NEOPLASIA
www.neoplasia.com

Volume 21 Number 6 June 2019 pp. 571–581 571

Gene Expression Profiling Reveals Distinct Molecular Subtypes of Esophageal Squamous Cell Carcinoma in Asian Populations[1]

Fengling Wang[*], Zhongyi Yan[*], Jiajia Lv[*], Junfang Xin[*], Yifang Dang[*], Xiaoxiao Sun[*], Yang An[*], Yijun Qi[†], Qiying Jiang[*], Wan Zhu[‡], Yongqiang Li[*], Ao Li[§] and Xiangqian Guo[*]

[*]Department of Preventive Medicine, Institute of Biomedical Informatics, Joint National Laboratory for Antibody Drug Engineering, Cell Signal Transduction Laboratory, School of Basic Medical Sciences, Henan University, Kaifeng 475004, China; [†]Henan Key Laboratory of Cancer Epigenetics, Cancer Hospital, The First Affiliated Hospital, College of Clinical Medicine, Medical College of Henan University of Science and Technology, Luoyang 471003, China; [‡]Department of Anesthesia, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA; [§]School of Information Science and Technology, Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, China

Abstract

Esophageal squamous cell carcinoma (ESCC) is one of the most common cancers worldwide, particularly in Asian populations, and responds poorly to conventional therapy. Subclassification of ESCCs by molecular analysis is a powerful strategy in extending conventional clinicopathologic classification, improving prognosis and therapy. Here we identified two ESCC molecular subtypes in Chinese population using gene expression profiling data and further validated the molecular subtypes in two other independent Asian populations (Japanese and Vietnamese). Subtype I ESCCs were enriched in pathways including immune response, while genes overexpressed in subtype II ESCCs were mainly involved in ectoderm development, glycolysis process, and cell proliferation. Specifically, we identified potential ESCC subtype-specific diagnostic markers (FOXA1 and EYA2 for subtype I, LAMC2 and KRT14 for subtype II) and further validated them in a fourth Asian cohort. In addition, we propose a few subtype-specific therapeutic targets for ESCC, which may guide future ESCC clinical treatment when further validated.

Neoplasia (2019) 21, 571–581

Introduction

Esophageal cancer, comprised of two main histopathological types, esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma [1], is the sixth most common cause of cancer mortality worldwide [2] and the fourth most lethal cancer in China [3]. ESCC is the most dominant esophageal carcinoma in Asian countries, especially in China [4]. It has been reported that smoking and alcohol abuse are the major contributors to ESCC burden in Western populations, but the attributable factors for ESCCs in non-Western countries were less studied [5,6]. Other factors, such as environmental factors, have also been reported to contribute to such high ESCC prevalence in China [7–9]. Currently, the viable therapeutic targets of ESCC remain to be identified, and ESCC patients suffer dismal outcomes from surgery, conventional chemotherapy, and radiotherapy [10–12].

Abbreviations: ESCC, esophageal squamous cell carcinoma; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; SD, standard deviation; SubMap, Subclass Mapping; GO, Gene Ontology; GSEA, Gene Set Enrichment Analysis; FDR, false discovery rate; TMA, tissue microarray; IHC, immunohistochemistry; CDF, cumulative distribution function; OS, overall survival; ADH, alcohol dehydrogenase; ALDH, aldehyde dehydrogenase; ADH1C*1, alcohol dehydrogenase1C*1.

Address all correspondence to: Prof. Xiangqian Guo, Department of Preventive Medicine, Institute of Biomedical Informatics, Joint National Laboratory for Antibody Drug Engineering, Cell Signal Transduction Laboratory, School of Basic Medical Sciences, Henan University, Kaifeng 475004, China. E-mail: xqguo@henu.edu.cn

[1] Conflict of Interest: None to declare.

Received 10 January 2019; Revised 15 March 2019; Accepted 29 March 2019

© 2019 The Authors. Published by Elsevier Inc. on behalf of Neoplasia Press, Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1476-5586
https://doi.org/10.1016/j.neo.2019.03.013

Molecular heterogeneity has been demonstrated to contribute to the cancer recurrence and metastasis after treatment, and to be even resistant to treatment. Gene expression-based molecular subtyping has been successful in identifying subtypes of a number of tumors (including breast cancer, colon cancer, melanoma, lung cancer, leiomyosarcoma, uterine carcinosarcoma, and uterine leiomyosarcoma), which subsequently led to remarkable benefits in clinical treatments and prognosis [13–19]. To date, a number of studies have investigated the genomic and molecular signatures of ESCC [2,4,11,20]; however, only few have attempted to stratify ESCC into different molecular subtypes [6,21]. The Cancer Genome Atlas (TCGA) study suggested that Western and Asian ESCCs exhibit different gene expression patterns [6], making it even harder to define universal subtypes suitable for both Western and Asian ESCC patients. Because of the high incidence of ESCC in Asia and the difference in the gene expression patterns between Asian and Western ESCC patients, accurate molecular subtyping specifically for Asian ESCC is essential to improve our understanding of ESCC molecular heterogeneity and to potentially guide future treatment strategies.

In the current study, two common ESCC molecular subtypes were measured, defined, and solidified using gene expression profile data of 179 Chinese ESCC, 274 Japanese ESCC, and 37 Vietnamese ESCC cases.

Material and Methods

Determination and Validation of Subtype in Each Asian ESCC Cohort

Expression profiling data of 179 Chinese ESCC cases [22], which are publicly available through accession number GSE53625 in the Gene Expression Omnibus (GEO), were used as the discovery cohort. Dataset GSE69925 including 274 Japanese ESCC cases [23] and dataset TCGA comprised of 37 Vietnamese ESCC cases [6] were used as the validation cohorts. The datasets of GSE53625 and GSE69925 have been properly normalized in the original studies [22,23]. As a result, we used the normalized data for following subtype analysis. The TCGA data were further normalized into transcripts per million reads on level 3 RNAseq data according to a previous study [17].

To define the ESCC molecular subtypes, Consensus clustering (R package Consensus Clustering Plus) [24] was ran on each dataset independently with parameters of "Distance – (1-Pearson correlation), 80% sample resampling, 80% gene resampling, a maximum evaluated k of 12, and agglomerative hierarchical clustering algorithm over 1000 iterations" after filtering the gene expression datasets with standard deviation (SD) and transforming the data by gene-based centering. Then silhouette width (R package cluster) [25] was calculated to evaluate the accuracy of assignments from Consensus Clustering Plus.

Measurement of the Reproducibility of ESCC Subtypes Among the Three Cohorts

Subclass Mapping (SubMap) [26] of the GenePattern was performed to determine the reproducibility of ESCC molecular subtypes between GSE53625, GSE69925, and TCGA RNASeq datasets. And the SubMap was ran with parameters of "num. marker. genes = 300, num. perm = 1000 and num. perm. fisher = 1000."

Differential, Gene Ontology (GO), and Gene Set Enrichment Analysis (GSEA)

To investigate the subtype specific genes, SAM [27] and SAMSeq [28] were applied between ESCC molecular subtypes with a false discovery rate (FDR) of 0.05. Subsequently, GO and KEGG pathway analyses were performed using DAVID Bioinformatics Resources online version 6.7 (https://david.ncifcrf.gov/). GSEA [29] was used to investigate the subtype specific gene expression patterns and pathways. The TARGET V2 database (https://www.broadinstitute.org/cancer/cga/target) was used to identify potential therapeutic targets in each ESCC subtype.

Statistical Analysis

Kaplan-Meier plots were used to compute the survival curves, and log-rank (Mantel-Cox) test was used to determine the statistical significance of survival between subtypes by using GraphPad Prism7 software. Significance of contingency analysis between subtypes was measured by the χ^2 test and Fisher exact test using GraphPad Prism7 software.

Immunohistochemical Analysis

A tissue Arrayer (Beecher, MTA-1) was used to construct the ESCC tissue microarray (TMA) comprising of 166 primary ESCC cases collected from Linzhou Cancer Hospital, Henan, China, between 2010 and 2011, as an independent cohort, which were obtained with institutional review board approval and a waiver of consent due to the archival nature of the specimens. For immunohistochemical staining, primary antibodies against FOXA1 (1:300, Abcam, Ab55178), EYA2 (1:400, Abcam, Ab95875), LAMC2 (1:66, Millipore, MAB19562), and KRT14 (1:1200, Abcam, Ab7800) were used. The immunohistochemistry (IHC) was performed as previously described [30]. In brief, the TMA sections were baked in an oven for 1 hour, deparaffinized, and rehydrated with xylene and graded alcohols. After blocking by 3% H_2O_2 for 15-minute incubation, the sections were boiled in antigen retrieval buffer. Then, the TMAs were incubated in primary antibodies and HRP-conjugated secondary antibodies (Boster). After developing with diaminobenzidine, the sections were counterstained with hematoxylin and dehydrated in graded alcohols and xylene. The staining results were scored as follows: 0 (–), absence of any staining; 1 (+), weak/faint staining (<10%); 2 (++), moderate staining (10%-30%); and 3 (+++), strong staining (>**30**% positive). The cases were divided into negative/low, when scored as 0 or 1 (+), and positive staining, when classified as 2 (++) or 3 (+++).

Results

Consensus Clustering Identifies Two Distinct ESCC Molecular Subtypes

The clinical information of 179 Chinese ESCC patients was shown in Table 1 and Supplementary Table S1. This ESCC cohort consists of 146 males (81.6%) and 33 females (18.4%), with a median age of 60 (range, 36-82) years. In this cohort, 20 (11.2%), 97 (54.2%), and 62 (34.6%) cases have ESCC in the upper, middle, and lower sites of esophagus, respectively. Approximately 63.7% patients of this cohort were documented with tobacco use, and 59.2% were alcohol users.

In order to identify the molecular subtypes of ESCC, Consensus Clustering was performed by genes with the most variable expression levels across the whole ESCC cohort. As shown (Figure 1, *A-C*), two molecular subtypes are the optimal number of clusters, determined with the greatest increase in the area under the empirical cumulative distribution function (CDF) curve. We next measured the confidence of subtype assignments with silhouette width (Figure 1*D*), showing

Table 1. Clinicopathologic Characteristics of the Chinese cohort (N = 179)

Characteristic	Patients, *n* (%)	Subtype I	Subtype II	Other	*P* Value
Total number	179	95	65	19	
Age (year)					.731
Median	60	60	59	60	
Range	36-82	36-81	39-82	45-72	
Sex					.007*
Female	33 (18.4)	23	5	5	
Male	146 (81.6)	72	60	14	
Tobacco use					.958
No	65 (36.3)	34	23	8	
Yes	114 (63.7)	61	42	11	
Alcohol use					.002*
No	73 (40.8)	48	17	8	
Yes	106 (59.2)	47	48	11	
Tumor location					.146
Upper	20 (11.2)	14	5	1	
Middle	97 (54.2)	54	33	10	
Lower	62 (34.6)	27	27	8	
Tumor grade					.086
Poorly	49 (27.4)	27	14	8	
Moderately	98 (54.7)	56	34	8	
Well	32 (17.9)	12	17	3	
T stage					.187
T1	12 (6.7)	9	2	1	
T2	27 (15.1)	10	13	4	
T3	110 (61.4)	59	38	13	
T4	30 (16.8)	17	12	1	
N stage					.553
N0	83 (46.4)	46	26	11	
N1	62 (34.6)	31	26	5	
N2	22 (12.3)	11	10	1	
N3	12 (6.7)	7	3	2	
TNM stage					.373
I	10 (5.6)	6	2	2	
II	77 (43.0)	43	25	9	
III	92 (51.4)	46	38	8	
Arrhythmia					.541
No	136 (76.0)	69	50	17	
Yes	43 (24.0)	26	15	2	
Pneumonia					.248
No	164 (91.6)	84	61	19	
Yes	15 (8.4)	11	4	0	
Anastomotic leak					.810
No	167 (93.3)	90	61	16	
Yes	12 (6.7)	5	4	3	
Adjuvant therapy					.503
No	45	28	14	3	
Yes	104	53	39	12	
Unknown	30	14	12	4	
Survival time (months)					.727
Median	34.7	39.2	32.1	25.5	
Range	0.1-72.6	0.1-72.5	1.4-72.6	0.6-69.6	
Event					.553
Dead	106 (59.2)	54	40	12	
Alive	73 (40.8)	41	25	7	

that 160 of the 179 ESCC cases had a positive silhouette value, of which 95 belonged to subtype I and 65 to subtype II, and these reliable "core" cases with positive silhouette width were then used for the subsequent analysis, unless otherwise indicated.

Validation of Subtypes in the Independent Datasets

To further validate and solidify the ESCC molecular subtypes as defined above and translate them into future clinical use, another two expression profiling datasets were collected from GEO and TCGA database as independent validation cohorts, including a cohort of 274 Japanese cases (GSE69925) and another cohort of 37 Vietnamese cases from TCGA. Regardless of different profiling methods (gene microarray and RNAseq) and different originations (China, Japan, and Vietnam) of ESCC patients in these three datasets, Consensus Clustering analysis consistently showed that there are two distinct molecular subtypes of ESCC in the Japanese and Vietnamese cohorts as well (Figure 2 and Supplementary Figure S1).

To further test the reproducibility of molecular subtypes between these three Asian datasets, SubMap analysis was performed among the three datasets. SubMap results showed that subtype A1 (Subtype I) in GSE53625, subtype B1 in TCGA dataset, and subtype C1 in GSE69925 were significantly correlated, and subtype A2 (Subtype II) in GSE53625, subtype B2 in TCGA dataset, and subtype C2 in GSE69925 were also significantly correlated, and demonstrated that both molecular subtypes were common and significantly reproduced in the three independent Asian cohorts, indicating their future potential clinical usage (Figure 3).

Molecular Subtypes and Their Correlation to Clinical Features

In order to look at the preference of ESCC molecular subtypes to clinical characteristics, the association between clinical data and molecular subtypes was tested in Chinese cohort. Interestingly, the subtype II ESCC included significantly more male patients (60/65, 92.3%) than subtype I ESCC (72/95, 75.8%) (Table 1, χ^2 test; P = .007). It is noteworthy that a large number of patients in subtype II (48/65, 73.8%) were alcohol drinkers compared with 49.5% alcohol drinkers in subtype I (47/95) (χ^2 test; P = .002). Subtype II ESCC patients had a median overall survival (OS) time of 32.1 months, and subtype I patients had a slightly longer median OS time of 39.2 months. However, survival curve analysis (Kaplan-Meier plots) showed that there was no significant difference in survival between the two subtypes. In addition, there was no significant difference between the two ESCC molecular subtypes for patient's age (P = .731), smoking (P = .958), tumor location (χ^2 test; P = .146), grade (χ^2 test; P = .086), and TNM stage (P = .373) in the Chinese, Japanese, and Vietnamese cohorts (Table 1, Supplementary Tables S2 and S3).

To build the general and comprehensive molecular subtypes of ESCCs, the ESCC cases after chemo/radio treatments are also included for the subtyping analysis in the discovery and validation cohorts, which is clinically meaningful and reasonable since the patients tolerant or resistant to previous chemo/radio treatments are much more eager to the subtype-targeted therapies. When comparing the status of chemo/radio treatments with ESCC molecular subtypes in the three cohorts, we found that there is no significant association between treatments and ESCC molecular subtypes in all three Asian cohorts (Table 1, Supplementary Tables S2 and S3), indicating that ESCC molecular subtypes are independent of previous chemo/radio treatments.

Functional Analysis of ESCC Subtype-Specific Genes

SAM analysis in Chinese cohort showed that a total of 8281 genes had a significant expression difference between the two molecular subtypes, of which 3065 genes were significantly overexpressed in subtype I ESCC than subtype II ESCC, while the other 5216 genes were significantly overexpressed in subtype II ESCC. The significantly overexpressed genes in each subtype were listed in Supplementary Table S4.

Subsequently, to gain insight into the biological meaning of the subtypes, GO and pathway analyses were performed on the top 3000 overexpressed genes in each ESCC subtype. In total, 432 biological processes terms and 33 pathway terms were enriched in subtype I ESCCs, while 356 biological processes and 27 pathway terms were

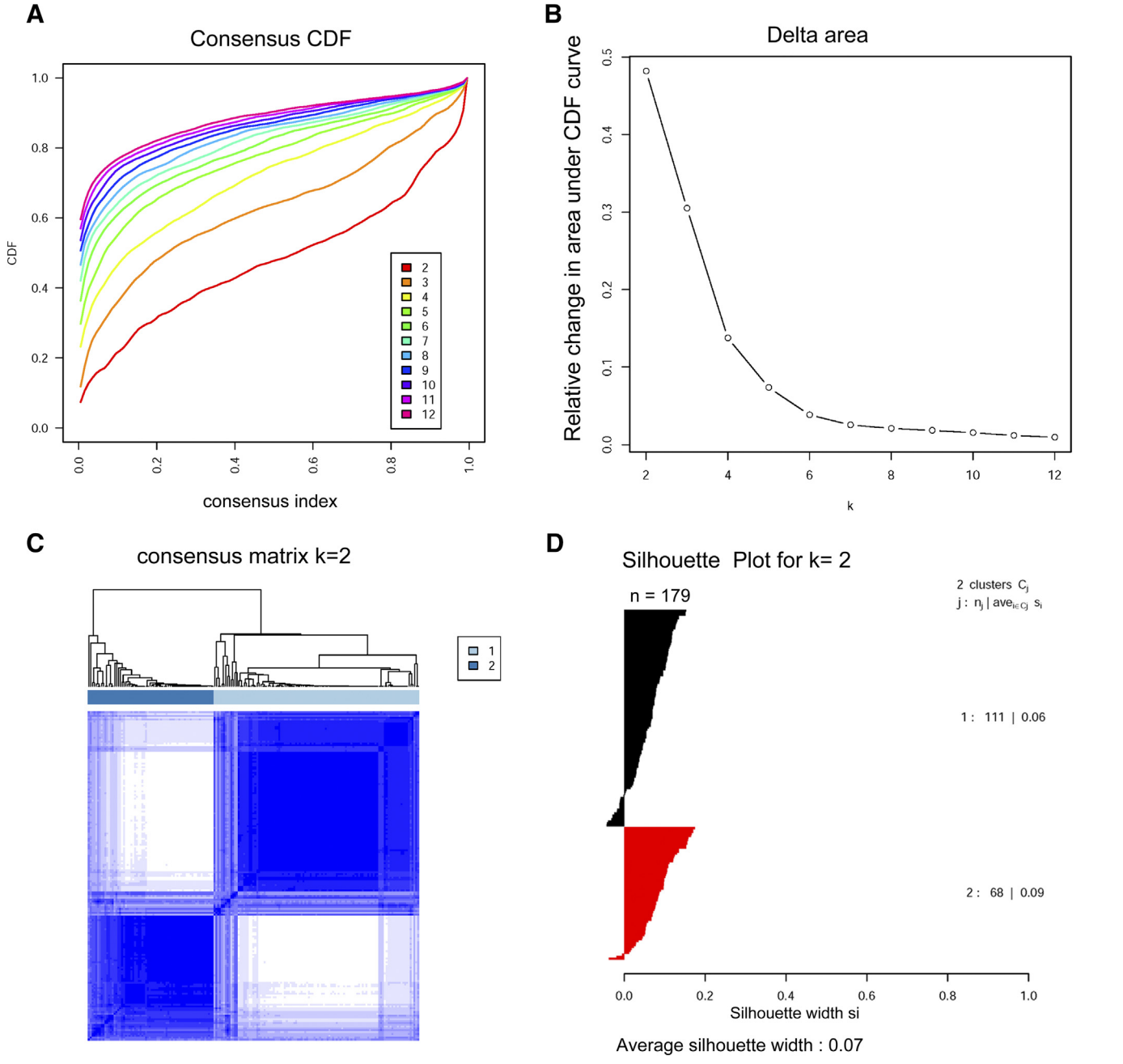

Figure 1. Identification of two molecular subtypes of ESCC in Chinese cohort. (A) Empirical cumulative distribution plot determines the optimal number of ESCC molecular subtypes. (B) Relative increase in the area under the CDF curve along with increasing assumed number of molecular subtypes. (C) Consensus clustering matrix of ESCC samples using two molecular subtypes. (D) Silhouette analysis of ESCC samples based on the assignments from Consensus Clustering.

identified in subtype II (Supplementary Tables S5 and S6). Interestingly, 220 of the top 3000 genes (7.3%) in subtype I ESCCs were related to immune response pathway, such as *CD5*, *CD27*, *CD38*, *CD83*, *CD37*, *CD74*, *CD28*, *CD19*, *CD226*, *IL2*, and *BCL6*, indicating that these immune-related genes are highly active in subtype I ESCC, potentially as a consequence of higher immunological activity in the subtype I ESCC (Figure 4*A* and Supplementary Table S5), and signifying that subtype I ESCC patients might benefit more from immunotherapeutic methods. Tanaka et al. have reported I-type ESCCs in their study and shown overexpression of 234 immune activation-related genes exhibited beneficial response from CRT (chemoradiotherapy) [23]. When comparing these 234 immune activation-related genes with genes highly expressed in subtype I in our study, we found that 119 of 234 immune activation-related genes are also significantly overexpressed in subtype I ESCCs in our study. It suggested that the subtype I ESCCs in our study may correspond to the I-type ESCCs as previously reported [23]. Other biological pathways enriched in subtype I included calcium signaling, drug metabolism, chemokine signaling, *etc.* (Figure 4*A*).

In comparison, those genes overexpressed in subtype II were associated with a different set of biological pathways, including focal

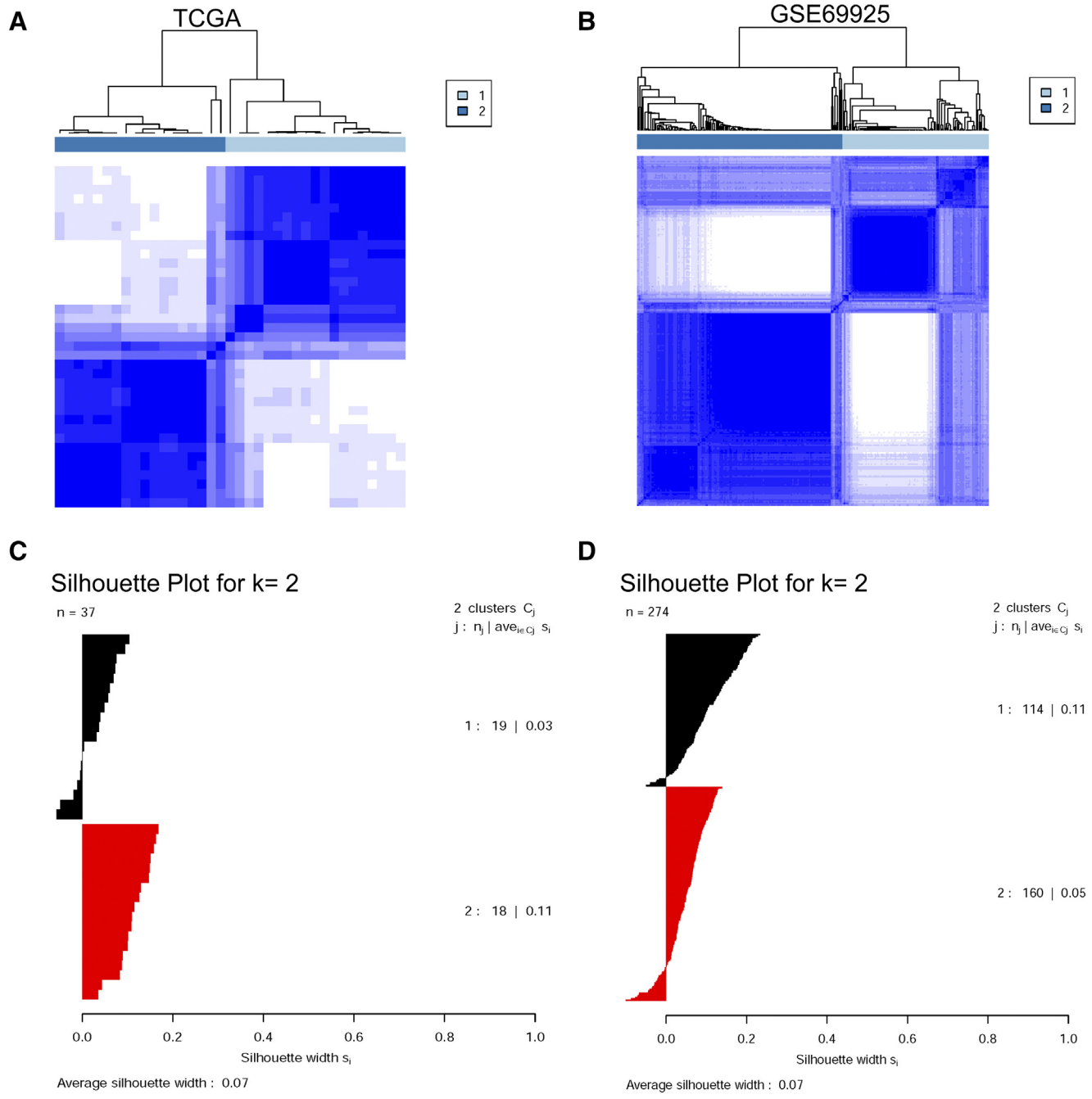

Figure 2. Identification of two molecular subtypes of ESCC in the Vietnamese (TCGA) and the Japanese cohorts (GSE69925). (A and B) Consensus clustering matrix shows two molecular subtypes of ESCC in TCGA and GSE69925 dataset, respectively. (C and D) Silhouette analysis for TCGA and GSE69925 datasets validates the subtype assignments from consensus clustering, respectively.

adhesion, regulation of actin cytoskeleton, epithelium development, glycolysis, MAPK pathway, regulation of cell cycle, apoptosis, programmed cell death, *etc.* (Figure 4*B*). In addition, well-described epithelium development genes, such as *E2F4*, *JUN*, *CFL1*, *VEGFA*, *KRT14*, *SFN*, *KRT5*, *LAMA3*, and *IRF6*, are highly expressed in the subtype II ESCC (Supplementary Table S5).

The additional GSEA in the Chinese cohort further demonstrated that gene sets including immune network were significantly enriched in subtype I ESCC, while gene sets including amplification of ribosome and galactose metabolism pathways were overrepresented in subtype II ESCC (Figure 5).

Identification of Diagnostic Biomarkers for Different Subtypes

Based on SAM and SAMseq analysis of gene expression profiling data from the three cohorts and the availability of commercial antibodies, we found that four genes are top-ranked in the SAM/SAMseq analysis in the three cohorts and high-quality antibodies against them are commercially available. In order to develop subtype diagnostic biomarkers as performed in others [17,31], these four genes, including *FOXA1* and *EYA2* overexpressed in subtype I ESCC, and genes *LAMC2* and *KRT14* overexpressed in subtype II ESCC, were selected to be further tested in another independent Chinese cohort, respectively (Supplementary Table S7). The fourth

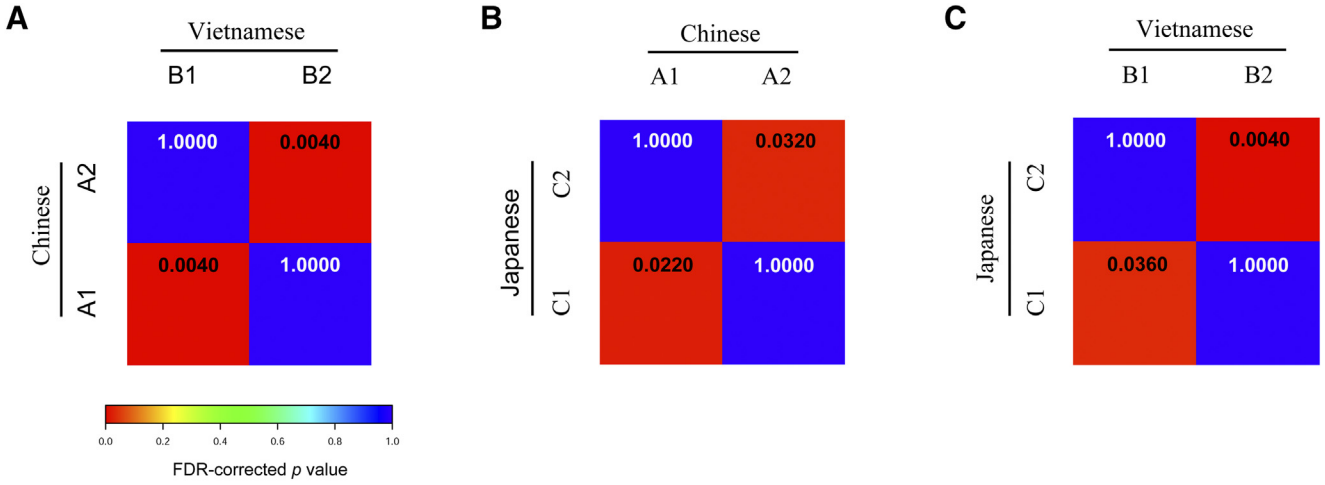

Figure 3. SubMap association matrix among Chinese (subtype A1 and subtype A2), Vietnamese (subtype B1 and subtype B2), and Japanese data (subtype C1 and subtype C2). *P* value is corrected with FDR.

independent Chinese cohort included 166 ESCC cases which were put on a tissue microarray for IHC staining. The clinical information of 166 Chinese ESCC patients was shown in Supplementary Table S8. IHC results showed that 64.2% (104/162) and 62.6% (102/163) of the ESCC cases were positive for FOXA1 and EYA2, respectively (Figure 6, *A* and *B*), and the IHC scores of FOXA1 and EYA2 staining were significantly correlated (r = 0.397, P < .001, Spearman correlation). The IHC staining results of LAMC2 and KRT14 showed that they are positive in 16.3% (27/166) (Figure 6*C*) and 41.4% (67/162) of ESCC cases (Figure 6*D*), respectively. Moreover, the IHC scores of LAMC2 and KRT14 were found to be also significantly correlated (r = 0.253, P = .001, Spearman correlation) as FOXA1 and EYA2 did.

Potential Clinical Implication of ESCC Subtyping

The ultimate goal of recognizing ESCC molecular subtypes is to develop subtype-specific therapeutic methods and translate them into clinical use; the importance of delineating these ESCC subtypes lies in proposing particular therapeutic approaches appreciated to each subtype.

In order to find potential novel therapeutic methods for each ESCC subtype, we compared our ESCC subtype-specific over-expressed genes with TARGET database [32], a database including gene targets and their functional inhibitors, to identify gene targets that have the possibility to be translated into meaningful clinical implication, as performed in other studies [33–35]. As shown in Table 2, seven genes significantly overexpressed in subtype I ESCC, including *BCL2*, *FLT3*, and *KIT*, and six genes overexpressed in subtype II ESCC, including *CDK6*, *CDKN1A*, *NRAS*, and *EGFR*, had potent target-oriented inhibitors available in clinic or clinical trials. All analyses indicated that above-stated targeted therapies may be potentially suitable and effective for ESCC patients from different subtypes when further validated.

Discussion

Esophageal squamous carcinoma is a major esophageal malignancy in Asian and African populations, particularly in China [36–39], accounting for about 90% of esophageal carcinoma [40]. ESCC, occurring three to four times more common in men than in women, is reported to be associated with environmental risk factors, including tobacco, alcohol, diet, human papillomavirus infection, esophageal diseases, oral health, *etc.* [41,42]. At present, the widely accepted ESCC clinical diagnosis, prognosis, and treatments are still based on the TNM staging system. However, due to the heterogeneity of cancers at the molecular and genetic levels [43,44], patients with the same stage may reach quite different clinical outcomes when given identical treatment. So in-depth exploration of the molecular heterogeneity of ESCC will not only help us understand the progression mechanisms of ESCC but also offer the opportunities to develop subtype-specific and personalized therapeutic strategies for ESCC patients.

Recently, TCGA demonstrated the remarkable molecular differences of ESCCs between Western and Vietnamese populations [6]. Liu et al. reported the molecular discrepancy of ESCC in sub-Saharan by comprehensive molecular analysis [21]. However, to date, few studies have focused on exploring the molecular heterogeneity of ESCC in Asian population, a population with the highest prevalence of ESCC.

In our study, two ESCC molecular subtypes were identified by analyzing the expression profile of 179 Chinese ESCCs and then validated in two independent Asian cohorts: a Japanese cohort with 274 ESCC cases and a Vietnamese cohort with 37 ESCC cases. Although samples in different cohorts were from different countries and different profiling methods were used in these three studies, the two ESCC molecular subtypes were significantly reproducible among these three Asian ESCC cohorts.

Exploring ESCC in the light of subtypes therefore may expand our understanding of ESCC pathology, and different subtypes of ESCC may develop or progress from the different genomic events or distinct cell processes. The functional annotation analysis showed that subtype I ESCC cases overpresented the genes in immune response process. For example, the receptors *CD96* and *TIGIT* which are expressed on the surface of T and NK cells and play important inhibitory roles in immune function were found to be significantly overexpressed in subtype I [45]. *CD96* has been identified to be highly expressed in acute myeloid leukemia and T-cell acute lymphoblastic leukemia [46] and also been proposed as a cancer stem cell marker in leukemia [47]. Interestingly, *TIGIT* is highly expressed on both $CD8^+$ T-cells and Tregs in many clinical tumor

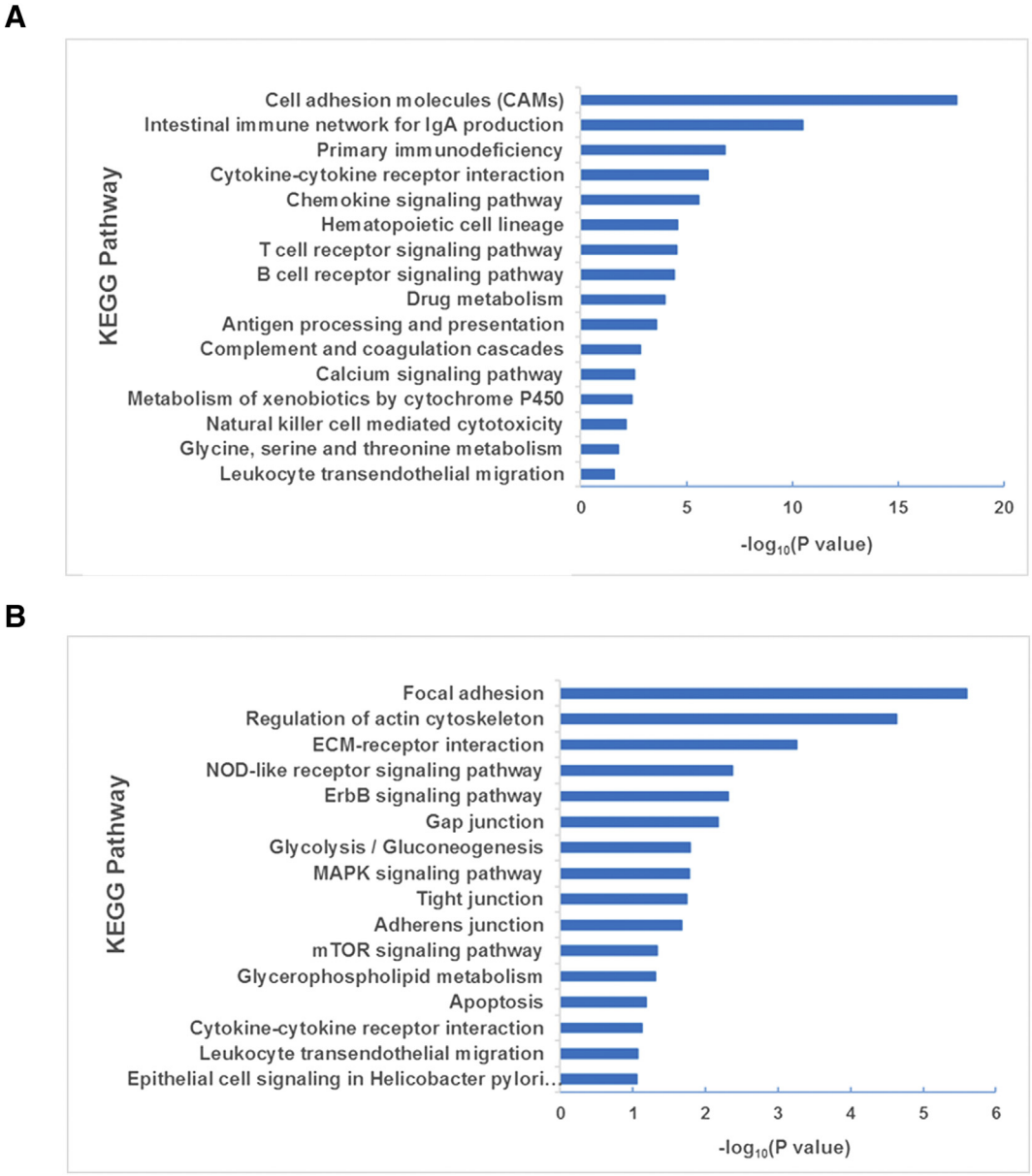

Figure 4. Enriched biological pathways by analyzing overexpressed genes in each ESCC subtype. (A) KEGG pathway terms by analyzing overexpressed genes in subtype I. (B) KEGG pathway terms by analyzing overexpressed genes in subtype II.

settings [48,49], and blockage of TIGIT can enhance T-cell function, particularly when used in combination with PD-1/PD-L1 blockade [49,50]. The high expression of immune response–related genes in subtype I suggests that this tumor subtype might be more likely to benefit from immunotherapy.

In contrast, the GO biological processes terms enriched in subtype II ESCC included ectoderm development, cell proliferation, regulation of apoptosis, cell cycle, cell motion, glycolysis process, *etc.* (Supplementary Table S5). The genes overexpressed in subtype II ESCC, including *FOSL1* [51], *PTHLH* [52], *CAV1* [53], and *LAMC2* [54], which have been reported to be overexpressed in a certain number of cancers, can promote cell proliferation, migration, and invasion. *LAMC2* was also reported to be associated with tumor metastasis, recurrence, and poor prognosis in different types of cancers [55–57]. *FOSL1*, a downstream effector of oncogene *KRAS* and a component of the mitotic machinery, can be used as a prognostic marker to identify mutant KRAS in lung and pancreatic cancer patients with the worst survival outcome [51]. *KRT14* and *PDPN* are tumor basal cell markers, both of which were significant overexpressed in subtype II ESCCs, while *SIX1* was reported to maintain or increase *PDPN*-positive cancer stem cells [58]; however, *SIX1* was only found to be significantly overexpressed in subtype II ESCCs from TCGA dataset in Vietnamese cohort in our study. In addition, subtype II ESCCs were characterized by increased expression of glycolytic pathway genes, including *PKM2*, *HK2*, *PFKP*, *PGAM4*, *ALDOA*, and *LDHA*. Cancer cells are addicted to glycolytic anaerobic pathways even under normal oxygen levels, a phenomenon first described by Otto Warburg [59]. *PKM2* can promote the Warburg effect, and its expression level is usually higher in human cancer biopsies than in adjacent normal tissues [60,61]. *HK2*, one member of hexokinase family, has also been widely studied in the cancer research, particularly for its roles in chemoresistance and prognosis [62,63].

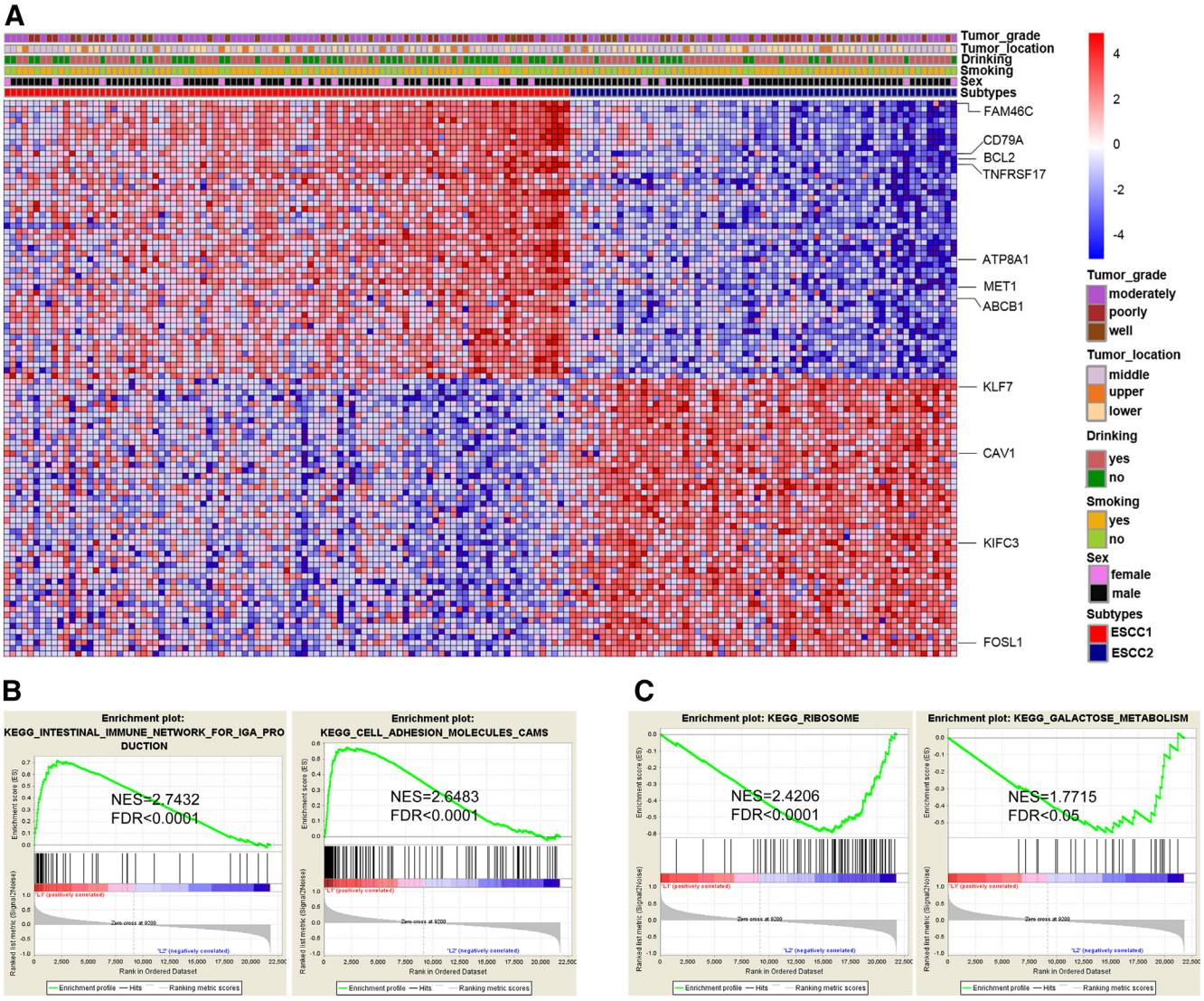

Figure 5. GSEA shows different gene expression signatures in distinct ESCC molecular subtypes. (A) Different gene expression patterns in subtype I and subtype II. Red, overexpressed genes; blue, downexpressed genes. (B) GSEA demonstrated enhanced activity of immune network and cell adhesion molecules pathways in subtype I. (C) GSEA showed enhanced activity of ribosome and galactose metabolism pathways in subtype II. Abbreviations: *NES*, normalized enrichment score; *FDR*, false discovery rate.

Alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) are the primary enzymes involved in alcohol metabolism; carrying these genetic variants, such as alcohol dehydrogenase1C*1 (ADH1C*1) homozygotes, often confers a higher risk for alcohol-related cancers, such as esophageal adenocarcinoma and gastric cancer risk [64,65]. In the Chinese cohort, subtype II cases include more male patients and alcohol drinkers. It should be noted that genes associated with alcohol metabolism were significantly underexpressed in molecular subtype II compared with subtype I (Supplementary Table S4), including *ADH1A*, *ADH1C*, *ADH4*, *ADH7*, *ALDH1A1*, and *CYP2E1*. Meanwhile, we analyzed the association between drinking and molecular subtypes in the Vietnamese dataset and found that subtype II also had a higher percentage (72.2%) of alcohol users compared to subtype I with 66.6%, but the ratio did not reach significance (χ^2 test; P = .745) (Supplementary Table S2).

FOXA1 is an upstream regulator of *KRT7* and *LOXL2*, which is expressed in a subset of poor prognostic ESCC patients; some of the FOXA1-positive ESCC cases show glandular structures. *FOXA1* siRNA can reduce the expression levels of both *KRT7* and *LOXL2*; knockdown of either *FOXA1* or *LOXL2* could reduce invasion and migration of ESCC cells [66]. In our study, we found that both *FOXA1* and *KRT7* are overexpressed in subtype I ESCCs, while *LOXL2* is significantly overexpressed in subtype II ESCCs. Some of the FOXA1-positive ESCC cases also show the glandular structures that Sano et al. previously reported [66]; however, there is no significant outcome difference between the two molecular subtypes. The discrepancy about *LOXL2* expression between the two studies indicates the molecular heterogeneity and complicated progression mechanisms of ESCCs.

Development of subtype-specific diagnostic IHC biomarkers will facilitate the clinical translation of ESCC molecular subtypes. To achieve this, we collected additional 166 Chinese ESCC cases as the fourth cohort and determined the expression levels of subtype-specific genes in the fourth cohort using IHC (FOXA1 and EYA2 for subtype I, LAMC2 and KRT14 for subtype II); the IHC results showed significant correlation between genes from the same subtype.

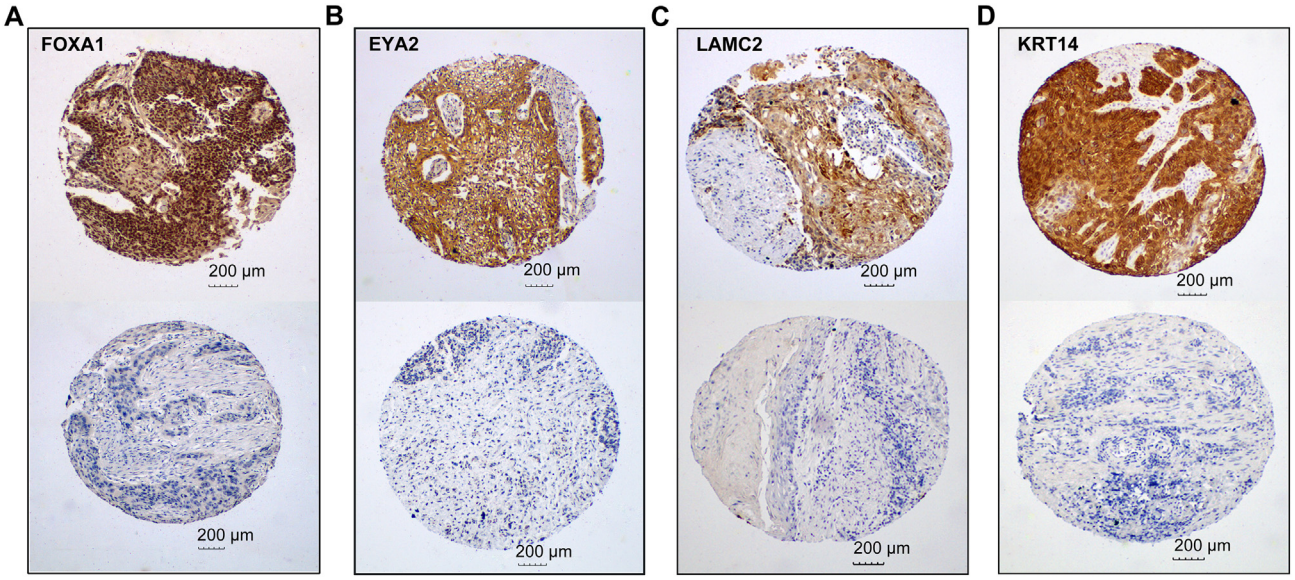

Figure 6. Immunohistochemical markers for subtype I and subtype II ESCC. (A, B, C and D) Representative staining for FOXA1, EYA2, LAMC2, and KRT14, respectively (positive, score 3; negative, score 0), 100×.

Currently, the therapeutic strategies available for ESCC patients were very limited. All novel drugs in clinical trials for ESCC are tested as if ESCC is a homogeneous group. In this scenario, the effect of a novel drug that affects only one molecular subtype may not become noticeable when cases belonging to other subtypes are included in the trial and not recognized as separate entities. Conducting molecular subclassification and developing subgroup targeted treatment methods have been shown to be very successful in other types of cancers, *e.g.*, breast cancer and lung cancer. A well-known example is the clinical trial of Herceptin treatment, in which HER2-positive breast cancer patients benefited from Herceptin treatment the most, whereas HER2-negative patients poorly responded to the treatment [67]. Therefore, molecular subtyping of ESCCs would shed light on the subtype-specific therapeutic methods as well. In our study, we demonstrated that a number of known targets were significantly and specifically overexpressed in subtype I and subtype II ESCCs, including *BCL2*, *FLT3*, *KIT*, and *FGFR2* in subtype I and *EGFR* in subtype II, respectively (Table 2). Amplification of antiapoptotic Bcl-2 family proteins has been reported to be related with chemotherapy resistance in various cancers, and successful development of drugs targeting Bcl-2 family members (oblimersen sodium, AT-101, ABT-263, GX15-070) might potentially overcome the chemoresistance for cancer patients, particularly for subtype I ESCC [68]. *EGFR*, a well-known gene whose mutation status has been successfully used to guide the clinical treatment for non–small cell lung cancer patients [35,69], might also serve as a therapeutic target for subtype II ESCC patients.

In conclusion, we defined two molecular distinct subtypes of ESCCs, which were common and reproduced in Japanese cohort and Vietnamese cohort. Our study opens the opportunities to explore the ESCC subtype-specific mechanisms for malignancy development and progression, and may shed light on the ESCC subtype-specific therapeutic treatments.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neo.2019.03.013.

Table 2. Potential Target Genes Enriched in Each Molecular Subtype

Gene Overexpressed		Examples of Potential Therapeutic Agents
	BCL2	BCL2 inhibitors
	DNMT3A	DNAMT inhibitors
	FGFR2	FGFR inhibitors
Subtype I	*FLT3*	Sunitinib, FLT3 inhibitors
	KIT	Imatinib, Sunitinib, Novel KIT inhibitors
	MITF	Vemurafenib, Dabrafenib, RAF inhibitors
	TMPRSS2	PARP inhibitors
	CDK6	CDK4/6 inhibitors
	CDKN1A	CDK inhibitors
Subtype II	*EGFR*	Erlotinib, Gefitinib, EGFR inhibitors
	MET	Gefitinib, Erlotinib, EGFR inhibitors, Crizotinib, MET inhibitors
	NRAS	Vemurafenib, Dabrafenib, RAF inhibitors, MEK inhibitors
	ROS1	Crizotinib

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 81602362 to X.G.; No. 61571414 to A.L.), the program for Science and Technology Development in Henan Province (No. 162102310391 to X.G.), the program for Young Key Teacher of Henan Province (No. 2016GGJS-214 to X.G.), the supporting grants of Henan University (No. 2015YBZR048 to X.G.; No. B2015151 to X.G.; No. 2015YBZR052 to F.W.), the program for Innovative Talents of Science and Technology in Henan Province (No. 18HASTIT048 to X.G.), and Yellow River Scholar Program (No. H2016012 to X.G.). The funding bodies were not involved in the study design, data collection, analysis and interpretation of data, or writing of this manuscript.

References

[1] Siewert JR and Ott K (2007). Are squamous and adenocarcinomas of the esophagus the same disease? *Semin Radiat Oncol* **17**, 38–44.

[2] Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, and Kong P, et al (2015). Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. *Am J Hum Genet* **96**, 597–611.

[3] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, and Bray F, et al (2016). Cancer statistics in China, 2015. *CA Cancer J Clin* **66**, 115–132.

[4] Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, and Sun ZM, et al (2014). Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet* **46**, 1097–1102.

[5] Engel LS, Chow WH, Vaughan TL, Gammon MD, Risch HA, and Stanford JL, et al (2003). Population attributable risks of esophageal and gastric cancers. *J Natl Cancer Inst* **95**, 1404–1413.

[6] Cancer Genome Atlas Research Network (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175.

[7] Tran GD, Sun XD, Abnet CC, Fan JH, Dawsey SM, and Dong ZW, et al (2005). Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. *Int J Cancer* **113**, 456–463.

[8] Wu C, Hu Z, He Z, Jia W, Wang F, and Zhou Y, et al (2011). Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. *Nat Genet* **43**, 679–684.

[9] Wu C, Kraft P, Zhai K, Chang J, Wang Z, and Li Y, et al (2012). Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nat Genet* **44**, 1090–1097.

[10] Wang LS, Chow KC, Chi KH, Liu CC, Li WY, and Chiu JH, et al (1999). Prognosis of esophageal squamous cell carcinoma: analysis of clinicopathological and biological factors. *Am J Gastroenterol* **94**, 1933–1940.

[11] Wu C, Li D, Jia W, Hu Z, Zhou Y, and Yu D, et al (2013). Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nat Genet* **45**, 632–638.

[12] Zhang Y (2013). Epidemiology of esophageal cancer. *World J Gastroenterol* **19**, 5598–5606.

[13] Higgins MJ and Baselga J (2011). Targeted therapies for breast cancer. *J Clin Invest* **121**, 3797–3803.

[14] Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, and Vescovo L, et al (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med* **10**e1001453.

[15] Cancer Genome Atlas Research Network (2015). Genomic classification of cutaneous melanoma. Cell **161**, 1681–1696.

[16] Chen F, Zhang Y, Parra E, Rodriguez J, Behrens C, and Akbani R, et al (2017). Multiplatform-based molecular subtypes of non–small cell lung cancer. *Oncogene* **36**, 1384–1393.

[17] Guo X, Jo VY, Mills AM, Zhu SX, Lee CH, and Espinosa I, et al (2015). Clinically relevant molecular subtypes in leiomyosarcoma. *Clin Cancer Res* **21**, 3501.

[18] An Y, Wang H, Jie J, Tang Y, Zhang W, and Ji S, et al (2017). Identification of distinct molecular subtypes of uterine carcinosarcoma. *Oncotarget* **8**, 15878–15886.

[19] An Y, Wang S, Li S, Zhang L, Wang D, and Wang H, et al (2017). Distinct molecular subtypes of uterine leiomyosarcoma respond differently to chemotherapy treatment. *BMC Cancer* **17**, 639.

[20] Sawada G, Niida A, Uchi R, Hirata H, Shimamura T, and Suzuki Y, et al (2016). Genomic landscape of esophageal squamous cell carcinoma in a Japanese population. *Gastroenterology* **150**, 1171–1182.

[21] Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, and Parker JS, et al (2016). Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. *JCI Insight* **1**e88755.

[22] Li J, Chen Z, Tian L, Zhou C, He MY, and Gao Y, et al (2014). LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700–1710.

[23] Tanaka Y, Aoyagi K, Minashi K, Komatsuzaki R, Komatsu M, and Chiwaki F, et al (2015). Discovery of a good responder subtype of esophageal squamous cell carcinoma with cytotoxic T-lymphocyte signatures activated by chemoradiotherapy. *PloS ONE* **10**, e0143804.

[24] Wilkerson MD and Hayes DN (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573.

[25] Rousseeuw PJ (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**, 53–65.

[26] Hoshida Y, Brunet JP, Tamayo P, Golub TR, and Mesirov JP (2007). Subclass mapping: identifying common subtypes in independent disease data sets. *PLoS ONE* **2**e1195.

[27] Tusher VG, Tibshirani R, and Chu G (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116–5121.

[28] Li J and Tibshirani R (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res* **22**, 519–536.

[29] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, and Gillette MA, et al (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* **102**, 15545–15550.

[30] Kadota K, Nitadori J, Rekhtman N, Jones DR, Adusumilli PS, and Travis WD (2015). Reevaluation and reclassification of resected lung carcinomas originally diagnosed as squamous cell carcinoma using immunohistochemical analysis. *Am J Surg Pathol* **39**, 1170–1180.

[31] Jönsson G, Busch C, Knappskog S, Geisler J, Miletic H, and Ringnér M, et al (2010). Gene expression profiling–based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin Cancer Res* **16**, 3356–3367.

[32] Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, and Marlow S, et al (2014). Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682–688.

[33] Moroni M, Veronese S, Benvenuti S, Marrapese G, Sartore-Bianchi A, and Di Nicolantonio F, et al (2005). Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. *Lancet Oncol* **6**, 279–286.

[34] Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, and Shyr Y, et al (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* **121**, 2750–2767.

[35] Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, and Gabriel S, et al (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500.

[36] Brown LM, Devesa SS, and Chow W-H (2008). Incidence of adenocarcinoma of the esophagus among white americans by sex, stage, and age. *J Natl Cancer Inst* **100**, 1184–1187.

[37] Cook MB (2011). Editorial: non-acid reflux: the missing link between gastric atrophy and esophageal squamous cell carcinoma? *Am J Gastroenterol* **106**, 1930–1932.

[38] Dawsey SM, Lewin KJ, Wang GQ, Liu FS, Nieberg RK, and Yu Y, et al (1994). Squamous esophageal histology and subsequent risk of squamous cell carcinoma of the esophagus. A prospective follow-up study from Linxian, China. *Cancer* **74**, 1686–1692.

[39] Lin Y, Totsuka Y, He Y, Kikuchi S, Qiao Y, and Ueda J, et al (2013). Epidemiology of esophageal cancer in Japan and China. *J Epidemiol* **23**, 233–242.

[40] Hiyama T, Yoshihara M, Tanaka S, and Chayama K (2007). Genetic polymorphisms and esophageal cancer risk. *Int J Cancer* **121**, 1643–1658.

[41] Castro C, Peleteiro B, and Lunet N (2018). Modifiable factors and esophageal cancer: a systematic review of published meta-analyses. *J Gastroenterol* **53**, 37–51.

[42] Chen X, Yuan Z, Lu M, Zhang Y, Jin L, and Ye W (2017). Poor oral health is associated with an increased risk of esophageal squamous cell carcinoma — a population-based case-control study in China. *Int J Cancer* **140**, 626–635.

[43] Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, and Endesfelder D, et al (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883–892.

[44] Navin N, Kendall J, Troge J, Andrews P, Rodgers L, and McIndoo J, et al (2011). Tumor evolution inferred by single cell sequencing. *Nature* **472**, 90–94.

[45] Blake SJ, Dougall WC, Miles JJ, Teng MW, and Smyth MJ (2016). Molecular pathways: targeting CD96 and TIGIT for cancer immunotherapy. *Clin Cancer Res* **22**, 5183–5188.

[46] Meyer D, Seth S, Albrecht J, Maier MK, du Pasquier L, and Ravens I, et al (2009). CD96 interaction with CD155 via its first Ig-like domain is modulated by alternative splicing or mutations in distal Ig-like domains. *J Biol Chem* **284**, 2235–2244.

[47] Hosen N, Park CY, Tatsumi N, Oji Y, Sugiyama H, and Gramatzki M, et al (2007). CD96 is a leukemic stem cell–specific marker in human acute myeloid leukemia. *Proc Natl Acad Sci U S A* **104**, 11008–11013.

[48] Kong Y, Zhu L, Schell TD, Zhang J, Claxton DF, and Ehmann WC, et al (2016). T-cell immunoglobulin and ITIM domain (TIGIT) associates with CD8+ T-cell exhaustion and poor clinical outcome in AML patients. *Clin Cancer Res* **22**, 3057–3066.

[49] Johnston RJ, Comps-Agrar L, Hackney J, Yu X, Huseni M, and Yang Y, et al (2014). The immunoreceptor TIGIT regulates antitumor and antiviral CD8+ T cell effector function. *Cancer Cell* **26**, 923–937.

[50] Chauvin JM, Pagliano O, Fourcade J, Sun Z, Wang H, and Sander C, et al (2015). TIGIT and PD-1 impair tumor antigen–specific CD8(+) T cells in melanoma patients. *J Clin Invest* **125**, 2046–2058.

[51] Vallejo A, Perurena N, Guruceaga E, Mazur PK, Martinez-Canarias S, and Zandueta C, et al (2017). An integrative approach unveils FOSL1 as an oncogene vulnerability in KRAS-driven lung and pancreatic cancer. *Nat Commun* **8**14294.

[52] Chang WM, Lin YF, Su CY, Peng HY, Chang YC, and Hsiao JR, et al (2017). Parathyroid hormone-like hormone is a poor prognosis marker of head and neck cancer and promotes cell growth via RUNX2 regulation. *Sci Rep* **7**41131.

[53] Joshi B, Strugnell SS, Goetz JG, Kojic LD, Cox ME, and Griffith OL, et al (2008). Phosphorylated caveolin-1 regulates Rho/ROCK-dependent focal adhesion dynamics and tumor cell migration and invasion. *Cancer Res* **68**, 8210–8220.

[54] Moon YW, Rao G, Kim JJ, Shim HS, Park KS, and An SS, et al (2015). LAMC2 enhances the metastatic potential of lung adenocarcinoma. *Cell Death Differ* **22**, 1341–1352.

[55] Takahashi S, Hasebe T, Oda T, Sasaki S, Kinoshita T, and Konishi M, et al (2002). Cytoplasmic expression of laminin γ2 chain correlates with postoperative hepatic metastasis and poor prognosis in patients with pancreatic ductal adenocarcinoma. *Cancer* **94**, 1894–1901.

[56] Yamamoto H, Itoh F, Iku S, Hosokawa M, and Imai K (2001). Expression of the γ2 chain of laminin-5 at the invasive front is associated with recurrence and poor prognosis in human esophageal squamous cell carcinoma. *Clin Cancer Res* **7**, 896–900.

[57] Shou J, Hu N, Takikita M, Roth MJ, Johnson LL, and Giffen C, et al (2008). Over-expression of CDC25B and LAMC2 mRNA and protein in esophageal squamous cell carcinomas and pre-malignant lesions in subjects from a high-risk population in China. *Cancer Epidemiol Biomarkers Prev* **17**, 1424–1435.

[58] Nishimura T, Tamaoki M, Komatsuzaki R, Oue N, Taniguchi H, and Komatsu M, et al (2017). SIX1 maintains tumor basal cells via transforming growth factor-β pathway and associates with poor prognosis in esophageal cancer. *Cancer Sci* **108**, 216–225.

[59] Warburg O, Wind F, and Negelein E (1927). The metabolism of tumors in the body. *J Gen Physiol* **8**, 519–530.

[60] Bluemlein K, Gruning NM, Feichtinger RG, Lehrach H, Kofler B, and Ralser M (2011). No evidence for a shift in pyruvate kinase PKM1 to PKM2 expression during tumorigenesis. *Oncotarget* **2**, 393–400.

[61] Luo W and Semenza GL (2012). Emerging roles of PKM2 in cell metabolism and cancer progression. *Trends Endocrinol Metab* **23**, 560–566.

[62] Peschiaroli A, Giacobbe A, Formosa A, Markert E, Bongiorno-Borbone L, and Levine A, et al (2013). miR-143 regulates hexokinase 2 expression in cancer cells. *Oncogene* **32**, 797–802.

[63] Palmieri D, Fitzgerald D, Shreeve SM, Hua E, Bronder JL, and Weil RJ, et al (2009). Analyses of resected human brain metastases of breast cancer reveal the association between up-regulation of hexokinase 2 and poor prognosis. *Mol Cancer Res* **7**, 1438–1445.

[64] Seitz HK and Stickel F (2007). Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat Rev Cancer* **7**, 599–612.

[65] Terry MB, Gammon MD, Zhang FF, Vaughan TL, Chow WH, and Risch HA, et al (2007). Alcohol dehydrogenase 3 and risk of esophageal and gastric adenocarcinomas. *Cancer Causes Control* **18**, 1039–1046.

[66] Sano M, Aoyagi K, Takahashi H, Kawamura T, Mabuchi T, and Igaki H, et al (2010). Forkhead box A1 transcriptional pathway in KRT7-expressing esophageal squamous cell carcinomas with extensive lymph node metastasis. *Int J Oncol* **36**, 321–330.

[67] Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, and Smith I, et al (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *N Engl J Med* **353**, 1659–1672.

[68] Kang MH and Reynolds CP (2009). Bcl-2 inhibitors: targeting mitochondrial apoptotic pathways in cancer therapy. *Clin Cancer Res* **15**, 1126–1132.

[69] Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, and Brannigan BW, et al (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *N Engl J Med* **350**, 2129–2139.