Translational Oncology
www.transonc.com

Volume 11 Number 6 December 2018 pp. 1283–1291

Mutations, Differential Gene Expression, and Chimeric Transcripts in Esophageal Squamous Cell Carcinoma Show High Heterogeneity[1]

Paulo Thiago de Souza-Santos[*], Sheila Coelho Soares Lima[*], Pedro Nicolau-Neto[*], Mariana Boroni[†], Nathalia Meireles Da Costa[*], Lilian Brewer[‡], Albert Nobre Menezes[¶], Carolina Furtado[§], Miguel Angelo Martins Moreira[§], Hector N. Seuanez[§], Tatiana de Almeida Simão[‡] and Luis Felipe Ribeiro Pinto[*,‡]

[*]Molecular Carcinogenesis Program, Instituto Nacional de Câncer–INCA, Rua Andre Cavalcanti, 37–6° andar, Centro, Rio de Janeiro, RJ, Brasil, 20231-050; [†]Bioinformatics and Computational Biology Laboratory, Instituto Nacional de Câncer–INCA, Rua Andre Cavalcanti, 37–1° andar, Centro, Rio de Janeiro, RJ, Brasil, 20231-050; [‡]Biochemistry Department, Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, Boulevard 28 de Setembro, 77-Maracanã, Rio de Janeiro, RJ, Brasil, 20551-030; [§]Genetics Program, Instituto Nacional de Câncer–INCA, Rua Andre Cavalcanti, 37–4° andar, Centro, Rio de Janeiro, RJ, Brasil, 20231-050; [¶]College of Medical and Dental Sciences, University of Birmingham, Vicent Drive, Edgbaston, Birmingham, B15 2TT, UK

Abstract

Esophageal squamous cell carcinoma (ESCC) is a frequent and lethal neoplasia. As recent advances in targeted therapy have not improved ESCC prognosis, characterization of molecular alterations associated to this tumor is of foremost relevance. In this study, we analyze, for the first time, the complete genomic profile of ESCC by RNA-seq. *TP53* was the most frequently mutated gene in the investigation and validation sets (78.6% and 67.4%, respectively). Differential expression analysis between tumor and nontumor adjacent mucosa showed 6698 differentially expressed genes, most of which were overexpressed (74%). Enrichment analysis identified overrepresentation of Wnt pathway, with overexpressed activators and underexpressed inactivators, suggesting activation of canonical and noncanonical Wnt signaling pathways. Higher *WNT7B* expression was associated with poor prognosis. Twenty-one gene fusions were identified in 50% of tumors, none of which involving the same genes in different patients; 71% of fusions involved syntenic genes. Comparisons with TCGA data showed co-amplification of seven gene pairs involved in fusions in the present study (~33%), suggesting that these rearrangements might have been driven by chromoanagenesis. In conclusion, genomic alterations in ESCC are highly heterogeneous, impacting negatively in target therapy development.

Translational Oncology (2018) 11, 1283–1291

Introduction

Esophageal carcinoma (EC) is a highly frequent and lethal tumor, representing the eighth most common and the sixth most fatal cancer worldwide [1]. Esophageal squamous cell carcinoma (ESCC) is the main histopathology subtype, accounting to approximately 80% of all EC cases, mainly in developing countries, such as Brazil [2]. ESCC patients show a poor prognosis, with a 5-year survival rate

Address all correspondence to: Luis Felipe Ribeiro Pinto, Rua Andre Cavalcanti, 37–6° andar, Centro, Rio de Janeiro, RJ, Brasil, 20231-050. E-mail: lfrpinto@inca.gov.br

[1] Declaration of Interest: none.

Received 2 August 2018; Revised 3 August 2018; Accepted 3 August 2018

1936-5233/18
https://doi.org/10.1016/j.tranon.2018.08.002

below 20% [3]. Regardless of the noteworthy advances in cancer diagnosis and therapy, current ESCC treatment approaches are frequently unsuccessful, and the outcome of ESCC patients remains unfavorable [4–6]. This makes the understanding of the molecular mechanisms involved in esophageal carcinogenesis a most relevant precondition for developing more efficient therapies.

A small number of genome-wide ESCC studies have been reported, mainly of patients from eastern countries. Recently, The Cancer Genome Atlas (TCGA) consortium published a genome-wide EC study that included ESCC and esophageal adenocarcinoma [7]. This report focused on copy number and mutation analyses revealing differences between patients from several geographic regions, albeit with limited data on gene expression profiles, probably due to lack of paired samples from tumor and normal, adjacent esophageal mucosa.

In this study, we report the ESCC transcriptome in a sample of patients from a Western population and analyze, for the first time, differentially expressed genes, mutations, and gene fusions, pointing to dysregulated signaling pathways potentially associated to ESCC carcinogenesis.

Materials and Methods

Patients and Sample Collection

A set of 55 paired biopsies of ESCC and nontumor adjacent mucosa was collected from 2006 to 2015 at the Endoscopy Service of the Instituto Nacional de Câncer (Rio de Janeiro, Brazil). Histopathology profiles were provided by the Pathology Department. Written informed consents were signed for using biological samples as well as clinical and pathology data from patient records. Samples, obtained before treatment, were separated in an investigation (14 paired samples) and a validation set (41 paired samples). This study was approved by the institution Ethics Committee and was conducted according to the Declaration of Helsinki.

RNA Isolation, Library Preparation, and Sequencing

Total RNA was isolated with the RNeasy Kit (Qiagen) following the manufacturer's instructions. The integrity of RNA was assessed with RNA 6000 Nano chip with a Bioanalyzer platform (Agilent), and only samples with RNA integrity number (RIN) ≥8 were included. RNA samples were used for constructing cDNA libraries with TruSeq RNA (Illumina) following the manufacturer's protocol. Libraries were sequenced in an Illumina HiSeq 2500 platform to produce 2×100 bp reads.

Data Processing and Read Mapping

Statistics and quality analyses of reads were generated with FastQC 0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) by Babraham Bioinformatics. RNA-Seq reads were trimmed to remove bad quality bases and reads with Prinseq [8]. Reads were subsequently aligned to a genomic reference sequence of *H. sapiens* (GRch37 version) downloaded from Ensembl database using TopHat2 mapper 2.0.9 [9] with default parameters. Mapped reads were filtered by quality and unique mapping features with Samtools 0.1.19 [10], with -bhq 20 -F 0×100 parameters.

Analysis of Differential Gene Expression

Raw counts for each gene were estimated for all datasets using the HTSeq Python package 0.5.3p3 [11] and the "–stranded=yes" option, as well as the "–mode=intersection-strict" option for exon-intersection counting. As cutoff, only genes with ≥10 read counts were considered for further analyses. Differential expression analysis was performed with DESeq2 package 1.2.5 [12]. Genes were considered to be differentially expressed with adjusted P value < .001 and |log fold-change| ≥ 1. Principal component analysis (PCA) was estimated, and heatmaps were constructed with regularized-logarithmic transformation (rlog) provided by the DESeq2 package.

Functional Annotation

Enrichment analysis was carried out with R package KEGG.db [13]. Following application of Fisher exact test, only pathways with P < 0.1 adjusted by Benjamin-Hochberg's procedure were considered to be enriched.

qPCR Validation

Four genes related to the WNT-signaling pathway (Supplementary Table 1) were selected for validation by quantitative PCR (qPCR) as described previously [14].

Variant Calling

RNA-seq reads were initially mapped to the reference genome and to all known splice junctions using STAR [15]. Uniquely mapped reads were subsequently used for calling the initial set of candidate variants with Genome Analyses Toolkit (GATK) following the best practices recommended by the Broad Institute. Subsequently, these candidates were subjected to several filtering procedures, including removal of variations present in the nontumor adjacent mucosa; InDel alterations and A:T>G:C conversions were excluded in view of their high probability of representing RNA editing products [16]. We used ANNOVAR [17] for annotating variants based on gene models from GENCODE, RefSeq, Ensembl, and the UCSC Genome Browser. All RNA-seq variants present in 1000 genomes were defined as SNPs and excluded from the analyses.

Validation of TP53 Mutations

Genomic DNA was isolated using the DNeasy Blood and Tissue kit (Qiagen) as recommended by the manufacturer. *TP53* was amplified using three primer pairs (Supplementary Table 1), 50 ng of DNA and Taq Platinum (Invitrogen), and amplified products purified with Purelink genomic DNA purification kit (Invitrogen). Subsequently, DNA libraries were constructed with 1 ng of PCR product and Nextera XT DNA sample preparation (Illumina) following the manufacturer's instructions and sequenced in a HiSeq 2500 platform. Reads were submitted to the above described quality control for cDNA reads. Good quality reads were mapped to the human genome with BWA aligner [18].

Fusion Analysis

The FusionMap package [19] was used with the following parameters: Sample.SeedCount > 3, SplicePatternClass = "CanonicalPattern[Major]" or "CanonicalPattern[Minor]", Filter = empity, FrameShiftClass = "InFrame" and OnExonBoundary = "Both".

Comparisons with TCGA data

TCGA data were used to compare our findings with other populations and datasets. The cBioportal for cancer genomics [20,21] was used for selecting and analyzing ESCC from the TCGA provisional dataset.

Further Statistical Analyses

Differential expression of selected genes between ESCC and paired nontumor mucosa was evaluated with paired *t* test or Wilcoxon test.

Table 1. Sample Data and Clinical and Pathology Profiles of 55 ESCC Patients Included in This Study

	ESCC Patients
Gender	
Male	41 (74.5%)
Female	14 (25.4%)
Age (median and range, years)	59 (39-79)
Follow-up (median and range, months)	8 (1.2-41.7)
Tumor central localization	
Upper third	5 (9.1%)
Middle third	43 (78.2%)
Lower Third	7 (12.7%)
Differentiation	
Well	1 (1.8%)
Moderately	39 (70.9%)
Poorly	15 (27.3%)
Tumor stage	
I	2 (3.6%)
II	5 (9%)
III	9 (16.4%)
IV	18 (32.7%)
NA	21 (38.2%)

NA, not available.

Associations between gene expression and the *TP53* mutation status with clinical and pathology data were analyzed with *t* test, Mann-Whitney test, and chi-square. Survival analyses were carried out using "survival" package for R, and hazard ratios (HRs) were adjusted by Cox regression. All analyses were performed in the R environment.

Results

Clinical and Pathology Data

The 55 patients included in this study showed a median age of 59 years (39-79) (Table 1). Most patients were male (74.5%) and showed a median overall survival of 8 months (1.2-42.7). Tumors were moderately differentiated (70.9%), most frequently located in the middle third of the esophagus (78.2%), and diagnosed at late stages (80%).

High-Throughput Evaluation of Gene Expression in ESCC

Global gene expression of tumor and nontumor adjacent mucosa was analyzed in the investigation set (IS). RNA-seq provided an average of 37 million sequences for nontumor adjacent samples, 33 million (89%) of which of good quality. Comparatively, an average of 40 million sequences was generated from tumor samples, with 38.5 million (96%) of good quality. Sequences of good quality, aligned to the reference human genome, were correctly mapped, accounting for 80% and 87% of sequences from nontumor and tumor samples, respectively. Reads were mapped to the 63,677 genes of the Ensembl database, and 44,402 (70%) of them were considered for further analyses following mapping of at least 10 reads. Principal component analysis revealed a homogenous distribution among nontumor adjacent esophageal mucosa but a spread distribution among ESCC samples (Supplementary Figure 1). Comparisons between total paired samples revealed 6698 differentially expressed genes (DEG), 4966 of which (74%) were overexpressed and 1732 (26%) underexpressed in tumors. The differential expression of the 10 most commonly amplified or deleted genes in the TCGA study was analyzed. Four amplified genes (*FGF3*, *FGF19*, *TP63*, and *TFRC*) were found to be overexpressed in ESCC with respect to the nontumor adjacent mucosa, while *CDKN2B*, found to be deleted in TCGA, was underexpressed in our dataset (Supplementary Table 2).

Based on our DEG dataset, pathway enrichment analysis using the KEGG database retrieved a list of 71 overrepresented pathways (adjusted *P* value < .1), among which the "Wnt signaling pathway" showed a total of 54 DEG of the 150 genes included in this pathway. Their expression patterns in tumors and nontumor adjacent mucosa are shown in Figure 1*A*. Most genes (87%) were found to be overexpressed in tumors, with a median log fold-change of 1.75 (1.08-7.55). Most underexpressed genes in tumors corresponded to pathway inhibitors, like *CTNNBIP1* and *DKK1*, indicating reinforcement of Wnt pathway activation. Additionally, we found that, in tumors, activation was not restricted to the canonical pathway represented by the *WNT7B* ligand and its downstream *CCND2* target leading to cell cycle progression, but also involved the noncanonical pathway activated by *WNT16* leading to *MAPK10* activation and apoptosis (Figure 1*B*). Validation of these findings confirmed overexpression of *WNT7B* (*P* < .001), *CCND2* (*P* = .041), *WNT16* (*P* = .037), and *MAPK10* (*P* < .001) in tumors with respect to the nontumor adjacent mucosa (Figure 1*C*). The Wnt signaling pathway was also found to be deregulated in the TCGA dataset, although this finding was not analyzed in the TCGA report [7]. Reanalysis of TCGA data revealed that, among the Wnt-related DEG herein analyzed, *NKD2* and *LRP5* were mutated or amplified (23% and 21%, respectively), while *SFRP5*, *WNT7B*, *CDC25C*, *RHOC*, *PAX2*, and *SOST* did not show alterations (Supplementary Figure 2).

The expression of genes selected for validation was subsequently analyzed in association to age and clinical and pathology variables (Table 2). Low *MAPK10* expression was associated with local/distant metastases (*P* = .046), while associations with age, tumor stage, or lymph node metastases were not observed for any other gene. Interestingly, patients whose tumors showed higher *WNT7B* expression showed a decreased overall survival (median = 7.73 months) when compared with patients with lower *WNT7B* expression in tumors (median = 17.17 months) (Figure 1*D*). Multivariate analysis further revealed that the impact of *WNT7B* expression on the prognosis of ESCC patients was independent of age and tumor stage (HR = 5.5 (1.7-18.0), *P* = .005).

Mutational ESCC Landscape

RNA-seq data from IS revealed a median of 65 point mutations per tumor (33-199) following removal of AT>GC conversions (Figure 2*A*). GC>AT was the most common conversion (57.2%), followed by GC>CG (17.7%), GC>TA (12.9%), AT>TA (7.5%), and AT>CG (4.7%) (Figure 2*B*). Among GC>AT, 58.6% occurred in a CpG context. *TP53* was the most commonly mutated gene, in 78.6% of tumors (Figure 2*C*), while mutations were not detected in the nontumor adjacent mucosa. Other frequently mutated genes were *LOC389831* (42.9%), *PI4KA* (21.4%), *MST1L* (21.4%), *HERC2* (21.4%), and *NBPF9* (21.4%). Comparisons between our findings and the TCGA dataset revealed that, except for *TP53*, mutations occurred in different genes. In our tumor samples, expression analyses of the 10 most frequently TCGA mutated genes (Supplementary Table 2) showed that 4 of them (*CSMD3*, *MUC16*, *DNAH5* and *PKHD1L1*) were overexpressed, while *NFE2L2* was underexpressed (adjusted *P* value < .05).

TP53 was selected for further analysis by DNA sequencing in the validation set (VS) where its mutation frequency was slightly lower (67.4%) than in IS, with single mutations representing the most frequent class in both cases (Figure 3*A*). *TP53* analysis identified four tumors with two mutations, and three and six mutations in single

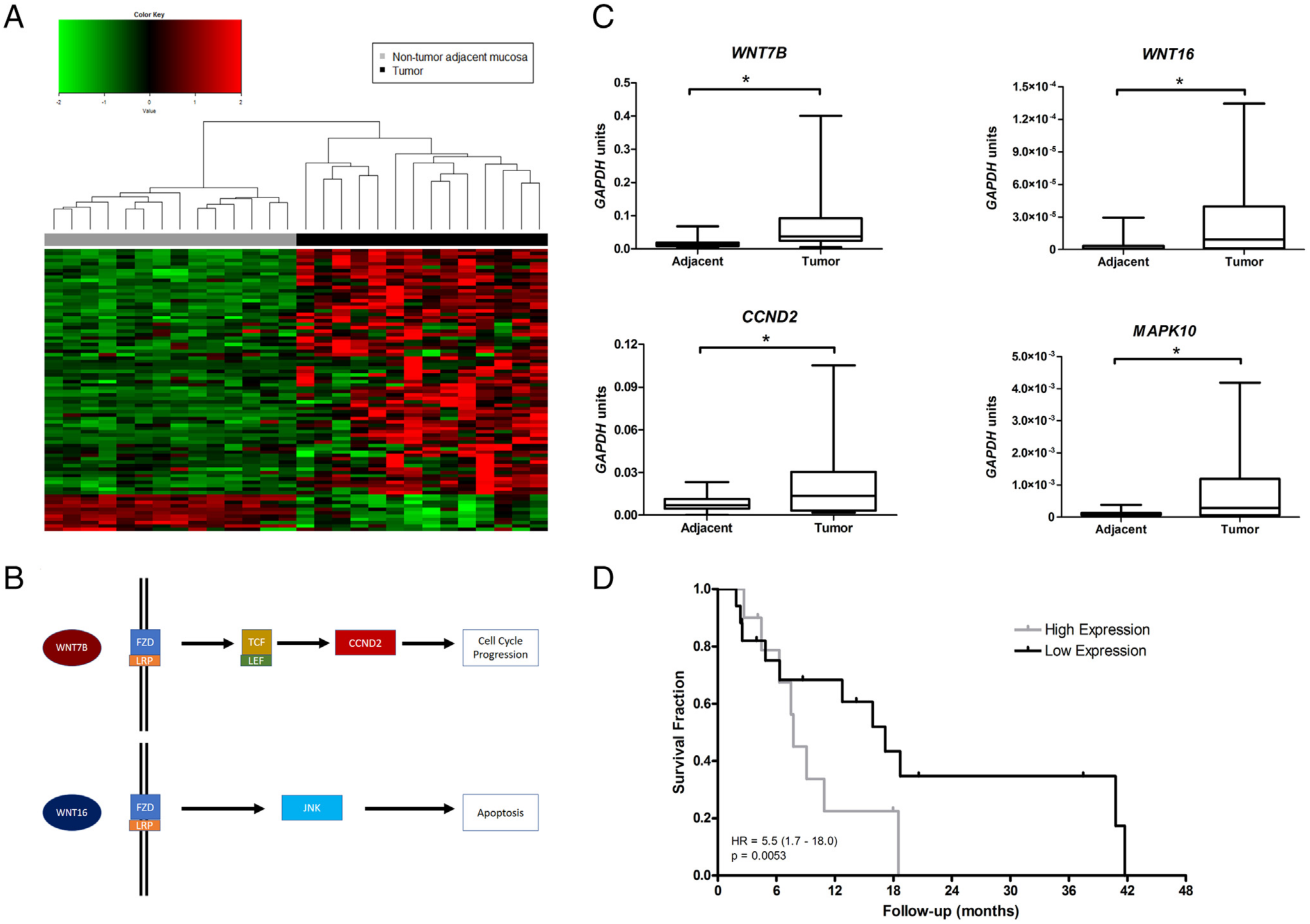

Figure 1. Dysregulation of the Wnt signaling pathway in ESCC. (A) Heatmap showing all differentially expressed genes (DEG) of the Wnt signaling pathway detected by RNA-seq analysis. Black: tumor samples; gray: nontumor adjacent mucosa. Each column represents a single sample; each line represents a single DEG. Red and green indicate high and low gene expression, respectively. (B) Schematic representation of the Wnt-signaling pathway; with overexpressed genes in the canonical (*WNT7B* and *CCND2*) and in the noncanonical pathway (*WNT16* and *JNK*) selected for validation by RT-qPCR. (C) Validation analysis of the Wnt signaling pathway selected targets: *WNT7B*, *WNT16*, *CCND2*, and *MAPK10.* (D) Kaplan-Meier plot of overall survival of patients of the validation set showing the prognostic impact of *WNT7B* expression. Patients with low *WNT7B* expression (<0.077 *GAPDH* units) showed a more favorable survival than those with a high expression (≥0.077 *GAPDH* units); **P* < .05.

tumors (Figure 3*A*). The most common conversion was AT>GC (31.7%), not considered in IS (Figure 3*B*). GC>AT was the second most frequent conversion (29.3%) in VS and the most frequent one in IS (45.5%); AT>TA and GC>TA were found in at least 10% of cases in both sets; InDel and GC>CG and AT>CG conversions showed smaller frequencies (Figure 3*B*). Figure 3*C* shows the distribution of *TP53* mutations per exon in IS and VS. The vast majority of mutations was found in exons 5 to 8 (altogether comprising 82% in the IS and 90% in the VS), while mutations were also observed in exons 4 (3% in the VS only), 9 (9% the IS only), 10 (9% the IS and 3% in the VS), and 11 (3% in the VS only). In IS and VS, 64% and 86% of missense mutations were observed, respectively, while all others comprised nonsense mutations in both datasets (Supplementary Table 3). Mutations were also identified in *TP53* introns in VS, always coexisting with exonic mutations. The presence of *TP53* mutations was not associated with clinical and pathology data (data not shown).

Gene Fusions Analysis in ESCC

RNA-seq data were further used for investigating the presence of gene fusions in ESCC. Interestingly, although 21 fusions were detected, none of them was shared by different tumors, and the vast majority of them (71%) involved syntenic genes (Table 3). These fusions had not been previously described in the Fusion Cancer Database and Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.

Five of seven patients of IS showed one or two fusions in tumors, and two patients showed a higher number of fusions. Patient 4 was shown to carry eight gene fusions, including two involving the same genes, *MAP4* and *SPINK8*, but with different breakpoints (Table 3), while patient 14 showed six fusions. Mutations were not detected in genes involved in fusions, and fusions were not associated with clinical and pathology data (not shown). However, TCGA data of all genes herein involved in fusions showed that seven gene pairs were co-amplified, albeit with frequencies ranging from 1% to 22%, while five gene pairs showed co-deletion (in 1%-4% of samples), and one gene pair was co-amplified and co-deleted in different samples (Supplementary Figure 3). Moreover, as fusions might be a consequence of dysregulated DNA repair, mainly by nonhomologous end-joining (NHEJ), analysis of 14 NHEJ-related genes from the KEGG database showed that only *ATR* was overexpressed in tumor samples with fusions (*P* = .034) (Figure 4).

Table 2. Association Between Characteristics of the ESCC Patients Included in This Study and Gene Expression Levels of the Wnt Signaling Pathway Components, Assessed by RT-qPCR in the Validation Set (*n*=28)

Clinical and Pathology Data	*WNT7B* Expression	*P* Value	*WNT16* Expression	*P* Value	*CCND2* Expression	*P* Value	*MAPK10* Expression	*P* Value
Age								
<60	4.9×10^{-2} (5.2×10^{-3} to 3.2×10^{-1})	1.0	4.2×10^{-6} (5.4×10^{-8} to 1.3×10^{-4})	.722	7.2×10^{-3} (5.7×10^{-4} to 1.0×10^{-1})	.602	7.5×10^{-4} (3.6×10^{-5} to 1.5×10^{-3})	.763
≥60	3.7×10¯2 (2.2×10^{-2} to 4.0×10^{-1})		7.4×10^{-6} (3.1×10^{-8} to 1.0×10^{-4})		1.4×10^{-2} (1.0×10^{-3} to 5.7×10^{-2})		1.8×10^{-4} (1.4×10^{-5} to 4.1×10^{-3})	
Tumor stage								
I/II	8.0×10^{-2} (5.2×10^{-3} to 3.2×10^{-1})	.72	9.9×10^{-5} (2.4×10^{-6} to 1.3×10^{-4})	.099	2.0×10^{-2} (2.4×10^{-3} to 1.0×10^{-1})	.183	9.7×10^{-4} (6.1×10^{-5} to 1.5×10^{-3})	.389
III/IV	3.7×10¯2 (1.3×10^{-2} to 1.4×10^{-2})		8.4×10^{-6} (3.1×10^{-8} to 9.2×10^{-5})		8.5×10¯3 (5.7×10^{-4} to 5.7×10^{-2})		1.7×10^{-4} (1.4×10^{-5} to 4.1×10^{-3})	
Lymph node metastases								
No	5.9×10^{-2} (5.2×10^{-3} to 4.0×10^{-1})	1.0	2.5×10^{-5} (2.8×10^{-7} to 1.3×10^{-4})	.859	3.8×10^{-2} (1.5×10^{-2} to 1.0×10^{-1})	.149	1.0×10^{-3} (6.1×10^{-5} to 4.1×10^{-3})	.818
Yes	6.2×10^{-2} (2.3×10^{-2} to 3.2×10^{-1})		3.4×10^{-2} (1.1×10^{-7} to 9.9×10^{-5})		7.1×10^{-3} (5.7×10^{-4} to 5.7×10^{-2})		1.0×10^{-3} (1.4×10^{-4} to 1.5×10^{-3})	
Local/distant metastases								
No	3.7×10^{-2} (5.2×10^{-3} to 4.0×10^{-1})	.56	5.7×10^{-6} (1.1×10^{-7} to 1.3×10^{-4})	.370	1.8×10^{-2} (5.7×10^{-4} to 5.7×10^{-2})	.105	9.7×10^{-4} (6.1×10^{-5} to 4.1×10^{-3})	.046
Yes	7.3×10^{-2} (2.2×10^{-2} to 1.4×10^{-1})		6.3×10^{-6} (3.1×10^{-8} to 9.2×10^{-5})		3.2×10^{-3} (1.6×10^{-3} to 3.4×10^{-2})		1.4×10^{-4} (1.4×10^{-5} to 1.2×10^{-3})	
Cox regression	HR = 5.5 (1.7-18.0)	.0053	HR = 1.8 (0.5-6.2)	.33	HR = 0.4 (0.1-1.4)	.57	HR = 0.7 (0.2-2.3)	.16

Discussion

This study reports, for the first time, a comprehensive analysis of the ESCC transcriptome, mutational landscape, and fusion events. It shows that the Wnt signaling pathway was frequently activated through dysregulation of gene expression. Furthermore, *TP53* was the only gene with a high mutation frequency, while gene fusions were apparently random.

RNA-seq is a high-throughput procedure that provides extensive data on genomic alterations with numerous advantages over hybridization-based transcriptome analysis, like detection of rare

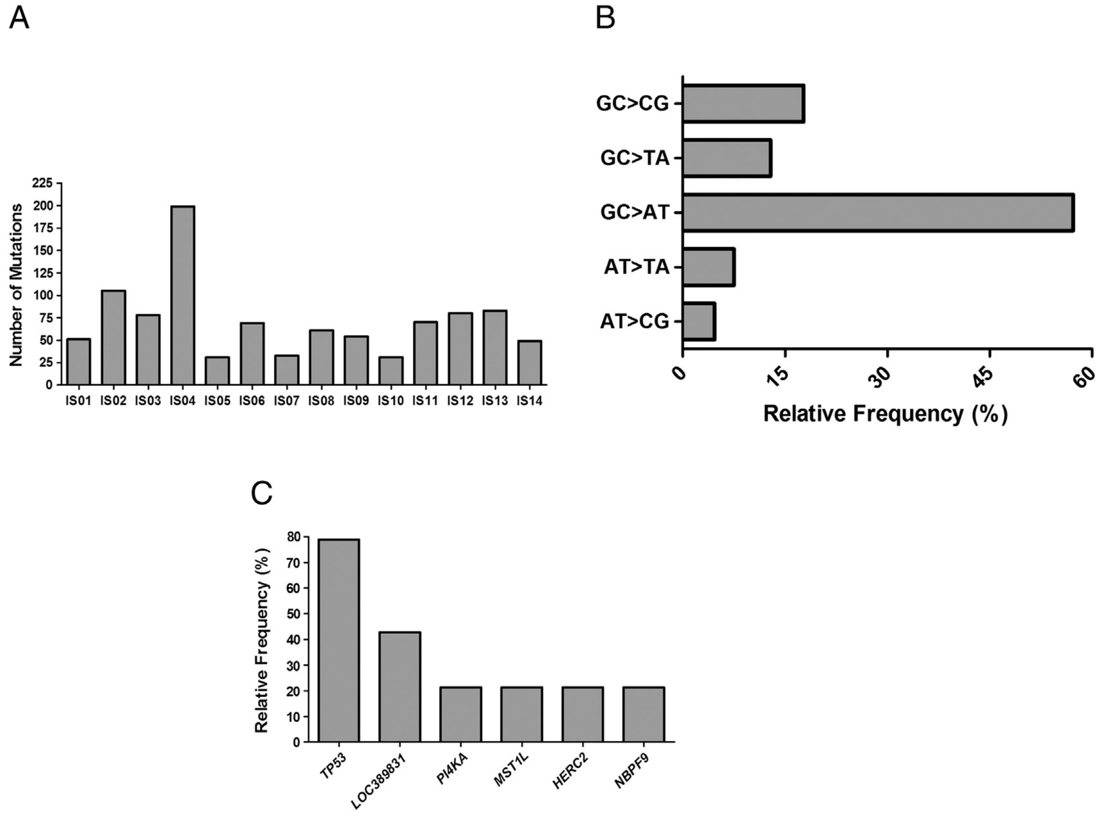

Figure 2. RNA-seq mutational analysis in ESCC. Only missense, nonsense, and synonymous point mutations were considered. SNPs, InDel, and A>T to G>C conversions were removed due to presumptive association with RNA edition. (A) Graphical representation of the number of mutations per tumor sample in the investigation set. (B) Graphical representation of the mutational profiles. (C) Graphical representation of the mutation frequency of the most commonly altered genes in tumors of the investigation set. *IS*, investigation set.

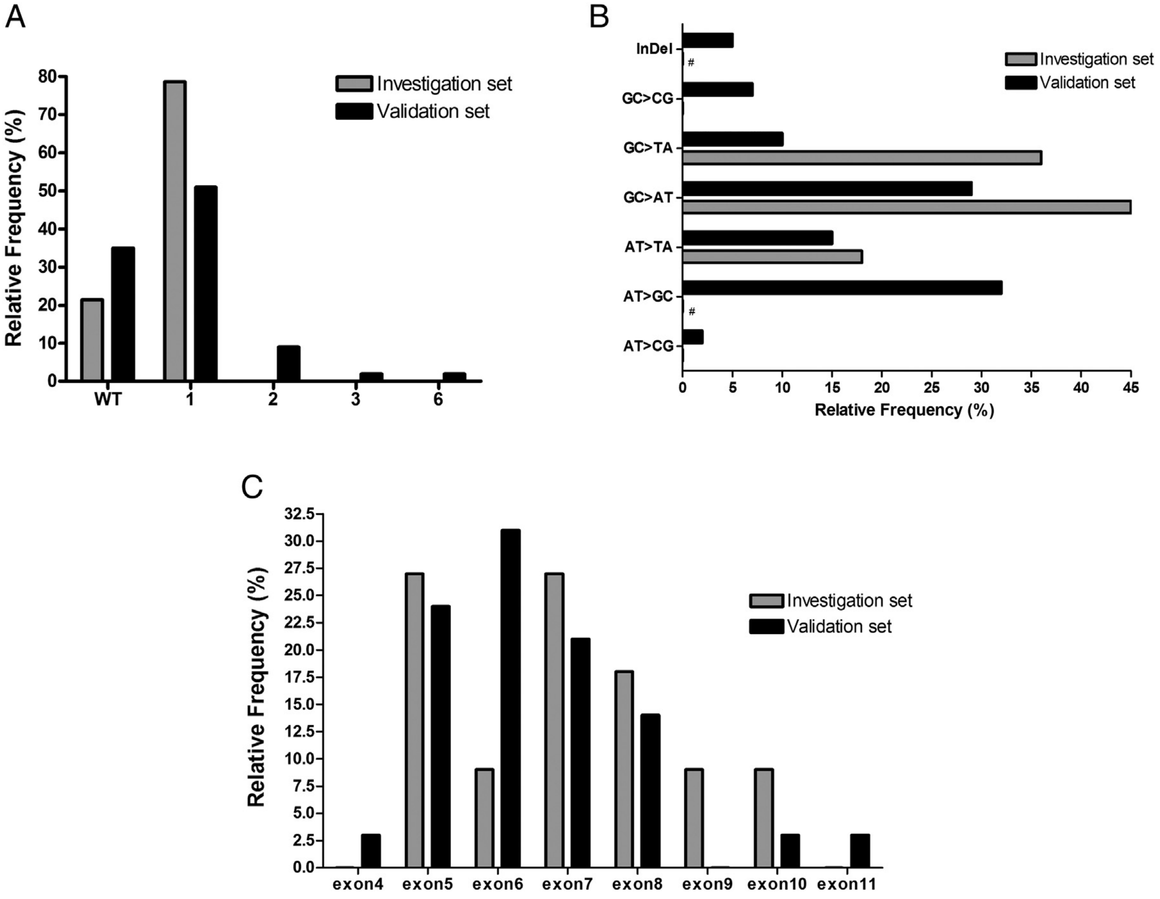

Figure 3. *TP53* mutational analysis by RNA-seq and DNA sequencing. (A) Graphical representation of the frequency of samples without mutations and with different numbers of *TP53* mutations in the investigation set (RNA-seq; n = 14) and validation set (DNA-seq; n = 41). (B) *TP53* InDel and conversions in the investigation and validation sets. (C) Distribution of *TP53* mutations per exon, in the investigation and validations sets. WT = wild type. # not applicable; InDel and A>T to G>C conversions were removed from analysis in the investigation set due to presumptive association with RNA edition.

Table 3. Gene Fusions (n = 21) Identified in the 14 Samples of the Investigation Set by RNA-seq

Sample	5' Gene	3' Gene	5' Breakpoint Position (chr:nt)	3' Breakpoint Position (chr:nt)
IS01	*PLS3*	*IKZF3*	X:114795587	17:37949186
	SLC25A43	*IMPG1*	X:118540664	6:76640868
IS02	*CCDC127*	*AHRR*	5:216844	5:376725
IS04	*RAB3IP*	*CCT2*	12:70132811	12:69985839
	HNRNPC	*CHD8*	14:21731470	14:21869672
	POLR2A	*NLGN2*	17:7416242	17:7317663
	NDUFA10	*MYEOV2*	2:240957970	2:241070505
	MAP4	*SPINK8*	3:48130263	3:48351436
	MAP4	*SPINK8*	3:48130263	3:48361108
	FLNB	*SLMAP*	3:58134579	3:57893611
	ZNF3	*TAF6*	1082:38:00	7:99711891
IS07	*SPAG9*	*CA10*	17:49197715	17:49825178
IS12	*NDUFAF2*	*ZSWIM6*	5:60241209	5:60768508
IS13	*SGMS1*	*CACNB2*	10:52220433	10:18439812
	CTBP2	*EFCAB5*	10:126848888	17:28378136
	LOC100133315	*DEFB131*	11:71627120	4:9452086
	YWHAZ	*DACH1*	8:101964157	13:72256042
	TRAF3	*CDK12*	14:103244012	17:37665958
	PHF2	*C9orf102*	9:96429522	9:98718196
IS14	*CTTNBP2NL*	*ST7L*	1:112958886	1:113098640
	SFRP1	*SLC20A2*	8:41160980	8:42302280

Chr, chromosome; *nt*, nucleotide; *IS*, investigation set.

and alternative transcripts, splice variants, mutations, and quantification of transcript expression [22]. Although mutational analysis of transcripts is restricted following removal of InDel mutations and A:T>G:C conversions largely produced by RNA editing [16], all other mutations in coding regions can be successfully analyzed, as was the case in liver and prostate cancer [23,24].

Less than 10 studies of ESCC alterations by RNA-seq have been reported in the literature, most of which of Chinese patients and with similar number of patients to those included in our study [25–31]. These studies also showed different but comparable estimates of DEG (1425 to 6150) to our findings (6698), depending on sample size, fold-change and P value cutoffs, with a predominance of overexpression. The difference in DEG among the different studies may, at least in part, be caused by the heterogeneity among tumor samples, as shown in this study through the comparison of principal component analysis between nontumor adjacent mucosa and tumor samples. Nevertheless, RNA-seq experiments with at least 12 biological replicates were found to be satisfactory for identifying differentially expressed genes for all fold changes [32]. Recently, the most comprehensive RNA-seq study on ESCC carried out by the TCGA consortium [7] included 90 tumors and 3 samples of nontumor esophageal mucosa from Asian, North American, South American, and European populations. Although this study did not compare

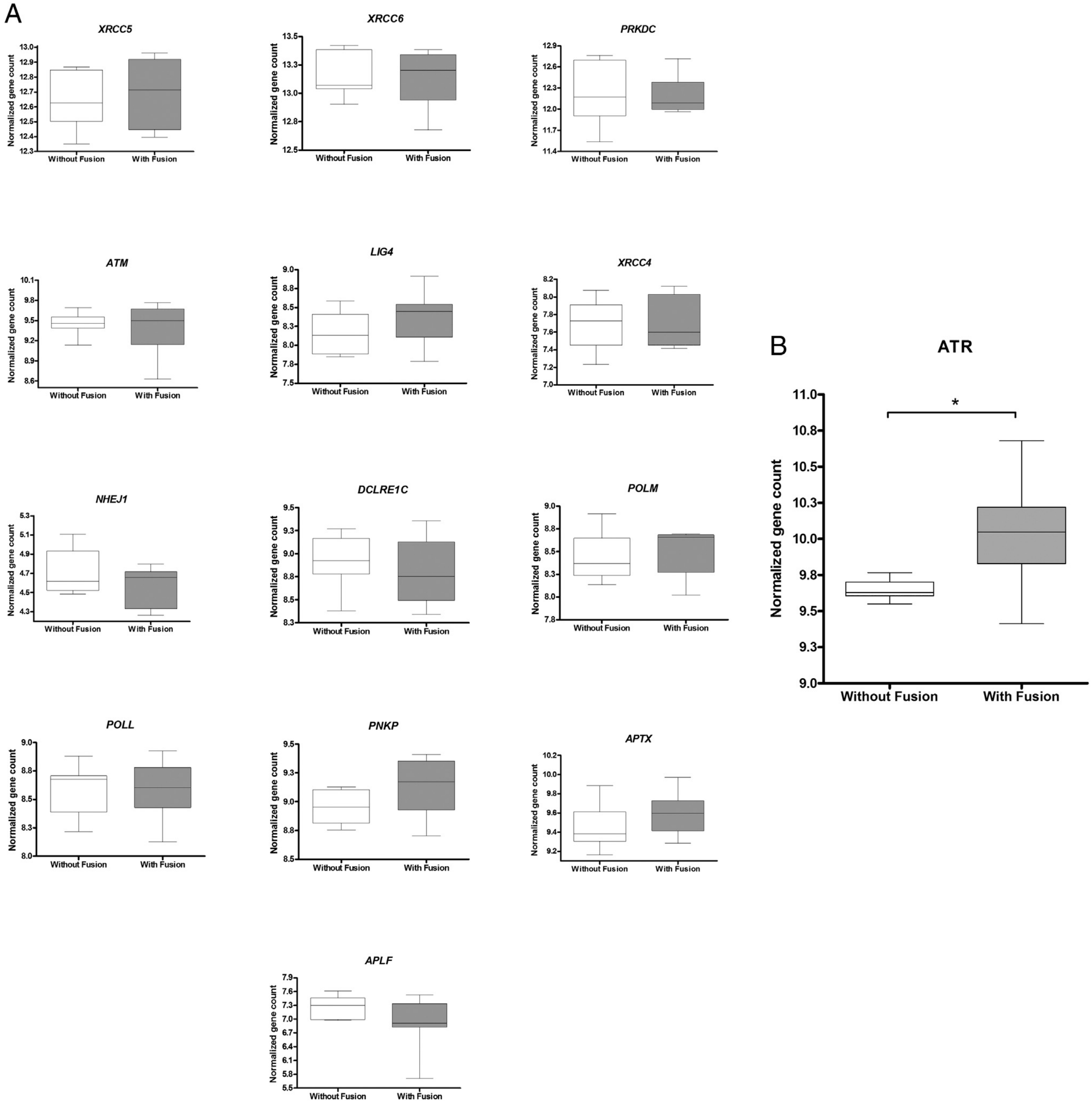

Figure 4. Association between gene fusions and nonhomologous end joining (NHEJ) DNA repair. (A) Boxplots showing expression patterns of genes involved in NHEJ DNA repair pathway (KEGG database) in tumors with and without gene fusion in the investigation set ($n = 14$). (B) *ATR* expression in the same set of samples with and without gene fusions. *$P < .05$.

ESCC tumor with nontumor adjacent tissues, three different molecular subtypes (ESCC1, ESCC2, and ESCC3) were suggested. The ESCC2 subtype, more common among Eastern European and South American patients, was characterized by loss of function of *KDM6A*, *KTM2D*, *PTEN* and *PIK3R1*, *CDK6* amplification and *BST2* overexpression. Interestingly, in our dataset, *CDK6* and *BST2* overexpression was observed in tumors with respect to the nontumor adjacent mucosa (logFC of 1.86 and 2.28, respectively), while *PTEN* was significantly underexpressed (logFC of −0.47). These findings suggested that these alterations might play a role in ESCC development in South American patients.

Although enrichment of alterations of the Wnt pathway has been reported in ESCC studies with RNA-seq [7,26,27], this pathway had not been further explored. Our findings showed 54 DEG in the Wnt pathway, most of which showing overexpression and associated with activation of Wnt signaling. Deregulation of Wnt signaling has been found in other tumors, albeit by different mechanisms. Recently, Guinney et al. [33] proposed a new molecular classification for colorectal cancer, showing that *APC* mutations were present in 70% of cases, being more common (83%) in the consensus molecular subtype 2 (CMS2), followed by CMS3 (78%), CMS4 (66%), and CMS1 (40%). Conversely, breast cancer showed a more similar scenario to ESCC because Wnt activation, preferentially observed in triple-negative invasive carcinomas, was mediated by differential expression of pathway members rather than by mutations [34–36]. Deregulation of *CTNNB1*, *APC*, and *DVL1* expression was found in 21% of all invasive breast carcinomas, while this frequency was much higher (56%) in PAM50 Basal subtype [36]. Although different genes were found to be involved in ESCC and breast cancer, Wnt pathway activation was associated with poor outcome in both tumors [present data; 34-36]. In this study, *WNT7B* overexpression was shown to be an independent prognostic marker in ESCC (HR = 5.5). *WNT7B* overexpression has been reported in breast cancer at mRNA and protein levels, with immunoreactivity in tumor and myeloid cells [37]. These authors showed that *WNT7B* silencing in myeloid cells resulted in reduction of breast tumor mass, volume, and lung metastases in a murine model. Additionally, *WNT16*, a gene involved in Wnt pathway activation by a noncanonical mechanism leading to JNK activation [38], was also found to be overexpressed in our study, as well as its downstream target *MAPK10*, also known as *JNK3*. The activation of this noncanonical pathway has been shown to increase keratinocyte proliferation [38], while JNK has been shown to be necessary for UV-mediated apoptosis [39].

Mutation analysis revealed that *TP53* was the most frequently mutated gene in ESCC, with approximately 70% of tumors showing at least one mutation, mainly missense (86%), and in exons 5 to 8 (90%). Most mutations resulted in a nonfunctional protein. Our RNA-seq analyses were more sensitive than previous ones based on Sanger sequencing pointing to *TP53* mutations in 35% of Brazilian patients with ESCC [40]. It also confirmed TCGA and the International Cancer Genome Consortium (ICGC) studies showing *TP53* as the only gene that was very frequently affected by mutations in ESCC (93% and 62%, respectively). *LOC389831,* a gene coding for uncharacterized isoforms, was the second most frequently mutated gene (43%) in our study, followed by *PI4KA*, *HERC2, NBPF9,* and *MST1L*, each with 21.4%. *Phosphatidylinositol 4-Kinase Alpha* (*PI4KA*) catalyzes the first step of the biosynthesis of phosphatidylinositol 4,5-bisphosphate. It has been found to be overexpressed in hepatocarcinoma, showing a positive and significant correlation with expression of PCNA and KI67 proliferation markers and associated to a poor prognosis [41]. *HECT Domain And RCC1-Like Domain-Containing Protein 2* (*HERC2*) facilitates the assembly of the RNF8-UBC13 complex to recruit BRCA1 to DNA damaged sites, and it also plays a role in p53 oligomerization [42]. Its overexpression has been previously associated with a favorable prognosis in non–small cell lung cancer [43]. *Neuroblastoma Breakpoint Family Member 9* (*NBPF9*) has been found to be overexpressed in lung adenocarcinoma [44], while *Macrophage Stimulating 1 Like* (*MSTIL*) codes for a serine-type endopeptidase of unknown status in tumors (www.genecards.org, GC01M016754).

Our findings were not coincident with TCGA and ICGC studies because none of these four above-mentioned genes were included among the 20 most mutated genes in these databases. A likely explanation for this discrepancy might be the exclusion of A:T>G:C conversions and InDel in our RNA-seq mutational analysis and the ESCC heterogeneity between populations since only 4 of the 10 most frequently mutated genes were shared by TCGA and ICGC data, albeit with different mutation frequencies, probably due to the different ethnic composition of patients included in these studies, *viz.*, Chinese in ICGC and a heterogeneous set of patients in TCGA. These findings pointed to the heterogeneous mutational landscape of ESCC, without a major mutation accounting for the likely activation of a driver oncogene, thus making the development of a target-specific tyrosine kinase inhibitor difficult.

Although half of the tumors herein studied showed fusions, none of them was recurrent or involved the same genes in different patients, and most of them involved syntenic loci. Approximately 33% of the gene pairs involved in fusions were co-amplified in TCGA, like the syntenic genes *CCDC127* and *AHRR*. This pointed to a likely occurrence of chromoanagenesis and formation of micronuclei initiated by error in mitotic segregation, a common event in malignancies with a highly dysregulated cell cycle like ESCC. Newly formed micronuclei frequently show a reduced nuclear import consequently to which defective response to DNA damage signaling and delayed DNA replication occur. Cells may thus enter in mitosis with micronuclei still undergoing DNA replication, resulting in chromosome fragmentation and repair of double-strand breaks by nonhomologous end joining (NHEJ), leading to deletions, translocations, and formation of double minute chromosomes. These small, circular DNA fragments might contain one or more genes, usually oncogenes, while lacking centromeres and telomeres [45], often present at many copies per cell. Double minute chromosomes have been reported in ESCC in association to amplification of two candidate oncogenes, *FGFR1* and *LETM2* [46]. Furthermore, the association between *ATR* overexpression and presence of fusions in our dataset further suggested the occurrence of chromoanagenesis in ESCC, although further studies are necessary for confirming this phenomenon.

This study showed that there was a high heterogeneity among ESCC patients, either for each type of genomic alteration or between them, and also when compared to other studies, such as TCGA. Therefore, we conclude that genomic alterations in ESCC are highly heterogeneous, impacting negatively in target therapy development for ESCC.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tranon.2018.08.002.

Acknowledgements

The authors would like to thank the National Tumor Bank from Instituto Nacional de Câncer (BNT/INCA) for sample collection and the Pathology Division (DIPAT/INCA) for diagnosis confirmation.

Funding

This work was supported by Swiss Bridge Foundation, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), and Ministério da Saúde (INCA). The sponsors had no involvement in either study conception or execution.

References

[1] Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, and Bray F (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386.

[2] Rustgi AK and El-Serag HB (2014). Esophageal carcinoma. *N Engl J Med* **371**, 2499–2509.

[3] Cohen DJ and Ajani J (2011). An expert opinion on esophageal cancer therapy. *Expert Opin Pharmacother* **12**, 225–239.

[4] Mauer AM, Kraut EH, Krauss SA, Ansari RH, Kasza K, Szeto L, and Vokes EE (2005). Phase II trial of oxaliplatin, leucovorin and fluorouracil in patients with advanced carcinoma of the esophagus. *Ann Oncol* **16**, 1320–1325.

[5] Homs MY, Steyerberg EW, Eijkenboom WM, Siersema PD, and D.S.S. Group (2006). Predictors of outcome of single-dose brachytherapy for the palliation of dysphagia from esophageal cancer. *Brachytherapy* **5**, 41–48.

[6] Napier KJ, Scheerer M, and Misra S (2014). Esophageal cancer: a review of epidemiology, pathogenesis, staging workup and treatment modalities. *World J Gastrointest Oncol* **6**, 112–120.

[7] C.G.A.R. NetworkA.W.G.A. UniversityB.C. AgencyB.a.W.s. HospitalB. InstituteB. UniversityC.W.R. UniversityD.-F.C. InstituteD. University, et al (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175.

[8] Schmieder R and Edwards R (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864.

[9] Trapnell C, Pachter L, and Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111.

[10] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and G.P.D.P. Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.

[11] Anders S, Pyl PT, and Huber W (2015). HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169.

[12] Anders S and Huber W (2010). Differential expression analysis for sequence count data. *Genome Biol* **11**, R106.

[13] Carlson M (2016). KEGG.db: a set of annotation maps for KEGG. R package version 3.2.3; 2016.

[14] Nicolau-Neto P, Da Costa NM, de Souza Santos PT, Gonzaga IM, Ferreira MA, Guaraldi S, Moreira MA, Seuánez HN, Brewer L, and Bergmann A, et al (2018). FOXM1 on patient outcome through novel PIK3R3 mediated activation of PI3K signaling pathway. *Oncotarget* **9**, 16634–16647.

[15] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21.

[16] Picardi E, Manzari C, Mastropasqua F, Aiello I, D'Erchia AM, and Pesole G (2015). Profiling RNA editing in human tissues: towards the inosinome atlas. *Sci Rep* **5**, 14941.

[17] Wang K, Li M, and Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**e164.

[18] Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM; 2013 [arXiv:1303.3997v1 [q-bio.GN]].

[19] Ge H, Liu K, Juan T, Fang F, Newman M, and Hoeck W (2011). FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics* **27**, 1922–1928.

[20] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, and Larsson E, et al (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1.

[21] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, and Larsson E, et al (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404.

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

[23] Choi JH, Kim MJ, Park YK, Im JY, Kwon SM, Kim HC, Woo HG, and Wang HJ (2017). Mutations acquired by hepatocellular carcinoma recurrence give rise to an aggressive phenotype. *Oncotarget* **8**, 22903–22916.

[24] Ma Y, Miao Y, Peng Z, Sandgren J, De Ståhl TD, Huss M, Lennartsson L, Liu Y, Nistér M, and Nilsson S, et al (2016). Identification of mutations, gene expression changes and fusion transcripts by whole transcriptome RNAseq in docetaxel resistant prostate cancer cells. *Springerplus* **5**, 1861–1872.

[25] Ma S, Bao JYJ, Kwan PS, Chan YP, Tong CM, Fu L, Zhang N, Tong AHY, Qin YR, and Tsao SW, et al (2012). Identification of PTK6, via RNA sequencing analysis, as a suppressor of esophageal squamous cell carcinoma. *Gastroenterology* **143**, 675–686.e612.

[26] Tong M, Chan KW, Bao JY, Wong KY, Chen JN, Kwan PS, Tang KH, Fu L, Qin YR, and Lok S, et al (2012). Rab25 is a tumor suppressor gene with antiangiogenic and anti-invasive activities in esophageal squamous cell carcinoma. *Cancer Res* **72**, 6024–6035.

[27] Jiang YZ, Li QH, Zhao JQ, and Lv JJ (2014). Identification of a novel fusion gene (HLA-E and HLA-B) by RNA-seq analysis in esophageal squamous cell carcinoma. *Asian Pac J Cancer Prev* **15**, 2309–2312.

[28] Chen Y, Yin D, Li L, Deng YC, and Tian W (2015). Screening aberrant methylation profile in esophageal squamous cell carcinoma for Kazakhs in Xinjiang area of China. *Mol Biol Rep* **42**, 457–464.

[29] Wei G, Luo H, Sun Y, Li J, Tian L, Liu W, Liu L, Luo J, He J, and Chen R (2015). Transcriptome profiling of esophageal squamous cell carcinoma reveals a long noncoding RNA acting as a tumor suppressor. *Oncotarget* **6**, 17065–17080.

[30] Fu JH, Wang LQ, Li T, and Ma GJ (2015). RNA-sequencing based identification of crucial genes for esophageal squamous cell carcinoma. *J Cancer Res Ther* **11**, 420–425.

[31] Li CQ, Huang GW, Wu ZY, Xu YJ, Li XC, Xue YJ, Zhu Y, Zhao JM, Li M, and Zhang J, et al (2017). Integrative analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma. *Oncogenesis* **6**e297.

[32] Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, and Owen-Hughes T, et al (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839–851.

[33] Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, and Angelino P, et al (2015). The consensus molecular subtypes of colorectal cancer. *Nat Med* **21**, 1350–1356.

[34] Geyer FC, Lacroix-Triki M, Savage K, Arnedos M, Lambros MB, MacKay A, Natrajan R, and Reis-Filho JS (2011). β-Catenin pathway activation in breast cancer is associated with triple-negative phenotype but not with CTNNB1 mutation. *Mod Pathol* **24**, 209–231.

[35] Dey N, Barwick BG, Moreno CS, Ordanic-Kodani M, Chen Z, Oprea-Ilies G, Tang W, Catzavelos C, Kerstann KF, and Sledge GW, et al (2013). Wnt signaling in triple negative breast cancer is associated with metastasis. *BMC Cancer* **13**, 537–551.

[36] De P, Carlson JH, Wu H, Marcus A, Leyland-Jones B, and Dey N (2016). Wnt-beta-catenin pathway signals metastasis-associated tumor cell phenotypes in triple negative breast cancers. *Oncotarget* **7**, 43124–43149.

[37] Yeo EJ, Cassetta L, Qian BZ, Lewkowich I, Li JF, Stefater JA, Smith AN, Wiechmann LS, Wang Y, and Pollard JW, et al (2014). Myeloid WNT7b mediates the angiogenic switch and metastasis in breast cancer. *Cancer Res* **74**, 2962–2973.

[38] Teh MT, Blaydon D, Ghali LR, Briggs V, Edmunds S, Pantazi E, Barnes MR, Leigh IM, Kelsell DP, and Philpott MP (2007). Role for WNT16B in human epidermal keratinocyte proliferation and differentiation. *J Cell Sci* **120**, 330–339.

[39] Tournier C, Hess P, Yang DD, Xu J, Turner TK, Nimnual A, Bar-Sagi D, Jones SN, Flavell RA, and Davis RJ (2000). Requirement of JNK for stress-induced activation of the cytochrome c-mediated death pathway. *Science* **288**, 870–874.

[40] Rossini A, de Almeida Simão T, Marques CB, Soares-Lima SC, Herbster S, Rapozo DC, Andreollo NA, Ferreira MA, El-Jaick KB, and Teixeira R, et al (2010). TP53 mutation profile of esophageal squamous cell carcinomas of patients from Southeastern Brazil. *Mutat Res* **696**, 10–15.

[41] Ilboudo A, Nault JC, Dubois-Pot-Schneider H, Corlu A, Zucman-Rossi J, Samson M, and Le Seyec J (2014). Overexpression of phosphatidylinositol 4-kinase type IIIα is associated with undifferentiated status and poor prognosis of human hepatocellular carcinoma. *BMC Cancer* **14**, 7–14.

[42] Cubillos-Rojas M, Amair-Pinedo F, Peiró-Jordán R, Bartrons R, Ventura F, and Rosa JL (2014). The E3 ubiquitin protein ligase HERC2 modulates the activity of tumor protein p53 by regulating its oligomerization. *J Biol Chem* **289**, 14782–14795.

[43] Bonanno L, Costa C, Majem M, Sanchez JJ, Rodriguez I, Gimenez-Capitan A, Molina-Vila MA, Vergnenegre A, Massuti B, and Favaretto A, et al (2016). Combinatory effect of BRCA1 and HERC2 expression on outcome in advanced non-small-cell lung cancer. *BMC Cancer* **16**, 312–317.

[44] Zhang Y, Wang H, Wang J, Bao L, Wang L, Huo J, and Wang X (2015). Global analysis of chromosome 1 genes among patients with lung adenocarcinoma, squamous carcinoma, large-cell carcinoma, small-cell carcinoma, or non-cancer. *Cancer Metastasis Rev* **34**, 249–264.

[45] Holland AJ and Cleveland DW (2012). Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat Med* **18**, 1630–1638.

[46] Cheng C, Zhou Y, Li H, Xiong T, Li S, Bi Y, Kong P, Wang F, Cui H, and Li Y, et al (2016). Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. *Am J Hum Genet* **98**, 256–274.