www.nature.com/cr
www.cell-research.com

Cell Research

ARTICLE

Whole-genome sequencing of 508 patients identifies key molecular features associated with poor prognosis in esophageal squamous cell carcinoma

Yongping Cui[1], Hongyan Chen[2], Ruibin Xi[3], Heyang Cui[4,5], Yahui Zhao[2], Enwei Xu[1,6], Ting Yan[1], Xiaomei Lu[7], Furong Huang[2], Pengzhou Kong[1], Yang Li[2], Xiaolin Zhu[2], Jiawei Wang[8], Wenjie Zhu[8], Jie Wang[8], Yanchun Ma[1], Yong Zhou[1,5], Shiping Guo[9], Ling Zhang[1,5], Yiqian Liu[1,5], Bin Wang[5], Yanfeng Xi[6], Ruifang Sun[10], Xiao Yu[2], Yuanfang Zhai[1,5], Fang Wang[1], Jian Yang[1], Bin Yang[1,9], Caixia Cheng[1,11], Jing Liu[1], Bin Song[1], Hongyi Li[1], Yi Wang[1,5], Yingchun Zhang[1,5], Xiaolong Cheng[1], Qimin Zhan[4,5], Yanhong Li[12] and Zhihua Liu[2]

Esophageal squamous cell carcinoma (ESCC) is a poor-prognosis cancer type with limited understanding of its molecular etiology. Using 508 ESCC genomes, we identified five novel significantly mutated genes and uncovered mutational signature clusters associated with metastasis and patients' outcomes. Several functional assays implicated that *NFE2L2* may act as a tumor suppressor in ESCC and that mutations in *NFE2L2* probably impaired its tumor-suppressive function, or even conferred oncogenic activities. Additionally, we found that the *NFE2L2* mutations were significantly associated with worse prognosis of ESCC. We also identified potential noncoding driver mutations including hotspot mutations in the promoter region of *SLC35E2* that were correlated with worse survival. Approximately 5.9% and 15.2% of patients had high tumor mutation burden or actionable mutations, respectively, and may benefit from immunotherapy or targeted therapies. We found clinically relevant coding and noncoding genomic alterations and revealed three major subtypes that robustly predicted patients' outcomes. Collectively, we report the largest dataset of genomic profiling of ESCC useful for developing ESCC-specific biomarkers for diagnosis and treatment.

Cell Research (2020) 30:902–913; https://doi.org/10.1038/s41422-020-0333-6

INTRODUCTION

Esophageal cancer (EC) remains the seventh most common cancer and the sixth leading cause of cancer deaths globally,[1] with over 477,900 new cases and 375,000 annual deaths occurring in China.[2] Esophageal squamous cell carcinomas (ESCC), the predominant histological subtype, is characterized by its aggressive clinical course and constitutes a poor-prognosis subgroup within EC. Epidemiologically, the incidence rate of ESCC shows marked geographical variation across China's regions. It predominates in Taihang Mountain, Xinjiang province and Chaoshan district compared to the rest of China, and rapidly increases in other regions in China.[3] ESCC is a highly heterogeneous disease with unclear molecular classifications and with variable clinical outcomes, and no prognostic biomarkers are available.[4] The precise molecular events underlying ESCC etiology are only partially understood, leading to limited targeted therapies and insufficient clinical management in ESCC patients.

Recent whole-genome and -exome analyses in ESCC revealed a complex mutational landscape and identified significantly mutated genes (SMGs) including *TP53*, *ZNF750*, *NOTCH1*, *FAT1*, *NFE2L2*, recurrent copy number amplifications occurring in *SOX2*, *TERT*, *FGFR1*, *MDM2*, and common deletions of *RB1*.[5–10] However, previous genomic studies in ESCC had several limitations including small sample sizes for the identification of rare driver genes, limited patient outcome information and the focus of targeted protein-coding sequence with little characterization of noncoding regions across ESCC genome.

Despite potential therapeutic targets, few genes altered in ESCC are clinically actionable and a very limited number of inhibitors are currently approved for treatment of advanced

[1]Department of Pathology & Shanxi Key Laboratory of Carcinogenesis and Translational Research on Esophageal Cancer, Shanxi Medical University, Taiyuan, Shanxi 030001, China; [2]State Key Laboratory of Molecular Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China; [3]School of Mathematical Sciences, Center for Statistical Science and Department of Biostatistics, Peking University, Beijing 100871, China; [4]Key laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Laboratory of Molecular Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China; [5]Cancer Institute, Peking University Shenzhen Hospital, Shenzhen Peking University-The Hong Kong University of Science and Technology (PKU-HKUST) Medical Center, Shenzhen, Guangdong 518035, China; [6]Department of Pathology, Shanxi Cancer Hospital, Taiyuan, Shanxi 030001, China; [7]Tumor Hospital affiliated to Xinjiang Medical University, Urumqi, Xinjiang 830011, China; [8]WuXi NextCODE, Shanghai 200131, China; [9]Department of Tumor Surgery, Shanxi Cancer Hospital, Taiyuan, Shanxi 030001, China; [10]Tumor Biobank, Shanxi Cancer Hospital, Taiyuan, Shanxi 030001, China; [11]Department of Pathology, the First Hospital, Shanxi Medical University, Taiyuan, Shanxi 030001, China and [12]Baidu, Beijing 100085, China
Correspondence: Qimin Zhan (zhanqimin@bjmu.edu.cn) or Yanhong Li (robin@baidu.com) or Zhihua Liu (liuzh@cicams.ac.cn)
These authors contributed equally: Yongping Cui, Hongyan Chen, Ruibin Xi, Heyang Cui, Yahui Zhao, Enwei Xu

Received: 21 October 2019 Accepted: 17 April 2020
Published online: 12 May 2020

ESCC. Although the classification of ESCC has been recently revised, the prognostic value of the molecular classification has not been fully elucidated.[11] Hence, there is a clear need to identify novel cancer driver genes based on whole-genome sequencing (WGS) data and to explore additional prognostic biomarkers in a larger set of ESCC patients with available clinical data.

To overcome these challenges, we performed deep WGS in microneedle-punctured formalin-fixed paraffin-embedded (FFPE) tumor tissues and matched adjacent noncancerous specimens from 508 ESCC patients with clinical follow-up data. Herein, we classify ESCC into three main subtypes based on clinically relevant genomic alterations observed across 508 genomic profiles and implicate the association of these subtypes with patients' outcomes. Our genomic study uncovers an extensive landscape of driver genetic alterations across coding and noncoding regions, and defines their potential molecular pathology in ESCC.

RESULTS

Genomic profiling of ESCC

Two clinical centers (Shanxi Cancer Hospital and Tumor Hospital affiliated to Xinjiang Medical University) in Shanxi and Xinjiang provinces, the districts with higher incidence of ESCC in China, participated in this study. WGS was performed on 508 pairs of ESCC tumors and matched adjacent noncancerous esophagus tissues (referred to as 508-WGS cohort). The cohort contains tumor samples collected at the time of diagnosis and includes 437 cases from Han population of Shanxi cohort and 71 cases from Kazak population of Xinjiang cohort (Supplementary information, Tables S1−S3). All tumors were therapy naïve.

The mean sequencing coverages for WGS were 98× and 44× for tumor tissues and matched nontumor samples, respectively. Sequence coverage exceeding 20× was 98.0% for tumors and 92.6% for normal samples. We detected 7,630,294 somatic mutations including single nucleotide variants (SNVs) and small insertions and deletions (InDels) with a median of 12,877 mutations per patient (Supplementary information, Tables S4, S5). We selected 153 mutations for further validation and 78/79 (98.7%) selected coding mutations and 69/74 (93.2%) selected noncoding mutations were confirmed by Sanger sequencing (Supplementary information, Tables S6, S7). A total of 66,260 (0.87% of total mutations) candidate somatic mutations occurred within coding regions linked to 14,971 genes; of which, 61.57% were missense mutations and 5.82% were InDels (Supplementary information, Fig. S1a). The median number of coding mutations was 105.5 (range 1−1217). Among the 13,801 genes with amino acid altering changes (nonsynonymous/truncation), 65 shcowed mutation rates of > 5% in our cohort (Supplementary information, Fig. S1b).

Mutational signatures and association of APOBEC signature with patient stage

To better understand the contribution of these mutations for ESCC etiology, we investigate mutational signatures. Using a modified nonnegative matrix factorization (NMF) algorithm,[12,13] we identified 11 mutational signatures (S1−S11) in the 508-WGS cohort (Fig. 1a; Supplementary information, Fig. S2a). Other than S7 and S10, all other signatures corresponded to mutation signatures in the COSMIC (Catalogue of Somatic Mutations in Cancer) database[14] (Supplementary information, Fig. S2b and Table S8a). Comparing with the COSMIC signatures, we found that S1 and S2 were related with APOBEC (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) activity, S3 with DNA mismatch repair deficiency (dMMR), S4 with age, S8 with aristolochic acid, S9 with alcoholic consumption, and S11 with homologous recombination deficiency. S6 was similar to COSMIC signature S17 and recent studies implicated its association with gastric acid reflux.[15,16]

Next, we correlated the proportions of different mutational signatures with patients' clinical characteristics. Patients at advanced stages of ESCC had significantly higher proportions of S1, S2, and S11 than those at earlier stages (Fig. 1b). Conversely, the proportions of S3, S4, S9 and S10 were significantly larger in patients at earlier stages (Fig. 1b). These data suggest that multiple factors may contribute to the initiation of ESCC, whereas APOBEC activity may continuously contribute to mutagenesis during tumor progression as illustrated in a recent study using cancer cell lines.[17] Moreover, we observed a strong correlation of S9, a signature similar to the drinking-related COSMIC-Signature 16, with both drinking (Cochran−Armitage-trend-test, $P = 0.0005$) and smoking ($P = 0.00049$) in 508-WGS cohort (Supplementary information, Fig. S2c, d). Drinking and smoking status were significantly interrelated in our cohort (Fisher Exact Test, $P = 2.2e-16$), indicating that the significant association of S9 with smoking may be partially attributed to the correlation between smoking and drinking.

Interestingly, clustering analysis of NMF signatures displayed three distinct subtypes across 508 ESCC patients, denoted NMF-cluster 1–3 (Fig. 1c). Cluster 1 (100 patients) was dominated by APOBEC-associated signatures (S1 and S2) while cluster 3 (312 patients) was correlated with mismatch repair deficiency (S3) and spontaneous deamination of 5-methylcytosine (S4), and cluster 2 (96 patients) shared features with deficient homologous recombination repair signatures (S11 and S6). To further investigate APOBEC-associated signatures across 508 ESCC genomes, we examined the occurrence of C > T transition in TpCpW motifs.[18] Notably, all samples in cluster 1 were APOBEC signature-enriched whereas only a small proportion of samples in cluster 2 and cluster 3 were APOBEC signature-enriched (Supplementary information, Fig. S2e). The mean APOBEC signature enrichment score of cluster 1 was dramatically larger than that of cluster 2 and cluster 3 (2.83 vs. 1.60, Student's *t*-test, $P = 2.2e-16$). In addition, we found that *ZNF750* mutations were highly enriched in cluster 1 (Fisher Exact Test, $P = 2.42e-09$, FDR = 4.35e−08) whereas mutations of *FAT2* (Fisher Exact Test, $P = 0.0117$, FDR = 0.21) and *CASP8* (Fisher Exact Test, $P = 0.041$, FDR = 0.39) showed a weak depletion in cluster 2 (Fig. 1d).

The overall survival (OS) rates were significantly different among three clusters, with APOBEC signature-enriched cluster 1 exhibiting worse prognosis (Kaplan−Meier analysis, log rank $P = 0.029$, FDR = 0.087, hazard ratio (HR) = 0.71, 95% confidence interval (CI), 1.007–1.746, Fig. 1e). Consistent with our finding, overexpression of *APOBEC3B* was associated with poor prognosis or treatment resistance in ER^+ breast cancer.[19,20] Moreover, NMF-cluster 1 was significantly associated with lymph node metastasis and advanced stages (Fisher Exact Test, $P = 0.004$, 0.001, respectively, Table 1). Therefore, our results suggest that cluster 1 may be a reliable clinical predictor of metastasis and poor outcome in ESCC.

Hypermutations and its clinical correlation

Recently, tumor mutational burden (TMB) was proved to be an effective biomarker for immunotherapy. Patients with high TMB (TMB-H) generally had a better response to immunotherapy than those with low TMB.[21] Of 508 ESCC patients, we found that 30 (5.9%) were hypermutated (TMB-H, > 10 mutations per Mb). As expected, TMB-H tended to occur in patients with nonsynonymous mutations in DNA mismatch repair (MMR) genes (Fisher Exact Test, $P = 3.72e-5$) or patients with high microsatellite instability (MSI-H) (Fisher Exact Test, $P = 0.00102$) (Fig. 2a; Supplementary information, Table S8b). In addition, TMB-H patients also exhibited a significantly higher proportion of the APOBEC-associated S2 mutations (Wilcoxon Test, $P = 3.999e-05$, Fig. 2b). Furthermore, the TMB was significantly larger in patients enriched with APOBEC-associated mutations within the TpCpW motif (Wilcoxon Test, $P = 2.2e-16$, Supplementary information, Fig. S2f), supporting that the APOBEC activity may be closely

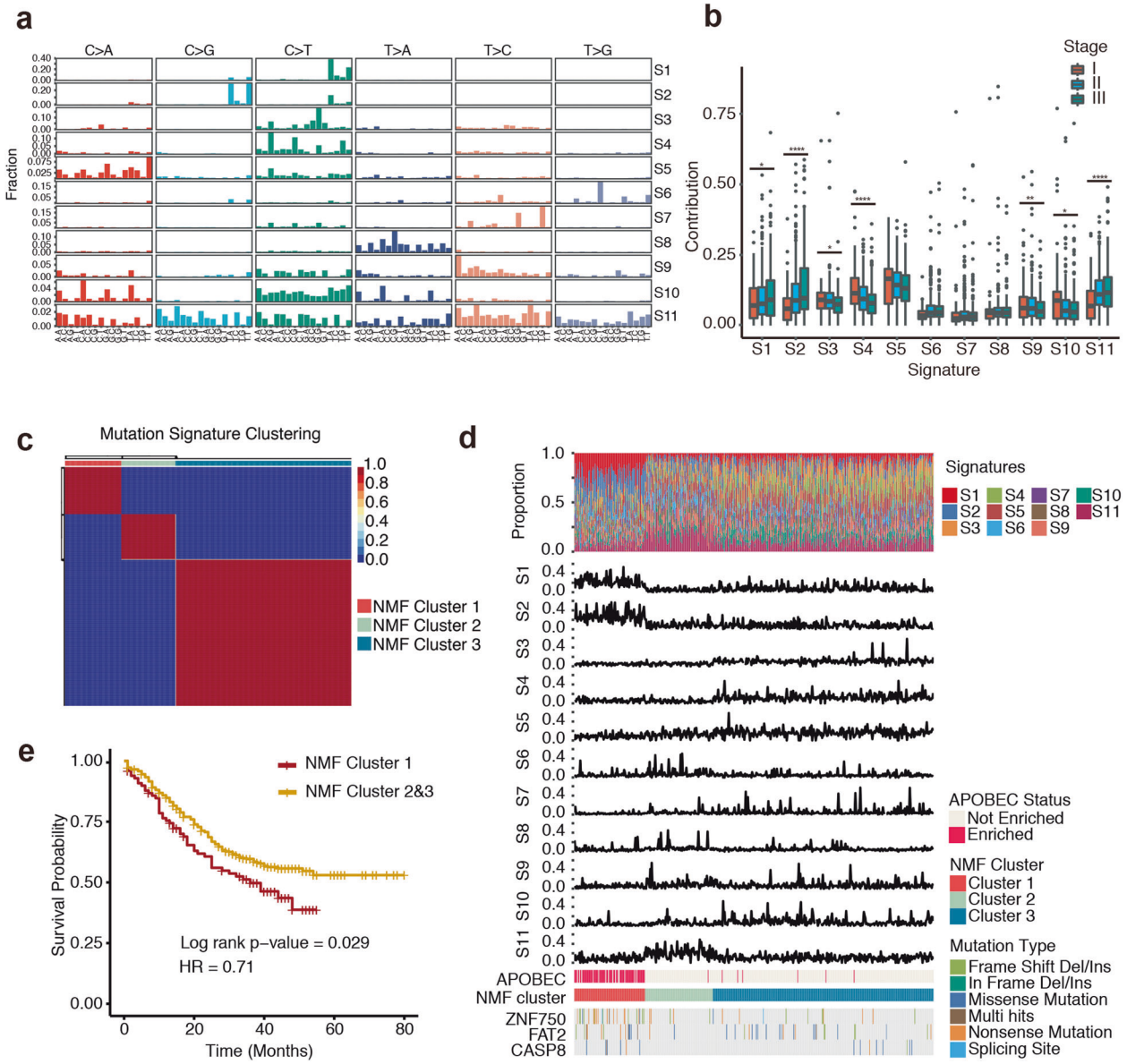

Fig. 1 Profile of mutational signatures in 508 ESCC patients. a Eleven mutation signatures detected in ESCC (S1−S11). **b** Boxplots of the contributions of 11 mutation signatures in tumors at different stages. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$. **c** Three clusters identified by NMF. **d** Upper, proportions of somatic mutations in 11 mutation signatures for each individual. Lower, status of APOBEC enrichment, signature clusters and somatic mutations in *ZNF750*, *FAT2* and *CASP8*. **e** Kaplan–Meier curve of cluster 1 and cluster 2&3.

related to the hypermutation phenotype. Among 30 TMB-H patients, 24 exhibited either MSI-H/dMMR mutation or APOBEC enrichment. Logistic regression coefficient analyses indicated that the hypermutation phenotype was related to MSI ($P = 1.34e{-}05$) and APOBEC ($P = 5.30e{-}08$) activity in ESCC (Supplementary information, Table S8c). Importantly, the TMB-H patients had significantly worse OS than other patients (Kaplan−Meier analysis, log rank $P = 0.011$, HR = 1.87, 95% CI, 1.141–3.078, Fig. 2c).

Significantly mutated genes (SMGs) in ESCC

Analysis of the 508-WGS data using the MutSigCV algorithm[22] and oncodriveFML[23] identified 22 candidate driver genes ($q < 0.1$, Fig. 3a). Analysis of six independent ESCC expression datasets showed that all of these 22 genes had moderate to high expression in ESCC (Supplementary information, Table S9). Of these 22 genes, 17 were reported as SMGs in at least one previous ESCC genome study, including *TP53* (74.80%), *FAT1* (15.94%), *NOTCH1* (15.35%), *KMT2D* (14.76%), *CDKN2A* (10.04%), *FBXW7* (8.66%), *ZNF750* (8.27%), *FAT2* (7.68%), *PIK3CA* (7.48%), *EP300* (6.69%), *NFE2L2* (6.10%), *AJUBA* (4.53%) *RB1* (4.53%), *KMT2C* (3.94%), *CREBBP* (3.54%), *KMD6A* (3.54%), and *TGFBR2* (3.15%). *NFE2L2* (also known as *NRF2*) was predicted to have a total of 31 missense mutations, with the most frequent hotspot at p.R34P and R34Q, followed by p.E79K and E79Q within the DLG and ETGE protein domains, respectively, that bind to KEAP1[24] (Fig. 3b). Although the oncogenic role of *NFE2L2* mutations that inactivate KEAP1-binding site was well established in several squamous cell carcinomas,[25] the biological function of *NFE2L2* and its mutations in ESCC have not been systematically determined.[26] Among the 22 SMGs, we found that patients with *NFE2L2* mutations had a much worse prognosis (Kaplan−Meier analysis, log rank $P = 0.00035$, FDR = 0.0081; Cox regression, $P = 0.001$, HR = 2.21, 95% CI, 1.349–3.325, Fig. 3c), which was further supported by a joint multivariate regression analysis (Cox regression, $P = 1.70e{-}5$, HR = 2.76, 95% CI, 1.74–4.38, Supplementary information, Table S10). Immunohistochemistry (IHC) analysis based on tissue microarray (TMA) that included sequenced tumors and matched normal tissues

Table 1. The association between clusters of mutational signatures and clinical phenotypes of ESCC patients.

Phenotype	NMF Cluster 1	NMF Cluster 2&3	*P* value	FDR
Population (Han)	88 (88.0%)	349 (85.5%)	0.630	1
Population (Kazak)	12 (12.0%)	59 (14.5%)		
Gender (Male)	72 (72.0%)	263 (64.5%)	0.160	0.96
Gender (Female)	28 (28.0%)	145 (35.5%)		
Tumor location lower	32 (32.0%)	128 (31.4%)	0.904	1
Tumor location middle	63 (63.0%)	259 (63.5%)		
Tumor location upper	5 (5.0%)	21 (5.1%)		
Never smoking	41 (41.0%)	213 (52.2%)	0.759	1
Light smoking	4 (4.0%)	17 (4.2%)		
Medium smoking	41 (41.0%)	111 (27.2%)		
Heavy smoking	14 (14.0%)	67 (16.4%)		
Never drinking	65 (65.0%)	282 (69.1%)	0.517	1
Light drinking	16 (16.0%)	57 (14.0%)		
Medium drinking	16 (16.0%)	51 (12.5%)		
Heavy drinking	3 (3.0%)	18 (4.4%)		
Invasion degree 1	5 (5.0%)	38 (9.3%)	0.626	1
Invasion degree 2	22 (22.0%)	112 (27.5%)		
Invasion degree 3	73 (73.0%)	258 (63.2%)		
Not lymphatic metastasis	47 (47.0%)	258 (63.2%)	0.004	0.028
Lymphatic metastasis	53 (53.0%)	150 (36.8%)		
Stage I and Stage II	48 (48.0%)	273 (66.9%)	0.001	0.008
Stage III	52 (52.0%)	135 (33.1%)		

showed lower expression of *NFE2L2* in tumors compared to normal tissues (paired *t*-test, $P = 2.0e{-}06$, Fig. 3d; Supplementary information, Fig. S3a). Cell proliferation assay showed that a reduction of endogenous wild-type *NFE2L2* (*NFE2L2*-wt) level in KYSE410 and KYSE450 cells led to increased cell proliferation whereas ablation of mutant *NFE2L2* (*NFE2L2*-mut) in KYSE70 and KYSE180 cells caused decreased cell proliferation (Supplementary information, Fig. S3b–e). Further studies verified the promotion of cell proliferation by stable *NFE2L2* knockdown in KYSE450 cells (Supplementary information, Fig. S3f). In contrast, overexpression of exogenous *NFE2L2*-wt in KYSE150 cells with low-expression of *NFE2L2*-wt significantly attenuated cell proliferation. Notably, *NFE2L2* mutants (p.R34Q, p.E79K, p.K438E, p.R569H) completely interfered with the suppressive activity of *NFE2L2*, and some *NFE2L2* mutants (p.R34Q, p.E79K) even exerted an oncogenic role (Supplementary information, Fig. S3g). Consistently, xenograft mouse model showed that *NFE2L2*-wt ablation promoted tumor growth whereas *NFE2L2*-wt overexpression suppressed tumor growth. Some *NFE2L2* mutants (p.R34Q, p.E79K) significantly promoted tumor growth, while other *NFE2L2* mutants (p.K438E, p.R569H) had no significant effect on tumor growth compared with the vector control (Fig. 3e, f). Taken with our genetic observations, these functional data indicate that *NFE2L2* may act as a tumor suppressor in ESCC; mutations in *NFE2L2*, which serve as a poor prognosis biomarker, probably impaired its tumor-suppressive function, or even conferred oncogenic activities.

Beyond the known SMGs, we identified five novel SMGs including *KRT5*, *CDH10*, *LILRB3*, *YEATS2* and *CASP8* in ESCC (Supplementary information, Table S11). Among them, *CASP8* was listed as a cancer consensus gene in the COSMIC database. The other genes including *CDH10*, and *YEATS2* have been reported as drivers in other cancer types.[27,28]

Copy number alterations (CNAs)

Analysis of somatic CNAs (SCNAs) identified 8 recurrent arm-level amplifications, 4 arm-level deletions, 24 focal amplifications and 28 focal deletions (Supplementary information, Fig. S4a). In addition to known focal events, such as the 11q13.3 amplification containing *CCND1* and the 9q21.3 deletion containing *CDKN2A/B*,[6] we identified two novel focal deletions at 11p15.5 and 22q13.33, encompassing the interferon regulatory factor *IRF7*[29] and a component of the MOZ/MORF acetyltransferase complex *BRD1*,[30] respectively. Putative SCNAs were validated in 22 cases within *CCND1*, *CTTN*, *MYEOV*, *CNTN5*, *TRPC6* and *PGR* genes using *HBB* as an internal control and 95.5% were confirmed by quantitative PCR (Supplementary information, Table S12). Amplification of 11q13.3 was significantly associated with patients' poor outcomes (Kaplan–Meier analysis, log rank $P = 0.016$, HR = 1.38, 95% CI, 1.058–1.805, Fig. 4a). Intriguingly, we identified, for the first time, that losses of 13q12.11, 13q14.2, 17q25.3 and 22q13.33 were significantly associated with patients' poor outcomes (log rank *P* in Supplementary information, Table S13). Moreover, four amplified and ten deleted regions were associated with tumor stage and lymph node metastasis (Supplementary information, Table S13). Finally, we performed hierarchical clustering based on the SCNAs and identified three CNA clusters. CNA-cluster 3 had the highest level of SCNA, followed by CNA-cluster 2 and 1 (Fig. 4b) and 11q13.3 amplification was more enriched in CNA-cluster 3 (Fisher Exact Test, $P < 2.2e{-}16$). Patients in CNA-cluster 2 and CNA-cluster 3 tended to have a poorer prognosis than CNA-cluster 1 (Kaplan–Meier analysis, log rank $P = 0.08$, Supplementary information, Fig. S4b).

Genome-wide analysis of noncoding regulatory mutations

We examined noncoding somatic mutations located at gene promoter regions, 5′ and 3′ untranslated regions (UTRs), long noncoding RNAs (lncRNAs), introns, and intergenic regions. As expected, intergenic and intron regions carried the highest mutational burden (Supplementary information, Fig. S4c). Recurrent noncoding hotspot mutation analysis identified 13 hotspots including hotspots in the promoter of WDR74 that were previously reported in multiple human cancer types[31] (Fig. 4c upper panel; Supplementary information, Table S14). Survival analysis revealed that the hotspot mutation at the promoter of *SLC35E2* was correlated with a worse prognosis (Kaplan–Meier analysis, log rank $P = 0.0025$, FDR = 0.0058, HR = 3.24, 95% CI, 0.984–3.507, Fig. 4d). Meanwhile, 3.9% (20 of 508) ESCC samples had the A30G hotspot mutation at the transcript lncRNA *RP11-69I8.2-003* that has not been previously linked to cancer (Supplementary information, Fig. S4d). We found that inhibition of *RP11-69I8.2-003* significantly suppressed cell proliferation whereas its overexpression enhanced cell proliferation (Supplementary information, Fig. S4e–h). Notably, the *RP11-69I8.2-003* A30G hotspot mutation dramatically promoted cell proliferation (Supplementary information, Fig. S4g, h). These data suggest that the hotspot mutation of *RP11-69I8.2-003* may be a gain-of-function mutation and its mutant may exert a tumor-promoting property.

Since the recurrent hotspot analysis was conservative, we next searched for recurrently mutated noncoding elements by investigating their mutation frequencies. We identified 112 lncRNAs, 225 3′-UTRs, 34 5′-UTRs, and 627 promoters that were significantly frequently mutated than expected (FDR < 0.05, Fig. 4c lower panel; Supplementary information, Table S15). *NEAT1* and *MALAT1* were the most frequently mutated lncRNAs. We identified 107 *NEAT1* mutations in 94 patients (FDR = 7.38e−16) and 67 *MALAT1* mutations in 54 patients (FDR = 7.38e−16). Both lncRNAs were found to be recurrently mutated in multiple tumor types and play important roles in tumorigenesis.[32,33] Further study needs to be done to explore the role of *NEAT1* and *MALAT1* in ESCC tumorigenesis. Among the recurrently mutated noncoding elements, one lncRNA, one 3′-UTR and eight promoters

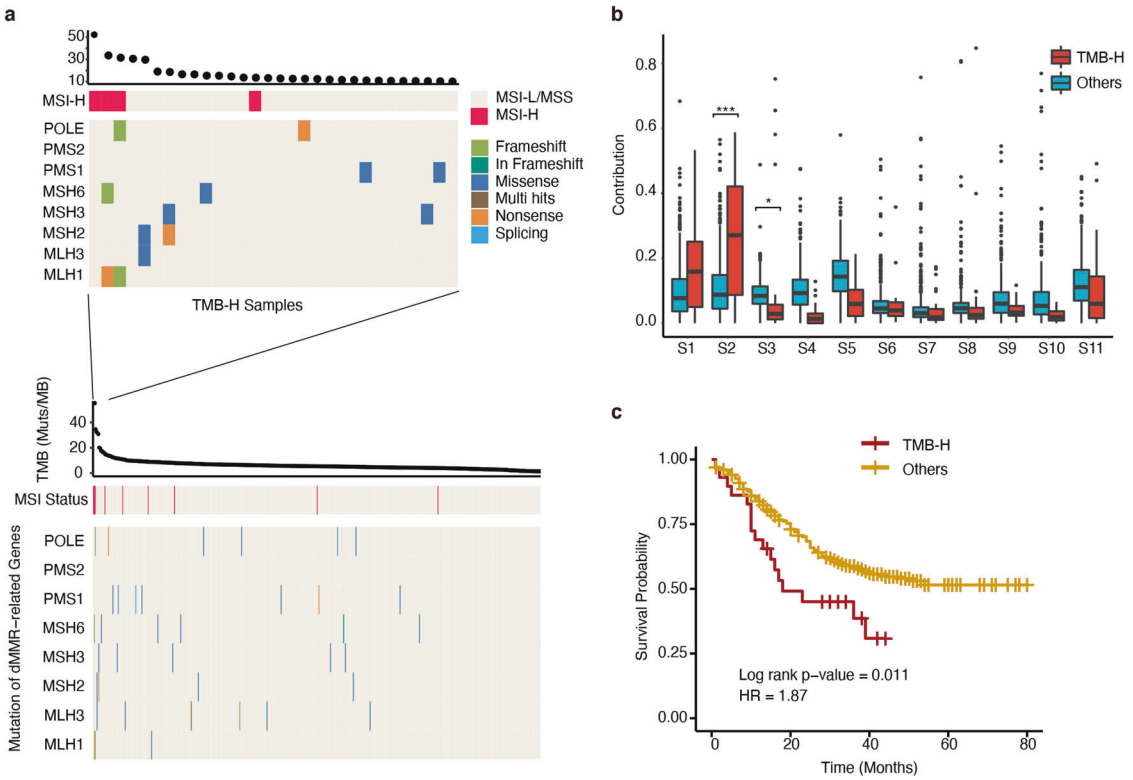

Fig. 2 TMB and MSI analyses. a The upper panel and lower panel are for the TMB-H patients and all patients, respectively. Each panel shows TMB, MSI-H status, and somatic mutation status in MMR-related genes, respectively. Patients are ordered decreasingly by their TMB values. **b** The contributions of 11 mutation signatures in TMB-H tumors and other tumors. *$P < 0.05$, ***$P < 0.001$. **c** Kaplan–Meier curve of TMB-H patients and other patients.

were significantly correlated with survival, such as the lncRNA *LINC00966* and the promoter of *CCND1* (FDR < 0.05, Fig. 4e).

Potential actionable targets and associated pathways in ESCC

ESCC patients are diagnosed at advanced stages of the disease and only a small group of patients will benefit from standard of care therapies. Significant efforts have been made to identify molecular-targeted therapies for ESCC patients. Here, we found that 77 out of 508 (15.2%) patients had at least one genomic alteration among the 40 targetable alterations in the curated precision oncology knowledge base (oncoKB[34]) (Supplementary information, Fig. S5a). *EGFR* amplification was the most prevalent actionable alteration (31 cases, 6.1%), followed by *FGFR1* amplification (26 cases, 5.12%). Interestingly, *EGFR* and *FGFR1* amplifications showed mutual exclusivity. Although not significant, patients harboring either *EGFR* or *FGFR1* amplification (57 patients) showed poorer prognosis (Supplementary information, Fig. S5b). The RTK-RAS pathway showed a high frequency of alterations in pan-cancer studies[35] and was also altered most frequently (257 out of 508 samples) in ESCC cohort (Fig. 5a). Importantly, amplifications in the RTK-RAS pathway, a pathway containing many actionable alterations including *EGFR* and *FGFR1*, were significantly correlated with the patients' worse survival (Kaplan−Meier analysis, log rank $P = 0.0032$, HR = 1.54, 95% CI, 1.178–2.009, Fig. 5b upper panel). Correlation of the RTK-RAS signaling-related amplifications with worse survival was also observed in gastric cancer.[36] Amplifications of *EGFR*, *FGFR1*, *KRAS*, *MET*, and *ERBB2* were largely mutually exclusive among 132 amplified samples ($P = 0.002$). Collectively, our results provide potential prognostic biomarkers and probably narrow down future target choices on the precision treatment in ESCC.

The RTK-RAS pathway-related amplifications showed significant co-occurrence with amplifications of the MYC pathway and cell cycle pathway; meanwhile, amplifications of the RTK-RAS pathway tended to co-occur with deletion of cell cycle pathway (Fisher Exact Test, all $P < 1e{-}7$). Amplification of the MYC pathway was also significantly correlated with patient's survival (Kaplan−Meier analysis, log rank $P = 0.017$, HR = 1.40, 95% CI, 1.058–1.847, Fig. 5b middle panel). For cell cycle pathway, *CDKN2A/B* were mostly altered by deletions whereas *CCND1*, located in 11q13.3, was frequently amplified. Interestingly, *CCND1* also harbored promoter mutations in 17 samples and patients with promoter mutations showed much worse prognosis than those with amplifications (Supplementary information, Fig. S5c). Additionally, within the NRF2 pathway, *NFE2L2* was mutated in an exclusive manner with *CUL3* and *KEAP1* as previously reported.[37] *HES1* and *NOV* ranked the two most amplified genes within the NOTCH pathway.[38] For the Wnt pathway, *LRP5*, *GSK3B*, *LRP6*, *SFRP1*, and *DKK4*[39] were amplified while mutations were not frequently detected.

Integrative analysis of clinically relevant genomic alterations across 508 ESCC genomes

To build a robust predictive model, we used the stability to perform feature selection[40] in the Cox regression and found that *NFE2L2* mutation and the RTK-RAS-related amplification were the two most stable molecular features, with selection probabilities of 0.90 and 0.83, respectively (Supplementary information, Table S16a). Although the MYC pathway-related amplification was significantly correlated with survival, the selection probability of MYC pathway was only 0.53, partly due to its high co-occurrence with RTK-RAS amplification (Fisher Exact Test, $P = 2e{-}9$). We thus combined amplification of RTK-RAS and MYC pathways and found that the new feature, RTK-RAS-MYC pathway amplification, had a selection probability of 94%, greater than that of RTK-RAS amplification (83%). The RTK-RAS-MYC amplification was also significantly correlated with patients' survival (Fig. 5b lower panel, Kaplan−Meier analysis,

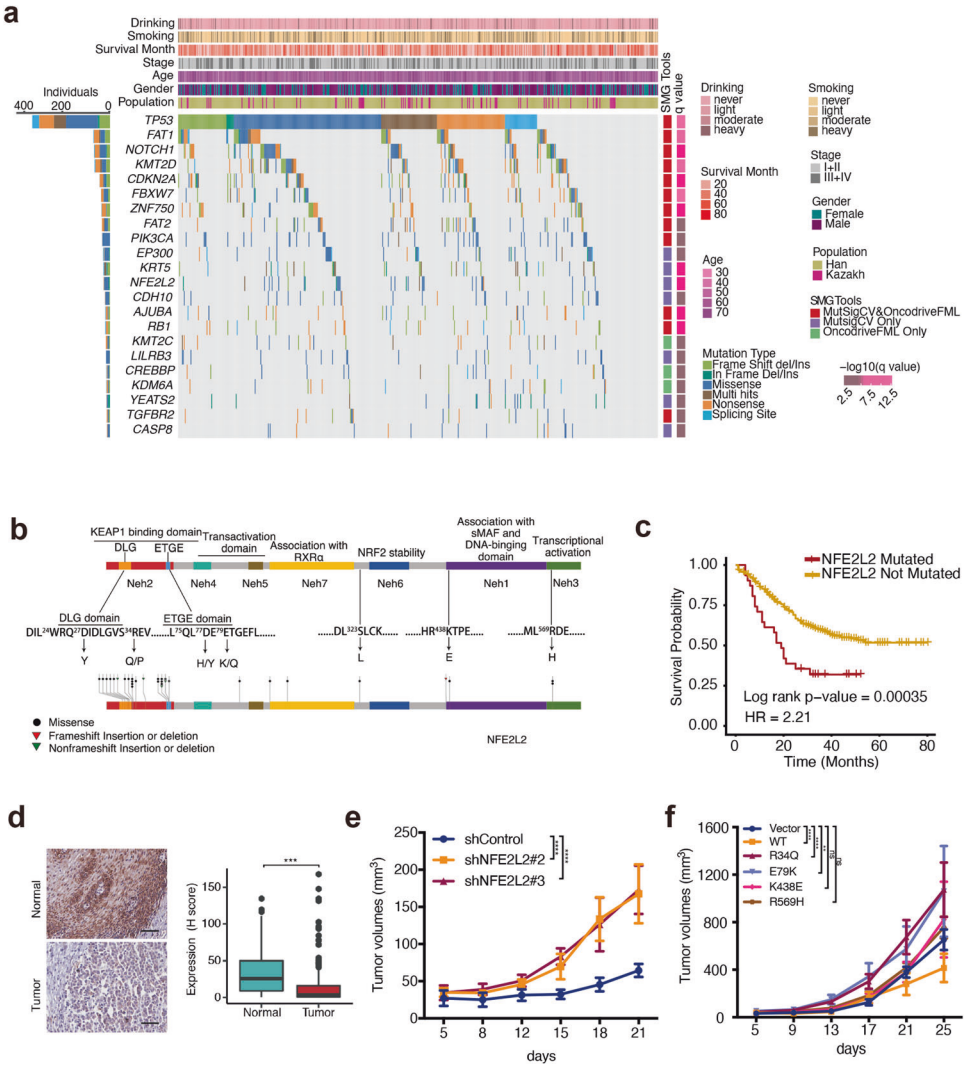

Fig. 3 Mutational landscape of somatic alterations across 508 ESCC genomes and oncogenic mutations of *NFE2L2* identified from SMGs. a SMGs identified by MutSigCV and OncodriveFML with *q* value < 0.1. Rows are genes and columns are tumor samples. The patients' phenotypic information is shown in the upper panel. The right panel shows the significance of each SMG. **b** Somatic mutations affecting *NFE2L2*. **c** Kaplan–Meier curve of *NFE2L2*-mutated and nonmutated samples. **d** Representative images (left) and statistical analysis (right) of IHC for NFE2L2. Scale bars, 50 μm. **e** Tumor volumes of *NFE2L2* shRNA and control KYSE450 cells. $n = 6$ mice per group. **f** Tumor volumes of *NFE2L2-wt*, mutant, and the corresponding control KYSE150 cells. $n = 6$ mice per group. Data in (**d**–**f**) represent mean ± SD. Data were analyzed by unpaired two-tailed Student's *t*-test (**d**) or two-way ANOVA with Bonferroni correction (**e**, **f**). ns no significant, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

log rank $P = 0.00018$, HR = 1.66, 95% CI, 1.27–2.17). Replacing the RTK-RAS amplification with the RTK-RAS-MYC amplification in the Cox regression model controlling clinical features led to an increase of the *R*-square from 0.165 to 0.173, and both *NFE2L2* mutation and RTK-RAS-MYC amplification were highly significant ($P = 1.25e{-}05$ and 5.11e−4, respectively, Supplementary information, Table S16b). We classified the tumors into three distinct subtypes, *NFE2L2*-mutated, RTK-RAS-MYC-amplified, and double-negative subtypes (Fig. 5c). The *NFE2L2*-mutated subtype consisted of tumors with *NFE2L2* mutations, the RTK-RAS-MYC-amplified subtype included tumors with the RTK-RAS-MYC amplifications but without *NFE2L2* mutations, and the double-negative subtype was all other tumors. The survival of these three subtypes was significantly different from each other, even after controlling for other covariates (Supplementary information, Table S16b). The *NFE2L2*-mutated subtype had the worst survival, followed by the RTK-RAS-MYC-amplified subtype (Kaplan−Meier analysis, log rank $P = 9.1e{-}07$, Fig. 5d; Supplementary information, Table S16c).

DISCUSSION

This study provides the most comprehensive characterization of ESCC genomes, to date, using deep WGS in a large Chinese patient cohort with clinical follow-up data. In addition to potential driver mutations reported in other studies,[5–10] this massive amount of WGS data allowed us to identify novel driver-coding (e.g., *KRT5, YEATS2*) and noncoding mutations (e.g., *SLC35E2* and *RP11-69I8.2-003*). A set of functional assays confirmed *NFE2L2* and *RP11-69I8.2-003* mutations as oncogenic drivers for ESCC progression. We also uncovered previously unexplored features in ESCC genomes, including the three mutation signature clusters, the discovery of TMB-H/MSI-H ESCC genomes, and the three CNV clusters. We found that a significant portion of ESCC patients had actionable mutations and may potentially benefit from targeted therapies. Most importantly, available clinical data allowed us to identify potential prognostic biomarkers such as *NFE2L2* mutations, *SLC35E2* promoter mutations, clusters of mutation signatures, TMB-H, and RTK-RAS-MYC amplification. Integrative analysis

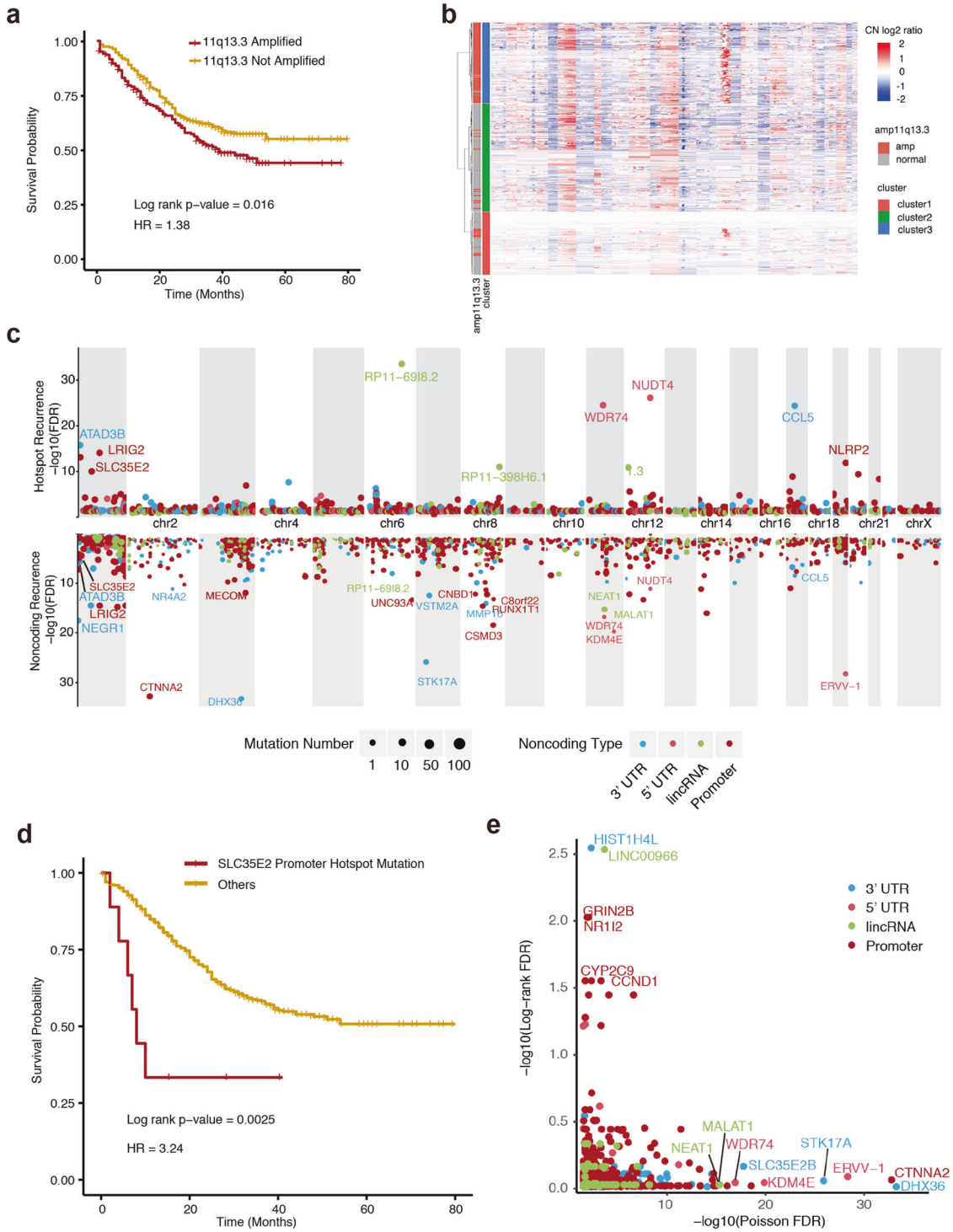

Fig. 4 CNA profiling and noncoding mutations. a Kaplan–Meier curve of 11q13.3-amplified samples and others. **b** Hierarchical clustering of tumors based on log2 ratio of copy number. Rows represent tumors and columns are genomic positions. The patients are grouped into three clusters. Amplifications are marked by red and deletions are marked by blue. **c** The significance of noncoding hotspot mutations (upper) and frequent mutated noncoding elements (lower). The *y*-axes are the log10 FDRs. Circle size represents the number of mutations. **d** Kaplan−Meier curve of patients with or without mutations at the *SLC35E2* promoter region. **e** Scatter plot of the significance of mutation frequencies in noncoding elements given by the Poisson model (*x*-axis) against the significance of the noncoding mutations with prognosis given by the log rank test (*y*-axis).

of these clinical relevant features revealed three subtypes of ESCC (*NFE2L2*-mutated, RTK-RAS-MYC-amplified, and double-negative) that may robustly predict patients' outcomes. We provide a public genotype/phenotype-coupled resource that represents a further step toward targeting the ESCC genome for clinical purposes.

Compared with prior ESCC studies, our larger number of microneedle-punctured tumor samples provided additional power to identify cancer driver genes, leading to the discovery of five novel driver genes (*KRT5*, *CDH10, LILRB3*, *YEATS2*, and *CASP8*), as well as recurrent CNAs. Previous studies mostly used whole-exome

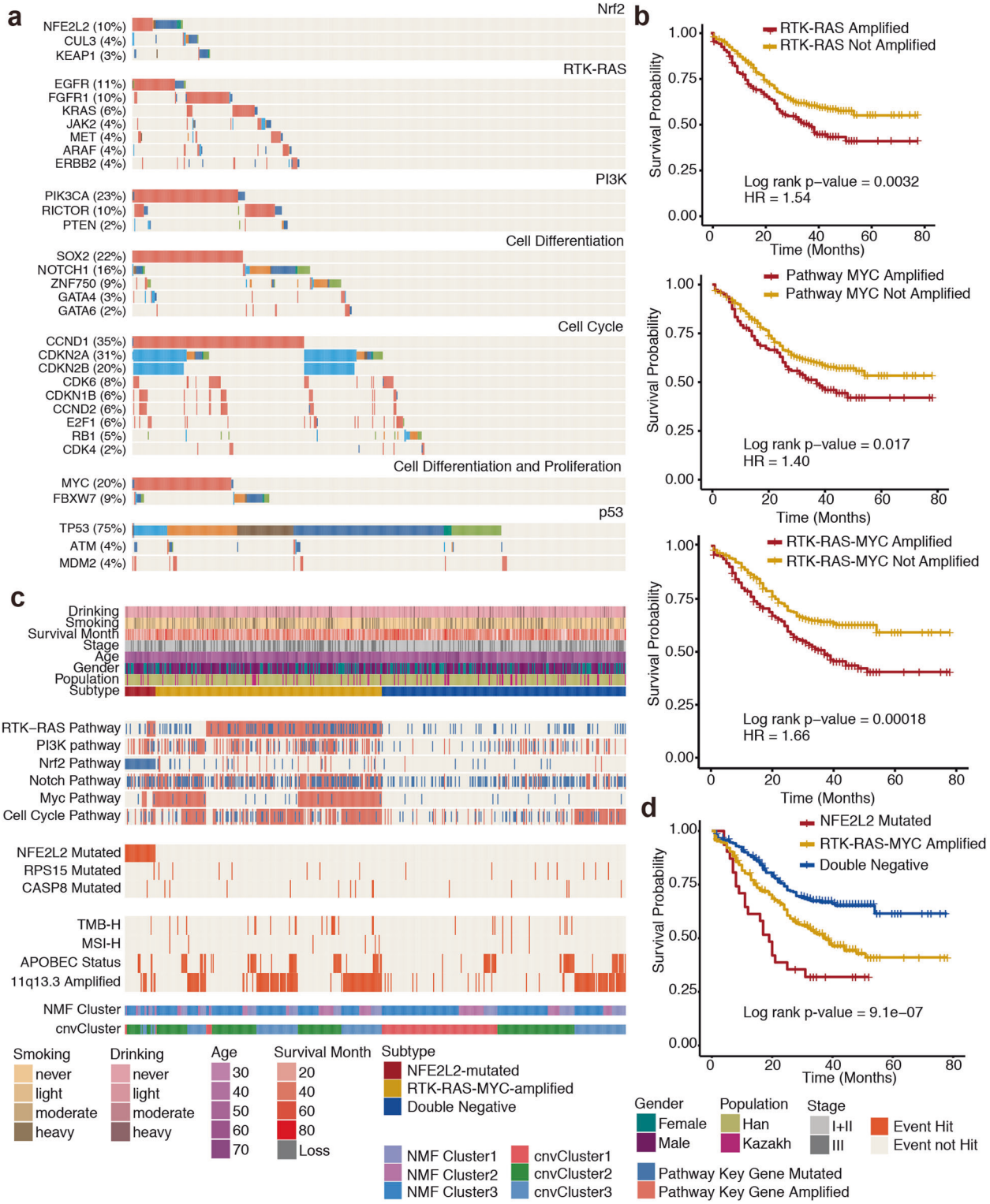

Fig. 5 Genomic alterations in actionable targets and cancer pathways, and an integrative model of the alterations. a Copy number alterations and nonsilent somatic mutations of key genes in five cancer pathways. Genes are ordered by alteration frequency. Wide bars represent the amplifications (red) and deletions (blue) while narrow bars represent various types of somatic mutations. **b** Kaplan–Meier curves of RTK-RAS amplification (upper), MYC amplification (middle) and RTK-RAS-MYC amplification (lower). **c** Integrative profiling of multiple genomic alterations in 508 patients. **d** Kaplan–Meier curve of the *NFE2L2*-mutated, RTK-RAS-MYC-amplified and double-negative ESCC subtypes.

sequencing, leaving the nonprotein-coding part of ESCC genomes remains widely unexplored. Our WGS analysis identified many novel recurrent- and prognostic-related noncoding mutations (e.g., *SLC35E2* and *CCND1* promoter mutations). Functional analysis clearly demonstrated that *RP11-69I8.2-003* hotspot mutation exerted gain-of-function properties. These noncoding mutations and rare driver-coding mutations may improve the understanding of ESCC tumorigenesis and can be used for genetic counseling, to

predict patients' outcome and to determine the most optimal treatment for ESCC patients.

APOBEC cytidine deaminase activity is an endogenous mutagen and APOBEC signature mutations can accumulate by the ongoing APOBEC activity, consistent with our observation that late-stage tumor tended to have more APOBEC signature mutations. Paradoxically, therapies targeting APOBEC may be achieved by either enhancing the mutagenic effect of APOBEC or inhibiting APOBEC activity. In fact, recent studies implicated that inhibition of DNA damage response (e.g., inhibition of PARP and ATR) in cells with high APOBEC expression promoted apoptosis and cell death.[41,42] Meanwhile, inhibiting APOBEC expression led to prolonged tamoxifen responses for ER^+ breast cancer in murine xenografts.[20] On the other hand, APOBEC signature was associated with high mutation burden and a large number of neoantigens may be generated due to APOBEC mutagenic effect. Hence, immunotherapy may be a better therapeutic option for patients enriched with APOBEC signature mutations.

Alterations of genes involved in the RTK-RAS pathway were largely mutually exclusive, as observed in previous pan-cancer studies.[35,43] This may be explained by (1) the second hit within the same pathway provides no further survival advantages or (2) the second hit results in survival disadvantage and even synthetic lethality. Forced expression of mutant *EGFR* and *KRAS* can be synthetic lethal,[44] implying that the second explanation may be more likely for the mutual exclusivity of RTK-RAS-related alterations. The co-occurring alteration pattern of the RTK-RAS pathway with cell cycle-related signaling and the MYC pathway may have important therapeutic implications in ESCC. In lung cancer, co-occurring cell cycle alterations were significantly associated with the lack of response to osimertinib treatment in *EGFR*-mutated patients.[45] In breast cancer, JQ1, a BET bromodomain inhibitor that decreases *MYC* expression, sensitized *ERBB2*-amplified breast cancer cells to lapatinib.[46] These mutual exclusive and co-occurring patterns in ESCC may provide new treatment options for ESCC patients.

Although recent efforts have focused on characterizing ESCC genomic alterations,[5–10] the number of clinically relevant biomarkers is still limited. In this study, we uncover *NFE2L2* mutation and the RTK-RAS-MYC pathway amplification as better molecular features in predicting ESCC patients' poor survival. In line with our finding, the RTK-RAS amplification has been associated with adverse prognosis in gastric cancer.[36] After further validation of its prognostic value and therapeutic implication, our molecular subtypes may help in selection for administration of therapeutic trials.

Recent research systematically characterized the genomic landscape of 551 esophageal adenocarcinomas (EAC) and defined genomic biomarkers to be pursued in the clinic.[47] Although EAC and ESCC share a number of recurrent driver alterations, such as frequent *TP53* mutation and *CDKN2A* deletion, their genomic landscapes are substantially different.[10] *NFE2L2* mutations were only present in ESCC. *GATA4/6* and *SMAD4* alterations were much more frequent in EAC. *EGFR* and *FGFR1* were the most often amplified RTK/RAS-related genes in ESCC, but *KRAS* and *ERBB2* amplifications were more common in EAC. Interestingly, clinical biomarkers were also different, such as *NFE2L2* mutation and RTK-RAS-MYC amplification in ESCC, and *GATA4* amplification and *SMAD4* alteration in EAC. These data show that ESCC and EAC were distinct diseases and their effective therapies may be totally different.

Our study has a few limitations. Firstly, the relationship between genomic landscape of ESCC and epidemiological and transcriptomic backgrounds remains unraveled. Complementary to this study, sequencing approaches that comprehensively determine transcriptome, epigenome and the omics-interactions will be an important next step to provide further insight into the molecular changes underlying the pathogenesis in patients with ESCC. Secondly, hypotheses generated from this study will require clinical validation. Going forward, it would be important to evaluate the candidate predictive biomarkers identified from this study in clinical trials. In summary, this in-depth analysis of all classes of somatic mutations demonstrates a unique molecular profile and provides clues to the etiology of ESCC. Our study also sets the stage for improved prognosis prediction of ESCC based on their unique genetic alterations with potential clinical relevance.

MATERIALS AND METHODS

Sample collection and clinicopathological features of ESCC patients

Two clinical centers from Shanxi and Xinjiang provinces participated in this study. A total of 508 patients diagnosed with ESCC who had received no prior treatment were recruited. Informed consent was obtained from all subjects, and this study was approved by the ethical committees of the Shanxi Medical University. Each tumor specimen had a companion normal tissue specimen. All cases were classified according to WHO criteria. Hematoxylin and eosin (H&E)-stained sections from each sample were subjected to review by at least three independent pathologists with no clinical or molecular information to confirm that the tumor specimen was histologically consistent with ESCC and that the adjacent tissue specimen contained no tumor cells. The clinical backgrounds for this cohort are shown in Supplementary information, Tables S1–S3.

DNA extraction and whole-genome sequencing

Experimental DNA samples were obtained from microneedle-punctured FFPE tissues which have high quality to ensure tumor purity. Briefly, our pathologist examined the pathological section of all tumors by H&E staining first, then selected and marked the area with the most abundant tumor cells and the least stromal components on the slide, avoiding necrosis and inflammatory cells. Then our pathologist marked the same area on the wax block according to the slide, punctured at the marked area of wax block with a hollow core puncture needle. The inner diameter of hollow core puncture needle was about 2 mm, and the thickness of wax block was about 2–3 mm. Tumor purity of all samples was evaluated by pathologist based on H&E staining (Supplementary information, Table S1). After microneedle-punctured procedure, high-quality total DNA from 10 mg tissues was extracted by Maxwell 16 Tissue DNA Purification Kit (Promega) according to the manufacturer's instructions. Approximately 300 ng high-quality DNA sample ($OD_{260}/_{280} = 1.8$~2.0) was sheared with Covaris S220 Sonicator (Covaris) to ~350 bp. Fragmented DNA was purified using Sample Purification Beads (Illumina). Adapter-ligated libraries were prepared with the TruSeq Nano DNA Sample Prep Kits (Illumina) according to Illumina's protocol. Sequencing was performed using an Illumina HiSeq system for 2×150 paired-end sequencing in WuXi NextCODE at Shanghai, China. For variant calls, please see Supplementary information, Data S1.

Somatic mutation calling and SMG identification

High-quality reads were aligned to the UCSC human reference genome (hg19) using Burrows–Wheeler Aligner (BWA v.0.7.12)[48] with default parameters. For each paired sample, somatic SNVs and InDels were detected by Sentieon TNseq.[49] Mutations in low complexity regions such as tandem repeat regions and highly homologous regions in the genome were filtered out. Low confidence variants were removed if any one of the following criteria is not satisfied: total depth > 10, alternative allele depth > 3 and mutation frequency > 0.01. All high-confident mutations were then annotated with ANNOVAR (Version 2016-02-01).[50] SMGs were detected using MutSigCV (v1.4)[22] and OncodriveFML (v2.0.2)[23] with default settings. Genes with FDR < 0.1 were considered to be significantly mutated.

Mutation signature and mutation signature cluster analysis
SomaticSignatures[12,13] was used to identify de novo mutation signatures. Number of signatures was determined by choosing the reflection point in the curves of explained variance and residual sum of squares (RSS) (Supplementary information, Fig. S2a). De novo mutational signatures were then compared to curated signatures in COSMIC using cosine similarity. The Cochran−Armitage trend test was used to examine the mutation signature contribution among various groups. For each tumor, we calculated the frequencies of the 96 mutation subtypes and obtained a 96×508 mutation subtype frequency matrix. Then NMF (https://cran.r-project.org/web/packages/NMF/index.html) in combination of the consensus clustering is applied to cluster the tumors.

APOBEC enrichment analysis
We used the method as previously described[51] to examine the APOBEC enrichment. This method was implemented in Maftools[52] and we used this tool for our APOBEC enrichment analysis. Tumors with enrichment scores > 2.5 were considered as APOBEC enriched.

Noncoding elements mutation analysis
We followed the previously described method[31] to identify hotspot in noncoding regions. Noncoding elements regions were downloaded from the git repository OncodriveFML (v2.0.2).[23] Promoter regions were defined as 1000 bp upstream and downstream around transcription start site. For lncRNAs, only exonic mutations were retained for downstream analysis. We categorized these mutations following the priority hierarchy, promoter > 5′ UTR > 3′UTR > lncRNA. To evaluate the significance of mutation frequency in noncoding elements, we first fitted a Poisson regression of mutation frequency against the length of noncoding elements. More specifically, let Y_i be the number of mutations detected on the ith noncoding element and L_i be its length. We modeled Y_i following a Poisson distribution with a mean parameter λ_i. We assumed that $E(Y_i|L_i) = \lambda_i = \beta_0 + \beta_1 \log L_i$. After obtaining the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we could calculate an expected number of mutations $E_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 \log L_i)$. Finally, we determined the P value of the ith noncoding element as $P(Y_i > E_i)$ using the Poisson distribution with parameter E_i. This significance analysis was performed separately for promoters, 5′ UTRs, 3′UTR and lncRNAs. In survival analysis, the genes having both coding and noncoding mutations were excluded.

Somatic copy number variation calling and profiling
We used BIC-seq2[53] to analyze SCNAs. SCNAs were called by BIC-seq2-seg by comparing the normalized tumor and normal data. Regions with absolute log2 copy number ratios at least 0.58 (= log2(1.5)) were viewed as losses or gains. For actionable amplification identification, we used a more stringent cutoff with log2 copy number at least 1 (log2(2)). Hierarchical clustering was performed to cluster tumors' SCNA profiles. GISTIC2.0 [54] was used to identify recurrent arm-level and focal SCNA segments.

Cox regression and survival analysis
The OS distributions were described by the Kaplan−Meier curves, and statistical significance was calculated using the log rank test. Univariate and multivariate analyses with the Cox proportional hazards model were used to examine the association between overall survival and clinical phenotypes. For multivariate analysis, in addition to the covariates with univariate P values of < 0.05, stages and lymphatic metastasis were adjusted for the analysis of the prognosis. Cox regression and survival analysis were performed with R package survival (https://github.com/therneau/survival) and survminer (https://cran.r-project.org/web/packages/survminer/index.html).

Integrative analysis of molecular features
We used lasso regularized Cox regression model coupled with the stability feature selection procedure[40] to select the most stable features that could predict the patient's outcome. In the Cox regression model, we considered molecular and clinical phenotypic features. The molecular features included the 22 SMGs, the 5 prognosis-related focal deletions/amplifications, TMB-H status, MSI-H status, somatic mutations at dMMR-related genes, APOBEC enrichment status, the NMF cluster, the CNV cluster and mutation status, amplification status and deletion status of the seven key cancer pathways (Nrf2, RTK-RAS, Myc, Cell cycle, Notch, Wnt, Chromatin remodeling). The clinical phenotypic features included TNM stages, lymphatic metastasis, tumor grade, gender, age, drinking, smoking and population. We varied the lasso penalty parameter λ from 0.03 to 0.2 stepped by 0.01 and defined the selection probability of each feature as its maximum of selection probabilities for different penalty λ.

TMA and immunohistochemical staining
TMA blocks were made using the TMA builder (TC IV, Chloe, China) with 508 sequenced tumors and matched normal tissues. Immunohistochemical analysis was performed as previously described.[55] The images were captured by Aperio Scan Scope (Leica, Nussloch, Germany). The immunoreactive H score was determined by Aperio ImageScope Cytoplasma 2.0 software.

Sanger sequencing
The somatic mutation was detected by PCR amplification and Sanger sequencing. We randomly selected 100 coding mutations and 100 noncoding mutations for validation. Considering the sensitivity of Sanger sequencing, we further filtered 175 mutations with a frequency of over 10% for validation. Among them, PCR amplification failed in 22 cases and PCR amplicon sequencing succeeded in 153 cases. Mutations of *NFE2L2* gene in cell lines were detected by Sanger sequencing. Sequences of primers are available in Supplementary information, Tables S6, S7, S17.

Cell culture
Immortalized human normal esophageal epithelial cell line Het-1A was purchased from ATCC (Manassas, VA). Immortalized esophageal epithelial NE-2 cells were gifted from Dr. Enmin Li. ESCC cell lines (KYSE series)[56] were provided by Dr. Y. Shimada. ZEC014 and ZEC134 cells were gifted from Dr. Dan Su. Het-1A cells were cultured in bronchial epithelial basal medium with growth supplements (Clonetics, USA). NE-2 cells were cultured in a 1:1 mixture of EpiLife and dKSFM (Gibco). KYSE series cells were cultured in RPMI 1640 medium supplemented with 10% fetal bovine serum (FBS) and antibiotics. And ZEC014 and ZEC134 cells were cultured in DMEM-F12 medium containing 10% FBS, 1% nonessential amino acid (NEAA) and antibiotics. All cell lines were maintained at 37 °C in humidified atmosphere containing 5% CO_2. All of the cells were authenticated by short tandem repeat (STR) analysis and regularly tested for mycoplasma contamination.

siRNA transfection
siRNAs were purchased from GE Dharmacon. Cells were transfected with 40 nM of siRNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. Sequences of siRNAs are available in Supplementary information, Table S17.

Vector construction
The cDNA was cloned into the pLVX-IRES-Neo vector. Site-directed mutagenesis was performed based on overlap extension by PCR. shRNA oligos were inserted into lentiviral vector pSIH1-H1. Sequences of primers and shRNAs were listed in Supplementary information, Table S17. All constructs were verified by DNA sequencing.

Lentiviral transduction
The lentiviral vector together with packaging vectors was transfected into HEK293T cells according to the manufacturer's instructions. Lentivirus infected cells in media containing polybrene (8 μg/mL).

Cell proliferation assay
Cells were plated on 96-well plates. As the indicated time points, the viability of cells was determined by Cell Counting Kit 8 (Roche) and measured at OD 450 nm with the BioTek Gen5 system (BioTeck, USA). The experiments were repeated three times with the representative experiment shown here.

Western blotting
Cells were lysed in RIPA buffer supplemented with protease inhibitors and subjected to Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis. Proteins were transferred onto polyvinylidene fluoride membranes (ImmobilonP). Blots were blocked and incubated overnight at 4 °C with the primary antibodies listed in Supplementary information, Table S17. Secondary antibodies were incubated and then washes were done. Blots were developed with chemiluminescent reagents from Pierce.

Quantitative real-time PCR (qRT-PCR) and qPCR copy number analysis
QRT-PCR analysis was performed and analyzed as previously described.[9] For copy number analysis, copy number was assessed in paired sequenced tumors and adjacent noncancerous tissues. DNA (10 ng) was amplified by real-time PCR with Power SYBRs Green. *HBB* was used as a diploid control. Data were analyzed via the comparative (delta-Ct) Ct method. The primers are listed in Supplementary information, Table S17.

Animal studies
All animal procedures were reviewed and approved by the Institutional Animal Care and Use Committee of Chinese Academy of Medical Sciences Cancer Hospital. Cells were subcutaneously injected into the flanks of nude mice. Tumor growth was monitored weekly by caliper measurements of the tumor length and width and calculated individually as the formula: volume $= a \times b^2/2$ (a represents length and b represents width).

Experimental statistical analyses
Data are presented as mean ± SD. Statistical analyses were performed with Graphpad Prism 7 software. Student's *t*-test, one-way ANOVA or two-way ANOVA (two-sided) were used. Statistical significance was assessed at $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

Accession codes
FASTQ files will be uploaded to Genome Sequence Archive (GSA) in BIG Data Center (http://bigd.big.ac.cn/gsa), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, with an accession number HRA000021.

ACKNOWLEDGEMENTS

A special thanks to Mr. Yanhong Li (Baidu) for his generous support of this project. We would like to thank Drs. Y. Shimada (Kyoto University), Enmin Li (Medical College of Shantou University) and Dan Su (Zhejiang Cancer Hospital) for kindly providing cell lines. This study was also supported by funding from the National Key R&D Program of China (2016YFC1302100), the CAMS Innovation Fund for Medical Sciences (2016-I2M-1-001, 2019-I2M-1-003), the National Natural Science Foundation of China (81490753, 81330063, 11971039), the fund of "San-ming" Project of Medicine in Shenzhen (No. SZSM201812088), and the Guangdong Basic and Applied Basic Research Foundation (2019B030302012).

AUTHOR CONTRIBUTIONS

Z.L., Yanhong L. and Q.Z. supervised the study, designed the experiments, and edited the manuscript. Y.C. supervised all data analyses and experimental studies, and wrote the manuscript. H. Chen, Y. Zhao and F.H. conceived the experimental studies and performed laboratory analyses. H. Chen interpreted the functional results and reviewed the manuscript. Y. Zhao, F.H., Yang L., X.Z. and X.Y. performed the functional experiments. R.X. oversaw the bioinformatics and statistics analyses, interpreted the results and wrote and reviewed the manuscript. H. Cui, Jiawei W., Y. Zhou, W.Z., Jie W. and B.W. provided medical records and survival data, performed bioinformatics analyses, validation of variations, and statistics analyses. T.Y., E.X., X.L., Y.M., S.G., L.Z., Y.X., R.S., B.Y. and X.C. provided clinical samples and data. T.Y., P.K., Y.M., J.L. and B.S. reviewed the histopathology. T.Y., E.X., P.K., Y.M., Y. Liu., F.W., J.Y., H.L., Y.W. and Y. Zhang prepared microneedle-punctured FFPE tissues. Y. Zhai and C.C. performed the immunohistochemistry analyses.

ADDITIONAL INFORMATION

Supplementary information accompanies this paper at https://doi.org/10.1038/s41422-020-0333-6.

Competing interests: The authors declare no competing interests.

REFERENCES

1. Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
2. Chen, W. et al. Cancer statistics in China, 2015. *CA Cancer J. Clin.* **66**, 115–132 (2016).
3. Lin, Y. et al. Esophageal cancer in high-risk areas of China: research progress and challenges. *Ann. Epidemiol.* **27**, 215–221 (2017).
4. Lin, D. C., Wang, M. R. & Koeffler, H. P. Genomic and epigenomic aberrations in esophageal squamous cell carcinoma and implications for patients. *Gastroenterology* **154**, 374–389 (2018).
5. Gao, Y. B. et al. Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 1097–1102 (2014).
6. Song, Y. et al. Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95 (2014).
7. Lin, D. C. et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* **46**, 467–473 (2014).
8. Zhang, L. et al. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.* **96**, 597–611 (2015).
9. Chang, J. et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. *Nat. Commun.* **8**, 15290 (2017).
10. Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
11. Qin, H. D. et al. Genomic characterization of esophageal squamous cell carcinoma reveals critical genes underlying tumorigenesis and poor prognosis. *Am. J. Hum. Genet.* **98**, 709–727 (2016).
12. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
13. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
14. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
15. Secrier, M. et al. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* **48**, 1131–1141 (2017).
16. Weaver, J. M. J. et al. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nat. Genet.* **46**, 837–843 (2014).
17. Petljak, M. et al. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell* **176**, 1282–1294 (2019).
18. Seplyarskiy, V. B., Andrianova, M. A. & Bazykin, G. A. APOBEC3A/B-induced mutagenesis is responsible for 20% of heritable mutations in the TpCpW context. *Genome Res.* **27**, 175–184 (2017).
19. Law, E. K. et al. The DNA cytosine deaminase APOBEC3B promotes tamoxifen resistance in ER-positive breast cancer. *Sci. Adv.* **2**, e1601737 (2016).
20. Sieuwerts, A. M. et al. Elevated APOBEC3B correlates with poor outcomes for estrogen-receptor-positive breast cancers. *Horm. Cancer* **5**, 405–413 (2014).
21. Chan, T. A. et al. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* **30**, 44–56 (2019).

22. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
23. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* **17**, 128 (2016).
24. Rojo de la Vega, M., Chapman, E. & Zhang, D. D. NRF2 and the hallmarks of cancer. *Cancer Cell* **34**, 21–43 (2018).
25. Shibata, T. et al. Cancer related mutations in NRF2 impair its recognition by Keap1-Cul3 E3 ligase and promote malignancy. *Proc. Natl. Acad. Sci. USA* **105**, 13568–13573 (2008).
26. Shibata, T. et al. NRF2 mutation confers malignant potential and resistance to chemoradiation therapy in advanced esophageal squamous cancer. *Neoplasia* **13**, 864–873 (2011).
27. Totoki, Y. et al. Unique mutation portraits and frequent COL2A1 gene alteration in chondrosarcoma. *Genome Res.* **24**, 1411–1420 (2014).
28. Yu, J. et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut* **64**, 636–645 (2015).
29. Papatriantafyllou, M. Cytokines: IRF7 lost in translation. *Nat. Rev. Immunol.* **13**, 221 (2013).
30. Mishima, Y. et al. The Hbo1-Brd1/Brpf2 complex is responsible for global acetylation of H3K14 and required for fetal liver erythropoiesis. *Blood* **118**, 2443–2453 (2011).
31. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
32. Adriaens, C. et al. p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat. Med.* **22**, 861–868 (2016).
33. Kim, J. et al. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat. Genet.* **50**, 1705–1715 (2018).
34. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* https://doi.org/10.1200/PO.17.00011 (2017).
35. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337 (2018).
36. Deng, N. et al. A comprehensive survey of genomic alterations in gastric cancer reveals systematic patterns of molecular exclusivity and co-occurrence among distinct therapeutic targets. *Gut* **61**, 673–684 (2012).
37. Sawada, G. et al. Genomic landscape of esophageal squamous cell carcinoma in a Japanese population. *Gastroenterology* **150**, 1171–1182 (2016).
38. Gupta, R., Hong, D., Iborra, F., Sarno, S. & Enver, T. NOV (CCN3) functions as a regulator of human hematopoietic stem or progenitor cells. *Science* **316**, 590–593 (2007).
39. Nusse, R. & Clevers, H. Wnt/beta-catenin signaling, disease, and emerging therapeutic modalities. *Cell* **169**, 985–999 (2017).
40. Nicolai Meinshausen, P. B. H. Stability selection. *J. R. Statist. Soc. B* **72**, 417–473 (2010).
41. Green, A. M. et al. Cytosine deaminase APOBEC3A sensitizes leukemia cells to inhibition of the DNA replication checkpoint. *Cancer Res.* **77**, 4579–4588 (2017).
42. Nikkila, J. et al. Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. *Br. J. Cancer* **117**, 113–123 (2017).
43. Mina, M. et al. Conditional selection of genomic alterations dictates cancer evolution and oncogenic dependencies. *Cancer Cell* **32**, 155–168 (2017).
44. Unni, A. M., Lockwood, W. W., Zejnullahu, K., Lee-Lin, S. Q. & Varmus, H. Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. *Elife* **4**, e06907 (2015).
45. Blakely, C. M. et al. Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat. Genet.* **49**, 1693–1704 (2017).
46. Stuhlmiller, T. J. et al. Inhibition of lapatinib-induced kinome reprogramming in ERBB2-positive breast cancer by targeting BET family bromodomains. *Cell Rep.* **11**, 390–404 (2015).
47. Frankell, A. M. et al. The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. *Nat. Genet.* **51**, 506–516 (2019).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows−Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Weber, J. A., Aldana, R., Gallagher, B. D. & Edwards, J. S. Sentieon D. N. A. pipeline for variant detection-Software-only solution, over 20× faster than GATK 3.3 with identical results. *PeerJ PrePrints* https://doi.org/10.7287/peerj.preprints.1672v2 (2016).
50. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
51. Roberts, S. A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
52. 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).
53. Xi, R., Lee, S., Xia, Y., Kim, T. M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **44**, 6274–6286 (2016).
54. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
55. Cheng, C. et al. Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. *Am. J. Hum. Genet.* **98**, 256–274 (2016).
56. Shimada, Y., Imamura, M., Wagata, T., Yamaguchi, N. & Tobe, T. Characterization of 21 newly established esophageal cancer cell lines. *Cancer* **69**, 277–284 (1992).