Computational and Structural Biotechnology Journal 21 (2023) 2845–2857

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Integrative big transcriptomics data analysis implicates crucial role of MUC13 in pancreatic cancer

Anupam Dhasmana [a,b,c], Swati Dhasmana [a,b], Shivangi Agarwal [d], Sheema Khan [a,b], Shafiul Haque [e,f,g], Meena Jaggi [a,b], Murali M. Yallapu [a,b], Subhash C. Chauhan [a,b,*]

[a] *Department of Immunology and Microbiology, School of Medicine, University of Texas Rio Grande Valley, McAllen, USA*
[b] *South Texas Center of Excellence in Cancer Research, School of Medicine, University of Texas Rio Grande Valley, McAllen, USA*
[c] *Himalayan School of Biosciences and Cancer Research Institute, Himalayan Institute of Medical Sciences, Swami Rama Himalayan University, Dehradun, India*
[d] *Department of Pathology, University of Illinois, Chicago, USA*
[e] *Research and Scientific Studies Unit, College of Nursing and Allied Health Sciences, Jazan University, Jazan, Saudi Arabia*
[f] *Gilbert and Rose-Marie Chagoury School of Medicine, Lebanese American University, Beirut, Lebanon*
[g] *Centre of Medical and Bio-Allied Health Sciences Research, Ajman University, Ajman, United Arab Emirates*

ARTICLE INFO

Article history:
Received 3 April 2023
Received in revised form 28 April 2023
Accepted 29 April 2023
Available online 1 May 2023

Keywords:
Pancreatic cancer
Mucins
MUC13
MUC13 genomic forms
Bigdata
Transcriptomics
Structural-Genomics & functional analysis

ABSTRACT

Big data analysis holds a considerable influence on several aspects of biomedical health science. It permits healthcare providers to gain insights from large and complex datasets, leading to improvements in the understanding, diagnosis, medication, and restraint of pathological conditions including cancer. The incidences of pancreatic cancer (PanCa) are sharply rising, and it will become the second leading cause of cancer related deaths by 2030. Various traditional biomarkers are currently in use but are not optimal in sensitivity and specificity. Herein, we determine the role of a new transmembrane glycoprotein, MUC13, as a potential biomarker of pancreatic ductal adenocarcinoma (PDAC) by using integrative big data mining and transcriptomic approaches. This study is helpful to identify and appropriately segment the data related to MUC13, which are scattered in various data sets. The assembling of the meaningful data, representation strategy was used to investigate the MUC13 associated information for the better understanding regarding its structural, expression profiling, genomic variants, phosphorylation motifs, and functional enrichment pathways. For further in-depth investigation, we have adopted several popular transcriptomic methods like DEGseq2, coding and non-coding transcript, single cell seq analysis, and functional enrichment analysis. All these analyzes suggest the presence of three nonsense MUC13 genomic transcripts, two protein transcripts, short MUC13 (s-MUC13, non-tumorigenic or ntMUC13), and long MUC13 (L-MUC13, tumorigenic or tMUC13), several important phosphorylation sites in tMUC13. Altogether, this data confirms that importance of tMUC13 as a potential biomarker, therapeutic target of PanCa, and its significance in pancreatic pathobiology.

1. Introduction

Big data analysis portrays a decisive role in developing biomedical health science by accelerating the advancement of personalized medicine, enhancing diagnostic and treatment resources, and advising in public health guidelines. As the amount of data created in biomedicals and healthcare industries is increasing, the role of big data analysis is becoming highly significant right from the prediction of current trends of certain parameters to future events of disease(s).

Cancer is a complicated, multi-stepped and progressive disease that contains distinct cellular and molecular processes. Subsequently, the cancer exploring groups produce huge amounts of molecular and phenotypic data to investigate cancer traits [1]. The large-scale omics data generation is channelized by the discoveries of high-throughput technologies. These breakthroughs have given the concept of 'big data' in cancer. The perfect example of big data accumulation is the assembling of The Cancer Genome Atlas (TCGA). TCGA contains 2.5 petabytes of raw data. This huge amount of data is 2500 times higher than advanced personal computers [2]. Certainly,

[*] Corresponding author at: South Texas Center of Excellence in Cancer Research, School of Medicine, University of Texas Rio Grande Valley, McAllen, USA.
E-mail address: subhash.chauhan@utrgv.edu (S.C. Chauhan).

https://doi.org/10.1016/j.csbj.2023.04.029

the blend of big data, bioinformatics and artificial intelligence has managed to lead significant developments to our basic understanding of cancer mechanism and its translational uses. Further rigorous effort needed among basic molecular biologists, data scientists, physicians, and lawmakers for the best utilization of these data in the development of diagnostic and therapeutic protocols [3].

Pancreatic cancer (PanCa) is a devastating and aggressive malignancy with the worst prognosis. By 2030 PanCa is projected to be the second prominent cause of cancer related deaths in the United State. PanCa has a very poor prognosis as only 11 % of PanCa patients have a 5-years of survival rate, as a result, the mortality rate of PanCa is almost equal to the incidence rate. The main clinical challenge with PanCa is poor treatment outcomes due the late diagnosis. Although there is availability of conventional biomarkers like CA19-9, CA125, Mucin 1 (MUC1) and carcinoembryonic antigen (CEA), but these biomarkers do not have optimal sensitivity and specificity for PanCa [4].

Mucins are a group of glycoproteins primarily engaged in the lubrication of epithelial cell surface, barrier protection and cell signaling. However, the aberrant expression of mucins has been implicated in various pathological conditions, including cancer, and cancer diagnosis. Mucins have been found to have diverse expression profiles amongst the gastro-intestinal cancers like colon, intestine, gastric and pancreas [5,6]. Among all mucins, MUC1 is recognized as reference gene in gastro-intestinal cancers and its unusual upregulation is observed in more than 60% of PanCa incidents, which corelates with the poor prognosis of the patients. It has been observed that MUC1 upregulation is also linked to the progression of different types of cancer, such as ovarian, lung, liver, cholangiocarcinoma, gallbladder, thyroid, colorectal, pancreatic and breast cancer [7].

Due to lack of specific onco-antigen, the current scenario urgently needs a biomarker for the clinical diagnosis of PanCa. Recent reports have proposed that Mucin 13 (MUC13), a novel oncogene exhibits high expression in PanCa while minimum expression in normal pancreas. MUC13 is a high-molecular-weight transmembrane glycoprotein. Like other mucins, MUC13 can be distinguished by a tandem repeat (TR) domain, which is one of the key features of mucins. TRs domain comprises of numerous serine and threonine residues that act as glycosylation sites. MUC13 also contains 3 epidermal growth factor (EGF)-like domains and a highly versatile, unique functional cytoplasmic tail which includes prospective phosphorylation sites to control the cellular signal transduction pathways [8]. The overexpression of MUC13 is also responsible for augmenting tumorigenic and metastatic phenotypes. The characteristics of PanCa cells may be mediated by physical interactions between MUC13 and HER2/Neu receptor via tyrosine kinase activity. MUC13 expression instigate the nuclear translocation of NF-κB p65 and phosphorylation of IκB, which upregulates the expression of GLUT-1, c-Myc, and Bcl-2, that are implicated in glucose metabolism. MUC13 has a direct crucial role in glucose metabolism, it stabilizes the GLUT1 receptor in pancreatic cancer cells. [9,10]. But still there are lots of opportunities to excavate the information about MUC13 from various OMICS databases.

2. Methods

2.1. Structural elucidation of MUC13

The homology modeling of human MUC13 (Uniprot ID: Q9H3R2) protein was done using SPARKS-X, one of the prudent single-method fold recognition and structure prediction server [11]. ModRefiner, a high resolution protein structure refinement server was used for the refinement of best modeled protein [12]. The best refined model was selected on the basis of the more 90 % of residues fall in the most favored region of Ramachandran plot by using ProCheck [13]. For identification of highly conserved, exposed, and functional residues ConSurf server was used [14,15]. ConSurf functions on the algorithm, for distinguishing functional residues of proteins by assessing the degree of conservation of residues sites among their identical sequence's homologs. Open target platform and previously published article [8] was used to describe and elaborate different domains, regions, unique and antigenic sequences of MUC13 genomic and protein structure.

2.2. Normal organ specific protein expression coverage, tissue specific gene expression analysis of MUC13 and MUC1 with its correlation

The organ specific protein expression coverage and tissue specific gene level comparison of MUC13 and MUC1 in normal conditions were performed by utilizing Human Protein Atlas and Genotype-Tissue Expression (GTEx) server [16,17]. Clustal Omega 2.1 was used for the sequence alignment between MUC13 and MUC1 (Uniprot ID: P15941) [18]. The comparative gene expression of MUC1 (reference gene) and MUC13 (test gene) in normal and pancreatic adenocarcinoma (PAAD) condition was performed using GEPIA2 [19]. The expressions were measured on the scale of log2(TPM + 1). The total sample size of tumor and normal samples was 179 and 171, respectively. The log2FC cutoff value was 1 and p value cutoff was 0.01. The pairwise gene correlation analysis was done between MUC13, MUC1 and S100A4 by using GEPIA2 correlation analysis plugin. The spearman correlation coefficient was used to compute the correlation. PAAD TCGA tumor, TCGA normal and GTEx databases were used to get the correlation between both genes.

2.3. Data collection and processing for DEG analysis

The mRNA raw count profiles of PDAC patients were downloaded from TCGA dataset (https://tcga-data.nci.nih.gov/tcga/). A total of 168 (164 PDAC and 04 normal) samples were available on TCGA, and the full manifest files were procured from this portal. The count data were normalized and visualized by using R packages "DEseq2 package" and "Enhanced Volcanic graph". The differentially expressed genes (DEG) in PDAC samples and control tissues were identified using R package "DESeq2" with a cut-off of |log2-fold change| > 1 and Padj < 0.05 (P-value adjusted for multiple testing using Benjamini-Hochberg method.

2.4. Copy number variation, gene expression by pathological stage and disease-free survival plot

ONCOMINE gene expression array datasets (https://www.oncomine.org/; last date of access 2nd March 2020), was a cancer microarray database used to analyze the mRNA expression and copy number variation of oncogenes like MUC13 in various clinical conditions like blood, normal pancreas and pancreatic cancer condition using a Student's t test to generate a p value. The cutoff of p value and fold change were defined as 0.01 and 2, respectively [20]. MUC13 gene expression in various pathological staging like stage I–IV and disease-free survival were done by using GEPIA2. In pathological staging the MUC13 expression was recorded on log2(TPM + 1) level. In disease free survival analysis median was the group cutoff with 95 % of confidence interval and hazard ratio.

2.5. Location of MUC13 in single cell type cluster of pancreatic cells

The location of MUC13 in pancreas cell cluster was assessed by single cell RNA sequencing (scRNA-seq) using Uniform Manifold Approximation and Projection (UMAP) plot, from the Protein Atlas database [4,17] based on healthy human tissues. Using this method, the foretelling location of MUC13 was detected in specific cell clusters of pancreatic cells from the range of 0–60 nTPM value and the number of included cells were expressed in the form of read count and cell count.

2.6. Analysis of MUC13 isoforms and phosphorylation sites

The expression distribution (violin plot) and isoform structure of MUC13 in PAAD condition was assessed by using isoform details sub module of GEPIA2. The genomic transcript data of GEPIA2 was retrieved from the TCGA database. In GEPIA2 isoform classes module exhibited total 60,498 various genes and theirs' 198,619 isoforms details like protein coding, retained intron and processed transcript etc. Clustal Omega was used for the sequence alignment between long form (Uniprot ID: Q9H3R2) and short form of MUC13 (Uniprot ID: C9IZG1) [18]. PhosphoSitePlus was used to locate phosphorylation and ubiquitylation sites on MUC13.

2.7. Impact of socio-behavioral and demographic aspect on MUC13 expression in pancreatic cancer

Socio-behavioral aspects were recorded by using University of ALabama at Birmingham CANcer data analysis Portal (UALCAN). UALCAN is a broad, user-friendly, and interactive web resource for analyzing cancer OMICS data. The backend programming is based on PERL-CGI with high quality graphics using javascript and CSS. This web server is intended to offer easy entry to publicly available cancer OMICS data (TCGA, MET500, CPTAC and CBTTC) [21]. The variation of MUC13 expression level in PAAD condition was assessed based on race, age, gender, diabetes status, alcohol habit and pancreatitis conditions. All input data was extracted from TCGA sample sets.

2.8. Functional enrichment of MUC13 and associated co-expressed genes

LinkedOmics is an incredibly effective publicly accessible portal that comprises multi-omics data from all 32 TCGA cancer types and 10 Clinical Proteomics Tumor Analysis Consortium (CPTAC) cancer cohorts. The submodules of LinkedOmics like LinkFinder (for gene association) and LinkInterpreter (for enrichment analysis) submodules of LinkedOmics [16] were used for exploring the genes that exhibited disparity in association with MUC13 in Pancreatic Cancer. From the UNC Institute's module, RNA-seq datasets, with 178 sample sizes were selected and scrutinized using pearson correlation coefficient and characterized in volcano plots. The positive or negative correlated genes were screened by the p-value ($p < 0.01$). The co-expressed genes were utilized for the gene enrichment analysis by using the Gene Set Enrichment Analysis (GSEA) database (gsea_result_134411_1677868274.rnk.), under the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathways plugin. The parameters for enrichment analysis were lowest no. of IDs in the 25 categories, lowest and highest no. of IDs in the categories were 3 and 2000 respectively, top 25 significance levels, 5000 number of permutations [4,22].

3. Results

3.1. Structural elucidation of MUC13

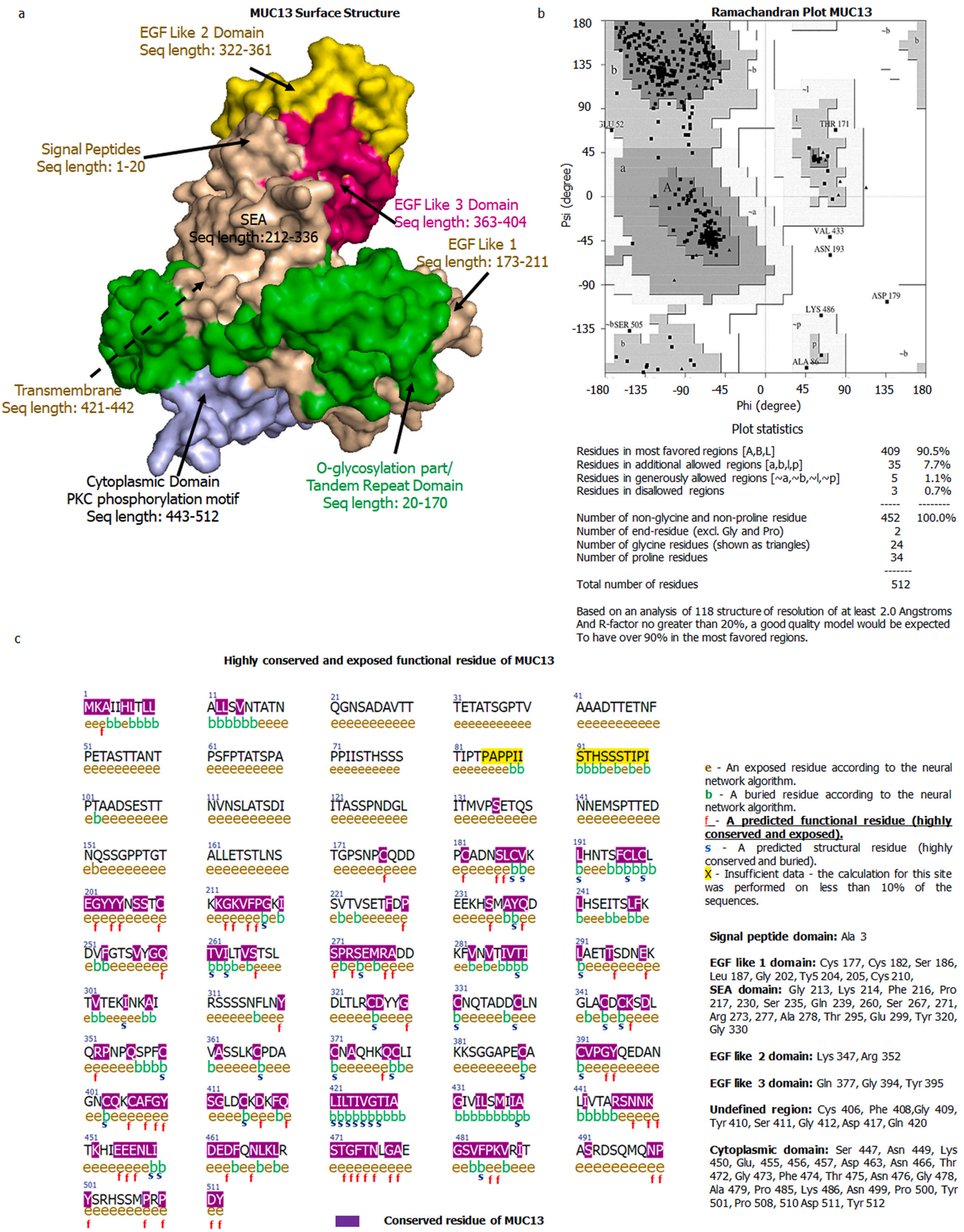
The amino acid sequence of MUC13 was obtained from the uniport database (Uniprot ID: Q9H3R2). The obtained sequence was considered as the input query sequence for the SPARKS-X (fold recognition and structure prediction tool, details mentioned in Supplementary Table 1) to model MUC13 protein. The selection of best structure among all 10 predicted proteins was performed based on z score. The highest ranked model (z score 8.17) was considered for the energy minimization and refinement of protein. ModRefiner was used to stabilize and refine the protein (Fig. 1a). After that the refined model of protein was uploaded for the quality check using ProCheck software. The uploaded protein scored more than 90 % of residues in most favored regions of the plot (90.5 % amino acids), 7.7 % of residues in additional allowed regions, 1.1 % of residues generously allowed regions and only 0.7 % of residues in disallowed regions (Fig. 1b). The comparison of structural validation was also performed between SPARKS-X and AlphaFold2 generated MUC13 models using Ramachandran plot (Procheck). Subsequent validation was found that SPARKS-X generated model has better quality, which showed more than 90 % of residues in favored region, whereas AlphaFold2 generated MUC13 model observed only 78.5 % of residues were in favored or core region, that's why SPARKS-X model was preferred over AlphaFold2 generated MUC13 model. ConSurf was used for the identification of highly conserved, exposed, and functional residues. After uploading MUC13 sequence, ConSurf server predicted 63 out of 512 amino acids as highly conserved, exposed, and functional residues (Fig. 1c). Most of the highly conserved, exposed, and functional residues are in cytoplasmic domain of MUC13, followed by SEA domain, EGF like domains (1–3) and undefined region. The structure of MUC13 was minutely scrutinized by the Open Target platform for the better elucidation of various important domains, regions, and fragments (Fig. 2). In this elucidation MUC13 structure was fragmented in various domains with exact location like Signal peptides, Tandem repeat region, EGF like domain, SEA domain and cytoplasmic domain. Apart from that glycosylated residues, disulfide bonds, antigenic regions, and unique peptide of MUC13 were precisely mentioned in the figure.

3.2. Normal organ-specific protein coverage, tissue specific gene expression and protein sequence similarity analysis of MUC13 with MUC1

Human Protein Atlas (HPA) was used for spotting the coverage of MUC13 and MUC1 in various organs at normal condition (Fig. 3a and b). Total 30 organs (14 high expression, 08 medium expression and 08 low expression) were identified with the expression of MUC1. Whereas only 05 organs have been noticed with the expression of MUC13. GTEx server was used to compare the expression level of MUC13 with a reference gene MUC1 in cancer and normal/disease free conditions. Among various organs, normal pancreas was aimed for comparative gene expression of both of genes. In this analysis MUC1 showed 57.03 TPM, whereas MUC13 demonstrated a very weak expression (only 1.603 TPM) in normal pancreas (Fig. 4a). Further, to evaluate the sequence homology between MUC13 (test protein) and MUC1 (reference protein), Clustal Omega was used for the multiple sequence alignment analysis, which showed only 22.96 % sequence similarity, which is not a significant homology among these two genes (Fig. 4b).

3.3. Differential gene expression (DEG) and correlation analysis between MUC13 and MUC1

For preprocessing of DEG analysis, mRNA raw counts of 168 (164 PDAC and 04 normal) samples were extracted from TCGA database. DEseq2 package was utilized for the spotting of differentially expressed genes in PDAC condition (273 upregulated genes and 207 negative regulated genes). MUC13 was found to be significantly upregulated gene with 3.73 log2 fold change, p adj value 0.001442265 and 92 ranked among all upregulated genes. Whereas MUC1 scored 2.52 log2 fold change, p adj value 0.031440648 and 192 ranked out of 273 upregulated proteins (list of all upregulated genes along with their statistics mentioned in Supplementary Table 2) (Fig. 5a). The comparative gene expression of MUC1 and MUC13 in normal and pancreatic adenocarcinoma (PAAD) tumor condition was done using GEPIA2, here MUC13 found around 6 time


a
MUC13 Surface Structure
EGF Like 2 Domain
Seq length: 322-361
Signal Peptides
Seq length: 1-20
SEA
Seq length:212-336
EGF Like 3 Domain
Seq length: 363-404
EGF Like 1
Seq length: 173-211
Transmembrane
Seq length: 421-442
Cytoplasmic Domain
PKC phosphorylation motif
Seq length: 443-512
O-glycosylation part/
Tandem Repeat Domain
Seq length: 20-170
b
Ramachandran Plot MUC13
180
135
90
45
0
-45
-90
-135
-180
-135
-90
-45
0
45
90
135
180
Psi (degree)
Phi (degree)
b
~b
A
a
~a
l
~l
p
~p
GLU 52
THR 171
VAL 433
ASN 193
ASP 179
LYS 486
~bSER 505
ALA 86
c
Highly conserved and exposed functional residue of MUC13
1 MKAIIHLTLL eeebbebbbb
11 ALLSVNTATN bbbbbbeeee
21 QGNSADAVTT eeeeeeeeeee
31 TETATSGPTV eeeeeeeeeee
41 AAADTTETNF eeeeeeeeeee
51 PETASTTANT eeeeeeeeee
61 PSFPTATSPA eeeeeeeeee
71 PPIISTHSSS eeeeeeeeee
81 TIPTPAPPII eeeeeeeebb
91 STHSSSTIPI bbbbebebeb
101 PTAADSESTT ebeeeeeeee
111 NVNSLATSDI eeeeeeeeee
121 ITASSPNDGL eeeeeeeeee
131 ITMVPSETQS eeeeeeeeee
141 NNEMSPTTED eeeeeeeeee
151 NQSSGPPTGT ebeeeeeeee
161 ALLETSTLNS eeeeeeeeee
171 TGPSNPCQDD eeeeeeeeee
181 PCADNSLCVK eeeeeeebbe
191 LHNTSFCLCL beeebbbbbb
201 EGYYYNSSTC eeeeeeeeee
211 KKGKVFPGKI eeeeeeebeb
221 SVTVSETFDP ebebeeeeee
231 EEKHSMAYQD eeeeeebbee
241 LHSEITSLFK beeebbebbe
251 DVFGTSVYGQ ebbeeeeeee
261 TVILTVSTSL bbbbebeeee
271 SPRSEMRADD ebebebeeee
281 KFVNVTIVTI eebebebbbb
291 LAETTSDNEK beeeeeeeee
301 TVTEKINKAI ebbeeeeebb
311 RSSSSNFLNY eeeeeebeee
321 DLTLRCDYYG eeeeebeeee
331 CNQTADDCLN beeeeebeee
341 GLACDCKSDL ebebebeeee
351 QRPNPQSPFC eeeeeebbbb
361 VASSLKCPDA beeeeebeee
371 CNAQHKQCLI beeeeeebbe
381 KKSGGAPECA eeeeeeeebe
391 CVPGYQEDAN beeeeeeeee
401 GNCQKCAFGY eebeeeeeee
411 SGLDCKDKFQ eeeebeeebe
421 LILTIVGTIA bbbbbbbbbb
431 GIVILSMIIA bbbbbbbbb
441 LIVTARSNNK bbbbbeeeee
451 TKHIEEENLI eeeeeeeebb
461 DEDFQNLKLR eeebeebebe
471 STGFTNLGAE eeeeeeeeee
481 GSVFPKVRIT eebbeebeee
491 ASRDSQMQNP eeeeeeeeee
501 YSRHSSMPRP eeeeeeeeee
511 DY ee
Conserved residue of MUC13

Plot statistics

Residues in most favored regions [A,B,L]	409	90.5%
Residues in additional allowed regions [a,b,l,p]	35	7.7%
Residues in generously allowed regions [~a,~b,~l,~p]	5	1.1%
Residues in disallowed regions	3	0.7%
	-----	--------
Number of non-glycine and non-proline residue	452	100.0%
Number of end-residue (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	24	
Number of proline residues	34	

Total number of residues	512	

Based on an analysis of 118 structure of resolution of at least 2.0 Angstroms And R-factor no greater than 20%, a good quality model would be expected To have over 90% in the most favored regions.

e - An exposed residue according to the neural network algorithm.
b - A buried residue according to the neural network algorithm.
f - **A predicted functional residue (highly conserved and exposed).**
s - A predicted structural residue (highly conserved and buried).
X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

Signal peptide domain: Ala 3

EGF like 1 domain: Cys 177, Cys 182, Ser 186, Leu 187, Gly 202, Ty5 204, 205, Cys 210,
SEA domain: Gly 213, Lys 214, Phe 216, Pro 217, 230, Ser 235, Gln 239, 260, Ser 267, 271, Arg 273, 277, Ala 278, Thr 295, Glu 299, Tyr 320, Gly 330

EGF like 2 domain: Lys 347, Arg 352

EGF like 3 domain: Gln 377, Gly 394, Tyr 395

Undefined region: Cys 406, Phe 408,Gly 409, Tyr 410, Ser 411, Gly 412, Asp 417, Gln 420

Cytoplasmic domain: Ser 447, Asn 449, Lys 450, Glu, 455, 456, 457, Asp 463, Asn 466, Thr 472, Gly 473, Phe 474, Thr 475, Asn 476, Gly 478, Ala 479, Pro 485, Lys 486, Asn 499, Pro 500, Tyr 501, Pro 508, 510 Asp 511, Tyr 512

Fig. 1. Structure and features of MUC13: a) Surface 3D structure of MUC13 along with various domains of MUC13. b) Ramachandran Plot of MUC13 indicates that more than 90 % of residues are in the most favored region, which is considered as good quality of model. c) Positioning of highly conserved, exposed and functionally (notation by f) active residues of MUC13.

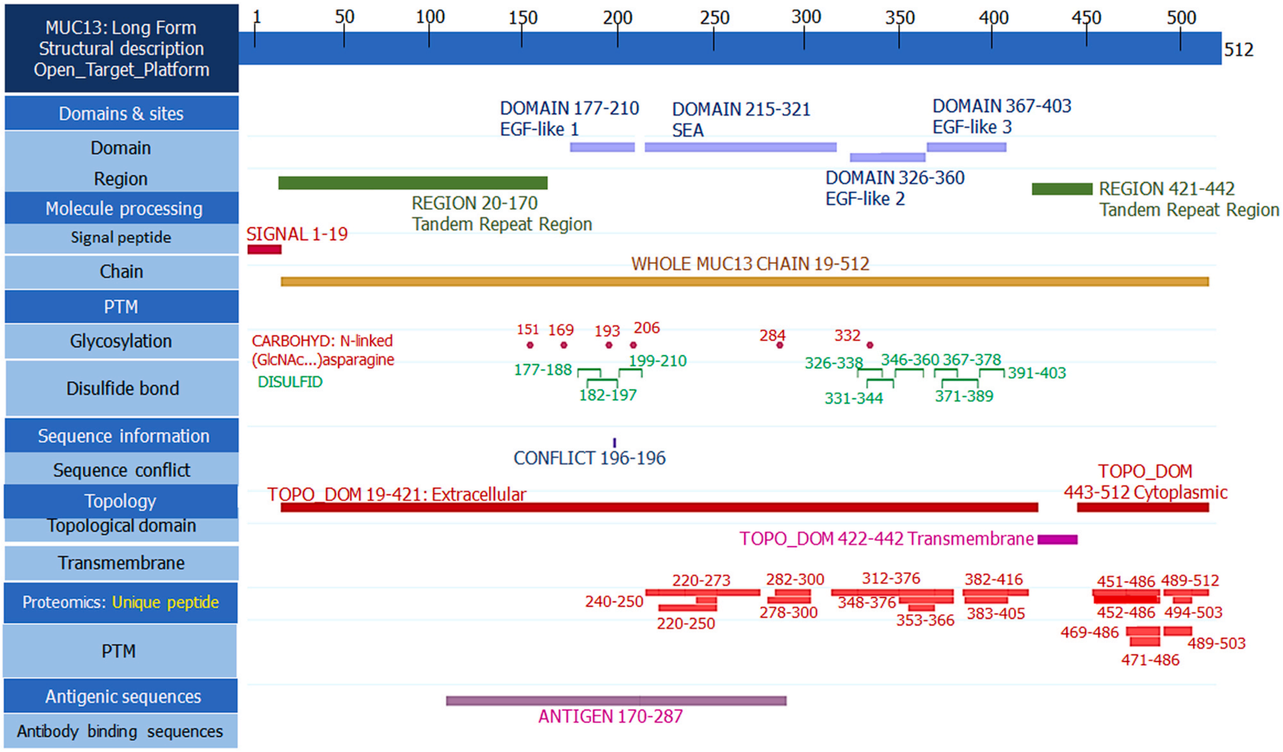


Fig. 2. Structural elucidation of various important fragments, domains and regions of MUC13.

over expressed in PAAD condition (Fig. 5b) while, MUC1 was found only 1.5 times over expressed in PAAD condition (Fig. 5c) in comparison of normal/disease free state. The pairwise gene correlation analysis was done between MUC13, MUC1 and S100A4 in PAAD condition. The computed value found a strong positive correlation between MUC13 and MUC1 with 0.71 R value and 5.9e^{-54} p value (Fig. 5d) and observed strong positive correlation between MUC13 and S00A4 with 0.75 R value and 3.9e^{-64} (Fig. 5e).

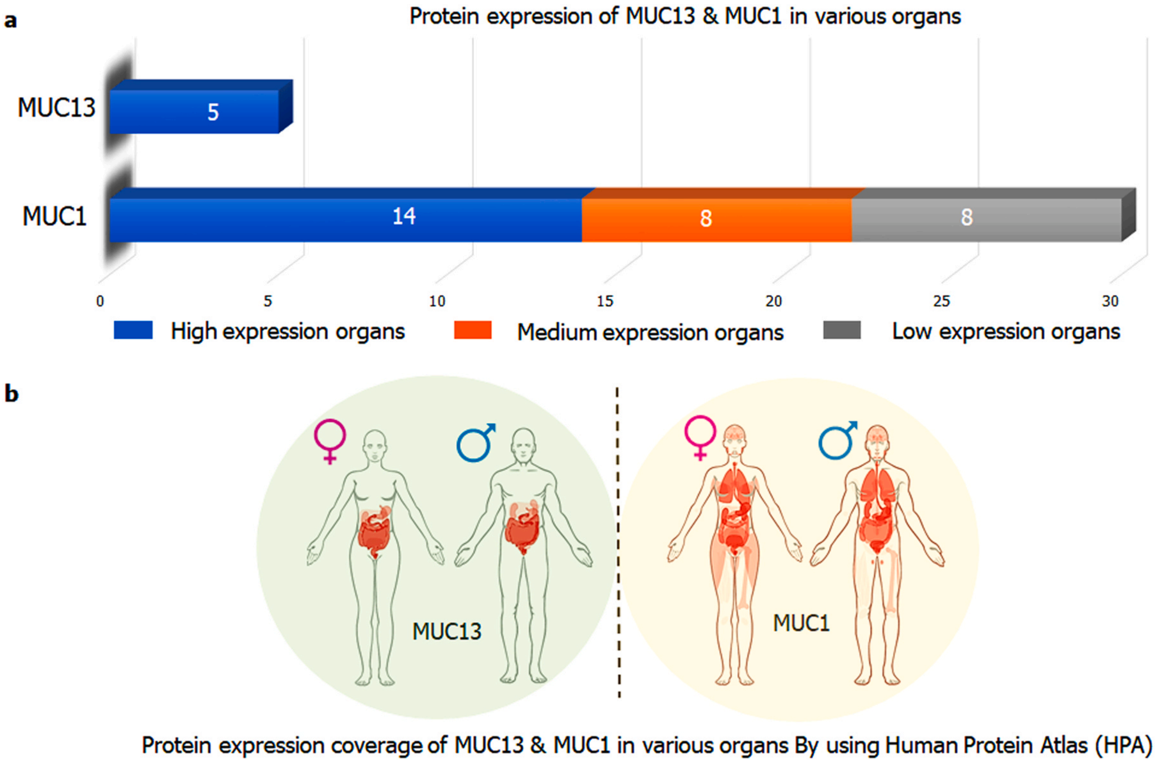


Fig. 3. Protein expression coverage and positioning of MUC13 and MUC1: a) Degree of protein expression coverage of MUC13 and MUC1. b) Positioning of MUC13 and MUC1 coverage.

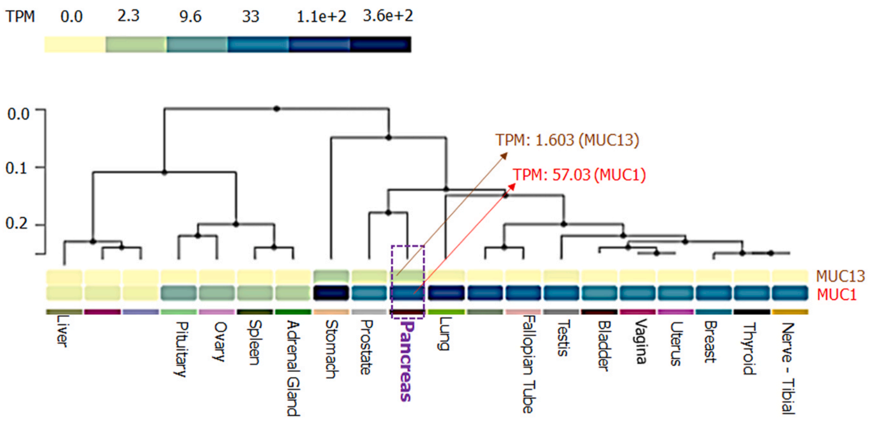


Fig. 4. Expression and sequence similarity analysis of MUC13 and MUC1: GTEx server-based comparison of the mRNA expression level of test gene MUC13 with a reference gene MUC1 in normal/disease free conditions.

3.4. *Copy number variation, gene expression by pathological stage and disease-free survival plot*

The oncomine data was extracted to identify the copy number of MUC13 in normal pancreas, and pancreatic cancer. In this analysis MUC13 copy number is significantly higher in pancreatic cancer while it was almost negligible in normal pancreas (Fig. 6a). The initial staging of PAAD showed very high expression of MUC13, specially in IA, IB, IIA and IIB stage, while among higher stages (III- IV) it has shown relatively lesser MUC13 expression (Fig. 6b). In addition, we determined the relationship of higher MUC13 expression (with Hazard Ratio 0.65, Logrank p = 0.05, among patient's sample size 178)) with pancreatic cancer patients' survival. In this analysis the higher expression of MUC13 correlated with diminished or poor patient's survival (Fig. 6c).

3.5. *Location of MUC13 in single cell type cluster of pancreatic cells*

The probable location of MUC13 in the pancreatic cell cluster was demonstrated in this section. The maximum expression was observed in the ductal cells followed by exocrine cells and mixed cell types. These single cell type clusters were indicating the higher expression of MUC13 in ductal cell clusters like C0 (15.8 nTPM), C2 (50.5 nTPM), C6 (6.0 nTPM), C12 (1.6 nTM) and C13 (36.6 nTPM). Remaining clusters like C11 belonged to pancreatic exocrine cells, and C15 belonged to mixed cell types. This graph was plotted based on a UMAP plot as shown in Fig. 7.

3.6. *Isoforms and phosphorylation sites of MUC13*

Total 05 transcripts of MUC13 were identified during the isoform analysis of MUC13. ENST00000616727.4 and ENST00000478191.1 were protein coding transcripts, remaining ENST00000497378.1 was processed transcript, ENST00000490147.1 and ENST00000462728.1 were retained intron or nonsense transcripts (Fig. 8a). Among all 05 transcript only 02 transcripts were protein coding transcripts those were accountable for synthesis of long (Isoform ID: ENST00000616727.4 and Isoform symbol: MUC13–001 or L-MUC13) and short (Isoform ID: ENST00000478191.1 and Isoform symbol: MUC13–003 or s-MUC13) form of MUC13. For better structural understanding of both forms, multiple sequence alignment (MSA) was performed (Fig. 8b). This analysis clearly cited that long and short form of MUC13 have 92.39 % of sequence homology. As comparing the structural features of both forms of MUC13 we found that short form has 184 amino acid sequences length and only three domains (small TR domain, EGF Like domain and SEA domain), while long form of MUC13 has 512 long amino acid sequences with all 5 domains (Signal peptide, TR domain, EGF like domain 1–3, SEA domain and Cytoplasmic domain) as shown in Fig. 8c. Interestingly, the long form of MUC13 was much more prevalent in tumors as compared to the short form (Fig. 8a), thus we considered long MUC13 isoform as tumorigenic (tMUC13) and short form as non-tumorigenic (ntMUC13) PhosphoSitePlus was used to locate phosphorylation sites and we identified that S471, S482, T490, S495, Y501, S502, 506 & Y512 residues are involved in phosphorylation and K308 and 468 residues are ubiquitylation sites on MUC13.

3.7. *Impact of socio-behavioral and demographic aspect on MUC13 expression in pancreatic cancer*

The impact of socio-behavioral factors on MUC13 expression in pancreatic cancer patients was evaluated by using UALCAN. In this segment various factors were analyzed including race/ethinicity (Fig. 9a). Interestingly, MUC13 expression was significantly higher in African Americans (TPM median 173.66, p value 2.729700E-02), and Asian (TPM median 133.24, p value 6.812700E-04) as compared to the Caucasian (TPM median 67.80, p value 7.21E-12). The age factor (Fig. 9b) is also an important factor. Higher expression of MUC13 was noticed at age 61–80 years (TPM median 81.2, p value 3.76E-10) followed by age range from 81 to 100 years (TPM median 78.52, p value 3.48E-03) and lowest among age range from 41 to 60 years (TPM median 76.4, p value 1.87E-09). Additionally, male pancreatic cancer patients have shown higher MUC13 expression (TPM median 87.1, p value 5.58E-11) as compared to the female patients (TPM median 69.1, p value 3.76E-09) as shown in Fig. 9c. Moreover, MUC13 expression is also likely to be influenced by life style factors such as alcohal consumption (Normal-vs-Daily Drinker p value 8.383E-05, TPM median 63.85; Normal-vs-Weekly Drinker p value 1.001800E-02, TPM median 83.98; Normal-vs-Occasional Drinker p value 6.204100E-03, TPM median 94.93, Normal-vs-Social Drinker p value 1.354160E-02, TPM median 30.05) as shown in Fig. 9d and some pathological conditions like diabetes (Normal-vs-Diabetic p value 4.839900E-04, TPM median 54.86) and pancreatitis satuts (Normal-vs-Pancreatitis p value 7.93E-03, TPM median 83.92) as shown in Fig. 9e and f.

3.8. *Functional enrichment of MUC13 and associated co-expressed genes*

LinkedOmic, was used to investigate the functional and pathways enrichment analysis of MUC13 along with its associated genes in a cohort of 178 patient sample size, total of 19,774 genes were enriched in this study, of them, 11,057 genes showed negative correlations (green dots) with MUC13 and the remaining 8717 genes

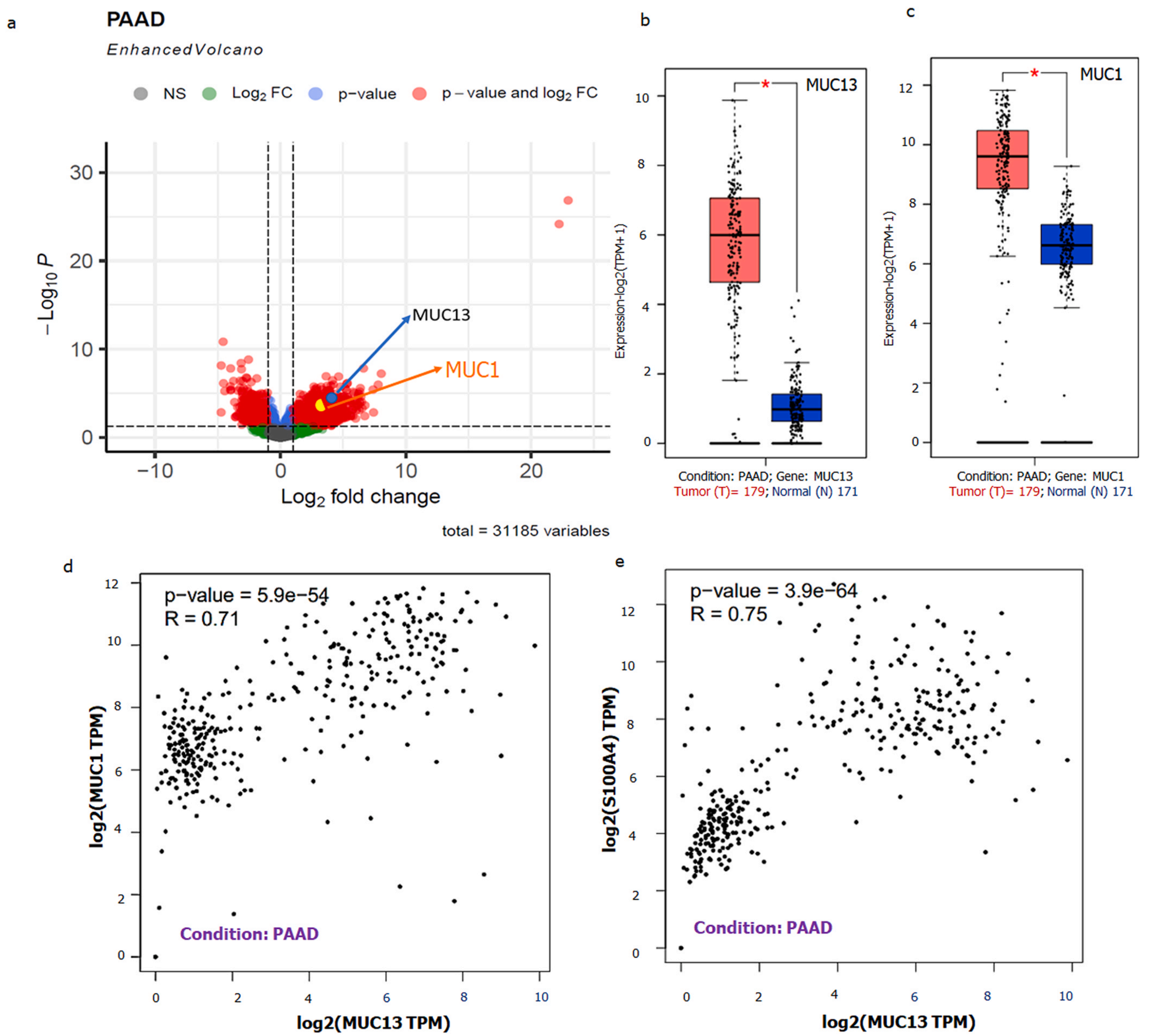


Fig. 5. Transcriptomics of MUC13 and MUC1: a) Differentially expressed genes (DEG) among 168 patients (164 PDAC + 04 Normal tissues). Among all 60,660 transcripts that are screened, total 273 genes are significantly upregulated in this analysis. MUC13 has secured 92nd ranked with a 3.73 Log2Fold change and 0.001442 p value, while ranked 192 with 2.52 Log2Fold change and 0.031441 p value. b and c) GEPIA generates box plots with 0.4 jitter value and p cut off 0.01 for comparing the expression in PAAD and normal pancreas expression for MUC13 & MUC1. Total sample size normal n = 171, and Tumor n = 179). d) Correlation analysis of MUC1 and MUC13, R value was 0.71 and p value 5.9 e^{-59}, which showed strong positive association. e) Correlation analysis of MUC13 and S100A4, R value was 0.75 and p value 3.9 $^{e-64}$, which showed strong positive association. (Correlation coefficients guidelines: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html).

showed significant correlation (red dots) with MUC13 (Fig. 10a). The list of gene set along with their respective pathways was listed in Supplementary Table 3. The most positive and negative corelated genes asociated with MUC13 are represented in heatmap (Fig. 10 b and c). For functional and pathway enrichment of the corelated genes with MUC13, enrichment studies were executed using GSEA method and KEGG pathways category in the LinkInterpreter submodule of LinkedOmics. As exhibited in Fig. 10d, the MUC13 corelated genes have positive involvement in chemical carcinogenesis, lipid and glucose metabolism (ether lipid metabolism, glycerolipid metabolism mucin type O-glycan biosynthesis, pentose and glucuronate interconversions, sphingolipid metabolism), pancreatic secretion, bile secretion and maturity onset of diabetes among the young patients. Whereas, DNA replication, mismatch repair, nucleotide excision repair, RNA polymerase, Fanconi anemia pathway, FoxO signaling pathway, RNA surveillance pathways showed a negative correlation with MUC13. The enrichment plots of some important positive and negative enriched functional and pathways are demonstrated in Fig. 10e.

4. Discussion

Pancreatic cancer (PanCa) patients suffer from extremely low survival and high mortality rate, due to the late diagnosis, lack early and specific detection biomarkers, appropriate molecular level understanding of the disease progression, and availability of effective therapeutic modalities to manage the advance disease conditions. Thus, urgent investigations are highly desirable in these areas to

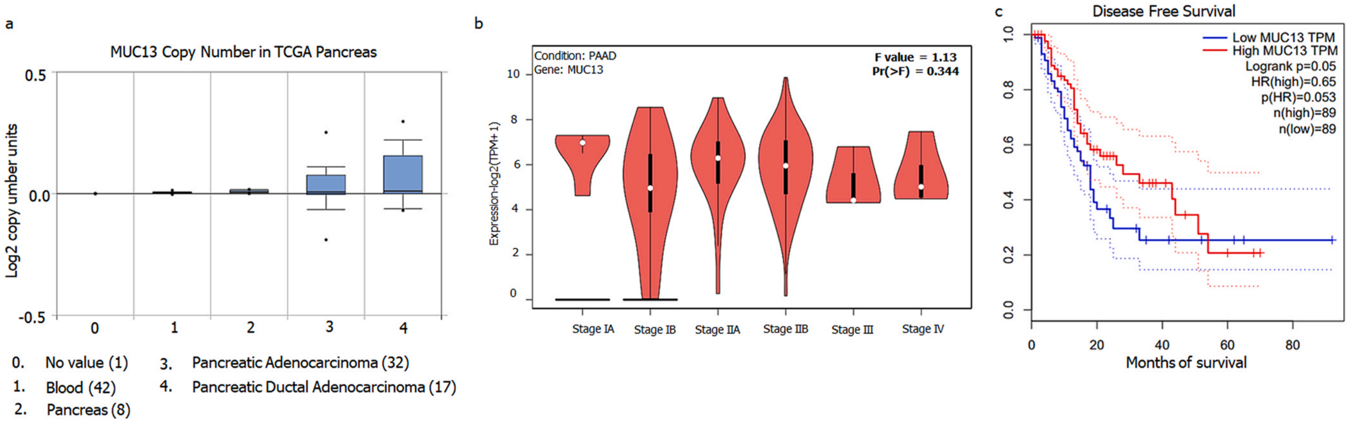
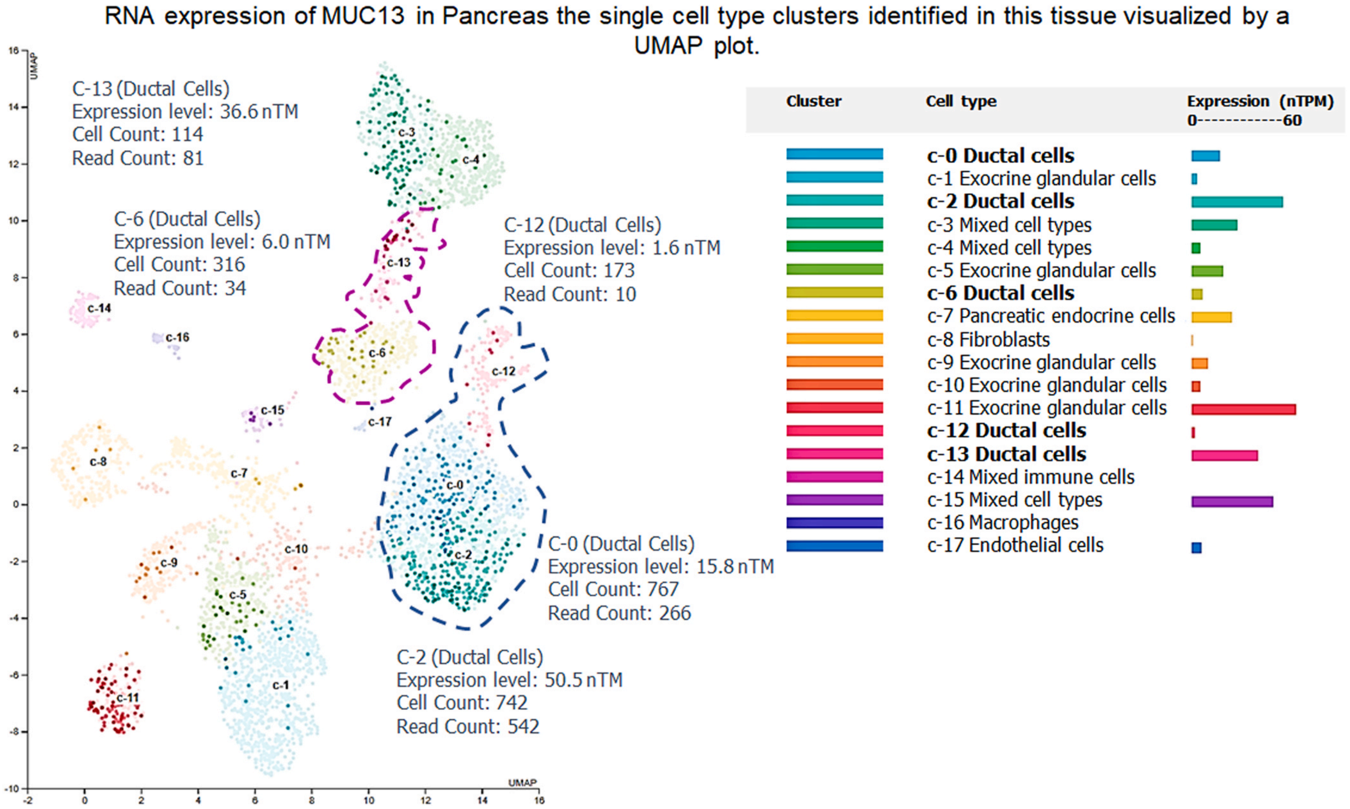


Fig. 6. Expression level of MUC13 in various cancer stages and survival analysis: a) MUC13 expression in non-pathological and pathological conditions. b) mRNA expression level of MUC13 in different stages of pancreatic cancer. c) Disease free survival plot based on the MUC13 expression level.

improve the management of PanCa. Towards this effort, herein, we described the role of a new transmembrane glycoprotein, MUC13, as a potential biomarker of PanCa, defined its structural features, genomic variance, and involvement in disease progression by using big transcriptomic data mining and computational systems biology approaches. Investigations published from our lab and others have suggested oncogenic role of MUC13. The high expression of MUC13 is observed in PanCa, while undetectable in normal pancreatic tissue [9].

Through literature survey, the comparative ROC of MUC1, MUC13 and CA-19.9 were also observed. Published literature suggests ROC value of CA-19.9 around 72 % sensitivity, 62 % specificity [23], MUC1 87 % sensitivity, 75.5 % specificity [24]. However, in a small cohort though, MUC13 has shown sensitivity 90, specificity 100 % in early lesions of pancreatic cancer (PanINs) and 94 sensitivity, 90 % specificity in pancreatic ductal adenocarcinoma samples in relation to healthy controls [9]. This observation reinforce the oncogenic alignment of MUC13 in PanCa among other popular biomarkers.

MUC13 has very unique structural features as compared to other transmembrane mucins like MUC1 and MUC4. The observed structural differences was observed between MUC1 (Uniprot ID: P15941), MUC4 (Uniprot ID: Q99102) and MUC13 (Uniprot ID: Q9H3R2) using Uniprot database. As per this analysis, MUC1 contains TR domain, SEA module and a functional cytoplasmic domain, while MUC4 contains TR domain, EGF like domains with a very short, almost non-functional cytoplasmic tail. MUC13, however holds all the unique



Cluster	Cell type	Expression (nTPM) 0------------60
	c-0 Ductal cells	
	c-1 Exocrine glandular cells	
	c-2 Ductal cells	
	c-3 Mixed cell types	
	c-4 Mixed cell types	
	c-5 Exocrine glandular cells	
	c-6 Ductal cells	
	c-7 Pancreatic endocrine cells	
	c-8 Fibroblasts	
	c-9 Exocrine glandular cells	
	c-10 Exocrine glandular cells	
	c-11 Exocrine glandular cells	
	c-12 Ductal cells	
	c-13 Ductal cells	
	c-14 Mixed immune cells	
	c-15 Mixed cell types	
	c-16 Macrophages	
	c-17 Endothelial cells	

Fig. 7. MUC13 positioning in different pancreatic cell types: RNA expression of MUC13 in pancreas tissues. The single cell type clusters identified in pancreatic tissue were visualized by a UMAP plot.

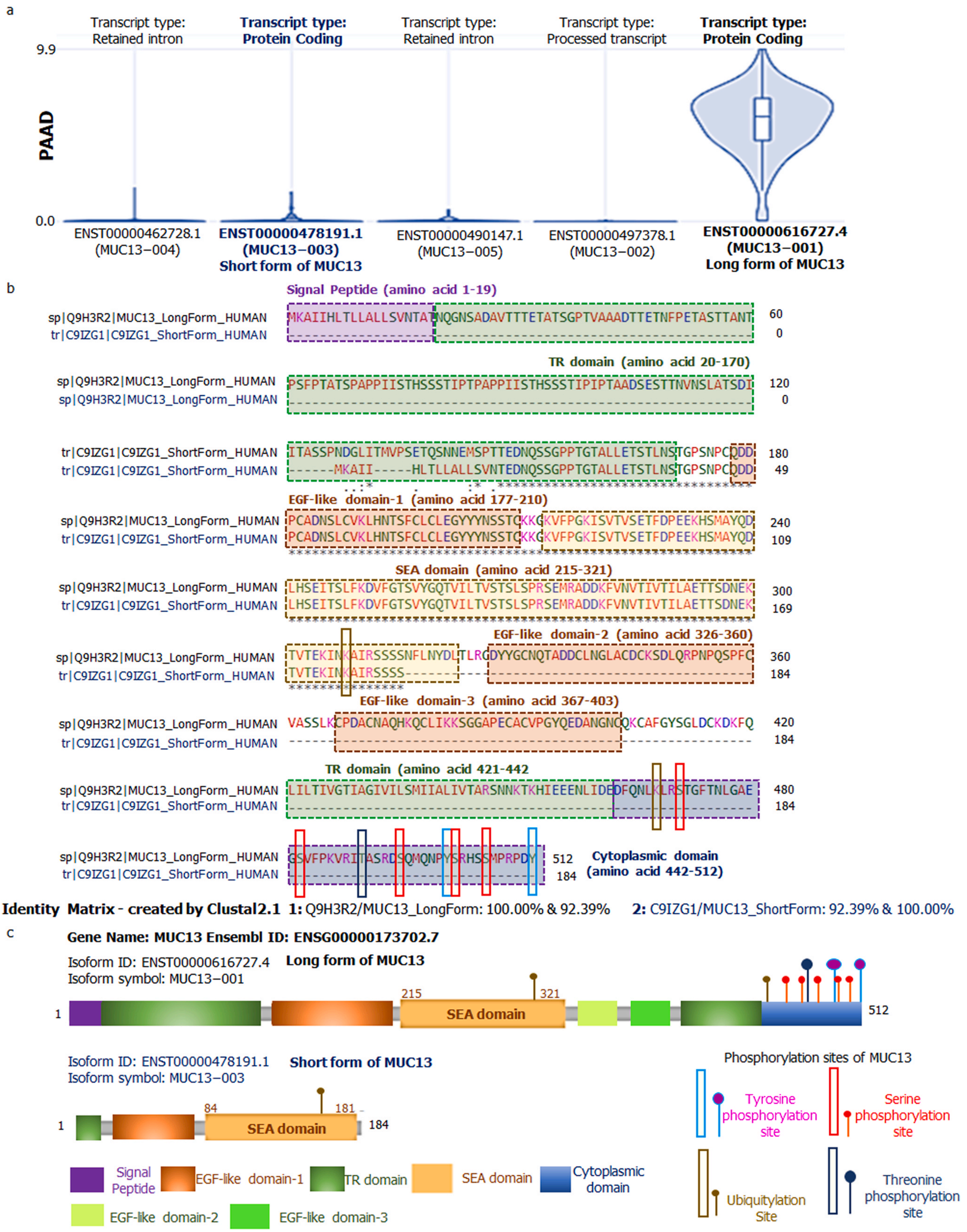

a
Transcript type: Retained intron
Transcript type: Protein Coding
Transcript type: Retained intron
Transcript type: Processed transcript
Transcript type: Protein Coding
9.9
PAAD
0.0
ENST00000462728.1 (MUC13−004)
ENST00000478191.1 (MUC13−003) Short form of MUC13
ENST00000490147.1 (MUC13−005)
ENST00000497378.1 (MUC13−002)
ENST00000616727.4 (MUC13−001) Long form of MUC13
b
Signal Peptide (amino acid 1-19)
sp|Q9H3R2|MUC13_LongForm_HUMAN MKAIIHLTLLALLSVNTATNQGNSADAVTTTETATSGPTVAAADTTETNFPETASTTANT 60
tr|C9IZG1|C9IZG1_ShortForm_HUMAN 0
TR domain (amino acid 20-170)
sp|Q9H3R2|MUC13_LongForm_HUMAN PSFPTATSPAPPIISTHSSSTIPTPAPPIISTHSSSTIPIPTAADSESTTNVNSLATSDI 120
sp|Q9H3R2|MUC13_LongForm_HUMAN 0
tr|C9IZG1|C9IZG1_ShortForm_HUMAN ITASSPNDGLITMVPSETQSNNEMSPTTEDNQSSGPPTGTALLETSTLNSTGPSNPCQDD 180
tr|C9IZG1|C9IZG1_ShortForm_HUMAN ------MKAII-----HLTLLALLSVNTEDNQSSGPPTGTALLETSTLNSTGPSNPCQDD 49
EGF-like domain-1 (amino acid 177-210)
sp|Q9H3R2|MUC13_LongForm_HUMAN PCADNSLCVKLHNTSFCLCLEGYYYNSSTCKKGKVFPGKISVTVSETFDPEEKHSMAYQD 240
tr|C9IZG1|C9IZG1_ShortForm_HUMAN PCADNSLCVKLHNTSFCLCLEGYYYNSSTCKKGKVFPGKISVTVSETFDPEEKHSMAYQD 109
SEA domain (amino acid 215-321)
sp|Q9H3R2|MUC13_LongForm_HUMAN LHSEITSLFKDVFGTSVYGQTVILTVSTSLSPRSEMRADDKFVNVTIVTILAETTSDNEK 300
tr|C9IZG1|C9IZG1_ShortForm_HUMAN LHSEITSLFKDVFGTSVYGQTVILTVSTSLSPRSEMRADDKFVNVTIVTILAETTSDNEK 169
EGF-like domain-2 (amino acid 326-360)
sp|Q9H3R2|MUC13_LongForm_HUMAN TVTEKINKAIRSSSSNFLNYDLTLRGDYYGCNQTADDCLNGLACDCKSDLQRPNPQSPFC 360
tr|C9IZG1|C9IZG1_ShortForm_HUMAN TVTEKINKAIRSSSS-- 184
EGF-like domain-3 (amino acid 367-403)
sp|Q9H3R2|MUC13_LongForm_HUMAN VASSLKCPDACNAQHKQCLIKKSGGAPECACVPGYQEDANGNCQKCAFGYSGLDCKDKFQ 420
tr|C9IZG1|C9IZG1_ShortForm_HUMAN -- 184
TR domain (amino acid 421-442)
sp|Q9H3R2|MUC13_LongForm_HUMAN LILTIVGTIAGIVILSMIIALIVTARSNNKTKHIEEENLIDEDFQNLKLRSTGFTNLGAE 480
tr|C9IZG1|C9IZG1_ShortForm_HUMAN 184
sp|Q9H3R2|MUC13_LongForm_HUMAN GSVFPKVRITASRDSQMQNPYSRHSSMPRPDY 512
tr|C9IZG1|C9IZG1_ShortForm_HUMAN 184
Cytoplasmic domain (amino acid 442-512)
% Identity Matrix - created by Clustal2.1 1: Q9H3R2/MUC13_LongForm: 100.00% & 92.39% 2: C9IZG1/MUC13_ShortForm: 92.39% & 100.00%
c
Gene Name: MUC13 Ensembl ID: ENSG00000173702.7
Isoform ID: ENST00000616727.4 Long form of MUC13
Isoform symbol: MUC13−001
1
215
SEA domain
321
512
Isoform ID: ENST00000478191.1 Short form of MUC13
Isoform symbol: MUC13−003
1
84
SEA domain
181
184
Phosphorylation sites of MUC13
Tyrosine phosphorylation site
Serine phosphorylation site
Ubiquitylation Site
Threonine phosphorylation site
Signal Peptide
EGF-like domain-1
TR domain
SEA domain
Cytoplasmic domain
EGF-like domain-2
EGF-like domain-3

Fig. 8. Transcripts of MUC13 and their structures: a) Various MUC13 transcripts in PAAD condition. Only two transcripts ENST00000616727.4 and ENST00000478191.1 were protein coding isoforms of MUC13. b) Multi sequence alignment of MUC13 Long and Short form with similarity index 92.39 %. c) Protein structural features of long and short form of MUC13.

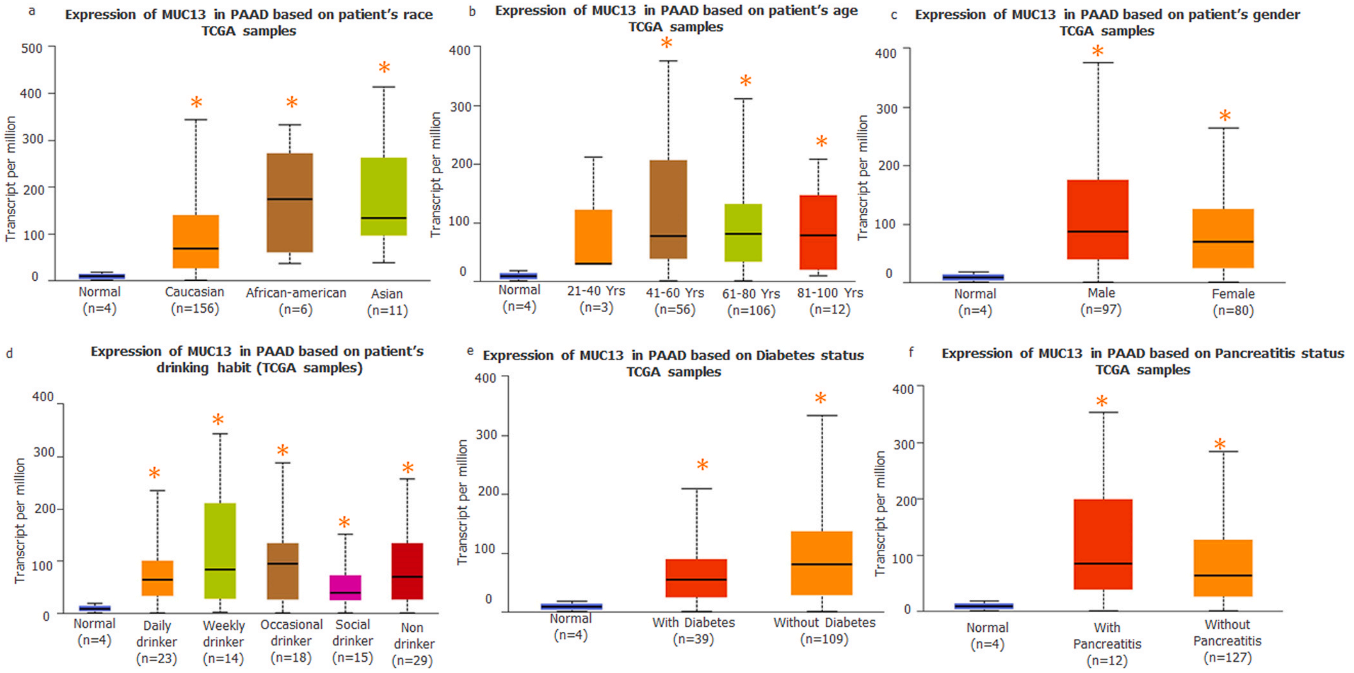


Fig. 9. Impact of socio-behavioral and demographic factors on MUC13 expression in Pancreatic Cancer: a) Impact of race/ethinicity on MUC13 expression, b) Impact of age on MUC13 expression, c) Impact of gender on MUC13 expression, d) Impact of alcohal habit on MUC13 expression, e and f) Impact of diabetes and pancreatitis satuts on MUC13 expression.

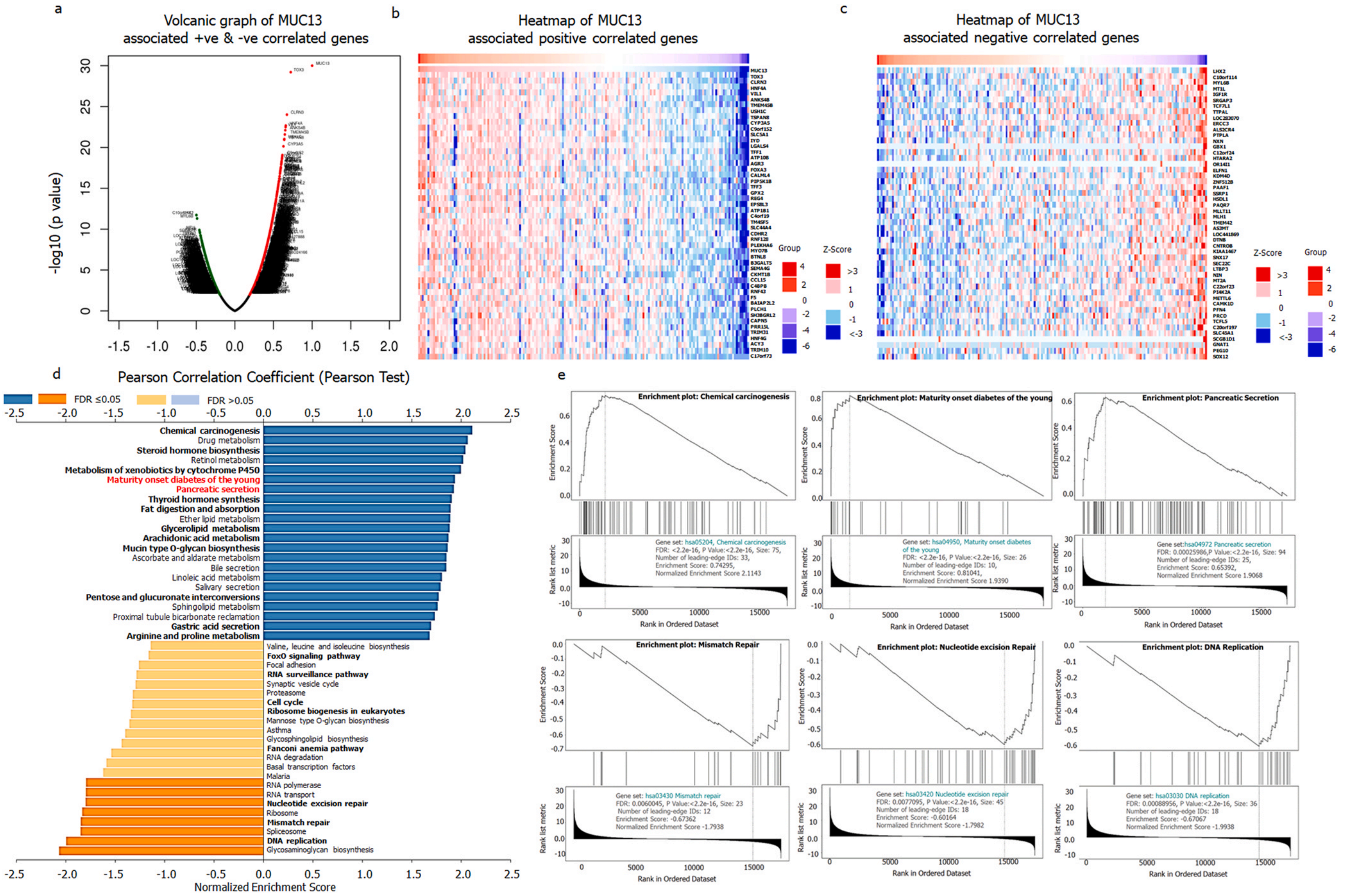


Fig. 10. Functional enrichment of MUC13: a) Volcanic graph of MUC13 representating positive (red) and negative (green) correlated genes. b) Heatmap of MUC13 assocated positive correlated genes. c) Heatmap of MUC13 assocated negative correlated genes. d) Functional and pathways enrichment analysis of MUC13 associated positive (blue) and negative (yellow) correlated genes. e) Enrichment plot of important enriched functional pathways.

features of both MUC1 and MUC4 mucins, like TR, SEA, EGF like domains and a long functional cytoplasmic tail with several phosphorylation sites. Due to these unique features, MUC13 appears to have a different place among all the transmembrane mucins. In this article, big data analysis, was utilized to made an attempt to comprehensively arrange all the scattered information about mucin MUC13. This study was initiated by elucidating the putative structural features of MUC13. Since there is no crystal structure available for MUC13 in public domain, we have modeled it through computational and systems biology approach to visualize its various domain and identify various exposed and functional residues. Approximately 63 highly conserved, exposed and functionally active residues are identified in MUC13 through this analysis. Most of these residues are present in cytoplasmic domain, which clearly suggest the crucial role of MUC13 cytoplasmic tail in cellular signaling. The structure of MUC13 was minutely scrutinized and observed for the precise location of all domains and regions. Total 6 glycosylation sites were observed, out of which, 2 residues in Tandem repeat region (151, 169), 2 residues in EGF like-1 domain (193, 206), 1 residue in SEA domain (284) and 1 residue in EGF like 2 domain (332). Hyper glycosylation in tandem repeat region leads to bulky MUC13 structure which will be equal in weight to heavy Mucins (MUC1, MUC4, MUC15 & MUC16) [8]. However, EGF like domains were recognized as, rich in disulfide bonds. The same domains along with SEA and cytoplasmic domains also hold some unique natured peptides, which create the unique positioning of MUC13 in proteomics. The antigenic sequence found from 107 to 287 was also scattered from tandem repeat domain to SEA domain. The antigenic region of tandem repeat domain potentially will be responsible for the antibody recognition.

This study indicates high expression level of MUC1 in normal pancreas which is almost 50 times higher than the MUC13 expression. MUC1 is one of the oncogene in PanCa [25] and other cancers [7], thus MUC1 holds the gold standard cancer biomarker status. Therefore, MUC1 was used as reference gene for the correlation analysis with MUC13. S100A4 is known metastatic biomarker that why MUC13 correlation was done t identify the metastatic potential of MUC13 in pancreatic cancer condition. In this investigation, MUC13, MUC1 and S100A4 have shown strong and positive expression correlation. Following this, multiple sequence alignment was performed between MUC1 and MUC13 and found only 22.96 % sequence similarity. This analysis clearly revealed that MUC1 and MUC13 had diverse sequences, even in global protein sequence BLAST, MUC13 did not get homology with any other proteins listed in non-redundant protein sequences database. In the PanCa condition expression level (via DEGseq2 of TCGA-PAAD data set) of both proteins was cross validated, interestingly, a better expression profile of MUC13 (~ 3.73 fold differential expression) was observed in PanCa patients as compared to MUC1 (~ 2.52 fold differential expression). Moreover, MUC13 has shown almost six fold higher expression in pancreatic tumors as compared to normal pancreas tissues, while MUC1 was only 1 fold higher. This differential expression profile suggest better specificity of MUC13 in PanCa with respect to MUC1. The copy number of MUC13 was found relatively higher in PanCa condition as compared to normal healthy condition. MUC13 expression analysis in the different tumor staging also provides a clear view regarding relatively higher MUC13 expression in early stages (IA, IB, IIA and IIB) of PanCa, as compared to the advance stages (III and IV). The higher expression of MUC13 has also shown a positive correlation with lower disease-free survival in PanCa patients. This data indicates that the higher expression of MUC13 in initial stage can be harnessed as an early detection biomarker of PanCa.

The protein atlas based single cell seq analysis clearly presents the MUC13 position over ductal cells in pancreas. The detailed isoform analysis concluded that MUC13 has total five transcripts, among them, two transcripts (ENST00000616727.4 and ENST00000478191.1) of MUC13 are coding transcripts, while remaining are non-coding. Interestingly, ENST00000616727.4 transcript which is coding for the long form of MUC13 (L-MUC13) with 512 residues is the prevalent form of MUC13 that is expressed in tumors, thus we considered it as tumorigenic MUC13 (tMUC13). While, ENST00000478191.1 transcript encodes for the short form of the MUC13 (s-MUC13) with a protein of 184 residues, which has shown fairly less expression in tumors. After deep comparative analysis of long and short form of MUC13, we found that short form of MUC13 had missing parts in tandem repeat, EGF like 1,2,3 domains and also does not contain cytoplasmic tail which would be responsible for downstream cellular signal transductions as it contains all important phosphorylation sites. The MUC13 cytoplasmic domain contains total of 8 phosphorylation sites, among these 2 are the most important tyrosine phosphorylation sites (501 & 512) remaining others were serine (482, 495, 502), threonine (490) and ubiquitylation phosphorylation sites. The cytoplasmic tail of the long form of MUC13 could be responsible for the oncogenesis, as tyrosine phosphorylation sites synchronize various cellular functions related to cellular growth and differentiation [26]. The tumor-antigenicity of both forms of MUC13 were evaluated using VaxiJen portal which exhibited that the long form of MUC13 was probable-tumor antigenic by crossing the threshold level of 0.4, while short form of MUC13 was found to be probable non tumor antigen with scoring 0.2133. Based on these analysis, we considered short form of MUC13 as non-tumorigenic (nt-MUC13) isoform of MUC13.

In this study we also considered to investigate impact of the socio-behavioral and demographic factors on MUC13 expression. It was observed that, ethnicity, age and gender are important factors for higher expression of MUC13 in PanCa condition. Our analysis suggests that African American and Asian PanCa patients expressed relatively higher MUC13 as compared to Caucasian patients. Additionally, male patients have shown higher expression level of MUC13 as compared to female patients. The reason for this male vs female disparity could be due to protective female hormones [27]. According our analysis, alcohol consumption, diabetic status and inflammation in pancreas may also be responsible for the higher MUC13 expression in PanCa patients. So lifestyle and socio-behavioral factors could be important risk factors in pancreatic pathobiology. Finally, the functional enrichment analysis was done to understand the involvement of MUC13 and its associated genes in various functional pathways. As per the system biology approach single gene or protein cannot work in the isolation. Fluctuation in expression of a particular protein/gene mediates sweeping change in entire gene/protein circuit [28]. Thus, the functional and pathways enrichment analyses had provided us with the cumulative and collective changes in entire MUC13 associated interactome. Furthermore, it was elucidated that the higher expression of MUC13 leads to modulation of several important pathways like upregulation of chemical carcinogenesis, maturity onset diabetes of the young, pancreatic-bile secretion and several glucoses, lipid metabolism associated pathways. Some studies suggest that around 80 % of PanCa patients suffer with either early onset of diabetes or deregulation of glucose metabolism at the time of PanCa diagnosis [4]. Consequently, it can be assumed that in future, MUC13 and its associated genes can be deemed as biomarkers to ascertain the diabetes associated PanCa. Meanwhile, DNA repair related pathways like mismatch repair, nucleotide excision repair, Fanconi anemia and FoxO signaling pathways were down regulated in this enrichment analysis. This point clearly cited that the upregulation of MUC13 in PanCa may lead the deregulation of DNA repair pathways, which may accumulate and lead to the progression of cancer [29]. However, we completely understand that being a transmembrane mucin it might have certain limitations as it has shown relatively lower expression in poorly differentiated or late stage of PDACs in comparison to the well differentiated and moderately differentiated or early stage of cancers. Despite that, we strongly believe MUC13 has very high potential to effectively aid pancreatic cancer diagnosis and treatment in future. Although the big transcriptomics analysis is very fascinating and generates huge data sets which is helpful to scan every aspect of the study. But still, Omics data are inconsistent between cohorts, it may vary from batch to batch, thats why to cross validate the omics data; specific experimental platforms are mandatory to make this methodology more significant and reliable. The data scale gap is one of the most important problem of omics studies because most of clinical trials do not collect omics data just to reduce the cost of the overall study. Decreasing the cost of the sequencing, the omics data collection can become mandatory and standard requirement of any clinical study. This action will open huge avenues to explore and investigate the data set in various aspects. Additionally, existing data repositories, like as ClinicalTrials.gov and NCBI GEO, should come up with the common metalanguage standards for the better data integration [3,30].

5. Conclusion

This work gives insight into the use of big data analyses that could be applied to improve diagnosis of cancers as well as clarify the hot point indicators like structural, functional, and genomic properties of any biomarker at a mechanistic level. This study clearly demonstrates that MUC13 can be a new early diagnostic biomarker for PanCa, and it also has potential to improve the efficacy of the existing biomarker panel. Current analysis will improve the understanding about the role of MUC13 in PanCa that will ultimately help develop additional therapeutic strategies for the management of PanCa and reduce its burden. Further, the big data analysis approach is opening an important avenue for the discoveries of specific and significant biomarkers not only for PanCa but can also be useful for other malignancies.

CRediT authorship contribution statement

Anupam Dhasmana: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Swati Dhasmana:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Shivangi Agarwal:** Software, Validation, Formal analysis, Investigation, Data curation. **Sheema Khan:** Methodology, Software, Validation, Formal analysis, Writing – review & editing. **Shafiul Haque:** Methodology, Software, Validation, Formal analysis, Writing – review & editing. **Meena Jaggi:** Supervision, Project administration, Writing – review & editing, Visualization, Funding acquisition. **Murali M Yallapu:** Supervision, Project administration, Writing – review & editing, Visualization, Funding acquisition. **Subhash C Chauhan:** Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors gratefully acknowledge the start-up research grants from the Dept. of Immunology and Microbiology, School of Medicine, University of Texas Rio Grande Valley awarded to Meena Jaggi, Murali M. Yallapu, and Subhash C Chauhan. This work was supported by the NIH/NCI's funding R01s: CA210192; CA206069; CA204552; SC1GM139727 and Herb Kosten Foundation, Memphis. This study utilizes core facilities of CPRIT "Integrated Cancer Research Core (ICRC) (RP210180)" & "RP230419" and UT-System.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.04.029.

References

[1] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646–74.

[2] Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research. Nat Rev Cancer 2022;22(11):625–39.

[3] Jiang P, Sinha S, Aldape K, Hannenhalli S, Sahinalp C, Ruppin E. Big data in basic and translational cancer research. Nat Rev Cancer 2022;22(11):625–39.

[4] Dhasmana A, Dhasmana S, Kotnala S, Laskar P, Khan S, Haque S, et al. CEACAM7 expression contributes to early events of pancreatic cancer. J Adv Res 2023.

[5] van Putten JPM, Strijbis K. Transmembrane mucins: signaling receptors at the intersection of inflammation and cancer. J Innate Immun 2017;9(3):281–99.

[6] Cox KE, Liu S, Lwin TM, Hoffman RM, Batra SK, Bouvet M. The mucin family of proteins: candidates as potential biomarkers for colon cancer. Cancers 2023;15:5.

[7] Lan Y, Ni W, Tai G. Expression of MUC1 in different tumours and its clinical significance (review). Mol Clin Oncol 2022;17(6):161.

[8] Maher DM, Gupta BK, Nagata S, Jaggi M, Chauhan SC. Mucin 13: structure, function, and potential roles in cancer pathogenesis. Mol Cancer Res 2011;9(5):531–7.

[9] Khan S, Zafar N, Khan SS, Setua S, Behrman SW, Stiles ZE, et al. Clinical significance of MUC13 in pancreatic ductal adenocarcinoma. HPB: J Int Hepato Pancreato Biliary Assoc 2018;20(6):563–72.

[10] Kumari S, Khan S, Gupta SC, Kashyap VK, Yallapu MM, Chauhan SC, et al. MUC13 contributes to rewiring of glucose metabolism in pancreatic cancer. Oncogenesis 2018;7(2):19.

[11] Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27(15):2076–82.

[12] Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 2011;101(10):2525–34.

[13] Mukherjee S, Das S, Sriram N, Chakraborty S, Sah MK. In silico investigation of the role of vitamins in cancer therapy through inhibition of MCM7 oncoprotein. RSC Adv 2022;12(48):31004–15.

[14] Yariv B, Yariv E, Kessel A, Masrati G, Chorin AB, Martz E, et al. Using evolutionary data to make sense of macromolecules with a "face-lifted" ConSurf. Protein Sci: Publ Protein Soc 2023;32(3):e4582.

[15] Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19(1):163–4.

[16] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science 2020;369(6509):1318–30.

[17] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347(6220):1260419.

[18] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 2011;7:539.

[19] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019;47(W1):W556–60.

[20] Sun CC, Li SJ, Chen ZL, Li G, Zhang Q, Li DJ. Expression and prognosis analyses of runt-related transcription factor family in human leukemia. Mol Ther Oncol 2019;12:103–11.

[21] Chandrashekar DS, Karthikeyan SK, Korla PK, Patel H, Shovon AR, Athar M, et al. UALCAN: an update to the integrated cancer data analysis platform. Neoplasia 2022;25:18–27.

[22] Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 2018;46(D1):D956–63.

[23] Azizian A, Rühlmann F, Krause T, Bernhardt M, Jo P, König A, et al. CA19-9 for detecting recurrence of pancreatic cancer. Sci Rep 2020;10(1):1332.

[24] Gold DV, Karanjawala Z, Modrak DE, Goldenberg DM, Hruban RH. PAM4-reactive MUC1 is a biomarker for early pancreatic adenocarcinoma. Clin Cancer Res: J Am Assoc Cancer Res 2007;13(24):7380–7.

[25] Zhao P, Meng M, Xu B, Dong A, Ni G, Lu L. Decreased expression of MUC1 induces apoptosis and inhibits migration in pancreatic cancer PANC-1 cells via regulation of Slug pathway. Cancer Biomark: Sect A Dis Mark 2017;20(4):469–76.

[26] Singh V, Ram M, Kumar R, Prasad R, Roy BK, Singh KK. Phosphorylation: implications in cancer. Protein J 2017;36(1):1–6.

[27] Dhasmana S, Dhasmana A, Kotnala S, Mangtani V, Narula AS, Haque S, et al. Boosting mitochondrial potential: an imperative therapeutic intervention in amyotrophic lateral sclerosis. Curr Neuropharmacol 2022.

[28] Dhasmana A, Uniyal S, Anukriti, Kashyap VK, Somvanshi P, Gupta M, et al. Topological and system-level protein interaction network (PIN) analyses to deduce molecular mechanism of curcumin. Sci Rep 2020;10(1):12045.

[29] Dhasmana A, Jamal QM, Gupta R, Siddiqui MH, Kesari KK, Wadhwa G, et al. Titanium dioxide nanoparticles provide protection against polycyclic aromatic hydrocarbon BaP and chrysene-induced perturbation of DNA repair machinery: a computational biology approach. Biotechnol Appl Biochem 2016;63(4):497–513.

[30] Jiang P, Zhang Y, Ru B, Yang Y, Vu T, Paul R, et al. Systematic investigation of cytokine signaling activity at the tissue and single-cell levels. Nat Methods 2021;18(10):1181–91.