# Whole-Exome Sequencing Reveals Critical Genes Underlying Metastasis in Esophageal Squamous Cell Carcinoma

Wei Dai[1], Josephine Mun Yee Ko[1], Sheyne Sta Ana Choi[1], Zhouyou Yu[1], Luwen Ning[1], Hong Zheng[1], Vinod Gopalan[2], Nikki Pui-yue Lee[3], Kwok Wah Chan[4], Simon Ying-Kit Law[3], Alfred King-Yin Lam[2], Maria Li Lung[1]*

[1]Department of Clinical Oncology, University of Hong Kong, Hong Kong (SAR), People's Republic of China; [2]Department of Cancer Molecular Pathology, Griffith Medical School and Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia; [3]Department of Surgery, University of Hong Kong, Hong Kong (SAR), People's Republic of China; [4]Department of Pathology, University of Hong Kong, Hong Kong (SAR), People's Republic of China

***Corresponding author:** Maria Li Lung
Address: Department of Clinical Oncology, University of Hong Kong, Room L6-43, 6/F, Laboratory Block, Faculty of Medicine Building, 21 Sassoon Road, Pokfulam, Hong Kong

Tel: (852) 3917 9783 Fax: (852) 2816 6279

**Abstract**

**Background and aims:** Esophageal squamous cell carcinoma (ESCC) is one of the most lethal cancers due to the high frequency of local metastasis. However, the genomic landscape of the metastatic ESCC has not been well-characterized. We aim to identify the genetic alterations that underlie metastasis in ESCC. **Methods:** We performed whole-exome sequencing (WES) for 42 patients with primary ESCC and 15 with metastatic ESCC in lymph nodes. Eleven ESCC cases had triple samples including matched primary cancer, metastatic cancer in LN and non-neoplastic mucosa. **Results:** In metastatic ESCC, there is a tendency for a weaker APOBEC-mediated signature. Mutations of four genes, namely *TP53*, *KMT2D*, *ZNF750* and *IRF5,* were frequently found in the metastatic ESCC. Importantly, loss-of-function mutations in *ZNF750* recurrently occurred in the metastatic ESCC and were strongly associated with lymph node metastasis in the primary tumors, suggesting a role of *ZNF750* as a metastasis suppressor. In addition, mutations of epigenetic regulators, including *KMT2D*, *TET2* and *KAT2A* as well as deletion of histone variants on chromosomes 6p22 and 11q23 were detected in over half of the metastastic ESCCs. Also, copy number gain of the *TERT* region was less frequently observed in these cases. **Conclusions:** A number of critical genetic events including *TP53* putative gain-of-functionmissense mutations, mutations of differentiation regulators *ZNF750* and *IRF5,* as well as nucleosome disorganization caused by genetic lesions, may play an important role in pathogenesis of metastases in ESCC.

**(249 words)**

## Introduction

Esophageal cancer is one of the most aggressive cancers and often presents with local invasion or regional lymph node (LN) metastasis at diagnosis[1]. The prognosis of patients with esophageal cancer is poor with 5-year survival rates <50%, after curative resection[2]. The cancer incidence exhibits remarkable worldwide geographical differences. In the Taihang Mountains of Northern China, the incidence rate is up to 133 per 100,000[3], while Hong Kong is a moderate risk region[4] with an incidence rate in males of 10 per 100,000[5]. In the endemic regions, esophageal squamous cell carcinoma (ESCC) is the most common histological type, accounting for approximately 90% of all cases.

The primary ESCC tumors have been well-characterized by whole-exome sequencing (WES) and/or whole-genome sequencing (WGS) in high-risk (Taihang Mountains and Chaoshan District)[6, 7] and low-risk (Beijing)[8, 9] regions of Mainland China, and in Japan[10]. The recent landscape paper from The Cancer Genome Atlas (TCGA) study included 90 ESCC cases from Asian and European populations[11]. A number of driver mutations were identified. However, the molecular profiles of the metastatic ESCC remain unclear and the genetic alterations contributing to metastasis have not yet been well-characterized.

In this study, we present the landscape of ESCC somatic alterations from patients of Hong Kong, characterize the metastatic ESCC in LNs by WES analysis and make a direct comparison between the metastasis and matched primary ESCC. This study expands the registry of somatically-disrupted driver genes and reveals the critical genetic events underlying ESCC metastasis.

## Methods and materials

### *Study subjects*

Tumor samples and histologically non-neoplastic mucosae at surgical resection margins were obtained from 46 ESCC patients from Queen Mary Hospital. Institutional Review Board (IRB) approval was obtained and all the study subjects signed the consent form. **Supplementary Table S1** details a description of the clinical characteristics of the cases. Fresh tumor samples from untreated patients were obtained at the time of surgery. Tumor cell contents from hematoxylin and eosin-stained frozen sections were estimated by a pathologist (AKY Lam). Only samples with a high percentage of tumor cells were selected for this study (**Supplementary Table S2**). Over 75% of the samples contained at least 60% tumor. A total of 57 tumors from 46 ESCC cases were sequenced, including 42 primary carcinomas and 15 LNs with metastatic carcinoma. For 11 of 46 ESCC cases, paired primary and matched LNs with metastatic carcinomas were available.

### *Whole-exome sequencing*

Genomic DNAs were extracted from tumors and non-neoplastic mucosae at proximal surgical resection margins using the AllPrep DNA/RNA Mini kit (Qiagen). The extracted DNA was analyzed on 0.7% agarose gels and the quality and quantity were examined by Nanodrop 1000 (Thermo Scientific), Qubit (Life Technologies) and a bioanalyzer (Agilent), respectively. The library preparation, capture, and sequencing were performed by the Centre for Genomic Sciences at the University of Hong Kong. In brief, 250 ng genomic DNA were fragmented by an ultrasonicator (Covaris). These fragments were amplified using NEBNext Ultra$^{TM}$ DNA library Prep Kit (NEB) and then hybridized to the Illumina TruSeq capture kit for enrichment. Paired end, 100 bp read-length sequencing was performed using the HiSeq 1500 sequencer (Illumina).

### *WES data analysis: detection of SNVs and small indels*

Clean sequencing reads were aligned to the human genome (hg19) with BWA[12]. Picards were applied to sort reads and mark duplicates. The paired or triple samples from the same cases were merged together. Subsequently, GATK[13] was applied for local realignment around indels with base quality recalibration according to GATK Best

Practices recommendations[14, 15]. After base quality recalibration, the merged reads were split for individual samples. The germline variants were identified by GATK and then examined for the number of singletons and to determine correct matching of samples from the same cases. SNVs were detected by Mutect[16] and Varscan2[17]. Small indels were detected by Varscan2. In Mutect, we removed the SNVs with less than 5 reads supporting a mutant allele in tumor samples. In Varscan2, we only considered the high-confidence SNVs and indels with variant allele frequency in the tumors >10%. SNVs with minor allele frequency (MAF) >1% in 1000 genomes and ESP6500 databases were further removed from the analysis. Recurrent somatic mutations and low-quality mutations (mutant allele frequency <10%) were manually checked in IGV alignment to further remove false positives based on mapping quality, base Phred quality score, and quality of adjacent regions. Loss-of-function (LOF) mutations were defined as stopgain, splicing, and frameshift indels.

***Detection of SCNVs***

We performed ADTEx[18] to detect the somatic copy number variations (SCNVs). ADTEx is tailored for WES data, to infer SCNVs in ESCC genomes on the basis of normalized ratio of WES data from tumor and matched non-neoplastic mucosae.

***RNASeq and data analysis***

Four tumor pairs were subjected to RNASeq. Total RNAs were extracted by the AllPrep DNA/RNA Micro Kit (Qiagen). The ribosomal RNAs were depleted using the Ribo-Zero Magnetic Kit (Illumina). The Kapa Stranded RNA-seq Library Preparation kit was used for the library preparation and the libraries were subjected to high-throughput sequencing using the HiSeq 1500 sequencer with 100 bp paired-end reads.

Clean reads were aligned to the human reference genome (hg19) using Tophat (version 2.1.0 with Bowtie version 2.2.4), and the gene expression level (FPKM, fragments per kilobase per million fragments mapped) was calculated by Cufflinks[19]. For each gene, the median FPKM value is summarized across all the samples.

***Statistical analysis***

Chi-square or Fisher's exact test was used to examine the difference of mutation frequency between groups. The difference of the quantitative data was tested by Mann-Whitney U test. In survival analysis, Kaplan-Meier analysis was used for survival curves with log-rank test for comparison between groups. A univariate Cox proportional hazards model was used to examine associations between overall survival and genetic alterations, as well as other clinical parameters. In the multivariable survival analysis, this association between genetic alteration and overall survival was adjusted by clinical parameters including stage and age. A p value $<0.05$ or false discovery rate (FDR) $<0.20$ in the case of multiple test correction was considered statistically significant. The power estimation for the analysis, publicly available microarray and mutation data analyses, identification of driver mutations and enrichment test utilized are detailed in the **supplementary methods**.

***Sanger sequencing***

Sanger sequencing was used to verify the selected mutations identified from the WES data. PCR amplification was performed using FastStart Taq DNA polymerase with 20 ng of genomic DNA as template. The primers are listed in **Supplementary Table S3**.

***qPCR copy number analysis***

Copy number of *CCND1*, *RNF168* and histone variants on 6p22.1 was assessed using frozen tumors and matched non-tumor tissues. Copy number was determined by quantitative PCR with DNA binding dye SYBR green and using the specific primer pairs (**Supplementary Table S3**). Gene *METTL21A* was used as a diploid control. We selected this control according to the WES data (**Supplementary Figure S1**). The copy number changes were estimated using the delta-delta Ct method[6].

**Results:**

***Identification of somatic mutations in the primary tumors and metastatic LNs***

The median average on-target coverage was 76-fold, 79-fold and 71-fold for the non-neoplastic mucosae, primary tumors and LNs with metastatic ESCC, respectively (**Supplementary Table S2**). In 41 primary tumors, the median mutation rate is 3.6 mutations per megabase (Mb) (range 1.5-11.2) in protein-coding regions and 2.8 non-silent mutations per Mb (range 0.9-7.2) (**Supplementary Table S4**). This non-silent mutation rate is comparable to recently published mutation rates in ESCC from Chinese patients (2.4-2.9 mutations per Mb)[6, 7, 9]. In 15 LNs with metastatic ESCC, the median mutation rate is 3.6 per Mb (range 1.1-7.6) in protein-coding regions and 2.6 non-silent mutations per Mb (range 0.7-5.4) (**Supplementary Table S5**). There is no statistically significant difference in the mutation rate between primary cancers and metastatic LNs (**Figure 1A**, **Supplementary Figure S2**). Sanger sequencing on a subset of variants confirmed 96.4% (54 variants) somatic variants identified by WES (**Supplementary Table S6**).

In the primary tumors, *TP53* (90.5%), *KMT2D* (23.8%), *CDKN2A* (11.9%) and *ZNF750* (9.5%) were significantly mutated with at least three non-silent mutations (MutSigCV $p<0.001$) (**Figure 1B, Supplementary Table S7**). In addition, multiple genes reported previously, including *FAM135B* (11.9%), *NOTCH1* (9.5%), *TET2* (9.5%), *NFE2L2* (7.1%), *RB1* (7.1%), *FAT1* (7.1%) and *NSD1* (7.1%) were mutated in at least 5% of the cases. In the metastatic LNs, four genes *TP53* (67%), *KMT2D* (40%), *ZNF750* (13%). and *IRF5* (13%) were mutated in at least two cases with MutSigCV $p<0.001$ (**Figure 1B, Supplementary Table S8**). There is no pronounced difference in the mutation frequency of these genes between the primary and metastatic ESCCs (**Figure 1C**). The gene mutation frequencies were also similar between the primary tumors with and without LN metastasis (**Figure 1D**).

We further examined the mutation profiles of eleven cases with matched primary tumors and metastatic ESCC in LNs. The results showed that 50-80% of the protein-altered mutations in the metastatic ESCC can be detected in the matched primary tumors (**Supplementary Figure S3**). The mutations of the putative driver genes were generally found in both primary and metastatic ESCCs (**Supplementary Figure S4**). *SYNE1* was

recurrently mutated in metastatic ESCCs, while the reads supporting the mutations can be detected in the primary tumors at a much lower allele frequency and, thus, were rejected by Mutect and Varscan2 (**Supplementary Figure S5**). This result shows that the tumor cells in the metastatic ESCCs with *SYNE1* mutations are likely to originate from a small subclone of the primary tumors.

***Mutational signatures in ESCC primary and metastatic LNs***

Similar to the previous findings for ESCC, two predominant signatures were identified in the primary tumors. Signature A is characterized as substitutions in CpG dinucleotides due to an increased rate of spontaneous 5-methyl-cytosine deamination. Signature B is characterized as substitutions involving C to G/T in TpCpX trinucleotides (**Figure 2A and B**), which was associated with the APOBEC-mediated mutations[20]. Interestingly, the APOBEC-mediated signature was weaker in the metastatic ESCC, when compared to primary tumors (p=0.025, **Figure 2C and D**, **Supplementary Figure S6**). We did not detect decreased tumor contents or coverage in metastatic ESCC LNs , when compared to primary tumors (**Supplementary Figure S7**). Thus, the weaker APOBEC-mediated signature in the metastatic ESCCs was not caused by differences in tumor contents and sequencing depth.

***TP53 mutations in ESCC primary and metastatic LNs***

*TP53* was the most significantly mutated gene in the primary (90.5%) and metastatic ESCC (66.7%). Recurrent missense mutations p.R175H, p.Y220C, p.R248Q/W and p.R273C/H occurred in at least 5% of the primary tumors (**Figure 3A**). The hot spots for *TP53* missense mutation in the Hong Kong cohort were slightly shifted, when compared to the ESCC cases from high-risk regions (Taihang mountains and Chaoshan district), but were similar to the combined ESCC cases from high- and low- risk regions (**Figure 3B**, **Supplementary Figure S8**), suggesting that the ESCC of Hong Kong cases are a mixed cohort from different risk regions. Interestingly, the patients with *TP53* missense mutations had much shorter survival times than patients with LOF mutations or without mutation (**Figure 3C**). The association between TP53 missense mutations and overall survival is independent from clinical parameters including stage and age (adjusted

HR=3.36, 95% CI: 1.51-7.49, p=0.003), while there is no statistical difference of survival between cases with LOF mutations or other mutations (**Supplementary Table S9**). In the metastatic ESCC in LNs, missense mutations p.R273H, p.R273C, p.V272M, p.I255F, p.V157A, p.F109C, truncating mutations p.R209* and p.S183*, frameshift indels p.T230fs and p.L93fs were identified (**Figure 3A**).

***ZNF750 LOF mutations in LN metastasis***

*ZNF750* is a putative tumor suppressor in ESCC[6, 8]. In this study, it was mutated in both primary (9.5%) and metastatic ESCC (13%) (MutSigCV p<0.001). We identified three LOF and one missense mutation at *ZNF750*. We further examined 33 LOF *ZNF750* mutations identified from a total of 630 primary tumors in current and previous studies[6-10], accounting for 6.5% of the cases (**Figure 4A**). The result confirmed that these mutations occurred at a higher frequency in primary ESCCs with LN metastasis (10% vs. 2.25%, p=0.0004, **Figure 4B**). In our samples, the *ZNF750* LOF mutations including p.Y302* and p.F213fs were found in two metastatic ESCCs and matched primary tumors. *ZNF750* deletion was observed in these two ESCCs carrying *ZNF750* LOF mutations (**Figure 4C**), suggesting this gene was subjected to biallelic inactivation via mutation and deletion. At the mRNA level, downregulation of *ZNF750* was found in over 80% of the cases (**Supplementary Figure S9**). Gene set enrichment analysis (GSEA) using the publicly available microarray data (GSE23400, n=53) indicates that *ZNF750* expression correlated with the gene sets upregulated after knockdown of the polycomb repressive complex (PRC) 1 protein BMI1 and PRC2 protein SUZ2 (**Figure 4D**). This result implies its role in inhibition of cancer cell stemness. In addition, the ESCC survival analysis in from the TCGA study showed that the cases with genetic alterations including point mutations and copy number variations (CNVs) at *ZNF750* had shorter disease relapse-free survival, when compared to the other cases (log-rank test p=0.0006, **Supplementary Figure S10**).

***Identification of SCNAs in the primary and metastatic LNs***

High-frequency recurrent SCNAs (>45% of the cases affected) were identified including 705 genes amplified and 84 genes deleted in 42 primary tumors (**Supplementary Table S10**) and 1106 genes amplified and 208 genes deleted in 15 metastatic ESCC in LNs (**Supplementary Table S11**). The deletions frequently occurred in 3p21, 6p22 (histone variants cluster 1), 9p21 (*CDKN2A*/*CDKN2B*), and 15q13-15, and amplifications were often identified in 3q12-13, 3q21-29 (*PIK3CA/TERC/SOX2/TP63*), 5p14-15 (*TERT*), 8q21-24 (*MYC*), 11q13 (*CCND1*/*FGFs*/*SHANK2*) and 18p11 in both primary and metastatic ESCCs (**Figure 1C and Figure 5**). The qPCR copy number analysis was used to evaluate the accuracy of the SCNAs in *CCND1* and histone variants from 6p22, and we confirmed these SCNAs in 83.3% of the samples (**Supplementary Table S12**).

***Nucleosome core deletion in late-stage primary tumors and metastatic LNs***

We further investigated the genes frequently deleted in metastatic ESCC in LNs. Out of 208 genes deleted in at least 7 out of 15 cases, genes encoding nucleosomal histones were significantly enriched (adjusted $p=7.0\times10^{-9}$), but this enrichment was not observed in 84 genes deleted in the primary tumors. These genes include a number of histones mainly from histone cluster 1 on 6p22 and H2A histone variant (*H2AFX*) on 11q23.3. We detected the 6p22.1-22.2 deletion in 38.1% of the primary tumors and 53.3% of ESCC in LNs **(Figure 1C**). A significant increase of deletion frequency was observed in the late-stage primary, as well as metastatic ESCCss when compared to the early-stage primary ESCCs (Fisher's exact test p=0.007 and p=0.012, respectively) (**Figure 6A, B and C**). The cases harboring the 6p22 deletion or 11q23 deletion at *H2AFX* had much shorter overall survival than those without deletion in these two regions (**Figure 6D**), but this association is not independent from stage (**Supplementary Table S9**).

***Copy number gain in telomerase subunit genes in primary tumors and metastatic LNs***

Telomerase is an enzyme complex that adds telomeric repeats to the ends of chromosomes. The core components of telomerase include telomerase reverse transcriptase (TERT) and the telomerase RNA component (TERC)[21]. *TERT* is located at chromosome 15p13.33 and amplification of *TERT* region has been reported in ESCC[11].

*TERC* maps to chromosome 3q26. In the Hong Kong cohort, copy number gain at *TERT* and *TERC* was observed in 66.7% and 59.5% of the primary tumors, respectively. Noticeably, the frequency of *TERT* copy number gain was reduced to 20% of the metastatic ESCC in LNs (p=0.028, FDR=0.18). On the other hand, *TERC* copy number gain was frequently observed in 66.7% of the metastatic cases (**Supplementary Table S10**).

**Discussion**

In this study, we investigated the genomic changes in the primary tumors and LNs with metastatic cancers in Hong Kong patients with ESCC. The mutation rates in the metastatic cancers were similar to those in the primary tumors. Generally, the mutations of the putative driver genes in the metastatic ESCC can be detected in the matched primary tumors. This could partially explain why ESCC is biologically so aggressive. The APOBEC-mediated signature was weaker in the metastatic cancers, although it is predominantly found in the primary tumors as reported in previous studies[6, 8], as well as by us. The role of APOBEC-mediated signature in cancer metastasis has not been well-studied. It is reported that the APOBEC-attributable mutations may be early as well as late events during tumor evolution depending on the cancer type[22]. Several APOBEC members, including *APOBEC3A*, *APOBEC3B* and *APOBEC1,* may contribute to this mutation signature[20]. In colon cancer, *APOBEC3G*, a member of a cluster of proteins of which APOBEC1 is a prototype, was highly expressed in hepatic metastasis. It enhanced colon cancer cell migration and invasion. Thus, APOBEC1promoted the liver metastasis in colon cancer[23]. In ESCC, dramatic increase of *APOBEC3B* expression was observed[8], while both *APOBEC3A* and *APOBEC3G* expression showed no difference between the tumor and non-neoplastic tissues (**Supplementary Figure S11**). Our data suggest that *APOBEC3B* may play a minor role for pathogenesis of metastasis in ESCC.

*TP53* is frequently mutated in the primary and metastatic ESCCs. In ESCC, the hot spots for *TP53* missense mutations are within the DNA binding domain. Previously, Lam *et al*. characterized *TP53* mutations in patients with primary ESCC from Hong Kong by Sanger sequencing[24]. Similar to their findings, the mutations at codons p.R248 and p.R273 were common in our cohort. In addition, we found frequent mutations at codons p.R175 and p.Y220 in 7% of the cases. Lam et al. reported that *TP53* mutations appeared to have a role in predicting clinical outcome of patients with ESCC, although their result did not reach statistical significance[24]. In this study, we now further categorized *TP53* mutations as missense or LOF and demonstrated a significant association between *TP53* missense mutations and shorter overall survival of patients with ESCC. The association between the TP53 missense mutations and shorter overall survival indicate the important functional impact of these missense mutations in ESCC disease progression. Validation

of this association in a larger cohort of patients with ESCC in Hong Kong is warranted. Two missense mutations identified in the primary tumors were also found in the metastatic ESCC at codon p.R273 (p.R273C and p.R273H). Previous studies showed that the p.R273H and p.R273C mutants enhanced cell proliferation, invasion and drug resistance in lung and breast cancer cell lines *in vitro*[25], and knock-in of p.R273H increased incidence of highly metastatic carcinomas for ovarian cancer *in vivo*[26]. The functional roles of these two mutations in ESCC remain unclear. Our data and previous studies imply that these two mutations are likely to be GOF mutations and contribute to metastasis in ESCC.

*IRF5* is a transcription factor and plays an important role to regulate B-cell differentiation[27] and to modify the tumor immune microenvironment by regulating lymphocyte infiltration[28]. Mutation of *IRF5* has not been emphasized in ESCC previously as its frequency is rather low in the primary tumors (<5%). We found this gene is recurrently mutated in metastatic ESCCs and is predicted to be one of the driver genes in metastasis. In breast cancer, reduced expression of IRF5 is reported in the metastatic cancer in LNs from invasive ductal carcinoma and IRF5 inhibits invasion and metastasis *in vitro*[29]. IRF5 expression in tumor cells induces the expression of a number of cytokines and chemokines, including CXCL13 that plays an important role to recruit the tumor infiltration lymphocytes to the tumor site[30]. Use of a modified CXCL13 targeting both the regulator and effector of the immune system to control tumor cell escape can abrogate lung metastasis in the aggressive breast cancer model[31]. In ESCC, according to the microarray data (GSE23400), downregulation of *IRF5* at the mRNA level was also observed (p=0.0015). Previous studies and our data suggest that IRF5 may also have a role in regulating a network of genes important for immune responses to ESCC cells and *IRF5* mutations are likely to abrogate this function to promote cancer invasion and metastasis in ESCC.

*ZNF750* LOF mutations were discovered in the metastatic LN with tumors. An important association between *ZNF750* mutations and LN metastasis was observed by combining the results of ESCC from previous studies. The genetic alterations at *ZNF750* were associated with shorter disease relapse-free survival of patients with esophageal cancer in the TCGA study. ZNF750 plays an essential regulatory role in controlling

epithelial homeostasis by inhibiting progenitor genes, while inducing differentiation[32, 33]. In squamous cell carcinomas, *ZNF750* mutated lesions were exclusively observed; it drives differentiation and regulates cell growth and migration[34]. In ESCC, evidence shows that *ZNF750* knockdown strongly promotes cell proliferation, migration and invasion[6]. Gene set enrichment analysis further revealed that *ZNF750* was associated with suppression of PRC1 protein BMI1 and PRC2 protein SUZ12. Expression of *BMI1* and *SUZ12* was significantly elevated in ESCC (**Supplementary Figure S12**). BMI1, as an epigenetic regulator, maintains the self-renewal and proliferative capacity of the cells[35]. Previous study showed that BMI1 promotes metastasis and chemoresistance in melanoma[36]. In tongue SCC, BMI1 mediates the podocalyxin (PODXL)-enhanced cisplatin chemoresistance[37]. Taken the data together, loss of wild-type *ZNF750* may promote persistence of cancer-initiating cells, increased stemness, drug resistance and cell migration. Therefore, we hypothesize that ZNF750 is a metastasis suppressor in ESCC. Mutations of *ZNF750*, especially the LOF mutations, abrogate inhibition of metastasis.

Epigenetic regulators *KMT2D*, *KAT2A* and *TET2* were mutated in 60% of the metastatic ESCC, suggesting an important role of these histone modifiers for metastasis in ESCC. Recently, Sawada *et al*. reported that increased invasive activity of ESCC was observed after knockdown of *TET2*, which encodes a methylcytosine dioxygenase catalyzing the conversion of methylcytosine to 5-hydrxymethylcytosine[10]. *KMT2D* encodes a histone methyltransferase that methylates the lysine-4 position of histone H3. Mutation of *KMT2D* was frequently reported in the primary ESCC tumors and most of the mutations are inactivating mutations[6-10]. We identified multiple mutations of *KMT2D* in the metastatic cancer in LNs of ESCC patients. Moreover, over 50% of the metastatic ESCC in LNs had 6p22 or 11q23 deletion and the cases with these deletions were associated with advanced cancer stage. Deletion of histone cluster 1 on 6p22.1 is not frequently reported in cancers, except in acute lymphoblastic leukemia in Downs' syndrome, in which about 22% of the cases carry the deletion in this region[38]. Frequent deletion of 11q23 has been reported in several cancers including hematological malignancies and solid tumors[39-41]. Variant histone *H2AFX* on 11q23, also known as *H2AX*, plays essential roles in DNA double-strand break repair and genomic stability, and

thus, is well-accepted as a "histone guardian"[42]. In colon cancer cells, loss of *H2AFX* induces mesenchymal-like characteristics[43]. These studies suggest a role of *H2AFX* in cancer initiation, progression and metastasis. Our results suggest that genetic lesions of chromatin regulators and histone variants may function collectively leading to chromatin disorganization and abnormal DNA packaging, which may promote metastasis in ESCC.

Telomeres are composed of the repetitive $(TTAGGG)^n$ repeats bound by shelterins, which protect them from being recognized as DNA double-strand breaks. The maintenance of telomere length is crucial for tumor cell survival. One of the mechanisms for maintenance of telomere length is through the reactivation of telomerase. Our WES data analysis in primary ESCC reinforces this idea by demonstration of frequent amplification of *TERT* at 5p15.33 (66.7%) in primary ESCC, as well as seen in a previous study[11], regardless of early or late stage primary tumors. Interestingly, *TERT* amplification at 5p15.33 was detected with a reduced frequency (20%) in the metastatic ESCC. Our data suggest that *TERT* amplification at 5p15.33 may be one of the major mechanisms operating in primary ESCC for TERT reactivation for maintenance of telomere length, while *TERT* amplifications are negatively selected in metastatic ESCC. However, we cannot completely rule out that *TERT* promoter mutations or structural rearrangements exist in the metastatic ESCC due to limitations of WES analysis. Further studies are needed to better answer the role of telomere biology in ESCC metastasis.

In summary, this study characterized the genomic landscape of the primary tumors and metastatic cancers in patients with ESCC in Hong Kong. Our study highlights the role of *TP53* GOF mutations, genetic lesions in differentiation regulators *ZNF750* and *IRF5*, epigenetic regulators and histone variants in metastatic ESCC, which are expected to aid in identifying the therapeutic and actionable targets and development of precision medicine for this deadly cancer.

**Acknowledgements:**

The study is funded by seed funding from the University of Hong Kong to MLL.

## References


1. Enzinger PC, Mayer RJ. Esophageal cancer. N Engl J Med 2003;349:2241-52.
2. Rice TW, Rusch VW, Apperson-Hansen C, et al. Worldwide EC collaboration. Dis Esophagus 2009;22:1-8.
3. Zhang Y. Epidemiology of esophageal cancer. World J Gastroenterol 2013;19:5598-606.
4. Lam KY, Ma L. Pathology of esophageal cancers: local experience and current insights. Chin Med J (Engl) 1997;110:459-64.
5. Hong Kong Cancer Registry, Hospital Authority. 2014;http://www3.ha.org.hk/cancereg/.
6. Zhang L, Zhou Y, Cheng C, et al. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. Am J Hum Genet 2015;96:597-611.
7. Song Y, Li L, Ou Y, et al. Identification of genomic alterations in oesophageal squamous cell cancer. Nature 2014;509:91-5.
8. Lin DC, Hao JJ, Nagata Y, et al. Genomic and molecular characterization of esophageal squamous cell carcinoma. Nat Genet 2014;46:467-73.
9. Gao YB, Chen ZL, Li JG, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet 2014;46:1097-102.
10. Sawada G, Niida A, Uchi R, et al. Genomic Landscape of Esophageal Squamous Cell Carcinoma in a Japanese Population. Gastroenterology 2016;150:1171-82.
11. Cancer Genome Atlas Research N, Analysis Working Group: Asan U, Agency BCC, et al. Integrated genomic characterization of oesophageal carcinoma. Nature 2017;541:169-175.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
13. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.
14. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491-8.
15. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;11:11 10 1-11 10 33.
16. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.
17. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 2012;22:568-76.
18. Amarasinghe KC, Li J, Hunter SM, et al. Inferring copy number and genotype in tumour exome data. BMC Genomics 2014;15:732.
19. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012;7:562-78.
20. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature 2013;500:415-21.
21. Cao Y, Bryan TM, Reddel RR. Increased copy number of the TERT and TERC telomerase subunit genes in cancer cells. Cancer Sci 2008;99:1092-9.
22. Swanton C, McGranahan N, Starrett GJ, et al. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov 2015;5:704-12.

23. Ding Q, Chang CJ, Xie X, et al. APOBEC3G promotes liver metastasis in an orthotopic mouse model of colorectal cancer and predicts human hepatic metastasis. J Clin Invest 2011;121:4526-36.
24. Lam KY, Tsao SW, Zhang D, et al. Prevalence and predictive value of p53 mutation in patients with oesophageal squamous cell carcinomas: a prospective clinico-pathological study and survival analysis of 70 patients. Int J Cancer 1997;74:212-9.
25. Li J, Yang L, Gaur S, et al. Mutants TP53 p.R273H and p.R273C but not p.R273G enhance cancer cell malignancy. Hum Mutat 2014;35:575-84.
26. Kang HJ, Chun SM, Kim KR, et al. Clinical relevance of gain-of-function mutations of p53 in high-grade serous ovarian carcinoma. PLoS One 2013;8:e72609.
27. Lien C, Fang CM, Huso D, et al. Critical role of IRF-5 in regulation of B-cell differentiation. Proc Natl Acad Sci U S A 2010;107:4664-8.
28. Garaud S, Willard-Gallo K. IRF5: a rheostat for tumor-infiltrating lymphocyte trafficking in breast cancer? Immunol Cell Biol 2015;93:425-6.
29. Bi X, Hameed M, Mirani N, et al. Loss of interferon regulatory factor 5 (IRF5) expression in human ductal carcinoma correlates with disease stage and contributes to metastasis. Breast Cancer Res 2011;13:R111.
30. Pimenta EM, De S, Weiss R, et al. IRF5 is a novel regulator of CXCL13 expression in breast cancer that regulates CXCR5(+) B- and T-cell trafficking to tumor-conditioned media. Immunol Cell Biol 2015;93:486-99.
31. Bodogai M, Lee Chang C, Wejksza K, et al. Anti-CD20 antibody promotes cancer escape via enrichment of tumor-evoked regulatory B cells expressing low levels of CD20 and CD137L. Cancer Res 2013;73:2127-38.
32. Cohen I, Birnbaum RY, Leibson K, et al. ZNF750 is expressed in differentiated keratinocytes and regulates epidermal late differentiation genes. PLoS One 2012;7:e42628.
33. Boxer LD, Barajas B, Tao S, et al. ZNF750 interacts with KLF4 and RCOR1, KDM1A, and CTBP1/2 chromatin regulators to repress epidermal progenitor genes and induce differentiation genes. Genes Dev 2014;28:2013-26.
34. Hazawa M, Lin DC, Handral H, et al. ZNF750 is a lineage-specific tumour suppressor in squamous cell carcinoma. Oncogene 2016; [Epub ahead of print].
35. Maynard MA, Ferretti R, Hilgendorf KI, et al. Bmi1 is required for tumorigenesis in a mouse model of intestinal cancer. Oncogene 2014;33:3742-7.
36. Ferretti R, Bhutkar A, McNamara MC, et al. BMI1 induces an invasive signature in melanoma that promotes metastasis and chemoresistance. Genes Dev 2016;30:18-33.
37. Zhou Y, Zhang L, Pan H, et al. Bmi1 essentially mediates podocalyxin-enhanced Cisplatin chemoresistance in oral tongue squamous cell carcinoma. PLoS One 2015;10:e0123208.
38. Loudin MG, Wang J, Leung HC, et al. Genomic profiling in Down syndrome acute lymphoblastic leukemia identifies histone gene deletions associated with altered methylation profiles. Leukemia 2011;25:1555-63.
39. Srivastava N, Gochhait S, Gupta P, et al. Copy number alterations of the H2AFX gene in sporadic breast cancer patients. Cancer Genet Cytogenet 2008;180:121-8.
40. Ma SK, Wan TS, Au WY, et al. Chromosome 11q deletion in myeloid malignancies. Leukemia 2002;16:953-5.
41. Lee AS, Seo YC, Chang A, et al. Detailed deletion mapping at chromosome 11q23 in colorectal carcinoma. Br J Cancer 2000;83:750-5.
42. Bonner WM, Redon CE, Dickey JS, et al. GammaH2AX and cancer. Nat Rev Cancer 2008;8:957-67.

43. Weyemi U, Redon CE, Choudhuri R, et al. The histone variant H2A.X is a regulator of the epithelial-mesenchymal transition. Nat Commun 2016;7:10711.

**Figure legends**

**Figure 1: Genomic alterations in Hong Kong ESCC.** A) Mutation rate and number of protein-altered mutations in primary and metastatic tumors. B) Mutation landscape of significantly mutated genes and relevant pathways in primary and metastatic tumors. * indicates the genes significantly mutated in the primary tumors (MutSigCV $p<0.001$) and the genes significantly mutated in the metastatic LNs are highlighted in orange (MutSigCV $p<0.001$). C) Frequency of SNVs and SCNVs in the primary and metastatic tumors. D) Frequency of SNVs and SCNVs in the stage I-II (early) and III-IV (late) primary tumors. **$p<0.01$ in Fisher's exact test.

**Figure 2: Mutational signature in Hong Kong ESCC.** A and B) Ninety-six substitution classifications from WES data derived from primary and metastatic tumors, respectively. Substitution types are displayed in different colors on the horizontal axis. The vertical axis indicates the proportion of each substitution pattern. Trinucleotide contexts of mutations occurring at cytosine nucleotides in primary and metastatic tumors are on the top right corner. Font size of the bases at 5' and 3' positions are proportional to the frequencies. C) The contribution of mutation signatures to the total variant number in each tumors. D) The contribution of mutation signatures in the stage I-II (early), III-IV (late) primary tumors and metastatic LNs (MET).

**Figure 3: TP53 mutations in ESCC**. A) *TP53* mutations in primary tumors and metastatic cancer in LNs from patients with ESCC in Hong Kong. The amino acids with mutation frequencies >5% in primary tumors are highlighted. B) Distribution of *TP53* missense mutations in Hong Kong, high-risk and low-risk regions of mainland China, and Japan, respectively (top), and mixed cohorts of patients with ESCC from mainland China and Japan (bottom). X-axis shows the position of amino acid in TP53. Y-axis shows the density of missense mutations. C) Kaplan-Meier analysis of survival curves for the cases with missense, LOF or no mutations.

**Figure 4: *ZNF750* in ESCC.** A) *ZNF750* mutations in ESCC from current and previous studies (n=630). Asterisk indicates the truncating mutations. B) ZNF750 mutation in ESCC with and without LN metastasis in the current and previous study cohort. ****p<0.0001; ***p<0.001; $^{+}$p<0.1. C) ZNF750 copy number in ESCC with LOF mutation. D) Gene Set Enrichment Analysis (GSEA) in microarray data (GSE23400) between the ESCC cases with high and low *ZNF750* expression (median as cut-off).

**Figure 5: Copy number variations in ESCC primary tumors and metastatic LN, respectively.** Red color indicates copy number loss (CN loss), blue color indicates copy number gain (CN gain). Green color indicates no copy number changes.

**Figure 6: Deletion of histone variants on 6p22 in ESCC.** A) Deletion frequency in ESCC. p value was estimated by Fisher's exact test. B) ESCCs with and without 6p22 deletion. Red color indicates copy number loss. Green color indicates no copy number changes. C) Genes within the deleted region on 6p22. D) Survival curves of the patients with and without deletion on 6p22 (histone cluster 1) or on 11q23 (*H2AFX*). p value was estimated by log-rank test.