Research paper

Molecular analysis of Chinese oesophageal squamous cell carcinoma identifies novel subtypes associated with distinct clinical outcomes

Meng Liu[a,d], Haiyin An[b], Yuan Zhang[a], Wei Sun[a], Shujun Cheng[b], Ruozheng Wang[a,c], Xiyan Wang[a,**], Lin Feng[b,**]

[a] Affiliated Tumor Hospital of Xinjiang Medical University, Urumqi 830011, Xinjiang, China
[b] State Key Laboratory of Molecular Oncology, Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100021, China
[c] Key Laboratory of Cancer Immunotherapy and Radiotherapy, Chinese Academy of Medical Sciences, Urumqi 830011, Xinjiang, China
[d] State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, Xinjiang Medical University, Urumqi 830011, Xinjiang, China

ARTICLE INFO

ABSTRACT

Background: Oesophageal squamous cell carcinoma (ESCC) is a highly heterogeneous cancer with a distinct incidence and prognosis. Molecular events driving ESCC subtypes and prognosis have not been established, and little is known regarding Chinese ESCC patients in Xinjiang, China.
Methods: Here, we first integrated the genomic and transcriptomic data of 125 Chinese ESCC patients from Xinjiang Tumor Hospital (Urumqi, China). Two independent datasets of GSE53624 and The Cancer Genome Atlas (TCGA) ESCC were used to confirm the results of this study. DNA mutation and overall survival (OS) were analysed independently in the Chinese ESCC cohorts.
Findings: Genomic analyses revealed a consistent mutation signatures and discordance among mutated genes across the different ESCC cohorts. In addition, transcriptomic profiling identified three Chinese ESCC subtypes associated with clinical and molecular attributes, including patient survival, lymph node status and genetic profile. Moreover, Chinese ESCC subtypes have distinct metabolic, inflammatory, metastatic, and cell proliferation features and unique potential therapeutics. Furthermore, the expression of cell cycle- and/or cell proliferation-related genes was higher in cyclin D1 (CCND1)-amplified tumours than in CCND1-normal tumours from Chinese ESCC patients, suggesting that CCND1 amplification promoted cell proliferation.
Interpretation: Our findings provide a framework to facilitate the rational categorization of ESCC in Chinese patients and a foundation for new therapies.
Funding: This study was supported by the Research Fund of Key Laboratory of Xinjiang oncology (Grant no.2017D04006) and the Outstanding Youth Science and technology training project fund of Xinjiang, China (Grant no. 2017Q058).

1. Introduction

Oesophageal cancer is among the most common malignancies in the world [1], and the 5-year survival rate in Western populations is 12−20% [2]. Histologically, oesophageal squamous cell carcinoma (ESCC) is the major type in Asian populations, while oesophageal adenocarcinoma (EAC) is the predominant type in Western populations [3]. Approximately 477,900 newly diagnosed oesophageal cancer patients and 375,000 disease-related deaths are estimated every year in China [4]. Several studies have used next-generation sequencing technology to examine ESCC tumours [5−9]. However, the molecular events underlying the initiation and progression of ESCC in Chinese populations, especially in the high-risk Chinese Uygur population, remain poorly understood.

Despite advances in diagnostic and treatment methods, the survival rates of ESCC patients remain largely unchanged due to ESCC's heterogeneity and invasive nature [10]. Therefore, the genomic and molecular characterization of ESCC should be clarified, and new strategies for classification and individualized therapy are urgently

** Corresponding author at: Lin Feng, State Key Laboratory of Molecular Oncology, Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100021, China.
E-mail addresses: liuxiaomeng198154@163.com (M. Liu), anhaiyin@163.com (H. An), 18099606776@163.com (Y. Zhang), sunweixinjiang2020@163.com (W. Sun), chengshj@cae.cn (S. Cheng), wrz8526@vip.163.com (R. Wang), wangxiyan2020@163.com (X. Wang), fenglin@cicams.ac.cn (L. Feng).

Research in context

Evidence before this study

Oesophageal squamous cell carcinoma (ESCC) is a highly heterogeneous cancer. The Cancer Genome Atlas (TCGA) facilitated the integrative clustering of ESCC by comprehensively classifying ESCC into three molecular subtypes. However, the molecular classification of ESCC in Chinese populations has not been elucidated. In addition, the genomic signature of Chinese Uygur ESCC cohorts in Xinjiang, China remain poorly understood.

Added value of this study

Our study found that three ESCC subtypes were associated with a distinct prognosis and lymph node metastasis in ESCC patients and may serve as independent risk factors. The three ESCC subtypes have been named the metabolic subtype, the inflammatory and metastatic subtype, and the cell proliferation subtype. We also found that specific molecules that define each subtype, thus serving as subtype-specific biomarkers as well as potential targets for the treatment of ESCC. Specific biomarkers and targets include the GPR98, DDX60 and DDX60L for the metabolic subtype, immune inhibitors for the inflammatory and metastatic subtype, and CDK4/6 inhibitors for the cell proliferation subtype. In addition, we identified the mutated genes, including FAT1, ADGB, DOPEY1, HECW1, LAMA1, THSD7A, USP9Y, PRKAG2, PLCH2, NPIPA5 and MIB2, in Chinese Uygur ESCC that may be potential targets for effective treatment of ESCC.

Implications of all the available evidence

We found that the three subtypes were associated with a distinct prognosis in 125 Chinese ESCC patients from Xinjiang Tumor Hospital (Urumqi, China) and were an independent risk factors. In addition, our three subtypes can be reproduced in other ESCC cohorts by building a linear discriminant analysis (LDA) classifier model. These findings all suggested the potential clinical guidance and application value. In conclusion, our study performed in ethnically diverse Chinese Uygur and Han ESCC cohorts provides further evidence to support substantial biological heterogeneity in ESCC, which may facilitate personalized treatment for ESCC patients.

needed. Recently, TCGA facilitated the integrative clustering of ESCC by comprehensively classifying ESCC into three molecular subtypes, which were associated with geographic differences [11]. However, this study was performed in small cohorts, and the molecular classification of ESCC in Chinese populations has not been elucidated. Moreover, whether such molecular subtypes are correlated with targeted therapy remains unexplored.

To date, several whole-exome/genome sequencing studies of ESCC have been performed in China [6−8, 12, 13]. However, the genomic and transcriptomic profiling of Chinese Uygur ESCC (CU-ESCC) cohorts in Xinjiang, China, has not been reported. Here, we first performed a genomic and transcriptomic analysis of CU-ESCC cohorts from Xinjiang, China, to reveal the mutational landscape of CU-ESCC patients. We used an unsupervised approach to analyse RNA sequencing and whole-exome sequencing (WES) data from 125 ESCC samples in Xinjiang, China, to identify the molecular subtypes and reveal molecular signatures and genomic alterations. Collectively, our results not only provide a genomic resource for CU-ESCC in Xinjiang, China, but also reveal distinct Chinese ESCC subtypes characterized by specific RNA and DNA markers, thereby providing a molecular subtyping framework and candidate pathways that define molecular

subtypes and offering potential targets for the effective treatment of ESCC.

2. Materials and methods

2.1. Specimen collection

Fresh-frozen tumour samples from patients who were not previously treated with chemotherapy or radiation therapy were obtained from the Bank of Tumor Resource, Institute of Oncology, Xinjiang Uygur Autonomous Region in China. This research was approved by the Institutional Review Board of the Affiliated Tumour Hospital of Xinjiang Medical University. The patients provided informed consent for the use of their tumour specimens. Each frozen primary tumour specimen had a corresponding normal tissue specimen (or peripheral blood samples). The tumour tissues and adjacent normal oesophageal tissues stained with haematoxylin and eosin were assessed by two pathologists to identify the tumour purity. Tumour tissues exhibiting >50% tumour cells were analysed, whereas the adjacent normal oesophageal tissues contained no tumour cells. All oesophageal cancers were regarded as squamous, according to the World Health Organization Classification of Tumors of the Digestive System, 4th edition. Overall survival (OS) was defined as the interval between surgery and death.

2.2. DNA/RNA extraction, WES and RNA sequencing

Genomic and transcriptomic analyses of the samples were performed according to the following procedures (Supplementary Fig. 1). DNA and RNA were co-isolated, and quality was tested as described previously [13]. Briefly, genomic DNA was extracted from tumour and matched blood samples using the traditional phenol-chloroform method. DNA was quantified using a Qubit 3.0 fluorometer (Invitrogen) and a NanoDrop 2000 spectrophotometer (Thermo Scientific), and the integrity was assessed with a TapeStation (Agilent Technologies). The fragmented DNA (150-200 bp) was captured using a SureSelect Human All Exon V6 Kit (Agilent Technologies) according to the manufacturer's instructions. DNA libraries with 150-bp paired-end reads were sequenced using an Illumina HiSeq 4000 system.

Total RNA was extracted from fresh-frozen tissues using TRIzol reagent (Invitrogen). RNA integrity was analysed on an Agilent 2100 Bioanalyzer (Agilent Technologies), and only samples with an integrity number >8.0 were used to prepare the transcriptome library using an NEBNext Ultra RNA Library Prep Kit for Illumina according to the manufacturer's instructions. Paired-end libraries were sequenced on an Illumina HiSeq X Ten platform (2×150-bp paired-end read length).

2.3. WES data analysis

2.3.1. Somatic mutation calling and filtering

WES reads after the exclusion of low-quality reads were mapped to the UCSC hg19 reference sequence with Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/) [14]. PCR duplications were removed with the Picard tool (http:// broadinstitute.github.io/picard/) and recalibrated with the BaseRecalibrator tool from the Genome Analysis Toolkit [15] (GATK, https://software. broadinstitute.org/gatk/). Somatic mutations were called using Mutect2 [16], SpeedSeq [17], VarDict [18] and VarScan2 [19] from exome data of tumour and matched blood samples. Low coverage and strand-biased mutations were removed. Any germline mutation, if found in more than one normal sample, was removed from our final list of somatic mutations. To further validate our somatic mutation calls, we selected 350 mutation positions for further validation on the same ESCC samples from Chinese Uygur patients. In total, 330 selected somatic mutations were confirmed by PCR-based Sanger sequencing, resulting in a true-

positive validation rate of 94.3%. Functional annotation of the mutations was performed with Oncotator [20] (https://software.broadinstitute.org/cancer/cga/oncotator) on the RefSeq gene.

2.3.2. Mutational signature analysis

Mutational signatures were displayed with 96-context classification. A nonnegative matrix factorization (NMF) approach was applied to estimate the 96-substitution pattern with 30 known Catalogue of Somatic Mutations in Cancer (COSMIC) cancer signatures [21] (https:// cancer.sanger.ac.uk/cosmic/signatures) and infer their exposure contributions. The most frequently mutated genes within the CU-ESCC (from Xinjiang, China) and TCGA ESCC cohorts were compared.

2.3.3. Exome-based somatic copy number alterations (SCNAs) analysis

SCNAs were inferred by CNVkit [22] (https://cnvkit.readthedocs.io/en/stable/pipeline.html) using the circular binary segmentation algorithm with default parameters. Segment-level ratios were calculated and $\log_2$ transformed. Significant focal SCNAs across all samples were identified by Genomic Identification of Significant Targets in Cancer (GISTIC) [23]. A $\log_2$ ratio cut-off of $\pm$ 0.9 was used to define SCNAs (amplifications and deletions).

2.4. RNA-seq data analysis

Adaptor contamination and polyA and polyC contamination were removed using Cutadapt [24] and Sickle (http://github.com/najoshi/sickle/), and sequencing reads were aligned to the human reference sequence (GRCh37 assembly) using Salmon [25]. Gene expression values were quantitated as transcripts per million reads (TPM) values. The $\log_2$-transformed TPM counts were used for subsequent analysis.

2.5. Correlations between the Chinese ESCC subtypes and clinical features

Detailed clinicopathologic features are summarized in Supplementary Data 3. Associations between the clinical information and Chinese ESCC subtypes were examined using Pearson's chi-square test for categorical data or the Kruskal-Wallis test for continuous data. A log-rank test and Kaplan-Meier survival curves were applied to compare the OS among the Chinese, GSE53624 and TCGA ESCC subtypes. To evaluate the prognostic power of the ESCC subtypes, a Cox proportional hazards regression model adjusted or not adjusted for available prognostic clinical covariates was used to determine hazard ratios (HRs) and 95% confidence intervals (CIs).

2.6. Consensus clustering

mRNA gene sequences were sorted by the aggregate rank of the median absolute deviations (MADs) across all samples, and the top 1500 most variable genes were used for clustering. To identify the Chinese ESCC subtypes, the R package *ConsensusClusterPlus* (number of repetitions=1000 bootstraps, pItem=0.8, pFeature=1) method was used [26]. The average pairwise consensus matrix within consensus clusters, a delta plot of the relative change in the area under the consensus cumulative distribution function (CDF) curve, and the average silhouette distance for the consensus clusters were used to determine the number of clusters. We selected 3 clusters as the optimized solution for the consensus matrix. Then, differentially expressed genes among the three ESCC subtypes were identified by the *limma* package in R (Benjamini-Hochberg false discovery rate (FDR) <0.001), and $\log_2$ fold change (FC) values were computed.

2.7. Validation of the clinical significance of the Chinese ESCC subtypes

To demonstrate the performance of the Chinese ESCC subtypes, we collected publicly available gene expression profiles associated with clinical information from the Gene Expression Omnibus (GEO) database (accession number GSE53624, http://www.ncbi.nlm.nih.gov/geo) and the TCGA ESCC dataset (http://cancergenome.nih.gov/). Prediction of patient class in 2 independent cohorts was performed as described previously [27]. Briefly, gene expression data from the training set (the Chinese ESCC cohort) were combined to form a classifier according to LDA in BRB-Array Tools (http://linus,nci.nih.gov/BRB-ArrayTools.html) software, and the power of the classifier was estimated using a misclassification rate determined during the leave-one-out cross-validation (LOOCV) in the training set. When a classifier was used on the 2 independent validation datasets, its prognostic significance for the three predicted GSE53624 and TCGA ESCC subtypes was analysed by Kaplan-Meier plots and log-rank tests.

2.8. Pathway and Gene Ontology (GO) functional enrichment analyses

We identified the upregulated pathways and GO terms among the different Chinese ESCC subtypes by performing gene set enrichment analysis (GSEA) (http://www.broadinstitute.org/gsea/index.jsp). *P* values less than 0.05 were considered significant. Gene sets were downloaded from the Molecular Signatures Database (MSigDB) v.5.2 (http://software.broadinstitute.org/gsea/msigdb/index.jsp). We included broad hallmarks, GO gene sets and specific pathways of interest from the curated collection of gene sets/canonical pathways, including KEGG, BIOCARTA, REACTOME and PID pathways (Supplementary Data 7.1 and Data 7.2). The normalized enrichment score (NES) reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes. The size of the circle represents the NES. Gene sets with *P* values <0.05 were plotted as a function of the NES, and different-coloured circles represent different *P* values.

2.9. Identification of copy-number alterations (CNAs) associated with mRNA abundance among the Chinese ESCC subtypes

First, we identified SCNAs affecting mRNA abundance in either "cis" (within the same aberrant locus) or "trans" (remote locus) mode, which were visualized with the multiOmicsViz package. Spearman's correlation coefficients and adjusted *P* values were calculated for all CNA-mRNA pairs for 23,109 genes (Supplementary Data 11). Next, to identify specific mutations significantly associated with the Chinese ESCC subtypes, Fisher's exact test was applied to assess differences in somatic mutations across the ESCC subtypes according to the mutation status (mutant versus wild-type). Likewise, significant differences in amplified or deleted genes among the Chinese ESCC subtypes were identified by Pearson's chi-square test. Somatic mutation and CNAs data were analysed and visualized using OncoPrint (https://cbioportal.org).

2.10. Statistical analysis

The normality of the variables was estimated by the Shapiro-Wilk normality test [28]. For comparisons of non-normally distributed continuous variables, the Wilcoxon rank-sum test was used to compare two groups, while the Kruskal-Wallis test was utilized to compare three groups. The Nemenyi post hoc test was used for the pairwise multiple comparisons. Fisher's exact test was employed for comparisons of unordered categorical variables, and Pearson's chi-square test was used for comparisons of ordered categorical variables. Correlation coefficients were computed by Spearman correlation analyses. Survival curves were generated using the Kaplan-Meier method, and the log-rank test was used to determine the

statistical significance of differences. *P* values were corrected to the false discovery rate (FDR) by the Benjamini-Hochberg method. All statistical analyses were performed by R packages version 3.6 (https://www.r-project.org/), and a significance level of 0.05 was used.

2.11. Data availability

Genomic and transcriptomic data have been uploaded to the Genome Sequence Archive of Beijing Institute of Genomics, Chinese Academy of Sciences (http://gsa.big.as.cn/) under accession number HRA000178. Data that support the findings of this study are available from the TCGA (http://cancergenome.nib.gov) and GEO databases (http://www.ncbi.nlm.nih.gov/geo).

3. Results

3.1. Genomic landscape of Chinese Uygur ESCC in Xinjiang, China

We collected tumour tissues, distant normal tissues, peripheral blood samples and clinical information from 125 individuals, including 80 Uygur and 45 Han patients with ESCC, based on stringent criteria (Supplementary Fig. 1). WES of the tumour and blood samples was performed in 50 Uygur ESCC patients. Moreover, mutational signatures, single nucleotide variants (SNVs) and CNAs were analysed.

Consequently, WES identified 36,512 somatic mutations, including 35,355 non-silent point mutations and 1,157 small insertions-deletions (Indels). The mutational landscape of CU-ESCC in Xinjiang, China, is summarized in **Fig. 1a**. Importantly, we found that the significantly mutated genes, including TP53 (88%) and PIK3CA (20%), among the 50 Uygur ESCC patients in Xinjiang, China, were similar to those describes in previous reports [13]. The mutation frequencies of several genes, including FAT1 (18% versus 5.3%), ADGB (10% versus 1%), DOPEY1 (10% versus 1%), HECW1 (12% versus 1%), LAMA1 (10% versus 1%), THSD7A (12% versus 1%), USP9Y (12% versus 1%), PRKAG2 (10% versus 0%), PLCH2 (10% versus 0%), RBPJ (8% versus 0%), NPIPA5 (8% versus 0%) and MIB2 (8% versus 0%), were relatively higher in the CU-ESCC cohorts than in the ESCC cohorts from the TCGA, while the mutation frequency of DNAH5 (4% versus 17.9%) was slightly lower (**Fig. 1d** and Supplementary Data 1). Our comparative analysis suggested that these significantly mutated genes in CU-ESCC may be new candidate genes. We further determined the number of SNVs in the matrix of 96 mutations occurring as a trinucleotide in each CU-ESCC sample and found that the predominant mutational spectra were C>T and C>G transitions (**Fig. 1b**), which is consistent with the previous studies [13]. From the 50 CU-ESCC samples, we identified three mutational signatures with high similarity to COSMIC [21] signatures 1, 13 and 4 (cosine similarity 0.902, 0.894 and 0.833, respectively) (**Fig. 1c**). COSMIC signature 1 has been identified as an age-dependent mutational signature. COSMIC signature 13 is believed to
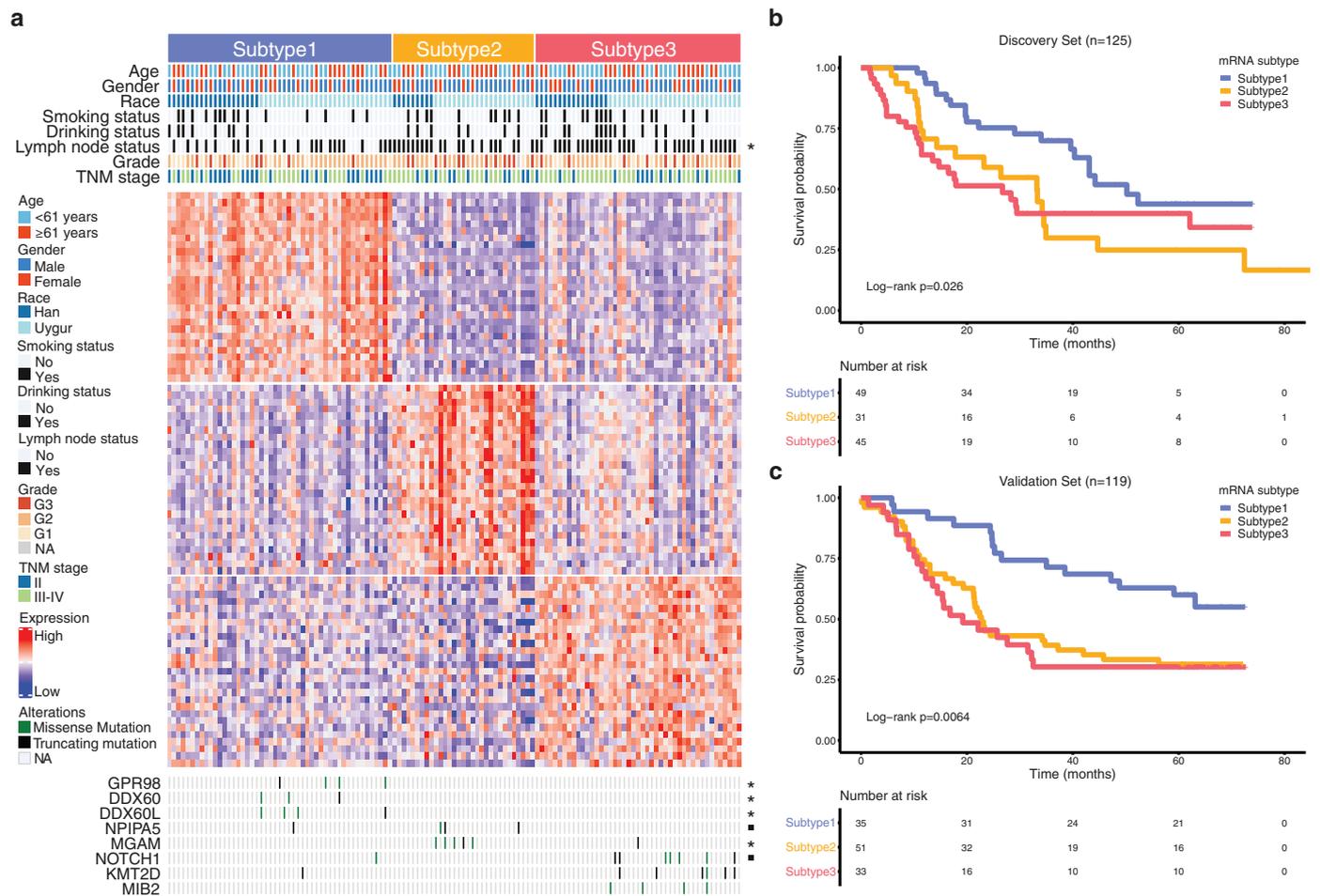
Fig. 2. Transcriptomic subtypes of ESCC in Chinese populations. (a) Heatmap of unsupervised consensus clusters of RNA sequencing gene expression data. The tumours were divided into three clusters by the consensus clustering method. One hundred twenty-five tumours are displayed as columns, and eighty-one differentially expressed genes are displayed as rows. The top tracks show important clinical variables. The bottom tracks represent the eight most commonly mutated genes in CU-ESCC patients. Age was tested by the Kruskal-Wallis test. Gender, race, smoking status, drinking status, lymph node status, tumour grade and TNM status were tested by Pearson's chi-square test. Somatic mutations were tested by Fisher's exact test. Significant associations between covariates and subtypes are indicated by □*P*<0.1, **P<0.05. **(b), (c)** Kaplan−Meier curves for OS based on mRNA subtypes in the Chinese ESCC and GSE53624 ESCC cohorts (the log-rank test was used to determine the statistical significance of differences).
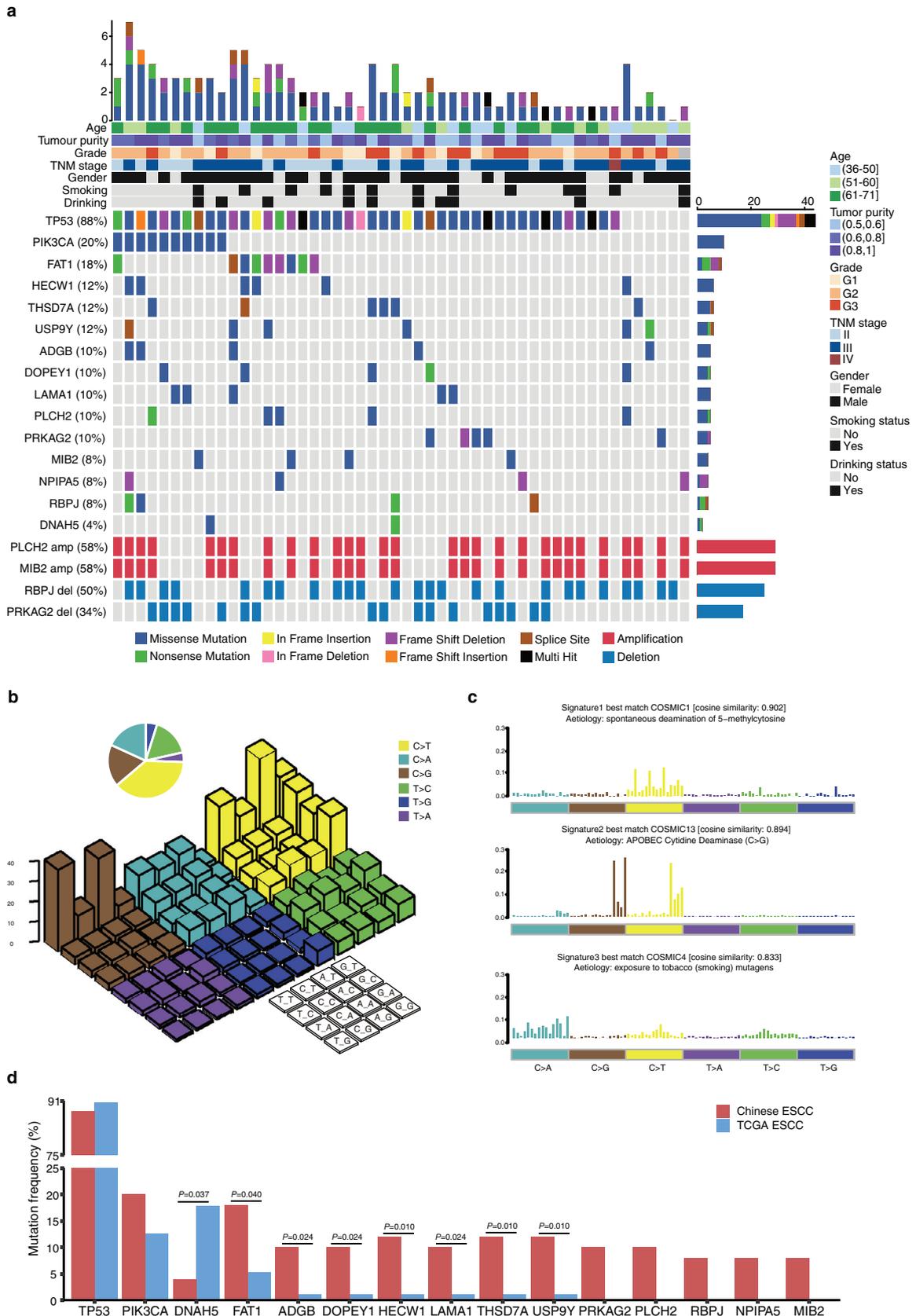
Fig. 1. The genomic landscape and mutational signatures of CU-ESCC in Xinjiang, China. (a) Genetic profiles and associated clinical features of 50 CU-ESCC patients. (b) Lego plots of mutation frequencies calculated from 50 CU-ESCC samples. Base substitutions were divided into 96 patterns based on the mutation type and nucleotides flanking the mutated base. The height of the bar represents the proportion of each substitution pattern. Insert pie charts show the distribution of all mutations for a given middle mutated base across the territory being evaluated. (c) Mutational signatures of CU-ESCC in Xinjiang, China. Each signature is displayed according to the 96 substitution classifications. The vertical axis represents the mutation fractions of each substitution classification. (d) Comparison of the mutated genes between the CU-ESCC (from Xinjiang, China) and TCGA ESCC cohorts.

be related to the over-activity of APOBEC [29], a member of the cytidine deaminases family. COSMIC signature 4 is associated with smoking and has been frequently reported in lung cancer patients [21]. Overall, these results highlight that the APOBEC signature may be a potential oncogenic pathway underlying mutational mechanisms in CU-ESCC development in Xinjiang, China.

3.2. Identification of molecular subtypes by transcriptome analysis

We employed unsupervised clustering based on the 1,500 most variable genes and identified three subtypes among the 125 ESCC tumours from Chinese patients by the CDF, delta area, and average silhouette (Supplementary Fig. 2). Principal component analysis (PCA) demonstrated a distinction among the transcriptomes of tumour samples subjected to consensus clustering, further highlighting the high heterogeneity among tumour samples. Next, we compared transcriptomic profiles among the three Chinese ESCC subtypes with the *limma* package in R, revealing 81 differentially expressed genes in each Chinese ESCC subtype (Supplementary Data 2). In addition, a heatmap was generated using the *Complexheatmap* package in R (**Fig. 2a**). To validate the robustness of the 3 molecular subtypes, we examined the gene expression profiles associated with each cluster in two independent datasets: an independent cohort of 119 tumours (GSE53624) and a TCGA ESCC dataset of 90 tumours. Eighty-one differentially expressed genes were used to build an LDA classifier to predict a cluster label for each sample. We found concordant gene expression patterns for all 3 clusters across the training and validation sets (Supplementary Fig. 3). These observations indicate that the molecular subtypes identified in our study are reliable and stable.

3.3. Clinical and molecular relevance of the Chinese ESCC subtypes

Patients with the three Chinese ESCC subtypes also showed distinct clinical characteristics (Table 1). Specifically, the lymph node status (*P*=0.037) and tumour grade (*P*=0.061) differed among the three Chinese ESCC subtypes (Pearson's chi-square test). In addition, alcohol consumption (31.1%) and TNM III/IV tumours (68.9%) were more frequent among patients with ESCC subtype 3 than among patients with the other subtypes. Survival analysis showed that ESCC subtype 1 was associated with the best OS, followed by subtype 2, and subtype 3 (log-rank test, *P*= 0.013, **Fig. 2b**). Associations between the Chinese ESCC subtypes (between subtype 1 and subtypes 2/3) remained significant in the univariable (Cox *P* = 0.027, hazard ratio (HR) = 2.03 and Cox *P*=0.016, HR=2.05, respectively) and multivariable (Cox *P* = 0.054, HR = 1.88 and Cox *P*=0.043, HR=1.87, respectively) analyses after adjusting for several covariates listed in Supplementary Data 4.1. Furthermore, the clustering of 2 independent ESCC cohorts also resulted in 3 subtypes with similar survival differences (**Fig. 2c**). Consistent with our ESCC subtypes, analysis of OS (log-rank test *P* = 0.0064 for GSE53624 and log-rank test *P*=0.1 for TCGA ESCC) and Cox analysis of the GSE53624 and TCGA ESCC cohorts confirmed the differences observed (Supplementary Data 4.2 and 4.3) among the three ESCC subtypes, demonstrating that our molecular subtypes can be reproduced in other ESCC cohorts.

Each of the three Chinese ESCC subtypes had unique molecular features. ESCC subtype 1 was more likely to have GPR98, DDX60 and DDX60L mutations (*P* =0.014, *P* =0.047 and *P* =0.014, respectively, Fisher's exact test). ESCC subtype 2 was more likely to have NPIPA5 and MGAM mutations (*P* =0.091 and *P* =0.018, respectively, Fisher's exact test). ESCC subtype 3 was more likely to have NOTCH1, KMT2D and MIB2 mutations (*P* =0.090, *P* =0.180 and *P* =0.173, respectively, Fisher's exact test) (Supplementary Data 5). Taken together, these

data indicate molecular heterogeneity within each Chinese ESCC subtype.

3.4. Chinese ESCC subtype-specific enrichment of molecular pathways and GO functions

We found that different Chinese ESCC subtypes showed distinct sets of molecular pathways and GO terms by GSEA. Chinese ESCC subtype 1 was characterized by the highest level of metabolism-related pathways, including retinol metabolism, glycosphingolipid biosynthesis (lacto) and drug metabolism cytochrome p450 signalling pathways (**Fig. 3a** and Supplementary Data 6.1). GO functional enrichment also showed that subtype 1 was enriched in retinol metabolism process and drug metabolism processes (**Fig. 3b** and Supplementary Data 6.2). In addition, ALDH3A2, ADH7, CYP2C18 and CYP3A5 are highly expressed in this subtype (Supplementary Fig. 4a-c, Supplementary Fig. 5 and Supplementary Data 9). Therefore, we termed Chinese ESCC subtype 1 the metabolic subtype.

GSEA revealed that the epithelial-to-mesenchymal transition (EMT) was the most significantly activated pathway and GO term in Chinese ESCC subtype 2 (**Fig. 3a,b** and Supplementary Data 6.1 and Data 6.2), as evidenced by the increased expression of genes directly regulated by EMT markers [30], such as VIM, ZEB1, CDH2 and CDH11, whereas epithelial cell markers such as PPL and JUP were significantly downregulated in subtype 2 (Supplementary Fig. 4D-I). In addition, ESCC subtype 2 exhibited upregulated cell adhesion molecules and integrin and extracellular matrix (ECM) receptor interaction pathways. The transforming growth factor β (TGF-β) pathway was also activated in subtype 2 (**Fig. 4** and Supplementary Data 8), providing further evidence that Chinese ESCC subtype 2 tumours lose epithelial characteristics and acquired a mesenchymal phenotype. Additionally, the expression of immune inhibitors, particularly B7-H3, HVEM, VEGFB, EDNRB, LAG3, CTLA-4, PD-L2 and PD1, was significantly upregulated in ESCC subtype 2 (Supplementary Fig. 6 and Supplementary Data 10), providing additional rationale for the use of immune checkpoint blockade as a therapeutic approach. Given this finding, we termed Chinese ESCC subtype 2 the inflammation and cell metastasis subtype.

Chinese ESCC subtype 3 was characterized by both an increase in cell-cycle related pathways (such as G2M, cell cycle, and MYC) and the GO biological process of ubiquitin-mediated proteolysis (**Fig. 3a, b**). Furthermore, we found that cell cycle- and/or cell proliferation-related genes, such as AURKA, NCAPG, PLK1, PCNA, and CDK1, were also upregulated in Chinese ESCC subtype 3 (Supplementary Fig. 4k-m, Supplementary Fig. 5 and Supplementary Data 9). Consistent with these results, WNT pathway activation, which may promote tumour cell growth and proliferation by upregulating these genes, was identified in Chinese ESCC subtype 3 (**Fig. 4** and Supplementary Data 8). Thus, we termed Chinese ESCC subtype 3 the cell proliferation subtype. Overall, these results further confirmed that molecular diversity existed in each Chinese ESCC subtype.

3.5. Cell cycle-related genes and signalling pathways

Aberrant activation of the TGF-β and WNT pathways has been observed in Chinese ESCC. As reported previously [31], activation of the WNT signalling pathway and inactivation of the TGF-β signalling pathway result in the activation of MYC, promoting cell proliferation. In accordance with this finding, activation of the WNT pathway and the upregulation of cell cycle regulators were observed in Chinese ESCC subtype 3 (**Fig. 4** and Supplementary Data 8), indicating that activation of the WNT pathway exerted important functions in cell cycle regulation. Together, these observations suggest that enhanced

---

Mutated genes are listed on the horizontal axis. The vertical axis represents the mutation frequencies obtained from different cohorts. Somatic mutations were tested by Fisher's exact test.

Table 1
Clinical characteristics of patients with ESCC in discovery cohorts and validation cohorts.

Characteristics	Discovery Set (n=125)				P value	Validation Set (n=119)			
	Subtype1 (N=49)	Subtype2 (N=31)	Subtype3 (N=45)	Total (N=125)		Subtype1 (N=35)	Subtype2 (N=51)	Subtype3 (N=33)	Total (N=119)
Age, years					0.403 [a]				
Median	62	58	61	61		59	60	57	59
Range	36−77	42−77	39−76	36−77		41−76	36−76	39−82	36−82
Gender					0.795				
Female	17 (34.7%)	9 (29.0%)	13 (28.9%)	39 (31.2%)		4 (11.4%)	15 (29.4%)	2 (6.1%)	21 (17.6%)
Male	32 (65.3%)	22 (71.0%)	32 (71.1%)	86 (68.8%)		31 (88.6%)	36 (70.6%)	31 (93.9%)	98 (82.4%)
Race					0.563				
Chinese Han	20 (44.4%)	9 (20.0%)	16 (35.6%)	45 (100%)		35 (29.4%)	51 (42.9%)	33 (27.7%)	119 (100%)
Chinese Uyghur	29 (36.3%)	22 (27.5%)	29 (36.2%)	80 (100%)		-	-	-	0
Smoking status					0.876				
Non-smoker	34 (69.4%)	21 (67.7%)	29 (64.4%)	84 (67.2%)		13 (37.1%)	19 (37.3%)	7 (21.2%)	39 (32.8%)
Smoker	15 (30.6%)	10 (32.3%)	16 (35.6%)	41 (32.8%)		22 (62.9%)	32 (62.7%)	26 (78.8%)	80 (67.2%)
Drinking status					0.236				
Non-drinker	41 (83.7%)	24 (77.4%)	31 (68.9%)	96 (76.8%)		13 (37.1%)	27 (52.9%)	5 (15.2%)	45 (37.8%)
Drinker	8 (16.3%)	7 (22.6%)	14 (31.1%)	29 (23.2%)		22 (62.9%)	24 (47.1%)	28 (84.8%)	74 (62.2%)
Tumor location					0.553				
Upper	8 (16.3%)	3 (9.7%)	8 (17.8%)	19 (15.2%)		6 (17.1%)	4 (7.8%)	4 (12.1%)	14 (11.8%)
Middle	28 (57.1%)	19 (61.3%)	30 (66.7%)	77 (61.6%)		19 (54.3%)	33 (64.7%)	17 (51.5%)	69 (58.0%)
Lower	13 (26.5%)	9 (29.0%)	7 (15.6%)	29 (23.2%)		10 (28.6%)	14 (27.5%)	12 (36.4%)	36 (30.3%)
Tumor grade					0.061				
G1	11 (22.4%)	2 (6.7%)	10 (22.2%)	23 (18.5%)		7 (20.0%)	10 (19.6%)	6 (18.2%)	23 (19.3%)
G2	31 (63.3%)	16 (53.3%)	26 (57.8%)	73 (58.9%)		20 (57.1%)	25 (49.0%)	19 (57.6%)	64 (53.8%)
G3	7 (14.3%)	12 (40.0%)	9 (20.0%)	28 (22.6%)		8 (22.9%)	16 (31.4%)	8 (24.2%)	32 (26.9%)
T stage					0.374				
T1	2 (4.1%)	0 (0.0%)	1 (2.2%)	3 (2.4%)		6 (17.1%)	1 (2.0%)	1 (3.0%)	8 (6.7%)
T2	9 (18.4%)	2 (6.5%)	8 (17.8%)	19 (15.2%)		5 (14.3%)	8 (15.7%)	7 (21.2%)	20 (16.8%)
T3	33 (67.3%)	28 (90.3%)	31 (68.9%)	92 (73.6%)		18 (51.4%)	29 (56.9%)	15 (45.5%)	62 (52.1%)
T4	5 (10.2%)	1 (3.2%)	5 (11.1%)	11 (8.8%)		6 (17.1%)	13 (25.5%)	10 (30.3%)	29 (24.4%)
N stage					0.054				
N0	26 (53.1%)	10 (32.3%)	13 (28.9%)	49 (39.2%)		17 (48.6%)	22 (43.1%)	15 (45.5%)	54 (45.4%)
N1	10 (20.4%)	15 (48.4%)	20 (44.4%)	45 (36.0%)		12 (34.3%)	16 (31.4%)	14 (42.4%)	42 (35.3%)
N2	9 (18.4%)	4 (12.9%)	11 (24.4%)	24 (19.2%)		3 (8.6%)	8 (15.7%)	2 (6.1%)	13 (10.9%)
N3	4 (8.2%)	2 (6.5%)	1 (2.2%)	7 (5.6%)		3 (8.6%)	5 (9.8%)	2 (6.1%)	10 (8.4%)
TNM stage					0.121				
I-II	25 (51.0%)	11 (35.5%)	14 (31.1%)	50 (40.0%)		20 (57.1%)	19 (37.3%)	14 (42.4%)	53 (44.5%)
III-IV	24 (49.0%)	20 (64.5%)	31 (68.9%)	75 (60.0%)		15 (42.9%)	32 (62.7%)	19 (57.6%)	66 (55.5%)
Lymph Node status					0.037				
N0	26 (53.1%)	10 (32.3%)	13 (28.9%)	49 (39.2%)		17 (48.6%)	22 (43.1%)	15 (45.5%)	54 (45.4%)
N1,2,3	23 (46.9%)	21 (67.7%)	32 (71.1%)	76 (60.8%)		18 (51.4%)	29 (56.9%)	18 (54.5%)	65 (54.6%)
Survival, months					0.012 [b]				
Median	34.4	22.6	16.5	26.3		61.0	22.7	19.2	32.2
Range	6.7−73.9	5.7−87.3	1.7−73.8	1.7−87.3		5.8−72.5	0.1−72.1	1.4−72.6	0.1−72.6

Data are shown as n (%). *P* values are calculated by Person's Chi-squared test.
[a] Kruskal Wallis rank sum test was used.
[b] Log rank test was used.

cellular proliferation in Chinese ESCC subtype 3 may further promote cancer progression and growth.

3.6. Effects of CNAs on cell cycle-related genes

SCNAs based on WES data showed the most frequent gains in chromosomes 8q and 3q and losses in chromosomes 9p, 18q, 5q and 11q, similar to previous findings [13]. Likewise, CNAs may affect mRNA abundance in either "cis" or "trans" modes, corresponding to the diagonal and vertical patterns, respectively (Supplement Fig. 7).

To determine oncogenic events across the different ESCC subtypes, we investigated the frequencies of SNVs, indels, and CNAs and the mRNA expression of cell cycle-related genes. CCND1 amplifications were frequent in Chinese ESCC subtype 3 tumours compared with the other subtypes (77% versus 69%, respectively). Transcriptomic analyses also showed increased mRNA expression levels of CCND1 in Chinese ESCC subtype 3 tumours (**Fig. 5a**). Specifically, CCND1 amplification and overexpression were associated with lymph node metastasis and a poor prognosis in ESCC [32].

Although the frequencies of mutations in cell cycle-related genes were not significantly different across subtypes, we found a significant increase in cell cycle-related gene expression in samples with CCND1 amplification (Wilcoxon rank sum test *P*=0.0049; **Fig. 5b**). A

strong correlation was found between the expression of cell cycle-related genes and CCND1 in Chinese ESCC subtype 3 (Spearman's correlation ρ=0.71, *P*=0.0002). In conclusion, these findings reveal that CCND1 copy number gains likely contribute to the mRNA expression upregulation of cell cycle- and/or cell proliferation-related genes in Chinese ESCC subtype 3 via possible cis and trans effects, thus facilitating ESCC progression.

4. Discussion

In our study, we first provided a genomic landscape of CU-ESCC patients in Xinjiang, China, and revealed that the mutational signatures of CU-ESCC in Xinjiang, China, are similar to those reported in other studies. Interestingly, we found that the mutated genes in CU-ESCC, including TP53, PIK3CA, NOTCH1, FAT1, FAT2 and KMT2D, were similar to those in Chinese Han ESCC (CH-ESCC). Additionally, the predominant mutational spectra of CU-ESCC were C>T and C>G transitions, consistent with those of CH-ESCC. The CNAs' profiles were similar between CU-ESCC and CH-ESCC, including losses at 9p, 5q, and 4q and gains at 8q, 3q, 5p, and 7p. In addition, we found that the RNA expression profiles in CU-ESCC were also similar to those in CH-ESCC in terms of ECM regulators, cell adhesion, cell differentiation and epithelial cell proliferation (Supplementary Data 12).
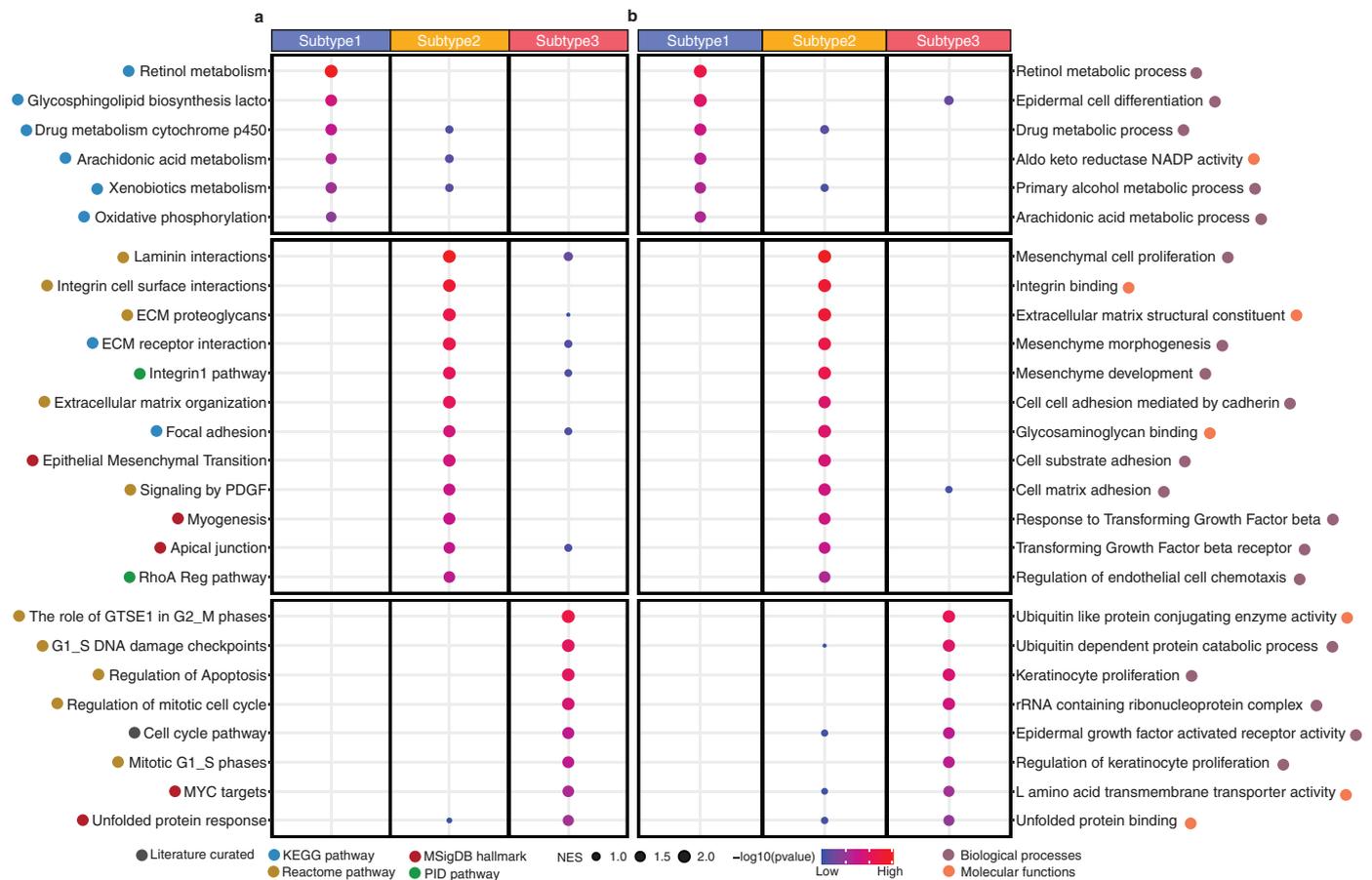
Fig. 3. Pathway and GO term enrichment in the Chinese ESCC subtypes. (a), Pathways and, **(b),** GO terms of different Chinese ESCC subtypes are grouped according to broad biological processes. Each run was performed with 1000 permutations. The x-axis represents the three Chinese ESCC subtypes, while the y-axis represents the pathways (left) and/or GO terms (right). The size of the circle represents the NES (GSEA enrichment score). The colour of the circle represents the P value (P value provided by GSEA <0.05).

Although a molecular characterization study has been conducted in ESCC patients [33], clinically relevant subtypes with molecular heterogeneity that can be used in preclinical and clinical research have not been reported. Therefore, we first used transcriptomic clustering to identify three distinct subtypes associated with unique survival outcomes, molecular features, genomic alterations and personalized treatments. Subsequently, we validated the Chinese ESCC subtypes and their prognostic significance in GSE53624 and TCGA ESCC cohorts. Genomic data also revealed subtype-specific gene mutations, including mutations in GPR98, DDX60, DDX60L, MGAM, NPIPA5, NOTCH1 and KMT2D mutation. Consequently, we performed an integrated analysis and revealed that the molecular classification of ESCC in Chinese patients was associated not only with distinct signatures of activated key signalling pathways and genomic alterations, but also with prognoses and therapeutic strategies across the Chinese ESCC subtypes.

Our study suggested that the upregulated genes and pathways associated with Chinese ESCC subtype 1 tumours were mainly cellular metabolism enzymes and regulators, indicating that metabolic activation is associated with Chinese ESCC subtype 1. Notably, GPR98 [34], DDX60 [35] and DDX60L were more frequently mutated in Chinese ESCC subtype 1 than that in the other subtypes, indicating that they may serve as potential targets in molecular-based therapies.

We found that Chinese ESCC subtype 2 displayed characteristics of inflammation and cell metastasis associated with high immune cell signalling and cytokine signalling gene expression. Three reasons may support the characteristics of Chinese ESCC subtype 2. First, the global gene expression patterns of Chinese ESCC subtype 2 tumours were highly similar to those associated with EMT. Second, adhesion

and chemokine pathways and GO biological processes were highly active in Chinese ESCC subtype 2 tumours. Third, mRNA expression levels of immune checkpoint inhibitor genes were significantly increased in Chinese ESCC subtype 2 tumours. Notably, we further identified novel genes, and MGAM mutation was significantly associated with an increased response and higher PD-L1 expression, indicating that MGAM may be an important component of the immunogenetic landscape and contribute to immune cell infiltration [36]. Overall, the high expression of immune-related signatures reveals that patients with Chinese ESCC subtype 2 tumours might potentially benefit from immune checkpoint inhibitors.

Furthermore, Chinese ESCC subtype 3 was characterized by upregulation of cell cycle- and/or cell proliferation-related genes, MYC oncogene activation and ubiquitin-mediated proteolysis, all of which are defining properties of highly proliferating tumours. We also identified frequently mutated genes, such as KMT2D, NOTCH1, MIB2 and E2F1, in Chinese ESCC subtype 3 tumours. KMT2D mutations are predicted to result in protein products lacking the key methyltransferase domain, supporting a tumour suppressor role for KMT2D in ESCC [37]. NOTCH1 mutations in ESCC are thought to result in a loss of function, supporting the tumour suppressive roles of the NOTCH pathway in squamous cell carcinoma [38]. NOTCH1 has been reported to be a marker of poor survival marker in human ESCC [39].

The cell cycle regulation pathway is frequently altered in ESCC through the deletion of CDKN2A or amplification of CCND1, thus promoting proliferation [40], which is consistent with the observations in our cohort. Our study revealed that ESCC tumours with CCND1 amplifications exhibited high expression of cell cycle-related genes, indicating that CCND1 amplification promotes proliferation, and that
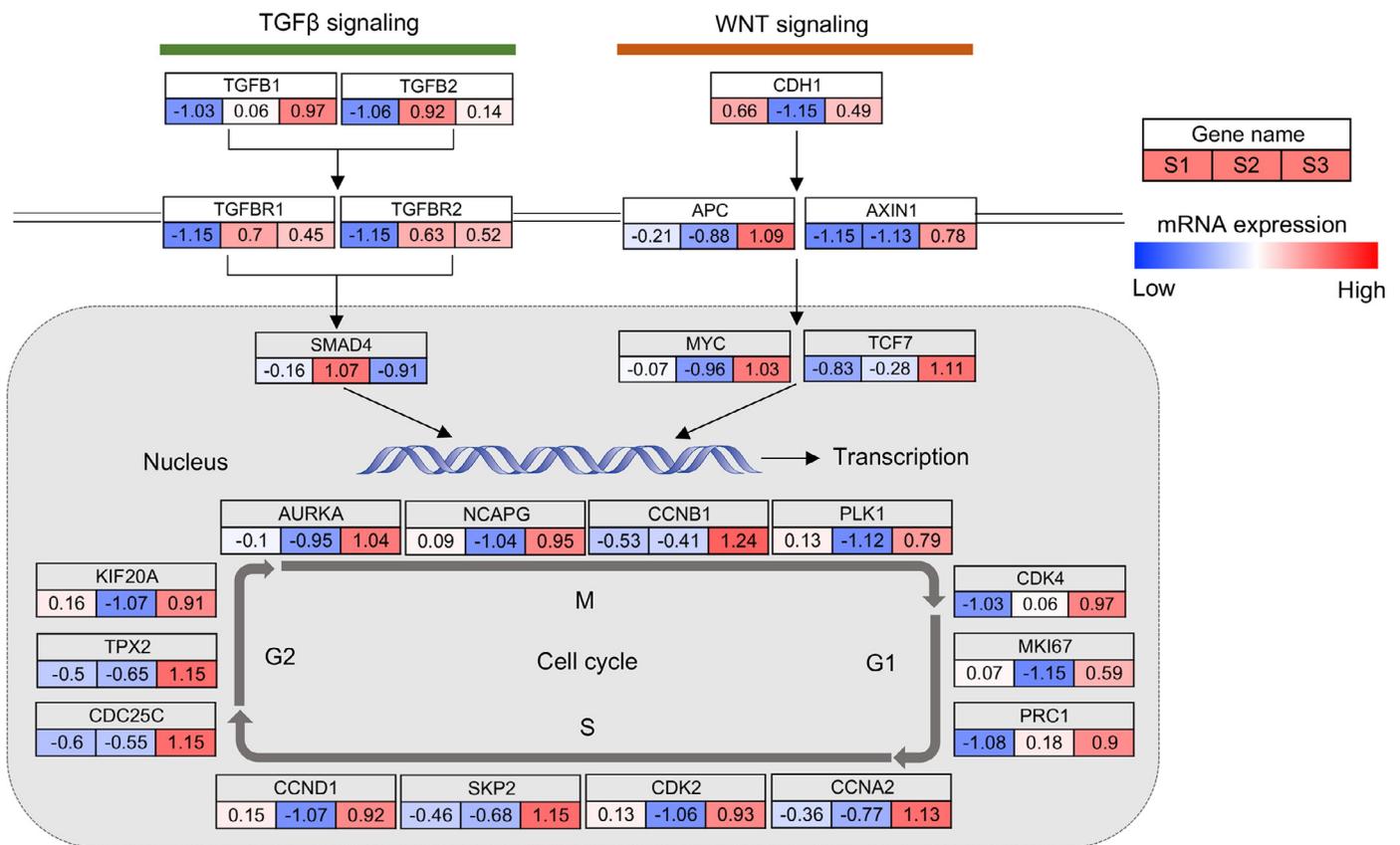
Fig. 4. Overview of cell cycle and signalling pathway cascades in the Chinese ESCC subtypes. Cell cycle-related genes and the TGFβ and WNT signalling pathways were altered based on the transcriptomic analysis. The mRNA abundance of tumours in the three Chinese ESCC subtypes is indicated. Red indicates upregulated genes, and blue indicates down-regulated genes.

these tumours are vulnerable to pharmacologic CCND1 inhibition. Accordingly, patients diagnosed with Chinese ESCC subtype 3 tumours may be potential candidates for treatment with CDK4/6 inhibitors [41] or other cell cycle-related inhibitors [42]. We also found that the expression levels of aldehyde dehydrogenase 2 (ALDH2) were significantly downregulated in ESCC subtype 3 compared with the other subtypes (Supplementary Data 13). According to the reports, acetaldehyde is detoxified to acetic acid by ALDH2, and exposure of human normal fibroblasts to acetaldehyde contributes to the development of ESCC [13, 43]. These findings might account for the poor prognosis and ESCC progression in patients with Chinese ESCC subtype 3.

We compared differences and similarities between Chinese ESCC subtypes and TCGA ESCC subtypes. The similarities showed that leukocyte infiltration and immunomodulatory molecules were enriched in both Chinese and TCGA ESCC subtype 2. In addition, the mutation frequencies of TP53 and PIK3CA were similarity in the Chinese ESCC and TCGA ESCC subtypes (88% versus 90.5%; 20% versus 12.6%, respectively). However, we found several differences, such as differences in ESCC cohorts, molecular pathways, driver genes and prognostic associations. TCGA ESCC subtypes demonstrated trends for geographic associations, while Chinese ESCC subtypes were associated with clinical prognosis and lymph node status. TCGA ESCC subtype 3 demonstrated no evidence for genetic deregulation of the cell cycle, while Chinese ESCC subtype 3 showed alterations and activation of cell cycle-related genes and signalling pathways. NFE2L2 mutations and TP63 amplification were highly associated with TCGA ESCC subtype 1 in the Vietnamese cohorts, but NFE2L2 mutations (24% versus 6%, P=0.04, Fisher's exact test) and TP63 amplification (60% versus 0%) were infrequent in the Chinese ESCC subtypes. In this study, we first showed the genomic landscape in 50 CU-ESCC patients

in Xinjiang, China, by WES profiling. Importantly, we identified three Chinese ESCC subtypes using transcriptomic data, which were more likely to demonstrate the cellular phenotypes. Furthermore, we employed clinical and follow-up data as well as genomic data and provided a Chinese ESCC subtypes' framework that can potentially be used in preclinical and clinical studies of ESCC.

Our study has several limitations. First, large ESCC cohorts from different medical centres and different geographic regions in China should be included to help reduce the possibility of selection bias. Second, mRNA expression analysis was performed by sampling only small peripheral portions of the tumours. Therefore, our results cannot account for the possibility of intra-tumoural heterogeneity in ESCC. Another potential limitation is related to the cost of using ESCC molecular signatures in practice, which is expensive. Multiplexed analyses such as Mammaprint or PAM50 [44] can be developed and applied to ESCC in the future. Additionally, immunohistochemistry and RNA-*in situ* hybridization methods can also be used to reveal the ESCC subtypes.

In summary, our current work in ethnically diverse Chinese Uygur and Han ESCC cohorts provides further evidence to support the generalizability of genomic signatures for the clinical molecular classification of ESCC. Moreover, the generalizability of these signatures will support their use in multicentre efforts aiming to develop targeted therapies for ESCC. Furthermore, we identified three novel molecular subtypes of ESCC in Chinese populations associated with distinct molecular features and survival outcomes, demonstrating substantial biological heterogeneity in ESCC. Therefore, the recognition of these molecular classifications provides a basis for detecting potential biomarkers or therapeutic strategies for specific subtypes in preclinical trials, which will ultimately contribute to personalized treatment for ESCC patients.
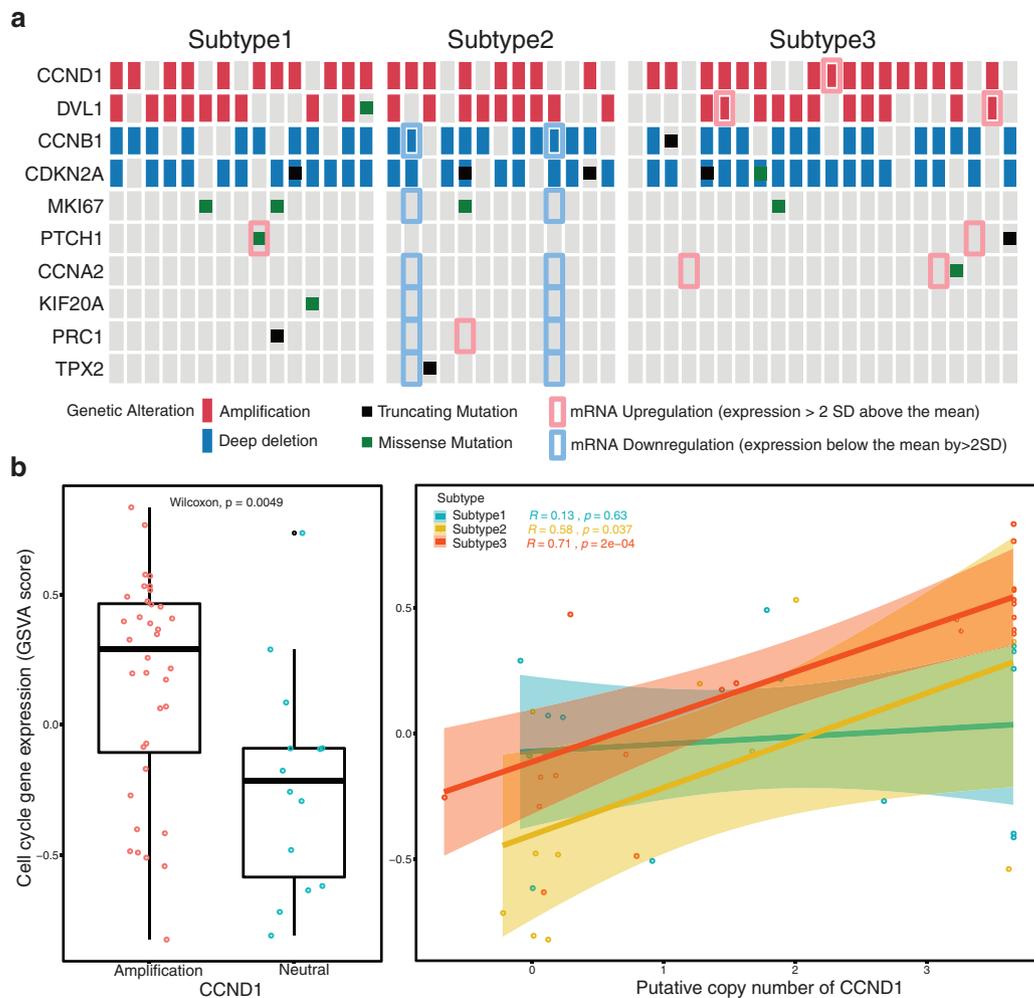
Fig. 5. Alterations of cell cycle-related genes in Chinese ESCC patients. (a) Specific alterations (mutations, SCNAs, and/or relative changes in the mRNA level) in each cell cycle-related genes across the Chinese ESCC subtypes. **(b)** Box plot illustrating the median expression levels of cell cycle-related genes in CCND1 copy number amplification and neutral samples (left); the Wilcoxon rank-sum test was used to compare two groups. Scatter plot depicting the correlations between the median expression levels of cell cycle-related genes and CCND1 copy number amplification (right). Correlation coefficients between CNAs and the median expression levels of cell cycle-related genes among the three subtypes were calculated by Spearman's correlation test.

Declaration of Competing Interest

The authors declare no conflict of interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ebiom.2020.102831.

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 2018;68 (6):394–424.

[2] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: Cancer J Clin 2016;66 (1):7–30.

[3] Arnold M, et al. Global incidence of oesophageal cancer by histological subtype in 2012. Gut 2015;64(3):381–7.

[4] Chen W, et al. Cancer statistics in China, 2015. CA: A Cancer J Clin 2016;66 (2):115–32.

[5] Song Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. Nature 2014;509(7498):91–5.

[6] Gao YB, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet 2014;46(10):1097–102.

[7] Lin DC, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. Nat Genet 2014;46(5):467–73.

[8] Qin HD, et al. Genomic characterization of esophageal squamous cell carcinoma reveals critical genes underlying tumorigenesis and poor prognosis. Am J Hum Genet 2016;98(4):709–27.

[9] Zhang L, et al. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. Am J Hum Genet 2015;96 (4):597–611.

[10] Hao J-J, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. Nat Genet 2016;48(12):1500–7.

[11] Kim JB, R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, Shih J, Cibulskis C, Dunford A, Meier SR, Kim J, Raphael BJ, Wu HT, Wong AM, Willis JE, Bass AJ, Derks S, Garman K, McCall SJ, Wiznerowicz M, Pantazi A, Parfenov M, Thorsson V. Integrated genomic characterization of oesophageal carcinoma. Nature 2017;541 (7636):169–75.

[12] Song Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. Nature 2014;509(7498):91–5.

[13] Chang J, et al. Genomic analysis of oesophageal squamous-cell carcinoma identifies alcohol drinking-related mutation signature and genomic alterations. Nat Commun 2017;8(1).

[14] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–60.

[15] DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491–8.

[16] Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31(3):213–9.

[17] Chiang C, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat Methods 2015;12(10):966–8.

[18] Lai Z, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucl Acids Res 2016;44(11):e108.

[19] Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22(3):568–76.

[20] Ramos Alex H, Manaswi Gupta LL, Lawrence Michael S, Pugh Trevor J. Gordon saksena, matthew meyerson and gad getz, oncotator- cancer variant annotation tool. Hum Mutat 2015;36(4):E2423–9.

[21] Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–21.

[22] Talevich E, et al. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol 2016;12(4):e1004873.

[23] Mermel Craig H, Barbara Hill SES, Meyerson Matthew L, Beroukhim Rameen, Getz Gad. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011;12: R41.

[24] Martin M. Cutadapt removes adapte sequences from high-throughput sequencing reads. EMBnet J 2011;17:10–2.

[25] Patro R, et al. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 2017;14(4):417–9.

[26] Burstein MD, et al. Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin Cancer Res 2014;21(7):1688–98.

[27] Lee J-S, et al. Expression signature ofE2F1and Its associated genes predict superficial to invasive progression of bladder tumors. J Clin Oncol 2010;28(16):2660–7.

[28] Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10(2):486–9.

[29] Swanton C, et al. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov 2015;5(7):704–12.

[30] de Wever O, et al. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. PLoS ONE 2012;7(12).

[31] Cancer Genome Atlas, N.. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487(7407):330–7.

[32] Chen J, et al. Esophageal squamous cell carcinoma (ESCC): advance in genomics and molecular genetics. Dis Esophagus 2015;28(1):84–9.

[33] Sasaki Y, et al. Genomic characterization of esophageal squamous cell carcinoma: Insights from next-generation sequencing. World J Gastroenterol 2016;22 (7):2284–93.

[34] Hynes N, et al. GPCRs show widespread differential mRNA expression and frequent mutation and copy number variation in solid tumors. PLOS Biol 2019;17 (11).

[35] Miyashita M, et al. DDX60, a DEXD/H Box helicase, is a novel antiviral factor promoting RIG-I-like receptor-mediated signaling. Mol Cell Biol 2011;31(18):3802–19.

[36] Yu Y, et al. Association of survival and immune-related biomarkers with immunotherapy in patients with non-small cell lung cancer: a meta-analysis and individual patient-level analysis. JAMA Netw Open 2019;2(7): e196879.

[37] Morin RD, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature 2011;476(7360):298–303.

[38] Cheng C, et al. Genomic analyses reveal FAM84B and the NOTCH pathway are associated with the progression of esophageal squamous cell carcinoma. Giga-Science 2016;5(1).

[39] Ogawa R, et al. NOTCH1 expression predicts patient prognosis in esophageal squamous cell cancer. Eur Surg Res 2013;51(3-4):101–7.

[40] Leemans CR, Braakhuis BJM, Brakenhoff RH. The molecular biology of head and neck cancer. Nat Rev Cancer 2010;11(1):9–22.

[41] O'Leary B, Finn RS, Turner NC. Treating cancer with selective CDK4/6 inhibitors. Nat Rev Clin Oncol 2016;13(7):417–30.

[42] Yan Wusheng, Michael IIW, Emmert-Buck R, Erickson Heidi S. Squamous cell carcinoma — similarities and differences among anatomical sites. Am J Cancer Res 2011;1(3):275–300.

[43] Sawada G, et al. Genomic landscape of esophageal squamous cell carcinoma in a japanese population. Gastroenterology 2016;150(5):1171–82.

[44] Parker JS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27(8):1160–7.