were generated using similar strategies as described above for differential analyses. Dysregulated SEs were defined using DESeq2 v1.36.0 with the thresholds of $log_2FC > 1.5$ and Padj < 0.01 for activated SEs (n = 234) and $log_2FC < -1.5$ and Padj < 0.01 for inactivated SEs (n = 86), respectively. Dysregulated SEs are listed in Supplementary Data 6.

Identification of dysregulated retrotransposon-derived CREs. Analyses were restricted to RepeatMasker-annotated retrotransposons (LTRs, LINEs, and SINEs) that are at least 100 bp long and mapped to chr1-22 and chrX (n = 3,195,734). Dysregulated RT-typical enhancers were defined as retrotransposons with at least 30% of its length overlapping with the dysregulated enhancers and at least 15% of the dysregulated enhancers also overlapping with the retrotransposons, respectively (activated RT-typical enhancers, n = 2333; inactivated RT-typical enhancers, n = 834). Dysregulated RT-promoters were similarly defined (activated RT-promoters, n = 1166; inactivated RT-promoters, n = 195). Dysregulated RT-typical enhancers and RT-promoters are listed in Supplementary Data 7

Over-represented subfamilies from activated RT-typical enhancers and activated RT-promoters were identified using the thresholds of observed/expected ratio≥4, Q value (FDR-corrected one-tailed hypergeometric test) <0.01 and had at least 30 copies in the genome. The observed/expected ratio was calculated using the formula (n/nall)/(N/Nall), where n is the number of a subfamily in activated RT-typical enhancers or activated RT-promoters, nall is the total number of activated RT-typical enhancers or activated RT-promoters, N is the number of a subfamily in the genome, and Nall is the total number of retrotransposons in the genome. Over-represented subfamilies from activated RT-typical enhancers and activated RT-promoters are listed in Supplementary Data 7.

Motif analyses. Motifs in activated and inactivated typical enhancers were analyzed with HOMER v4.1.1 findMotifsGenome.pl program using all non-redundant enhancers in HCC and TAN tissues (n = 46,500) as background[125]. Similar analyses were performed for activated RT-typical enhancers and activated RT-promoters using all RT-typical enhancers (n = 24,596) and RT-promoters (n = 10,261), respectively.

Hi-C data analyses

HepG2 Hi-C data were downloaded from 4DN[126]. Significant chromatin interactions were called by FitHiC2 v2.0.8 at 1 kb resolution with the threshold $Q < 0.05$[127].

TCGA data analyses

Data download. TCGA Pan-Cancer cohort gene-level raw coverage count matrices, bigWig files, and clinical metadata were downloaded using Recount3 v1.6.0[128]. TCGA-LIHC HM450 array data were downloaded from UCSC Xena[129]. HM450 probe annotation was downloaded from https://zwdzwd.github.io/InfiniumAnnotation. TCGA-LIHC *GPC3* copy-number and mutation data were downloaded from cBioPortal[130–132]. TCGA-LIHC survival data were downloaded from TCGA PanCanAtlas Publications page at https://gdc.cancer.gov/about-data/publications/pancanatlas.

WGCNA of TCGA-LIHC cohort. WGCNA v1.71 was used to construct gene expression network in the TCGA-LIHC cohort[133]. Analyses were restricted to genes that have at least 10 read counts in 90% of the TAN or HCC samples (n = 26,108 across 408 samples). Briefly, the network was constructed using the blockwiseModules function using the parameters -power = 14, mergeCutHeight = 0.1 and deepSplit = 4, generating 84 modules. Genes associated with 3 modules with high correlation to either TAN or HCC were visualized using Gephi v0.10.1[134].

DNA methylation level of *GPC3* CGIs and their correlations with *GPC3* expression within the TCGA-LIHC cohort. β-values for probes falling within the *GPC3* intragenic CGI (cg11442732, cg02442693, and cg09760728) and promoter CGI (cg05445331, cg06384623, cg07297906, cg27496708, cg16272791, and cg14867863) were extracted using the HM450 probe annotation and correlated with *GPC3* expression in matched HCC tissues.

HERVE-int expression in the TCGA-PanCancer cohort. Raw reads falling on HERVE-int was quantified from TCGA Pan-Cancer bigWig files using Megadepth v1.8.0 and subsequently converted to TPM for analyses[135].

Kapan-Meier survival analyses. Survival analyses for HERVE-int and THE1D were restricted to HCC patients with survival data. DESeq2-normalized counts of HERVE-int and THE1D were used for the analyses. For HERVE-int, patients were stratified into high-expression (top 25% percentile) and low-expression (bottom 25% percentile) groups. For THE1D, patients were stratified into high-expression (top 30% percentile), intermediate-expression (30%x70% percentile) and low-expression (bottom 30% percentile) groups. Kaplan-Meier curves were generated using survminer v0.4.9 (https://github.com/kassambara/survminer) and survival v3.5.7 (https://github.com/therneau/survival). Patient stratification results are presented in Supplementary Data 8.

Multivariate cox regression analyses. Patients of the HERVE-int-High and HERVE-int-Low groups were binary categorized in the following clinicopathological features: alcohol consumption (no vs yes), HBV/HCV infection (no vs yes), NAFLD (no vs yes), tumor stage (I/II vs III/IV), and tumor grade (G1/2 vs G3/G4). Patients without the corresponding clinical annotation were assigned NA. Multivariate cox regression analyses were performed using survminer v0.4.9 (https://github.com/kassambara/survminer) and survival v3.5.7 (https://github.com/therneau/survival). Patients with low HERVE-int expression, no history of alcohol consumption, no HBV/HCV infection, no NAFLD, early tumor stage (I/II) or low tumor grade (G1/2) were used as the reference level for the respective prognostic factor in the model.

TMB analyses. TMB for HERVE-int-High and HERVE-int-Low patients was calculated with maftools v2.14.0 using the parameters: captureSize=35.8 and logScale=TRUE[136].

Statistics and reproducibility

No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment. The experiments were not randomized. Luciferase assays were repeated twice with similar results obtained. Statistical analyses were performed using R v4.20. Statistical tests used are indicated in the figures or figure legends. *P* values and *Q*values are provided as exact values where possible or otherwise reported as a range.

Ethics approval and consent to participate

Patient samples were obtained from Prince of Wales Hospital (Hong Kong) with study protocol approved by the Joint Chinese University of Hong Kong-New Territories East Cluster Research Ethics Committee (CRE-2014.076).

Data availability

Epigenomic and transcriptomic datasets for the patient samples were deposited to EGA under the accession ID EGAD50000000058. H3K27ac and H3K4me3 ChIP-seq datasets for HepG2 and PLC were deposited to GEO under the accession code GSE276132. RNA-seq datasets for the CRISPRi experiments for HERVE-int and THE1D were deposited to GEO under the accession code GSE276133. All publicly available datasets used in this study are listed in Supplementary Data 10. All source data of this manuscript can be found in Supplementary Data 11 and 12.

Received: 3 February 2025; Accepted: 30 October 2025;
Published online: 13 December 2025

References

1. Rumgay, H. et al. Global burden of primary liver cancer in 2020 and predictions to 2040. *J. Hepatol.* **77**, 1598–1606 (2022).
2. Llovet, J. M. et al. Hepatocellular carcinoma. *Nat. Rev. Dis. Prim.* **7**, 6 (2021).
3. Yarchoan, M. et al. Recent developments and therapeutic strategies against hepatocellular carcinoma. *Cancer Res*. **79**, 4326–4330 (2019).
4. Kudo, M. et al. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. *Lancet* **391**, 1163–1173 (2018).
5. Llovet, J. M. et al. Sorafenib in advanced hepatocellular carcinoma. *N. Engl. J. Med.* **359**, 378–390 (2008).
6. Cheng, A. L. et al. Updated efficacy and safety data from IMbrave150: Atezolizumab plus bevacizumab vs. sorafenib for unresectable hepatocellular carcinoma. *J. Hepatol.* **76**, 862–873 (2022).
7. Finn, R. S. et al. Atezolizumab plus bevacizumab in unresectable hepatocellular carcinoma. *N. Engl. J. Med.* **382**, 1894–1905 (2020).
8. Ren, Z. et al. Sintilimab plus a bevacizumab biosimilar (IBI305) versus sorafenib in unresectable hepatocellular carcinoma (ORIENT-32): A randomised, open-label, phase 2-3 study. *Lancet Oncol.* **22**, 977–990 (2021).
9. Llovet, J. M. et al. Immunotherapies for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* **19**, 151–172 (2022).
10. Goyal, A. et al. DNMT and HDAC inhibition induces immunogenic neoantigens from human endogenous retroviral element-derived transcripts. *Nat. Commun.* **14**, 6731 (2023).
11. Brocks, D. et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet* **49**, 1052–1060 (2017).
12. Chiappinelli, K. B. et al. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **162**, 974–986 (2015).
13. Roulois, D. et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell* **162**, 961–973 (2015).
14. Kong, Y. et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.* **10**, 5228 (2019).

15. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet* **21**, 71–87 (2020).
16. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
17. Consortium, E. P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
18. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
19. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
20. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
21. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
22. Zhang, X. et al. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat. Genet* **48**, 176–182 (2016).
23. Mansour, M. R. et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
24. Jeon, A. J. et al. A genomic enhancer signature associates with hepatocellular carcinoma prognosis. *JHEP Rep.* **5**, 100715 (2023).
25. Wu, F. et al. Sirtuin 7 super-enhancer drives epigenomic reprogramming in hepatocarcinogenesis. *Cancer Lett.* **525**, 115–130 (2022).
26. Tsang, F. H. et al. Aberrant super-enhancer landscape in human hepatocellular carcinoma. *Hepatology* **69**, 2502–2517 (2019).
27. Du, A. Y., Chobirko, J. D., Zhuo, X., Feschotte, C. & Wang, T. Regulatory transposable elements in the encyclopedia of DNA elements. *Nat. Commun.* **15**, 7594 (2024).
28. Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V. & Wang, T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat. Commun.* **10**, 5640 (2019).
29. Sundaram, V. et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).
30. Shah, N. M. et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nat. Genet* **55**, 631–639 (2023).
31. Jang, H. S. et al. Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet* **51**, 611–617 (2019).
32. Smith, C. C. et al. Endogenous retroviral signatures predict immunotherapy response in clear cell renal cell carcinoma. *J. Clin. Invest* **128**, 4804–4820 (2018).
33. Panda, A. et al. Endogenous retrovirus expression is associated with response to immune checkpoint blockade in clear cell renal cell carcinoma. *JCI Insight* **3**. https://doi.org/10.1172/jci.insight.121522 (2018).
34. Hashimoto, K. et al. CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* **25**, 1812–1824 (2015).
35. Unfried, J. P. et al. Long noncoding RNA NIHCOLE promotes ligation efficiency of DNA double-strand breaks in hepatocellular carcinoma. *Cancer Res.* **81**, 4910–4925 (2021).
36. Wu, Y. et al. An LTR retrotransposon-derived long noncoding RNA lncMER52A promotes hepatocellular carcinoma progression by binding p120-catenin. *Cancer Res.* **80**, 976–987 (2020).
37. Deng, B. et al. An LTR retrotransposon-derived lncRNA interacts with RNF169 to promote homologous recombination. *EMBO Rep.* **20**, e47650 (2019).
38. Shi, D. et al. Chimeric antigen receptor-glypican-3 T-cell therapy for advanced hepatocellular carcinoma: Results of phase I trials. *Clin. Cancer Res.* **26**, 3979–3989 (2020).
39. Gao, H. et al. Development of T cells redirected to glypican-3 for the treatment of hepatocellular carcinoma. *Clin. Cancer Res*. **20**, 6418–6428 (2014).
40. Yamauchi, N. et al. The glypican 3 oncofetal protein is a promising diagnostic marker for hepatocellular carcinoma. *Mod. Pathol.* **18**, 1591–1598 (2005).
41. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*. **22**, 1813–1831 (2012).
42. Li, X., Liu, Y., Salz, T., Hansen, K. D. & Feinberg, A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res*. **26**, 1730–1741 (2016).
43. Schreiber, J., Durham, T., Bilmes, J. & Noble, W. S. Avocado: A multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol.* **21**, 81 (2020).
44. Salhab, A. et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. *Genome Biol.* **19**, 150 (2018).
45. Hon, G. C. et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res*. **22**, 246–258 (2012).
46. Berman, B. P. et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat. Genet* **44**, 40–46 (2011).
47. Ohm, J. E. et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet* **39**, 237–242 (2007).
48. Schlesinger, Y. et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet* **39**, 232–236 (2007).
49. Hama, N. et al. Epigenetic landscape influences the liver cancer genome architecture. *Nat. Commun.* **9**, 1643 (2018).
50. Cancer Genome Atlas Research Network. Electronic address, w. b. e. & Cancer Genome Atlas Research, N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341 e1323 (2017).
51. Song, M. A. et al. Elucidating the landscape of aberrant DNA methylation in hepatocellular carcinoma. *PLoS One* **8**, e55761 (2013).
52. Shen, J. et al. Genome-wide DNA methylation profiles in hepatocellular carcinoma. *Hepatology* **55**, 1799–1808 (2012).
53. Hoshida, Y. et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* **69**, 7385–7392 (2009).
54. Chiang, D. Y. et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res.* **68**, 6779–6788 (2008).
55. Liu, Y. et al. TET2 is recruited by CREB to promote Cebpb, Cebpa, and Pparg transcription by facilitating hydroxymethylation during adipocyte differentiation. *iScience* **26**, 108312 (2023).
56. Schmidt, L., Heyes, E. & Grebien, F. Gain-of-function effects of N-terminal CEBPA mutations in acute myeloid leukemia. *Bioessays* **42**, e1900178 (2020).
57. Varma, S. et al. The transcription factors Grainyhead-like 2 and NK2-homeobox 1 form a regulatory loop that coordinates lung epithelial cell morphogenesis and differentiation. *J. Biol. Chem.* **287**, 37282–37295 (2012).
58. Loven, J. et al. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
59. Yuan, J. H. et al. The MBNL3 splicing factor promotes hepatocellular carcinoma by increasing PXN expression through the alternative splicing of lncRNA-PXN-AS1. *Nat. Cell Biol.* **19**, 820–832 (2017).

60. Li, N., Gao, W., Zhang, Y. F. & Ho, M. Glypicans as cancer therapeutic targets. *Trends Cancer* **4**, 741–754 (2018).
61. Hsu, H. C., Cheng, W. & Lai, P. L. Cloning and expression of a developmentally regulated transcript MXR7 in hepatocellular carcinoma: Biological significance and temporospatial distribution. *Cancer Res.* **57**, 5179–5184 (1997).
62. Tsuchiya, N. et al. Immunological efficacy of glypican-3 peptide vaccine in patients with advanced hepatocellular carcinoma. *Oncoimmunology* **6**, e1346764 (2017).
63. Igarashi, Y. et al. Identification and clinical implications of endogenous retrovirus elements suppressed by SETDB1 in hepatocellular carcinoma. *JHEP Rep.* **7**, 101307 (2025).
64. Sawada, Y. et al. Phase II study of the GPC3-derived peptide vaccine as an adjuvant therapy for hepatocellular carcinoma patients. *Oncoimmunology* **5**, e1129483 (2016).
65. Zhang, Q. et al. (American Society of Clinical Oncology, 2024).
66. Zhao, J. & Faryabi, R. B. Spatial promoter-enhancer hubs in cancer: Organization, regulation, and function. *Trends Cancer* **9**, 1069–1084 (2023).
67. Uyehara, C. M. & Apostolou, E. 3D enhancer-promoter interactions and multi-connected hubs: Organizational principles and functional roles. *Cell Rep.* **42**, 112068 (2023).
68. Zhu, X. T., Yuan, J. H., Zhu, T. T., Li, Y. Y. & Cheng, X. Y. Long noncoding RNA glypican 3 (GPC3) antisense transcript 1 promotes hepatocellular carcinoma progression via epigenetically activating GPC3. *FEBS J.* **283**, 3739–3754 (2016).
69. Shen, M. et al. Hepatic ARID3A facilitates liver cancer malignancy by cooperating with CEP131 to regulate an embryonic stem cell-like gene signature. *Cell Death Dis.* **13**, 732 (2022).
70. Feng, M. et al. Sox13 promotes hepatocellular carcinoma metastasis by transcriptionally activating Twist1. *Lab Invest* **100**, 1400–1410 (2020).
71. Rhee, C. et al. ARID3A is required for mammalian placenta development. *Dev. Biol.* **422**, 83–91 (2017).
72. Frau, M. et al. Mybl2 expression is under genetic control and contributes to determine a hepatocellular carcinoma susceptible phenotype. *J. Hepatol.* **55**, 111–119 (2011).
73. Lu, G. D. et al. C/EBPalpha is up-regulated in a subset of hepatocellular carcinomas and plays a role in cell growth and proliferation. *Gastroenterology* **139**, 632–643 (2010).
74. Huber, R. et al. DNA methylation in transcriptional repression of two differentially expressed X-linked genes, GPC3 and SYBL1. *Proc. Natl. Acad. Sci. USA* **96**, 616–621 (1999).
75. Karttunen, K. et al. Transposable elements as tissue-specific enhancers in cancers of endodermal lineage. *Nat. Commun.* **14**, 5313 (2023).
76. Grillo, G. et al. Transposable Elements Are Co-opted as Oncogenic Regulatory Elements by Lineage-Specific Transcription Factors in Prostate Cancer. *Cancer Discov.* **13**, 2470–2487 (2023).
77. Deniz, O. et al. Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nat. Commun.* **11**, 3506 (2020).
78. Gao, L. et al. Single-cell analysis reveals transcriptomic and epigenomic impacts on the maternal-fetal interface following SARS-CoV-2 infection. *Nat. Cell Biol.* **25**, 1047–1060 (2023).
79. Frost, J. M. et al. Regulation of human trophoblast gene expression by endogenous retroviruses. *Nat. Struct. Mol. Biol.* **30**, 527–538 (2023).
80. Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9**, e1003470 (2013).
81. Zhang, X. et al. Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature* **575**, 385–389 (2019).
82. Weintraub, A. S. et al. YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588 e1528 (2017).
83. Monahan, K. et al. Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-alpha gene expression. *Proc. Natl. Acad. Sci. USA* **109**, 9125–9130 (2012).
84. Babbin, B. A. et al. Annexin I regulates SKCO-15 cell invasion by signaling through formyl peptide receptors. *J. Biol. Chem.* **281**, 19588–19599 (2006).
85. Paweletz, C. P. et al. Loss of annexin 1 correlates with early onset of tumorigenesis in esophageal and prostate carcinoma. *Cancer Res* **60**, 6293–6297 (2000).
86. Chen, K. et al. Dual-specificity phosphatase 9 regulates cellular proliferation and predicts recurrence after surgery in hepatocellular carcinoma. *Hepatol. Commun.* **5**, 1310–1328 (2021).
87. Schmitt-Graff, A. et al. Cellular retinol-binding protein-1 in hepatocellular carcinoma correlates with beta-catenin, Ki-67 index, and patient survival. *Hepatology* **38**, 470–480 (2003).
88. Yarchoan, M., Hopkins, A. & Jaffee, E. M. Tumor mutational burden and response rate to PD-1 inhibition. *N. Engl. J. Med.* **377**, 2500–2501 (2017).
89. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 e814 (2018).
90. Horikawa, N. et al. Expression of vascular endothelial growth factor in ovarian cancer inhibits tumor immunity through the accumulation of myeloid-derived suppressor cells. *Clin. Cancer Res.* **23**, 587–599 (2017).
91. Voron, T. et al. VEGF-A modulates expression of inhibitory checkpoints on CD8+ T cells in tumors. *J. Exp. Med* **212**, 139–148 (2015).
92. Zhu, A. X. et al. Molecular correlates of clinical response and resistance to atezolizumab in combination with bevacizumab in advanced hepatocellular carcinoma. *Nat. Med* **28**, 1599–1611 (2022).
93. Hlady, R. A. et al. Integrating the epigenome to identify drivers of hepatocellular carcinoma. *Hepatology* **69**, 639–652 (2019).
94. Montironi, C. et al. Inflamed and non-inflamed classes of HCC: A revised immunogenomic classification. *Gut* **72**, 129–140 (2023).
95. Loukas, I. et al. Selective advantage of epigenetically disrupted cancer cells via phenotypic inertia. *Cancer Cell* **41**, 70–87 e14 (2023).
96. Yang, W. et al. A selective HDAC8 inhibitor potentiates antitumor immunity and efficacy of immune checkpoint blockade in hepatocellular carcinoma. *Sci. Transl. Med.* **13**. https://doi.org/10.1126/scitranslmed.aaz6804 (2021).
97. Sawada, Y. et al. Phase I trial of a glypican-3-derived peptide vaccine for advanced hepatocellular carcinoma: Immunologic evidence and potential for improving overall survival. *Clin. Cancer Res.* **18**, 3686–3696 (2012).
98. Abou-Alfa, G. K. et al. Randomized phase II placebo controlled study of codrituzumab in previously treated patients with advanced hepatocellular carcinoma. *J. Hepatol.* **65**, 289–295 (2016).
99. Jiang, Q. et al. HIF regulates multiple translated endogenous retroviruses: Implications for cancer immunotherapy. *Cell* **188**, 1807–1827 e1834 (2025).
100. Greten, T. F. et al. Biomarkers for immunotherapy of hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* **20**, 780–798 (2023).
101. Davis, A. A. & Patel, V. G. The role of PD-L1 expression as a predictive biomarker: An analysis of all US Food and Drug Administration (FDA) approvals of immune checkpoint inhibitors. *J. Immunother. Cancer* **7**, 278 (2019).
102. Zhu, X., Fang, H., Gladysz, K., Barbour, J. A. & Wong, J. W. H. Overexpression of transposable elements is associated with immune evasion and poor outcome in colorectal cancer. *Eur. J. Cancer* **157**, 94–107 (2021).
103. Dahl, J. A. et al. Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature* **537**, 548–552 (2016).

104. Matsui, T. et al. Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. *Nature* **464**, 927–931 (2010).
105. Concordet, J. P. & Haeussler, M. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* **46**, W242–W245 (2018).
106. Zhou, X., Sam, T. W., Lee, A. Y. & Leung, D. Mouse strain-specific polymorphic provirus functions as cis-regulatory element leading to epigenomic and transcriptomic variations. *Nat. Commun.* **12**, 6462 (2021).
107. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
108. Akalin, A. et al. methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* **13**, R87 (2012).
109. Wang, H. Q., Tuominen, L. K. & Tsai, C. J. SLIM: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics* **27**, 225–231 (2011).
110. He, Y. et al. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* **583**, 752–759 (2020).
111. Xie, W. et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
112. Decato, B. E. et al. Characterization of universal features of partially methylated domains across tissues and species. *Epigenetics Chromatin* **13**, 39 (2020).
113. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
114. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
115. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187–W191 (2014).
116. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
117. Zhou, Y. et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).
118. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
119. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
120. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
121. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
122. Stovner, E. B. & Saetrom, P. epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics* **35**, 4392–4393 (2019).
123. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).
124. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
125. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
126. Reiff, S. B. et al. The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. *Nat. Commun.* **13**, 2365 (2022).
127. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. *Nat. Protoc.* **15**, 991–1012 (2020).
128. Wilks, C. et al. recount3: Summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).
129. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
130. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR Project GENIE biopharma collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
131. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).
132. Cerami, E. et al. The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
133. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinforma.* **9**, 559 (2008).
134. Bastian, M., Heymann, S. & Jacomy, M. in *Proceedings of the international AAAI conference on web and social media*. 361-362.
135. Wilks, C. et al. Megadepth: Efficient coverage quantification for BigWigs and BAMs. *Bioinformatics* **37**, 3014–3016 (2021).
136. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).

Acknowledgements

This work is supported by the Hong Kong Epigenome Project (Lo Ka Chung Charitable Foundation) and the Croucher Foundation (CIA16SC02). It is also partially supported by the University Grants Committee through the Collaborative Research Fund (C407-14G), General Research Fund (GRF 16104122 and 16103721), and the CUHK Strategic Seed Funding for Collaborative Research Scheme. We thank Prof. Nathalie Wong for sharing the PLC cell line, Dr. Gary Hon for sharing the Lenti-dCas9-KRAB-blast plasmid, Dr. Ting Wang for sharing the pHRdSV40-dCas9-22aalinker-GCN4_V4-P2A-BFP, scFv-sfGFP-DNMT3A1, and scFv-sfGFP-TET1CD plasmids, and all staff in the HKUST Biosciences Central Research Facility for their assistance in carrying out the study. We also acknowledge the Center for PanoOmic Sciences (CPOS) at the University of Hong Kong for generating a subset of the HCC WGBS data.

Author contributions

C.C.Y.C., K.Y.Y., A.S.C., and D.L. conceived the study. P.B.S.L. and W.Y. collected and processed the patient samples. M.F.C., A.Y.L., C.C.Y.C., G.L., and J.Y.J.A. performed all the experiments. C.C.Y.C., I.R.M., M.F.C., Q.W., H.W., and S.C.H.C. conducted all data analyses. C.C.Y.C., K.Y.Y., A.S.C., and D.L. prepared the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-09154-w.

Correspondence and requests for materials should be addressed to Kevin Y. Yip, Alfred S. Cheng or Danny Leung.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Bibekanand Mallick and Johannes Stortz. [A peer review file is available].