Methods and Materials

Patient population

This study consisted of 327 individuals above 18 years and comprised blood samples from patients with esophageal cancer (n=150) and healthy controls (n=177 plus 6 replicative samples). Both Patients and controls were enrolled in this study between September 2016 and July 2017. Peripheral blood samples from esophageal cancer patients were obtained preoperatively from department of thoracic surgery, the First Affiliated Hospital of Zhengzhou University, Zhengzhou, China. Patients treated with surgery, chemotherapy, radiation or immunotherapy, were excluded from this study. Cancer-free control subjects were recruited from population receiving routine physical examinations at the First Affiliated Hospital of Zhengzhou University. The study was reviewed and approved by the Institutional Review Board of the First Affiliated Hospital of Zhengzhou University. Written informed consent was obtained from each participating subject.

Sample characteristics

Detailed clinical characteristics of the study subjects are shown in Table S1, including gender, age, clinical diagnosis, staging categories, and cancer biomarkers. A total of 150 esophageal cancer patients were enrolled, including 137 squamous carcinoma, 9 adenocarcinoma, 3 small cell cancer and 1 neuroendocrine carcinoma patients. The average age of esophageal cancer patients and healthy controls were 63 and 46, respectively. Male percent of patients and healthy controls were similar, 65.3% and 66.1, respectively. Cancer stage was classified according to staging of cancer of the esophagus and esophagogastric junction of the eighth edition of the AJCC/UICC cancer staging manuals[1]. 31.3% (47 out of 150) of esophageal cancer patients were at advanced stages (stage III and IV). 47 esophageal cancer patients showed lymph node metastasis, and 2 patents had distal metastasis. Seven conventional cancer biomarkers, CEA, AFP, CA72-4, CA19-9, CA15-3, CK19 and CA125, were measured. Only 5, 7, 2 and 4 esophageal cancer patients showed positive for CEA, CA72-4, CA19-9 and CK19 respectively, while 2 healthy subjects showed positive for CA125.

Preparation of cfDNA samples

Circulating cell-free DNA (cfDNA) was isolated from peripheral blood of esophageal cancer patients and healthy individuals. Briefly, peripheral blood (about 4 ml per sample) was collected into EDTA anticoagulant tubes, and plasma was prepared by centrifuging twice at 1,350× g for 12 min, and 13,500 × g for 12 min at 4 °C. cfDNA was isolated using QIAamp Circulating Nucleic Acid Kit (Qiagen) following the company manual.

5hmC library construction and sequencing

5hmC library was constructed according to our previousmethods[2]. Briefly, the cfDNA was repaired and installed with the Illumina compatible adaptors. Ligated DNA was incubated in a 25 μl reaction solution containing 50 mM HEPES buffer (pH 8.0), 25 mM MgCl2, purified DNA, 100 μM N3-UDP-Glc and 1 μM βGT for 1 h at 37 °C. Subsequently, 1 μl DBCO-PEG4-DBCO (Click Chemistry Tools, 4.5 mM stock in DMSO) was added to the reaction mixture and incubated for 2 h at 37 °C. Next, the DNA was purified by a Micro Bio-Spin 30 Column (Bio-Rad). The purified DNA was incubated with 5 μl C1 Streptavidin beads (Life Technologies) in 2× buffer (1× buffer: 5 mM Tris pH 7.5, 0.5 mM EDTA, 1 M NaCl) for 15 min. Then, the beads underwent eight 5 min washes with 1× buffer. The captured DNA fragments were amplified by 14-16 cycles of PCR amplification. The PCR products were purified using AMPure XP beads. Finally, sequencing was performed on the Illumina NextSeq 500 platform.

.

Mapping and detecting differentially modified regions

For the sequencing data, adaptor sequences were trimmed off for all raw reads and reads less than 35 nt in length or containing an ambiguous nucleotide were discarded by Trimmomatic (version 0.33)[3]. The remaining reads were aligned to the human genome (version hg19) using Bowtie 2 (version 2.2.9)[4]. Only uniquely-mapped reads with mapping quality score ≥ 20 were kept for the subsequent analysis for each sample.

The identification of 5hmC-enriched regions (peaks) in each sample was performed using MACS2 (version 2.1.1)[5]. Peaks with high enrichment and significance (q

value<1e-12; fold enrichment > 10) in all samples were considered as high confident peaks. High confident peaks from all samples were combined into one unified catalog for each study separately by ‘‘mergePeak’’ functionality from HOMER (v4.3) (merged peaks: 177484)[6], tag counts in merged high-confident peaks were calculated by BEDTools (v2.26.0)[7] in all samples. The Bioconductor DESeq2 package was used to detect the DhMRs between tumor and control samples (|FC| > 1.2 and p.adj < 1e-7). The HOMER software was used to perform de novo motif analysis around DhMRs.

Detection of Differentially modified Genes and the subsequent analysis

The mapped bam files were handled by Homer software and 5hmC raw counts in genes across samples were calculated[6]. After filtering genes in chromosome X and Y, differentially modified genes in autosome between esophageal carcinoma and healthy control samples were detected using the negative binomial generalized linear model in DESeq2 package[8] (|logFC| >= 0.5 and p.adj =< 0.05). Furthermore, significantly differential genes were analyzed by principle component analysis (PCA) using prcomp function. Hierarchical clustering and heatmap analysis were done by pheatmap package (v1.0.8). GO (Gene ontology) and KEGG pathway analysis were performed by Metascape[9]. The normalized counts in all samples were calculated by rlog function in DESeq2 package. The differential genes, PCA as well as clustering and heatmap were analyzed in R (v4.4.1).

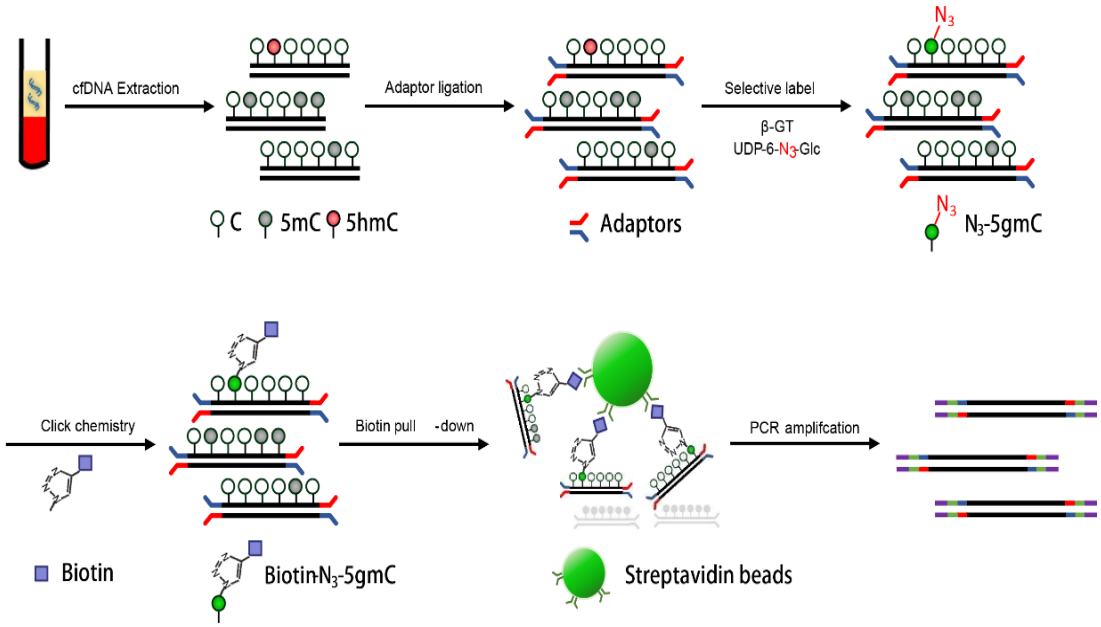
Cancer diagnosis and feature selection

Samples were separated into three subsets (training set 60%, validation set 20% and testing set 20%), normalized 5hmC counts of all genes were analyzed by XGBoost[10] classifier with python API to predict cancer and control. All genes were used to train our model. Reducing the number of gene to carry out training and testing (Figure 1K), feature selections were conducted by 199 training samples only. Validation set was used to tune the hyperparameters of classifiers. Test set used to access the performance is independent of the training and validation set. The classified results were evaluated by AUC (area under curve). Every randomly selected samples or genes set was trained 20

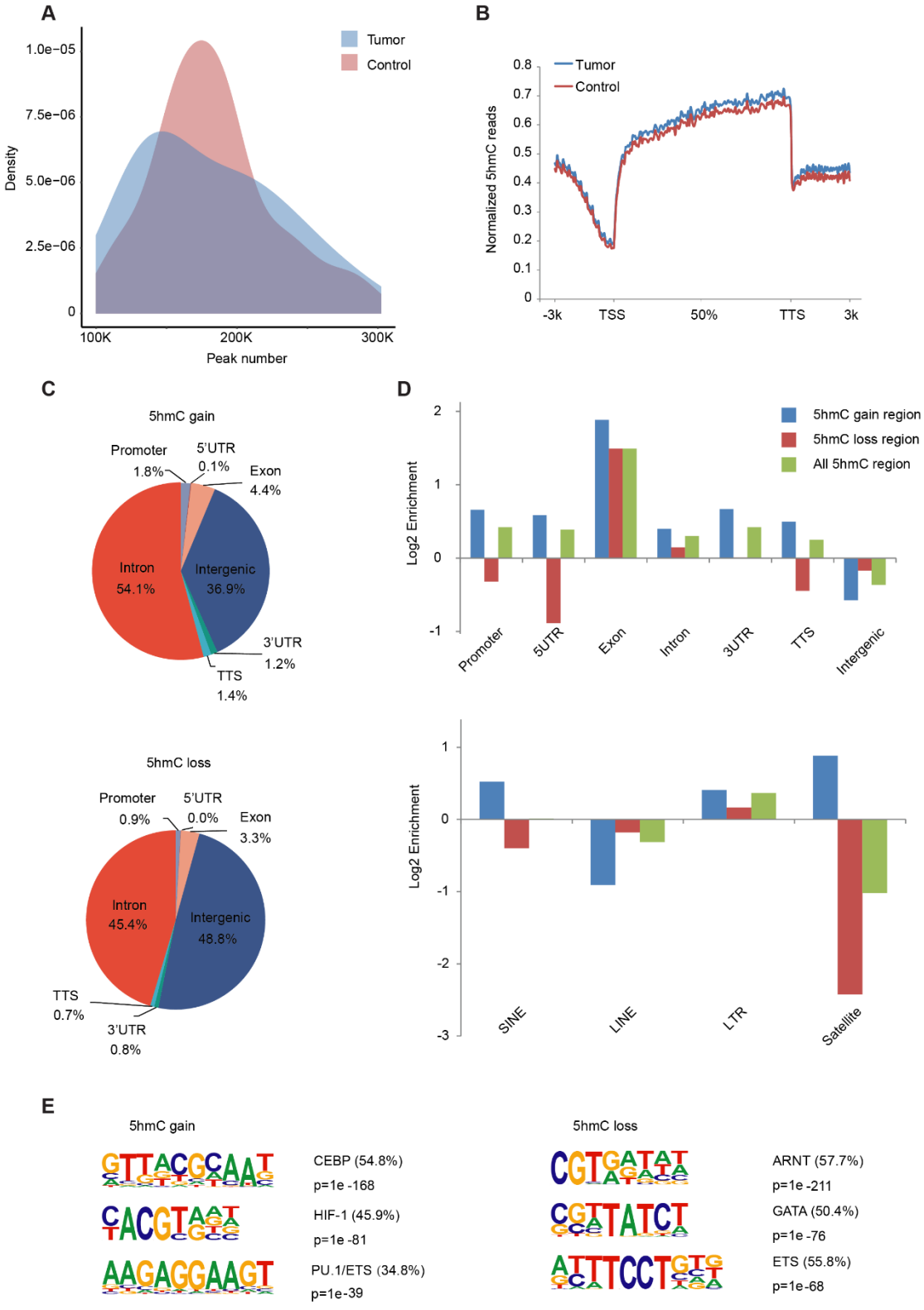
times independently and produced 20 independent models. Top important features (genes) were selected by the F score (feature importance) in model training. The average F score of each gene was calculated after independent repeat of random sub-sampling validation for 100 times.

Data availability statement

The 5hmC-seq data were deposited in the Genome Sequence Archive[11] in BIG Data Center[12], Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number PRJCA000646.



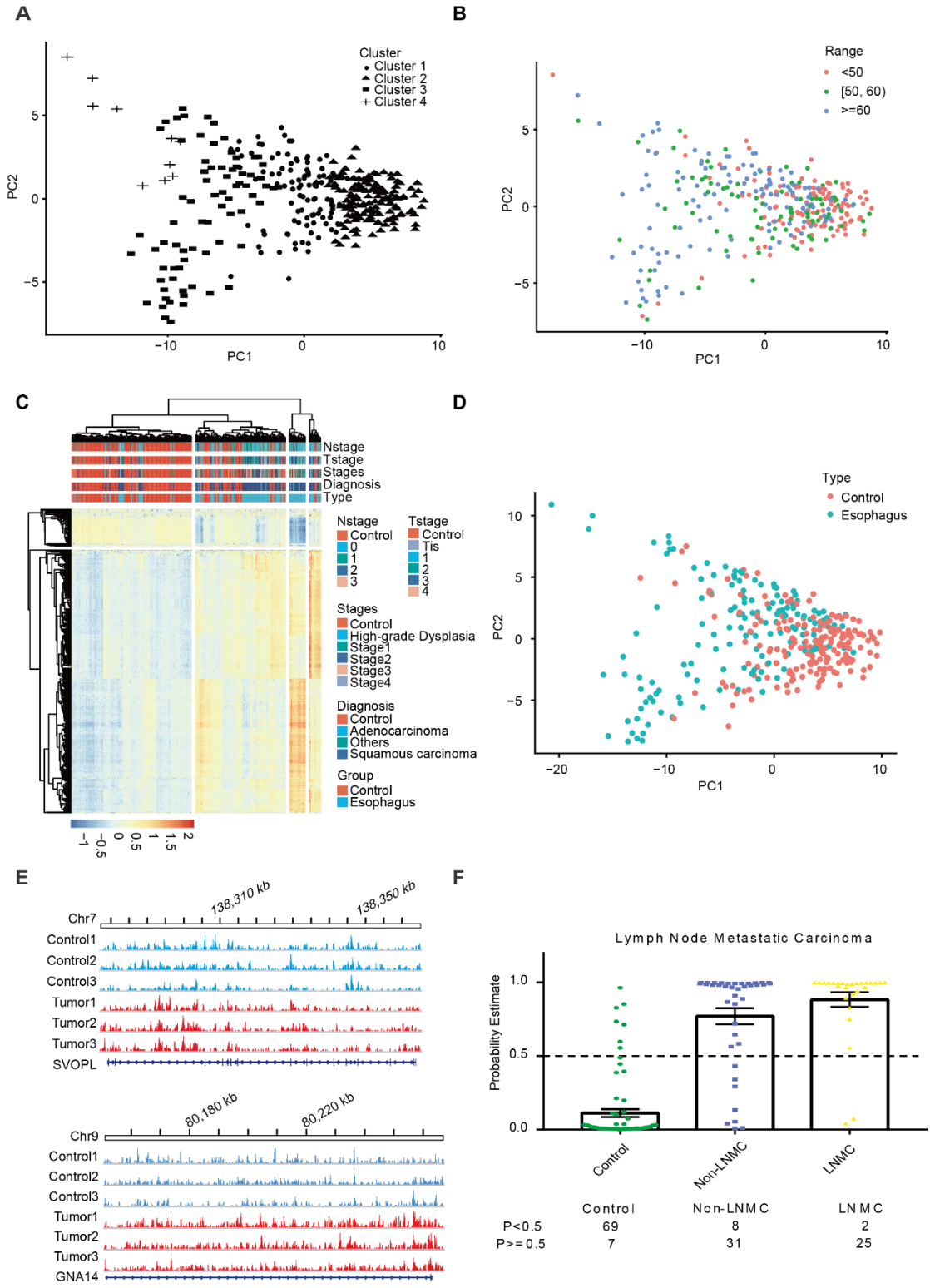
Supplementary information, Figure S1. Workflow of library preparation.
(A) highly sensitive and robust approach for whole-genome 5hmC sequencing was used. Firstly, cfDNA was ligated with sequencing compatible adaptors. And then biotin was used to enrich 5hmC-containing cfDNA fragments by selectively labeling on 5hmC. Finally, 5hmC-containing DNA fragments with biotin tags were efficiently captured by streptavidin beads, followed by PCR amplification and high-throughput sequencing.



Supplementary information, Figure S2. Genome-wide distribution of 5hmC regions in both control and tumor samples.

(A) Density distribution of peak number in control and tumor samples. 137 tumor and 163 control samples with peak number ranging from 10 thousand to 30 thousand were

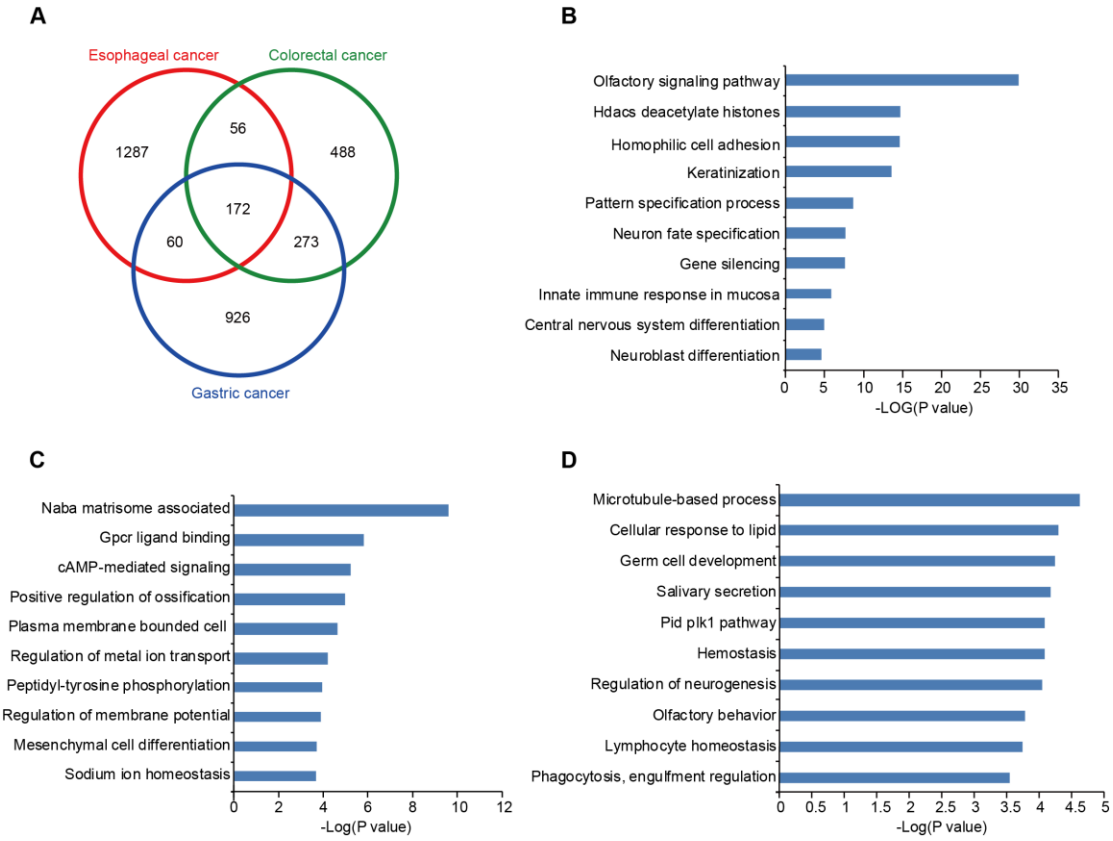
used in the analysis.
(B) Normalized 5hmC reads distributed across meta-gene region in control and tumor samples. Normalized 5hmC reads are the same as average 5hmC enrichment signal, meta-gene was calculated by peak calling with normalization and average signals in the control and tumor groups.
(C) Distribution of differential hydroxymethylated regions (DhMRs) (top: 5hmC gain; bottom: 5hmC loss) in genomic elements. The top panel showed the genomic elements distribution of 5hmC-gain regions in tumor samples versus control samples, and the lower panel showed the distribution of 5hmC losing regions.
(D) Normalized enrichment score of 5hmC regions in different genomic elements. Enrichment scores of 5hmC-gain and -loss regions in tumor samples versus control samples were calculated. All 5hmC enriched regions were used as control.
(E) Top transcription factor binding motifs detected at DhMRs (left: 5hmC gain; right: 5hmC loss). Motif information was obtained from the Homer motif database.



Supplementary information, Figure S3. PCA and Classification based on clustering and metastasis, respectively.

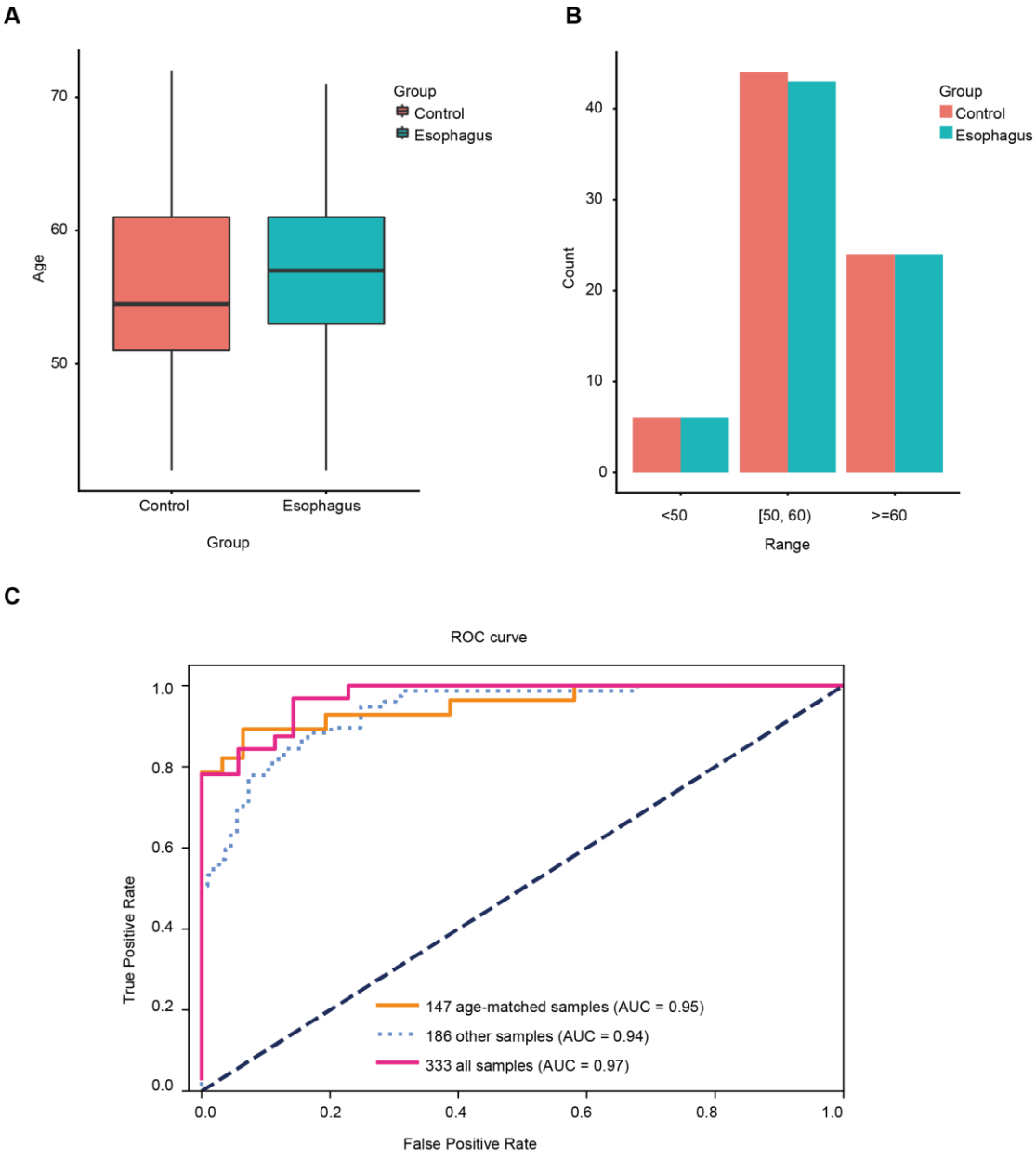
(A) PCA plot of normalized 5hmC reads from control and esophagus cancer samples. Four cluster categories were obtained from (**Figure 1D**).

(B) PCA plot of normalized 5hmC reads for control and cancer samples from different ages.
(C-D) Hierarchical clustering **(C)** and PCA plot **(D)** of 183 control and 150 esophagus cancer samples based on top 400 variance genes.
(E) The normalized 5hmC values of SVOPL and GNA14 in control and esophagus cancer samples.
(F) The predicted cancer probability for LNMC and non- LNMC patients.



Supplementary information, Figure S4. Venn plot and GO analysis of differentially regulated 5hmC genes in three cancers.

(A) Venn plot of differential 5hmC genes in esophageal, colorectal and gastric cancers. **(B-D)** GO enrichment analysis of uniquely differential 5hmC genes in esophagus cancer **(B)**, colorectal cancer **(C)** and gastric cancer **(D)**.



Supplementary information, Figure S5. Classifier with Age-matched samples
(A-B) The age distribution of Age-matched samples **(A)** and sample numbers in three age range **(B)**.
(C) ROC curve for age-matched classifier. The darkorange line indicates the classifier training with 147 age-matched samples. 147 samples were split into 88 training set and 59 validation set (6:4) randomly. The cornflowerblue line indicates the classifier using other 186 samples (of the 333 samples, 147 excepted) to test the model. The deeppink line indicates classifier using all samples.

Supplementary information, Table S1. Clinical characteristics of patients with esophageal cancer and healthy individuals.

	Esophageal Cancer (n=150)	**Healthy Controls (n=177)**
Age (year)		
Mean ± SD	63±8.1	46±11
Median	64	46
Min. to Max.	42-81	25-72
Gender		
Male	98 (65.3%)	117 (66.1%)
Female	52 (34.9)	60 (34.1%)
BMI (kg/m^2)		
Mean ± SD	22±3.1	26±3.5
Median	22	25
Min. to Max.	16-30	17-46
Cancer type		
Squamous carcinoma	137	
Adenocarcinoma	9	
Small Cell Cancer	3	
Neuroendocrine Carcinoma	1	
TNM stages		
High-grade dysplasia	4	
Ⅰ	43	
Ⅱ	56	
Ⅲ	37	
	10	
Depth of invasion		
Tis	4	
T1	44	
T2	30	
T3	64	
T4	8	
Lymphatic metastasis		
Positive	47	
Negative	103	
Distal metastasis		
Positive	2	
Negative	148	
Histologic grade		
G1	10	
G2	102	
G3	30	
Not available	11	
CEA		
Positive	5	0
Negative	78	145
Not available	67	32
AFP		
Positive	0	0

Negative	63	140
Not available	87	37
CA72-4		
Positive	7	0
Negative	70	3
Not available	73	174
CA19-9		
Positive	2	0
Negative	81	92
Not available	67	85
CA15-3		
Positive	0	0
Negative	20	69
Not available	130	108
CK19		
Positive	4	0
Negative	20	0
Not available	126	177
CA125		
Positive	0	2
Negative	82	114
Not available	68	61

Supplementary References

1 Rice TW, Ishwaran H, Ferguson MK *et al.* Cancer of the Esophagus and Esophagogastric Junction: An Eighth Edition Staging Primer. *Journal of Thoracic Oncology* 2017; **12**:36.

2 Han D, Lu X, Shih AH *et al.* A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations. *Molecular Cell* 2016; **63**:711.

3 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**:2114.

4 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 2012; **9**:357.

5 Feng J, Liu T, Qin B *et al.* Identifying ChIP-seq enrichment using MACS. *Nature Protocols* 2012; **7**:1728-1740.

6 Heinz S, Benner C, Spann N *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell* 2010; **38**:576.

7 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**:841-842.

8 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014; **15**:550.

9 Tripathi S, Pohl MO, Zhou Y *et al.* Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. *Cell Host & Microbe* 2015; **18**:723.

10 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016:785-794.

11 Wang Y, Song F, Zhu J *et al.* GSA: genome sequence archive. *Genomics, proteomics & bioinformatics* 2017; **15**:14-18.
12 Members BDC. The BIG Data Center: from deposition to integration to translation. *Nucleic acids research* 2017; **45**:D18.