Supplementary information

Large language models completely understand squamous cervical cancer

Weizhi Zhang[1,2,3,*], Funian Lu[1,*], Tianyu Qin[1,*], Yujie Gou[2,*], Ensong Guo[1], Di Peng[2], Li Zhang[1], Bin Yang[1], Si Liu[1], Cheng Han[2], Shanshan Fu[2], Kun Song[4,5,6], Bairong Xia[7], Dongling Zou[8], Yuanming Shen[9], He Huang[10], Shengtao Zhou[11], Cunzhong Yuan[4,5,6], Yao Shu[4,5,6], Yanan Pi[7], Shuxiang Wang[7], Wenjuan Chen[8], Haixia Wang[8], Lin Zhong[8], Li Yuan[8], Baogang Wen[8], Siqi Yang[9], Ting Wan[10], Junpeng Fan[1], Yu Fu[1], Dan Liu[2], Rourou Xiao[1], Chi Zhang[2], Yuxiang Wei[2], Wenju Peng[1], Xinhe Huang[2], Beibei Wang[1], Peng Wu[1], Beihua Kong[4,5,6], Gordon B. Mills[12], Ding Ma[1,13#], Gang Chen[1,#], Yu Xue[2,3,#], Chaoyang Sun[1,#]

[1]National Clinical Research Center for Gynecology and Obstetrics and Cancer Biology Research Center, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

[2]Department of Bioinformatics and Systems Biology, MOE Key Laboratory of Molecular Biophysics, Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

[3]Nanjing University Institute of Artificial Intelligence Biomedicine, Nanjing, 210031, China

[4]Department of Obstetrics and Gynecology, Qilu Hospital of Shandong University, Jinan 250000, China

[5]Gynecology Oncology Key Laboratory, Qilu Hospital of Shandong University, Jinan, 250000 China

[6]Division of Gynecology Oncology, Qilu Hospital of Shandong University, Jinan, 250000 China

[7]Department of Gynecology, The First Affiliated Hospital of USTC, Division of

Life Sciences and Medicine, University of Science and Technology of China, Hefei 230031, China

[8]Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, Chongqing University Cancer Hospital, Chongqing 400030, China

[9]Department of Gynecologic Oncology, Women's Hospital, Zhejiang University School of Medicine, Hangzhou 310000, China

[10]Department of Gynecologic Oncology, Sun Yat-sen University cancer center, 651 Dongfeng East Road, Guangzhou 510060, China

[11]Department of Obstetrics and Gynecology, Key Laboratory of Birth Defects and Related Diseases of Women and Children of MOE and State Key Laboratory of Biotherapy, West China Second Hospital, Sichuan University and Collaborative Innovation Center, Chengdu, 610000, China

[12]Division of Oncological Sciences, Knight Cancer Institute, Oregon Health and Sciences University, Portland, OR 97201, USA

[13]Lead contact

*These authors contributed equally

#These authors contributed equally

*Correspondence: dingma424@126.com (D.M.), tjchengang@hust.edu.cn (G.C.), xueyu@hust.edu.cn (Y.X.), suncydoctor@gmail.com (C.S.)

Supplementary results

Multi-omic profiling of Chinese SCC samples

For multi-omic profiling, 114 pairs of SCC tumors and NATs were taken from patients with detailed clinical data (Table S2). In our cohort, samples of patients without unambiguous clinical information were not considered for further multi-omic analysis. In total, we obtained a total data volume of 7.81 TB (~12.4 GB per WSI, size ranged from 36,352 × 106,240 pixels to 44,800 × 139,264 pixels) (Figure S1A, Table S1). To identify molecular alterations of SCC, we conducted comprehensive multi-omic profiling, including proteomics (*N* = 111), phosphoproteomics (*N* = 111), exomics (*N* = 89), and transcriptomics (*N* = 80), in paired tumors and NATs in 114 cases where tumors passed quality control (QC) (Methods, Table S3).

To quantify proteomic and phosphoproteomic alterations in SCC, the 222 tumors and paired NATs were categorized into 23 batches and separately subjected to tandem mass tag (TMT) labeling after cell lysis and peptide digestion. For each batch, individual samples were labeled with the TMT 11-plex reagent. Because the samples were not simultaneously obtained, we selected the cervical cancer cell line, HeLa, as an internal reference for all the batches to eliminate the batch effect of quantification. After fractionation, each batch of peptide or phosphopeptide mixtures was analyzed by liquid chromatography with tandem mass spectrometry (LC-MS/MS), respectively. From the proteomic profiling, we identified 10,032 proteins, including 6,749 proteins quantified in > 50% of all samples (Figure S1B, Table S3). Using the Wilcoxon rank-sum test, we identified 1,394 differentially expressed proteins (DEPs) with |$\log_2$(fold change [FC])| > 0.5 in tumors against NATs (Adjusted p < 0.01, Figure S1B, Table S4). From the phosphoproteomic profiling, we identified 46,713 phosphorylation sites (p-sites), including 36,891 phospho-serine (pS), 8,472 phospho-threonine (pT) and 1,350 phospho-tyrosine (pY) residues (Figure S1C, Table S4). In our phosphoproteomic data, there were 17,670 p-sites simultaneously quantified in > 50% of all samples (Figure S1C, Table S3). Again, we used Wilcoxon signed-rank test, and identified 4,598 differentially regulated p-sites (DRPs) with $|\log_2(FC)| > 0.5$ in tumors against NATs (Adjusted $p < 0.01$, Table S4). Furthermore, the p-sites from the same phosphoproteins were collapsed by calculating the median ratio, and 1,191 differentially phosphorylated proteins (DPPs) were identified by two-sided Wilcoxon signed-rank test with statistical significance (Adjusted $p < 0.01$ and $|\log2(FC)| > 0.5$).

To identify genomic and transcriptomic alterations in encoding sequences (CDSs) of SCC, both whole-exome sequencing (WES) and mRNA sequencing (RNA-seq) were conducted. WES was used for sequencing 89 pairs of SCC tumors and NATs, having a sequencing depth ranging from 133X to 252X, with an average depth of 173X in tumor or NAT samples (Figure S1D). The corresponding transcriptomes of 80 pairs of samples were then profiled by RNA-seq, obtaining clean reads ranging from 20.6 M to 68.3 M, with an average number of 38.8 M reads in tumor or NAT samples (Figure S1E). For each gene, the RNA-Seq by expectation maximization (RSEM) count was calculated as its expression level (Table S3). In total, we obtained 10,217 genes mutually quantified in all samples, including 2,424 differentially expressed genes (DEGs) with $|\log_2(FC)| > 1$ in SCC tumors, using the Wilcoxon signed-rank test (Adjusted $p < 0.01$, Figure S1F, Table S3, S4).

Using 3,837 proteins, 5,299 p-sites, and 10,217 genes mutually quantified in all samples, the principal component analysis (PCA) was performed for proteomic (Figure S1G), phosphoproteomic (Figure S1H), and transcriptomic data (Figure S1I). The results indicated that tumors and NATs could be clearly distinguished without any batch effect. An analysis of the abundances of the 3,837 proteins in tumors or NATs found a single peak, indicating protein degradation did not occur during sample preparation (Figure S1J).

Using 491 DEPs, 1,892 DRPs, and 1,088 DEGs significantly up-regulated in SCC tumors against NATs (Table S4), we performed the enrichment analysis of Gene Ontology (GO) biological processes ($p < 0.01$, Figure S1K). It could be found that different processes were preferentially regulated at different types of omic levels. One biological process, defense response to virus, was simultaneously regulated at proteomic and transcriptomic levels. Two processes, DNA replication and mRNA splicing via spliceosome were mutually regulated at both proteomic and phosphoproteomic levels. No process was simultaneously regulated at all three omic levels.

Additional results on molecular subtyping

Here, we applied 16 CFs (Table S2), all of which were tightly associated with SCC outcomes, to calculate the correlations between molecular subtypes and clinical relevance by Chi-square ($\chi 2$) tests. Using the k-means clustering method, we first employed 1,394 DEPs, 1,198 DPPs, and 2,424 DEGs altered in SCC tumors to conduct molecular subtyping for each omic layer (Figure S2A, B). However, none of the resultant omic subtypes were significantly associated with the 16 CF types (Figure S2C, Table S5).

Then we hypothesized that CFs and molecular subtypes that are able to predict outcomes should be consistent with each other, and such a consistency could be identified through optimization of consistency. We adopted consistant learning in subtyping module. After the embedding,The CF- and omics-embedded vectors were separately used for k-means clustering, and the statistical association between two different subtypes was tested. The weight value of each CF type was randomly changed, and one or more molecules in each omic data were randomly dropped. The iterative manipulation was terminated when the significance of the association between clinical and omic subtypes did not increase further (Figure 3A). The decoder module was used to ensure the fidelity of encoder-based embedding.

Functional enrichment analyses of GO biological processes or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [1,2] were also conducted for different subtypes clustered from phosphoproteomic, or transcriptomic data, respectively (Figure S2E-F). The results showed that phosphoproteomic subtypes were consistent with proteomic subtypes with a number of inflammation/complement-related processes being enriched in the phosphoproteomic subtype S-IIIp. We also found out that the image features of keratinization have a consistent trend with the classification (Figure S2G).

Additional results on potential biomarker proteins in SCC

To identify potential tumor- and subtype-specific biomarkers, we developed deepCMD marker module to select optimal biomarker combinations. We identified marker combinations able to classify tumor and NATs and importantly to identify the S-III inflammation/complement subtype, which is associated with clinical characteristics of worsened outcomes, with high fidelity.

There may be additional prognostic biomarkers and therapeutic options to explore in S-III tumors. Immune complexes such as elevated IGHG2, IGHG4 may initiate complement activation in S-III. In addition, upregulated members of coagulation (F2/prothrombin and KNG1) and fibrinolytic (PLG, PLGLA, and PLGLB1) system that we observed in S-III may also contribute to [3]. HRG, which enhances complement activation by interacting with immunoglobulins, was increased in S-III tumors [4]. Furthermore, PLG, HRG, SEPTIN5, and SERPINA4 are coordinately upregulated in S-III, supporting these as targets and potentially predictive biomarkers. Also, reverse phase protein arrays (RPPA) have been used to measure the expression or phosphorylation levels of hundreds of important proteins in SCC [5]. We obtained RPPA data of 171 cervical cancer tumors from The Cancer Proteome Atlas (TCPA) [6], in which the expression levels of 13 proteins were significantly correlated with OS (log-rank $p < 0.05$). Importantly 3 of the proteins associated with OS in the RPPA data including SCD, PARP1, and YAP1 were also identified as DEPs up-regulated in our SCC tumors (Figure S3I).

SCSP profile in cervical cancer

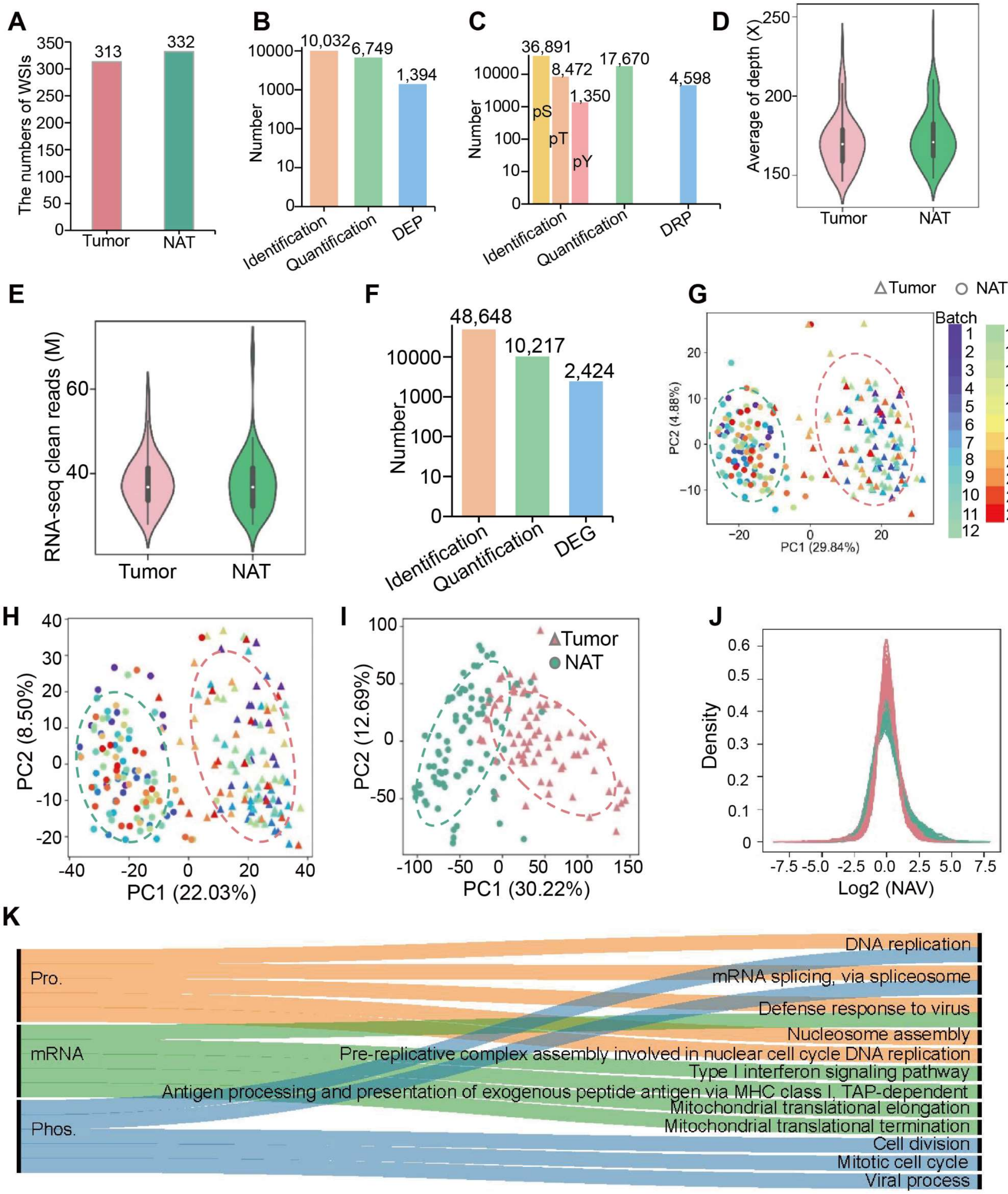
In total, we identified 3,541 unique proteins from at least one region, and up to 2,194 proteins were quantified in > 50% of all regions. Most proteins were identified with more than 2 peptides. (Figure S6D, E). And to improve our understanding of the single-cell proteomic heterogeneity of SCC, we estimated the proportions of the 29 deepCMD single-cell states from the corresponding WSI, and then the eigen matrix of expression levels of 3,541 proteins in the 29 states was determined by non-negative matrix factorization (NMF) from the SCSP data. Using such a matrix, NMF determined the proportions of the 29 single-cell states in each tumor sample based on bulk proteomic data. Subsequently, 6,694 expression levels of the protein not detected in SCSP profiling were then determined by NMF for each of the 29 single-cell states. We assumed that the known markers of immune cells should be enriched in corresponding immune cell subtypes. Based on this hypothesis, we iteratively optimized our immune cell subtypes until the enrichment of known immune cell markers did not increase.

We deconvoluted the spatial proteomes of the 29 single-cell states. For 29 single-cell states, Spearman's correlation between two proteomes calculated from spatial proteome and bulk proteomic data was analyzed, showing an overall significantly positive correlation in 29 single-cell states (Adjusted $p < 0.01$, Figure S6F). The consistency of results from different levels indicates the high accuracy and reliability of our results.

Supplementary References

1. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).
2. The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* **47**, D330-D338 (2019).
3. Morgan, B.P. & Harris, C.L. Complement, a target for therapy in inflammatory and degenerative diseases. *Nat Rev Drug Discov* **14**, 857-77 (2015).
4. Manderson, G.A. *et al.* Interactions of histidine-rich glycoprotein with immunoglobulins and proteins of the complement system. *Mol Immunol* **46**, 3388-98 (2009).
5. Cancer Genome Atlas Research, N. *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378-384 (2017).
6. Chen, M.M. *et al.* TCPA v3.0: An Integrative Platform to Explore the Pan-Cancer Analysis of Functional Proteomic Data. *Mol Cell Proteomics* **18**, S15-S25 (2019).
7. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).

Figure S1



A
The numbers of WSIs
350
300
250
200
150
100
50
0
313
332
Tumor
NAT
B
Number
10000
1000
100
10
0
10,032
6,749
1,394
Identification
Quantification
DEP
C
Number
10000
1000
100
10
0
36,891
8,472
1,350
17,670
4,598
pS
pT
pY
Identification
Quantification
DRP
D
Average of depth (X)
250
200
150
Tumor
NAT
E
RNA-seq clean reads (M)
60
40
Tumor
NAT
F
Number
10000
1000
100
10
0
48,648
10,217
2,424
Identification
Quantification
DEG
G
△Tumor ○ NAT
Batch
1 2 3 4 5 6 7 8 9 10 11 12
13 14 15 16 17 18 19 20 21 22 23
PC2 (4.88%)
20
10
0
−10
−20
0
20
PC1 (29.84%)
H
PC2 (8.50%)
40
30
20
10
0
-10
-20
-40
-20
0
20
40
PC1 (22.03%)
I
Tumor
NAT
PC2 (12.69%)
100
50
0
-50
-100
-50
0
50
100
150
PC1 (30.22%)
J
Density
0.6
0.5
0.4
0.3
0.2
0.1
0
-7.5 -5.0 -2.5 0.0 2.5 5.0 7.5
Log2 (NAV)
K
Pro.
mRNA
Phos.
DNA replication
mRNA splicing, via spliceosome
Defense response to virus
Nucleosome assembly
Pre-replicative complex assembly involved in nuclear cell cycle DNA replication
Type I interferon signaling pathway
Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-dependent
Mitochondrial translational elongation
Mitochondrial translational termination
Cell division
Mitotic cell cycle
Viral process

Supplementary Figure Legends

Figure S1. Multi-omic profiling of tumors and NATs from SCC patients

(A) The number of WSIs in tumors and NATs. **(B)** The numbers of identified proteins in at least one sample, mutually quantified proteins in > 50% of all samples, and DEPs with $|\log_2(FC)| > 0.5$ in tumors *vs*. NATs (adjusted p < 0.01). **(C)** The numbers of identified pS, pT, and pY residues in at least one sample, mutually quantified p-sites in > 50% of all samples, and DRPs with $|\log_2(FC)| > 0.5$ in tumors *vs*. NATs (adjusted $p < 0.01$). **(D)** The sequencing depths of WES in tumors and NATs. **(E)** The distribution of numbers of QC-passed reads in tumors and NATs from RNA-seq profiling. **(F)** The numbers of identified mRNAs in at least one sample, mutually quantified mRNAs in all samples, and DEGs with $|\log_2(FC)| > 1$ in tumors *vs*. NATs (adjusted p < 0.01). **(G-I)** PCA results of (G) proteomic data, (H) phosphoproteomic data, and (I) transcriptomic data, respectively. **(J)** The distribution of normalized abundance values (NAVs) of proteins in tumors and NATs. **(K)** The sankey plot of enriched GO biological processes of DEPs, DRPs, and DEGs significantly up-regulated in SCC tumors against NATs.

Figure S2

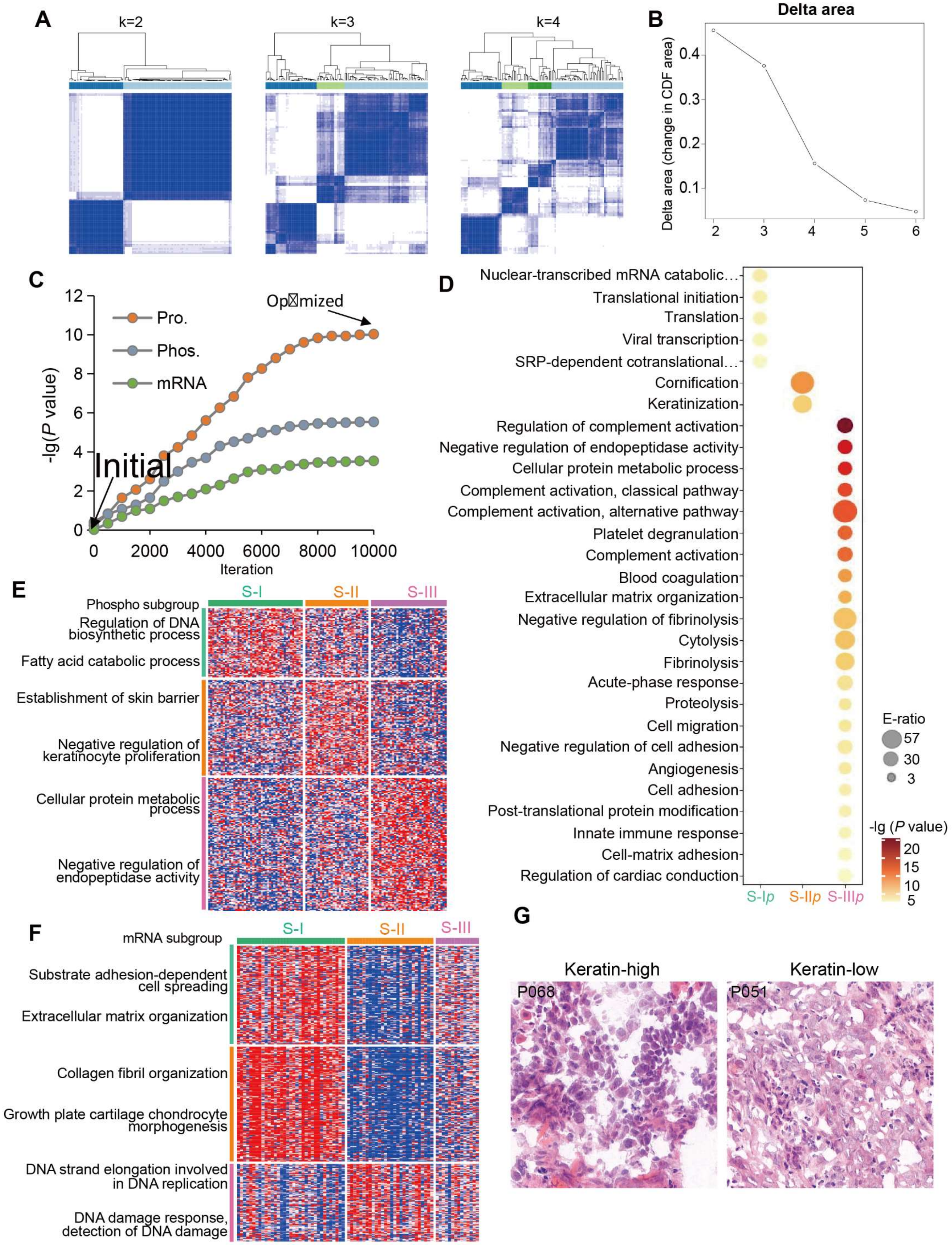

A
k=2
k=3
k=4
B
Delta area
Delta area (change in CDF area)
0.4
0.3
0.2
0.1
2
3
4
5
6
C
-lg(P value)
12
10
8
6
4
2
0
Pro.
Phos.
mRNA
Op⊠mized
Initial
0
2000
4000
6000
8000
10000
Iteration
D
Nuclear-transcribed mRNA catabolic…
Translational initiation
Translation
Viral transcription
SRP-dependent cotranslational…
Cornification
Keratinization
Regulation of complement activation
Negative regulation of endopeptidase activity
Cellular protein metabolic process
Complement activation, classical pathway
Complement activation, alternative pathway
Platelet degranulation
Complement activation
Blood coagulation
Extracellular matrix organization
Negative regulation of fibrinolysis
Cytolysis
Fibrinolysis
Acute-phase response
Proteolysis
Cell migration
Negative regulation of cell adhesion
Angiogenesis
Cell adhesion
Post-translational protein modification
Innate immune response
Cell-matrix adhesion
Regulation of cardiac conduction
S-Ip
S-IIp
S-IIIp
E-ratio
57
30
3
-lg (P value)
20
15
10
5
E
Phospho subgroup
S-I
S-II
S-III
Regulation of DNA biosynthetic process
Fatty acid catabolic process
Establishment of skin barrier
Negative regulation of keratinocyte proliferation
Cellular protein metabolic process
Negative regulation of endopeptidase activity
F
mRNA subgroup
S-I
S-II
S-III
Substrate adhesion-dependent cell spreading
Extracellular matrix organization
Collagen fibril organization
Growth plate cartilage chondrocyte morphogenesis
DNA strand elongation involved in DNA replication
DNA damage response, detection of DNA damage
G
Keratin-high
Keratin-low
P068
P051

Figure S2. Integrated molecular subtyping in SCC

(A) The schematic diagram of different solutions for the *k*-means clustering (*k* = 2, 3, or 4) of patients based on proteomic data. **(B)** The delta plot of the relative change in the area under the cumulative distribution function (CDF) curve of the *k*-means clustering with different k values. **(C)** The statistical significance values of the conventional clustering (initial) and cLCM-based clustering (optimized), for proteomic, phosphoproteomic and transcriptomic data. **(D)** Enriched GO biological processes associated with proteomic subtypes. (Hypergeometric test, $p < 0.05$). **(E,F)** Enriched GO biological processes associated with (E) phosphoproteome-, and (F) transcriptome-based subtypes, respectively (Hypergeometric test, $p < 0.05$). **(G)** The representative tiles from S-I and other types.

Figure S3

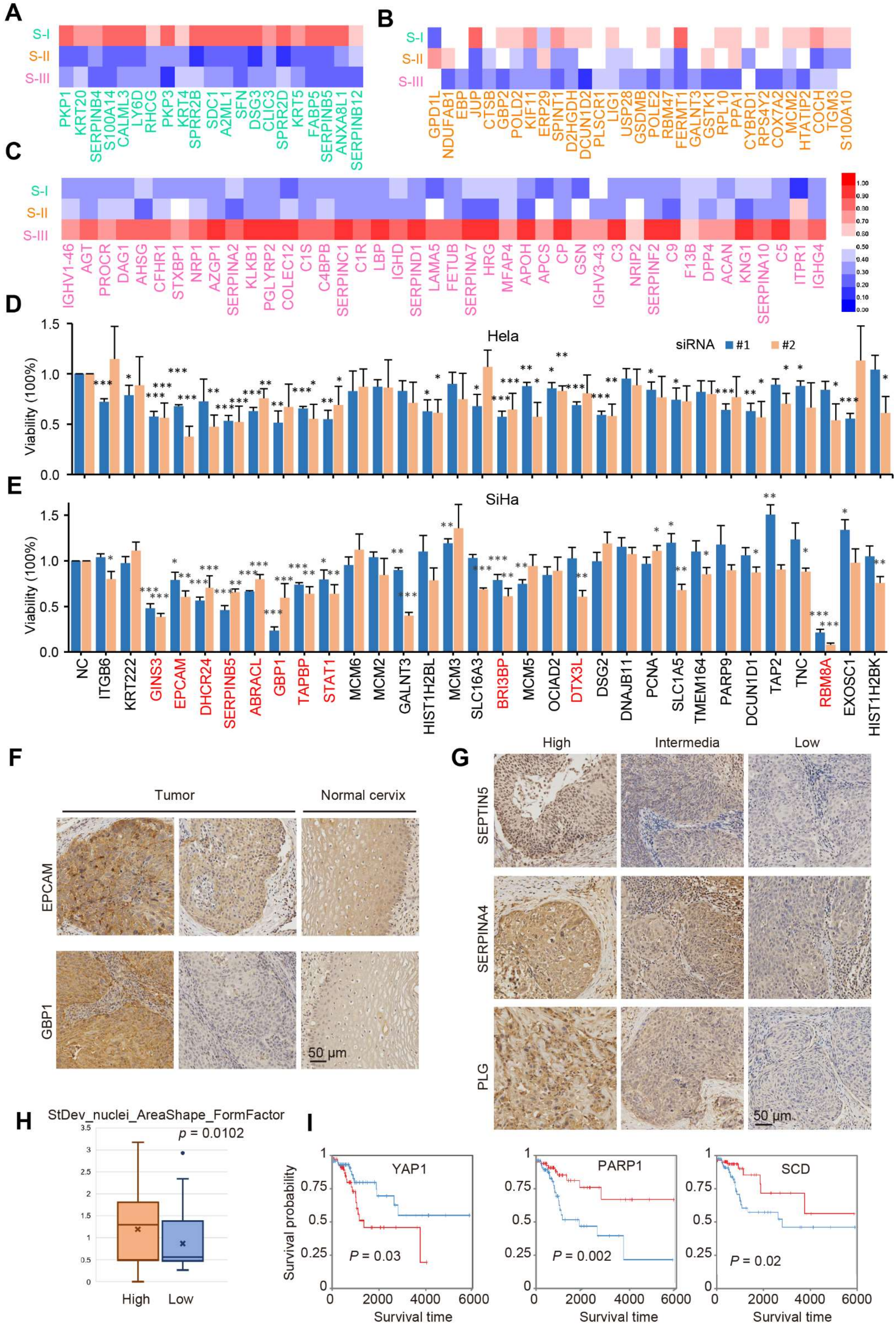
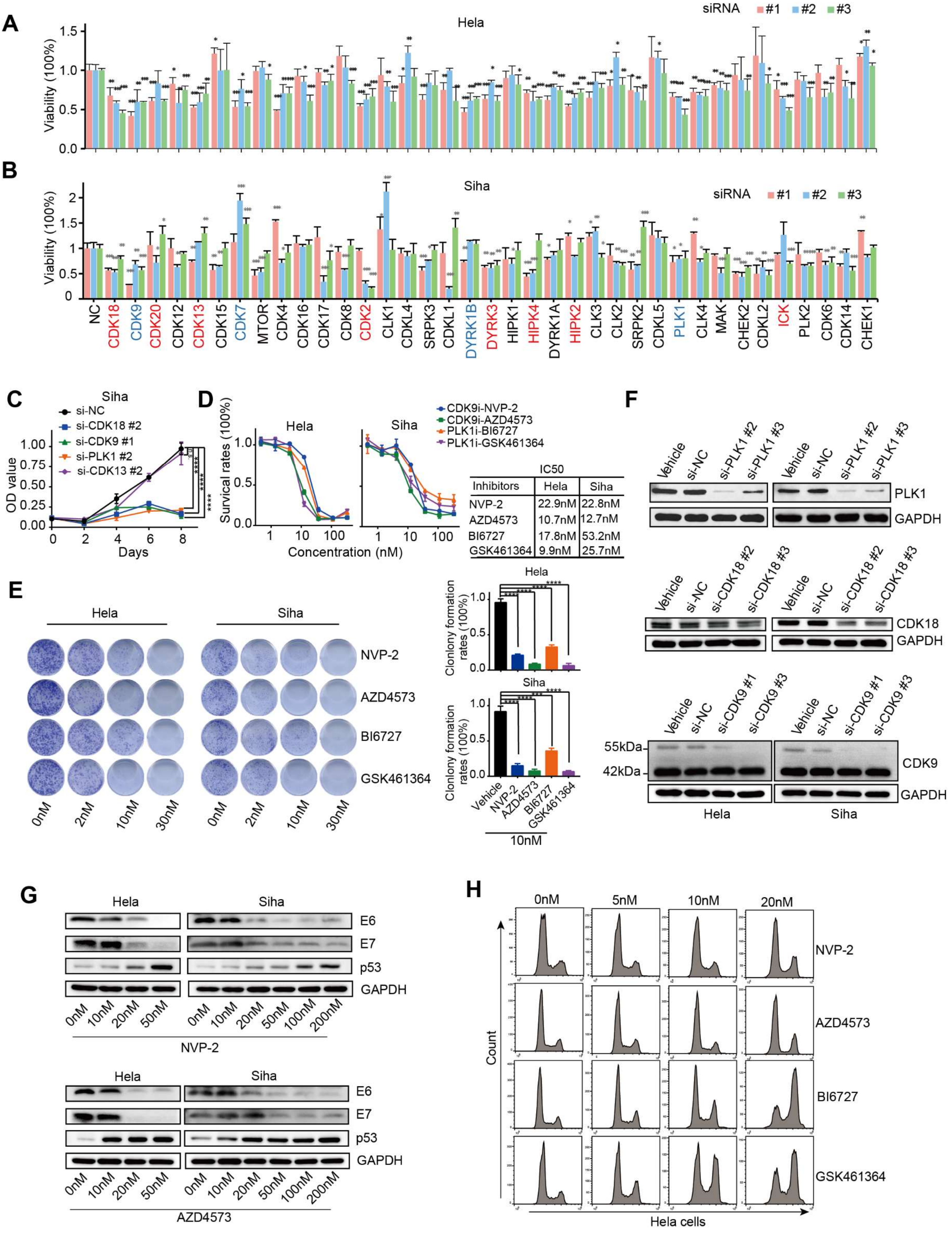

A
S-I
S-II
S-III
PKP1 KRT20 SERPINB4 S100A14 CALML3 LY6D RHCG PKP3 KRT4 SPRR2B SDC1 A2ML1 SFN DSG3 CLIC3 SPRR2D KRT5 FABP5 SERPINB5 ANXA8L1 SERPINB12
B
S-I
S-II
S-III
GPD1L NDUFAB1 EBP JUP CTSB GBP2 POLD2 KIF11 ERP29 SPINT1 D2HGDH DCUN1D2 PLSCR1 LIG1 USP28 GSDMB POLE2 RBM47 FERMT1 GALNT3 GSTK1 RPL10 PPA1 CYBRD1 RPS4Y2 COX7A2 MCM2 HTATIP2 COCH TGM3 S100A10
C
S-I
S-II
S-III
IGHV1-46 AGT PROCR DAG1 AHSG CFHR1 STXBP1 NRP1 AZGP1 SERPINA2 KLKB1 PGLYRP2 COLEC12 C1S C4BPB SERPINC1 C1R LBP IGHD SERPIND1 LAMA5 FETUB SERPINA7 HRG MFAP4 APOH APCS CP GSN IGHV3-43 C3 NRIP2 SERPINF2 C9 F13B DPP4 ACAN KNG1 SERPINA10 C5 ITPR1 IGHG4
1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
D
Hela
siRNA #1 #2
Viability (100%)
1.5 1.0 0.5 0.0
E
SiHa
Viability (100%)
1.5 1.0 0.5 0.0
NC ITGB6 KRT222 GINS3 EPCAM DHCR24 SERPINB5 ABRACL GBP1 TAPBP STAT1 MCM6 MCM2 GALNT3 HIST1H2BL MCM3 SLC16A3 BRI3BP MCM5 OCIAD2 DTX3L DSG2 DNAJB11 PCNA SLC1A5 TMEM164 PARP9 DCUN1D1 TAP2 TNC RBM8A EXOSC1 HIST1H2BK
F
Tumor
Normal cervix
EPCAM
GBP1
50 μm
G
High
Intermedia
Low
SEPTIN5
SERPINA4
PLG
50 μm
H
StDev_nuclei_AreaShape_FormFactor
p = 0.0102
3.5 3 2.5 2 1.5 1 0.5 0
High
Low
I
YAP1
P = 0.03
PARP1
P = 0.002
SCD
P = 0.02
Survival probability
1 0.75 0.5 0.25 0
0 2000 4000 6000
Survival time

Figure S3. Additional experiments to validate potential biomarkers specific for SCC tumors or proteomic subtype S-III

(A-C) Individual candidate biomarker proteins in the top 10 combinations specific for proteomic subtypes (A) S-I, (B) S-II, and (C) S-III. **(D,E)** The histograms showing the siRNA library screening against 33 DEPs in HeLa and SiHa cells. HeLa (3500 per well) and SiHa (3500 per well) are harvested 3d post transfection and cell viability was determined by CCK8 assay. **(F)** The representative images of immunohistochemistry analysis showing different expression level of EPCAM and GBP1 protein in SCC tumor tissue and normal cervix. Scale bar, 50μm. **(G)** The representative images of immunohistochemistry analysis showing SEPTINS, SERPINA4 and PLG protein expression heterogeneity among SCC tumor samples. Scale bar, 50 μm. **(H)** The statistics of image feature “StDev FormFactor” values between HRG high expressed and low expressed samples. (Two-sided t-test, *p* = 0.0124) **(I)** The prognostic powers of 3 DEPs identified in this study, including SCD, PARP1, and YAP1. The data of their correlation between RPPA-based expression levels to OS outcomes were directly taken from TCPA [6].

Figure S4



IC50		
Inhibitors	Hela	Siha
NVP-2	22.9nM	22.8nM
AZD4573	10.7nM	12.7nM
BI6727	17.8nM	53.2nM
GSK461364	9.9nM	25.7nM

Figure S4. Additional experiments to validate potential actionable PKs

(A,B) The histograms showing the siRNA library screening against 37 potential PKs in HeLa and SiHa cells. HeLa (3500 per well) and SiHa (3500 per well) are harvested 3 d post transfection and cell viability was determined by CCK8 assay. **(C)** The representative results of cell proliferation assay in SiHa cells while CDK18, CDK9 and PLK1 expression were diminished by siRNA. Each sample was assayed in quintuplicates in 96-well plate. Cell viability was measured with CCK8 kit at OD = 450 nm every 48 h consecutively. **(D)** The drug response curves of CDK9 and PLK-1 inhibitors in HeLa (left) and SiHa (right) cells. IC50 of each drug in two cell lines was calculated with Graphpad Prism software. **(E)** The representative images (left panel) and quantitative analysis at a concentration of 10nM (right panel) of colony formation assays of CDK9 and PLK-1 inhibitors treated HeLa and SiHa cells. **(F)**The western blot validation of the effect of PLK1 (top), CDK18 (middle) and CDK9 (bottom) siRNA interference in HeLa and SiHa cells. **(G)** The western blot analysis of HPV E6, E7 oncoprotein and p53 protein levels in HeLa and SiHa cells treated with increasing dose of two CDK9 inhibitor NVP-2 (above) AZD4573 (bottom). **(H)** The cell cycle analysis of HeLa cells treated with DMSO, NVP-2, AZD4573, BI6726 and GSK461346 at 5 nM, 10 nM and 20 nM respectively for 48 h. PLK-1 inhibitors mainly induced cell cycle arrest at G2/M phase while CDK9 inhibitors resulted a clear G1 phase arrest.

Figure S5

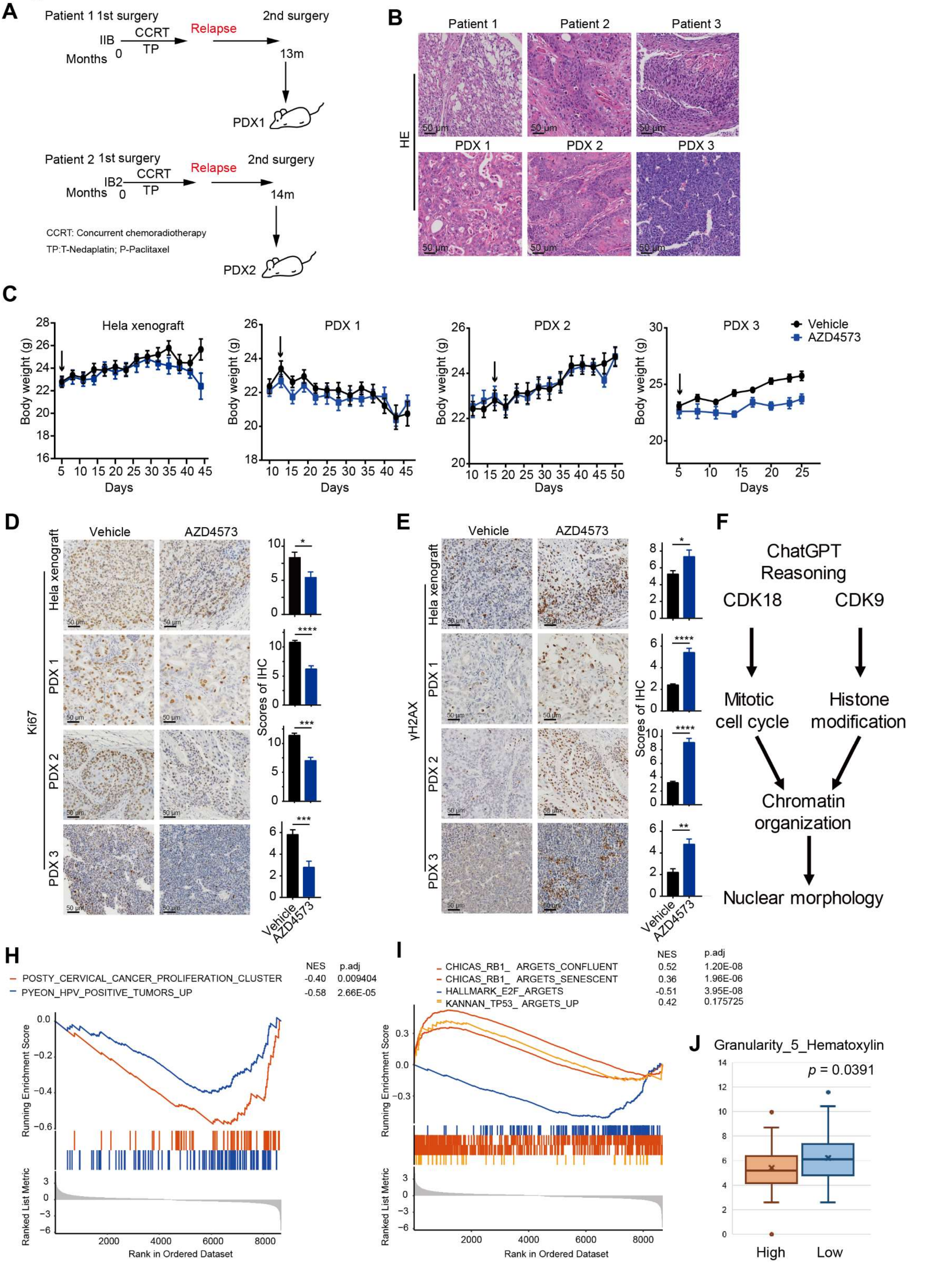
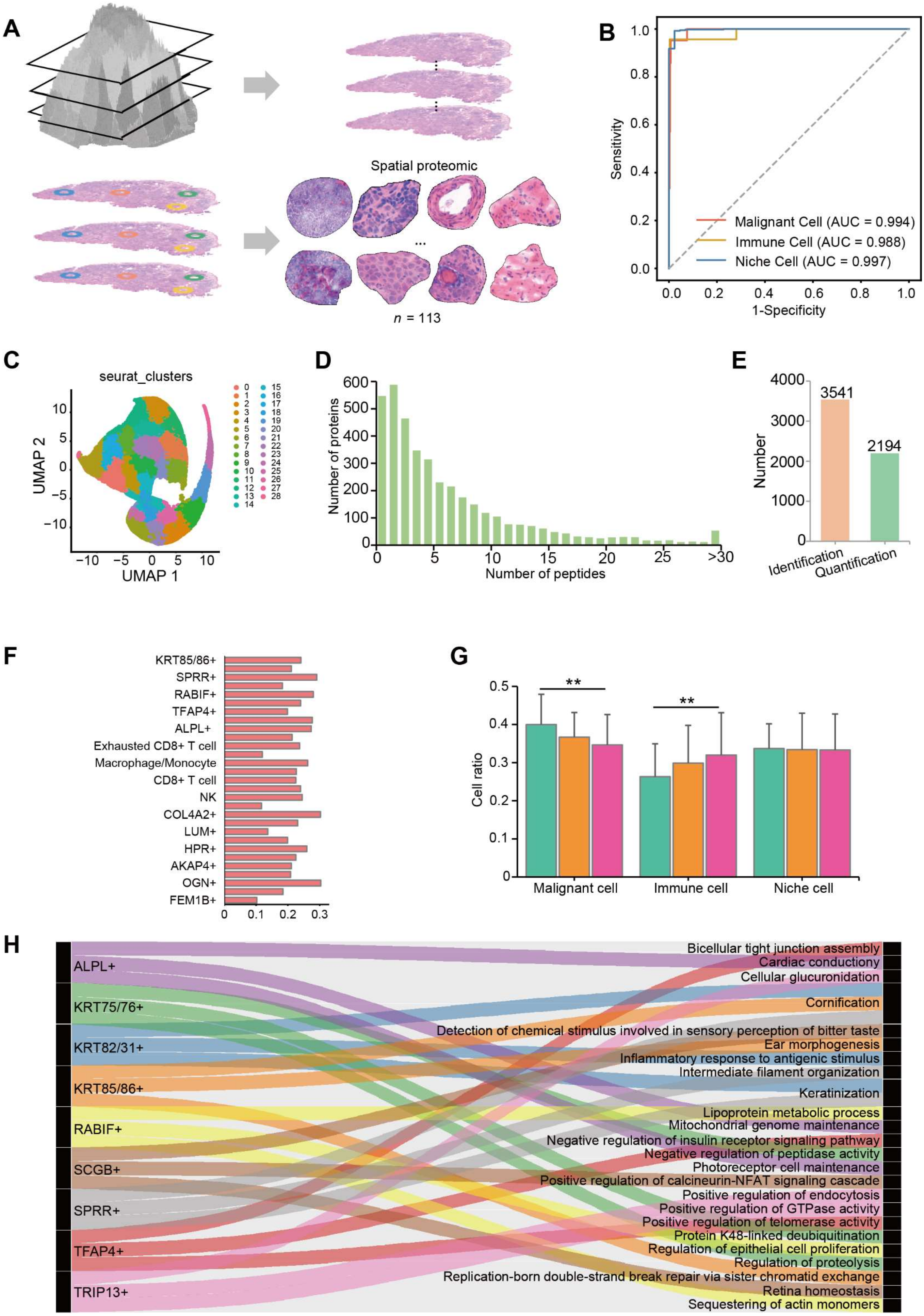

A
Patient 1 1st surgery
IIB
CCRT
TP
Relapse
2nd surgery
Months 0
13m
PDX1
Patient 2 1st surgery
IB2
CCRT
TP
Relapse
2nd surgery
Months 0
14m
PDX2
CCRT: Concurrent chemoradiotherapy
TP:T-Nedaplatin; P-Paclitaxel
B
Patient 1
Patient 2
Patient 3
HE
PDX 1
PDX 2
PDX 3
50 μm
C
Hela xenograft
PDX 1
PDX 2
PDX 3
Body weight (g)
Days
Vehicle
AZD4573
D
Vehicle
AZD4573
Hela xenograft
PDX 1
PDX 2
PDX 3
Ki67
Scores of IHC
*
****
***
***
Vehicle
AZD4573
E
Vehicle
AZD4573
Hela xenograft
PDX 1
PDX 2
PDX 3
γH2AX
Scores of IHC
*
****
****
**
Vehicle
AZD4573
F
ChatGPT Reasoning
CDK18
CDK9
Mitotic cell cycle
Histone modification
Chromatin organization
Nuclear morphology
H
NES p.adj
POSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER -0.40 0.009404
PYEON_HPV_POSITIVE_TUMORS_UP -0.58 2.66E-05
Running Enrichment Score
Ranked List Metric
Rank in Ordered Dataset
I
NES p.adj
CHICAS_RB1_ ARGETS_CONFLUENT 0.52 1.20E-08
CHICAS_RB1_ ARGETS_SENESCENT 0.36 1.96E-06
HALLMARK_E2F_ARGETS -0.51 3.95E-08
KANNAN_TP53_ ARGETS_UP 0.42 0.175725
Running Enrichment Score
Ranked List Metric
Rank in Ordered Dataset
J
Granularity_5_Hematoxylin
p = 0.0391
High
Low

Figure S5. Additional results for CDK18 and CDK9

(A) The schematic illustration of the PDX1-2 establishments from 2 recurrent cervical cancer patients. **(B)** The representative images showing the concordance of the histological morphology between SCC primary tumor and the corresponding PDX. Scale bar, 50 μm. **(C)** Line charts showing the body weight changes of nude mice in each treatment group during drug administration (n = 5 or 6). Black arrows indicate the starting point of drug administration. **(D,E)** The representative images and quantification of immunohistochemistry analysis of Ki67 (left panel) and γH2AX (right panel) in HeLa xenograft and three PDXs after indicated treatment (n = 5 or 6). Scale bar, 50 μm. An unpaired t test was performed for statistical analysis. **(F)** The CoT of reasoning the relationship between the expression of CDK18 and CDK9 with image feature. **(H,I)** GSEA plots of the indicated pathway when compared RNA-seq data of CDK18 knock-out cells to parental HeLa cells. **p* < 0.05, ***p* < 0.01, ****p* < 0.001. **(J)** The statistics of image feature “Granularity” values and CDK18/CDK9 kinase activities (Two-sided t-test, *p* = 0.0391).

Figure S6



A
Spatial proteomic
...
n = 113
B
1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
0.0 0.2 0.4 0.6 0.8 1.0
1-Specificity
Malignant Cell (AUC = 0.994)
Immune Cell (AUC = 0.988)
Niche Cell (AUC = 0.997)
C
seurat_clusters
UMAP 2
UMAP 1
10
5
0
−5
−10
−10 −5 0 5 10
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
D
Number of proteins
600
500
400
300
200
100
0
0 5 10 15 20 25 >30
Number of peptides
E
4000
3000
2000
1000
0
Number
3541
2194
Identification
Quantification
F
KRT85/86+
SPRR+
RABIF+
TFAP4+
ALPL+
Exhausted CD8+ T cell
Macrophage/Monocyte
CD8+ T cell
NK
COL4A2+
LUM+
HPR+
AKAP4+
OGN+
FEM1B+
0 0.1 0.2 0.3
G
Cell ratio
0.5
0.4
0.3
0.2
0.1
0
**
**
Malignant cell
Immune cell
Niche cell
H
ALPL+
KRT75/76+
KRT82/31+
KRT85/86+
RABIF+
SCGB+
SPRR+
TFAP4+
TRIP13+
Bicellular tight junction assembly
Cardiac conductiony
Cellular glucuronidation
Cornification
Detection of chemical stimulus involved in sensory perception of bitter taste
Ear morphogenesis
Inflammatory response to antigenic stimulus
Intermediate filament organization
Keratinization
Lipoprotein metabolic process
Mitochondrial genome maintenance
Negative regulation of insulin receptor signaling pathway
Negative regulation of peptidase activity
Photoreceptor cell maintenance
Positive regulation of calcineurin-NFAT signaling cascade
Positive regulation of endocytosis
Positive regulation of GTPase activity
Positive regulation of telomerase activity
Protein K48-linked deubiquitination
Regulation of epithelial cell proliferation
Regulation of proteolysis
Replication-born double-strand break repair via sister chromatid exchange
Retina homeostasis
Sequestering of actin monomers

Figure S6. Additional analyses of the single-cell heterogeneity in SCC.

(A) The basic procedure of sampling 113 heterogeneous regions from whole tumor sample. **(B)** The ROC curve of single-cell recognition on malignant, immune and other cells. **(C)** Seurat-based clustering [7] of single cells recognized by deepCMD-SCSP. **(D)** The distribution of peptide numbers of proteins identified from the spatial proteomic profiling. **(E)** The numbers of identified proteins and proteins quantified in > 50% of all small regions. **(F)** The Spearman's correlation between the two proteomes of 29 single-cell subtypes calculated from spatial proteome and bulk proteomic data. **(G)** The histogram showing the proportion of malignant cells, immune cells and other cells in S-I, S-II, and S-III. **(H)** The enrichment analysis of GO biological processes ($p <$ 0.01) of top 50 state-specific proteins of each malignant cell state.

Supplementary Table Legends

Table S1. The total ChatGPT questioning and answering pairs of interpretation, reasoning and new insights. Table S1A. First question in interpretation module of querying DEPs, DPPs, DEGs, FMGs and single-cell markers. Table S1B. Second question in interpretation module of querying DEPs, DPPs, DEGs, FMGs and single-cell markers. Table S1C. Third question in interpretation module of querying DEPs, DPPs, DEGs, FMGs and single-cell markers. Table S1D First question in reasoning module of querying image characteristics. Table S1E. Second question in interpretation module of querying image characteristic changing. Table S1F. Third question in interpretation module of querying molecular alteration with imaging characteristic changing. Table S1G. First question in new insight module of querying experimental verified moleclues. Table S1H. Second question in new insight module of querying experimental results of molecules. Table S1I. Third question in new insight module of querying reasoning confidence of molecule involvement in cervical cancer. Table S1J. Questions within signal webs.

Table S2. The Clinical Characteristics of 114 SCC Patients.

Table S3. The Multi-omic Profiling and the Spatial Proteomic Profiling, Related to STAR Methods and Figure S2.

Table S3A. The proteomic profiling of 111 paired samples. Table S3B. The phosphoproteomic profiling of 111 paired samples. Table S3C. The WES-based somatic variants. Table S3D. The transcriptomic profiling of 80 paired samples. Table S3E. The spatial proteomic profiling of 113 small regions.

Table S4. The DEPs, DRPs and DPPs in SCC Tumor against NAT Samples, Related to STAR Methods and Figure S1, S2.

Table S4A. The list of 1,394 DEPs. Table S4B. The list of 1,198 DPPs. Table S4C. The list of 4,598 DRPs. Table S4D. The list of 2,424 DEGs. Table S4E. The list of 2,073 omic DICs. Table S4F The list of 288 single-cell DICs.

Table S5. The Correlation between Clinical Subtypes and Molecular Subtypes, the List of Top 10 Tumor- and Subtype-specific Candidate Biomarker Combinations, 37 Actionable Kinases, and the Expression Matrix of 10,235 Proteins in 29 Single-cell states, Related to STAR Methods and Figure 6.
Table S6A. The correlation between clinical subtypes and proteome-, transcriptome- and phosphoproteome-based subtypes. Table S6B. The details of top 10 tumor- and subtype-specific biomarker combinations. Table S6C. The details of 37 predicted actionable kinases. Table S6D. The expression matrix of 10,235 proteins in 29 deconvoluted single-cell subtypes.

Table S6. The siRNAs for Candidate Protein Biomarkers and Candidate PKs, as well as 14 Kinase Inhibitors with Corresponding Doses for 8 PKs, Related to STAR Methods and Figure 5.
Table S7A. The list of siRNAs for candidate protein biomarkers. Table S7B. The list of siRNAs for candidate PKs. Table S7C. The 14 kinase inhibitors for 8 PKs.