Genetic Variations in Esophageal Cancer

Jin Qian    Jing-Yuan Fang
Division of Gastroenterology and Hepatology, Shanghai Institute of Digestive Disease, Renji Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China

Key Words
Esophageal cancer · Genetic variation · Single nucleotide polymorphism · Genome-wide association study

Abstract
Background: Esophageal cancer is one of the most prevalent cancers worldwide, and occurs at a relatively high frequency in China and other Asian countries. The etiology of esophageal cancer remains unclear, making its early diagnosis and treatment difficult. Summary: In recent decades, efforts using candidate gene approaches have been made to identify genetic susceptibility factors for esophageal cancer in a series of genome-wide association studies. Here, we review the latest progress in research on genetic variations in esophageal cancer cases. Key Message: Genetic variations partially account for esophageal cancer incidence and mortality rates. A comparative genomic analysis of esophageal squamous cell carcinomas and esophageal adenocarcinomas suggests little overlap genetically. Current integrated studies using high-throughput approaches have unmasked a number of novel genetic lesions in both esophageal cancer and its precursor lesions, although distinct results obtained from different regions and cohorts remain to be investigated. Practical Implications: Pooled analyses of genome-wide association studies should be conducted to establish unique susceptibility loci for specific districts and nationalities. Meanwhile, the identification of ‘driver mutations’ or mutations associated with prognosis in long-term follow-up studies is critical for the development of therapeutic and prognostic strategies.

Esophageal cancer ranks as one of the most deadly cancers, and it is the sixth leading cause of cancer death worldwide. Currently, there are limited clinical approaches for the early diagnosis and treatment of esophageal cancer, resulting in a 5-year survival rate of 15–30% [1]. Esophageal cancer has a notable geographic and racial landscape, with esophageal squamous cell carcinoma (ESCC) as the dominant histological type in China and worldwide,
while in Western countries, the rate of esophageal adenocarcinoma (EAC) has been increasing over the past decades. Global risk factors, such as cigarette smoking and alcohol consumption, only partially account for esophageal cancer incidence and mortality rates, reflecting the fact that some of the most prevalent locations in China are associated with relatively low levels of cigarette smoking and alcohol consumption. Genome-wide association studies (GWAS) have been performed to identify genomic abnormalities underlying esophageal cancer, to determine its molecular basis, and to guide the development of effective targeted therapies and diagnosis biomarkers. A comparative genomic analysis of ESCC and EAC suggests little overlap genetically; we aimed to define the mutational landscape and determine the molecular bases of EAC and ESCC, respectively, by summarizing the latest research progress.

Esophageal Adenocarcinoma

The incidence of EAC has increased 600% over the last 30 years [1]. The limited knowledge of the genomic aberrations underlying EAC has hindered the development of new therapies. EAC arises from a specialized metaplastic epithelium, which is diagnostic of Barrett’s esophagus (BE) in the context of chronic inflammation secondary to exposure to acid and bile. Consequently, a genetic component to the development of BE and EAC has long been suspected, based on previous studies in unrelated individuals and familial disease clusters.

The first GWAS on BE was conducted by Su et al. [2] in UK cohorts. They observed that 2 single nucleotide polymorphisms (SNPs), on chromosomes 6p21 (rs9257809) and 16q24 (rs9936833), were associated with BE. One of these loci (rs9257809) lies within the HLA region and the other (rs9936833) is close to FOXF1, which is involved in the esophageal structure and development. More importantly, both SNPs were also shown to be associated with a risk of EAC afterwards [3].

More recently, the Barrett’s and Esophageal Adenocarcinoma Consortium (BEACON) identified three new associations in a combined analysis of EAC and its precursor lesion BE, which indicated susceptibility SNPs within CRTC1 and BARX1, and near FOXP1 [3]. They also refined a previously reported association with BE near the putative tumor suppressor gene FOXF1 on 16q24, and extended previous findings to include EAC. Intriguingly, 2 of the other regions (BARX1/9q22.32 and FOXF1/16q24.1) contained risk-associated SNPs that disrupted the binding of FOXP1, a transcription factor that regulates esophageal development. Notably, the most significant results were for cancer and pre-cancer combined, and a large proportion of the genes affecting the risk of these 2 conditions was shared between them. Taken together, these findings suggest that much of the genetic basis for EAC lies in the development of BE rather than the progression from BE to EAC.

Based on this, Ek et al. [4] investigated the genetic architecture underlying chronic gastroesophageal reflux disease (GERD), BE, and EAC by considering all SNPs simultaneously, and found that the genetic correlation between BE and EAC was high ($r_g = \text{estimated genetic correlation} = 1.0, \ SE = \text{standard error} = 0.37$). They further estimated a statistically significant polygenic overlap between BE and EAC (one-sided $p = 1 \times 10^{-6}$), which suggests, together with the high genetic correlation, that shared genes underlie the development of BE and EAC. Conversely, no statistically significant results were obtained for GERD. Moreover, they estimated that the proportion of variation in BE risk explained by common variants was 35% ($H^2_g = \text{estimated genetic variance explained}; SE = 6\%; \ one-sided \ p = 1 \times 10^{-9}$), and in EAC, the proportion was 25% ($SE = 5\%; \ one-sided \ p = 2 \times 10^{-7}$).

To validate these findings, Palles et al. [5] performed a GWAS to identify variants associated with BE and further analyzed promising variants identified by BEACON by genotyping 10,158 patients with BE and 21,062 controls. As a result, 2 SNPs not previously associated
with BE were reported: rs3072 (2p24.1; OR = odds ratio = 1.14; 95% CI: 1.09–1.18; p = 1.8 × 10^{-11}) and rs2701108 (12q24.21; OR = 0.90; 95% CI: 0.86–0.93; p = 7.5 × 10^{-9}). The closest protein-coding genes were \textit{GDF7} (rs3072), which encodes a ligand in the bone morphogenetic protein pathway, and \textit{TBX5} (rs2701108), which encodes a transcription factor that regulates esophageal and cardiac development. Their data also confirmed the three risk SNPs identified by BEACON (rs2687201, rs11789015, and rs10423674) in BE cases. A meta-analysis of all data located another SNP associated with BE and EAC: rs3784262, within \textit{ALDH1A2} (OR = 0.90; 95% CI: 0.87–0.93; p = 3.72 × 10^{-9}).

Although the BE-EAC GWAS have not yet identified functional SNPs in each region or their gene targets, the information generated already permits the generation of hypotheses regarding processes that may be involved in BE. First, transcription factors involved in the development and structure of the thorax, diaphragm, and esophagus may be important: the SNPs near \textit{FOXF1}, \textit{FOXP1}, and \textit{TBX5} might act in this way and the genes appear to be functionally related. Second, the inflammatory response may be important: the SNPs within the HLA region might influence this pathway. A plausible testable hypothesis is that these 2 groups of SNPs influence the tendency to GERD, perhaps through the thoracic and diaphragmatic structure (hiatal hernia defect), and an inflammatory response to refluxed gastric acid.

Given the limited data on the molecular signatures and pathways involved in the tumorigenesis of EAC, it was intriguing to analyze the mutation spectra of EAC tumors. As Dulak et al. \cite{6} reported, 149 EAC tumor-control pairs were subjected to whole-exome sequencing, 15 of which had also been subjected to whole-genome sequencing. They firstly established that the highest rate of A>C transversions in EAC were in non-coding areas, and within coding areas, they were overrepresented in less expressed genes. A subsequent statistical analysis of the exome data identified 26 significantly mutated genes, the most significant of which were 2 known EAC tumor suppressors, \textit{TP53} and \textit{CDKN2A}. The novel significantly mutated genes included chromatin-modifying factors and regulators of invasion and motility: \textit{SPG20}, \textit{TLR4}, \textit{ELMO1}, and \textit{DOCK2}. Surprisingly, in many cases, it was impossible to generate hotspot mutations, such as KRAS or PIK3CA with an AA transversion. Functional analyses of EAC-derived mutations in \textit{ELMO1} revealed an increased cellular invasion. Given that EAC is a highly invasive tumor prone to early metastasis, alterations in the RAC1 pathway may be a contributor to EAC tumorigenesis. To the best of our knowledge, this was the first GWAS to investigate gene variations together with pathway abnormalities in EAC.

Data from high-density SNP arrays and exome-sequencing studies have accumulated, with a plethora of mutations identified in many different genes. However, little work has focused on the precise ordering of these alterations in large cohorts of patients with premalignant disease and associated clinical follow-up data. It is becoming increasingly clear that extensive genomic heterogeneity is present in the majority of advanced cancers, suggesting that the most appropriate therapeutic targets are those mutations that occur early in the development of disease and are thus clonal in the resulting malignancy. The identification of causative mutations occurring early in the pathogenesis is also pivotal to developing clinically useful biomarkers. In this context, mutations occurring at disease-stage boundaries, for example, the transition from nondysplastic epithelium to dysplasia and then to cancer, would be the most informative. To understand at what stage in the disease process mutations occur, Agrawal et al. \cite{7} performed exome sequencing on 11 EAC samples and 2 samples of BE adjacent to the cancer. Intriguingly, the majority of mutations were found to be present even in apparently normal BE. Consistent with these findings, using whole-genome sequencing and amplicon resequencing of 112 EAC cases, Weaver et al. \cite{8} screened their panel for the 26 most recurrently mutated genes in samples from different stages of carcinogenesis: 90 EAC, 66 never-dysplastic BE, and 43 high-grade dysplasia (HGD) cases. Contrary to expectations, the most prevalent gene mutations in EAC were also present at a similar frequency in...
HGD and never-dysplastic BE samples, including mutations within cancer-associated genes, for example, ARID1A and SMARCA4. Only mutations in TP53 and SMAD4 were confined to HGD and EAC cases, making them surveillance candidate markers for malignant progression risk. Given the low frequency of SMAD4 mutations (13%) and that the mutation was specific to EAC but not HGD samples, they focused on TP53 in clinical applications, i.e., using the noninvasive, nonendoscopic, cell-sampling device, the Cytosponge, coupled with high-throughput sequencing. This test showed 100% specificity and 86% sensitivity in differentiating between patients with HGD and those without dysplasia.

From a clinical perspective, this was an important and groundbreaking study. It was remarkable that Cytosponge could become an alternative candidate gene approach with the goal of identifying clinical biomarkers to complement histological examination, which is an approach fraught with difficulties. Besides, the fact that most 'passenger gene mutations' but not 'driver mutations' were present in lesions that appeared to be very benign and histologically stable also had wide implications for both diagnosis and therapy.

**Esophageal Squamous Cell Carcinoma**

About 50% of all ESCC cases occur in China. Overall, the somatic mutation rate for ESCC is less than that for EAC [7]. Moreover, the proportion of genetic variations in ESCC risk might differ by districts. Global risk factors, such as cigarette smoking and alcohol consumption, account for more than 90% of population-attributable risks in the United States, but these factors only account for 46% of ESCC incidence and mortality risks in China.

Initial findings through large-scale genome-wide interrogation have provided an understanding of the basic biology underlying ESCC susceptibility alleles and have highlighted gene variation interaction with environmental factors. To identify genetic susceptibility loci for ESCC, a GWAS was performed by Wu et al. [9] on 2,031 individuals with ESCC (cases) and 2,044 controls of Chinese descent using 666,141 autosomal SNPs, with a validation of promising associations in an additional 6,276 cases and 6,165 controls of Chinese descent from different areas of China. As a result, they identified 7 susceptibility loci on chromosomes 5q11, 6p21, 10q23, 12q24, and 21q22 (ranging from $p = 7.48 \times 10^{-12}$ to $p = 2.44 \times 10^{-31}$). Three variants in high linkage disequilibrium on 12q24 conferred their risks of ESCC in a gene–lifestyle interaction manner, with more pronounced risk enhancement seen in tobacco and alcohol users. Furthermore, the identified variants had a cumulative association with ESCC risk ($p_{\text{trend}} = 7.92 \times 10^{-56}$) [9]. To identify further genetic loci in ESCC, another GWAS [10] was conducted and a genome-wide gene–environment interaction analysis of ESCC in 2,031 affected individuals (cases) and 2,044 controls; the results were independently evaluated in 8,092 cases and 8,620 controls. Six new ESCC susceptibility loci were confirmed, 4 of which, at chromosomes 4q23, 16q12.1, 22q12, and 3q27, had a significant marginal effect ($p = 1.78 \times 10^{-39}$ to $p = 2.49 \times 10^{-11}$) and 2, at 2q22 and 13q33, had a significant association only in the gene–alcohol drinking interaction ($p_{G \times E} = \text{gene–environment interaction } p = 4.39 \times 10^{-11}$ and $4.80 \times 10^{-8}$, respectively). Variants at the 4q23 locus, which includes the ADH cluster, each had a significant interaction with alcohol drinking in their association with ESCC risk ($p_{G \times E} = 2.54 \times 10^{-7}$ to $p_{G \times E} = 3.23 \times 10^{-2}$). The authors confirmed the known association of the ALDH2 locus on 12q24 with ESCC, and a joint analysis showed that drinkers with both of the ADH1B and ALDH2 risk alleles had a 4-fold increased risk for ESCC compared with drinkers without these risk alleles. These serial researches underscore the direct genetic contribution to ESCC risk, as well as the genetic contribution to ESCC through interaction with environmental factors.

To date, few studies on genetic variants in ESCC have been published, which mainly focus on the identification of candidate genes in ESCC. The candidate gene approach relies on
whole-exome/whole-genome interrogation and an analysis of somatic copy number variants (SCNVs) as a powerful and successful tool to identify genetic variants and the related pathways. Indeed, using this approach, several loci, some of which are located in genes involved in cell cycle and apoptosis regulation, have been defined.

Wu et al. [11] carried out the first genome-wide interrogation of genetic variants associated with length of survival in 1,331 ESCC cases, for whom death or survival information was available, followed by replication in two independent data sets consisting of 1,962 cases. They identified rs1050631 in SLC39A6 as being associated with the survival times of affected individuals, with a hazard ratio of 1.30 for death from ESCC in the combined sample (95% CI: 1.19–1.43; p = 3.77 × 10⁻⁸). A G>A transition at rs724248, which is located in the 5′ untranslated region (5′ UTR) of the gene, leads to an increased constitutive expression of SLC39A6, probably owing to the disturbance of a transcriptional repressor binding site. Subsequent immunohistochemical staining of ESCC tissues and functional experiments in ESCC cells also indicated the involvement of SLC39A6 in ESCC tumorigenesis, suggesting that SLC39A6 may be a potential therapeutic target and prognostic marker of ESCC. Lin et al. [12] reported previously uncharacterized mutated genes such as FAT1, FAT2, ZNF750, and KMT2D, in addition to those already known (TP53, PIK3CA, and NOTCH1) using whole-exome or targeted deep sequencing of 139 paired ESCC cases and the analysis of SCNVs of over 180 ESCC cases from China. Further SCNV evaluations, immunohistochemistry and biological analyses suggested their functional relevance in ESCC. Their approaches also identified a potential therapeutic target, XPO1, because it showed both gene mutation and protein overexpression. As part of the International Cancer Genome Consortium research project, Song et al. [13] described a comprehensive genomic analysis of 158 ESCC cases in the Chaoshan district of Guangdong province, an area of high ESCC prevalence in China. They conducted whole-genome sequencing in 17 ESCC cases and whole-exome sequencing in 71 cases, of which 53 cases, plus an additional 70 ESCC cases not used in the whole-genome and whole-exome sequencing, were subjected to array comparative genomic hybridization analysis. By this means, they identified 8 significantly mutated genes, 6 of which are well-known tumor-associated genes (TP53, RB1, CDKN2A, PIK3CA, NOTCH1, and NFE2L2), 1 is a tumor-associated gene (ADAM29) that has not previously been described in ESCC, and 1 gene (FAM135B) has not been linked to cancer previously. Notably, FAM135B (family with sequence similarity 135, member B) was mutated in 6.8% (6 of 88) of cases and was associated with poor prognosis in ESCC (p = 0.026, log-rank test). It was also identified as a novel cancer-implicated gene as assayed for its ability to promote malignancy of ESCC cells. Gao et al. [14] defined the mutational landscape of ESCC in a larger cohort of northern Chinese cases. They performed exome sequencing on 113 tumor-control pairs, yielding a mean of 82 nonsilent mutations per tumor, in 8 cell lines. The mutational profile of ESCC closely resembled that of squamous cell carcinomas of other tissues, but differed from that of EAC. Genes involved in the cell cycle and apoptosis regulation were mutated in 99% of all cases by somatic alterations of TP53 (93%), CCND1 (33%), CDKN2A (20%), NFE2L2 (10%), and RB1 (9%).

Researchers have also investigated the mutational pathways of ESCC and highlighted mutations in epigenetic modulators with prognostic and potentially therapeutic implications. Lin et al. [12] reported that, in the RTK-MAPK-P13K pathways, the cell cycle and epigenetic regulation are frequently dysregulated by multiple molecular mechanisms in ESCC. Moreover, Song et al. [13] detected frequent nonsilent mutations in 48 histone modification-related genes in 53.4% of ESCC cases, within which several important histone regulator genes (MLL2 – also called KMT2D –, ASH1L, MLL3 – also called KMT2C –, SETD1B, CREBBP, and EP300) are frequently altered. Pathway assessment revealed that somatic aberrations are mainly involved in the Wnt, cell cycle and Notch pathways. According to the report by Gao et al. [14], cell cycle, apoptosis and DNA damage control pathways are ubiquitously dysregu-
lated, largely because of the contribution of mutations in TP53 in 93% of all cases, followed by mutations affecting the chromatin modification, Notch, phosphoinositide 3-kinase (PI3K), and Ras pathways. Histone-modifier genes are frequently mutated, including KMT2D (also called MLL2; 19%), KMT2C (also called MLL3; 6%), KDM6A (7%), EP300 (10%), and CREBBP (6%). The Hippo and Notch pathways are dysregulated by mutations in FAT1, FAT2, FAT3, or FAT4 (27%) or AJUBA (JUB; 7%) and NOTCH1, NOTCH2, or NOTCH3 (22%) or FBXW7 (5%), respectively. Cases with EP300-mutated tumors have a dismal overall survival rate. In summary, these studies provide a comprehensive mutational landscape of ESCC and identify mutations in histone modifiers and multiple pathways with potentially therapeutic and even prognostic implications.

Current integrated studies using high-throughput approaches have unmasked a number of novel genetic lesions in ESCC and have provided an important molecular foundation, including involved pathways, to understand esophageal tumors and to develop therapeutic targets. However, with respect to future GWASs on ESCC tumorigenesis, a few proposals should be taken into account. Firstly, sequencing of ESCC samples from different parts of China has pinpointed several, recurrently mutated, ESCC-relevant genes and led to the suggestion that ESCC cases from different geographic regions might differ genetically [15]. Given the distinct results among different cohorts, etiological heterogeneity might have an important role in interpreting GWAS results, and should be considered when GWASs are extended to study populations with distinct lifestyles. Consistently, environmental factors have been shown to have varied relevance for ESCC in different Chinese populations, and this variation might have led to current differential GWAS findings. In addition, a recent joint analysis of 3 GWASs of ESCC drawn from distinct locations in China uncovered 2 new loci which achieved genome-wide significance in the pooled data of individual genotypes for all 3 GWASs and 2 replication studies [15]. In a further analysis, a new locus in the HLA class II region at 6p21.32 (rs35597309) showed geographic differences in its effect. Its association with ESCC was restricted to the 2 populations in the Taihang Mountain region at highest risk of ESCC, where the total mortality rate from ESCC and gastric cardia cancer could exceed 20%. These results provide evidence that pooled analyses should be conducted to establish unique susceptibility loci for specific districts and nationalities. Secondly, until recently, little has been known about genetic loci associated with length of survival in ESCC, based on genome-wide examinations [11]. However, a genome-wide interrogation of genetic variants associated with length of survival in ESCC individuals might provide the most promising therapeutic targets and novel prognosis markers. Furthermore, the long-term follow-up of cases and the availability of detailed clinical information have increased the statistical power to detect a genetic effect and have decreased the potential influence of confounding factors on survival times. Lastly, although many gene variations have been identified as significantly associated with ESCC tumorigenesis, many SNPs in non-coding regions and the exact function of the identified mutations have not been thoroughly investigated, suggesting that there is a marked preference for the most common type of mutation among all alterations, C>T transition at CpG dinucleotides (comprising up to 85% of all CpG dinucleotide alterations) [13]. Thus, further studies should focus on the biological and therapeutic significance of both the newly discovered mutated/amplified genes and the non-coding SNPs, which may ultimately lead to the development of effective diagnostic and therapeutic approaches for ESCC.

**Disclosure Statement**

The authors declare no conflicts of interest.
References

1. Holmes RS, Vaughan TL: Epidemiology and pathogenesis of esophageal cancer. Semin Radiat Oncol 2007; 17: 2–9.
2. Su Z, Gay LJ, Strange A, Palles C, Band G, Whitemen DC, et al: Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett’s esophagus. Nat Genet 2012; 44: 1131–1136.
3. Levine DM, Ek WE, Zhang R, Liu X, Onstad L, Sather C, et al: A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett’s esophagus. Nat Genet 2013; 45: 1487–1493.
4. Ek WE, Levine DM, D’Amato M, Pedersen NL, Magnusson PK, Bresso F, et al: Germline genetic contributions to risk for esophageal adenocarcinoma, Barrett’s esophagus, and gastroesophageal reflux. J Natl Cancer Inst 2013; 105: 1711–1718.
5. Palles C, Chegwidden L, Li X, Findlay JM, Farnham G, Castro Giner F, et al: Polymorphisms near TBX5 and GDF7 are associated with increased risk for Barrett’s esophagus. Gastroenterology 2015; 148: 367–378.
6. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, et al: Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet 2013; 45: 478–486.
7. Agrawal N, Jiao Y, Bettegowda C, Hutfless SM, Wang Y, David S, et al: Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. Cancer Discov 2012; 2: 899–905.
8. Weaver JM, Ross-Innes CS, Shannon N, Lynch AG, Forshew T, Barbera M, et al: Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. Nat Genet 2014; 46: 837–843.
9. Wu C, Hu Z, He Z, Jia W, Wang F, Zhou Y, et al: Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat Genet 2011; 43: 679–684.
10. Wu C, Kraft P, Zhai K, Chang J, Wang Z, Li Y, et al: Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. Nat Genet 2012; 44: 1090–1097.
11. Wu C, Li D, Jia W, Hu Z, Zhou Y, Yu D, et al: Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. Nat Genet 2013; 45: 632–638.
12. Lin DC, Hao JJ, Nagata Y, Xu L, Shang L, Meng X, et al: Genomic and molecular characterization of esophageal squamous cell carcinoma. Nat Genet 2014; 46: 467–473.
13. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, et al: Identification of genomic alterations in oesophageal squamous cell cancer. Nature 2014; 509: 91–95.
14. Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, Sun ZM, et al: Genetic landscape of esophageal squamous cell carcinoma. Nat Genet 2014; 46: 1097–1102.
15. Wu C, Wang Z, Song X, Feng XS, Abnet CC, He J, et al: Joint analysis of three genome-wide association studies of esophageal squamous cell carcinoma in Chinese populations. Nat Genet 2014; 46: 1001–1006.