Heliyon 6 (2020) e04813

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Proteomic variations of esophageal squamous cell carcinoma revealed by combining RNA-seq proteogenomics and G-PTM search strategy

Pooja Ramesh, Vidhyavathy Nagarajan, Vartika Khanchandani, Vasanth Kumar Desai, Vidya Niranjan *

Department of Biotechnology, RV College of Engineering, Bangalore, Karnataka, India

ARTICLE INFO

Keywords:
Bioinformatics
Cancer research
Genetics
Oncology
Proteogenomics
Esophageal squamous cell carcinoma
PTM
RNA-seq proteogenomics
Bottom-up proteomics

ABSTRACT

Background: Cancer that arises from epithelial cells of the esophagus is called esophagus squamous cell carcinoma (ESCC) and is mostly observed in developing nations. Evaluation of cancer genomes and its regulation into proteins plays a predominant role in understanding the cancer progressions. Mass-spectrometry-based proteomics is a consequential tool to estimate proteomic variation and posttranslational modifications (PTMs) from standard protein databases. Post-translational modifications play a crucial role in protein folding and PTMs can be accounted for as a biological signal to interpret the structural changes and transition order of proteins. Functional validation of cancer-related mutations can explain the effects of mutations on genes and the identification of Oncogenes and tumor suppressor genes. Therefore, we present a study on protein variations to interpret the structural changes and transition order of proteins in ESCC carcinogenesis.

Methodology: We are using a bottom-up proteomics approach with Galaxy-P framework and RNA sequence data analysis to generate the sample-specific databases containing details of RNA splicing and variant peptides. Once the database generated with information on variable modification, only the curated PTMs at specific positions are considered to perform spectral matching. Proteogenomics mapping was performed to identify protein variations in ESCC.

Results: RNA-sequence proteogenomics with G-PTM (Global Post-Translational Modification) searching strategy has revealed proteomic events including several peptides that contain single amino acid variations, novel splice junction peptides and posttranslationally modified peptides. Proteogenomic mapping exhibited the splice junction peptides mapped predominantly for Malic enzyme exon type (ME-3) and MCM7 protein-coding genes that promote cancer progression, found to be exhibited in ESCC samples. Approximately 25 ± types of PTM modifications were recorded, and Protein Phosphorylation was largely noted.

Conclusion: ESCC cancer prognosis at the molecular level enables a better understanding of cancer carcinogenesis and protein modifications can be used as potential biomarkers.

1. Introduction

The incidences of esophageal squamous cell carcinoma (ESCC) are recorded mostly in developing nations like Asian countries. The rising incidences of ESCC are poorly understood, but the contributory factors may be listed as lifestyle behaviors, alcohol consumption, and tobacco smoking habits [1]. Advancements in technology like Next-generation sequencing of DNA and RNA helps in profiling the somatic alterations caused by mutations, and rapid diagnosis can vouch to personalized cancer treatment [2, 3]. Post-translational modifications (PTMs) in peptide sequences have a significant impact on structure, function, and localization of proteins in a cell [4]. It has been well acknowledged that the PTMs are critical molecular events that alter the protein conformations and modulating the stability of proteins [5]. More than 300 different types of PTMs have been observed till date, few among them are methylation, acetylation, phosphorylation, ubiquitination, glycosylation, and sumoylation.

Classic illustrations like regulation of chromatin folding and gene expressions [6] by histone PTMs and the Phosphorylation of the tumor suppressor protein p53 in response to DNA damage. These PTMs are responsible for modulation of DNA binding properties and proteasomal degradation signaling [7]. Another example of PTM regulations of

* Corresponding author.
E-mail address: vidya.n@rvce.edu.in (V. Niranjan).

https://doi.org/10.1016/j.heliyon.2020.e04813
Received 24 May 2020; Received in revised form 10 July 2020; Accepted 25 August 2020

proteins can be understood in epigenetic regulator EZH2 (Enhancer of zeste homolog 2), which are known to regulate the wide variety of human cancers in both transcriptional and post-transcriptional levels [8]. The promoter E2Fs binds to EZH2 and regulates a large number of micro-RNAs at PTM level in the development of cancer [9]. Similarly, in sickle cell anemia, caused by amino acid substitutions in the beta-hemoglobin gene, the amino acid can efficaciously influence the structure of proteins, point mutations such as the change of glutamic acid to valine mutation in hemoglobin causing the disorder [10].

The remarkable proteomic variants with important biological consequences have remained unexplored in standard bottom-up proteomics analysis. We combined two methods to uncover the variants, first the generation of the sample-specific database using RNA-seq data analysis of ESCC cancer samples. And secondly, the global posttranslational modification search strategy to understand the unknown information of sequence variations and PTMs [10, 11]. In a classic bottom-up proteomics method, the identification and characterization of proteins are based on amino acid sequences and PTMs by proteolytic digestion of proteins. First, the proteins to be characterized are purified using gel electrophoresis method and separated by liquid chromatography method coupled to mass spectrometry, this method is known as shotgun proteomics [12]. MS/MS-based proteomics typically involves enzymatic digestion, fragmentation, and separation in a spectrometer. Analyzing the output data obtained from the spectrometer is performed computationally and is identified based on masses of peptides [13].

The database search strategy determines the measurements of peptides that read in mass spectra of shotgun proteomics search by conducting an Insilco tryptic digestion of amino acid sequences. Later it matches the experimental spectra to theoretical spectra and finds the small range of measured peptide mass [14]. The query spectrum is compared with spectrum libraries to assess the possible peptides using scoring algorithms. The peptide spectral matching is a scoring function that assigns numerical values to peptide-spectrum (P–S) pair, and the highest-scoring match is called as peptide spectral match (PSM) [15]. Hence the database search strategy is restricted to the search-database and there is a requirement for special approaches to detect amino acid variant peptides and PTM [16].

RNA-Seq proteomics is a powerful analysis method to build the sample-specific peptide database that is invisible to the standard mass spectrometry proteome database [17]. Transcriptomics, along with proteomics, improves the sample-specific proteome characterization and peptide variants in the database. This process is called Proteome informed by Transcriptomics (PIT), where the sample-specific database generated for peptides and protein identification informed by transcriptomics [18]. Using the Galaxy platform to perform the PIT analysis is more established and provides the workflow for performing tasks such as standard search against reference proteome, protein identification with a reference genome and genome annotations via simple web interface. RNA seq data includes peptides with a single amino acid variation called SAV peptides [19] and peptides with novel splice junction peptides as NSJ peptides [20] into the customized protein sequence database. Construction of sample-specific proteome databases for SAV peptides and NSJ peptides involves a combination of two workflows. Directly employing the RNA seq data provides the flexibility of working with samples that have not yet been characterized in a large consortium like TCGA and CanvarPro database [21].

In the present study, we employ publicly available MS/MS from the Pride database for ESCC and Transcriptomics data for 10 ESCC samples from NCBI-SRA database to evaluate various changes using RNA-Seq data for constructing proteomic sample-specific databases. Considering the error rate things that should be taken care of are False discovery rates (FDR) while filtering the sequence variants for SAV and NSJ peptides. In a classic PTM-identification and enrichment procedures like combinatorial explosion of all possible potential molecular states, the search leads to the discovery of new PTM sites and also contains high false discovery rates due to large search space [22]. Currently, we are using the site-specific PTM annotated database and open search models to reduce the high FDR [23]. Mapping of PTMs to the database is an issue in current proteomics as the PTM types have to be explicitly mentioned by the users, but in this case, much novel information is lost. Hence to solve this problem we perform the open search method where the peptide mass tolerance is expanded precisely for example +79.97Da for Phosphorylation. This approach does not require the user to specify the PTM type, the search will be able to detect all possible spectra from peptides that are unparticipated PTM and identify new PTMs of various types 13). The second approach is site-specific PTM annotated database which involves modifications of annotated PTMs at specific residues for each sample. This approach decreases the combinatorial explosion of peptides and thus high FDR and yields similar results as the open model search method [24].

In this study, we have employed a Global-Post translational Modification (G-PTM) search strategy using the Metamorepheus tool. G-PTM tool expands the scope of peptide identification by including the site-specific PTM search. The PTM information obtained from Uniprot for further validation. In the current work, we searched RNA-Seq proteogenomic databases with the G-PTM strategy to uncover variant protein information.

2. Experimental details

2.1. RNA sequencing data

The RNA-Sequence data for 10 ESCC samples (Table 1) were obtained from publicly available Sequence Read Archive (NCBI-SRA) (https://www.ncbi.nlm.nih.gov/sra) (Table-S1) and used for the generation of the sample-specific databases. The experiments conducted for sequencing of these reads involved three main steps: isolation of RNA, preparation of cDNA libraries and sequencing of the reads using Illumina HiSeq 2500. Initially, RNA was isolated and analyzed on gels for degradation. Then mRNA was purified using poly T oligos attached to magnetic beads. The purification step was followed by fragmentation, after which synthesis of the first and second cDNA strands took place. The final cDNA library was prepared by ligating to adaptors and further enrichment using PCR. The prepared libraries were quantified for purity, analyzed for insert size and concentration. Finally, the inserts were sequenced on Illumina HiSeq 2500 and the raw reads are generated. The read count (number of molecules sequenced) and read length (length of each sequence) varies across the RNA-sequence data used in this study (Table 1). The performance of RNA-sequences also relies on important parameters: sequence depth or read depth which describes the number of times the particular nucleotide is being read in an experiment. The optimal sequencing depth for the RNA-sequence employed in the present work has 400 million reads to 650 million reads, with a read length of 100bp (base pair) long. The sequence coverage of the RNA-seq ranges from 13x-21x coverage. The RNA-sequences are downloaded from NCBI Sequence Read Archive under accession number SRP064894 [25].

2.2. Mass spectrometry data

The MS/MS spectra data from the analysis of protein expression in esophageal squamous cell carcinoma tissue sections was obtained from the pride archive database. Proteins were extracted using pressure cycling technology (barocycler) by orbitrap fusion mass spectrometer. The clarified proteins were then digested and fractionalized. Later LC/MS analysis was done, followed by data analysis. The MS data was pre-processed using the Mascot and Sequest search algorithm, using Human RefSeq75 protein databases as a reference. The data set consisted of 12 raw files included in the study with 6 SCX.raw and 6 RPS.raw files (Table S-2) [26].

Table-1. High through-put RNA-seq data for ESCC samples obtained from NCBI-SRA repository. Repository information is listed in Table S-1. Abbreviations: poly(A)+: Polyadenylation tail to mRNA.

Cell Line	SRR ID	RNA-Library Preparation	Sequence Type	Sample Type	Read length (bp)	Read count	Size of Sample (GB)
esophageal squamous cell carcinoma	SRR2678176	Poly(A)+	Paired	Adult Male	100	$6.5x10^9$	3.8Gb
esophageal squamous cell carcinoma	SRR2678178	Poly(A)+	Paired	Adult Female	100	4.2×10^9	2.5Gb
esophageal squamous cell carcinoma	SRR2678180	Poly(A)+	Paired	Adult Male	100	4.8×10^9	2.9Gb
esophageal squamous cell carcinoma	SRR2678170	Poly(A)+	Paired	Adult Male	100	3.9×10^9	2.5Gb
esophageal squamous cell carcinoma	SRR2678360	Poly(A)+	Paired	Adult Male	100	4.1×10^9	2.5Gb
esophageal squamous cell carcinoma	SRR2678324	Poly(A)+	Paired	Adult Male	100	5.5×10^9	3.3Gb
esophageal squamous cell carcinoma	SRR2678264	Poly(A)+	Paired	Adult Male	100	5×10^9	3Gb
esophageal squamous cell carcinoma	SRR2678182	Poly(A)+	Paired	Adult Female	100	5.4×10^9	3.2Gb
esophageal squamous cell carcinoma	SRR2678172	Poly(A)+	Paired	Adult Male	100	5.5×10^9	3.3Gb
esophageal squamous cell carcinoma	SRR2678162	Poly(A)+	Paired	Adult Female	100	4.4×10^9	2.6Gb

2.3. Sample specific database construction for proteomic variations

Proteomics variations such as SAV peptides (single amino acid variations), NSJ peptides (Novel splice junctions) and, PTM (Post-translational Modifications) are widely uncovered by utilization of high-throughput RNA-seq data [27]. Galaxy server is a multi-omics interface that allows users to utilize tools and programs to analyze genomics and proteomics data (Figure 1) [28].

RNA-sequence data obtained from the NCBI-SRA database are checked for quality of reads produced by the sequencer and pre-processing of data is done by the removal of adapter sequences using a Trimmomatic tool. Further the RNA-seq reads are aligned to the reference genome GRCh 38 (version 87) based on identification of poly (A) tail and locating the intron-exon regions in the sequence, further helps in determination of splice patterns and variations [29]. The Galaxy work processes are used to create SAV and NSJ peptide FASTA database as described by Sheynkman et al. Each database comprises of three segments, the single amino acid variation peptides, the novel splice junction peptides, and Uniprot reference proteome database along with site-specific PTM annotations for Homo sapiens. The SAV and NSJ peptides are obtained in FASTA format, whereas the UniProt reference genome with PTM annotations is recorded in UniProt XML language.

2.4. Single-amino acid peptide database

Exploring amino acid variants such as single nucleotide polymorphism (SNP), multiple nucleotide polymorphisms (MNP), insertions and, deletions that lead to nucleic acid variations can be detected using a spectral search of SAV database. Galaxy server is used for the alignment of ESCC RNA-seq reads to reference genome GRCh38 version 86 using the HISTAT2 tool for the construction of the SAV database. HISAT2 is a fast sensitive alignment tool used for mapping RNA-seq reads to reference genome. HISAT2 has hierarchical indexing of transcripts and aligning will be done by graph FM index (small local indexes), combining several alignment strategies for rapid and precise alignment of sequence reads. HISTAT2 alignments for the paired-end RNA sequence with reference genome and ensemble gene model (GRCh38) that contains splicing exon patterns within genes discussed in Table S-3. The aligned RNA-seq reads are obtained in BAM files. The Bam files are converted to Variant Calling Files (VCF) using the Bcftools mpile up tool on the galaxy server. Bcftools manipulate bam files into a vcf file format which are later used to extract out the missense SNV and nucleotide variants using a program called SnpEff. The reference genome GRCh38.86 is obtained from SnpEff prebuilt database on the galaxy server. The nucleic acid variants and missense SNV recorded in VCF files are further used in customized database construction using a software Sample specific Database generator (https://sourceforge.net/projects/samplespecificdbgenerator/). The SAV peptides with complete tryptic fragments along with the allowance of two missed cleavage are counted as variant amino acids in standard Peptide spectral match (PSM) [19].

2.5. Novel splice junction database

NSJ workflow is executed using the Tophat alignment tool on the galaxy server. Tophat is a fast splice junction read mapper to reference genome to identify splice junctions between exon regions. The NSJ workflow has two sets of Tophat alignment for RNA-seq reads with reference genome GRCh86 version 86. The first alignment is to identify known splice junctions with the addition of gene annotations and excluding the 'coverage search' in Tophat tool alignment advance settings (Table S-4). The second alignment search is to identify novel with known splice junctions in annotated genes that include the 'coverage search' are recorded in BED (Browser Extensible files) files. The BED files contain information on splice junction flanking regions between two exon and its genomic coordinates. Now, the known BED files are subtracted from the known and novel BED files to reveal the Novel junctions. The Novel Bed files are later translated using a tool called Translate BED files on the galaxy server to translate each entry to nucleotides from exon into NSJ peptides. The NSJ peptides recorded in BED files are entered into a customized sample-specific database. The NSJ peptide BED files contain tryptic peptides to reduce the false discovery rates during PSM identifications. The UniProt XML file of human reference proteome was downloaded from Ensemble database and protein accession number found on Table S-5 [19].

2.6. Spectral matching using morpheus

Spectral matching and database searching using the G-PTM strategy was achieved by the Metamorephus software tool. Metamorephus is bottoms-up proteomics database search software that discovers PTM (post-translational modifications) along with proteomic variations using a robust search algorithm. Global Post-Translational Modification (G-PTM) strategy utilizes PTM site-specific annotation information obtained from Uniprot reference human proteome for database generation [29]. G-PTM strategy provides a more capable search method by narrowing down the mass tolerance coverage and thus including a wide variety of peptide modifications in the database. A Metamorephus tool considers all the possible peptide isoforms combinations for given site-specific PTM annotation for each of peptide entry in the database [30]. For each peptide, the maximum allowed variable modification isoforms were set to the default value of 1024. The following settings were used in searches:

In-Silico digestion parameters:
Protease = Trypsin
Maximum Missed Cleavages = 2
Precursor Mass Tolerance = 10 ppm
Product Mass Tolerance = 20ppm
Top N peptides per precursor = 300
Maximum Modification Isoforms = 1024
Fragment Ion search parameters:
Product Mass tolerance = 20ppm

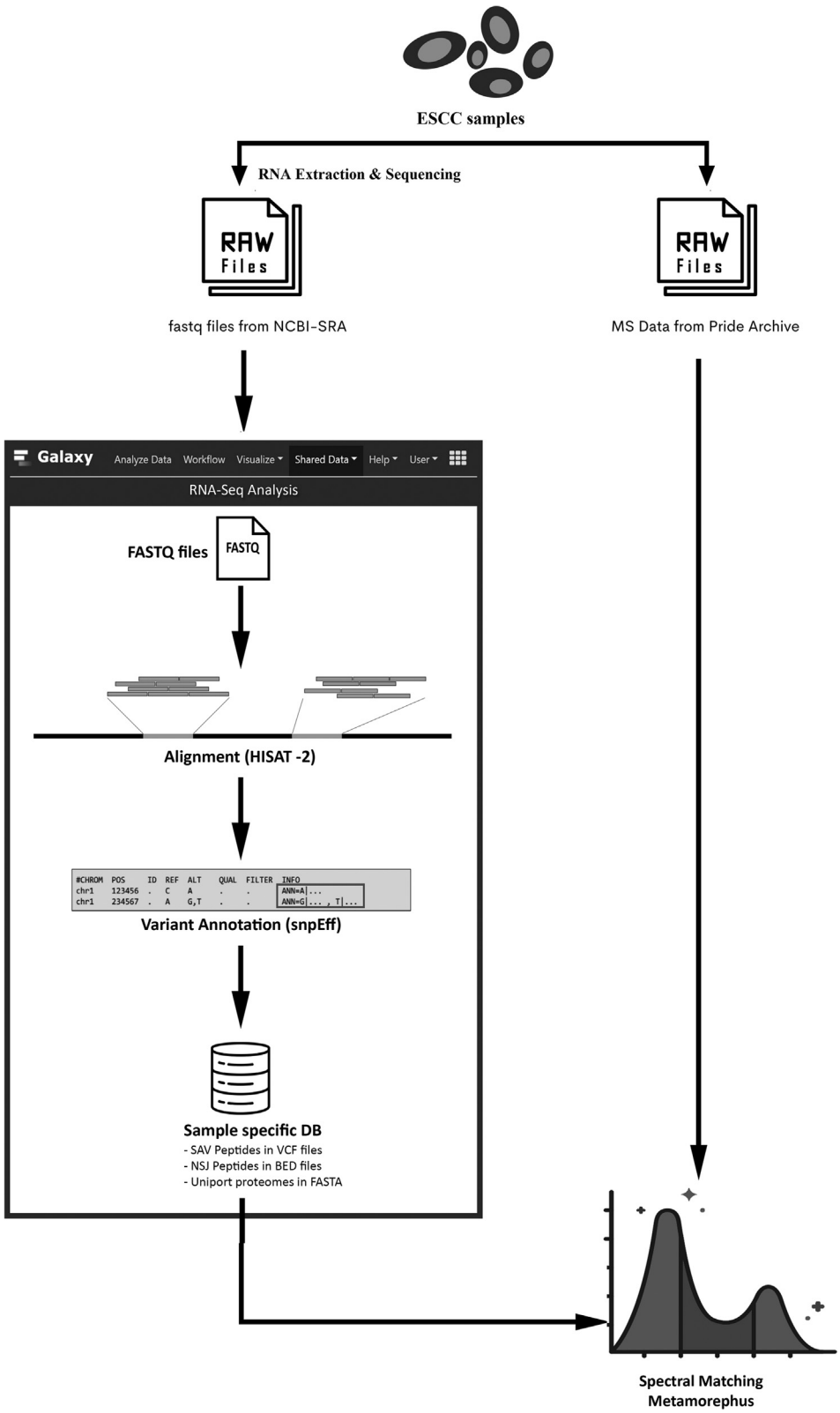

Figure 1. Overview of the RNA-seq Proteogenomics workflow with G-PTM search strategy for identification of proteomic variants and PTM peptides. RNA-seq is analyzed using Galaxy package for sequence variants and sample-specific database is generated containing SAV peptides, NSJ peptides and uniprot protein sequence with site-specific PTM annotations. Spectral matching is performed using the Metamorephus tool using G-PTM search strategy.

Maximum threads = 3
Modifications:
Fixed Modification = common fixed
Variable Modification = common variable
Maximum False Discovery Rate = 1%
Spectral Matching was performed on HPC Intel Xeon processor with 80GB RAM and 1TB storage

2.7. Global false discovery rates

Millions of tandem mass spectra are generated in Mass spectrometry-based proteomics study. The peptides fragments are read under spectrometry and produce spectra.RAW files. These spectra are later used to search against the sample-specific database using the Metamorephus software tool. The peptide matching to spectra is known as a peptide-spectral match (PSMs). The scores are produced for each matching and Meatmorephus tool uses a simple scoring algorithm for producing PSMs and false discovery rates (FDRs).

3. Results

3.1. Impact of RNA sequencing data

The high throughput data consideration in the present study plays a vital role in the identification of nucleotide variants. The impact of RNA-seq data reads, and sequence type is directly proportional to the quality of sequence variants in customized sample-specific databases for each of the ESCC samples. RNA-seq data with sequencing depth ranging from $3.0*10^9$ to $6.5*10^9$, a total of 604 SAV peptides has been identified at 1% FDR rates (Table S-6) and 37 NSJ peptides have been recorded at 1% FDR (Table S-7).Generating more sequence reads and using fast alignment tools like Tophat and HISTAT2 helps in the detection of nucleotide sequence variants at 1% false discovery rates. In this study, we have employed larger ESCC RNA-seq data sets with an average file size of 10GB with 5.8 G reads. Overall, the deep sequencing of RNA reads has been beneficial for the detection of NSJ and SAV peptides.

3.2. Impact of RNA sequence type

We have selected paired-end (PE) data over single-end data for analysis as paired-end data provides adequate advantages over single-end data for variant peptide identification [19, 37]. For NSJ peptides, the splicing patterns are better understood in paired-end data. The splice junctions, aligned between the exon regions and genomic coordinated are translated into NSJ peptides leads to better identification and matching with spectral data with low FDR rates (>1%). It is highly recommended to apply paired-end data over single-end data for proteogenomics analysis.

3.3. Database search results

The search results for each of the ten ESCC sample files are elaborately listed containing SAV peptides along with codon change and chromosome locations of SAV at a 1% FDR threshold (Table S-6). It is also observed that most of the SAV occurrences are found in chromosome 10(Table S-6, Table S-6A). NSJ peptides information is discussed in Table-S7 (Table S-7A & TableS-7B), Table S-7A summarizes the exon type and further categorized into the protein-coding gene or pseudogenes. Table S–7B lists the exon types concerning chromosomes and ME3 (Malic Enzyme-3) exon type is observed to be predominant. A study [31] suggests that the Malic Enzymes are observed to be associated with the progression of cancer in human oral squamous cell carcinoma and hence suggesting proof of targeting the ME3 splice exon. Splice exon peptides also show the presence of novel splice junction MCM7 (Minichromosome Maintenance Complex Component 7) which is a protein-coding exon type. MCM7 exon plays a direct role in amplifying DNA synthesis and increase the rate of cell invasion in cancer cells [32]. The unique peptides for SAV and NSJ were separately counted and for peptide residues that are not cleaved by trypsin, each sample was allowed for two missed tryptic cleavages for identifying arginine and lysine residues during database search. Table S-8 contains a list of unique PTM peptides at 1% FDR rates.

In our study, we have identified a novel exon type associated to ESCC has been listed in Table S-7A. On mapping the genomic coordinates to human reference genome hg38 using USCS genome browser, it is observed that the novel exon belongs to PGGHG (protein-glucosylgalactosylhydroxylysine glycosidase) gene with ensemble ID: ENSG00000142102 is protein-coding gene found on 11th chromosome. The function of this gene Catalyzes involves in the hydrolysis of glucose from the disaccharide unit linked to hydroxylysine residues of collagen and collagen-like proteins. This gene is also studied to have a site-directed mutagenesis [38].

3.4. PTM peptides identification

A total of 161 unique PTM peptides were identified (Table S-8), and the search results comprise a diverse range of PTM peptides illustrated in (Figure 2). Each cell line has identified different types of PTM peptides and has been illustrated in graph and Table S-9. All unique types of data including asymmetric and symmetric types of PTM peptides have also been included in Uniprot PTM database annotation and have been listed. Threshold False discovery rates were in the range of 1.0%–1.6% FDR for globally identifying the peptides.

3.5. Role of PTMs in ESCC

The role of post-translational modifications in Cancer could range from modulating signaling cascades, cell division, DNA repair, impacting protein expression, localization, or protein-protein interactions. Kinase driven Phosphorylation induces genetic level modifications leading to a proliferation of cell growth in cancer cells [33]. The alterations in signal transduction pathways due to Phosphorylation has observed to produce a cascade of reaction pathways including MAP kinase, tyrosine kinase and other kinase-dependent pathways which plays a major role in cancer cell growth and progression [34]. A well-known example to understand the implications of PTM's in cancer is p53 indicating the post-translational modification influences the expression of its target genes, further affecting different pathways in cancer. Phosphorylation of threonine and/or serine residues, acetylation of lysine residues are the most frequent modifications observed in the p53 gene. These two modifications are known to activate and stabilize the p53 gene under the influence of cellular stress [35]. It has also been suggested that phosphorylation of the mutant p53 gene contributes to cancer progression. In ESCC, it is observed that the p38 MAP kinase pathway agitates cell proliferation and cell growth [36].

Lysine-acetylation is observed to regulate gene transcription by targeting histone and various transcription factors (TFs) in a cell. A study has reported the influence of lysine acetylation affecting the modification protein activity and their implications in cellular functions of esophageal SCC (squamous cell carcinoma) [39, 40]. However, currently, there are no studies to have looked at over-all post-translational modifications in ESCC, and hence we using proteogenomic mapping analysis to exhibit the amount of phosphorylation and acetylation expressed in a cancer cell.

4. Discussion

Initial proteogenomics approaches hampered by the technical challenges of implying high-throughput data and the sensitivity of proteomics data. Advancement in Mass spectrometry data has improved technically over past years by an increase in sensitivity and accuracy of data. In this study, we address applying a combined approach of RNA-seq Proteogenomics technology along with the G-PTM strategy for the

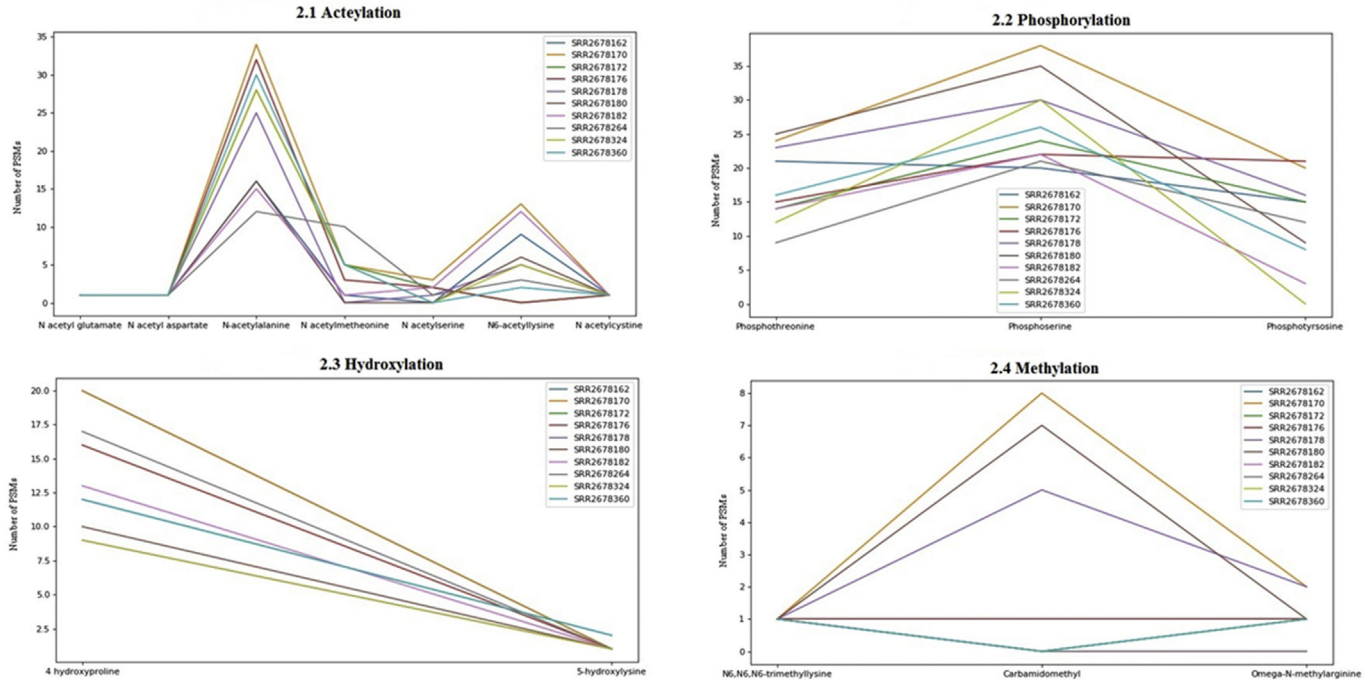

Figure 2. Diverse types of PTMs obtained by G-PTM search strategy for 10 ESCC samples. Figure 2.1: Various types of Acetylation and most numbers of PSMs accounted for N-acetylalanine PTM. Figure 2.2 In Phosphorylation, Phosphoserine PTM has the maximum number of PSMs. Figure 2.3: 4-Hydroxyproline shows most of PSMs in Hydroxylation. Figure 2.4: illustrates most of PTMs are recorded for Carbamidomethyl PTM.

annotation of protein variants. We employed a single pass bottom-up MS/MS proteomic search approach to identify diverse range of proteomic variants including Single-amino acid variants, Novel splice junction variants and different types of PTM peptides. The depth of RNA-seq data integrated with Proteomics data provides greater information on proteomic variants like PTM events and therefore potentially informing investigation on biochemical regulations and protein expression in cell.

Considering RNA-seq paired-end data and analysis using software tools with crucial parameters has achieved in a self-assured diverse range of protein variants. We have observed that implying the paired-end RNA sequence data has improved the detection of variant peptides. The quality of the sequence data is inversely proportional to the identification of protein peptides and the depth coverage ratio of the protein variants with 1.0%–1.5% FDR for SAV peptides and 5.0%–10.5% for NSJ peptides.

The identification of protein variants in 10 RNA-seq ESCC samples is reported in supplementary data. The studies identified several tens of NSJ peptides, several hundreds of SAV peptides, and thousands of PTM peptides with various types of modifications for 10 different samples. Approximately 25 ± types of modifications were recorded for each ESCC samples using G-PTM strategy. G-PTM search strategy along with RNA-seq proteogenomics techniques was applied in identification of amino acid variants including SAV peptides and NSJ peptides with lower FDR rates. Cumulative studies on proteomic variants decipher better understanding of ESCC with cell regulations and cancer progression.

Declarations

Author contribution statement

Pooja Ramesh: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Vidhyavathy Nagarajan, Vartika Khanchandani: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Vasanth Kumar Desai: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Vidya Niranjan: Conceived and designed the experiments; Performed the experiments.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e04813.

Acknowledgements

We would like to thank Mr.Akshay Uttarkar, Senior Research Fellow at the Centre of Excellence Computational Genomics-RVCE, for a critical review of the manuscript.

References

[1] K. Higuchi, W. Koizumi, S. Tanabe, T. Sasaki, C. Katada, M. Azuma, K. Nakatani, K. Ishido, A. Naruke, T. Ryu, Current management of esophageal squamous-cell carcinoma in Japan and other countries, Gastroint. Cancer Res. GCR 3 (4) (2009) 153–161.

[2] C. Meldrum, M.A. Doyle, R.W. Tothill, Next-generation sequencing for cancer diagnostics: a practical perspective. The Clinical biochemist, Review 32 (4) (2011) 177–195.

[3] M. Morash, H. Mitchell, H. Beltran, O. Elemento, J. Pathak, The role of next-generation sequencing in precision medicine: a review of outcomes in Oncology, J. Personalized Med. 8 (3) (2018) 30.

[4] K.W. Barber, J. Rinehart, The ABCs of PTMs, Nat. Chem. Biol. 14 (3) (2018) 188–192.

[5] X. Wang, Q. Liu, B. Zhang, Leveraging the complementary nature of RNA-Seq and shotgun proteomics data, Proteomics 14 (23-24) (2014) 2676–2687.

[6] K. Prakash, D. Fournier, Histone code and higher-order chromatin folding: a hypothesis, Genom. Comp. Biol. 3 (2) (2017) e41.

[7] M. Grunstein, Histone acetylation in chromatin structure and transcription, Nature 389 (6649) (1997) 349–352.

[8] C.J. Chang, M.C. Hung, The role of EZH2 in tumour progression, Br. J. Cancer 106 (2) (2012) 243–247.

[9] L. Lu, R.J. Millikin, S.K. Solntsev, Z. Rolfs, M. Scalf, M.R. Shortreed, L.M. Smith, Identification of MS-cleavable and noncleavable chemically cross-linked peptides with MetaMorpheus, J. Proteome Res. 17 (7) (2018) 2370–2376.

[10] T. Carlice-Dos-Reis, J. Viana, F.C. Moreira, G.L. Cardoso, J. Guerreiro, S. Santos, Â. Ribeiro-Dos-Santos, Investigation of mutations in the HBB gene using the 1,000 genomes database, PloS One 12 (4) (2017), e0174637.

[11] M.P. Washburn, D. Wolters, J.R. Yates, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, Nat. Biotechnol. 19 (3) (2001) 242–247.

[12] K.L. Huang, S. Li, P. Mertins, S. Cao, H.P. Gunawardena, K.V. Ruggles, D.R. Mani, K.R. Clauser, M. Tanioka, J. Usary, S.M. Kavuri, L. Xie, C. Yoon, J.W. Qiao, J. Wrobel, M.A. Wyczalkowski, P. Erdmann-Gilmore, J.E. Snider, J. Hoog, P. Singh, L. Ding, Corrigendum: proteogenomic integration reveals therapeutic targets in breast cancer xenografts, Nat. Commun. 8 (2017) 15479.

[13] J.K. Eng, A.L. Mccormack, J.R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectr. 5 (1994) 976–989.

[14] J.A. Alfaro, A. Ignatchenko, V. Ignatchenko, A. Sinha, P.C. Boutros, T. Kislinger, Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines, Genome Med. 9 (1) (2017) 62.

[15] A.M. Frank, A ranking-based scoring function for peptide-spectrum matches, J. Proteome Res. 8 (5) (2009) 2241–2252.

[16] Y.C. Wang, S.E. Peterson, J.F. Loring, Protein post-translational modifications and regulation of pluripotency in human stem cells, Cell Res. 24 (2) (2014) 143–160.

[17] K. Ning, A.I. Nesvizhskii, The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment, BMC Bioinf. 11 (Suppl 11) (2010) S14.

[18] S. Saha, D.A. Matthews, C. Bessant, High throughput discovery of protein variants using proteomics informed by transcriptomics, Nucleic Acids Res. 46 (10) (2018) 4893–4902.

[19] A.J. Cesnik, M.R. Shortreed, G.M. Sheynkman, B.L. Frey, L.M. Smith, Human proteomic variation revealed by combining RNA-seq proteogenomics and global post-translational modification (G-PTM) search strategy, J. Proteome Res. 15 (3) (2016) 800–808.

[20] S. Prabakaran, G. Lippens, H. Steen, J. Gunawardena, Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding, Wiley Interdiscip. Rev. Syst. Biol. Med. 4 (6) (2012) 565–583.

[21] M.R. Shortreed, C.D. Wenger, B.L. Frey, G.M. Sheynkman, M. Scalf, M.P. Keller, A.D. Attie, L.M. Smith, Global identification of protein post-translational modifications in a single-pass database search, J. Proteome Res. 14 (11) (2015) 4714–4720.

[22] D. Ye, Y. Fu, R.X. Sun, H.P. Wang, Z.F. Yuan, H. Chi, S.M. He, Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate, Bioinformatics (Oxford, England) 26 (12) (2010) i399–i406.

[23] E. Afgan, D. Baker, B. Batut, et al., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Res. 46 (W1) (2018) W537-W544.

[24] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions, Genome Biol. 14 (4) (2013) R36.

[25] C.Q. Li, G.W. Huang, Z.Y. Wu, Y.J. Xu, X.C. Li, Y.J. Xue, Y. Zhu, J.M. Zhao, M. Li, J. Zhang, J.Y. Wu, F. Lei, Q.Y. Wang, S. Li, C.P. Zheng, B. Ai, Z.D. Tang, C.C. Feng, L.D. Liao, S.H. Wang, L.Y. Xu, Integrative analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma, Oncogenesis 6 (2) (2017) e297.

[26] Vinuth N. Puttamallesh, Krishna Patel, Gowda. Harsha, Rapid processing of archival tissue samples for proteomic analysis using pressure-cycling technology, J. Protein Proteonomics 8 (2) (2017) 121–125.

[27] J.M. Proffitt, J. Glenn, A.J. Cesnik, A. Jadhav, M.R. Shortreed, L.M. Smith, K. Kavanagh, L.A. Cox, M. Olivier, Proteomics in non-human primates: utilizing RNA-Seq data to improve protein identification by mass spectrometry in vervet monkeys, BMC Genom. 18 (1) (2017) 877.

[28] Y. Han, S. Gao, K. Muegge, W. Zhang, B. Zhou, Advanced applications of RNA sequencing and challenges, Bioinf. Biol. Insights 9 (Suppl 1) (2015) 29–46.

[29] G.M. Sheynkman, M.R. Shortreed, B.L. Frey, M. Scalf, L.M. Smith, Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences, J. Proteome Res. 13 (1) (2014) 228–240.

[30] H. Lu, G. Li, C. Zhou, W. Jin, X. Qian, Z. Wang, H. Pan, H. Jin, X. Wang, Regulation and role of post-translational modifications of enhancer of zeste homologue 2 in cancer development, Am. J. Cancer Res. 6 (12) (2016) 2737–2754.

[31] C. Nakashima, K. Yamamoto, R. Fujiwara-Tani, Y. Luo, S. Matsushima, K. Fujii, H. Ohmori, T. Sasahira, T. Sasaki, Y. Kitadai, T. Kirita, H. Kuniyasu, Expression of cytosolic malic enzyme (ME1) is associated with disease progression in human oral squamous cell carcinoma, Cancer Sci. 109 (6) (2018) 2036–2045.

[32] K. Qu, Z. Wang, H. Fan, J. Li, J. Liu, P. Li, Z. Liang, H. An, Y. Jiang, Q. Lin, X. Dong, P. Liu, C. Liu, MCM7 promotes cancer progression through cyclin D1-dependent signaling and serves as a prognostic marker for patients with hepatocellular carcinoma, Cell Death Dis. 8 (2) (2017), e2603.

[33] T.M. Karve, A.K. Cheema, Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease, J. Amino Acids 2011 (2011) 207691.

[34] F. Ardito, M. Giuliani, D. Perrone, G. Troiano, L. Lo Muzio, The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review), Int. J. Mol. Med. 40 (2) (2017) 271–280.

[35] A. Bode, Z. Dong, Post-translational modification of p53 in tumorigenesis, Nat. Rev. Cancer 4 (2004) 793–805.

[36] S. Zheng, C. Zhang, X. Qin, et al., The status of phosphorylated p38 in esophageal squamous cell carcinoma, Mol. Biol. Rep. 39 (2012) 5315–5321.

[37] J. Salzman, H. Jiang, W.H. Wong, Statistical modeling of RNA-seq data, Stat. Sci. Rev. J. Inst. Math. Stat. 26 (1) (2011).

[38] H. Hamazaki, M.H. Hamazaki, Catalytic site of human protein-glucosylgalactosylhydroxylysine glucosidase: three crucial carboxyl residues were determined by cloning and site-directed mutagenesis, Biochem. Biophys. Res. Commun. 469 (2016) 357–362.

[39] M. Di Martile, D. Del Bufalo, D. Trisciuoglio, The multifaceted role of lysine acetylation in cancer: prognostic biomarker and therapeutic target, Oncotarget 7 (34) (2016) 55789–55810.

[40] J. Gil, A. Ramírez-Torres, S. Encarnación-Guevara, Lysine acetylation and cancer: a proteomics perspective, J. Proteom. 150 (2017 Jan) 297–309.