BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(2)

Computational analysis of Concanavalin A binding glycoproteins of human seminal plasma

Anil Kumar Tomar[1], Balwinder Singh Sooch[2], Savita Yadav[1]*

[1]Department of Biophysics, All India Institute of Medical Sciences, New Delhi India; [2]Department of Biotechnology, Punjabi University, Patiala, India; Savita Yadav – Email: savita11@gmail.com; Phone: +91-11-26596445; Fax: +91-11-26588641; *Corresponding author

Received August 27, 2011; Accepted August 30, 2011; Published September 06, 2011

Abstract:
Glycoproteins have immense clinical importance and comparative glycoproteomics has become a powerful tool for biomarker discovery and disease diagnosis. Seminal plasma glycoproteins participate in fertility related processes including sperm-egg recognition, modulation of capacitation and acrosome reaction inhibition. Affinity chromatography using broad specificity lectin such as Con A is widely applied for glycoproteins enrichment. More notably, Con A-interacting fraction of human seminal plasma has decapacitating activity which makes this fraction critically important. In our previous study, we isolated Con A-interacting glycoproteins from human seminal plasma and subsequently identified them by mass spectrometry. Here, we report the computational analysis of these proteins using bioinformatics tools. The analysis includes: prediction of glycosylation sites using sequence information (NetNGlyc 1.0), functional annotations to cluster these proteins into various functional groups (InterProScan and Blast2GO) and identification of protein interaction networks (STRING database). The results indicate that these proteins are involved in various biological processes including transport, morphogenesis, metabolic processes, cell differentiation and homeostasis. The clusters illustrate two major molecular functions - hydrolase activity (6) and protein (4)/carbohydrate (1)/lipid binding (1). The large interactomes of proteins point towards their versatile roles in wide range of biological processes.

Background:
Glycosylation is one of the most common post-translational modifications and more than half of all mammalian proteins are glycosylated **[1]**. The studies towards isolation, discovery and subsequent identification of glycosylated proteins are becoming more and more important in glycoproteomics and disease diagnosis **[2]**. In particular, differential glycosylations (e.g. missing, aberrant or additional) are known to be linked to certain diseases and may be utilized as markers for diagnosis and/or therapeutic monitoring **[3]**. Glycoproteins play essential roles in controlling various biological processes in immunology, cancer, protein folding, host-pathogen interactions, human diseases and signal transduction etc. The broad specificity lectins, such as Concanavalin A (Con A), are widely applied for enriching serum glycoproteins **[4]**. Human seminal plasma contains a large array of proteins of clinical importance which are essentially needed to maintain the reproductive physiology of spermatozoa and for successful fertilization. Seminal plasma glycoproteins are known to participate in sperm-egg recognition **[5]**, modulation of capacitation **[6, 7]** and acrosome reaction inhibition **[8]**. Moreover, Con A-interacting fraction of human seminal plasma is reported to have decapacitating activity **[9]**. Thus, functional analysis of various proteins of this fraction is of immense importance for better understanding of fertility related processes.

We had isolated glycoproteins from human seminal plasma by lectin affinity chromatography using Con A – agarose. Overall ten proteins bands on SDS-PAGE gel, corresponding to nine different proteins, were identified by MALDI-TOF/MS analysis, viz. aminopeptidase N precursor (ANPEP), lactoferrin (LTF), prostatic acid phosphatase (ACPP), human zinc-alpha-2-glycoprotein (AZGP1), prostate specific antigen (KLK3), progestagen-associated endometrial protein (PAEP), kinesin light chain 4 (KLC4), izumo sperm-egg fusion protein 1 (IZUMO1) and prolactin inducible protein (PIP) **[10].** There are

a number of bioinformatics tools available for *in silico* analysis of proteins isolated and identified in protein chemistry laboratories. These analyses help us in better understanding of functional aspects of new proteins in various biological processes. Hence, we report the computational analysis of Con A binding glycoproteins, identified, using various bioinformatics tools. The objectives of this study include**,** (1) prediction of glycosylation sites using sequence information and to compare the results with available experimental data, (2) functional annotation studies using Interpro and Blast2GO to cluster these proteins into functional groups, and (3) identification of protein-protein interaction (PPI) networks.

Methodology:

Sequence Retrieval:

The amino acid sequences of glycoproteins - ANPEP (P15144), LTF (P02788), ACPP (P15309), AZGP1 (P25311), KLK3 (P07288), PAEP (Q5T6T6), KLC4 (Q9NSK0), IZUMO1 (Q8IYV9) and PIP (P12273) were retrieved in FASTA format from Protein Knowledgebase (UniProt KB) **[11].**

Prediction of glycosylation sites and comparison with known sites:

The possible N-glycosylation sites were predicted using NetNGlyc 1.0 program **[12]**. This program predicts N-glycosylation sites in human proteins using artificial neural networks that examine the amino acid sequence of N-X-S/T (Asn-Xaa-Ser/Thr). The predicted sites were compared with the known sites in these proteins, as evidenced by direct experiments **[13]**.

Statistical analysis of amino acid content:

The statistical analysis of amino acid content of each protein was done using program Pepstat **[14].** This is a basic statistical tool which calculates the % share of individual amino acids in a protein sequence as well as shares of nine specific types of amino acid groups, such as tiny, small, aliphatic, aromatic, polar, non-polar, charged, basic and acidic.

Functional annotations and clustering using Blast2GO and InterProScan:

InterProScan is a popular program suite for protein sequence analysis and classification **[15]**. It classifies sequences at various levels such as superfamily, family and subfamily and predicts the occurrence of functional domains and repeats. InterPro analysis was performed for glycoproteins to identify their subcellular locations and functions. The functional annotations were also carried out using Blast2GO and subsequently proteins were grouped into functional clusters **[16]**. All protein sequences were arranged in a single file in FASTA format and uploaded to the Blast2Go software suite **[17]** to facilitate batch handling of sequence data. The file was processed by implementing batch mode blastp, mapping to retrieve GO terms associated with each blast hit and Gene Ontology annotations. The program finally provides refined functional terms to each query based on their functions, statistical testing and InterProScan analysis. Finally, the retrieved information was used for graphical representation of results (cellular components, biological processes and molecular functions) in the form of pie charts.

Prediction of protein-protein interaction (PPI) networks:

PPI networks for each protein were retrieved from STRING database **[18, 19]**. This database consists of known and predicted protein interactions collected from direct (physical) and indirect (functional) associations. This database quantitatively integrates interaction data from four sources - genomic context, high-throughput experiments, co-expression and previous knowledge from research publications.

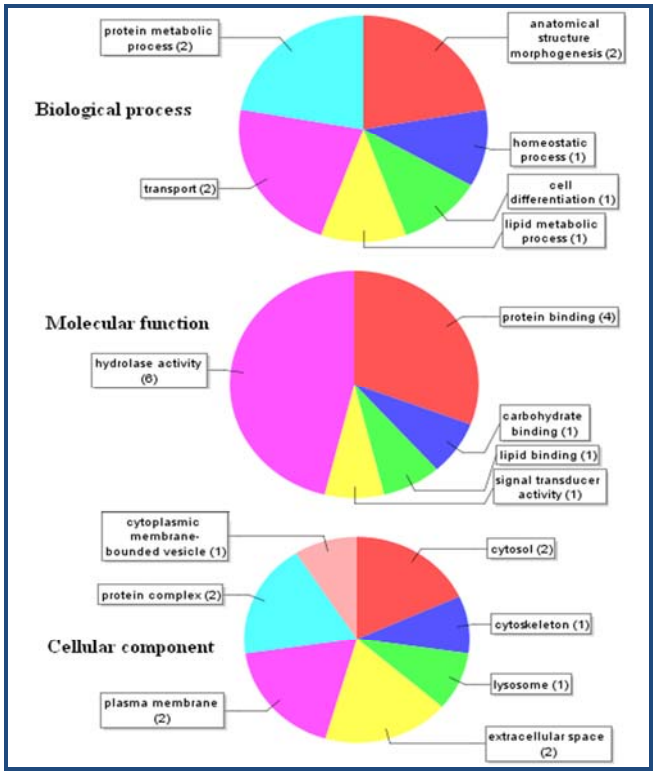

Figure 1: Blast2GO analysis of Con A binding glycoproteins of human seminal plasma **(see, Table 2 in supplementary material)**

Results and Discussion:

The predicted N-glycosylation sites in Con A binding glycoproteins were compared with experimentally known sites in these proteins **(see, Table 1 in supplementary material)**. AZGP1, LTF, ACPP, KLK3 and PIP have 4,3,3,1 and 1 known N-glycosylation sites respectively, which were accurately predicted by NetNGlyc 1.0 program. This program predicts that ANPEP has 11 potential N-glycosylation sites (N42, N128, N234, N265, N319, N527, N573, N625, N681, N735, and N818), of which six are previously known (N128, N234, N265, N573, N681 and N818). IZUMO1 has one known N-glycosylation site at position N204 and predictions indicate that it may have another potential glycosylation site at position N239. PAEP and KLC4 have no known N-glycosylation sites and NetNGlyc 1.0 predicts that they may have 2 (N33, N55) and 1 (N4) glycosylation sites respectively. The Pepstat results show that nine glycoprotein sequences have the following mean mole percentage values of different types of residues: Aliphatic (A+I+L+V) = 29.13±6.81; Aromatic (F+H+W+Y) = 11.25 ±2.37; Non-polar (A+C+F+G+I+L+M+P+V+W+Y) = 53.73±4.12; Polar (D+E+H+K+N+Q+R+S+T+Z) = 46.16±4.12; Charged

(B+D+E+H+K+R+Z) = 25.46±3.83; Basic (H+K+R) = 13.30±1.80; Acidic (B+D+E+Z) = 12.13±2.36.

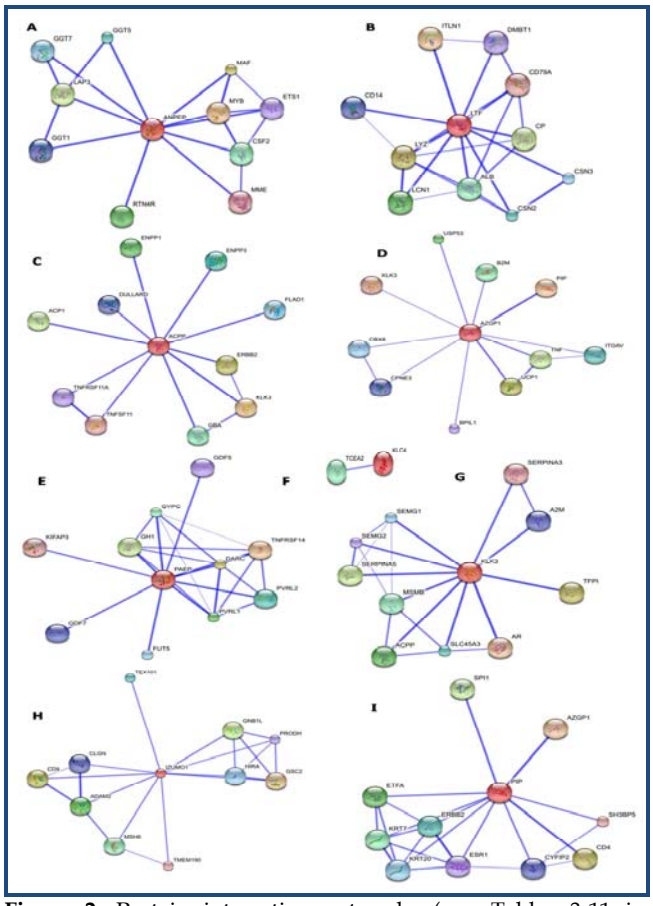

Figure 2: Protein interaction networks (see, Tables 3-11 in supplementary material)

InterProScan results were integrated to the Blast2GO analysis to increase the confidence level of functional clustering. The final outputs of functional annotation studies are shown in **Figure 1 and Table 2 (see Supplementary material)**. The annotations specify that these proteins are involved in various biological processes including transport (LTF, PAEP), morphogenesis (ANPEP, KLK3), metabolic processes (ANPEP, AZGP1, KLK3), cell differentiation (ANPEP) and homeostasis (LTF). ACPP and IZUMO1 have reported roles in hydrolysis and reproduction (sperm-egg fusion) respectively, but the exact biological processes, they are involved, are still unknown. The functional clusters show that Con A – binding glycoproteins have two major molecular functions - hydrolase activity (ANPEP, LTF, ACPP, AZGP1, KLK3, PAEP) and binding - protein (LTF, AZGP1, KLC4, PIP)/carbohydrate (LTF)/lipid (AZGP1). The subcellular localizations of these proteins are also shown in the results, indicating that most of them originate from different cellular components. These proteins play imperative roles in various biological processes related to fertility/infertility and their expression regulates the processes essentially required for successful fertilization. Their key roles in reproductive physiology are well discussed **[10]**. PPI networks for these glycoproteins are shown in **Figure 2(A-I)**. The large interactomes for most of the proteins point towards their versatile roles in wide range of biological processes. Thus, in depth characterization of these proteins may reveal that these are more important and multifaceted entities than what we are assuming about them for long.

Conclusion:

The computational tools aid to the functional characterization of biomolecules by identifying their homologs in the biological databases and retrieving information from the research articles published worldwide. It helps researchers to guide their future studies towards *in vivo* or *in vitro* functional characterization. We have successfully identified the N-glycosylation sites of Con A binding glycoproteins isolated from human seminal plasma, clustered them into functional groups and mapped their interactomes. Thus, it is of importance in better understanding of functional aspects of these proteins in reproductive physiology.

Acknowledgement:

This work is supported by grant from Department of Science and Technology (DST), New Delhi INDIA. Anil Kumar Tomar also thanks DST for his fellowship.

References:

[1] Apweiler R *et al. Biochim Biophys Acta.* 1999 **1473**: 4 [PMID: 10580125]
[2] Sparbier K *et al. J Biomol Tech.* 2005 **16**: 407 [PMID: 16522863]
[3] Durand G & Seta N. *Clin Chem.* 2000 **46**: 795 [PMID: 10839767]
[4] Madera M *et al. J Sep Sci.* 2008 **31**: 2722 [PMID: 18623281]
[5] Iborra A *et al. Am J Reprod Immunol.* 1996 **36**: 118 [PMID: 8862257]
[6] Thérien I *et al. Biol Reprod.* 1995 **52**: 1372 [PMID: 7632845]
[7] Calvete JJ *et al. Biol Chem.* 1996 **377**: 521 [PMID: 8922287]
[8] Drisdel RC *et al. Biol Reprod.* 1995 **53**: 201 [PMID: 7669849]
[9] Marquínez AC *et al. J Protein Chem.* 2003 **22**: 423 [PMID: 14690244]
[10] Tomar AK *et al. Dis Marker.* 2011 **31**
[11] http://www.uniprot.org
[12] http://www.cbs.dtu.dk/services/NetNGlyc/
[13] www.expasy.ch
[14] www.ebi.ac.uk/Tools/emboss/pepinfo/
[15] www.ebi.ac.uk/Tools/pfa/iprscan/
[16] Götz S *et al. Nucleic Acids Res.* 2008 **36**: 3420 [PMID: 18445632]
[17] http://www.blast2go.org
[18] Szklarczyk D *et al. Nucleic Acids Res.* 2011 **39**: D561 [PMID: 21045058]
[19] http://string-db.org

Edited by P Kangueane
Citation: Tomar *et al.* Bioinformation 7(2): 69-75 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Prediction of N-glycosylation sites on Con A binding glycoproteins of human seminal plasma

Protein	Predicted N-glycosylation sites		Experimentally known N-glycosylation sites (www.expasy.ch)	
	Number of sites	Positions in sequence	Number of sites	Positions in sequence
Aminopeptidase N precursor	11	42, 128, 234, 265, 319, 527, 573, 625, 681, 735, 818	6	128, 234, 265, 573, 681, 818
Lactoferrin	3	156, 497, 642	3	156, 497, 642
Prostatic acid phosphatase	3	94, 220, 333	3	94, 220, 333
Human Zinc-Alpha-2-Glycoprotein	4	109, 112, 128, 259	4	109, 112, 128, 259
Prostate specific antigen	1	69	1	69
Progestagen-associated endometrial protein	2	33, 55	Not reported	
Kinesin light chain 4	1	4	Not reported	
Izumo sperm-egg fusion protein 1	2	204, 239	1	204
Prolactin inducible protein	1	105	1	105

Table 2: Integration of InterProScan and Blast2GO results

Protein	Subcellular location	Biological Process	Molecular Function	GO Stat IDs
Aminopeptidase N precursor	Organelle, cytosol	anatomical structure morphogenesis, proteolysis, cell differentiation, metabolic process	receptor activity, metallopeptidase activity, zinc ion binding	30154,42277,46872,8237,8233, 6725,4872,4177,7275,44419, 1525,6508,5737,12506,5793, 5625,31983,5829,8270,5887, 5886,16021,16020,16787
Lactotransferrin	cytoplasmic membrane-bounded vesicle, extracellular region	response to stress & biotic stimulus, cellular homeostasis, ion transport	protein binding, ferric iron binding, peptidase activity, carbohydrate binding	30141,4252,8199,6811,5737, 5576,42742,8233,6826,5515, 6959,8201,46872,6879
Prostatic acid phosphatase	extracellular region, lysosome	hydrolase activity	acid phosphatase activity	5765,16787,3993,16020,5576
Human Zinc-Alpha-2-Glycoprotein	extracellular region, protein complex, plasma membrane	cell proliferation, lipid metabolic process, immune response, antigen processing and presentation	protein binding, transporter activity, nuclease activity, lipid binding	5504,16020,4540,16042,8320, 42612,5576,8285,6955,19882, 7155,5615
Prostate specific antigen	extracellular space, cytoplasm	protein metabolic process, anatomical structure morphogenesis, response to stress	serine-type endopeptidase activity, catalytic activity	6954,51919,16525,4252,5737, 7596,8236,2542,5576,8233, 6508,31639,31638,5615,16787, 3824
Progestagen-associated endometrial protein	-	multicellular organismal development, transport	binding	5488,7275,5215,6810,19841, 5576
Kinesin light chain 4	Cytoplasm, cytoskeleton, protein complex	-	motor activity, protein binding	5871,5829,3777,3774,5515, 5874,5488
Izumo sperm-egg fusion protein 1	Cell, integral to membrane	cellular component organization, reproduction	fusion of sperm to egg plasma membrane	7155,16021,16020,7342

Prolactin inducible protein	extracellular region	-	actin binding	3779,8150,5576,5515

Table 3: Aminopeptidase N (ANPEP) interacting proteins (Reference: Figure 2A)

MYB	v-myb myeloblastosis viral oncogene homolog ; Transcriptional activator; DNA-binding protein that specifically recognize the sequence 5'-YAAC[GT]G-3'. Plays an important role in the control of proliferation and differentiation of hematopoietic progenitor cells
MAF	v-maf musculoaponeurotic fibrosarcoma oncogene homolog; Acts as a transcriptional activator or repressor. Involved in embryonic lens fiber cell development
LAP3	leucine aminopeptidase 3; Presumably involved in the processing and regular turnover of intracellular proteins. Catalyzes the removal of unsubstituted N-terminal amino acids from various peptides
RTN4R	reticulon 4 receptor; Receptor for RTN4, OMG and MAG. Mediates axonal growth inhibition and may play a role in regulating axonal regeneration and plasticity in the adult central nervous system
CSF2	colony stimulating factor 2 (granulocyte-macrophage); Cytokine that stimulates the growth and differentiation of hematopoietic precursor cells from various lineages, including granulocytes, macrophages, eosinophils and erythrocytes
GGT5	gamma-glutamyltransferase 5; Cleaves the gamma-glutamyl peptide bond of glutathione conjugates, but maybe not glutathione itself. Converts leukotriene C4 (LTC4) to leukotriene D4 (LTD4)
GGT7	gamma-glutamyltransferase 7; Cleaves glutathione conjugates
GGT1	gamma-glutamyltransferase 1; Initiates extracellular glutathione (GSH) breakdown, provides cells with a local cysteine supply and contributes to maintain intracelular GSH level
ETS1	v-ets erythroblastosis virus E26 oncogene homolog ; Transcription factor
MME	membrane metallo-endopeptidase; Thermolysin-like specificity

Table 4: Lactoferrin **(**LTF**)** interacting proteins (Reference: Figure 2B)

ITLN1	intelectin 1 (galactofuranose binding); Has no effect on basal glucose uptake but enhances insulin-stimulated glucose uptake in adipocytes
LYZ	lysozyme (renal amyloidosis); Lysozymes have primarily a bacteriolytic function
CP	ceruloplasmin (ferroxidase); Ceruloplasmin is a blue, copper-binding
LCN1	lipocalin 1 (tear prealbumin); Could play a role in taste reception
ALB	Serum albumin, the main protein of plasma, has a good binding capacity for water, Ca^{++}, Na^{+}, K^{+}, fatty acids, hormones, bilirubin and drugs
CSN2	casein beta; Important role in determination of the surface properties of the casein micelles
CSN3	casein kappa; Kappa-casein stabilizes micelle formation, preventing casein precipitation in milk
CD14	CD14 molecule; Cooperates with MD-2 and TLR4 to mediate the innate immune response to bacterial lipopolysaccharide
DMBT1	deleted in malignant brain tumors 1; May be considered as a candidate tumor suppressor gene
CD79A	CD79a molecule, immunoglobulin-associated alpha; Required in cooperation with CD79B for initiation of the signal transduction cascade activated by binding of antigen to the B-cell antigen receptor complex (BCR) which leads to internalization of the complex, trafficking to late endosomes and antigen presentation

Table 5: Prostatic acid phosphatise (ACPP) interacting proteins (Reference: Figure 2C)

KLK3	kallikrein-related peptidase 3; Hydrolyzes semenogelin-1 thus leading to the liquefaction of the seminal coagulum
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene
ACP1	acid phosphatase 1, soluble; Acts on tyrosine phosphorylated proteins
ENPP1	ectonucleotide pyrophosphatase/phosphodiesterase 1; Involved primarily in ATP hydrolysis
GBA	glucosidase, beta; acid (includes glucosylceramidase)
ENPP3	ectonucleotide pyrophosphatase/phosphodiesterase 3; Cleaves a variety of phosphodiester and phosphosulfate bonds including deoxynucleotides, nucleotide sugars, and NAD
FLAD1	FAD1 flavin adenine dinucleotide synthetase homolog (S. cerevisiae); Catalyzes the adenylation
DULLARD	dullard homolog (Xenopus laevis); Serine/threonine phosphatase which may be required for proper nuclear membrane morphology. Involved in LPIN1 dephosphorylation
TNFRSF11A	tumor necrosis factor receptor superfamily, member 11a, NFKB activator
TNFSF11	tumor necrosis factor (ligand) superfamily, member 11

Table 6: Zinc alpha-2-glycoprotein (AZGP1) interacting proteins (Reference: Figure 2D)

PIP	prolactin-induced protein
UCP1	uncoupling protein 1 (mitochondrial, proton carrier); UCP are mitochondrial transporter protein
TNF	tumor necrosis factor (TNF superfamily, member 2)
USP53	ubiquitin specific peptidase 53
B2M	beta-2-microglobulin; Component of the class I major histocompatibility complex (MHC)
ITGAV	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)
CBX8	chromobox homolog 8 (Pc class homolog)
CPNE3	copine III; May function in membrane trafficking
BPIL1	bactericidal/permeability-increasing protein-like 1
KLK3	kallikrein-related peptidase 3; Hydrolyzes semenogelin-1 thus leading to the liquefaction of the seminal coagulum

Table 7: Progestagen-associated endometrial protein (PAEP) interacting proteins (Reference: Figure 2E)

TNFRSF14	tumor necrosis factor receptor superfamily, member 14 (herpesvirus entry mediator)
DARC	Duffy blood group, chemokine receptor; Non-specific receptor for many chemokines such as IL-8
GH1	growth hormone 1; Plays an important role in growth control
PVRL1	poliovirus receptor-related 1 (herpesvirus entry mediator C)
GYPC	glycophorin C (Gerbich blood group)
PVRL2	poliovirus receptor-related 2 (herpesvirus entry mediator B); Probable cell adhesion protein
FUT5	fucosyltransferase 5 (alpha (1,3) fucosyltransferase); May catalyze alpha-1,3 glycosidic linkage
GDF7	growth differentiation factor 7
GDF5	growth differentiation factor 5; Could be involved in bone formation
KIFAP3	kinesin-associated protein 3; Involved in tethering the chromosomes to the spindle pole

Table 8: Kinesin light chain 4 (KLC4) interacting protein (Reference: Figure 2F)

TCEA2	transcription elongation factor A (SII), 2; Necessary for efficient RNA polymerase II transcription

Table 9: Prostate specific antigen (KLK3) interacting proteins (Reference: Figure 2G)

AR	androgen receptor; Steroid hormone receptors are ligand-activated transcription factors
TFPI	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)
SERPINA5	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 5
ACPP	acid phosphatase, prostate
MSMB	beta-microseminoprotein
SLC45A3	solute carrier family 45, member 3
SEMG1	semenogelin I; Predominant protein in semen
A2M	alpha-2-macroglobulin; Is able to inhibit all four classes of proteinases
SEMG2	semenogelin II; Participates in the formation of a gel matrix (sperm coagulum)
SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin)

Table 10: Izumo sperm-egg fusion protein 1 (IZUMO1) interacting proteins (Reference: Figure 2H)

GSC2	goosecoid homeobox 2; May have a role in development
CD9	CD9 molecule; Involved in platelet activation and aggregation
GNB1L	guanine nucleotide binding protein (G protein)
ADAM2	ADAM metallopeptidase domain 2; Sperm surface membrane protein that may be involved in sperm-egg plasma membrane adhesion and fusion during fertilization
MSH6	mutS homolog 6 (E. coli); Component of the post-replicative DNA mismatch repair system (MMR)
TEX101	testis expressed 101; May play a role in signal transduction
HIRA	HIR histone cell cycle regulation defective homolog A
CLGN	calmegin; Probably plays an important role in spermatogenesis, Binds calcium ions
PRODH	proline dehydrogenase (oxidase) 1; Converts proline to delta-1-pyrroline-5-carboxylate
TMEM190	transmembrane protein 190

Table 11: Prolactin inducible protein (PIP) interacting proteins (Reference: Figure 2I)

AZGP1	alpha-2-glycoprotein 1, zinc-binding; Stimulates lipid degradation in adipocytes
CD4	CD4 molecule; Accessory protein for MHC class-II antigen/T-cell receptor interaction
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene spi1
ETFA	electron-transfer-flavoprotein, alpha polypeptide
KRT7	keratin 7; Blocks interferon-dependent interphase and stimulates DNA synthesis in cells
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2
KRT20	keratin 20; Plays a significant role in maintaining keratin filament organization
CYFIP2	cytoplasmic FMR1 interacting protein 2; Involved in T-cell adhesion and p53-dependent induction of apoptosis.
ESR1	estrogen receptor 1; Nuclear hormone receptor
SH3BP5	SH3-domain binding protein 5 (BTK-associated)