Materials and methods

EC dataset source and preprocessing

The workflow of this study was shown in Figure S1 of Supplementary materials. Public gene-expression data and full clinical annotation were searched in the Cancer Genome Atlas (TCGA) database. As to datasets in TCGA, RNA sequencing data (FPKM value) of gene expression were downloaded from the Genomic Data Commons (GDC, https://portal.gdc.cancer.gov/) using the R package TCGA bio links, which was specifically developed for integrative analysis with GDC data [1]. Batch effects from non-biological technical biases were corrected using the "Combat" algorithm of sva package. The baseline information of all eligible EC datasets was summarized in Table S1. The somatic mutation data was acquired from TCGA database. The AS data for each EC patients were analyzed by SpliceSeq. The percent spliced-in (PSI) value calculated by SpliceSeq is used to indicate the reliability of each AS event, and the missing PSI values were imputed using missForest (version 1.4). Data were analyzed with the R (version 3.6.1) and R Bioconductor packages.

Unsupervised clustering for 22 m^6A regulators

A total of 22 regulators were extracted from Cancer Genome Atlas (TCGA) database for identifying different m^6A modification patterns mediated by m^6A regulators. These 22 m^6A regulators included 7 writers (METTL3, METTL14, METTL16, RBM15, RBM15B, WTAP and ZC3H13), 2 erasers (ALKBH5 and FTO) and 13 readers (RBMX, YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, HNRNPA2B1, HNRNPC, FMR1, LRPPRC, IGF2BP1, IGF2BP2, IGF2BP3) (Tables. S2 and S3). Unsupervised clustering analysis was applied to identify distinct m^6A modification patterns based on the expression of 22 m^6A regulators and classify patients for further analysis. The number of clusters and their stability were determined by the consensus clustering algorithm. We used the ConsensuClusterPlus package to perform the above steps and 1000 times repetitions were conducted for guaranteeing the stability of classification.

Gene set variation analysis (GSVA) and functional annotation

To investigate the difference on biological process between m^6A modification patterns, we performed GSVA enrichment analysis using "GSVA" R packages. GSVA, in a non-parametric and unsupervised method, is commonly employed for estimating the variation in pathway and biological process activity in the samples of an expression dataset. The gene sets of "c2.cp.kegg.v6.2.-symbols" and "h.all.v7.4-symbols" were downloaded from GSEA database for running GSVA analysis. Adjusted p with value less than 0.05 was considered as statistically

significance. The cluster Profiler R package was used to perform functional annotation for m^6A-related genes, with the cutoff value of $p < 0.05$.

Identification of differentially expressed genes (DEGs) between m^6A distinct phenotypes

To identify m^6A-related genes, we classified patients into two distinct m^6A modification patterns based on the expression of 22 m^6A regulators. The empirical Bayesian approach of limma R package was applied to determine DEGs between different modification patterns. The significance criteria for determining DEGs was set as adjusted p value < 0.05.

Generation of m^6A gene signature

To quantify the m^6A modification patterns of individual tumor, we constructed a set of scoring system to evaluate the m^6A modification pattern of individual patients with EC—the m^6A gene signature, and we termed as m^6Ascore. The procedures for establishment of m^6A gene signature were as follows:

The DEGs identified from different m^6A clusters were firstly normalized among all ACRG samples and the overlap genes were extracted. The patients were classified into several groups for deeper analysis by adopting unsupervised clustering method for analyzing overlap DEGs. The consensus clustering algorithm was utilized for defining the number of gene clusters as well as their stability. Then, we performed the prognostic analysis for each gene in the signature using univariate Cox regression model. The genes with the significant prognosis were extracted for further analysis. We then conducted principal component analysis (PCA) to construct m^6A relevant gene signature. Both principal component 1 and 2 were selected to act as signature scores. This method had advantage of focusing the score on the set with the largest block of well correlated (or anticorrelated) genes in the set, while down-weighting contributions from genes that do not track with other set members. We then define the m^6Ascore using a method similar to GGI: $m^6Ascore=\sum(PC1_i+PC2_i)$, where i is the expression of m^6A phenotype-related genes [2,3].

Correlation between m^6A gene signature and other related biological processes

We constructed a set of gene sets that stored genes associated with some biological processes, including (1) DNA replication; (2) DNA−dependent DNA replication; (3) chromosome segregation; (4) catalytic activity, acting on DNA; (5) catalytic activity, acting on RNA; (6) Spliceosome; (7) Cell cycle; (8) Ribosome biogenesis in eukaryotes; (9) Mismatch repair; (10) Metabolism of RNA; (11) cellular response to DNA damage stimulus; (12) RNA localization. We then performed a correlation analysis to further reveal the association between m^6A gene signature

and some related biological pathways.

Identification of differentially expressed AS events and enrichment analysis

Differentially expressed AS events (DEAS) and differentially expressed genes (DEG) were analyzed through the limma package (version 3.42.0). Adj. *p*-value < 0.05 was used as the threshold to prevent skipping significant changes. The interactive sets between the seven types of reliable DEAS events were illustrated by the distinguishable visualization Upset plot (UpSetR, version 1.3.3) and the differences among DEAS and DEG were illustrated using Venn diagrams. Subsequently, the parent genes of these significantly DEAS were used as the background in enrichment analysis using Metascape [4]. Adj. *p*. value< 0.05 was statistically significant.

Survival Analysis

According to the survival information of two groups, univariate multi-Cox regression and lasso regression were used to determine DEAS, which was significant to overall survival (OS). Further, to establish a rigorous prognostic model, signature DEAS with an area under the receiver operating characteristic (ROC) curve (AUC) greater than 0.6 was selected as candidates for multivariate Cox regression. Then, the risk score for each sample was calculated based on the PSI values of the prognostic DEAS signatures and the corresponding coefficients. EC samples were subsequently divided into two subgroups by the median risk score: high risk group and low risk group. Kaplan-Meier analysis was used to test the model's ability to distinguish patient's survival. All the reported *p*-values were less than 0.05. All the analyses were performed using RStudio (version 3.5.2).

Development and Apparent Performance of a DEAS-Clinicopathologic Nomogram

We combined multivariable Cox regression analysis and all informative clinicopathologic variables described above to formulate a nomogram for the better prediction of the individualized survival rates of EC patients [5]. Backward stepwise variable selection with the Akaike information criterion (AIC) was performed to determine the variables included in the final nomogram [6]. Then, the predictive accuracy of the final nomogram was evaluated by Calibration curves. To further identify the predictive efficiencies of the model, the Uno's inverse-probability of censoring weighting estimation of the dynamic time-dependent ROC area under the curve (AUC) values (time span from 0 to 3 years) was calculated with time ROC package (version 0.3) [7].

Cell apoptosis assays

KYSE150 cells transfected with ABI1|11037|ES-WT or ABI1|11037|ES-MUT or the negative control were harvested and rinsed twice with pre-cooling PBS. The samples were diluted with 150 μl of 1×annexin-binding buffer, then 5 μl of FITC-labeled enhanced annexin V and 5 μl (20 μg/ml) of propidium iodide (PI, Beyotime, China) were added. Then the cells were incubated in the dark for 15 minutes at room temperature. Flow cytometry was conducted on a FACSCalibur instrument (BD, America).

5-Ethynyl-2′-deoxyuridine (EdU) assays

Proliferation of ESCC cells were monitored using BeyoClick™ EdU Cell Proliferation Kit with Alexa Fluor 488 (Beyotime, China) according to the manufacturer's instructions. KYSE150 cells transfected with ABI1|11037|ES-WT or ABI1|11037|ES-MUT or the negative control were seeded in 96-well plates. After 24 hours, cells were stained with 50 μM EdU for 2 hours. All the cells nuclei were stained with DAPI for one hour, then the cells were then examined using a florescence microscope (Olympus, Japan).

Statistical analysis

One-way ANOVA and Kruskal-Wallis tests were used to conduct difference comparisons of three or more groups. On the basis of the correlation between m^6Ascore and patients' survival, the cut-off point of each data set subgroup was determined using the survminer *R* package. The "surv-cutpoint" function, which repeatedly tested all potential cut points in order for finding the maximum rank statistic, was applied to dichotomize m^6Ascore, and then patients were divided into high and low m^6Ascore groups based on the maximally selected log-rank statistics to decrease the batch effect of calculation. The survival curves for the prognostic analysis were generated via the Kaplan-Meier method and log-rank tests were utilized to identify significance of differences. We adopted a univariate Cox regression model to calculate the hazard ratios (HR) for m^6A regulators and m^6A phenotype-related genes. The independent prognostic factors were ascertained through a multivariable Cox regression model. Patients with detailed clinical data were eligible for final multivariate prognostic analysis. The forest-plot R package was employed to visualize the results of multivariate prognostic analysis for m^6Ascore in TCGA-ESCA cohort. The specificity and sensitivity of m^6Ascore were assessed through receiver operating characteristic (ROC) curve, and the area under the curve (AUC) were quantified using ROC R package. The waterfall function of maftools package was used to present the mutation landscape in patients with high and low m^6Ascore subtype in TCGA-ESCA cohort. The R package of RCircos was adopted to plot the copy number variation landscape of 22 m^6A regulators in 23 pairs of chromosomes. All statistical

p value were two side, with $p < 0.05$ as statistically significance. All data processing was done in R 3.6.1 software.

References

1 Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research* **44**, e71, doi:10.1093/nar/gkv1507 (2016).

2 Choi, S., Lee, S., Kim, Y., Hwang, H. & Park, T. HisCoM-GGI: Hierarchical structural component analysis of gene-gene interactions. *Journal of bioinformatics and computational biology* **16**, 1840026, doi:10.1142/S0219720018400267 (2018).

3 Choi, S., Lee, S. & Park, T. HisCoM-GGI: Software for Hierarchical Structural Component Analysis of Gene-Gene Interactions. *Genomics & informatics* **16**, e38, doi:10.5808/GI.2018.16.4.e38 (2018).

4 Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature communications* **10**, 1523, doi:10.1038/s41467-019-09234-6 (2019).

5 Zhang, Z., Geskus, R. B., Kattan, M. W., Zhang, H. & Liu, T. Nomogram for survival analysis in the presence of competing risks. *Annals of translational medicine* **5**, 403, doi:10.21037/atm.2017.07.27 (2017).

6 Zweig, M. H., Broste, S. K. & Reinhart, R. A. ROC curve analysis: an example showing the relationships among serum lipid and apolipoprotein concentrations in identifying patients with coronary artery disease. *Clinical chemistry* **38**, 1425-1428 (1992).

7 Blanche, P., Dartigues, J. F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in medicine* **32**, 5381-5397, doi:10.1002/sim.5958 (2013).