Phenome-Genome Association Studies of Pancreatic Cancer: New Targets for Therapy and Diagnosis

RAMASWAMY NARAYANAN

Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, Boca Raton, FL, U.S.A.

Abstract. *Background: Pancreatic cancer, has a very high mortality rate and requires novel molecular targets for diagnosis and therapy. Genetic association studies over databases offer an attractive starting point for gene discovery. Materials and Methods: The National Center for Biotechnology Information (NCBI) Phenome Genome Integrator (PheGenI) tool was enriched for pancreatic cancer-associated traits. The genes associated with the trait were characterized using diverse bioinformatics tools for Genome-Wide Association (GWA), transcriptome and proteome profile and protein classes for motif and domain. Results: Two hundred twenty-six genes were identified that had a genetic association with pancreatic cancer in the human genome. This included 25 uncharacterized open reading frames (ORFs). Bioinformatics analysis of these ORFs identified putative druggable proteins and biomarkers including enzymes, transporters and G-protein-coupled receptor signaling proteins. Secreted proteins including a neuroendocrine factor and a chemokine were identified. Five out of these ORFs encompassed non coding RNAs. The ORF protein expression was detected in numerous body fluids, such as ascites, bile, pancreatic juice, milk, plasma, serum and saliva. Transcriptome and proteome analyses showed a correlation of mRNA and protein expression for nine ORFs. Analysis of the Catalogue of Somatic Mutations in Cancer (COSMIC) database revealed a strong correlation across copy number variations and mRNA over-expression for four ORFs. Mining of the International Cancer Gene Consortium (ICGC) database identified somatic mutations in a significant number of pancreatic patients' tumors for most of these ORFs. The pancreatic cancer-associated ORFs were also found to be genetically associated with other neoplasms, including leukemia, malignant melanoma, neuroblastoma and prostate carcinomas, as well as other unrelated diseases and disorders, such as Alzheimer's disease, Crohn's disease, coronary diseases, attention deficit disorder and addiction. Conclusion: Based on Genome-Wide Association Studies (GWAS), copy number variations, somatic mutational status and correlation of gene expression in pancreatic tumors at the mRNA and protein level, expression specificity in normal tissues and detection in body fluids, six ORFs emerged as putative leads for pancreatic cancer. These six targets provide a basis for accelerated drug discovery and diagnostic marker development for pancreatic cancer.*

Pancreatic cancer is currently the fourth leading cause of cancer-related death in the United States (US) and is anticipated to become the second by 2020. Pancreatic cancer accounts for about 3% of all cancers in the US and 7% of cancer deaths. The American Cancer Society's most recent estimates for pancreatic cancer in the US is that in 2014 about 46,420 people (23,530 men and 22,890 women) will be diagnosed with pancreatic cancer and 39,590 people will die of pancreatic cancer. Current therapeutics for pancreatic cancer still revolve around chemotherapy (1). Thus, novel molecular targets are urgently required for this cancer type. Rational drug discovery for pancreatic cancer requires new drug targets to move further from the conventional chemotherapy (2-4). In addition, diagnosis of pancreatic cancer can be greatly aided by the discovery of pancreatic cancer-associated secreted proteins in body fluids.

The genetic association databases offer an attractive starting point for disease-relevant target discovery (5). The phenotypic traits associated with diseases can be readily enriched for associated genes using the Phenome Genome Integrator (PheGeni) tool (6) . The genes identified from the association studies can be verified for gene ontology and pathways using diverse bioinformatics tools. The expression

Correspondence to: Dr. Ramaswamy Narayanan, Department of Biological Sciences, Charles E. Schmidt College of Science, Florida Atlantic University, Boca Raton, FL 33431, U.S.A. Tel: +1 5612972247, Fax: +1 5612973859, e-mail: rnarayan@fau.edu

Key Words: Alzheimer's disease, biomarkers, body fluids, clinical variations, dark matter proteome, druggable targets, expression quantitative loci, gene ontology, genome-wide association studies, leukemia and lymphoma, malignant melanoma, neuroblastoma, pancreatic cancer, phenome-genome association, open reading frames.

specificity of the genes can be verified at the transcriptome and proteome level. The recent availability of human proteome expression datasets, the Human Proteome Map (HPM) (7) and the Proteomics DB (8), which encompasses protein expression data from diverse body fluids, provides a valuable avenue for verifying expression relevance of the targets identified.

In the present study, using the PheGenI genetic association tool, 226 pancreatic neoplasm-associated genes were identified in the human genome. Among this list of genes, 196 were known proteins. Twenty-five previously uncharacterized open reading frames (ORFs) were identified. A large number of the human proteins are uncharacterized ORFs; these ORFs have been termed as the Dark Matter of the human proteome (9-11). The Dark Matter offers an opportunity to discover new targets. Using a streamlined approach we have recently demonstrated discovery of cancer- (12-14) and diabetes-associated proteins (15) among the uncharacterized ORFs to aid in diagnosis and therapy. A novel secreted protein ORF, termed Secreted Glycoprotein in Chromosome X (SGPX) (14), and a putative druggable calcium-binding transporter Carcinoma Related EF-Hand Protein (CREF) (12) were discovered in these studies.

Reasoning that the pancreatic neoplasm-associated ORFs may offer a rationale for novel pancreatic cancer target discovery, detailed bioinformatics and proteomics approaches were undertaken. Results indicate a potential for druggableness in these ORFs, which encompass enzymes, transporters and receptor-binding classes of proteins. Further, ORFs with a signal peptide, including a chemokine, were identified from these efforts. The ORF expression was detected in diverse body fluids, including ascites, bile, plasma, saliva, milk and pancreatic juice. Lead ORFs were identified based on copy number variations (CNV), correlation of protein and mRNA expression and protein motif and domain analysis for druggableness. These results provide further evidence of the power of mining the human proteome for accelerated novel target discovery.

Materials and Methods

The bioinformatics and proteomics tools used in the study have been described elsewhere (12-14). The following genome-wide association tools were used: the Genetic Association Database (GAD) (5), the Catalogue of Somatic Mutations (COSMIC) (16), the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/), the cBioPortal (17), the Integrated Cancer Drug Discovery Platform (CanSar v2 database) (18), the Database for Annotation, Visualization and Integrated Discovery (DAVID v6.7 from the NCBI) (19), GeneALaCart (LifeMap discovery) from the GeneCards (20), the National Center for Biotechnology Information (NCBI) Phenotype-Genotype Integrator (PheGenI) (6), the Expression Quantitative Trait (eQTL) browser (21), the Database of Genomic Variants (DGV) (22), Clinical Variations (ClinVar) (23), the Oncomine database (24) and the International HapMap project (http://hapmap.ncbi.nlm.nih.gov/).

The entire database of GAD, Human Protein Atlas (HPA) (25) and UniGene was downloaded and the Excel filtering tool was used to scan for the ORFs. Batch analysis of the ORF database was performed for canSar, the Multi Omics Protein Expression Database (MOPED) (26), the DAVID annotation tool, the Human Proteome Map, Proteomics DB, the Human Proteins Reference Database (HPRD) (27), the PheGenI and the eQTL browser.

All of the bioinformatics mining was verified by two independent experiments. Big data were downloaded two independent times and the output verified for consistency. Only statistically significant results per each tool's requirement are reported. Prior to using a given bioinformatics tool, a series of control query sequences was tested to evaluate the predicted outcome of the results.

Results

Mining the human genome for pancreatic neoplasm association. The NCBI PheGenI association tool was used to establish an initial database of genes (Figure 1). A list of 226 genes was generated, including known proteins, uncharacterized genes and non-coding RNAs (ncRNAs). These genes showed positive genetic association for 233 single nucleotide polymorphisms (SNPs). The known protein classes included enzymes, growth factors and receptors, transcription factors, apoptotic proteins, oncogenes and tumor suppressor genes, ribosomal proteins, hormones and hormone receptors.

The study further identified 25 novel uncharacterized ORFs with genetic association evidence (see Table I for details on the ORFs). These ORFs included five ncRNAs including long intergenic RNA and antisense RNAs. As these uncharacterized ORFs offer a potential for new pancreatic cancer target discovery, detailed bioinformatics and proteomics characterization of these ORFs was undertaken.

Genome-Wide Association Studies (GWAS) of the pancreatic neoplasm-associated ORFs. To establish a definitive link for these ORFs with pancreatic neoplasms, diverse GWAS tools were used. The International Cancer Genome Consortium (ICGC) database was screened for the ORFs and the number of mutations present in pancreatic cancer patient samples were identified (Table II).

Somatic mutations were identified for 24/25 ORFs in pancreatic tumors. The mutations for six of these ORFs (C1orf95, C7orf8, C7orf10, C20orf39, C20orf45 and FAM19A5) were seen in more than 10% of the patients analyzed. The largest percentage of mutations was seen for the ORFs FAM19A5 (42.86%) and C7orf10 (38.1%).

The ICGC mutations were also verified by using the COSMIC database for copy number variations (CNV) and over-expression. Seven of the ORFs (C1orf94, C7orf4, C7orf10, C9orf123, C10orf67, FAM91A and KIAA1549) showed elevated copy number and over-expression in pancreatic tumor patients. The cBioPortal GWAS tool identified coding sequence mutations with somatic mutations

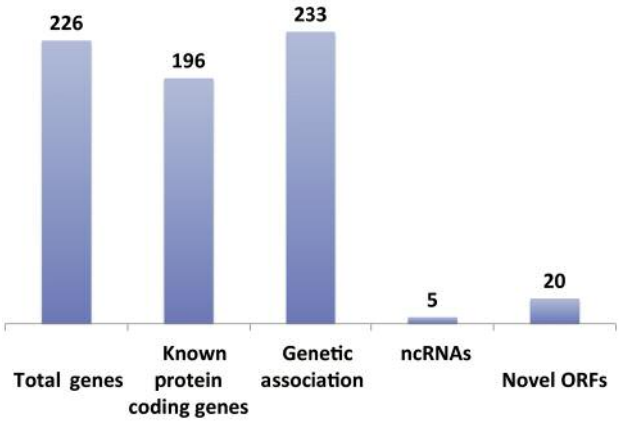

Figure 1. *Pancreatic cancer-associated genes in the human genome. The NCBI Phenome-Genome Integrator was used to identify genes showing pancreatic neoplasm association evidence. The number of genes for the indicated category is shown. ncRNA, Noncoding RNAs; ORF, open reading frames.*

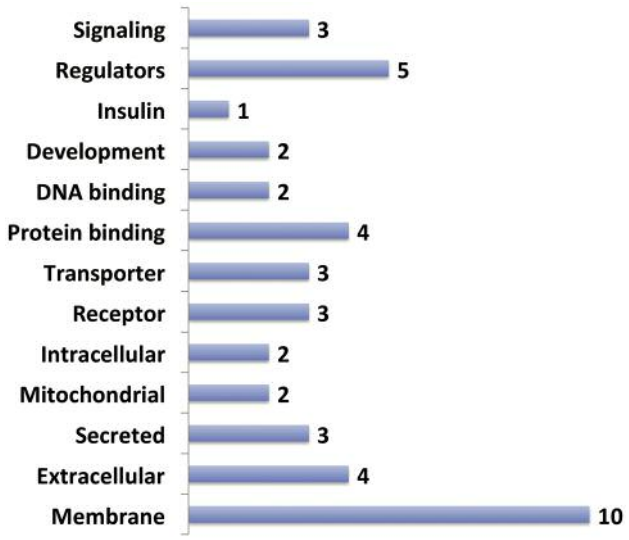

Figure 2. *Gene Ontology of the novel pancreatic neoplasm-associated proteins. The Gene Ontology (GO) inferred from the CanSar, DAVID and the GeneALaCart tools is shown. The numbers indicate the number of ORFs for each category shown.*

score 1 for C10orf67 (H147N), C20orf45 (R844C) and C20orf174 (E112K). The latter two mutations may be functionally important as the charge of the amino acids is changed due to mutations.

An additional hint of clinical relevance for these ORFs was established by mining the NCBI ClinVar database. The C7orf10 was associated with Glutaryl-coA oxidase deficiency (28). The C10orf67, FAM19A5, KIAA1217 and KIAA1549 were associated with lung cancer. The ORFs FAM149A, FAM19A5, FAM91A1, KIAA0232, KIAA1217 and KIAA1549 were associated with malignant melanoma. The C1orf44 was associated with left ventricular hypertrophy. These results helped establish a strong correlative evidence for the pancreatic association of the ORFs. Further, clinical relevance to other tumor types and disorders were also indicated.

Pancreatic neoplasm-associated ORFs characterization by Gene Ontology (GO). To develop a hint of function, cell location and the processes involved, the pancreatic neoplasm-associated ORFs were characterized using the CanSar and the GeneALaCart tools by batch analysis (Figure 2). ORFs related to signaling (including insulin and G-protein coupled receptor), nucleic acid and protein binding, transporters (vesicle and proteins), extracellular proteins (including secreted), transmembrane proteins (including receptors), cytoplasmic and mitochondrial proteins, developmental proteins (brain and embryo) and regulators (immune and neuronal) were identified by GO analysis.

Protein expression analysis in body fluids. The availability of diverse proteomics databases for expression studies, such as the MOPED (26), Human Proteome Map (7), the Proteomics DB (8), the HPRD (27) and the HPA (25) enabled this study for gene expression analysis of the pancreatic neoplasm-associated ORFs in diverse body fluids (Figure 3). In HPA tissue microarray samples, 18/20 protein coding ORFs were detected in pancreatic tumors by immunohistochemistry. ORFs expression were detected in ascites (C7orf49, FAM91A1), bile (KIAA1217), serum (KIAA1217), plasma (KIAA1217, KIAA1549, C1orf94), saliva (C7orf43), proximal fluid (C1orf39, C7orf8, FAM84B, FAM91A1, C14orf180) pancreatic juice (C7orf49) and milk (FAM84B, FAM91A1). In addition, membrane and nuclear localization was predicted for some of the ORFs.

The Human Proteome Map analysis helped establish a degree of selectivity in normal tissues. The C1orf94 and FAM19A5 expression was detected only in fetal and not in adult tissues. The C7orf orf9 expression was restricted to adult retina. The C10orf67, C11orf44 and C14orf180 expression was not detected in any tissue analyzed. The C20orf39 expression was restricted to adult frontal cortex. The C10orf84 expression was highly restricted to testis (fetal, adult) and B-cells. The C20orf174 expression was highly enriched in CD4+ T cells. None of the above-mentioned ORFs were detected in adult pancreas. The only ORFs detected in the normal adult pancreas included C1orf9, FAM84B, FAM91A1, KIAA0232 and KIAA1217. The lack of expression of these ORFs in diverse normal human tissues suggests a high level of specificity to the pancreatic neoplasms.

Table I. *Novel ORFs associated with pancreatic neoplasms*.

Gene name	Gene symbol	Aliases and descriptions	Refseq_Protein_ID	PubMed_IDs
C1ORF39	FNBP1L	Transducer of Cdc42-dependent actin assembly protein 1 | chromosome 1 open reading frame 39 | TOCA1 | C1orf39 | formin binding protein 1-like | toca-1 |formin-binding protein 1-like | Toca-1 | transducer of Cdc42-dependent actin assembly protein 1 | transducer of Cdc42-dependent actin assembly 1	NP_001020119.1|NP_001157945.1|NP_060207.2	14654988 | 19342671
C1ORF94	C1orf94	Chromosome 1 open reading frame 94 | uncharacterized protein C1orf94	NP_001128206.1| NP_116273.2	14702039 | 12477932
C6ORF155 ¶	LINC00472	C6orf155 |dJ288M22.3 | long intergenic non-protein coding RNA 472 | chromosome 6 open reading frame 155		14702039 | 14574404
C7ORF4 ¶	LINC00244	Chromosome 7 open reading frame 4 | long intergenic non-protein coding RNA 244 | NCRNA00244 | non-protein coding RNA 244 | C7orf4		10329000 | 12853948 |
C7ORF8	CTTNBP2	KIAA1758 | CORTBP2 | C7orf8 | cortactin binding protein 2 | CortBP2 | cortactin binding protein 2 | Orf4 | cortactin-binding protein 2	NP_219499.1	11707066 | 20379614
C7ORF9	NPVF	Neuropeptide NPVF | "FMRFamide-related peptide precursor" | "RFamide-related peptide precursor" | pro-FMRFamide-related neuropeptide VF | chromosome 7 open reading frame 9 | neuropeptide VF precursor | FMRFamide-related peptides | RFRP | C7orf9 | RFamide-related peptide | FMRFamide-related peptide	NP_071433.3	11951088 | 11025660
C7ORF10	SUGCT	Succinate--hydroxymethylglutarate CoA-transferase |ORF19 | SuccinylCoA:glutarate-CoA transferase | succinylCoA: glutarate-CoA transferase | "dermal papilla derived protein 13" | Dermal papilla-derived protein 13 | dermal papilla-derived protein 13 | "Russel-Silver syndrome candidate" | DERP13 | chromosome 7 open reading frame 10 | Russel-Silver syndrome candidate | C7orf10 |EC 2.8.3.13	NP_001180240.1| NP_001180241.1| NP_001180242.1 |NP_079004.1	23893049 | 11829489
C9ORF70 ¶	GLIS3-AS1	GLIS3 antisense RNA 1 | chromosome 9 open reading frame 70 | C9orf70 |GLIS3 antisense RNA 1 (non-protein coding)		12477932 | 15489334 |
C9ORF123	TMEM261	C9orf123 | chromosome 9 open reading frame 123 |transmembrane protein C9orf123 | transmembrane protein 261	NP_219500.1	21666724 | 15489334
C10ORF67	C10orf67	Chromosome 10 open reading frame 67 | uncharacterized protein C10orf67	NP_714925.2	18723019 | 15164054
C10ORF84	FAM204A	Protein FAM204A | chromosome 10 open reading frame 84 |bA319I23.1 | C10orf84 | family with sequence similarity 204, member A	NP_001128144.1| NP_071346.1	15489334 | 14702039
C10ORF109 ¶	ADARB2-AS1	ADARB2 antisense RNA 1 | NCRNA00168 | ADARB2 antisense RNA 1 (non-protein coding) | ADARB2 antisense RNA 1 | ADARB2 antisense gene protein 1 | C10orf109 | non-protein coding RNA 168 | chromosome 10 open reading frame 109 |bA466B20.1		16344560 | 15164054
C11ORF44	C11orf44	Uncharacterized protein C11orf44 | chromosome 11 open reading frame 44	NP_001258912.1	14702039 | 23251661 |
C14ORF77	C14orf180	Transmembrane protein C14orf180 |NRAC | chromosome 14 open reading frame 180 | nutritionally-regulated adipose and cardiac enriched protein homolog | C14orf77 | chromosome 14 open reading frame 77 | "nutritionally-regulated adipose and cardiac-enriched"	NP_001008404.1| NP_001273328.1| NP_001273329.1	23029450 | 14702039
C17ORF54 ¶	LINC00469	Long intergenic non-protein coding RNA 469 |chromosome 17 open reading frame 54 | C17orf54		15489334 | 14702039
C20ORF39	SYNDIG1	SynDIG1 | TMEM90B | synapse differentiation inducing 1 | "synapse differentiation induced gene 1" | transmembrane protein 90B | interferon induced transmembrane protein domain containing 5 | Dispanin subfamily C member 2 | C20orf39 | synapse differentiation-inducing gene protein 1 | chromosome 20 open reading frame 39 | dispanin subfamily C member 2 | synapse differentiation induced gene 1 | "interferon induced transmembrane protein domain containing 5" |IFITMD5 | Transmembrane protein 90B | DSPC2	NP_079169.1	22363774 | 20152115

Table I. *continued*

Table I. *continued*

Gene name	Gene symbol	Aliases and descriptions	Refseq_Protein_ID	PubMed_IDs
C20ORF45	GNAS	Guanine nucleotide regulatory protein | adenylate cyclase-stimulating G alpha protein | PHP1A | GSA | NESP | "secretogranin VI" | extra large alphas protein | alternative gene product encoded by XL-exon | protein ALEX | GNAS1 | GPSA | XLalphas | NESP55 | guanine nucleotide binding protein (G protein), alpha stimulating activity polypeptide 1 | AHO | GNAS complex locus |C20orf45 | Alternative gene product encoded by XL-exon | PHP1C | GSP | PHP1B | neuroendocrine secretory protein | Adenylate cyclase-stimulating G alpha protein | POH | secretogranin VI | Extra large alphas protein | guanine nucleotide-binding protein G(s) subunit alpha isoforms XLas	NP_000507.1| NP_001070956.1| NP_001070957.1| NP_001070958.1| NP_057676.1| NP_536350.2| NP_536351.1	| 12621129 | 9860993
C20ORF174	ZNF831	C20orf174 | chromosome 20 open reading frame 174 | dJ492J12.1 | zinc finger protein 831	NP_848552.1	19875103 | 19430479
FAM149A	FAM149A	Family with sequence similarity 149, member A | MSTP119 | protein FAM149A	NP_001006656.1| NP_056213.1	19460752 | 17974005
FAM19A5	FAM19A5	Protein FAM19A5 | chemokine-like protein TAFA-5 | TAFA-5 | TAFA protein 5 | Chemokine-like protein TAFA-5 | TAFA5 | family with sequence similarity 19 (chemokine (C-C motif)-like), member A5 | UNQ5208 |QLLK5208	NP_001076436.1| NP_056196.2	15028294 | 22158540
FAM84B	FAM84B	Breast cancer membrane protein 101 | "breast cancer membrane-associated protein 101" | neurological/sensory 2 | Breast cancer membrane protein 101 | family with sequence similarity 84, member B | breast cancer membrane-associated protein 101 | "neurological/sensory 2" | NSE2 | protein FAM84B | Protein NSE2 | BCMP101	NP_777571.1	20158306 | 19549893
FAM91A1	FAM91A1	Protein FAM91A1 | family with sequence similarity 91, member A1 | skeletal muscle cells re-entry induced	NP_659400.2	19734545 | 16421571
KIAA0232	KIAA0232	Uuncharacterized protein KIAA0232 | KIAA0232	NP_001094060.1| NP_055558.2	9039502 | 22139419
KIAA1217	KIAA1217	SKT |likely orthologue of Mus musculus enhancer trap locus 4 | sickle tail protein homolog | "sickle tail" | KIAA1217	NP_001091970.1| NP_001269696.1| NP_001269697.1| NP_001269698.1| NP_001269699.1| NP_062536.2	10574462 | 20379614
KIAA1549	KIAA1549	KIAA1549 |UPF0606 protein KIAA1549	NP_001158137.1| NP_065961.2	20379614 | 18974108

Uncharacterized proteins with positive genetic association are shown. The gene name, the Human Genome Nomenclature approved symbol, Reference Sequence, RefSeq and Two PubMed IDs are shown. ¶, ncRNAs. Data from GeneALaCart, GeneCards.

Protein classes of the pancreatic neoplasm-associated ORFs. The ORFs were next characterized for druggableness and secreted biomarker potential using diverse protein motif and domain analysis tools including the Meta Analysis tools GeneALaCart, the DAVID functional annotation tool, the protein Family (PFAM) (29), the InterPro (30), the Signal P (31), the SMART domain analysis tool, (32) the Prosite (33) and HPRD (27). A summary of these results is shown in Table III. Putative druggable proteins (enzymes, transporters, receptors, protein binding and transmembrane proteins) and biomarkers (secreted proteins and chemokines), signaling (hormonal, insulin), neuropeptides and transcription factor class was predicted for the ORFs. Five of the ORFs belong to the ncRNA class including long intergenic RNA (linc RNA) and antisense RNAs.

Correlation of gene expression: mRNA versus protein. The ORF expression in pancreatic tumors was investigated using the HPRD (27), the Oncomine Microarray Database (34), the HPA (34), the MOPED (26) and COSMIC database (16) (Table IV). Expression correlation at mRNA and protein levels was seen for 6/21 ORFs. This included up-regulation of gene expression (C1orf94, C7orf10, FAM19A5 and KIAA1217) and down-regulation (C7orf8, C20orf45). Analysis of the COSMIC database revealed a strong correlation across copy number variations and mRNA over-expression for four ORFs (C1orf94,

Table II. *Genome-Wide Association studies of the pancreatic cancer novel ORFs.*

Gene name	Gene symbol	# ICGC Mutations/% donors	cBIOPortal mutations. CDS/% of somatic mutation rate	NCBI ClinVar
C1ORF39 §★	FNBP1L	**5/14.29**		
C1ORF94+§★	C1orf94	12/8.04		
C6ORF155 ¶	LINC00472	3/8.57		
C7ORF4 ¶§★	LINC00244			
C7ORF8	CTTNBP2	**50/10.46**		
C7ORF9 §★	NPVF	3/1.79		
C7ORF10+§★	SUGCT	**569/38.1**		Glutaryl-CoA oxidase deficiency
C9ORF70 ¶	GLIS3-AS1	1/0.89		
C9ORF123+§★	TMEM261	35/6.89		
C10ORF67+§★	C10orf67	40/7.9	H147N/1	Lung cancer
C10ORF84	FAM204A	16/3.57		
C10ORF109¶	ADARB2-AS1	19/4.08		
C11ORF44	C11orf44	18/4.34		Left Ventricular Hypertrophy
C14ORF77	C14orf180	5/1.18		
C17ORF54¶	LINC00469	22/4.59		
C20ORF39	SYNDIG1	**86/15.56**		
C20ORF45§★	GNAS	**14/13.39**	R844C/1	
C20ORF174§★	ZNF831	44/8.93	E1112K/1	
FAM149A	FAM149A	7/6.25		Malignant melanoma
FAM19A5§★	FAM19A5	**24/42.86**		Lung cancer, Melanoma
FAM84B§★	FAM84B	4/2.68		
FAM91A1+§★	FAM91A1	17/3.32		
KIAA0232§★	KIAA0232	23/4.59		Malignant melanoma
KIAA1217§★	KIAA1217	**294/28.46**		Malignant melanoma, Lung cancer
KIAA1549+§★	KIAA1549	36/5.8		Malignant melanoma, Lung cancer

Pancreatic cancer-associated novel proteins were analyzed using the International Cancer Gene Consortium Portal (ICGC). The number of mutations affected in percent of donors is shown. Mutations seen in 10% or above donors are bolded. ¶, NcRNAs; +, mutations verified by Catalogue of Somatic Mutations (COSMIC) databases; §, copy number variations (CNV) and over-expression from the COSMIC database; ★, cBioPortal mutations; CDS, coding domain sequence. The NCBI Clinical Variations (ClinVar) are shown for indicated diseases.

Table III. *Motif and domain analysis of the pancreatic neoplasm-associated novel proteins.*

Putative Class	Open reading frames
Enzyme	C7ORF10
Secreted proteins	C11orf44, C20orf45, FAM19A5
Chemokine	FAM19A5
Nucleotide binding	C20ORF45, C20ORF174, KIAA0232
Protein binding	C1ORF94, C7ORF8, C20ORF39, FAM84B
Neuropeptide	C7ORF9
Transporters	C1ORF39, C20ORF39, C20ORF45
Signaling	C1ORF39, C7ORF9, C20ORF45
Insulin and Glucagon pathway	C20ORF45
Transcription factor	C20ORF174
Receptor binding	C20ORF39, C20ORF45
Transmembrane	C1ORF39, C7ORF9, C9ORF123,C14ORF77, C20ORF39, C20ORF45, FAM19A5, FAM84B, KIAA1549
Non coding RNAs	C6orf155, C7orf4, C9orf170, C10orf109, C17orf54

Protein motifs and domains were identified using diverse bioinformatics tools, as described in the Materials and Methods. The protein classes inferred are shown for the ORFs.

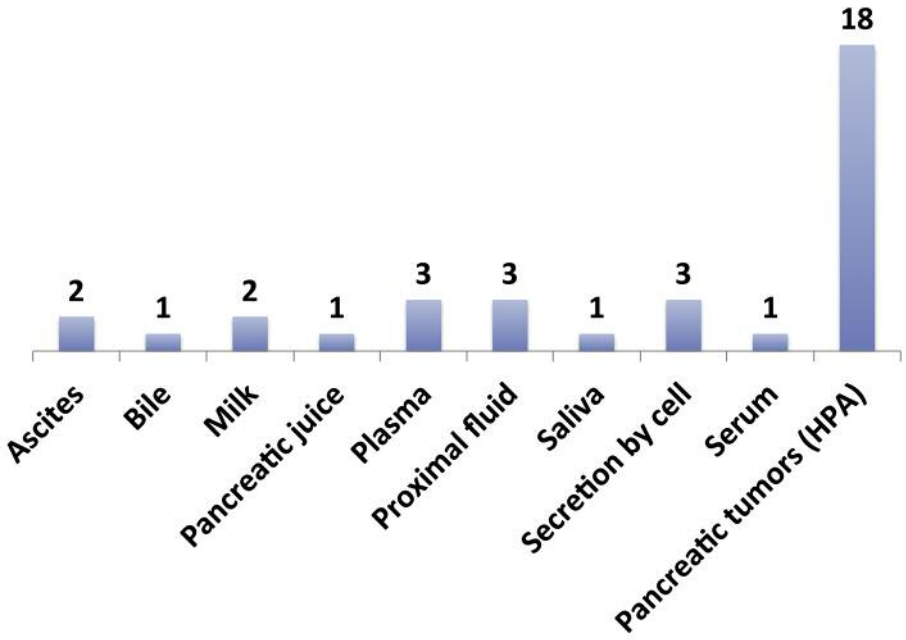

Figure 3. *Protein expression analysis of the pancreatic neoplasm-associated novel proteins. Protein expression in body fluids inferred from the Multi Omics Protein Expression Database, the ProteomesDB and the Human Protein Map is shown. The numbers indicate the number of ORFs for each of the body fluid. The number of ORFs detected in the Human Protein Atlas tissue microarray (34) is shown for comparison.*

C7orf10, C10orf67 and FAM91A1). These results are consistent with several findings that show limited correlation of gene expression between mRNA and protein (35-38).

Pancreatic ORFs in other diseases. The PheGenI tool allowed for identification of pancreatic neoplasm-associated traits in other diseases (Table V). Association was seen for other neoplasms (neuroblastoma, leukemia and lymphoma and prostate cancer); psychiatry (attention deficit disorder, neuropsychological tests); cardiac (coronary artery disease, echocardiography, left ventricular hypertrophy, heart failure); Alzheimer's disease and amyloid beta peptides; metabolic (cholesterol, potassium, ferritin, body mass, height and index, vitamin D, fibrinogen, breath tests) and insulin. These results demonstrate the involvement of the pancreatic neoplasm-associated ORFs in a complex landscape of diseases.

Lead ORFs associated with pancreatic neoplasms. Based on genome-wide association, mutational prevalence and expression correlation in pancreatic tumors, six ORFs emerged as putative lead ORFs (Table VI). This included two secreted products (neuroendocrine secretory protein C20orf45/GNAS and a chemokine, FAM19A5) and an enzyme (C7orf10/SUGCT). Further, two ORFs belonging to protein binding classes (C1orf94 and C7orf8/CTTNB2) might offer druggableness. The KIAA1217, a Sickle tail protein homolog, is a serine- and proline-rich protein involved in actin binding.

Discussion

We have recently embarked on a systematic approach to deciphering the Dark Matter proteome of the human genome for accelerated drug discovery and diagnostic marker development (12). Utilizing GWAS, proteomics and transcriptome approaches, we have identified a fingerprint of cancer and diabetes-associated ORFs (12-15). In the present study, an effort was made to identify novel molecular targets for pancreatic cancer. In view of the relative paucity of druggable targets and diagnostic markers for pancreatic cancer, which is rarely diagnosed at an early stage and has a very high mortality rate, new targets are urgently needed for this cancer (2-4).

Disease target discovery is greatly aided by the availability of Phenome to Genome information mining from the human genome (6). An advantage of target discovery using this approach is the ability to amass the genetic association evidence for the predicted genes. Using the NCBI PheGenI association tool, an initial list of pancreatic neoplasm-associated genes was predicted. Among the 226 genes found to be associated with pancreatic cancer, 196 known proteins (encompassing oncogenes, tumor suppressor genes, growth factors and receptors, enzymes and adhesion molecules), 20 previously uncharacterized protein-coding ORFs and 5 ncRNAs were identified.

Table IV. *Correlation of mRNA and protein expression*.

Gene name	Gene symbol	HPA pancreatic tumors	Oncomine pancreatic tumors	COSMIC pancreatic tumors	mRNA protein correlation
C1ORF39	*FNBP1L*				
C1ORF94	*C1orf94*	Medium to low		Over-expression	Y
C6ORF155 ¶	*LINC00472*				
C7ORF4 ¶	*LINC00244*				
C7ORF8	*CTTNBP2*	Low	Down-regulated		Y
C7ORF9	*NPVF*	Medium to low			
C7ORF10	*SUGCT*	Medium to low	Up-regulated	Over-expression	Y
C9ORF70 ¶	*GLIS3-AS1*				
C9ORF123	*TMEM261*	Medium to low			
C10ORF67	*C10orf67*	High to medium		Over-expression	Y
C10ORF84	*FAM204A*	Medium	Up-regulated		Y
C10ORF109¶	*ADARB2-AS1*				
C11ORF44	*C11orf44*				
C14ORF77	*C14orf180*	Medium to low			
C17ORF54¶	*LINC00469*				
C20ORF39	*SYNDIG1*	Medium to low			
C20ORF45	*GNAS*	Low	Down-regulated		Y
C20ORF174	*ZNF831*				
FAM149A	*FAM149A*	Medium to low			
FAM19A5	*FAM19A5*	High	Up-regulated		
FAM84B	*FAM84B*	Medium to low			
FAM91A1	*FAM91A1*	Medium to low		Over-expression	Y
KIAA0232	*KIAA0232*	High to medium			
KIAA1217	*KIAA1217*	High to medium	Up-regulated		Y
KIAA1549	*KIAA1549*	High to medium		Over-expression	Y

The mRNA expression was inferred from the Oncomine Microarray and the COSMIC datasets. The ORF protein expression as it results from the Human Protein Atlas (34) is shown. Y, positive correlation between mRNA and protein expression.

Table V. *Involvement of the pancreatic neoplasm-associated novel proteins in other diseases*.

Gene name	Associated traits
C1ORF94	Attention Deficit Disorder with Hyperactivity|Coronary Artery Disease|Heart Failure| Potassium|Cholesterol|Body Height|Echocardiography|**Neuroblastoma**
C6ORF155	Amyloid beta-Peptides
C7ORF10	Body Fat Distribution|Body Height|Coronary Artery Disease|Heart Failure|Vitamin D|Neuropsychological Tests| Body Mass Index|Cardiomegaly|Ferritins|Precursor Cell **Lymphoblastic Leukemia-Lymphoma**|Menarche|**Prostatic Neoplasms**
C9ORF70	Stroke
C9ORF123	Cholesterol, HDL|Body Height|Heart Rate|Respiratory Function Tests|Attention Deficit and Disruptive Behavior Disorders|Platelet Aggregation|Braces|**Prostatic Neoplasms**
C10ORF67	Crohn's disease and Sarcoidosis (combined)
C10ORF84	Alzheimer Disease|Tobacco Use Disorder
C11ORF44	Body Height|Hypertrophy, Left Ventricular
C14ORF180	Fibrinogen
C17ORF54	Breath Tests|Insulin

The Genetic Association database (June 2014 version) was used to identify other disease disorder-associated traits. Neoplasm traits are bolded.

Table VI. *Pancreatic cancer lead targets summary.*

Gene name	Gene symbol	mRNA/protein correlation	ICGC mutations %	Protein class of the ORFs
C1ORF94	*C1orf94*	Yes	8.04	Transcription (Regulator)|Topoisomerase II protein binding
C7ORF8	*CTTNB2*	Yes	10.46	Protein binding|protein-protein interaction
C7ORF10	*SUGCT*	Yes	38.1	CoA-transferase
C20ORF45	*GNAS*	Yes	13.39	Neuroendocrine Secreted|GTP Binding
FAM19A5	*FAM19A5*	Yes	42.86	Brain-specific Chemokine|Secreted
KIAA1217	*KIAA1217*	Yes	28.46	Skeletal Development protein|actin binding

The protein class for the six pancreatic cancer leads identified in the study is shown.

A comprehensive profiling of these twenty ORFs enabled for identification of six putative lead ORFs for drug discovery, as well as biomarker development.

The C7orf10, succinylCoA:Glutarate-CoA transferase, is a likely druggable target. Mutations in this gene are associated with glutaric aciduria type III (28, 39).

Inhibition of 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) by fluvastatin and lovastatin has been shown to reduce human pancreatic cancer cell invasion and metastasis (40). The novel association of the C7orf10 with pancreatic neoplasms identified in this study provides a basis for drug discovery. The C7orf10 is also associated with other diseases and traits, including precursor cell leukemia and lymphoma, prostate neoplasm, coronary artery disease and cardiomegaly.

A novel protein family with sequence similarity 19 (chemokine (C-C motif)-like), member A5) also emerged from this study. This gene, *FAM19A5*, is a member of the TAFA family, which encodes small-secreted proteins. These proteins contain conserved cysteine residues at fixed positions and are distantly related to MIP-1alpha, a member of the CC-chemokine family (41). The TAFA proteins are predominantly expressed in specific regions of the brain and are postulated to function as brain-specific chemokines or neurokines that act as regulators of immune and nervous cells. The *FAM19A5* was found to be one of the five loci associated with pancreatic cancer in a Chinese population (42). CC chemokines have been extensively implicated in pancreatic neoplasm (43). In the ICGC database, somatic mutations in pancreatic cancer patients were seen in 42.86% of the cases. Eventual identification of the cognate receptor for the FAM95A5 might lead to a drug therapy potential. Additional targets identified herein, C1orf94 (Topoisimerase II protein binding), C7orf8 (Contactin-Binding Protein 2 (CTTNBP2), C20orf45 (Guanine Nucleotide Binding Protein (G Protein), Alpha Stimulating Activity Polypeptide 1 (GNAS)) and KIAA1217, a skeletal developmental actin binding protein, expand the scope of druggable targets discovered in this study.

In summary, the results presented in this study demonstrate the feasibility of mining the human proteome from genetic association to target discovery for pancreatic neoplasms. The Phenome to Genome approach offers a rational basis for accelerated target discovery for drug therapy and diagnostic use.

Conflicts of Interest

None.

Data Availability

The detailed data of this study as a supplemental table is available upon request.

Acknowledgements

This work was supported, in part, by the Genomics of Cancer Fund, Florida Atlantic University Foundation. I thank Dr. Stein of the GeneCards team for generous permission to use the powerful GeneALaCart tool; Dr. Montague, Kolker Laboratory of the MOPED Team for batch analysis of the ORFs and the Human Proteome Map and ProteomesDB for the datasets. I thank Jeanine Narayanan for editorial assistance.

References

1 Warsame R and Grothey A: Treatment options for advanced pancreatic cancer: a review. Expert review of anticancer therapy *12(10)*: 1327-1336, 2012.

2 Paulson AS, Tran Cao HS, Tempero MA and Lowy AM: Therapeutic advances in pancreatic cancer. Gastroenterology *144(6)*:1316-1326, 2013.

3 Valsecchi ME, Diaz-Canton E, de la Vega M and Littman SJ: Recent treatment advances and novel therapies in pancreas cancer: a review. Journal of gastrointestinal cancer *45(2)*: 190-201, 2014.

4 Wolfgang CL, Herman JM, Laheru DA, Klein AP, Erdek MA, Fishman EK and Hruban RH: Recent progress in pancreatic cancer. CA: A Cancer Journal for Clinicians 63: 318-348, 2013.

5 Zhang Y, De S, Garner JR, Smith K, Wang SA and Becker KG: Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. BMC medical genomics *3*: 1, 2010.

6 Ramos EM, D Hoffman, HA Junkins, D Maglott, L Phan, ST Sherry, M Feolo and LA Hindorff: Phenotype-Genotype Integrator (PheGenI): Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet 22(1): 144-147, 2014.

7 Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H and Pandey A: A draft map of the human proteome. Nature *509(7502)*: 575-578, 2014.

8 Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F and Kuster B: Mass-spectrometry-based draft of the human proteome. Nature *509(7502)*: 582-587, 2014.

9 Martin L and Chang HY: Uncovering the role of genomic "dark matter" in human disease. The Journal of clinical investigation *122(5)*: 1589-1595, 2012.

10 Babcock JJ and Li M: Deorphanizing the human transmembrane genome: A landscape of uncharacterized membrane proteins. Acta pharmacologica Sinica 2013.

11 Blaxter M: Genetics. Revealing the dark matter of the genome. Science *330(6012)*: 1758-1759, 2010.

12 Delgado AP BP, Hamid S and Narayanan R: Mining the dark matter of the cancer proteome for novel biomarkers. Current Cancer Therapy Reviews *9(4)*: 265-277, 2013.

13 Delgado A, Brandao P, Chapado M, Hamid S and Narayanan R: Open reading frames associated with cancer in the dark matter of the human genome. Cancer genomics & proteomics *11(4)*: 201-213, 2014a.

14 Delgado A, Hamid S, Brandao P and Narayanan R: A novel transmembrane glycoprotein cancer biomarker present in the x chromosome. Cancer genomics & proteomics *11(2)*: 81-92, 2014b.

15 Delgado A, Brandao P and Narayanan R: Diabetes Associated Genes from the Dark Matter of the Human Proteome. MOJ Proteomics Bioinform *1(4)*: 00020, 2014c.

16 Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A, Teague JW, Campbell PJ, Stratton MR and Futrel PA: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research *39(Database issue)*: D945-950, 2011.

17 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C and Schultz N:The cBio cancer genomics portal: an open platform for exploring multi-dimensional cancer genomics data. Cancer discovery *2(5)*: 401-404, 2012.

18 Bulusu KC, Tym JE, Coker EA, Schierz AC and Al-Lazikani B: canSAR: updated cancer research and drug discovery knowledgebase. Nucleic acids research *42(Database issue)*: D1040-7, 2014.

19 Huang DW, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols *4(1)*: 44-57, 2008.

20 Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A and Lancet D: GeneCards Version 3: the human gene integrator. Database 2010; 2010.

21 Consortium GT: The Genotype-Tissue Expression (GTEx) project. Nature genetics *45(6)*: 580-585, 2013.

22 MacDonald JR, Ziman R, Yuen RK, Feuk L and Scherer SW: The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic acids research *42(Database issue)*: D986-992, 2014.

23 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM and Maglott DR: ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research *42(Database issue)*: D980-985, 2014.

24 Rhodes DR, S Kalyana-Sundaram, V Mahavisno, R Varambally, J Yu, BB Briggs, TR Barrette, MJ Anstet, C Kincead-Beal, P Kulkarni, S Varambally, D Ghosh and Chinnaiyan AM: Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia *2*: 166-80, 2007.

25 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Bjorling L and Ponten F: Towards a knowledge-based Human Protein Atlas. Nature biotechnology *28(12)*: 1248-1250, 2010.

26 Kolker E, Higdon R, Haynes W, Welch D, Broomall W, Lancet D, Stanberry L and Kolker N: MOPED: Model Organism Protein Expression Database. Nucleic acids research *40(Database issue)*: D1093-1099, 2012.

27 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R and Pande A: Human Protein Reference Database – 2009 update. Nucleic acids research *37(Database issue)*: D767-72, 2009.

28 Marlaire S, Van Schaftingen E and Veiga-da-Cunha M: C7orf10 encodes succinate-hydroxymethylglutarate CoA-transferase, the enzyme that converts glutarate to glutaryl-CoA. Journal of inherited metabolic disease *37(1):* 13-19, 2014.

29 Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A and Finn RD: The Pfam protein families database. Nucleic acids research *40(Database issue)*: D290-301, 2012.

30 Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF,

Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C and Yong SY: InterPro in 2011: new developments in the family and domain prediction database. Nucleic acids research *40(Database issue)*: D306-12, 2012.

31 Petersen TN, Brunak S, von Heijne G and Nielsen H: SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods *8(10)*: 785-786, 2011.

32 Letunic I, Doerks T and Bork P: SMART: recent updates, new developments and status in 2015. Nucleic acids research 2014.

33 Sigrist CJ, E de Castro, L Cerutti, BA Cuche, N Hulo, A Bridge, L Bougueleret and Xenarios I: New and continuing developments at PROSITE. Nucleic acids research *41(Database issue)*: D344-347, 2013.

34 Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A and Chinnaiyan AM: ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia *6(1)*: 1-6, 2004.

35 Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC, N.C. the, C. National Cancer Institute Clinical Proteomics Tumor Analysis, N.C. National Cancer Institute Clinical Proteomics Tumor Analysis Consortium and N.C. National Cancer Institute Clinical Proteomics Tumor Analysis Consortium: Proteogenomic characterization of human colon and rectal cancer. Nature *513*: 382-393, 2014.

36 Vogel C and Marcotte EM: Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature reviews Genetics *13(4)*: 227-232, 201.

37 Greenbaum D, Colangelo C, Williams K and Gerstein M: Comparing protein abundance and mRNA expression levels on a genomic scale. Genome biology *4(9)*: 117, 2003.

38 Narayanan R and Van De Ven WJ: Transcriptome and Proteome Analysis: A Perspective on Correlation. MOJ Proteomics Bioinform *1(5)*: 00027, 2014.

39 Sherman, EA, KA Strauss, S Tortorelli, Bennett MJ, Knerr I, Morton DH and Puffenberger EG EG: Genetic mapping of glutaric aciduria, type 3, to chromosome 7 and identification of mutations in c7orf10. American journal of human genetics *83(5)*: 604-609, 2008.

40 Kusama T, Mukai M, Iwasaki T, Tatsuta M, Matsumoto Y, Akedo H, Inoue M and Nakamura H: 3-hydroxy-3-methylglutaryl-coenzyme a reductase inhibitors reduce human pancreatic cancer cell invasion and metastasis. Gastroenterology *122(2)*: 308-317, 2002.

41 Tom Tang Y, P Emtage, WD Funk, T Hu, M Arterburn, EE Park and Rupp F: TAFA: a novel secreted family with conserved cysteine residues and restricted expression in the brain. Genomics *83(4)*: 727-734, 2004.

42 Wu C, Miao X, Huang L, Che X, Jiang G, Yu D, Yang X, Cao G, Hu Z, Zhou Y, Zuo C, Wang C, Zhang X, Zhou Y, Yu X, Dai W, Li Z, Shen H, Liu L, Chen Y, Zhang S, Wang X, Zhai K, Chang J, Liu Y, Sun M, Cao W, Gao J, Ma Y, Zheng X, Cheung ST, Jia Y, Xu J, Tan W, Zhao P, Wu T, Wang C and Lin D: Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. Nature genetics *44(1)*: 62-66, 2012.

43 Wormann SM, Diakopoulos KN, Lesina M and Algul H: The immune network in pancreatic cancer development and progression. Oncogene *33(23)*: 2956-2967, 2014.

Received November 8, 2014
Revised November 13, 2014
Accepted November 17, 2014