Current Topics in Peptide & Protein Research
Vol. 20, 2019

Original Article

An integrative proteogenomics study to identify peptides and protein-coding genes in esophageal squamous cell carcinoma

Pooja*,# and Vidya Niranjan*,$

Department of Biotechnology, RV College of Engineering, RV Vidyanikethan Post, Mysuru Road, Bengaluru - 560 059, Karnataka, India.

ABSTRACT

Esophageal squamous cell carcinoma (ESCC) is considered as the dominant type of malignancy in both Western and Asian countries. The increase in trends of smoking and drug addiction is proportional to cases of ESCC and this has to be addressed. The delay in the detection of ESCC can be countered by using the proteogenomics approach that helps in identifying novel genes and tracking them back to the genome. In this top-down proteomic approach, we propose a customized protein sequence database generated using genomics information of ESCC mapped to mass spectrometry-based proteomics data to derive peptides and protein-coding genes that are missing from the current annotations for esophageal cancer. We carried out global genomics and proteomics analysis using proteogenomics approach and identified unpredicted peptides and their biological functions. A customized database containing six-frame-translated whole-genome sequence data of esophageal cancer was produced to map with mass spectrometry-derived proteomic data. On mapping, we obtained a list of sixteen unique peptides and their respective genes involved in ESCC. On further investigation, the functional site prediction suggests detailed evidence of phosphorylation playing a major role in the modification of the identified peptides. Overall, our study suggested that further analysis of post-translational modifications (PTM) and single nucleotide polymorphism would give more comprehensive information that aids in understanding the disease mechanism and cancer prognosis. Hence the impacts of delay in the detection of esophageal cancer can be countered using proteogenomic approach.

KEYWORDS: proteomics, proteogenomics, top-down proteomics, mass spectrometry, esophageal cancer, whole genome sequencing.

1. INTRODUCTION

Esophageal cancer is a pestilent type of cancer affecting the esophagus. It is a process of tumor formation that occurs in the digestive tract (upper, middle, and lower) [1].The formed tumor leads to symptoms like difficulty in swallowing, hoarse voice, unexplained weight loss, and lymph gland enlargement, etc [2]. The esophageal squamous cell carcinoma is classified into four stages, Stage I, II, III, and IV. Stage I esophageal tumor is small and is limited to the esophagus. At Stage II, the cancer tumor grows larger but still remains limited to the esophagus. In Stage III, the esophageal tumor grows beyond the esophagus and extends to nearby tissues. And at stage IV, the tumor would have grown into bigger size and grown beyond the esophagus and spread to lymph nodes affecting distant sites like the liver and abdominal cavity. Further, the ESCC staging is proposed by the American Joint committee on Cancer (AJCC) that utilizes the TNM (tumor-node-metastasis) model

*Corresponding authors
#pooja.ramesh3990@gmail.com
$niranjan.vidya@gmail.com

classification based on the depth of invasion of the tumor. Stage T (T1-T4) describes the region of origin and size of the primary tumor and layers of the esophagus invaded by cancer (Figure 1). Stage N (N0-N3) describes the degree of cancer spread to the lymph node. And stage M (M0 & M1) refers to the presence of metastasis in distant organs, including lung, liver, and bone tissues.

Esophageal cancer has two sub-types based on the site of impact.

1. Esophegal squamous cell carcinoma
2. Esophegal adenocaricnoma

1.1. Risk factor and global statistics of esophageal cancer

The esophageal squamous cell carcinoma is stimulated by the lifestyle variations of the person. The lifestyle variations include incongruent intervals of food consumption, dipsomania leading to the mutation of enzymes, aldehyde metabolism [3], and tobacco addiction. In Asian countries, the main risk of squamous cell carcinoma is due to alcohol consumption, chewing betel nut, hot beverage drinking, and poor nutrition. Research data suggest that about half of all cases are due to tobacco and about one-third due to alcohol, and the observed cases in men are due to the combination of both smoking and heavy drinking which is a stimulation agent for ESCC [3, 4]. ESCC is considered to be the eighth most common cancer in the world. In Asia, the Northern part of Asia is considered to be most affected, compared to southern Asia. China is the highest affected country in Asia with regard to male population, accounting for a total of 27 out of 100000 males whereas Mongolia is the most affected country with regard to female population, summing to 12 out of 100000 females (Figure 2) [5, 6].

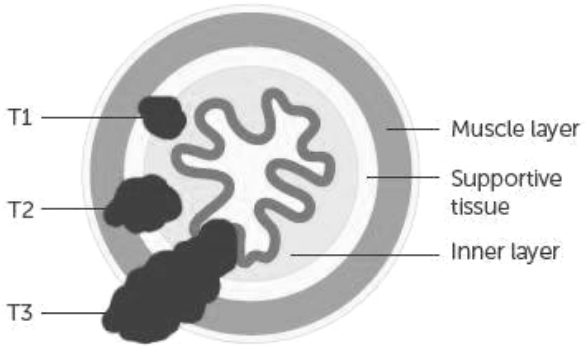

Figure 1. Tumor staging in esophageal cancer: cross-section of the esophagus; food passes through the mucosa (the inner layer); the next layer submucosa, surrounded by muscularis. Tumor growth in mucosa and submucosa is labeled as T1. A tumor that has grown through submucosa into the muscle is labeled as the T2 stage and finally, a tumor that has grown through all three layers of the esophagus into the outer layer adventitia is labeled as the T3 stage.

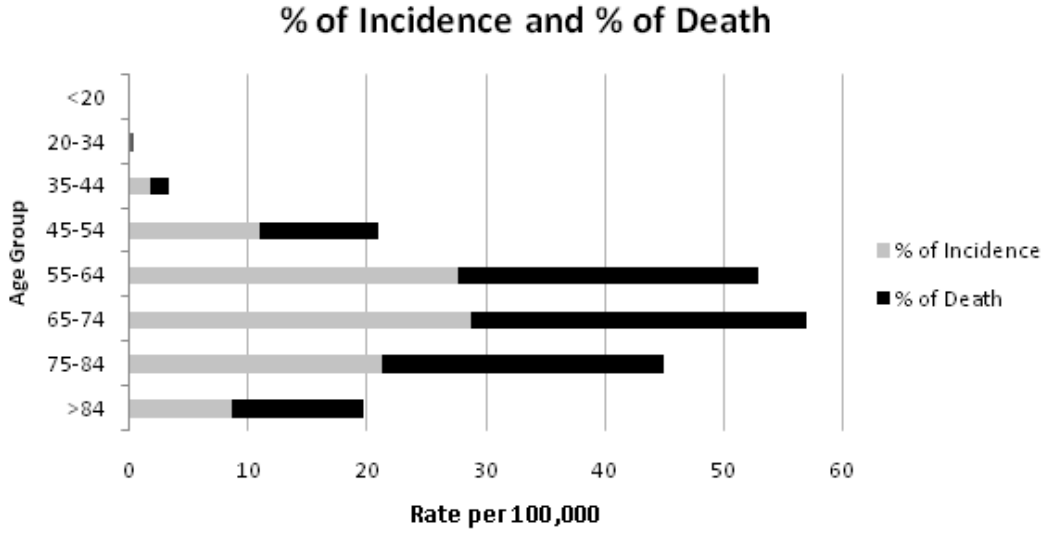

Figure 2. Statistics of occurrences of esophageal cancer in Asia. Age is standardized to the World standard population.

1.2. ESCC cancer genomics and proteomics

The recent advances in technology like high throughput deep sequencing aid in understanding the complexities of human genomics and proteomics thereby easing the effort of exemplifying and characterizing the human proteomics [7, 8]. Proteogenomics studies provide a crucial understanding of the expression patterns of proteins in normal and diseased cell states, capturing the overall proteomic profiles and providing a spotlight for researchers during the recent past. Nevertheless, the researchers are now focusing on protein target identification and verification using quantitative proteomic analysis in union with genomic analysis. However, the proteogenomics pipeline has to overcome several technological hindrances that include the construction of a customized database, speed, and accuracy of peptide identification. A comprehensive database is the significant element of this pipeline, enabling the efficient identification and mapping of predicted peptides to the genomic loci along with annotation information of genes. Software scripts allow the creation of automated genome annotation analysis reports [8-12]. The current work focuses on searching a six-frame translated database (Figure 3) to reveal a few of the novel peptides, which include coding regions, and open reading frames. [9, 13, 14].

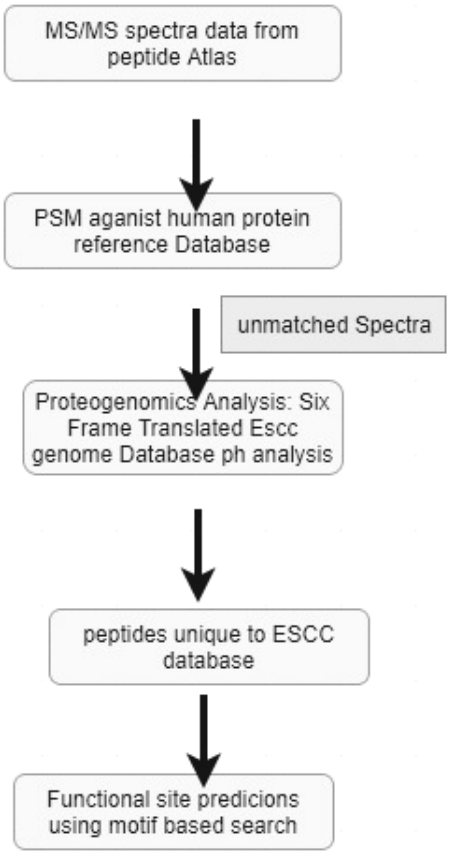

Figure 3. Overview of proteogenomics workflow that illustrates a top-down proteomics approach for the identification of peptides sequences.

2. MATERIALS AND METHODS

2.1. Whole-genome sequencing data

Whole-genome data for human esophageal squamous cell carcinoma were obtained from the NCBI-SRA database repository (Table 1) (Table S1) and used in the construction of a customized six-framed database for each sample. The sequencing was done using the Illumina Solexa platform. The Illumina approach uses sequencing by synthesis technology and clone-free amplification of single-strand DNA in a flow cell where the bridge amplification of single strands of DNA templates takes place. In the amplification process, the single strand of DNA is attached to an adapter in the flow cell and hybridized to synthesize a template strand that is complementary to a single strand DNA molecule. At the end of the amplification process, the flow cell will contain a thousand copies of a single strand DNA molecule. The templates are massively sequenced using sequencing by synthesis approach by utilizing fluorescent dyes and altered DNA polymerases that include reversible chain terminators into growing oligonucleotide chains. The reversible terminators are dyed using four different fluorescent colors to distinguish the bases added to the oligonucleotide chain. The number of molecules sequenced, length of the sequence, and the coverage of the sequence varies across the data.

2.2. Mass spectrometry data

Tandem mass spectrometry (MS/MS) data from the analysis of protein expression of ESCC cell lines were obtained from the Pride Archive, a proteomic data repository from EMBL-EBI (https://www.ebi.ac.uk/pride/archive/projects/PXD006255). These data sets have quantitative consistency and deep proteomic coverage (Table S2). The mass spectrometry-based proteomics is a

Table 1. Information about publicly available whole-genome sequencing data of esophageal cancer used for the construction of a six-frame translated database.

SRA DATA	Size(GB)	Cell Lines	Layout	Read length(bp)	Number of Bases(GB)
SRR1056628	10	Esophageal cancer	Paired end	90	16.5
SRR1056898	10	Esophageal cancer	Paired end	90	16.4
SRR1056899	9	Esophageal cancer	Paired end	90	14.8
SRR1057024	9	Esophageal cancer	Paired end	90	13.9
SRR1057043	4	Esophageal cancer	Paired end	90	5.3
SRR1057037	3	Esophageal cancer	Paired end	90	4.8
SRR1057044	4	Esophageal cancer	Paired end	90	5.6
SRR1057016	9	Esophageal cancer	Paired end	90	12.8

prominent tool in the emerging field of systems biology. Thermo fisher raw files obtained from Pride archive were converted into Mascot Generic Format (MGF) format using a raw converter tool, which was later used as an input file in X!Tandem for further proteogenomics analysis.

2.3. Proteomic analysis

Protein Database: a curated set of protein datasets for humans retrieved from Uniprot Proteomes (https://www.uniprot.org/proteomes). A total of 20395 entries of proteins were downloaded to create a protein database for Homo sapiens. In-house python scripts were used to search for the matched peptides obtained from mass spectrometry data. The results obtained were the unmatched spectra, which did not match with the existing curated protein database [9]. The unmatched spectra were later used to map with six-frame translated genome database to understand novel peptides [15-18].

2.4. Six-frame translated human ESCC genome database

Six-frame translation of the genome provides all possible set of protein-coding sequences. Incorporating a six-frame translated genome in proteogenomics studies aids in discovering novel peptide and open reading frames (ORF) [19]. Proteomic data obtained from tandem mass spectrometer were searched against six-frame translated ESCC genome to identify distinct peptides that mapped back to the genome [20]. The whole-genome sequence data for ESCC obtained from the NCBI-SRA database is translated into a six-frame database using a sequence translational tool EMBOSS Sixpack from EMBL- EBI website (https://www.ebi.ac.uk/Tools/st/emboss_sixpack/). The tool reads a DNA in forward sequences and reverse sense sequences with forward and reverse translations in display format. A standard codon table, along with the genetic code was specified for the translation of DNA sequence for six-frame translation. The algorithm uses fastq files as input to create a six-frame translated database including three forward and three reverse translations. The output is written in a display file and any ORFs that are longer than the specified minimum size (minimum size of ORF is specified as 1) are written to the output sequence file. Amino acid sequences with lesser than ten residues were excluded from the translated database. The whole-genome data was translated using HPC (High-performance computer) Intel Xeon E5 processor with 80 GB RAM and 1TB storage.

2.5. Proteogenomics analysis using X! Tandem

To enhance the annotations of the existing genome and to enable the identification of novel peptides, we searched eight different databases, generated using in-house python scripts and EMBOSS six-pack translational tool against the unmatched spectral proteomic data for proteogenomic analysis. Along with this we also added uniprot reference proteome downloaded from Uniprot. The six-frame translation included stop codon to terminate translation of the template sequence [21]. Translated peptide sequences smaller than 10 amino acids were not included in the database. The raw MS/MS files were obtained from Pride archive. All peak lists of individual runs were converted to Mascot Generic Format

(MGF) before submitting to X! Tandem along with Fasta files of six-frame translated ESCC genomic database.

The following parameters were common to all searches in X!Tandem:

Precursor mass error: 10 p.p.m.,
Fragment mass error: 0.05 Da,
Carbamidomethylation of cysteine: fixed modification,
Oxidation of methionine: variable modification.
Tryptic peptides: 2 missed cleavages.
Peptide sequences: one missed cleavage (length of 6–25 amino acids)
Quick acetyl search: 2 N-terminal amino acids clipping

3. RESULTS AND DISCUSSION

3.1. Impact of whole-genome sequence depth and read type

Whole-genome sequence read normally increases the chance of detection of peptides, novel to ESCC genome, upon searching the database. Eight SRA datasets from NCBI-SRA combined with the Uniprot XML proteome database were selected and shortlisted to create the six-frame translated database and searched against MS/MS data. The coverage of the dataset and the number of reads aligning to the reference genome has a direct influence on the identification of novel peptides. The selected datasets have 20X coverage to ensure the maximum matching of spectra and mining of novel peptides. The sequencing of both ends of cDNA fragments is termed as paired-end (PE) sequencing. The paired-end data provides an advantage of long reads and resolve splicing events. It is always a better approach to use paired-end data for improvised identification of peptides.

3.2. Six-frame translation ESCC genomic DNA

Genomic DNA sequences obtained from NCBI-SRA is converted to all six-reading frames to create a database in Fasta format as shown in Figure 4.

3.3. Protein identification using X! Tande

In the whole-genome proteogenomics mapping approach, peptides from six-frame translated genome were mapped onto the spectra for peptide identification and characterization. Tandem mass spectra data is not biased as it reads protein isoforms from the provided database. The true hits (matched peptides) are then mapped back to the genome to identify the genomic loci. The data consisted of 2,230,503 spectra and identified sixteen distinctive peptides (Table S3). X! Tandem tool was installed locally and eight customized six-frame translated database was created using python script submitted in Fasta format and mass spectrometry data submitted in MGF format. Searches were performed using a mass error tolerance of +/- 2.0 Daltons. The output files have a list of peptides, identified in the search process, along with distinct information on spectra matches for further understanding of genes (Table S4).

The identified peptides encodes for the DELC1 gene and the C16orf62 gene. Methylation of the DELC1 gene found in plasma DNA has been observed in lung cancer, gastric, and esophageal cancer. Methylation in DELC1 has been correlated with loss of gene expression and associated with mutations including inactivated mutations, nonsense mutations, deletions and inactivate functionality of proteins [22]. Located on chromosome 3, the promoter region of the gene is methylated in cancer cases. The alternative splicing transcripts disrupt coding and encode non-functional proteins, and

```
  R  F  S  S  I  S  L                                    F1
   G  S  P  P  S  R  X                                   F2
    V  L  L  H  L  A                                     F3
1 AGGTTCTCCTCCATCTCGCT 20
  ----:----|----:----|
1 TCCAAGAGGAGGTAGAGCGA 20
   L  N  E  E  M  E  S                                   F6
  X  T  R  R  W  R  A                                    F5
    P  E  G  G  D  R                                     F4
```

Figure 4. EMBOSS sixpack output file giving out the forward and reverse sense sequences with the three forward translations and three reverse translations.

hence the DELC gene is usually seen to be down-regulated in cancer. The main promoter in the DELC gene is methylation for the upregulation process [23]. C16orf62 gene encodes for VPS35L, the endosomal protein sorting factor. It has a function in the prediction of intracellular proteins and membrane proteins. It encodes 12 transcription factors is found on chromosome 16 [24].

3.4. Prediction of gene functions using ELM prediction tool

Eukaryotic linear motif predicts the functions of peptides and PTM predictions based on consensus patterns using annotated motif data. We preformed ELM searches for a predicted set of peptides to study the functional sites and post-translational modifications by applying a regular expression method by matching query peptide sequence to curated ELM database. We predicted a total of twenty-seven phosphorylated functional sites for our identified peptide set and other functional sites including ubiquitination, protein-protein interaction mediated by SH3 domains, and histone modifications (Figure 5) (Figure 6) (Table S5). Protein phosphorylation represents widely studied post-translational modifications, as it engages in cellular functions. Kinase-driven phosphorylation alterations at a genetic level lead to a proliferation of cell growth in tumor cells [25]. Alterations in signal transductions pathways due to phosphorylation have cascade reactions in pathways including MAP kinase, tyrosine kinase, and other kinase-dependent pathways which play major roles in cancer cell growth and progression [26, 27]. In esophageal squamous cell carcinoma, it is well-established that the p38 MAP kinase pathway stimulates cell proliferation, cell migration, and cellular transformations that reiterate cancer cell growth [28].

Phosphorylation plays a significant role in cell regulation, including cell cycle, signal transduction pathways, and apoptosis. Protein phosphorylation is a prominent cellular regulatory mechanism involving phosphorylation/dephosphorylation of receptor molecules in the cell cycle [29]. However, the alterations in the functions of the phosphorylation pathway have serious outcomes in cancer cells.

4. CONCLUSION

In the present research, we gauge the significance of peptides and searching peptide sequences in a customized database constructed using ESCC six-frame translated whole-genomic data. We report a top-down proteomics approach for the detection of a diverse range of peptide sequences. The selection of whole-genome sequence data is crucial for spectral matching. We examined the application of deep paired-end data, which has a

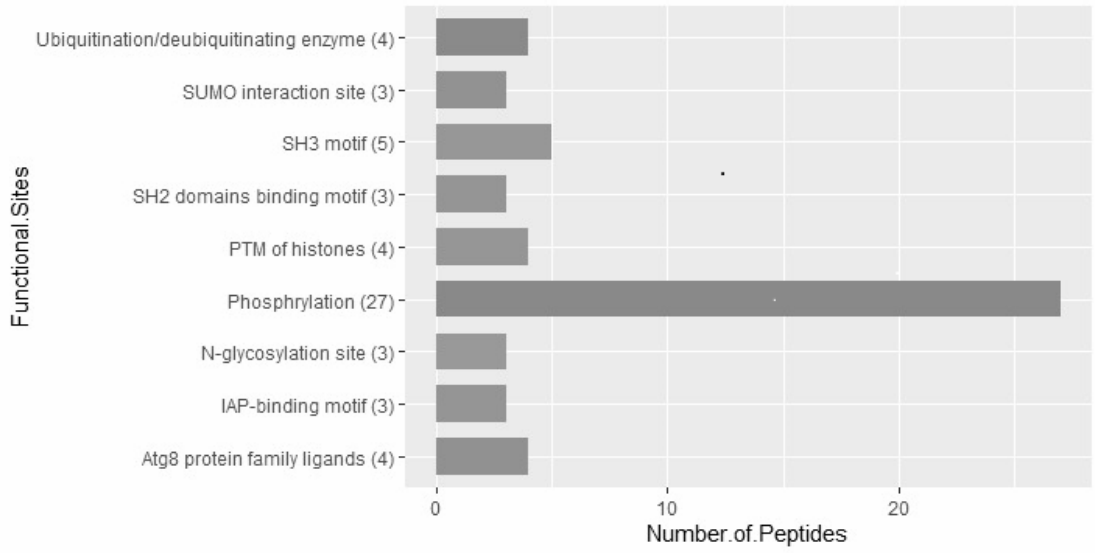

Figure 5. Functional site predictions using ELM prediction tool: Number of peptides mapped against the number of functional sites predicted based on the motif search. The bar graph displays the phosphorylation function that has been predominantly predicted for a large set of peptides.

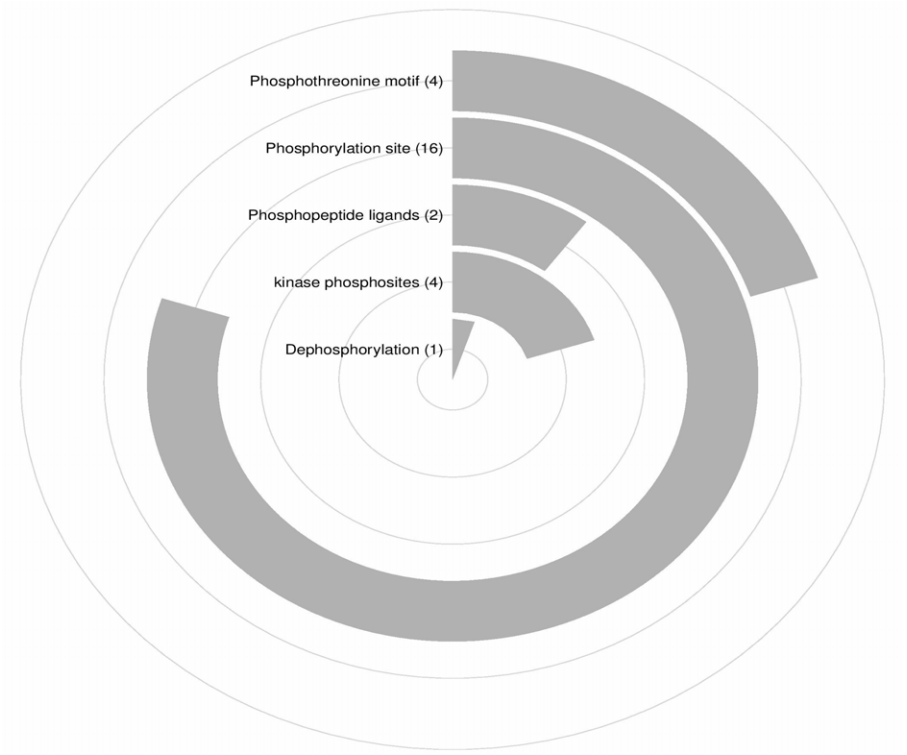

Figure 6. Functional site prediction - Phosphorylation/Dephosphorylation classification (Table S5). The classification representation graph was produced using R programming.

high impact on the identification of peptide sequences. We identified a set of sixteen peptide sequences that were not part of ESCC protein annotations earlier and it encodes for the DELC gene and C16orf62 gene. Besides, investigating the functional sites showed that phosphorylation was predominantly observed for most of the identified peptide sequences suggesting the role of phosphorylated sites and its impact on cancer cell proliferation Furthermore, our work demonstrates the practice of proteogenomic approach by the integration of multi-omics data to further improve our understanding of the complex proteomic variations such as SNPs, PTMs, and novel splice junctions for ESCC and the role of genomic and proteomic variations in human health.

SUPPLEMENTARY INFORMATION

Table S1. Hyperlinked repository information for the RNA-Seq datasets used in this project.

Repository	Accession Number	Links
NCBI GEO	SRR1056628	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1056628
NCBI GEO	SRR1056898	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1056898
NCBI GEO	SRR1056899	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1056899
NCBI GEO	SRR1057024	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1057024
NCBI GEO	SRR1057043	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1057043
NCBI GEO	SRR1057037	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1057037
NCBI GEO	SRR1057044	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1057044
NCBI GEO	SRR1057016	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1057016

Table S2. The MS/MS datasets for the 10 ESCC samples used in this study.

The dataset was obtained from Pride Archive Project Accession number PXD006255
FFPE_RPS_1
FFPE_RPS_2
FFPE_RPS_3
FFPE_RPS_4
FFPE_RPS_5
FFPE_RPS_6
FFPE_SCX_1
FFPE_SCX_2
FFPE_SCX_3
FFPE_SCX_4
FFPE_SCX_5
FFPE_SCX_6

Table S3. List of Novel peptides.

QSSLGLSFFNSILAHGDLR
MRPSESFLLEFLCNFFSTLLIVPDHPEHGVLFLVR
LNQLSVNLWHLAQR
LNLYLHSGQVALANQCLSQADAFFK
ISDCHIIEVEPGTGVIEPSEVGDFELNFTGGVPGPTS
AIATVGFVEQPPFGILPSVFELAPGHAILVEVL
SSELYESTMVVEGVLGEK
HGDTVQNQLVVQGVELPSYLPLYPPAMDWIFQ
GSVEPQVLLEPYALIIPGENYIGINVK
LAEFEDELDHTVDSLTWNLTPK
TNECQGTMWAPTSPPAGSSSPSQPTWK
SQPDPGELTALTAQHFIR
GIGDPLVSVYAR
IYTCVLHLLSAMSQETYLYHIDK
ACVAYCFITIPSLAGIFTR
DFEQQLSFYVESR

Table S4. List of Novel peptides with description.

Gene	Peptide Sequence	Genomic Coordinates of identified Peptides	X!Tandem score (log -e)	MH+1	Mass error (Da)	Charge	Name of Protein	Protien information
C16orf62	QSSLGLSFFNSILAHGDLR	16:19710836-19710892(+)	-4.8	1596.77	2237.30	2	VPS35 endosomal protein sorting factor	esophageal cancer associated protein
C16orf62	MRPSESFLLEFLCNFFSTLLIVPDHPEHGVLFLVR	16:19,680,553-19,680,621(+)	-11	2326.25	2045.99	2	VPS35 endosomal protein sorting factor	
C16orf62	LNQLSVNLWHLAQR	16:19710902-19710943(+)	-4.2	2633.28	1076.53	2	VPS35 endosomal protein sorting factor	
C16orf62	LNLYLHSGQVALANQCLSQADAFFK	16:19,566786-19711798(+)	-7.1	2137.87	1298.61	3	VPS35 endosomal protein sorting factor	esophageal cancer associated protein
DLEC1	ISDCHIIEVEPGTGVIEPSEVGDFELNFTGGVPGPTS	3:38138718-38139059(+)	-3.9	1821.86	958.53	2	DLEC	Deleted in lung & esophageal cancer protein
DLEC1	AIATVGFVEQPPFGILPSVFELAPGHAILVEVL	3:38138718-38139059(+)	-9.7	1873.01	1596.77	2	DLEC	Deleted in lung & esophageal cancer protein
DLEC1	SSELYESTMVVEGVLGEK	3:38163781-38163834(+)	-6.1	1175.65	2293.97	2	DLEC	Deleted in lung & esophageal cancer protein
C1orf62	HGDTVQNQLVVQGVELPSYLPLYPPAMDWIFQ	16:19628011-19628106(+)	-3.2	998.53	779.42	2	VPS35 endosomal protein sorting factor	esophageal cancer associated protein
DLEC1	SQPDPGELTALTAQHFIR	3:3809365-38093708(+)	-5.6	1119.62	1460.70	2	DLEC	Deleted in lung & esophageal cancer protein
C1orf62	GIGDPLVSVYAR	16:19610365-19610400(+)	-5	1210.63	1567.78	2	VPS35 endosomal protein sorting factor	esophageal cancer associated protein
C1orf62	IYTCVLHLLSAMSQETYLYHIDK	16:19,682,321-19,682,389 (+)	-8.7	1991.90	1034.53	2	VPS35 endosomal protein sorting factor	esophageal cancer associated protein
C1orf62	ACVAYCFITIPSLAGIFTR	16:19651976-19652032(+)	-10.1	1783.84	1873.01	1	VPS35 endosomal protein sorting factor	

Table S4 continued..

C1orf62	DFEQQLSFYVESR	16: 19659121-19659159(+)	-4.4	1385.70	779.42	1	VPS35 endosomal protein sorting factor	
DLEC1	GSVEPQVLLEPYALIIPGENYIGINVK	3:38,138,151-38,138,216(+)	-3	1008.61	1425.70	2	DLEC	Deleted in lung & esophageal cancer protein
DLEC1	LAEFEDELDHTVDSLTWNLTPK	3:38,103,764-38,103,829(+)	-6.3	1783.84	1152.66	2	DLEC	Deleted in lung & esophageal cancer protein
DLEC1	TNECQGTMWAPTSPPAGSSSPSQPTWK	3:38,080,762-38,080,842(+)	-3.8	1385.70	998.53	2	DLEC	Deleted in lung & esophageal cancer protein

Table S5. List of Functional sites predicted for peptide sequences.

Functional Site Prediction using ELM prediction tool	**Motif Function**	**ELM Name**
Peptide Sequence- QSSLGLSFFNSILAHGDLR	APC/C-binding proteins	DEG_APCC_TPR_1
	N-terminal protein motifs	DEG_Nend_UBRbox_3
	Canonical LIR motif that binds to Atg8 protein family to mediate processes involved in autophagy.	LIG_LIR_Gen_1
	LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	LIG_LIR_Nem_3
	phosphorylation site	MOD_GSK3_1
	phosphorylation site	MOD_NEK2_1
	kinase phosphosites-play key roles during multiple stages of mitosis including prophase, metaphase, anaphase, and cytokinesis	MOD_Plk_4
Peptide Sequence- MRPSESFLLEFLCNFFSTLLI VPDHPEHGVLFLVR	SUMO interaction site	LIG_SUMO_SIM_anti_2
	USP7 binding motif- deubiquitinating enzyme that cleaves ubiquitin moieties from its substrates	DOC_USP7_MATH_1
	Ca2+- and calmodulin-regulated serine/threonine protein phosphatase known to affect cell biological function	DOC_PP2B_LxvP_1
	SH3 motif is involved in protein-protein interaction mediated by SH3 domains.	LIG_SH3_3
	Pex14 ligand motif	LIG_Pex14_2
	FHA phosphopeptide ligands	LIG_FHA_1
	BRCT phosphopeptide ligands domains are protein modules mainly found in Eukaryota	LIG_BRCT_BRCA1_1
	SUMO interaction site	LIG_SUMO_SIM_par_1
	NEK2 phosphorylation site	MOD_NEK2_1
	Polo-like kinase phosphosites	MOD_Plk_4
	The Nuclear Export Signal (NES) is a linear motif involved in the regulated export of macromolecules from the nucleus via the nuclear pores	TRG_NES_CRM1_1
	PKA Phosphorylation site	MOD_PKA_2
Peptide Sequence- LNQLSVNLWHLAQR	ubiquitin-dependent proteasomal degradation	DEG_Nend_Nbox_1
Peptide Sequence- LNLYLHSGQVALANQCLS QADAFFK	Caspase cleavage motif important role in programmed cell death (apoptosis)	CLV_C14_Caspase3-7

Table S5 continued..

	ubiquitin-dependent proteasomal degradation	DEG_Nend_Nbox_1
	RIR motif-replicative bypass of DNA lesions	LIG_REV1ctd_RIR_1
	SH2 domains binding motif.	LIG_SH2_STAT5
	SH2 domains binding motif.	LIG_SH2_SRC
	Post-translational modification of histones	LIG_WD40_WDR5_VDV_2
	Glycosaminoglycan attachment site	MOD_GlcNHglycan
	N-glycosylation site	MOD_N-GLC_2
	PIKK phosphorylation site	MOD_PIKK_1
Peptide Sequence- ISDCHIIEVEPGTGVIEPSEVGDFELNFTGGVPGPTS	ubiquitin-dependent proteasomal degradation	DEG_Nend_Nbox_1
	Phosphothreonine motif	LIG_FHA_1
	Pex14 ligand motif	LIG_Pex14_2
	SH3 motif is involved in protein-protein interaction mediated by SH3 domains.	LIG_SH3_3
	SUMO interaction site	LIG_SUMO_SIM_par_1
	TRAF6 protein acts as intracellular adaptor that is recruited to different receptors through its C-terminal TRAF domain.	LIG_TRAF6
	Post-translational modification of histones	LIG_WD40_WDR5_VDV_2
	N-glycosylation site	MOD_N-GLC_1
	NEK2 phosphorylation site-NEK Serine/Threonine protein kinase family	MOD_NEK2_1
Protein Sequence-AIATVGFVEQPPFGILPSVFELAPGHAILVEVL	PP2A-mediated protein dephosphorylation is involved in a broad range of cellular processes including cell-cycle progression,	DOC_PP2A_B56_1
	IAP-binding motif (IBM)-Inhibitor of Apoptosis Proteins	LIG_BIR_II_1
	IAP-binding motif (IBM)-Inhibitor of Apoptosis Proteins	LIG_BIR_III_3
	SH3 motif is involved in protein-protein interaction mediated by SH3 domains.	LIG_SH3_3
	PDZ domain ligands	LIG_PDZ_Class_3
	TRFH domain docking motifs	LIG_TRFH_1
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
	CK2 Phosphorylation site	MOD_CK2_1
Protien Sequence- SSELYESTMVVEGVLGEK	IAP-binding motif (IBM)-Inhibitor of Apoptosis Proteins	LIG_BIR_II_1
	Phosphothreonine motif	LIG_FHA_1

Table S5 continued..

	NEK2 phosphorylation site-NEK Serine/Threonine protein kinase family	MOD_NEK2_1
	Ser/Thr residue phosphorylated by the Plk1 kinase	MOD_Plk_1
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
Protein Sequence- HGDTVQNQLVVQGVELPSY LPLYPPAMDWIFQ	Ca2+- and calmodulin-regulated serine/threonine protein phosphatase known to affect cell biological function	DOC_PP2B_LxvP_1
	Di-Tryptophan targeting motif to the Delta-COP MHD domain	LIG_deltaCOP1_diTrp_1
	Canonical LIR motif that binds to Atg8 protein family to mediate processes involved in autophagy.	LIG_LIR_Gen_1
	LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	LIG_LIR_Nem_3
	SH2 domains binding motif.	LIG_SH2_STAT5
	SH3 LIGAND is involved in protein-protein interaction mediated by SH3 domains.	LIG_SH3_3
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
	Y-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex	TRG_ENDOCYTIC_2
Protein Sequence- SQPDPGELTALTAQHFIR	Anaphase Promoting Complex -E3 ubiquitin ligase is an important regulator of the cell cycle.	DEG_APCC_TPR_1
	IAP-binding motif (IBM)-Inhibitor of Apoptosis Proteins	LIG_BIR_II_1
	Post-translational modification of histones	LIG_WD40_WDR5_VD V_2
	Ser/Thr residue phosphorylated by the Plk1 kinase	MOD_Plk_1
Protein Sequence- GIGDPLVSVYAR	USP7 binding motif- deubiquitinating enzyme that cleaves ubiquitin moieties from its substrates	DOC_USP7_MATH_1
	Post-translational modification of histones	LIG_WD40_WDR5_VD V_2
	NEK2 phosphorylation site-NEK Serine/Threonine protein kinase family	MOD_NEK2_1
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
Protein Sequence- **IYTCVLHLLSAMSQETYLY HIDK**	ubiquitin-dependent proteasomal degradation	DEG_Nend_Nbox_1
	Phosphothreonine motif	LIG_FHA_1
	SH2 domains binding motif.	LIG_SH2_SRC

Table S5 continued..

	SH2 domains binding motif.	LIG_SH2_STAT5
	PIKK phosphorylation site	MOD_PIKK_1
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
	CK1 Phosphorylation site	MOD_CK1_1
	Glycosaminoglycan attachment site	MOD_GlcNHglycan
Protein Sequence-ACVAYCFITIPSLAGIFTR	IAP-binding motif (IBM)-Inhibitor of Apoptosis Proteins	LIG_BIR_II_1
	SH2 domains binding motif.	LIG_SH2_STAT5
	Post-translational modification of histones	LIG_WD40_WDR5_VDV_2
	Immunoreceptor tyrosine-based motif-that are critical for the activation and termination of signal transduction pathways.	LIG_TYR_ITIM
	Tyrosine-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex	TRG_ENDOCYTIC_2
Protein Sequence-DFEQQLSFYVESR	SH2 domains binding motif.	LIG_SH2_SRC
	SH2 domains binding motif.	LIG_SH2_STAT5
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
	N-terminal motif that initiates protein degradation by binding to the UBR-box of N-recognins	DEG_Nend_UBRbox_2
Protein Sequence-GSVEPQVLLEPYALIIPGENYIGINVK	Ca2+- and calmodulin-regulated serine/threonine protein phosphatase known to affect cell biological function	DOC_PP2B_LxvP_1
	Canonical LIR motif that binds to Atg8 protein family to mediate processes involved in autophagy.	LIG_LIR_Gen_1
	LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	LIG_LIR_Nem_3
	SH2 domains binding motif.	LIG_SH2_STAT5
	Tyrosine-based sorting signal responsible for the interaction with mu subunit of AP (Adaptor Protein) complex	TRG_ENDOCYTIC_2
	Sorting and internalisation signal found in the cytoplasmic juxta-membrane region of type I transmembrane proteins	TRG_LysEnd_APsAcLL_1
	SH2 ligand domains recognize small motifs containing a phosphorylated Tyrosine residue	LIG_SH2_PTP2
Protein Sequence-LAEFEDELDHTVDSLTWNLTPK	N-terminal protein motifs	DEG_Nend_UBRbox_3

Table S5 continued..

	Phosphothreonine motif	LIG_FHA_1
	WW Domains are small but abundant domains found in diverse regulatory situations- Phyphorylation	DOC_WW_Pin1_4
	Cyclin-dependent kinases (CDK) Phosphorylation Site	MOD_CDK_SPK_2
	GSK3 phosphorylation site	MOD_GSK3_1
	N-glycosylation site	MOD_N-GLC_1
	Polo-like kinase phosphosites	MOD_Plk_1
Protein Sequence-TNECQGTMWAPTSPPAGSSSPSQPTWK	SPOP SBC docking motif-Ubiquitination	DEG_SPOP_SBC_1
	USP7 binding motif- deubiquitinating enzyme that cleaves ubiquitin moieties from its substrates	DOC_USP7_MATH_1
	WW Domains are small but abundant domains found in diverse regulatory situations- Phyphorylation	DOC_WW_Pin1_4
	SH3 LIGAND is involved in protein-protein interaction mediated by SH3 domains	LIG_SH3_3
	CK1 Phosphorylation site	MOD_CK1_1
	GSK3 phosphorylation site	MOD_GSK3_1
	PIKK phosphorylation site	MOD_PIKK_1
	Ser/Thr residue phosphorylated by Plk4	MOD_Plk_4
	MAPK Phosphorylation Site	MOD_ProDKin_1

ACKNOWLEDGEMENT

We thank Vasant Kumar, JRF at RV College of Engineering for helping to produce representation graphs using R programming.

CONFLICT OF INTEREST STATEMENT

The authors have no potential conflict of interest to declare.

REFERENCES

1. Napier, J. K., Scheerer, M. and Mishra, S. 2014, World J. Gastroinstest. Oncol., 6, 112.
2. Ferri, F. F. 2013, Ferri's clinical advisor. Elsevier, Philadelphia, 389.
3. Pennathur, A., Gibson, M. K., Jobe, B. A. and Luketich, J. D. 2013, Lancet, 381, 9864..
4. Yan, W., Wistuba, II., Emmert-Buck, M. R. and Erickson, H. S. 2011, Am. J .Cancer. Res., 3, 275.
5. Pakzad, R., Mohammadian-Hafshejani, A., Khosravi, B., Soltani, S., Pakzad, I., Mohammadian, M., Salehiniya, H. and Momenimovahed, Z. 2016, Ann. Transl. Med., 2, 29.
6. Pera, M. and Pera, M. 2001, Surg. Oncol., 10, 81.
7. Nesvizhskii, A. I. 2014, Nat. Methods, 11, 1114.
8. Ruggles, K. V., Krug, K., Wang, X., Clauser, K. R., Wang, J., Payne, S. H., Fenyö, D., Zhang, B. and Mani, D. R. 2017, Mol. Cell. Proteomics., 16, 959
9. Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D., Patil, A. H.,

Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. and Pandey, A. 2014, Nature, 7502, 575.

10. Aebersold, R. and Mann, M. 2003, Nature, 422, 198.
11. Walther, T. C. and Mann, M. 2010, J. Cell. Biol., 190, 491.
12. Kelkar, D. S., Kumar, D., Kumar, P., Balakrishnan, L., Muthusamy, B., Yadav, A. K., Shrivastava, P., Marimuthu, A., Anand, S., Sundaram, H., Kingsbury, R., Harsha, H. C., Nair, B., Prasad, T. S., Chauhan, D. S., Katoch, K., Katoch, V. M., Kumar, P., Chaerkady, R., Ramachandran, S., Dash, D. and Pandey, A. 2011, Mol. Cell. Proteomics., 10, 111.
13. Khatun, J., Yu, Y., Wrobel, J. A., Risk, B. A., Gunawardena, H. P., Secrest, A., Spitzer, W. J., Xie, L., Wang, L., Chen, X. and Giddings, M. C. 2013, BMC Genomics, 28, 141.
14. Fermin, D., Allen, B. B., Blackwell, T. W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G. S. and States, D. J. 2006, Genome. Biol., 7, 35.
15. Ning, K., Fermin, D. and Nesvizhskii, A. I. 2010, Proteomics, 14, 2712.
16. Yates, J. R. 2013, J. Am. Chem., 135, 1629.
17. Helmy, M., Sugiyama, N., Tomita, M. and Ishihama, Y. 2012, Genes. Cells., 7, 633.
18. Khatun, J., Yu, Y., Wrobel, J. A., Risk, B. A., Gunawardena, H. P., Secrest, A., Spitzer, W. J., Xie, L., Wang, L., Chen, X. and Giddings, M. C. 2013, BMC Genomics, 14, 141.
19. Sheynkman, G. M., Shortreed, M. R., Cesnik, A. J. and Smith, L. M. 2016, Annu. Rev. Anal. Chem. (Palo Alto Calif)., 12, 521.
20. Jaffe, J. D., Berg, H. C. and Chruch, G. M. 2004, Proteomics, 4, 59.
21. Craig, R. and Beavis, R. C. 2004, Bioinformatics, 20, 1466.
22. Sasaki, H., Hikosaka, Y., Kawano, O., Moriyama, S., Yano, M. and Fujii, Y. 2010, Oncol. Lett., 2, 283.
23. Ye, X., Feng, G., Jiao, N., Pu, C., Zhao, G. and Sun, G. 2014, Dis. Markers., 80, 4023.
24. Phillips-Krawczak, C. A., Singla, A., Starokadomskyy, P., Deng, Z., Osborne, D. G., Li, H., Dick, C. J., Gomez, T. S., Koenecke, M., Zhang, J. S., Dai, H., Sifuentes-Dominguez, L. F., Geng, L. N., Kaufmann, S. H., Hein, M. Y., Wallis, M., McGaughran, J., Gecz, J., Sluis, B. V., Billadeau, D. D. and Burstein, E. 2015, Mol. Bio. Cell., 26, 93.
25. Sever, R. and Brugge, J. S. 2015, Cold. Spring. Harb. Perspect. Med., 5, a006098.
26. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. and Lo Muzio, L. 2017, Int. J. Mol. Med., 40, 271.
27. Chen, B., Lam, T. C., Liu, L. and To, C. 2017, Mol. Med. Rep., 15, 3923.
28. Zheng, S., Zhang, C., Qin, X., Gen, Y., Liu, T., Sheyhidin, I., and Lu, X. 2011, Mol. Bio. Rep., 39, 5315.
29. Brognard, J. and Hunter, T. 2011, Curr. Opin. Genet. Dev., 21, 4.