A Novel Transmembrane Glycoprotein Cancer Biomarker Present in the X Chromosome

ANA PAULA DELGADO, SHEILIN HAMID, PAMELA BRANDAO and RAMASWAMY NARAYANAN

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

Abstract. *Background: The uncharacterized proteins of the human proteome offer an untapped potential for cancer biomarker discovery. Numerous predicted open reading frames (ORFs) are present in diverse chromosomes. The mRNA and protein expression data, as well as the mutational and variant information for these ORF proteins are available in the cancer-related bioinformatics databases. Materials and Methods: ORF proteins were mined using bioinformatics and proteomic tools to predict motifs and domains, and cancer relevance was established using cancer genome, transcriptome and proteome analysis tools. Results: A novel testis-restricted ORF protein present in chromosome X called CXorf66 was detected in the serum, plasma and neutrophils. This gene is termed secreted glycoprotein in chromosome X (SGPX). The SGPX gene is up-regulated in cancer of the brain, lung and in leukemia, and down-regulated in liver and prostate cancer. Brain cancer in female patients exhibited elevated copy numbers of the SGPX gene. Conclusion: The SGPX gene is a putative novel cancer biomarker. Our results demonstrate the feasibility of mining the 'dark matter' of the cancer proteome for rapid cancer biomarker discovery.*

The human genome project is an attractive starting point for cancer gene discovery (1-3). Numerous drug targets and biomarkers have emerged from the Genome Project (3-7). Expression specificity provides a strong rationale for finding targets that can lead to highly selective and less toxic therapeutics (6-8). An attractive area for mining the human genome resides in the uncharacterized proteome (9). Currently, over half of the predicted proteins in the human genome are of unknown nature (10). These proteins and the non-coding RNAs are called the 'dark matter' of the genome (11-13). Whereas in the past most gene discovery has revolved around known genes due to the ease of follow-up studies (14-17), the dark matter of the genome offers an untapped potential (18). Realizing the importance of this area, the US National Cancer Institute has recently announced a major initiative called illuminating the dark matter for druggable targets (http://commonfund.nih.gov/idg/index).

Establishing cancer relevance for novel or uncharacterized proteins is a crucial first step in lead discovery. The cancer genomes of patients from around the world can be readily mined using databases such as the cBioPortal (19), canSar (20), the Catalogue of Somatic Mutations in Cancer, COSMIC (21), The Cancer Gene Atlas, TCGA (http://cancergenome.nih.gov) and the International Cancer Genome Consortium, ICGC (https://www.icgc.org).

The availability of microarray databases such as Oncomine (22) and ArrayExpress (23), and protein expression analysis tools including the Human Protein Atlas (24), the Human Protein Reference Database, HPRD (25), and the Model Organism Protein Expression database, MOPED (26) can facilitate the cancer target verification.

Numerous proteomic analysis tools including Expasy (http://www.expasy.org), PredictProtein (27) and MESSA Meta analysis server (28) are available for predicting the putative motifs and domains of the novel proteins. These tools can be used to mine the dark matter of the proteome to identify motifs with biomarker and druggable potential (signal peptides, receptors, transporters and enzyme) signatures (29-31).

In the present report, we demonstrate the feasibility of discovering novel biomarkers from a database of cancer-related uncharacterized proteins we have recently developed (18). An X-chromosome specific ORF, CXorf66, was rapidly validated for secreted nature using the protein expression databases. Detailed bioinformatics and proteomics characterization of the *CXorf66* gene confirmed the cancer biomarker potential of this gene. Our results support the potential of the uncharacterized proteome to be tapped for cancer biomarker discovery.

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: +1 561 2972247, Fax: +1 5612973859, e-mail: rnarayan@fau.edu

Key Words: Signal peptide, ORF, 'dark matter' of the genome, serum protein, uncharacterized proteins, biomarkers, X-chromosome, cell trafficking, vesicular transport, secreted protein.

Materials and Methods

The bioinformatics tools used in the study are shown in Table I. All the bioinformatics mining was verified by two independent experiments. Only statistically significant results per each tool's requirement are reported. Prior to using a bioinformatics tool, a series of control query sequences were tested to evaluate the predicted outcome of the results.

Results

We recently established a database of expression verified uncharacterized ORFs by mining the cancer proteome (18). In that study, using a streamlined approach involving gene expression, protein motifs and domains analysis and Genome-Wide Analysis Studies (GWAS) analysis, we identified a novel cancer biomarker called carcinoma related EF-hand protein (CREF). The *CREF* gene (C1orf87) is a calcium-binding protein specific to breast, lung and liver cancer. These results supported our premise that it is possible to harness the dark matter of the cancer proteome systematically for cancer drug target and biomarker discovery. Reasoning that prediction of signal peptide motifs in these uncharacterized ORFs may lead to novel cancer diagnostic marker discovery, we have undertaken bioinformatics mining of the hits from the database. Preliminary experiments using the signal P tool (http://www.cbs.dtu.dk/services/SignalP/) identified an ORF, CXorf66, which may harbor a putative signal peptide sequence. Encouraged by this finding, we undertook a comprehensive bioinformatics and proteomics analysis of the *CXorf66* gene.

CXorf66 expression in normal tissues. Initially, protein expression analysis tools were used to verify the secreted nature of the *CXorf66* gene. Protein expression for the *CXorf66* gene was detected in the serum and testis using the HPRD (http://www.hprd.org) (Figure 1A). The GeneCards (http://www.genecards.org) summary of protein expression databases showed the presence of Secreted glycoprotein in chromosome X (SGPX) protein in hematopoietic tissues plasma and platelets (Figure 1B).

Additional evidence for expression of the CXorf66 protein in normal tissues was obtained from the Human Protein Atlas (http://www.proteinatlas.org). In tissue microarray sections, the CXorf66 protein was detected at a medium expression level in 7 out of 77 analyzed normal tissue cell types. Major normal tissues included salivary gland, spleen, lymph node, kidney and tonsil. A subset of leukocytes scattered throughout most tissues showed strong cytoplasmic positivity. Most remaining normal tissues were negative. Immunohistochemical (IHC) staining of normal testis with the antibody HPA048517 showed membranous/cytoplasmic staining (Figure 1C). Predominant staining was observed in the cells in seminiferous ducts (65%). CXorf66 protein expression in normal tissues was further correlated with mRNA expression. The Unigene EST expression tool (http://www.ncbi.nlm.nih.gov/unigene) indicated a restricted expression in testis. Developmentally, CXorf66 expression was detected in the fetus but not in adult tissues.

CXorf66 expression in tumors. We next investigated the CXorf66 expression in diverse tumors. The NextBio meta analysis tool (www.nextbio.com) indicated that the CXorf66 gene is up-regulated in brain and lung cancer, and in lymphoid leukemia (Figure 2A). In contrast, down-regulation of CXorf66 gene expression was seen in liver and prostate carcinomas. Uterine cancer association (most highly correlated) with the *CXorf66* gene was seen only at the somatic mutation level. Expression of the CXorf66 protein was seen in about 4% of tumors analyzed by tissue microarrays in the Human Protein Atlas tool (see Figure 2B). Elevated expression of the CXorf66 protein was seen in carcinoids (colon), lung, ovarian and urothelial cancers. An expanded view of the carcinoid IHC (Figure 2C) demonstrates a strong cytoplasmic and membranous staining. The current protein expression data for CXorf66 is available for a very limited set of patient samples. Additional verification is needed. The CGAP short SAGE tag (sTTTCAAGCAA) analysis of the CGAP tissue libraries (http://cgap.nci.nih.gov) showed down-regulation of the *CXorf66* mRNA in liver and prostate carcinomas, as well as up-regulation in brain and lung cancer. These results were consistent with the NextBio Meta analysis (Figure 2). Analysis of the NCI60 cancer cell lines for *CXorf66* mRNA expression using the NCI Developmental Therapeutics Molecular target database DTP (http://dtp.nci.nih.gov) indicated *CXorf66* mRNA expression in non-small cell lung carcinoma (NCI-H322M, NCI-H460, NCI-H52), breast (MDA-MBA-231, HS578T, BT-549 and T47D), CNS cancer (SF-268, SF-295, SF-539, SNB-19 and SNB-75) and ovarian carcinoma-derived (OVCAR 4) cell lines. We next performed an Oncomine microarray analysis (https://www.oncomine.org) for the mRNA expression of the *CXorf66* gene. The CXorf66 copy number was significantly elevated in anaplastic astrocytomas, anaplastic oligodendrogliomas and in primary and secondary glioblastomas (Figure 3A). Significant differences in the CXorf66 gene copy number were seen between males and females in these tumor types (Figure 3B).

Characterization of the SGPX gene. The molecular characterization of the *CXorf66* gene is shown in Table II. The *CXorf66* gene is present on chromosome X 27.1 and codes for an ORF of 361 amino acids (39944 Da). NCBI-AceView (http://www.ncbi.nlm.nih.gov/ieb/research/acembly/) predicted one primary transcript with three exons spread over 9,762 bp. This gene is present in the common ancestor of human and mouse. According to NCBI-

Table I. *Bioinformatics tools used in the study. Diverse bioinformatics tools used in the study to characterize the CXorf66 gene are shown.*

Bioinformatics tools used in the study	URL
Genome analysis	
UCSC Genome Browser	http://genome.ucsc.edu/
NCBI Gene	http://www.ncbi.nlm.nih.gov/gene/
NCBI Aceview	http://www.ncbi.nlm.nih.gov/ieb/research/acembl/
The Sanger Institute Catalogue Of Somatic Mutations In Cancer COSMIC	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/
cBioPortal	http://www.cbioportal.org/public-portal/
The CanSar	https://cansar.icr.ac.uk/
International Cancer Genome Consortium, ICGC	https://www.icgc.org/
The Roche Cancer Genome Database, Mutome DB	http://rcgdb.bioinf.uni-sb.de/MutomeWeb/
EBI-EMBL	http://www.ebi.ac.uk/
Transcriptome analysis	
NCBI-UniGene	http://www.ncbi.nlm.nih.gov/unigene
SAGE Digital Anatomical Viewer	http://cgap.nci.nih.gov/SAGE/AnatomicViewer
Cancer Genome Anatomy Project, CGAP	http://cgap.nci.nih.gov/
Oncomine microarray analysis tool	https://www.oncomine.org/resource/login.html
The Array express	https://www.ebi.ac.uk/arrayexpress/
The gene expression Omnibus, GEO	http://www.ncbi.nlm.nih.gov/geo/
Gene Indices from the Dana Farber Cancer Institute	http://compbio.dfci.harvard.edu/tgi/
Proteome analysis	
UniProt Knowledge base, UniProtKB	http://www.uniprot.org/
Swiss Expasy server	http://www.expasy.org/
Protein Database, PDB	http://www.rcsb.org/pdb/home/home.do
Post translational modification sites at Expasy	http://www.expasy.org/proteomics/post-translational_modification
PredictProtein	https://www.predictprotein.org/
MEta Sever for Sequence Analysis, MESSA	http://prodata.swmed.edu/MESSA/MESSA.cgi
I -TASSER server	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
Motif and domain analysis	
NCBI Conserved domain database, CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml/
The PFAM	http://pfam.sanger.ac.uk/
The ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
InterProscan4	http://www.ebi.ac.uk/Tools/pfa/iprscan/
HMMER	http://hmmer.janelia.org/
Signal P	http://www.cbs.dtu.dk/services/SignalP/
Prints	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php
Eukaryotic Linear Motif prediction, ELM	http://elm.eu.org/
Protein expression analysis	
The Human Protein Atlas, HPA	http://www.proteinatlas.org/
The Model Organism Proteins Expression database, MOPED	https://www.proteinspire.org/MOPED/mopedviews/
The Human Protein Reference Database, HPRD	proteinExpressionDatabase.jsf http://www.hprd.org/
Knowledge-based datamining	
GeneCards	http://www.genecards.org/
GeneAtlas	http://genatlas.medecine.univ-paris5.fr/
NextBio meta analysis tool	http://www.nextbio.com/b/nextbioCorp.nb
MalaCards	http://www.malacards.org/
On line Mendelian Inheritance in Man, OMIM	http://www.omim.org/
Human Genome Nomenclature Committee, HGNC	http://www.genenames.org/
Gene Ontology, amiGO	http://www.geneontology.org/
NCBI SNP database	http://www.ncbi.nlm.nih.gov/projects/SNP/
Ensembl	http://www.ensembl.org/index.html
The Strings Interactome	http://string-db.org/
BioGrid	http://thebiogrid.org/
IntAct	http://www.ebi.ac.uk/intact/
Source database	http://smd.princeton.edu/cgi-bin/source/sourceResult
mirDB	http://mirdb.org/miRDB/

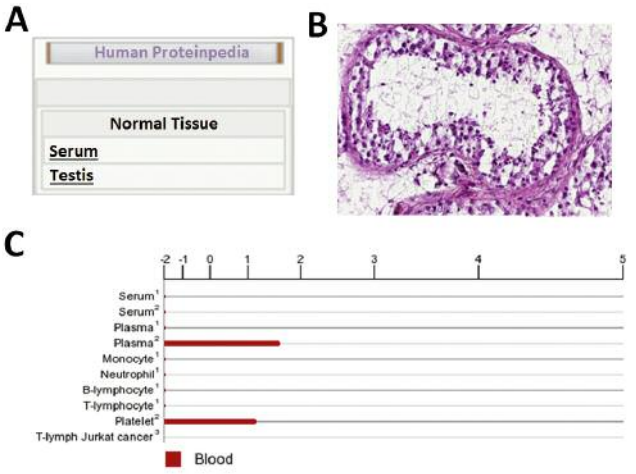


Figure 1. *Expression profile of CXorf66/Secreted glycoprotein in chromosome X (SGPX) gene in normal tissues. Protein expression is shown for the CXorf66 gene as analyzed by Human Protein Reference database, HPRD (A); Human Protein Atlas, Immunohistochemistry (IHC), antibody HPA048517, testis (B); and Model Organism Protein Expression database, MOPED (C).*

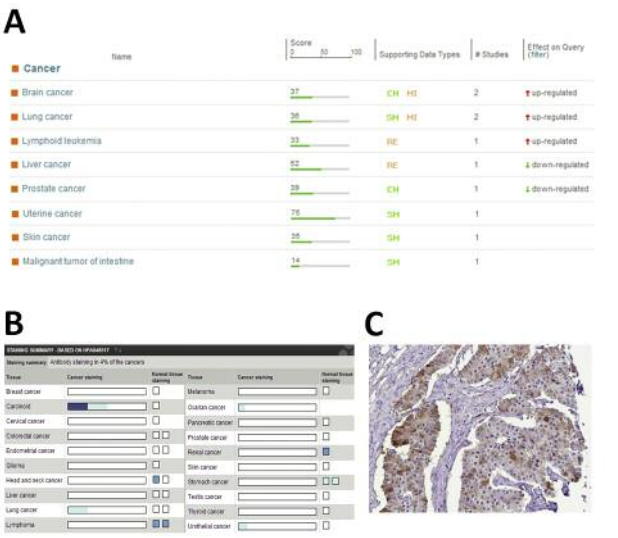


Figure 2. *Tumor expression profile of the CXorf66 Secreted glycoprotein in chromosome X (SGPX) gene. The mRNA expression profile of CXorf66 in tumors is shown. A: NextBio Meta-analysis tool. Score, correlation score; Data type, SM: somatic mutations, CN: copy number variation, RE: RNA expression, and MI: miRNA. B: The protein expression from the Human Protein Atlas, antibody HPA048517 (dark blue: high; blue: medium; light blue: low; white: not detected). C: Expression of CXorf66 protein in colon carcinoma (carcinoid) patient tissue, antibody ID HPA048517 (patient ID. 3043).*

homologue Gene (http://www.ncbi.nlm.nih.gov/homologene), orthologs include *Pan troglodytes, Mus musculus, Rattus norvegicus, Bos taurus* and *Canis familiaris*.

Putative transcription factor binding sites for the *SGPX* gene included Transcriptional repressor CTCF, CCAAT/enhancer-binding protein beta (CEBPB), Myc proto-oncogene protein (MYC), DNA-directed RNA polymerase II subunit RPB1 (POLR2A), Transcription factor E2F1, Proto-oncogene c-Fos and Transcription factor Sp1 (UCSC browser, http://genome.ucsc.edu). In fetal mouse stem cells, overexpression of *cMYC* caused down-regulation of the *SGPX* mRNA, suggesting a role of cMYC in the transcriptional regulation of the SGPX gene (NextBio, data not shown). The NexBio Meta analysis revealed three miRs that are implicated in the regulation of the *SGPX* gene. The most highly correlated miRs by Meta-analysis included hsa-miR-130b/a (sarcoidosis, lung, glioblastoma), hsa-130a (lung cancer) and hsa-130b (glioblastoma). An additional miR, hsa-miR-1290, was predicted by GeneCards.

Characterization of the CXorf66 protein. Using a streamlined approach that we recently developed to characterize the novel ORFs (18), a detailed motifs and domain analysis of the CXorf66 protein was undertaken (Table III). The UniProtKB database (http://www.uniprot.org) analysis showed that CXorf66 (Uniprot id Q5JRM2) is a single-pass type-I transmembrane protein with a signal peptide (amino acids 1-19). Topologically distinct domains, extracellular (amino acids 20-47), transmembrane helical (amino acids 48-68), cytoplasmic (amino acids 69-361) and serine rich (amino acids 91-177) sites were detected in the CXorf66 protein. The presence of signal peptide was further verified using the Signal P prediction tool for eukaryotic network (http://www.cbs.dtu.dk/services/SignalP/). This tool predicted a signal peptide sequence MNLVICVLLLSIWKNNCMT with most likely cleavage site between pos. 19 and 20: CMT-TN (Y-score 0.476 at amino acids 18). Secretome P (http://www.cbs.dtu.dk/services/SecretomeP/) analysis of the CXorf66 showed that it is not secreted by the non-classical secretary pathway (data not shown). These results, together with the protein expression in the plasma and serum (Figure 1), supported the premise that CXorf66 is a secreted transmembrane protein. Hence, CXorf66 was named secreted glycoprotein in chromosome X (SGPX).

To further characterize the SGPX protein's nature, we used diverse motif and domain analysis tools (Table III). The NCBI CDD analysis identified a superfamily (DUF 936) present in several hypothetical proteins from Arabidopsis (e value: 5.97e-03). The MESSA analysis tool predicted an additional conserved domain in the SGPX protein (KOGO566, inositol-1,4,5-triphosphatase-synaptojanin, INP51/INP52,INP53 family) with an e value of 7e-04. This family of proteins is involved in intra-cellular trafficking, secretion and vesicular transport (32).

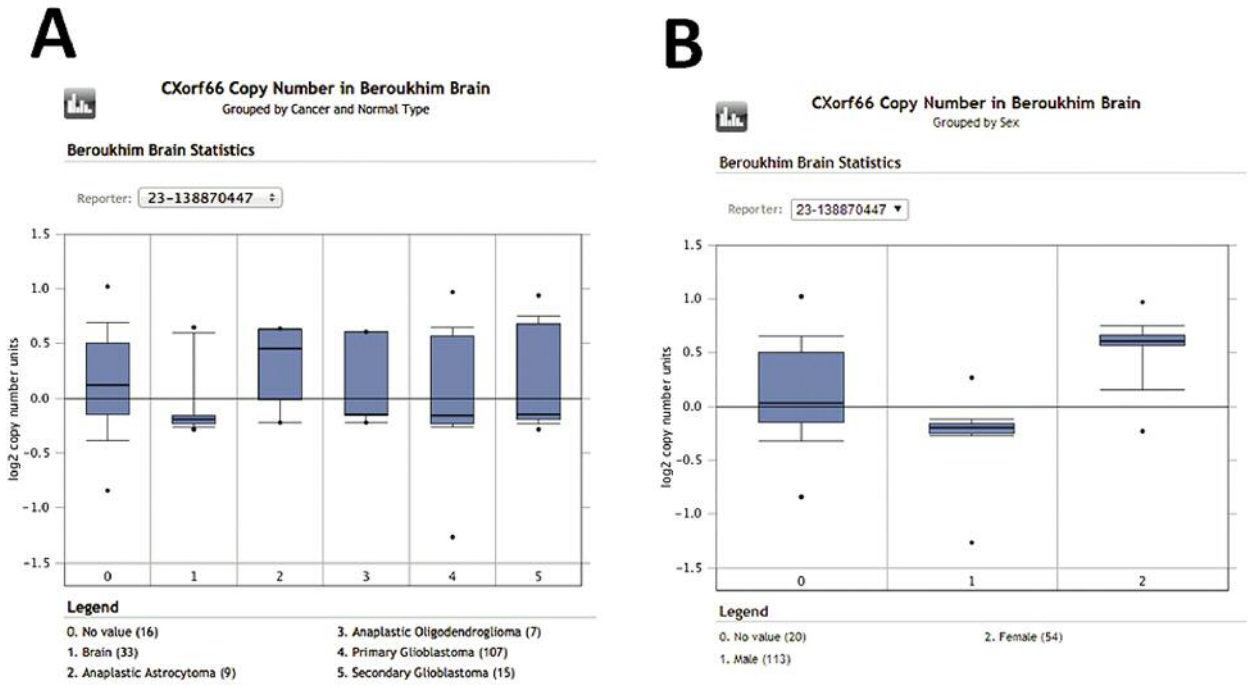


Figure 3. *CXorf66/ Secreted glycoprotein in chromosome X (SGPX) microarray Box Plot analysis. A: Copy numbers in normal brain (0-no value, 1, normal brain) vs. subtypes of brain tumors (2, Anaplastic Astrocytoma; 3, anaplastic Oiligodendroglioma; 4, primary Glioblastoma and 5, secondary Glioblastoma) are shown as indicated. B: Panel A data analyzed by sex difference (males vs. females). 0, control, no value; 1, males and 2, females. The number of independent samples analyzed in the study are shown in parentheses. Log2 copy number is shown for each tissue sample. p-Value-1E-4. Fold change>2-fold. Normalized copy number values including the maximum and minimum are shown as dots. The 90th and 10th percentiles are shown as bars. The center of the box plot shows the median values.*

The PFAM tool (http://pfam.sanger.ac.uk) identified the FAM 163 family signature, present in neuroblastoma-derived secretory protein (NDSP). This signature was further verified by Motif (http://www.genome.jp). The NDSP is highly expressed in neuroblastoma compared to other tissues, suggesting that it may be useful as a marker for metastasis in bone marrow (33). The HMMER tool (http:// hmmer.janelia.org) further verified the transmembrane and signal peptide domains. In addition, a ribosomal protein S2 PFAM domain was predicted at amino acids 18-113 in the SGPX protein.

The PRODOM domain analysis (http://prodom.prabi.fr/) identified two distinct domains, PDA8v7u5 (amino acids1-149, e value: 4e-48) and PDA3C6G2 (amino acids 186-356, e value: 1e-86). These two domains further verified that the SGPX is a transmembrane glycoprotein and suggested that the full-length protein may be a precursor to the secreted product.

The secreted nature of the SGPX protein was further verified using the Predict Protein meta-analysis tool. The sub-cellular localization for the eukarya domain was predicted as secreted (GO term ID: GO: 0005576, prediction confidence 77%). Three protein binding sites were identified (amino acids 184, 270 and 278) in the SGPX protein using the profisis (ISIS), a machine learning-based method (34). The secondary structure of the SGPX protein was classified as mixed.

The nature of the post-translational modification site was next investigated using diverse proteomic tools from the Swiss Expasy server http://www.expasy.org). The Prosite (http://prosite.expasy.org) identified a serine-rich motif. The SGPX protein is modified by phosphorylation at serine (114, 165) and theronine (169) as predicted by GeneCards. Three phosphorylation sites (Ser:41, Thr:2 and Tyr:5) were predicted by the NetPhos 2.0 server (http://www.cbs.dtu.dk/services/NetPhos/). Furthermore, a protein kinase C-specific protein phosphorylation site was predicted by NetPhosK 1.0 server (http://www.cbs.dtu.dk/services/NetPhosK/). The Myristylator tool (http://web.expasy.org/myristoylator/) indicated that the SGPX is not myristoylated. SGPX is, on the other hand, glycosylated (O-linked, at amino acids 92 and 94) and N-linked, (amino acids 24, NGSS) according to the NetoGlyc tool (http://www.cbs.dtu.dk/services/ NetNGlyc/).

We next performed 3-D modeling for the SGPX protein. The UCSC genome browser analysis of the *CXorf66* gene identified a protein model template in the Modbase comparative 3-D structural database. A 17% identity with tyrosine-protein phosphatase, auxilin (PDB code, 3n0a, e

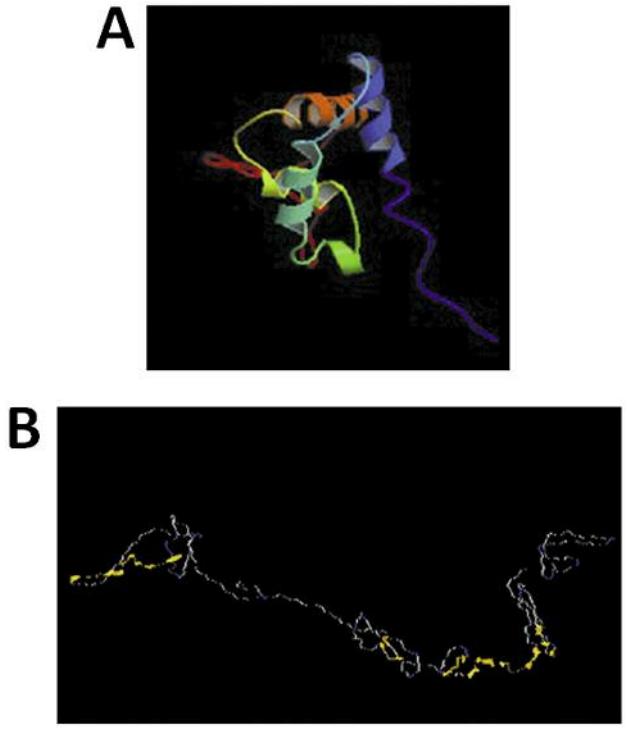


Figure 4. *CXorf66/ Secreted glycoprotein in chromosome X (SGPX) molecular modeling. A: CXorf66 3D modeling data from the UCSC genome browser. Template, auxilin (PDB code, 3n0a, e value, 0, reliable model). B: The I-TASSER model. Template Properdin, high TM score (0.77). Normalized Z- score, 3.40 (>1, good alignment).*

value, 0, reliable model) was detected (Figure 4A). Auxilin, a J-domain containing protein, is involved in the recruitment of the Hsc70 uncoating ATPase to newly-budded clathrin-coated vesicles (35). The top template used by I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) to generate the 3-D model was properdin, a glycoprotein (Figure B).

The Meta Server for Sequence Analysis (MESSA) and the Predict Protein tools were used to further characterize the structure of the SGPX protein. A disordered region lacking a stable tertiary structure was predicted using both of these Meta structural tools. Secondary structures included helical (amino acids 11-16 and 35-67) and strand (amino acids 3-10).

GWAS studies on the SGPX gene. The dbSNP database showed 230 variants with one missense variant (amino acids 233, P to L, Haploid frequency: p=67.9 and L=32.1). One deletion variant esv2672915 is present (36). To further establish a strong correlation of the *SGPX* gene with the cancer genome from different patients, we next performed GWAS analysis using multiple cancer genome analysis tools (Figure 5). Mining the COSMIC somatic mutations database (http://cancer.sanger.ac.uk) indicated that the *SGPX* mutations are largely missense and non-sense (Figure 5A). The cBioPortal (http://www.cbioportal.org/) identified in tumors SGPX gene amplifications (prostate, lung and gliomas), deletions (sarcomas), and mutations (lung, gliomas and pancreas) (Figure 5B). One frameshift insertion mutation (p.H273fs*8) was seen in a mucinous colon adenocarcinoma patient (TCGA-AA-A01R-01). Current COSMIC mutations largely include hematopoietic, lung, CNS, ovary, uterus, gastric, melanoma and kidney tumors.

We next performed a comprehensive mutational analysis for the *SGPX* gene using the ICGC world-wide cancer genome database. The data were compared with CanSar and COSMIC and common mutations were manually curated. Table IV shows the compilation of the current mutations from these databases. Predominant mutations centered around the lung and uterus.

CXorf66/SGPX-interacting proteins. Understanding the nature of the proteins that interact with an uncharacterized protein is critical to deciphering its putative pathways and mechanisms. Hence, the String interactome tool (http://string-db.org) was used to identify putative protein partners for the *SGPX* gene (Figure 6). Numerous Sperm protein associated with the nucleus on the X chromosome A (SPANX) family members are predicted to interact with the SGPX protein. The *SPANX* gene family members are located on the X chromosome. This gene family encodes proteins that play a role in spermiogenesis (37). These proteins represent a specific subgroup of cancer/testis-associated antigens (37). The SPANX family members are associated with prostate cancer (38). The involvement of the SPANX family of proteins in the SGPX pathway was also verified by co-expression analysis using the Oncomine Microarray database in breast tumors (data not shown).

Another predicted protein partner for SGPX interaction was FUN14 domain-containing protein-2 (FUNDC2), also known as cervical carcinoma oncogene 3. The FUNDC2 protein is involved in hepatitis C and cervical cancer (39). This gene also is present on chromosome X. Additional X chromosome-specific genes predicted to interact with the SGPX protein include paraneoplastic antigen-like (PNMA6B) and testis expressed 28 (TEX28P1) both of which are pseudogenes. In addition, ncRNAs were also predicted to interact with the SGPX protein.

Regulation of the SGPX gene. Regulation of gene expression occurs at different levels including promoter methylation, transcription factors, cell cycle and the ncRNAs (40-44). Hence we have attempted to develop an understanding of the SGPX gene regulation using the NextBio Meta-analysis tool. Two miRs (hsa-130a and hsa-130b) were found to be most highly correlated with the SGPX mRNA expression. The hsa-130a was up-regulated (16-fold) in glioblastoma (Bioset: Glioblastoma multiform WHO grade 4 *vs*. normal brain

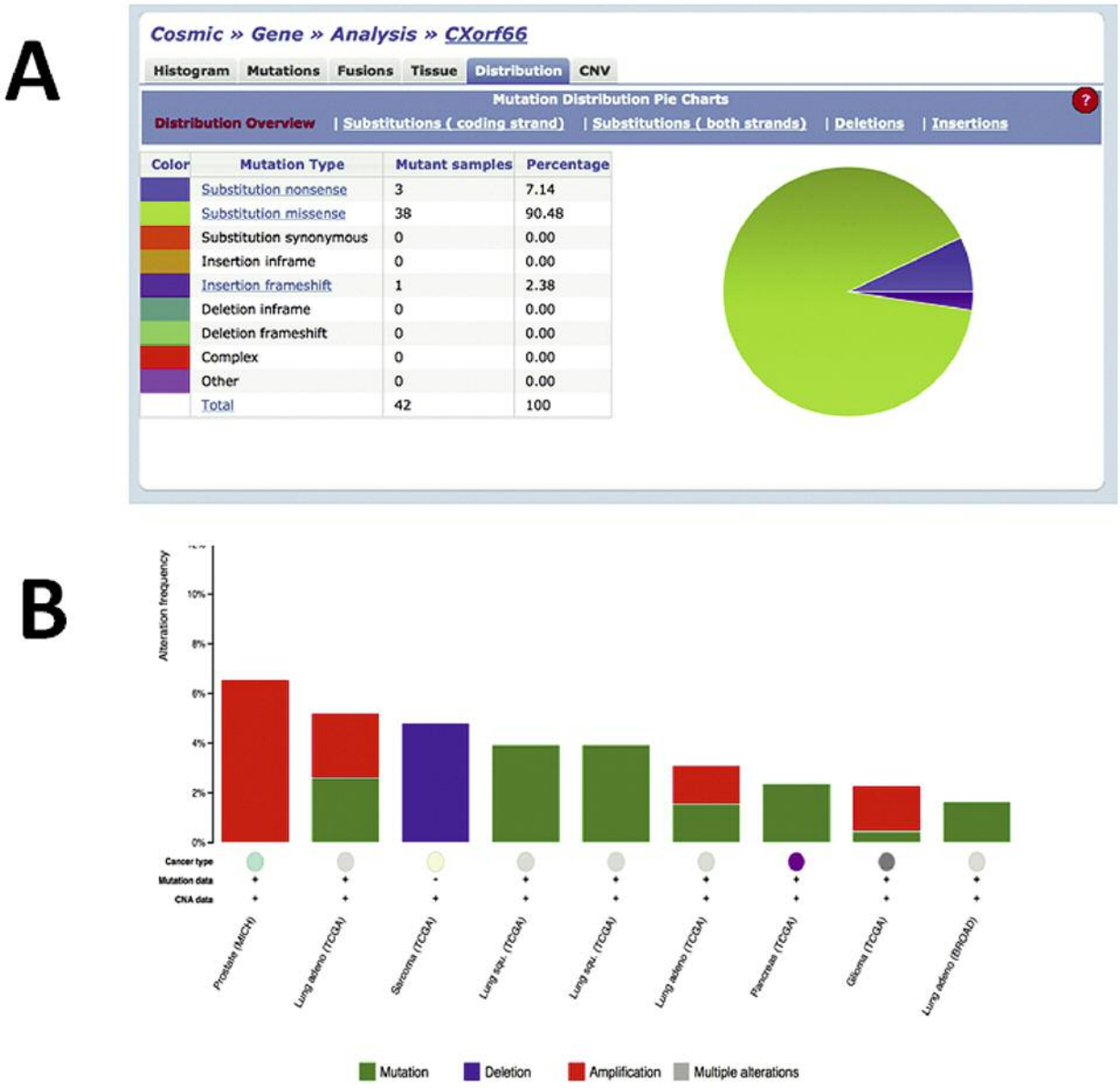


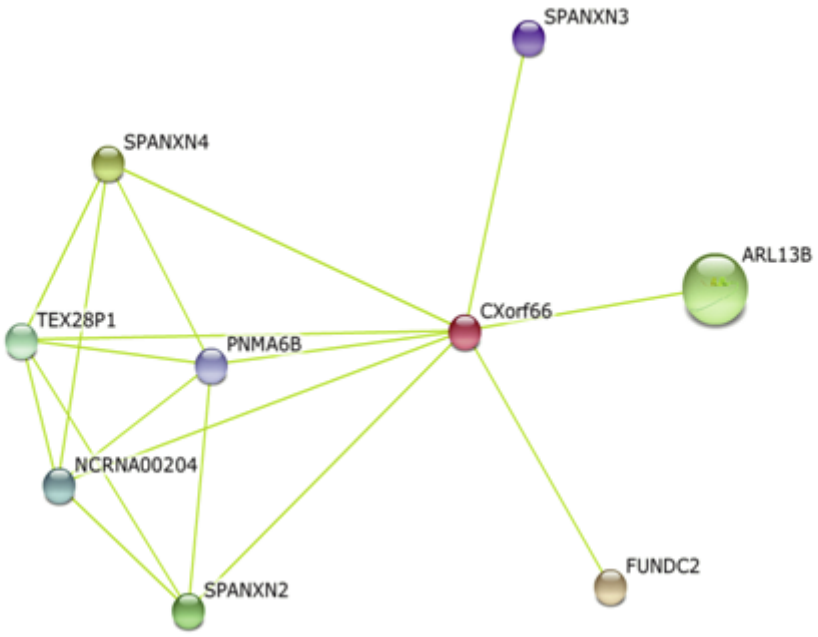
Figure 5. *CXorf66/SGPX alterations in cancer. A: Cosmic mutational class analysis. B: Amplifications, deletions and mutations of the SGPX gene as analyzed by cBioPortal.*

tissue, *p*-value 2.5e-21). The Hsa-130 b miR was up-regulated (2.36-fold) in blood from lung cancer patients (Bioset: Blood from patients from lung cancer *vs*. healthy control, *p*-value=0.0001).

The DNA methylation status of the *SGPX* gene was next investigated. In T-cell lymphoblastic leukemia, hypermethylation of the *SGPX* gene was seen as monitored by CpG island methylator phenotype (45). Lapatinib, a dual Epidermal growth factor (EGF) and Receptor tyrosine-protein kinase erbB-2 (HER2) kinase inhibitor, causes G_0/G_1 cell cycle arrest (46). In a breast cancer cell line, SKBR3, lapatinib treatment caused up-regulation of the *SGPX* mRNA, suggesting a G_1 regulation of *SGPX* gene. The G_1 regulation of the *SGPX* gene was further corroborated by NextBio Meta analysis. Mutations and overexpression of the G_1/S cyclin D1 (*CCND1*) (47) resulted in down-regulation of *SGPX* mRNA (data not shown).

Discussion

Discovery of novel secreted proteins offers a biomarker potential for non-invasive diagnosis of cancer. New diagnostic and therapeutic targets are likely to emerge among



Your Input:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
CXorf66	chromosome X open reading frame 66 (361 aa) *(Homo sapiens)*									

Predicted Functional Partners:

FUNDC2	FUN14 domain containing 2 (189 aa)							•		0.465
SPANXN4	SPANX family, member N4 (99 aa)							•		0.434
ARL13B	ADP-ribosylation factor-like 13B (428 aa)							•		0.430
SPANXN2	SPANX family, member N2 (180 aa)							•		0.430
TEX28P1	testis expressed 28 pseudogene 1 (410 aa)							•		0.429
NCRNA00204	non-protein coding RNA 204 (122 aa)							•		0.429
PNMA6B	paraneoplastic antigen like 6B (399 aa)							•		0.427
SPANXN3	SPANX family, member N3 (141 aa)							•		0.412

Figure 6. *CXorf66/SGPX interactome analysis. String interactome output for CXorf66 is shown. The bottom panel, predicted interacting partners. Output, medium confidence level. aa: Amino acids.*

the numerous uncharacterized proteins with putative ORFs. The results presented in the present study demonstrate the feasibility of mining the human proteome for cancer target discovery.

The CXorf66/*SGPX* gene is inferred to be a secreted protein product from the detection of the protein in the body fluids including plasma and serum. Motifs and domain analysis of the SGPX protein showed the presence of the classical signal peptide at the N-terminus. Transmembrane extracellular and cytoplasmic tail features were identified in the SGPX protein sequence. The SGPX protein harbors a serine rich motif (amino acids 91-177). This motif is present in numerous oncoproteins and is involved in binding to β-catenin (48).

The *SGPX* gene is a testis-restricted gene located on chromosome Xq27.1. Various cancer-associated genes are located in this region including members of the SPANX sperm protein associated with the nucleus mapped to the X chromosome (39). NCBI map positions MCF.2 cell line derived transforming sequence and testicular germ cell tumor

Table II. *Molecular characterization of the CXorf66 Secreted glycoprotein in chromosome X (SGPX) gene. Key findings from this study are shown in bold font.*

Characteristics	Gene Description	Tools
Gene aliases	Chromosome X Open Reading Frame 66; CXorf66; uncharacterized Protein CXorf66;RP11-35F15.2, LOC347487; hypothetical protein LOC347487, RP11-35F15.2	GeneCards, UniProt, AceView
Genomic RefSeq	NG_021221.1, 9.8kb	NCBI Gene
mRNA RefSeq	NM_001013403.2, 1,288 bp	NCBI Gene
Protein RefSeq	NP_001013421.1, 361 aa (precursor)	NCBI Protein
Gene summary	Membrane; Single-pass type I membrane protein (potential)	GeneCards, UniProt, Gene
Transcription factor binding sites	CTCF, E2F1, MYC, POLR2A, NR3C1, CEBPB, FOS	UCSC Genome Browser
Chromosome map	Xq27.1 (+3648486899712.00)	NCBI Map Viewer
Neighboring genes	MCF 2, ATP11C	AceView, NCBI Map Viewer
Variants	230: Natural variant-rS5955139 (P to L, aa. 233). Missense	dbSNP
Isoforms	One primary transcript	UniProt KB
Introns/ Exon	3 exons	Aceview
Homologues	7 homologs (Mammals) *H. sapiens, P. troglodytes, M. mulatta, C. lupus, B. taurus, M. musculus, R. norvegicus*	Homologue Gene
Ontology	Integral component of membrane, **secreted (putative)**	UnitProt, canSar, **Present study**
Putative class	**Cell trafficking, secretion, vesicular transport**	**Present study**
MicroRNAs	hsa-130a, hsa-130b,hsa-1290	NextBio, GeneCards
Motifs, Domains	DU936, Signal,Ribosomal S2, INP family, FAM163, serine-rich	Numerous
3D model templates	Auxilin, Properdin glycoprotein	UCSC, I-Tasser
Modification	Glycosylation, phosphorylation	Expasy
Expression of mRNA (normal)	Testis (enriched), fetus	UniGene, NextBio
Protein expression (normal)	Salivary gland, kidney, testis, lymph node, tonsils, spleen, leucocytes (strong staining), plasma, serum	Human Protein Atlas, MOPED, HPRD, GeneCards
Expression of mRNA (cancer)	Up-regulated: brain, lung, lymphoid leukemia. Down-regulated: liver, prostate	NextBio
Cell lines expression	Non-small cell lung cancer, breast, CNS, ovary	NCI 60, DTP
Protein Expression (cancer)	Up-regulated: colon, lung, ovarian, urothelial. Down-regulated: kidney, lymphoma, head and neck	Human Protein Atlas

Table III. *Characterization of the CXorf66/ Secreted glycoprotein in chromosome X (SGPX) protein. The indicated proteomic tools were used to analyze the SGPX protein. The canonical sequence NP_001013421 was used in these analyses.*

Motifs and Domains by Tool	Features
UniProt	Extracellular, helix, cytoplasmic, serine-rich
CDD	Uncharacterized domain DUF 936
HMMER	Ribosomal protein S2
Prosite/MotifScan	Serine-rich
Pfam/Motif Genomenet	Family 163, neuroblastoma-derived secretary protein
ProDom	PDA8V7U5 (4e-48), PDA3C6G2 (1e-86)
Signal P	Signal peptide, cleavage site located between aa.17-18, NNC-MT (Y-score, 0.476)
MESSA	KOGO566 (INP family), Signal, Disorder, Helix transmembrane
Predict protein	Protein binding sites (184, 270, 278), Secreted (GO:0005576), helix transmembrane, disorder
Post-translational modification:	
N-Glyc	aa. 24 (NGSS)
O-link glyc	aa. 92, 94
Myristylation	Non-myristoylated
Phosphorylation, GeneCards	S114, S165, T169 (Modified)
Phosphorylation, NetPhos 2.0	S41, T2, Y5
Kinase, NetPhos 1.0	PKC-specific at aa. 170

Table IV. *CXorf66/ Secreted glycoprotein in chromosome X (SGPX) cancer genome mutation summary. Mutations data from CanSar, COSMIC and ICGC databases is shown.*

Amino Acid Mutation	Mutation Type	Tissue
C105S	Missense	Breast
P158P	coding silent	Breast
S73C	Missense	Breast
V60V	c.180C>A	coding silent
Y350C	Missense	Colon
S122A	Missense	Endometrium
S187N	Missense	Glioma
L216F	Missense	Intestine
S119*	Nonsense	Intestine
H273fs*8	Insertion - Frameshift	Intestine
R311H	Missense	Kidney
L137I	Missense	Kidney
K40E	Missense	Large intestine
N325D	Missense	Large intestine
C105S	Missense	Large intestine
P239L	Missense	Lung
R316S	Missense	Lung
S170*	Nonsense	Lung
A237S	Missense	Lung
Y350N	Missense	Lung
S27C	Missense	Lung
P281H	Missense	Lung
A88S	Missense	Lung
K304N	Missense	Lung
L143L	coding silent	Lung
T294S	Missense	Lung
M18I	Missense	Lung
pE224Q	Missense	Lung
W13*	Nonsense	Lung
S355Y	Missense	Lung
H220N	Missense	Lung
A237S	Missense	Lung
N259K	Missense	Lung
P233L	Missense	Lung
M112L	Missense	Pancreas
P239L	Missense	Stomach
S168P	Missense	Stomach
A88V	Missense	Uterus
K166T	Missense	Uterus
K93T	Missense	Uterus
L9F	Missense	Uterus
K40E	Missense	Uterus
L137I	Missense	Uterus
K270N	Missense	Uterus
K331I	Missense	Uterus

susceptibility 1 genes to this locus. A list of all genes (Atlas of Genetics and Cytogenetics in Oncology and Hematology) present in chromosome X indicates the presence of numerous cancer-related genes (49).

The *SGPX* gene showed a complex pattern of expression in diverse tumors. It is found to be up-regulated in brain and lung tumors and in leukemia. On the other hand, it is down-regulated in liver and prostate carcinomas. Somatic mutations were found in glioma, lung, uterine and pancreatic cancer. Deletions were found in sarcoma and amplifications in prostate, lung and glioma. To date, however, no association has been established in testicular cancer, which is not yet adequately represented in the numerous cancer genome databases.

Microarray datamining showed interesting sex differences in patients with brain cancer. In different subtypes of brain cancer, female patients showed significantly elevated DNA copy numbers of the *SGPX* gene. In mammals, silencing of one of the two X chromosomes is required for dosage compensation (50). Some X-linked genes, however, escape such silencing. Hypermutation of the inactive X chromosome in females contributes to various cancers (51). It is tempting to speculate that the *SGPX* gene may play an important role in the development and progress of reproductive and urological cancers. Additional experiments are warranted to clarify the relevance of the *SGPX* gene in female cancer.

The function of the *SGPX* gene is unclear. However, our results with the 3-D modeling and the interactome analysis offer a clue. The structural homologues include the Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase (PTEN)-like region of auxilin, which is involved in vesicular budding (35). The SGPX protein shows regions of similarity to inositol-1,4,5-triphophate 5-phosphatase/synaptojanin. The family members of this protein are involved in intracellular trafficking, secretion and vesicular transport (52). We speculate that the *SGPX* gene function involves cellular trafficking, transport and budding.

HMMER identified a ribosomal S2 Protein Family (PFAM) domain in the SGPX protein (amino acids 18-113). This domain encompasses DS RNA and protein binding signatures and is involved in mitogenic fibroblast growth factor binding (GO). The *RPS2* mRNA was expressed in all cancer cells and non-malignant cell lines tested, but was not expressed in normal tissues except for the testis, muscle, and peripheral mononuclear leukocyte cells (53). The protein expression of RPS2 correlates with the tumor types for the *SGPX* gene (Human Protein Atlas, data not shown). The S2 domain similarity raises a possibility that the SGPX protein may be involved in mitogenic signaling pathways.

Additional insight into the SGPX protein comes from the NDSP family signature predicted by the PFAM tool. NDSP is a secretory protein highly specific to neuroblastoma (33). The NDSP protein expression pattern correlated with the SGPX in the tumors (Human Protein Atlas, data not shown). The precise function of NDSP, however, is not known. In view of the shared signature and expression profile with the SGPX protein, we postulate that the SGPX protein may share a functional role similar to the NDSP in brain tumors. Alternatively, NDSP and the SGPX could be interacting protein partners.

The interactome analysis predicted SGPX interactions with SPANX family members, which are involved in spermiogenesis (32). We postulate that the *SGPX* gene is involved in the trafficking and transport of sperm cells.

In conclusion, the discovery of the *SGPX* gene from the dark matter of the human proteome and the establishment of its relevance to major cancers underscore the power of bioinformatics in mining of the cancer genome. The *SGPX* gene offers a valuable biomarker potential for cancer.

Conflicts of Interest

None.

Contributions

RN was responsible for the overall execution of the project. Data generation and validation were performed by APD, PB and SH.

Acknowledgements

We thank the cancer genome databases, CanSar, cBioportal, the ICGC, TCGA and COSMIC for the mutations dataset. This work was supported in part by the Genomics of Cancer Fund, Florida Atlantic University Foundation. We thank Jeanine Narayanan for editorial assistance.

References

1 Lander ES, Linton LM, Birren B *et al*: and International Human Genome Sequencing, Consortium Initial sequencing and analysis of the human genome. Nature *409(6822)*: 860-921, 2001.
2 Venter JC, Adams MD, Myers EW *et al*: The sequence of the human genome. Science *291(5507)*: 1304-1351, 2001.
3 Narayanan R: Bioinformatics approaches to cancer gene discovery. Meth Mol Biol *360*: 13-31, 2007.
4 Scheurle D, DeYoung MP, Binninger DM, Page H, Jahanzeb M and Narayanan R: Cancer gene discovery using digital differential display. Cancer research *60(15)*: 4037-4043, 2000.
5 De Young MP, Damania H, Scheurle D, Zylberberg C and Narayanan R: Bioinformatics-based discovery of a novel factor with apparent specificity to colon cancer. In Vivo *16(4)*: 239-248, 2002.
6 Schmitt AO: Mining expressed sequence tag (EST) libraries for cancer-associated genes. Methods in molecular biology *76*: 89-98, 2010.
7 Alfoldi J and Lindblad-Toh K: Comparative genomics as a tool to understand evolution and disease. Genome Res *23(7)*: 1063-1068, 2013.
8 Lauriola M, Ugolini G, Rosati G, Zanotti S, Montroni I, Manaresi A *et al*: Identification by a Digital Gene Expression Displayer (DGED) and test by RT-PCR analysis of new mRNA candidate markers for colorectal cancer in peripheral blood. International J Oncol 37(2): 519-25, 2010.
9 Hopkins AL and Groom CR: The druggable genome. Nature reviews Drug Discovery *1(9)*: 727-730, 2002.
10 Pertea M and Salzberg SL: Between a chicken and a grape: estimating the number of human genes. Genome Biology *11(5)*: 206, 2010.
11 Nagano T and Fraser P: No-nonsense functions for long noncoding RNAs. Cell *145(2)*: 178-181, 2011.
12 Brylinski M: Exploring the "dark matter" of a mammalian proteome by protein structure and function modeling. Proteome Science *11(1)*: 47, 2013.
13 Blaxter M: Genetics. Revealing the dark matter of the genome. Science *330(6012)*: 1758-1759, 2010.
14 Hauptman N and Glavac D: MicroRNAs and long non-coding RNAs: prospects in diagnostics and therapy of cancer. Radiology and oncology *47(4)*: 311-318, 2013.
15 Mak L, Liggi S, Tan L, Kusonmano K, Rollinger JM, Koutsoukas A, Glen RC and Kirchmair J: Anticancer drug development: computational strategies to identify and target proteins involved in cancer metabolism. Current pharmaceutical design *19(4)*: 532-577, 2013.
16 Natrajan R and Wilkerson P: From integrative genomics to therapeutic targets. Cancer Res 7 *3(12)*: 3483-488, 2013.
17 Nevins JR and Potti A: Mining gene expression profiles: expression signatures as cancer phenotypes. Nature Reviews Genetics *8(8)*: 601-609, 2007.
18 Delgado A.P BP, Hamid S and Narayanan R: Mining the dark matter of the cancer proteome for novel biomarkers. Current Cancer Therapy Review 10, 2014 (in press).
19 Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA *et al*: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery *2(5)*: 401-404, 2012.
20 Halling-Brown MD, Bulusu KC, Patel M, Tym JE and Al-Lazikani B: canSAR: an integrated cancer public translational research and drug discovery resource. Nucleic Acids Res *40(Database issue)*: D947-956, 2012.
21 Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D *et al*: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids Res *39(Database issue)*: D945-950, 2011.
22 Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB *et al*: Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia *9(2)*: 166-180, 2007.
23 Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R *et al*: ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic acids Res *33(Database issue)*: D553-555, 2005.
24 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M *et al*: Towards a knowledge-based Human Protein Atlas. Nature biotechnology *28(12)*: 1248-1250, 2010.
25 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S *et al*: Human Protein Reference Database–2009 update. Nucleic acids Res *37(Database issue)*: D767-772, 2009.
26 Kolker E, Higdon R, Haynes W, Welch D, Broomall W, Lancet D *et al*: MOPED: Model Organism Protein Expression Database. Nucleic acids Res *40(Database issue)*: D1093-1099, 2012.
27 Rost B, Yachdav G and Liu J: The PredictProtein server. Nucleic acids research *32(Web Server issue)*: W321-326, 2004.

28 Cong Q and Grishin NV: MESSA: MEta-Server for protein Sequence Analysis. BMC Biol *10*: 82, 2012.
29 Workman P, Al-Lazikani B and Clarke PA: Genome-based cancer therapeutics: targets, kinase drug resistance and future strategies for precision oncology. Current opinion in pharmacology *13(4)*: 486-496, 2013.
30 Landry Y and Gies JP: Drugs and their molecular targets: an updated overview. Fundamental & Clinical Pharmacology *22(1)*: 1-18, 2008.
31 Orth AP, Batalov S, Perrone M and Chanda SK. The promise of genomics to identify novel therapeutic targets. Expert opinion on therapeutic targets *8(6)*: 587-596, 2004.
32 Zendman AJ, Zschocke J, van Kraats AA, de Wit NJ, Kurpisz M, Weidle UH *et al*: The human SPANX multigene family: genomic organization, alignment and expression in male germ cells and tumor cell lines. Gene *309(2)*: 125-133, 2003.
33 Vasudevan SA, Russell HV, Okcu MF, Burlingame SM, Liu ZJ, Yang J *et al*: Neuroblastoma-derived secretory protein messenger RNA levels correlate with high-risk neuroblastoma. Journal of pediatric surgery *42(1)*: 148-152, 2007.
34 Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics *23(2)*: e13-16, 2007.
35 Guan R, Dai H, Harrison SC and Kirchhausen T: Structure of the PTEN-like region of auxilin, a detector of clathrin-coated vesicle budding. Structure *18(9)*: 1191-1198, 2010.
36 Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA *et al*: Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501(7468)*: 506-511, 2013.
37 Westbrook VA, Schoppee PD, Diekman AB, Klotz KL, Allietta M, Hogan KT *et al*: Genomic organization, incidence, and localization of the SPAN-x family of cancer-testis antigens in melanoma tumors and cell lines. Clinical cancer research: an official journal of the American Association for Cancer Res *10(1 Pt 1)*: 101-12, 2004.
38 Salemi M, Calogero AE, Zaccarello G, Castiglione R, Cosentino A, Campagna C *et al*: Expression of SPANX proteins in normal prostatic tissue and in prostate cancer. European journal of histochemistry: EJH *54(3)*: e41, 2010.
39 Li K, Wang L, Cheng J, Zhang L, Duan H, Lu Y *et al*: Screening and cloning gene of hepatocyte protein interacting with hepatitis C virus core protein. Chin J Exp Clin Virol *16(4)*: 351-353, 2002. (in Chinese).
40 Ma X, Wang YW, Zhang MQ and Gazdar AF: DNA methylation data analysis and its application to cancer research. Epigenomics *5(3)*: 301-316, 2013.
41 Gnyszka A, Jastrzebski Z and Flis S: DNA methyltransferase inhibitors and their emerging role in epigenetic therapy of cancer. Anticancer Res *33(8)*: 2989-2996, 2013.
42 Villicana C, Cruz G and Zurita M: The basal transcription machinery as a target for cancer therapy. Cancer cell international *14(1)*: 18, 2014.
43 Qi W, Liang W, Jiang H and Miuyee Waye M: The function of miRNA in hepatic cancer stem cell. BioMed Research International *2*: 358902, 2013.
44 Vermeulen K, Van Bockstaele DR, Berneman ZN. The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. Cell Proliferation *36(3)*: 131-149, 2003.
45 Borssen M, Palmqvist L, Karrman K, Abrahamsson J, Behrendtz M, Heldrup J *et al*: Promoter DNA methylation pattern identifies prognostic subgroups in childhood T-cell acute lymphoblastic leukemia. PloS one *8(6)*: e65373, 2013.
46 Wainberg ZA, Anghel A, Desai AJ, Ayala R, Luo T, Safran B *et al*: Lapatinib, a dual EGFR and HER2 kinase inhibitor, selectively inhibits HER2-amplified human gastric cancer cells and is synergistic with trastuzumab in vitro and in vivo. Clinical cancer research: an official journal of the American Association for Cancer Research *16(5)*: 1509-1519, 2010.
47 Casimiro MC, Velasco-Velazquez M, Aguirre-Alvarado C and Pestell RG: Overview of cyclins D1 function in cancer and the CDK inhibitor landscape: past and present. Expert opinion on Investigational Drugs *23(3)*: 295-304, 2014.
48 Huang L, Chen D, Liu D, Yin L, Kharbanda S and Kufe D: MUC1 oncoprotein blocks glycogen synthase kinase 3beta-mediated phosphorylation and degradation of beta-catenin. Cancer Res *65(22)*: 10413-10422, 2005.
49 Huret JL, Ahmad M, Arsaban M, Bernheim A, Cigna J, Desangles F *et al*: Atlas of genetics and cytogenetics in oncology and haematology in 2013. Nucleic Acids Res *41(Database issue)*: D920-4, 2013.
50 Liao DJ, Du QQ, Yu BW, Grignon D and Sarkar FH: Novel perspective: focusing on the X chromosome in reproductive cancers. Cancer Investigation *21(4)*: 641-658, 2003.
51 Jager N, Schlesner M, Jones DT, Raffel S, Mallm JP, Junge KM *et al*: Hypermutation of the inactive X chromosome is a frequent event in cancer. Cell *155(3)*: 567-581, 2013.
52 Astle MV, Horan KA, Ooms LM and Mitchell CA: The inositol polyphosphate 5-phosphatases: traffic controllers, waistline watchers and tumour suppressors? Biochemical Society Symposium *74*: 161-181, 2007.
53 Koga M, Shichijo S, Yamada A, Ashihara J, Sawamizu H, Kusukawa J *et al*: Identification of ribosomal proteins S2 and L10a as tumor antigens recognized by HLA-A26-restricted CTL. Tissue Antigens *61(2)*: 136-145, 2003.

Received March 4, 2014
Revised March 12, 2014
Accepted March 13, 2014