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

Supplementary Figures 1-10

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[6], Kevin Y. Yip[4,5]*, Alfred S. Cheng[3]* and Danny Leung[1,2]*

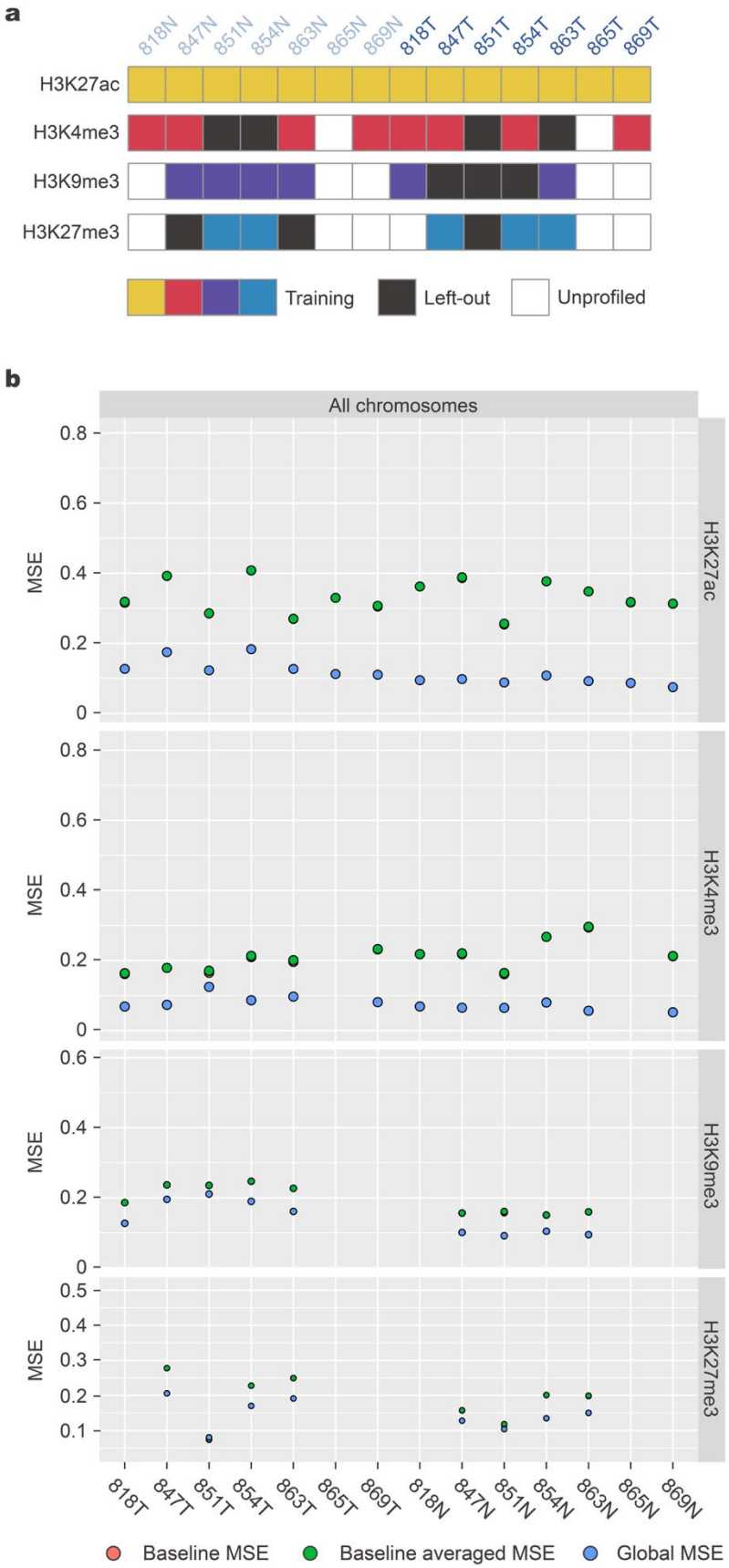
Affiliations

1. Division of Life Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
2. Center for Epigenomics Research, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
3. School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
4. Cancer Genome and Epigenetics Program, NCI-Designated Cancer Centre, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California 92037, USA
5. Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
6. Department of Surgery, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

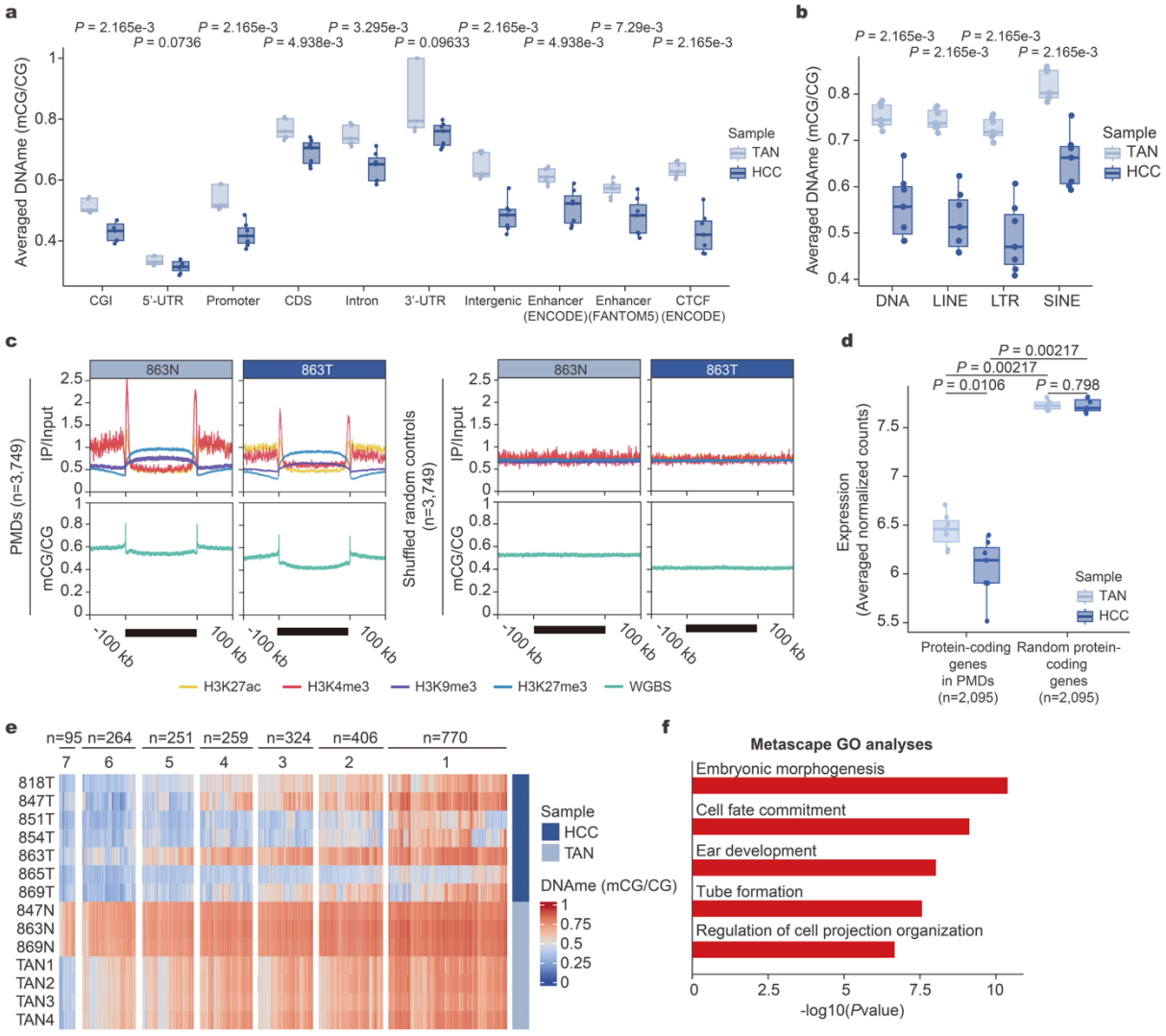
* Corresponding authors: kyip@sbpdiscovery.org (K.Y.Y.), alfredcheng@cuhk.edu.hk (A.S.C.), dcyleung@ust.hk (D.L.)

This PDF file includes:

Supplementary Figures and Legends 1-10

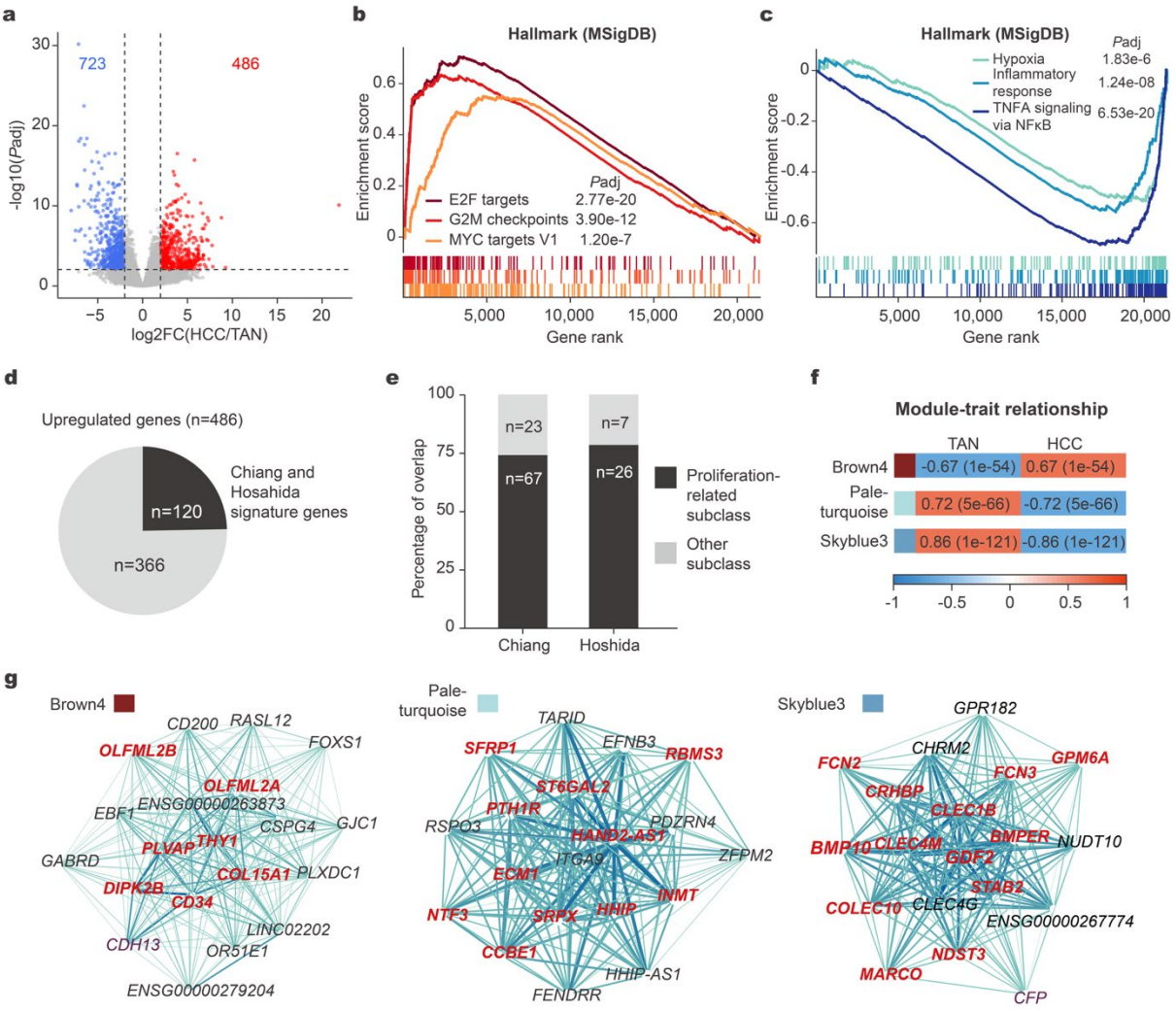

a
818N
847N
851N
854N
863N
865N
869N
818T
847T
851T
854T
863T
865T
869T
H3K27ac
H3K4me3
H3K9me3
H3K27me3
Training
Left-out
Unprofiled
b
All chromosomes
MSE
H3K27ac
H3K4me3
H3K9me3
H3K27me3
Baseline MSE
Baseline averaged MSE
Global MSE

Supplementary Fig. 1. Quality controls of Avocado-imputed ChIP-seq signals. **a** Schematic of model training strategy. Models were trained on a per chromosome basis and 10 out of the 43 samples were left-out to assess the per-sample global mean-square error (MSE) of the imputed signal. **b** Per-sample global MSE compared to per-sample baseline MSE and baseline MSE averaged across all other samples.



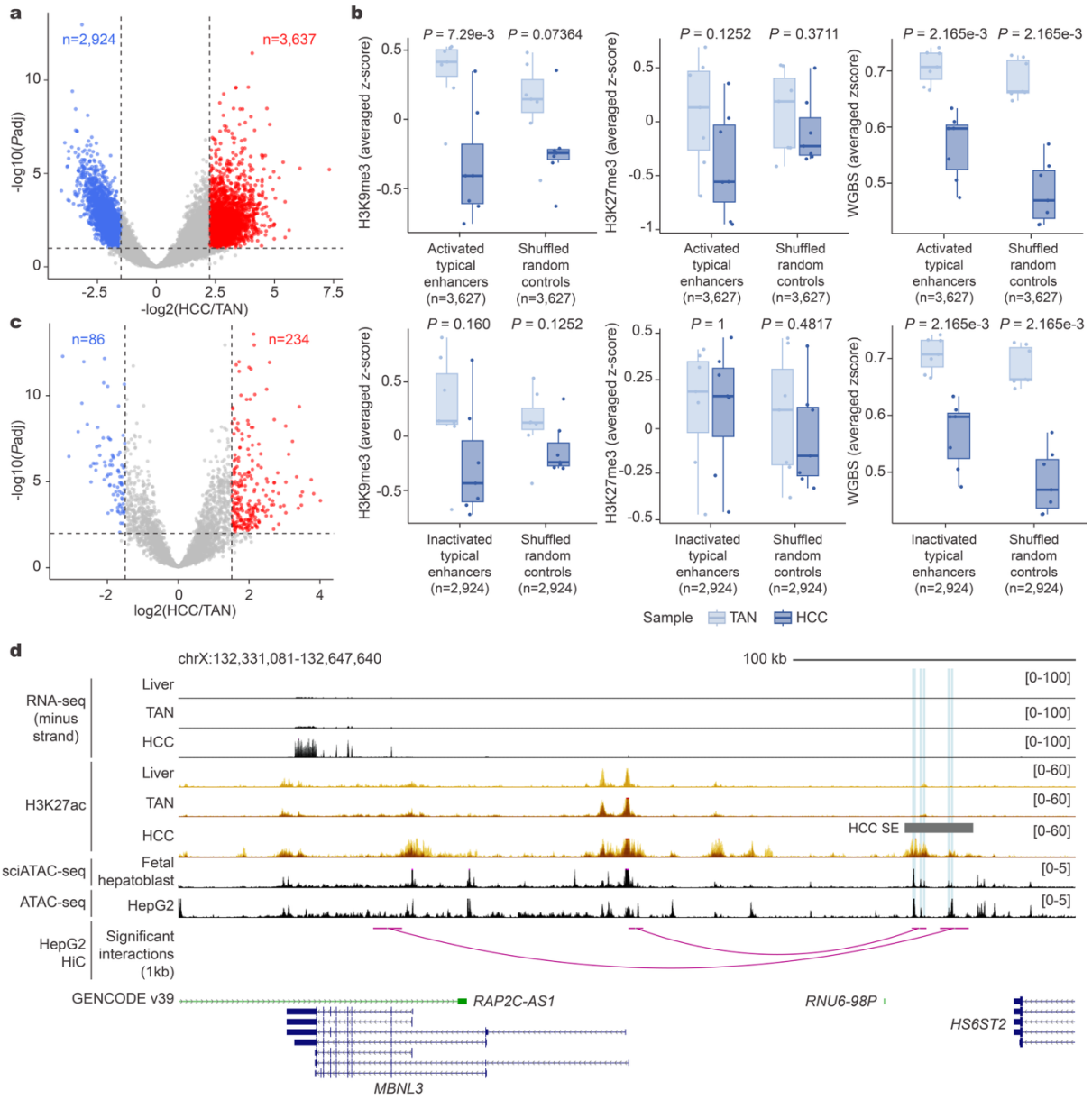
Supplementary Fig. 2. Global DNA hypomethylation and focal DNA hypermethylation in HCC. a-b Boxplots showing the averaged DNA methylation of CGI, genic features (5'-UTR, CDS, intron, and 3'-UTR), CREs (promoters and enhancers), and CTCF sites (**a**), and transposable elements (**b**) in HCC and TAN. **c** Aggregation plots of DNA methylation and histone mark signals in defined PMDs in patient 863. 100-kb flanking regions of defined PMDs are also shown. **d** Boxplots comparing the expression of protein-coding genes that reside within PMDs between HCC and TAN. **a**,**b**,**d** *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. **e** Heatmap showing the DNA methylation levels of PMDs in HCC and TAN tissues. PMDs are categorized into 7 clusters based on the number of PMDs shared among the HCC samples. **f** Gene ontology analyses of genes with promoter DNA hypermethylation. *P* values (one-tailed hypergeometric test) for the top relevant GO terms are shown.



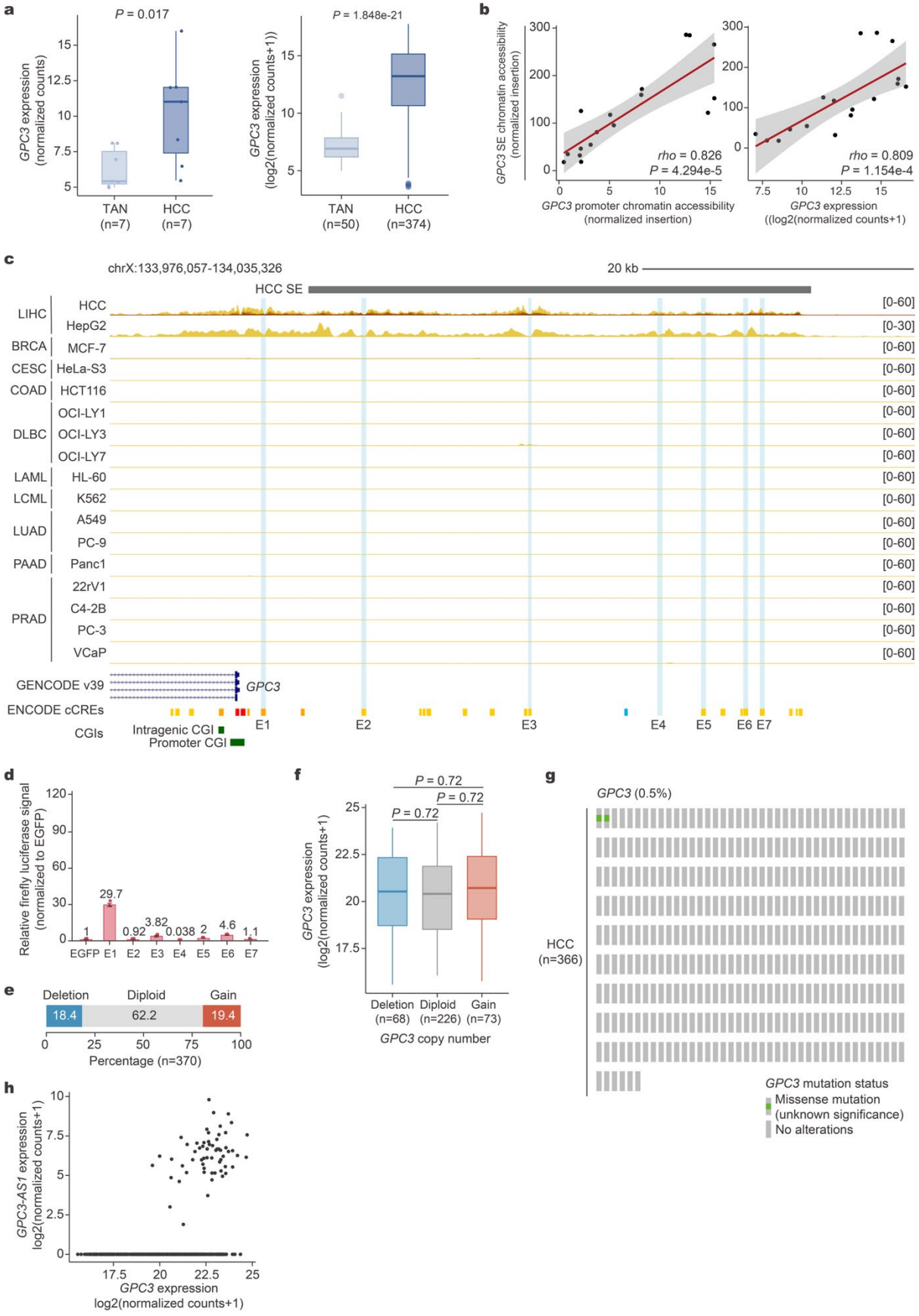
Supplementary Fig. 3. HCC exhibits transcriptomic signatures of the proliferation class. a Volcano plot showing dysregulated genes in HCC. For each gene, the -log10-transformed, two-tailed adjusted *P* values (*P*adj) is plotted against the log2-transformed fold change (log2FC). Upregulated (*P*adj < 0.01 and log2FC > 1; n=486) and downregulated (*P*adj < 0.01 and log2FC < -1; n=723) genes in HCC are labeled in red and blue, respectively. Dashed lines represent the indicated thresholds. **b-c** Gene set enrichment analyses showing the biological pathways positively (**b**) or negatively (**c**) associated with the dysregulated genes. **d** Pie chart showing the number of upregulated genes from previously defined molecular subtypes[1,2] **e** Barplots showing the number of upregulated genes from previously defined molecular subtypes in **d** that are associated with cell proliferation pathways. Note that 3 upregulated genes

were classified as signature genes under both molecular subtypes. **f** WGCNA modules significantly associated with HCC (Brown4) or TAN (Paleturqoise and Skyblue3). **g** Network graphs of modules significantly associated with HCC (Brown 4, left) or TAN (Paleturqoise, middle and Skyblue3, right). Dysregulated genes defined in our HCC samples are highlighted in red within each module. In the network, each node represents a gene and each edge represents the co-expression relationship between two genes.



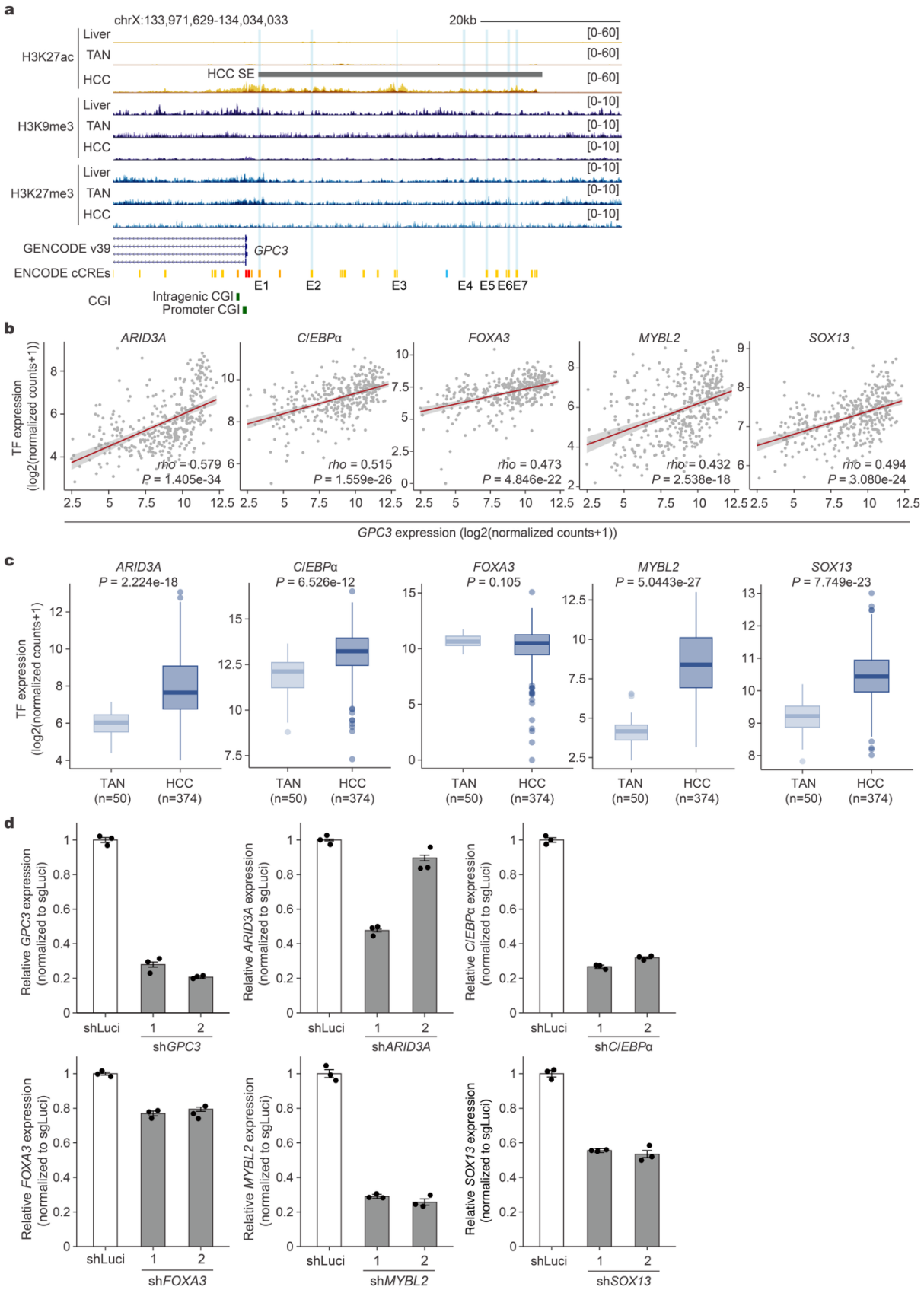
Supplementary Fig. 4. Characterization of dysregulated typical enhancers and SEs in HCC. a Volcano plot showing dysregulated typical enhancers based on differential H3K27ac enrichment in HCC. For each typical enhancer, the -log10-transformed, two-tailed adjusted *P* values (*P*adj) is plotted against the log2-transformed fold change (log2FC). Activated (*P*adj < 0.01 and log2FC > 2.25; n=3,637) and inactivated typical enhancers (*P*adj < 0.01 and log2FC < −1.5; n=2,924) in HCC are labeled in red and blue, respectively. Dashed lines represent the indicated thresholds. **b** Boxplots comparing the averaged H3K9me3 (left) and H3K27me3

(middle) signals and averaged DNA methylation (right) of dysregulated typical enhancers between HCC and TAN. Each dot represents a sample. *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. **c** Volcano plot showing dysregulated SEs based on differential H3K27ac enrichment in HCC. Activated (*P*adj < 0.01 and log2FC > 1.5, n=234) and inactivated SEs (*P*adj < 0.01 and log2FC < −1.5, n=86) in HCC are labeled in red and blue, respectively. Dashed lines represent the indicated thresholds. **d** Genome browser screenshot showing an activated SE and several activated typical enhancers in HCC at the *MBNL3* locus. The activated SE is annotated with a gray bar and the activated typical enhancers are highlighted by cyan shadings. Significant chromatin interactions are shown as arcs and the color denotes statistical significance. Stranded RNA-seq tracks are displayed as TPM, H3K27ac ChIP-seq and ATAC-seq tracks are displayed as fold enrichment over input, and sciATAC-seq track is shown as fragment count. Liver RNA-seq track is from Roadmap Epigenomics Project, sciATAC-seq for fetal hepatoblast is from CATlas, and ATAC-seq for HepG2 is from ENCODE. All tracks except Liver RNA-seq are shown as composite signals by overlaying tracks across samples.

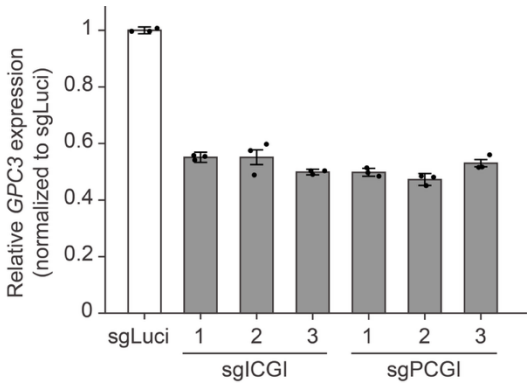


Supplementary Fig. 5. Aberrant reactivation of super-enhancer is a key

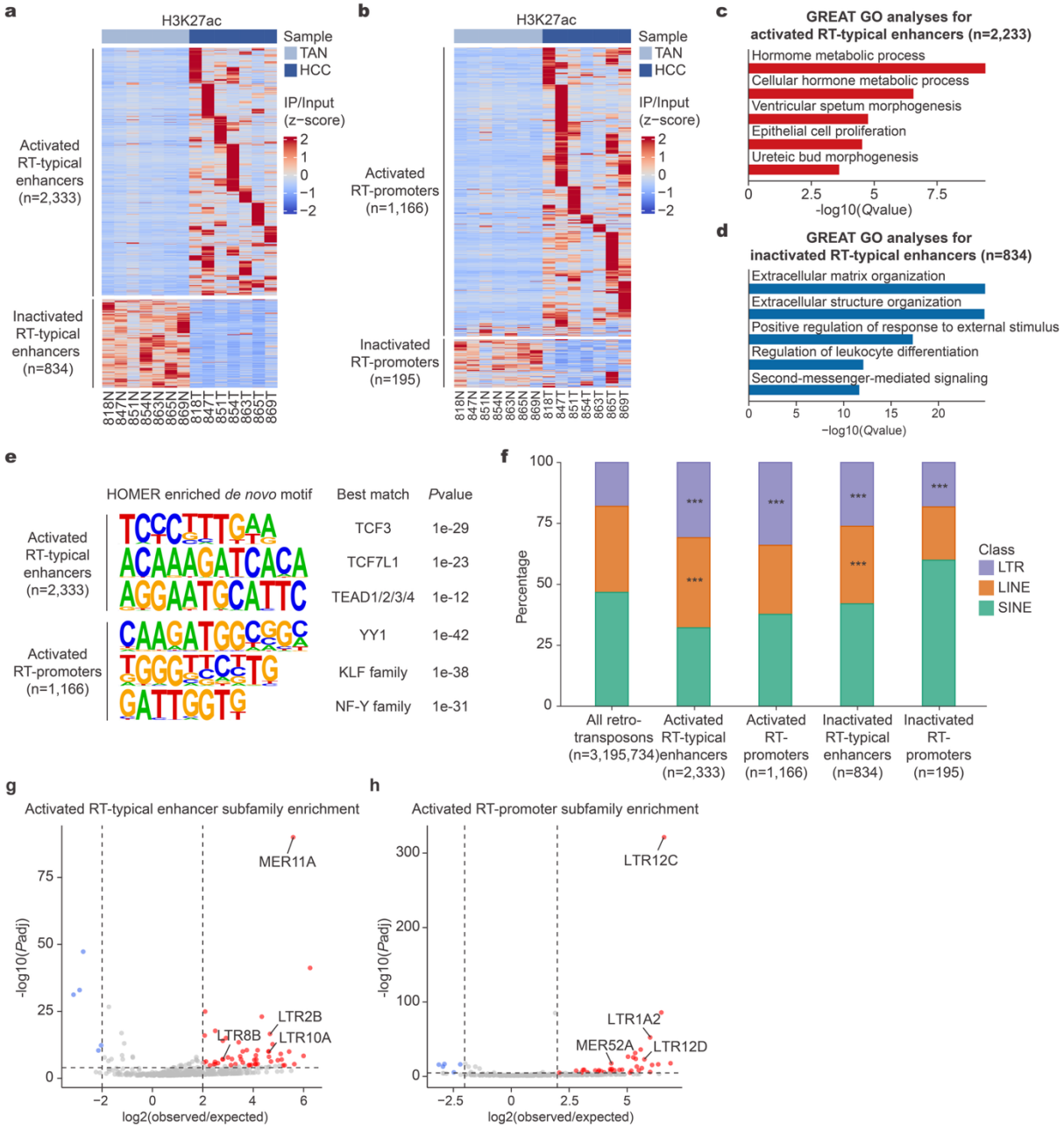
regulatory mechanism underlying *GPC3* overexpression in HCC. a Boxplots comparing the expressions of *GPC3* between TAN and HCC samples in our patient cohort (left) and in the TCGA-LIHC cohort (right). **b** Scatterplots showing the correlations between the chromatin accessibility of *GPC3* SE and *GPC3* promoter (left) or *GPC3* expression (right). Spearman's correlation coefficient (*rho*) and *P* values are indicated at the bottom right of each scatterplot. **c** Genome browser screenshot showing the H3K27ac signals at the *GPC3* SE in HCC and cancer cell lines representing different cancer types. Cancer type abbreviations follow TCGA definitions. *GPC3* constituent enhancers (E1-E7) are highlighted by cyan shading. **d** Luciferase assay assessing the promoter activities of E1-E7. Data represent the mean ± standard deviation of three technical replicates from one representative experiment (n=2 independent experiments). **e** Copy number status of *GPC3* in the TCGA-LIHC cohort. **f** Boxplots comparing the expression of *GPC3* among HCC patients with different copy number status of *GPC3* in the TCGA-LIHC cohort. **g** Mutation status of *GPC3* in the TCGA-LIHC cohort. **h** Scatterplot showing the correlation between the expressions of *GPC3* and *GPC3-AS1* in the TCGA-LIHC cohort. Each dot represents a patient. **a**,**f** *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 $1-1.5\times$the interquartile range and quartile $3+1.5\times$the interquartile range, respectively.


a
chrX:133,971,629-134,034,033
20kb
H3K27ac
Liver
TAN
HCC
HCC SE
[0-60]
H3K9me3
[0-10]
H3K27me3
GENCODE v39
GPC3
ENCODE cCREs
E1 E2 E3 E4 E5 E6E7
CGI
Intragenic CGI
Promoter CGI
b
ARID3A
C/EBPα
FOXA3
MYBL2
SOX13
TF expression (log2(normalized counts+1))
rho = 0.579
P = 1.405e-34
rho = 0.515
P = 1.559e-26
rho = 0.473
P = 4.846e-22
rho = 0.432
P = 2.538e-18
rho = 0.494
P = 3.080e-24
GPC3 expression (log2(normalized counts+1))
c
P = 2.224e-18
P = 6.526e-12
P = 0.105
P = 5.0443e-27
P = 7.749e-23
TAN (n=50)
HCC (n=374)
d
Relative GPC3 expression (normalized to sgLuci)
Relative ARID3A expression (normalized to sgLuci)
Relative C/EBPα expression (normalized to sgLuci)
Relative FOXA3 expression (normalized to sgLuci)
Relative MYBL2 expression (normalized to sgLuci)
Relative SOX13 expression (normalized to sgLuci)
shLuci
shGPC3
shARID3A
shC/EBPα
shFOXA3
shMYBL2
shSOX13

Supplementary Fig. 6. Loss of DNA methylation and increased binding of key TFs contribute to the aberrant reactivation of the *GPC3* SE. a Genome browser screenshot showing the subtle changes in H3K9me3 and H3K27me3 at the *GPC3* SE. The SE is annotated with a gray bar. Individual constituent enhancers (E1-E7) are highlighted by cyan shadings. ChIP-seq signals are displayed as fold enrichment over input. All tracks are shown as composite signals by overlaying tracks across samples. **b** Scatterplots showing the correlations between the expressions of five candidate TFs and that of *GPC3* in the TCGA-LIHC cohort. Spearman's correlation coefficient (*rho*) and *P* values are indicated at the bottom right of each scatterplot. **c** Boxplots comparing the expressions of the five candidate TFs between TAN and HCC tissues in the TCGA-LIHC cohort. *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. **d** RT-qPCR of the knockdown efficiency of each candidate TF. 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 expressions and the values are normalized to that of shLuci.

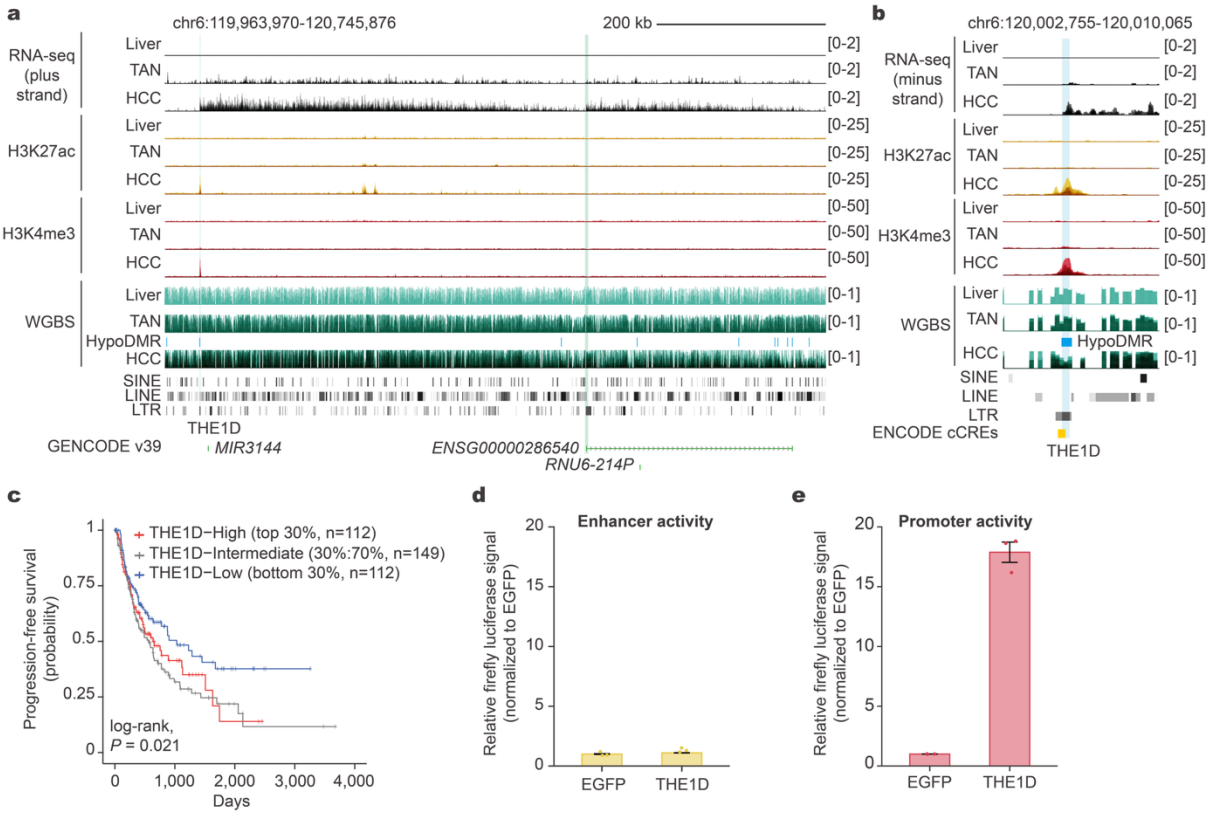


Supplementary Fig. 7. DNA methylation of either the intragenic CGI or promoter CGI is sufficient to reduce the expression of *GPC3*. b RT-qPCR of the relative expression of *GPC3* upon targeted DNA methylation of the intragenic CGI or promoter CGI. Each dot represents a technical replicate and error 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.



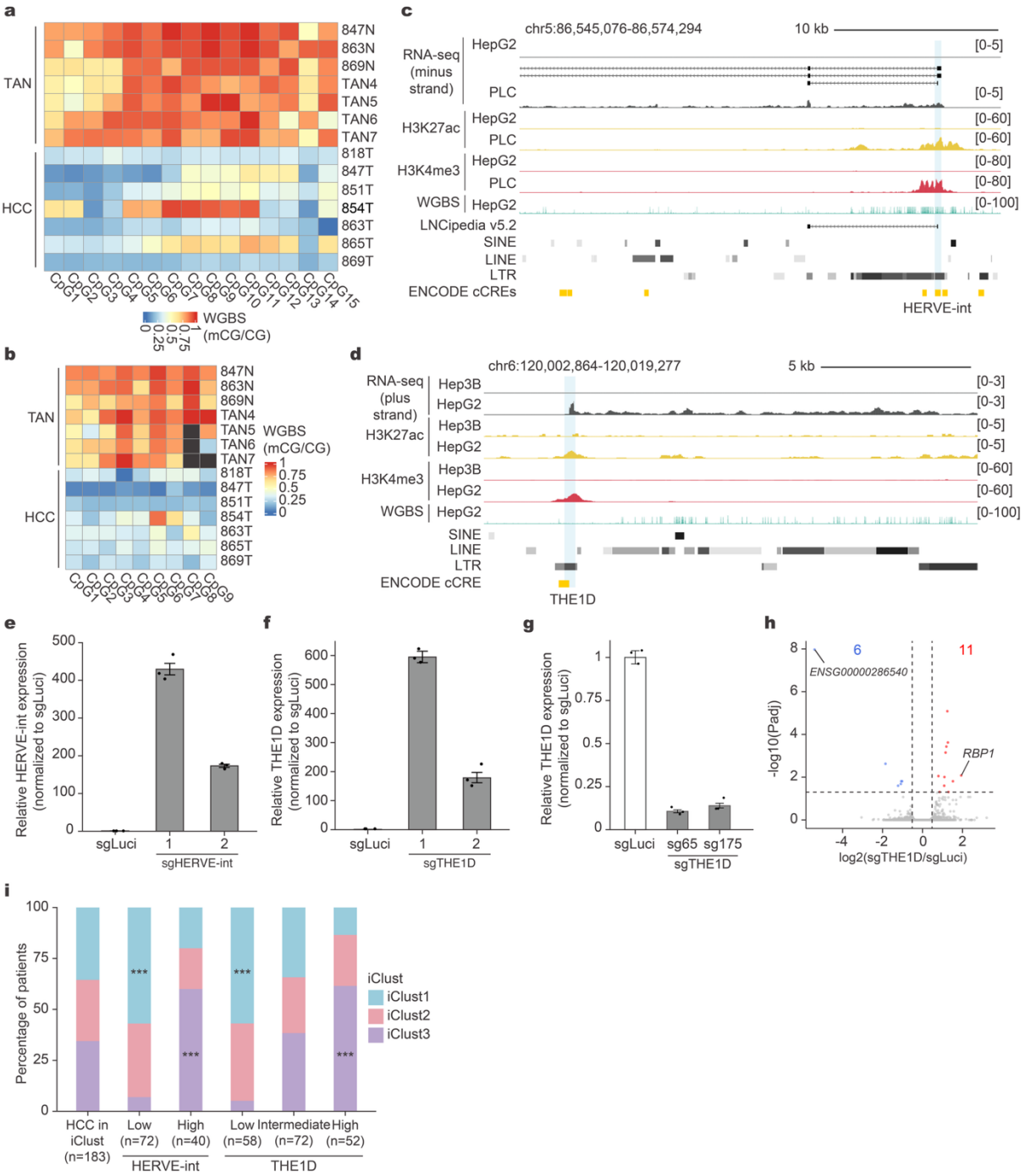
Supplementary Fig. 8. Epigenetic dysregulation of retrotransposon-derived CREs in HCC. a-b Heatmap of H3K27ac signal of dysregulated RT-typical enhancers (**a**) and RT-promoters (**b**) in HCC. Each row represents an element and each column represents a sample. **c-d** Top relevant GO terms associated with the activated (**c**) and inactivated (**d**) RT-typical enhancers. **e** HOMER motif analyses of activated RT-typical enhancers and activated RT-promoters. **f** Barplots comparing the percentage of the three classes of retrotransposon among the dysregulated retrotransposon-derived

CREs. Activated RT-typical enhancers: LTR (*P* = 2.41e-131), LINE (*P* = 1.12e-41); activated RT-promoters: LTR (*P* = 2.442e-87); inactivated RT-typical enhancers: LTR (*P* = 4.760e-51), LINE (*P* = 1.47e-4); inactivated RT-promoters: LTR (*P* = 2.487e-5). *P* values (one-tailed hypergeometric test corrected by the Benjamini-Hochberg approach) are shown. **g,h** Retrotransposon subfamily enrichment among activated RT-typical enhancers (**g**) and activated RT-promoters (**h**) in HCC. Over-represented (observed/expected>4 and *P*adj<0.01) and depleted subfamilies (observed/expected < 4 and *P*adj < 0.01) are labeled in red and blue, respectively. Selected over-represented subfamilies are indicated.



Supplementary Fig. 9. Epigenetic dysregulation of THE1D-driven transcript in HCC. a-b Genome browser screenshot showing THE1D serving as the alternative promoter for the pseudogene *ENSG00000286540* in HCC (**a**) and a zoom-in screenshot of the THE1D element gaining active promoter-associated histone modifications (**b**). THE1D is highlighted by cyan shading and the *ENSG00000286540* pseudogene promoter is highlighted by green shading, respectively. HypoDMR in HCC is annotated with a blue bar. 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. 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 THE1D expression among HCC patients in the TCGA-LIHC cohort. Based on the THE1D expression, patients were stratified into low (bottom 30%), intermediate (30%<x<70%) and high (top 30%) groups, respectively. *P* value (log-rank test) is

shown. **d**-**e** Luciferase assays assessing the enhancer (**d**) and (**e**) promoter activities of THE1D. Data represent the mean ± standard deviation of three technical replicates from one representative experiment (n=2 independent experiments).



Supplementary Fig. 10. DNA hypomethylation is sufficient for transcriptional de-repression of HERVE-int and THE1D in HCC. a,b Heatmap comparing the DNA methylation of HERVE-int (**a**) or THE1D (**b**) between HCC and TAN. Each row represents a sample and each column represents a CpG within HERVE-int or THE1D. **c,d** Genome browser screenshot showing the differential expression of HERVE-int in PLC cells (**c**) or THE1D in HepG2 cells (**d**). StringTie-assembled transcripts are shown

above the HCC RNA-seq track. HERVE-int and THE1D are highlighted by cyan shading. HepG2 WGBS track is from ENCODE and shown as averaged signals. Stranded total RNA-seq tracks are displayed as TPM, ChIP-seq tracks are displayed as fold enrichment over input, WGBS tracks are shown as averaged methylation levels. **e** RT-qPCR of relative HERVE-int expression upon targeted DNA demethylation in HepG2 cells. **f** As in **e** but for THE1D in Hep3B cells. **g** RT-qPCR of the relative expression of THE1D upon epigenetic silencing. sgRNA targeting luciferase (sgLuci) is used as the negative control. *GAPDH* serves as the internal control for the calculation of the relative THE1D expression and the values are normalized to that of sgLuci. **h** Volcano plot showing dysregulated genes in HepG2 cells depleted of THE1D. For each gene, the -log10-transformed, two-tailed adjusted *P*values (*P*adj) is plotted against the log2-transformed fold change (log2FC). Upregulated (*P*adj < 0.05 and log2FC > 0.5) and downregulated (*P*adj < 0.05 and log2FC < -0.5) are labeled in red and blue, respectively. Dashed lines represent the indicated thresholds. Selected dysregulated genes are labeled. **i** Barplots comparing the percentage of HERVE-int-Low/-High and THE1D-Low/-Intermediate/-High patients annotated with the iClust1-3 definitions from TCGA. HERVE-int-Low: iClust1 (*P* = 7.148e-6); HERVE-int-High: iClust3 (*P* = 4.915e-4); THE1D-Low: iClust1 (*P* = 2.045e-4). THE1D-High (*P* = 1.675e-5). *$P < 0.05$ and ***$P < 0.001$, one-tailed hypergeometric test corrected by the Benjamini-Hochberg approach. **e**,**f,g** 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 HERVE-int and THE1D expressions and the values are normalized to that of sgLuci.

REFERENCES

1 Chiang, D. Y. *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* **68**, 6779-6788 (2008). https://doi.org:10.1158/0008-5472.CAN-08-0742

2 Hoshida, Y. *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res* **69**, 7385-7392 (2009). https://doi.org:10.1158/0008-5472.CAN-09-1089

3 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). https://doi.org:10.1038/s41556-023-01169-x