RESEARCH ARTICLE

Identification of a Novel Fusion Gene (*HLA-E* and *HLA-B*) by RNA-seq Analysis in Esophageal Squamous Cell Carcinoma

Yu-Zhang Jiang[1]*, Qian-Hui Li[1], Jian-Qiang Zhao[2], Jun-Ji Lv[2]

Abstract

Esophageal squamous cell carcinoma (ESCC) is the most common histologic subtype of esophageal cancer and is characterized by a poor prognosis. Determining gene changes in ESCCs should improve understanding of putative risk factors and provide potential targets for therapy. We sequenced about 55 million pair-end reads from a pair of adjacent normal and ESCC samples to identify the gene expression level and gene fusion. Sanger sequencing was used to verify the result. About 17 thousand genes were expressed in the tissues, of which approximately 2400 demonstrated significant differences between tumor and adjacent non tumor tissue. GO and KEGG pathway analysis revealed that many of these genes were associated with cellular adherence and movement, simulation responses and immune responses. Notably we identified and validated one fusion gene, *HLA-E* and *HLA-B*, located 1 MB apart. We also identified thousands of remarkably expressed transcripts. In conclusion, a novel fusion gene *HLA-E* and *HLA-B* was identified in ESCC via whole transcriptome sequencing, which would be a biomarker for ESCC diagnosis and target for therapy, shedding new light for better understanding of ESCC tumorigenesis.

Keywords: RNA-Seq - ESCC - gene fusion - *HLA-E* - *HLA-B*

Asian Pac J Cancer Prev, **15 (5)**, 2309-2312

Introduction

Esophageal cancer is malignancy of the esophagus, which is the most common cancer and ranks the sixth major cause of cancer-related deaths worldwide. There are various subtypes. Esophageal squamous cell carcinoma (ESCC)account of approximately 90% of all cases esophageal cancer worldwide (Jemal et al., 2011). While the middle of China is in the "esophageal cancer belt", which has the highest risk incidence, and the fourth leading cause of cancer death (Lin et al., 2013). It was reported that genetics alteration at both DNA and RNA level contribution to initiation and development of ESCC (Gibb et al., 2011). As many other cancers, genomics instability is believed to be the main cause for genetic diversity, which foster multiple hallmarks functions, such as tumor invasion, metastasis and recurrent (Hanahan and Weinberg, 2011). To date, the exact genetics and molecular mechanism of ESCC have yet not classified, It also not practical to comprehensively illuminate whole map of genetic alteration in ESCC using the traditional methods such as Sanger sequencing.

With the remarkable advances in high-throughput sequencing over the last decade make possible to map the whole genetic variation in genome-wide scale. Transcriptome sequencing (RNA-Seq) has become a revolutionary tool for comprehensive study of the whole transcripts and with the merits to identify the gene structure variation, such as gene fusion, splicing variants (Wang et al., 2009). Recently limited study has performed in ESCC (Bandla et al., 2012; Ma et al., 2012; Zhang et al., 2013), which has elucidated some basis of the tumorigenesis and development in ESCC.

A major goal for cancer research was to find the tumor specific causal genetic aberrations. Gene fusions resulting from chromosomal rearrangements in cancer are believed to define the most prevalent category of 'cancer genes' (Futreal et al., 2004). While gene fusion has widely described in rare hematological malignancies, and recently some recurrent fusion gene were discovered in solid cancer by RNA-Seq (Maher et al., 2009; Ju et al., 2012; Ren et al., 2012), also the typically molecular BCR-ABL1 which was successfully re-discovered by RNA-Seq in chronic myeloid leukemia (Maher et al., 2009).

Here we reported a whole map of RNAs expressed profile in a pair of ESCC and adjacent normal tissue, meanwhile a novel fusion *HLA-E* and *HLA-B* was identified using RNA-Seq, and which is verified by Sanger sequencing. We observed a number of differentially expressed genes in ESCC. Integrative analysis with GO and KEGG, which revealed cellular adherent and movement as well as simulation response may be

*[1]Department of Medical Laboratory, [2]Department of Cardiothoracic Surgery, Huai'an First People's Hospital, Nanjing Medical University, Huai'an, China *For correspondence: jyz8848@163.com*

Table 1. Summary Statistics of RNA-Seq Data in This Study

	Normal tissue	Cancer tissue
Total Reads number	58, 180, 764	55, 074, 384
Reads map to reference genome	51, 505, 394	41, 865, 911
Mapped reads ratio (%)	88.53	76.02
Total base number	5, 818, 076, 400	5, 507, 438, 400
Bases map to reference genome	5, 150, 539, 400	4, 186, 591, 100
Average coverage	44.99	36.57

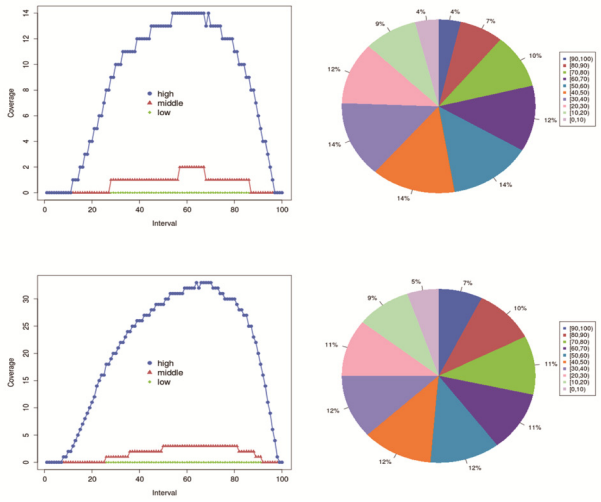

Figure 1. The Evenness (Left) and Integrity (Right) of Reads Distribution in cDNA Reference. The left figure is shown the reads distribution in multiple part in gene. In which X-axis is shown the gene body are divided in 100 parts from 5 to 3, Y-axis is the number of reads locate in certain part. High expression is marked in blue curve, red is medium expression, and green is low expression. The right picture is shown read coverage rate in genes. Coverage rate is count by reads number in exon divided by reads in all exons. Picture A and B is shown reads distribution in normal and cancer. C and D are shown the gene coverage in normal and cancer

associated with tumor development. Taken together, *HLA-E* and *HLA-B* fusion gene was discovered and may represent a novel class of molecular alteration in ESCC that could have important implications in understanding the tumorigenesis and early detection, may be prediction of prognosis in ESCC.

Materials and Methods

Samples

A pair of fresh frozen ESCC and its adjacent normal tissue specimen was a 63 male patient, who was diagnosed with a TNM grade of T2N0M0 (moderate differentiation and no lymph node metastasis). Samples used in this study were approved by the committee for ethical review of research involving human subjects.

For the samples subjected to RNA-seq, total RNA was isolated and its quality was assed using Agilent Bioanalyzer. CDNA library were prepared following the standard mRNA protocol by Illumina and sequenced (115-bp pair-end read length) using Genome Analyzer.

Bioinformatics analysis

Clean and high-quality sequencing reads were aligned against both genome hg19 and transcripts reference using Bowtie 2.0.0 (Langmead et al., 2009). Ribosomal RNA sequences were removed by aligning to 28S, 18S mitochondrial ribosome. TopHat (Trapnell et al., 2009) is a fast splice junction mapping software that uses Bowtie alignment to align RNA-Seq reads. Cufflinks (Trapnell et al., 2010) was used to count mRNA expressed level which was measured by RPKM by normalizing the number of exon reads to the length of exons within that gen and per million mapped reads. GO and KEGG pathway analysis were performed on the web tool DAVID based on the differential expressed gene level (Huang da et al., 2009). Only the FDR<5%are selected to show the gene function enrichment results. TopHat Fusion (Kim and Salzberg, 2011) were used to identify gene fusion.

Evenness distribution of reads on reference genes

We normalized locus positions in those genes that length above 300nt and expression level FPKM above 1 by dividing the gene bodies with 100 windows. High-, medium- and low-expressed group were divided by the expression level. Then reads mapped to every window are counted and then used to calculate read distribution along genes.

Gene fusion validation

To detect fusion transcripts, we design the forward primer targeting the 5' partner gene and reverse primer targeting the 3' partner. Primer pairs (Table 3) for the coding exons of the fusion genes were generated using Primer 5 software (PREMIER Biosoft International, Palo Alto, Calif.), and the PCR volume used comprised 10 μl sample, 1 μl 10× PCR buffer, 1 μl cDNA template, 0.2 μl dNTP, 0.2 μl Taq Enzyme (Dingguo), and 2 pmol/μl each oligonucleotide. PCR was performed using the following procedure: 94°C for 1 min, 40 cycles of 94°C for 20 sec, 55°C for 20 sec and 72°C for 15 sec, followed by 72°C for 1 min. The PCR products of the fusion genes were cloned in the pGEM®-T Easy Vector (Promega) and then sequenced with the T7 primer using a 3730 DNA Analyzer (ABI).

Results

RNA-Seq and sequencing reads alignment

We sequenced a paired normal and tumor tissues cDNA from a 63 male who diagnosed as ESCC. About ~55 million short reads were passed the QC.HG19 genome assembly and transcripts from the UCSC were used as reference sequence. Approximately 51.5M (88.53%) and 41.86M (76.02%) reads were aligned to the reference, which reached about 36X and 45Xcoverage in cancer and adjacent normal tissue, respectively. Besides, reads map to 28S, 18S rRNA and mitochondrial transcripts were removed to eliminate the non-nuclear transcripts effects. The detailed sequencing statistics were showed in Table1. Gene expression differences between ESCC and normal tissue

Expression levels were tabulated in accordance with the number of fragments per gene per kilobase exon per million mapped reads (RPKM). Using Cufflinks software,

Table 2. Summary of Candidate Fusion Genes Detected by RNA-Seq

Sample	Gene1	Gene2	Breakpoint of gene1	Breakpoint of gene2
Cancer	ENSG00000242058	CPT1A	chr18:9020357	chr11:68560779
Cancer	*HLA-E*	*HLA-B*	chr6:30458291	chr6:31323368
Normal	TMSB10	AGAP1	chr2:85133515	chr2:236791034

Table 3. The Primer Sequences Used in qPCR and Gene Fusion Validation

Gene	Forward (5'-3')	Reverse (5'-3')
GAPDH*	CATGAGAAGTATGACAACAGCCT	AGTCCTTCCACGATACCAAAGT
$*HLA-E*	GACGGCAAGGATTATCTCAC	
$*HLA-B*		CAGTGTCCTGAGTTTGGTCC

*the control; $primer used to amplify gene fusion

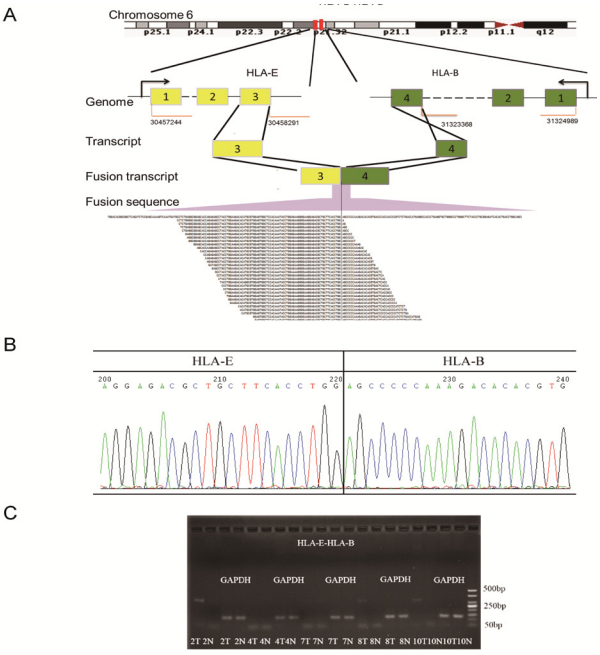

Figure 2. The Breakpoint of *HLA-E-HLA-B* Fusion Gene. 2A) *HLA-E* and *HLA-B* are located in chromosome 6 with the distance of 1 million bases. The breakpoint *HLA-E-HLA-B* is supported by both paired reads and single reads. 2B) Sanger sequencing of cloned fusion gene 2C) Validation of gene fusion in other ESCC patient by PCR, and sample 4, 7, 8, 10 are validated samples

we identify about 17 thousands expressed transcripts in paired samples, in which about 13 thousands (account 62% of known gene) genes FPKM value were above 1, in which about 1/3 (6490/17455 in cancer, 6535/17028 in normal) FPKM above 10.The FPKM value in cancer and control samples ranged from 0 to 657542 and 236728, and reached average value 171.68 and 109.The evenness and integrity of read distribution were evaluated (Figure 1). Besides, the correlation coefficient of the gene expression levels between the paired samples was 0.7298.

Differential gene function annotation

Differential gene was analysis by Cufflink with q value less than 0.01. 2419 differential expressed genes were identified, while 1190 of them were up regulated in tumor and 1229 were down regulated. 82 Go functions were enriched in those differential genes (P<0.001), which exhibited those differential genes were enriched in cellular adherent and movement, simulation response, immune response, multi-cellular metabolism and development, as well as enrichment in adherent point and extracellular receptor pathway.

Gene fusion analysis

Three fusion genes (Table 2) were identified by TopHat-fusion under such criteria: 1) above 8 span reads, 2) opposed reads by support reads less than 0.5 and 3) rRNA and srRNA removed.

We re-analyzed the fusion gene using TopHat which used the unmapped or mapped to other gene reads to realignment to those fusion genes. Finally, *HLA-E-HLA-B* was the most reliable fusion with 80 paired reads and 1, 309 single reads supports. *HLA-E* and *HLA-B*are located in chromosome 6 with the distance of 1 million base. The fusion point was in the border of the third exon of *HLA-E* and the fourth exon border of *HLA-B*, shown in Figure 2A. Fusion gene was amplified in plasmid and then performed Sanger sequencing (Figure 2B). Moreover, we amplified the gene fusion sequence by PCR in other patients, and recurrent gene fusion was found in tumor 8T and 10T (Figure 2C).

Discussion

It is well accepted that cumulative genetic mutation play great role in tumorigenesis and tumor development. Based on this theory we carried on the research on whole map genetic aberrant of ESCC by RNA-Seq. In our study we sequenced a paired cancer and adjacent tissue from a 63 male ESCC patient from middle of China. With the average coverage of ~36X and ~44X, we constructed the comprehensive profile of transcripts on ESCC. Moreover 1190 genes were up regulated and 1229 down regulated in cancer. GO function and KEGG pathway based on differential genes revealed that those gene were associated with cellular adherent and movement, simulation response, immune response, multi-cellular metabolism and development, as well as enrichment in adherent point and extracellular receptor pathway.

Previous study genetic aberrant in ESCC by next generation sequencing, discovered PTK6 was a suppressor in ESCC (Ma et al., 2012), however in our study FPKM of PTK6 were only 11.4 and 17.0 in cancer and control sample with no significant difference, it was hard to conclude PTK6 had similar function in our samples. KRT13, KRT4, SPRR3, SPRR2A, TGM3 were remarkable down

regulated in our result that were consistent with previous study (Luo et al., 2004), indicate RNA-Seq was a robust approach to study ESCC gene expression.

Moreover, we also found that both *HLA-E* and *HLA-B* expression level were markedly increase in ESCC, nevertheless *HLA-E* was down regulated in breast cancer (de Kruijf et al., 2010) and glioblastomas (Kren et al., 2011) as a prognostic marker, also weekly expressed in multiple cell line (Marin et al., 2003), exception the higher expression predicted better survival in cervical adenocarcinomas (Spaans et al., 2012) while *HLA-B* was down-regulated in ovarian cancers (Le et al., 2002). Therefore, *HLA-E* and *HLA-B* may have specific function pattern in ESCC.

Most importantly, we explored and further analysis confirmed a novel fusion gene *HLA-E* and *HLA-B*. In the target sequencing clone and Sanger sequencing also confirmed the fusion gene recurrent among patients, and was exclusively enriched in tumor tissues at stable expression levels. A fusion gene is a hybrid gene formed from two previously separate genes, which may occurred as a result of: translocation, interstitial deletion, or chromosomal inversion. Fusion genes may lead to a new gene product acquired new or different function from both fusions molecular. A series fusion genes have been characterized, such as BCR-ABL (Maher et al., 2009), KIF5B-RET (Ju et al., 2012). A recently reported fusion gene GOLM1-MAK10 in ESCC was not recurrent in this study, perhaps because of its low frequency in Chinese people. *HLA-E-HLA-B* was recurrent (3/5) in the ESCC sample from China, which could serve as biomarker even drug target in the future. *HLA-E-HLA-B* would benefit for the better understanding of the ESCC tumorigenesis and development, and might be a biomarker for ESCC diagnosis and therapy target.

Acknowledgements

The author (s) declare that they have no competing interests.

References

Bandla S, Pennathur A, Luketich JD, et al (2012). Comparative genomics of esophageal adenocarcinoma and squamous cell carcinoma. *Ann Thorac Surg*, **93**, 1101-6.

de Kruijf EM, Sajet A, van Nes JG, et al (2010). *HLA-E* and HLA-G expression in classical HLA class I-negative tumors is of prognostic value for clinical outcome of early breast cancer patients. *J Immunol*, **185**, 7452-9.

Futreal PA, Coin L, Marshall M, et al (2004). A census of human cancer genes. *Nat Rev Cancer*, **4**, 177-83.

Gibb EA, Enfield KS, Tsui IF, et al (2011). Deciphering squamous cell carcinoma using multidimensional genomic approaches. *J Skin Cancer*, **2011**, 541405.

Hanahan D, Weinberg RA (2011). Hallmarks of cancer: the next generation. *Cell*, **144**, 646-74.

Huang da W, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.

Jemal A, Bray F, Center MM, et al (2011). Global cancer statistics. *CA Cancer J Clin*, **61**, 69-90.

Ju YS, Lee WC, Shin JY, et al (2012). A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Res*, **22**, 436-45.

Kim D, Salzberg SL (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*, **12**, R72.

Kren L, Slaby O, Muckova K, et al (2011). Expression of immune-modulatory molecules HLA-G and *HLA-E* by tumor cells in glioblastomas: an unexpected prognostic significance? *Neuropathology*, **31**, 129-34.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.

Le YS, Kim TE, Kim BK, et al (2002). Alterations of HLA class I and class II antigen expressions in borderline, invasive and metastatic ovarian cancers. *Exp Mol Med*, **34**, 18-26.

Lin Y, Totsuka Y, He Y, et al (2013). Epidemiology of esophageal cancer in Japan and China. *J Epidemiol*, **23**, 233-42.

Luo A, Kong J, Hu G, et al (2004). Discovery of Ca2+-relevant and differentiation-associated genes downregulated in esophageal squamous cell carcinoma using cDNA microarray. *Oncogene*, **23**, 1291-9.

Ma S, Bao JY, Kwan PS, et al (2012). Identification of PTK6, via RNA sequencing analysis, as a suppressor of esophageal squamous cell carcinoma. *Gastroenterology*, **143**, 675-86 e1-12.

Maher CA, Kumar-Sinha C, Cao X, et al (2009). Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97-101.

Marin R, Ruiz-Cabello F, Pedrinaci S, et al (2003). Analysis of *HLA-E* expression in human tumors. *Immunogenetics*, **54**, 767-75.

Ren S, Peng Z, Mao JH, et al (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res*, **22**, 806-21.

Spaans VM, Peters AA, Fleuren GJ, Jordanova ES (2012). *HLA-E* expression in cervical adenocarcinomas: association with improved long-term survival. *J Transl Med*, **10**, 184.

Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-11.

Trapnell C, Williams BA, Pertea G, et al (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-5.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.

Zhang Q, Zhang J, Jin H, Sheng S (2013). Whole transcriptome sequencing identifies tumor-specific mutations in human oral squamous cell carcinoma. *BMC Med Genomics*, **6**, 28.