BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

Open access

www.bioinformation.net
Volume 12(2)

Hypothesis

Towards the construction of an interactome for Human WD40 protein family

Hulikal Shivashankara Santosh Kumar[1], Vadlapudi Kumar[2*], Sharath Pattar[3], Sandeep Telkar[4]

[1]Department of Biotechnology and Bioinformatics, Kuvempu University, Shankaraghatta-577451, Karnataka, India; [2]Department of Biochemistry, Davanagere University, Shivagangothri, Davanagere-577002, Karnataka, India; [3]National Bureau of Agriculturally Important Insects, Hebbal, Bengaluru, Karnataka, India; [4]Department of Biotechnology and Bioinformatics, Kuvempu University, Shankaraghatta-577451, Karnataka, India; Vadlapudi Kumar - Email: vadlapudikumar@gmail.com; *Corresponding author

Received February 17, 2016; Revised February 25, 2016; Accepted February 25, 2016; Published April 10, 2016

Abstract:
WD40 proteins are involved in a variety of protein-protein interactions as part of a multi-protein assembly modulating diverse and critical cellular process. It is known that several proteins of this family have been implicated in different disorders such as developmental abnormalities and cancer. However, molecular functions of many proteins in this family are yet unknown and it is of clinical interest. Therefore, it is of interest to define, construct, understand, analyze, evaluate, redefine and refine an interactome for WD40 protein family. We used data from literature mining using Cytoscape followed by linear regression analysis between Betweenness centrality and stress scores to define a model to filter the nodes in a representative WD40 interactome construction. We identified 10 ranked nodes in this analysis and subsequent microarray data selected three of them in insulin resistance that is further demonstrated in HepG2 cell culture models. We also observed the expression of GRWD1, RBBP5 and WDR5 genes during perturbation. Thus, we report hub nodes of WD40 interactome in insulin resistance. It should be noted that the pipeline using protein interaction network help find new proteins of clinical importance

Key Words: WDR5, Interactome, Betweenness Centrality, Stress, Insluin Resistance

List of abbreviations: BC- Betweenness Centrality; rt-PCR – Riverse Transcription Polymerase Chain Reaction; WDR5 – WD repeat containing protein 5; GRWD1- Glutamate Rich WD repeat containing protein 1; RBBP4 – Retinoblastoma binding protein 4; RBBP5 – Retinoblastoma binding protein 5; ASH2L– Set1/Ash2 histone methyltransferase complex subunit; KAT2B – Lysine Acetyl Transferase 2B

Background:

WD-repeats are minimally conserved domains of approximately 40–60 amino acids that are initiated by a Glycine-Histidine (GH) dipeptide 11 to 24 residue from the N terminus and end with a tryptophan-aspartic acid (WD) dipeptide at the C terminus **[1].** Studies suggests WD motifs act as a site for protein-protein interaction, and proteins containing WD repeats (WDRs) are known to serve as platforms for the assembly of protein complexes or mediators of transient interplay among other proteins. WDR proteins are intimately involved in a variety of cellular and organismal processes, including cell division and cytokinesis, apoptosis, light signalling and vision, cell motility, flowering, floral development, and meristem organization, to name a few. They are highly conserved in eukaryotes **[2].** According to Nocker and Ludwig (2003) since WDRs serve as platforms for the assembly of protein complexes, this characteristic of WDRs allows for three general functional roles. First, WDRs within one protein can provide binding sites for two or more other proteins and facilitate transient interactions among these other proteins. A second potential role of WDR proteins is as an integral component of multiple protein complexes. A third recognized role of the WDR is to act as a modular interaction domain of larger proteins. The presumed role of the WDR in these cases is to bring the protein and associated ancillary domains into proximity of its targets **[2].**

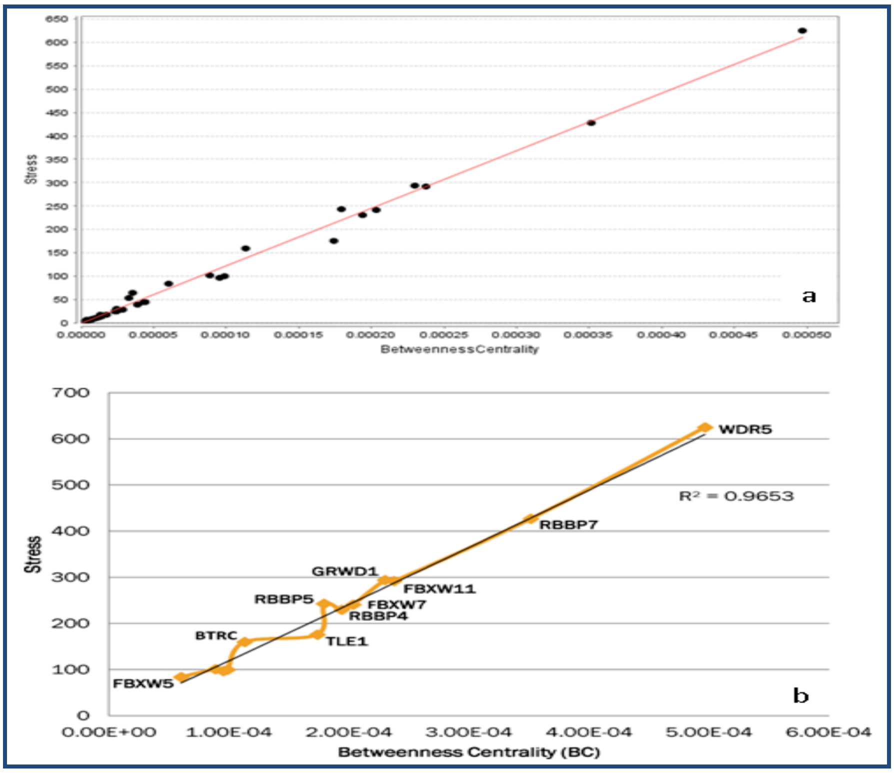

Figure 1: Graph depicting the scores of Stress and Betweenness Centrality: (a) This figure depicts Betweenness Centrality and Stress applied to WD40 interactome that resulted in projecting only few nodes with significant scores filtering all the less significant nodes; (b) Top 10 ranked nodes from the graph with node labels. It is noteworthy that WDR5 and RBBP7 are with significantly higher values that others the values are given in Table 1.

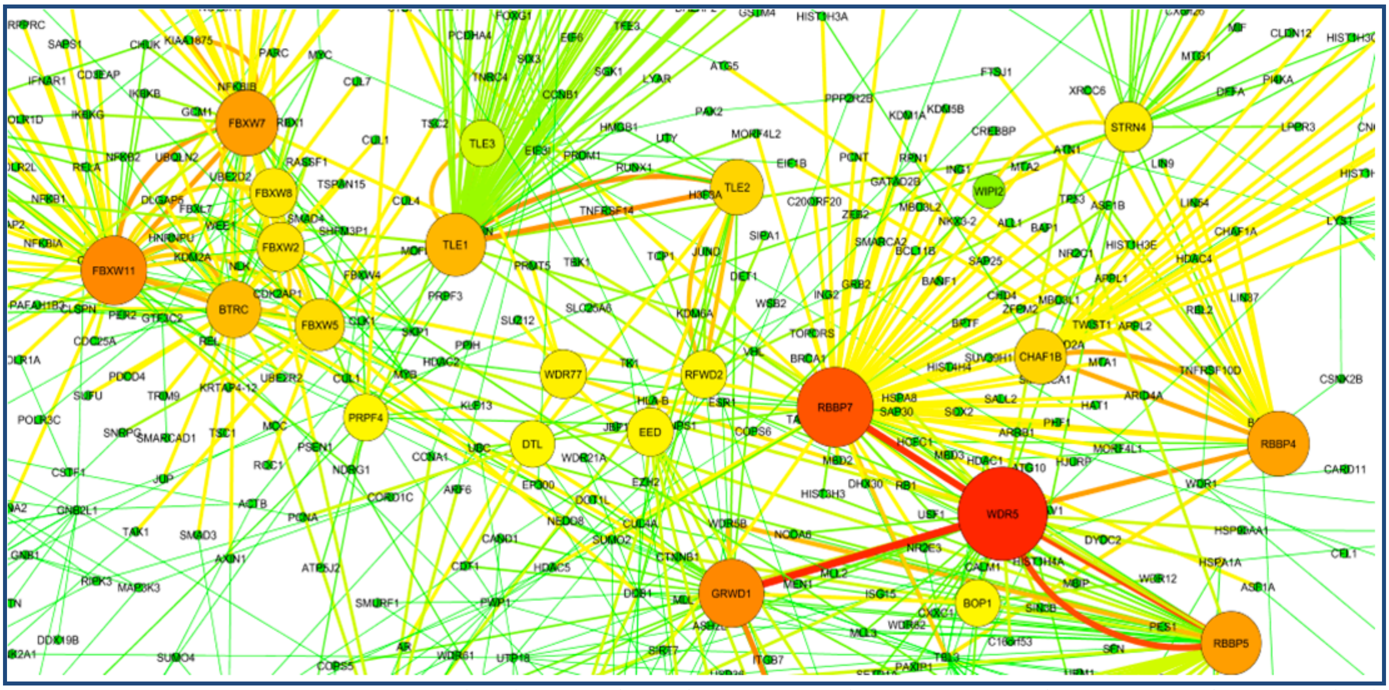

Figure 2: WD40 protein interaction network redrawn with emphasis to BC and stress score. High BC score attributes to red color and high stress score to bigger node size and interaction strength is designated by bolder lines attributing to edge betweenness.

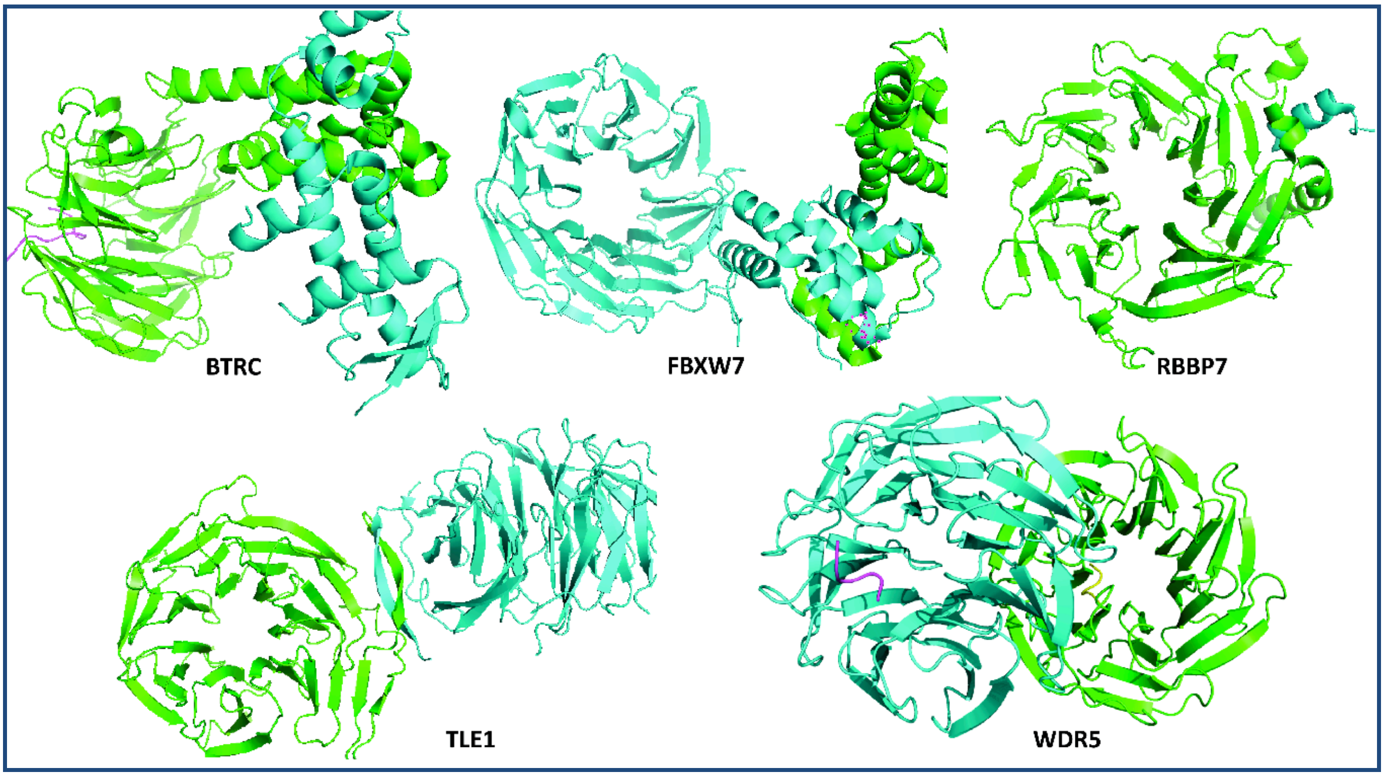

Figure 3: Available structures of top 10 ranked nodes from PDB database. PDBIDs are 1P22, 2OVP, 3CFS, 1GXR and 2G99 respectively from left to right in the above panel.

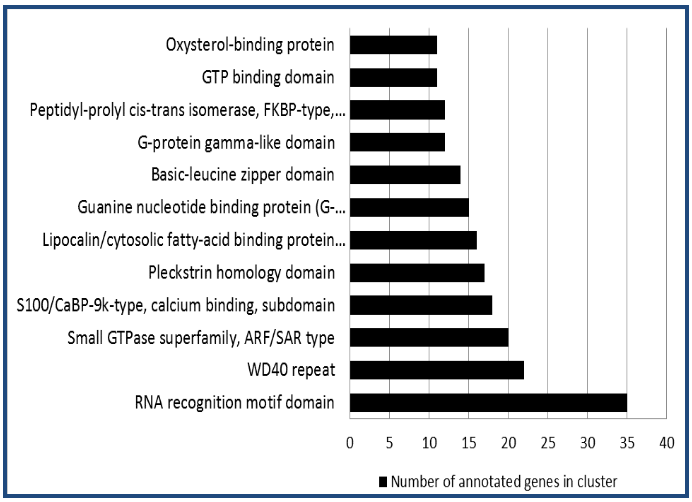

Figure 4: Domain based clustering of differentially expressed genes. WD40 proteins are second largest differentially expressed genes in the array. X axis denotes the number of annotated genes in the corresponding clusters present in Y axis. Datasets are available in supplementary material 2.

WD40 domains participate in a diverse set of interactions including those between globular proteins, 'induced-fit' interactions, but with conserved hydrogen bonds **[3]**, and those involving short peptides/linear motifs, which are most commonly involved in the assembly of transient complexes **[1]**. WD40-containing proteins can mediate interactions between DNA and histones, protein and RNA and host pathogen interaction complexes such as adhesion protein complexes of merozoites and human red cells **[4]** to name a few. Their diversity is illustrated also from their participation in complexes involved in a variety of processes in different cellular compartments, demonstrating that they are capable of adapting and interacting with the appropriate partner in entirely different functional contexts **[1]**.

Indeed, in the most recent yeast interactome, which is the most complete of all species, WD40 domains participate in more interaction pairs than any other domain. This is true both for datasets of yeast two-hybrid (i.e. binary interaction) **[5]** and of mass spectroscopy/tandem-affinity purification (MS/TAP) (i.e. multi-protein complex) experiments **[6]**. Notably, the number of WD40 domain-containing complexes found in the MS/TAP datasets easily surpasses the corresponding number of binary interactions in the yeast two-hybrid datasets, a finding that supports the notion that WD40 proteins act as scaffolds for assemblies of larger complexes. WD40 domains also rank among the top interacting domains in the available human interactome datasets. WD40 domains can thus act as large interaction platforms for multiple proteins, making them ideally suited to be hubs in cellular interaction networks **[8]**. Yet there is no work regarding the interactome of this family members.

Reports on interactome analysis such as the analysis of CBL family interactome and PDZ family interactome and their significance in control of immune systems, cell proliferation and cellular context based domain selectivity respectively **[6, 10]**. All these reports signify the role of interactome analysis in finding important hub proteins and network modules of biological importance. Reports propose that interactome analysis is also one of the methods for identification of therapeutic targets/markers since highly connected proteins with a central role in the network's architecture are three times more likely to be essential than proteins with only a small number of links to other proteins **[11].** Though the reports about the structural, functional and evolutionary aspects of WD proteins are available but insights about the protein interaction network is not available which may have resulted in many important nodes being missed from coming to light. Hence an interactome approach is essential that may uncover the important nodes in the family based on their centrality in the interactome. The present study propose a work pipeline where developing and analysis of curated interactome of WD40 family proteins is done leading to identification of highly connected nodes followed by clustering the interactome. In order to establish the importance of these nodes we performed microarray data analysis later validating expression data in cell culture models.

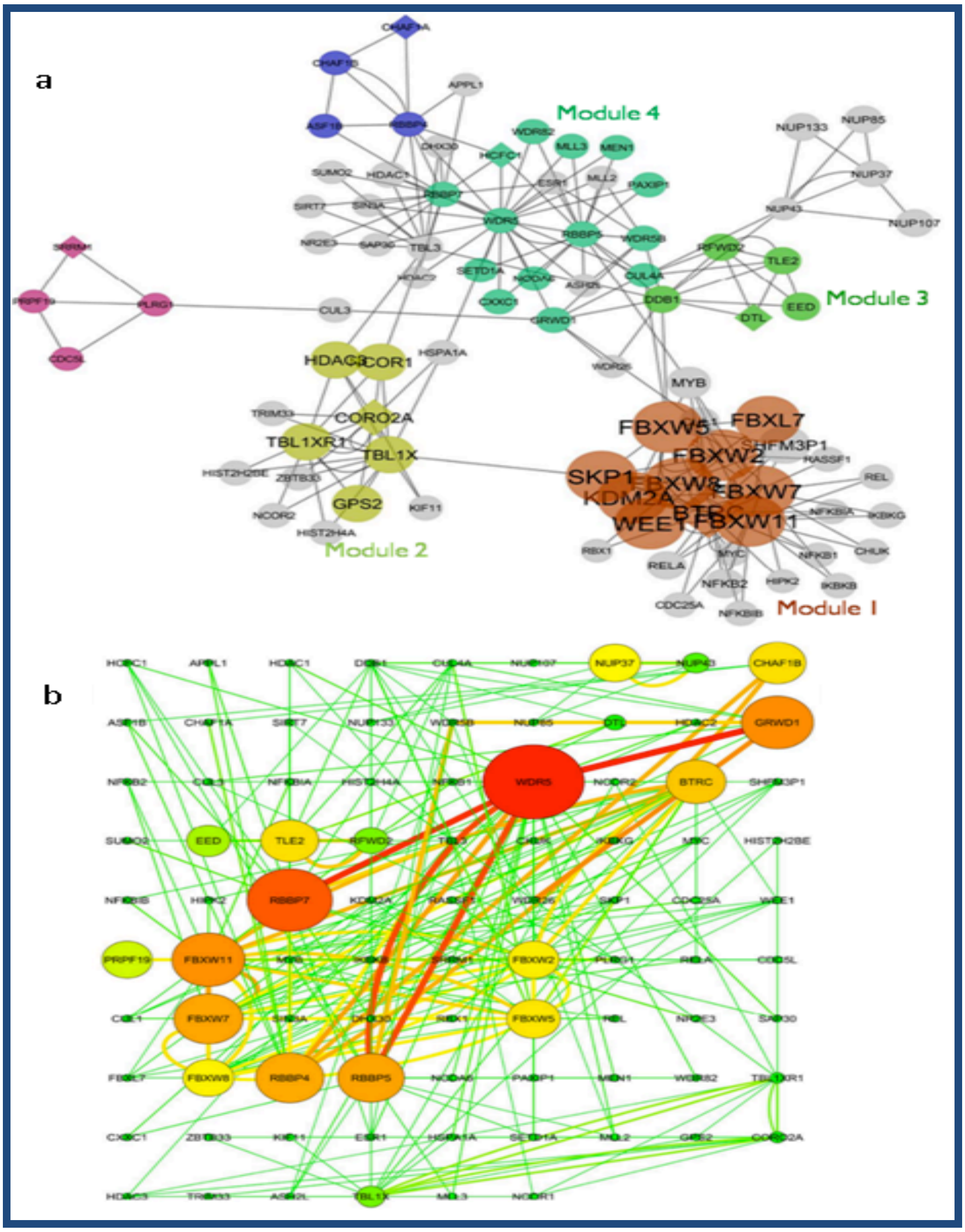

Figure 5: (a) clustering of interactome into modules using MCODE. Note that all the four clusters (each colored with different colour) are highly connected. Fuction of each module is given in table2; (**b)** The network when redrawn with BC and stress value highlights the top ten nodes. The bold and red colour line represent direct interaction based on edge betweenness scores. The figure also indicate that the WDR5 is central to all the clusters hence a true hub.

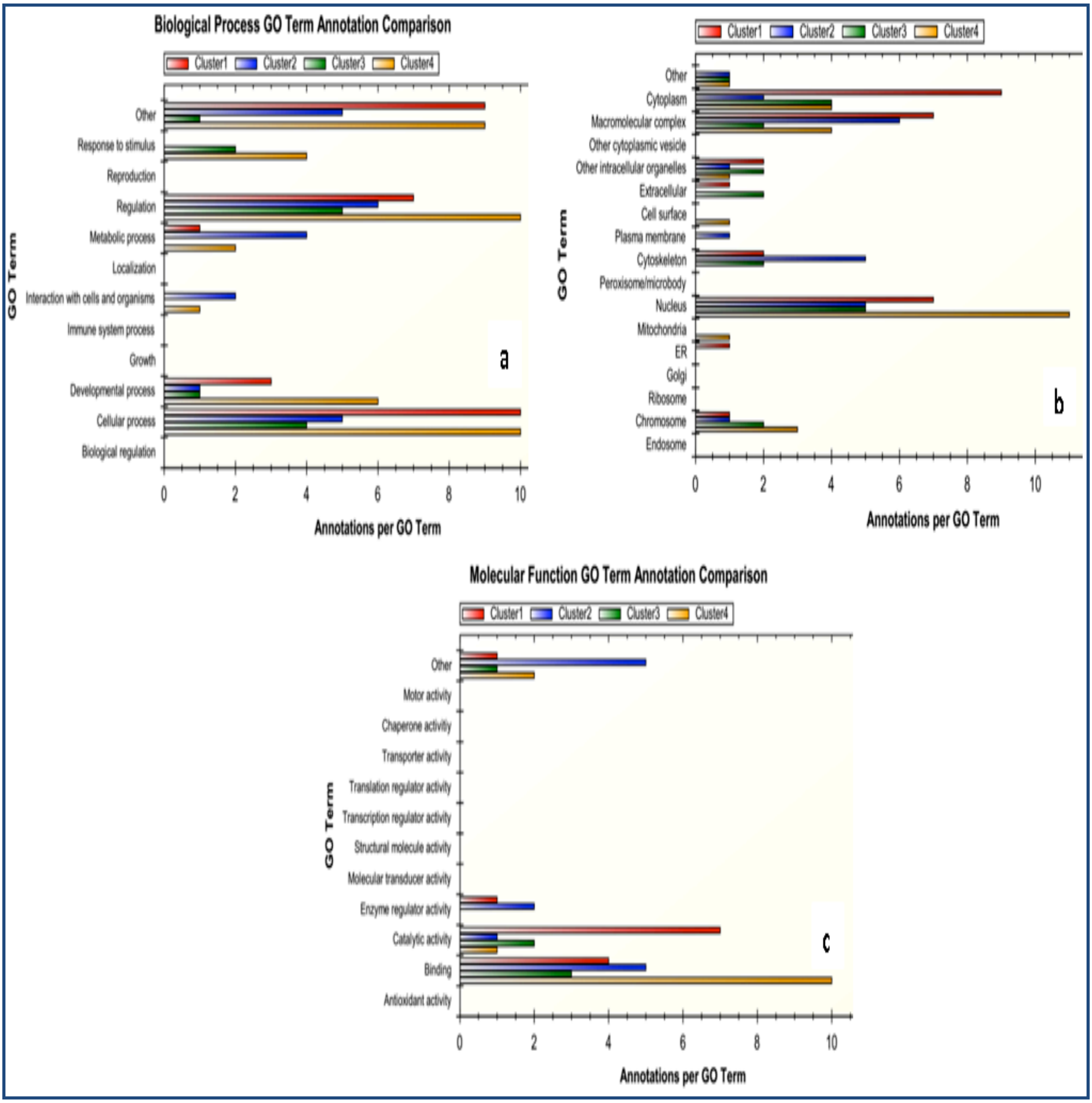

Figure 6: Cluster Level Gene Ontology analysis for MCODE derived clusters. X axis has number of annotation per GO term in a cluster and Y axis is GO term under consideration. Length of the bar in the graph designates the score towards particular GP term under consideration: **a)** GO of Biological Process; **b)** GO of Cellular Component and **c)** GO of Molecular Function. Note that, cluster 4 has maximum score for binding property, regulation activity and nuclear localization.

Methodology:

Collection of sequences and interaction data

WD sequences were retrieved by using the profile HMMs from Pfam database. Each sequence was manually curated for the gene name using NCBI GENE and Uniprot database. Sequence redundancy was removed manually. Gene IDs were used for querying against various interaction databases including STRING 8.0, BIOGRID, and HPRD to include all possible nodes reported in different repositories. The interactions were verified by gene name based literature survey using PUBMED and iHOP databases with keywords "binds" and "interact". The nodes which were represented with their aliases were renamed to the standard gene name to avoid false interaction data.

Network construction and analysis:

The networks for each WD protein were saved in "sif" format and were loaded into Cytoscape 2.8.0. Network was merged to obtain a union of the protein interaction. Network analysis was done with Centiscape 1.1 and Network Analyzer plug-in to analyse the Betweenness Centrality (BC) and Stress value. Since both BC and stress are good measures to identify the network hubs **[19].** We wanted to verify that whether the identified nodes are subjected to variation by isolated binary interactions, hence MCODE plug-in was used to dissect the interactome to locate the high density clusters**. [14].** Further, the nodes in the clusters were checked for their evidence at mRNA level in different tissues by searching literature and GENECARDS database. Gene ontology was analysed at cluster level using STRAP tool.

Gene expression analysis:

Ravnskjaer *et al* (2013) has shown that WDR5 induces insulin resistance in transfected liver cell models. Since WDR5 was found to be one of the hub protein in the network we wanted to check for the implication of other hub proteins in insulin resistance. After identification of most connected nodes, Gene expression profile GSM524162 was retrieved from NCBI GEO database and expression analysis was done with dCHIP software as described by Marselli *et al.* **[23]** Along with the calculation of fold change for genes, we also did the domain based clustering of significantly modulated genes.

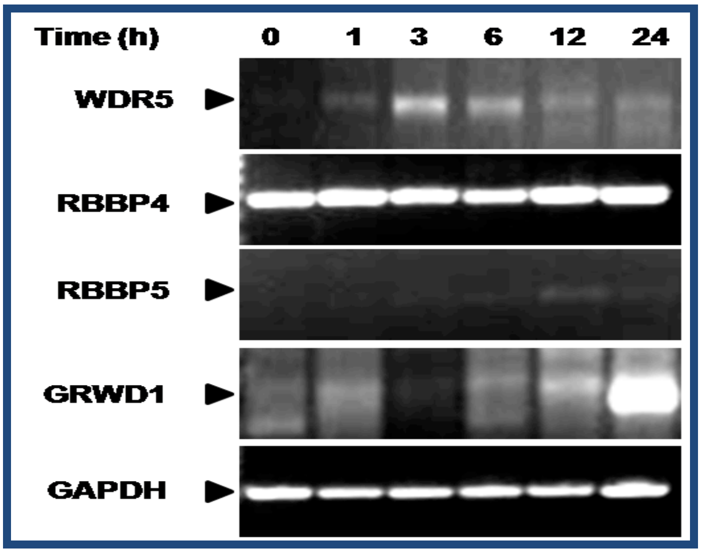

Figure 7: Gene expression validation by semi quantitative reverse transcription-PCR in FFA induced insulin resistant HepG2 model. WDR5 showed expression at 3Hr and 6hr, RBBP5 showed expression at 6Hr and 12hr and GRWD1 showed expression at 6hr, 12 Hr and 24 hr time points. Among all these genes WDR5 expressed early followed by RBBP5 and GRWD1. RBBP4 was found to be expressed at all-time points

Gene expression analysis in HepG2 cells

In order to validate the dry lab data of microarray and network analysis results, we carried out a simple expression study of top ranked nodes in the cell culture models by semi quantitative gel based method. HepG2 cells were cultured and perturbation was done according to Lee *et al* **[28].**

The following primers were used for reverse transcription semi quantitative gel based expression studies:

hWDR5,
sense 5′-TCCTCCGTGAAATTCAGCCC-3′ and
antisense 5′-ATCGATGAGCGTCTTCAGGC-3′;

hRBBP4,
sense 5′- CAAGACTGTTGCCTTGTGGG-3′ and
antisense 5′- GTCCTTCTGGATCCACGCTT-3′;

hRBBP5,
sense 5′- GCATCCATTTCCAGTGGAGT-3′ and
antisense 5′-TGGTGACATCCACTTCCTCA-3′;

hGRWD1
sense 5′- TGCAGACCACCAGATCACAC-3′ and
antisense 5′- AGATGGTGAAGCCTGACAGC-3′;

hGAPDH,
sense 5′-CGGAGTCAACGGATTTGGTCGTAT-3′ and
antisense 5′-AGCCTTCTCCATGGTGGTGAAGAC-3′;

Amplification conditions followed were: denaturation at 94 °C for 30 s, annealing at 55~60 °C for 20 s, and extension at 72 °C for 40 s for 30 cycles. The PCR products were resolved and visualized on 1.5% agarose gel and stained with ethidium bromide.

Results & Discussions:

WD proteins are among the top ten most abundant family of proteins in eukaryotes and among top interacting proteins **[9].** The growing number of available WD40 domain complex structures has led to a better understanding of function of WD40 domain **[8].** Construction and analysis of human interactome can be done by retrieving the total interaction data from public repositories and enriching with HGNC accession numbers **[7, 16].** However, these databases may contain the false positive interactions **[17],** where such interactions can be removed by curation improving the accuracy. Hence in the present study, the literature based curation of the interaction data using iHOP and PUBMED databases was done **[18].** The curated interactome contains 1110 proteins (nodes) having 1337 interactions (edges). The network presented here is not directional network, as there are no signaling or metabolic cascades involved and WD proteins are devoid of enzymatic activity **[9].** WD40 proteins play a crucial role in diverse protein-protein interactions by acting as scaffolding molecules and thus assisting the proper activity of the proteins **[20, 2]** and hence the network statistics has been worked out by considering the network as non-directional. Therefore, BC and stress are considered as ideal measure of centrality [19]. BC has a large influence on the transfer of items through the network, under the assumption that, item transfer follows the shortest paths. Stress is calculated by measuring the number of shortest paths passing through a node **[12, 13].**

We identified top 10 ranked nodes from the interactome through the scatter plot with very high degree of correlation between BC and stress (R² = 0.9653) for all the top 10 ranked nodes and taken for further analysis **Table 1 (see supplementary material) & (Figure 1a & 1b)** the available 3D structures for these proteins are given in **Figure 2**. The interactome was redrawn using the values for nodes given in graph **(Fig 1a)** with emphasis to node size for BC scores and node colour to stress scores to get the better representation of the interactome and edges were also given emphasis based on edge betweenness **(Figure 2).** We found WDR5, RBBP7, RBBP5, RBBP4 and GRWD1 are among the top ranked nodes experiencing the high BC and stress scores, there exists a high edge betweeness among these nodes. Higher the edge betweenness score bolder the edge. Edge betweenness represent the number of the shortest paths that go through an edge in a graph or network **[13].**

No	Nodes	(BC) Score	Stress Score	Conserved in
1	WDR5	4.97E-04	625	Eukaryota
2	RBBP7	3.51E-04	427	Euteleostomi
3	FBXW11	2.37E-04	292	Bilateria
4	GRWD1	2.30E-04	294	Eukaryota
5	FBXW7	2.03E-04	241	Bilateria
6	RBBP4	1.94E-04	230	Eukaryota
7	RBBP5	1.79E-04	243	Eukaryota
8	TLE1	1.74E-04	176	Tetrapoda
9	BTRC	1.13E-04	160	Bilateria
10	TLE2	9.87E-05	100	Boreoeutheria

Table 1: this table lists the top 10 ranked nodes in the interactome based on Betweenness Centrality (BC) and Stress Scores. This table also shows the level of conservation in the tree of life for each node. Note that only WDR5, GRWD1, RBBP4 and RBBP5 are conserved throughout eukaryotes. All the other nodes are present at the higher level in tree of life. Datasets can be found in supplementary material 1.

The network was dissected into four clusters by MCODE, each represent functional module, which were ranked according to network density **(Figure 5a) & Table 2 (see supplementary material).** Resultant clusters when analysed with BC and stress yet again WDR5 was found to be the most central and hence most significant node **(Figure 5b)** implying that, WDR5 is central to high density networks, hence is a true hub connecting the four modules. Searching in GENECARDS database revealed that member of all 4 clusters are found to be present in almost all organs with evidence at mRNA level. Member of cluster 4 are among the top 10 ranked nodes which are participating in regulatory activity **(Figure 6a)** being most of them inside nucleus **(Figure 6b)** and having binding activity **(Figure 6c)** according to STRAP score. All the above observations add to the importance of WDR5 and other top ranked nodes.

WDR5 is known for its role in different types of cancer, but recently it has also been demonstrated in rat models to regulate the expression of genes involved in gluconeogenesis in response to glucagon signalling. It has been shown that, WDR5 stimulated the gluconeogenic program through a self-reinforcing cycle and thereby promoting insulin resistance. **[21].** In view of these facts on importance of WDR5, in the present study, the microarray data analysis was carried for the Geo profile GSM524162 **[23].** The importance of WD40 repeats is very well demonstrated by domain based clustering of the significantly modulated genes showing the second largest group of modulated genes that belong to WD40 family **(Figure 4).** Similar work has been done in foxtile millet where *Insilico* gene expression analysis and subsequent PCR based validation of WD40 proteins have been done that has led to identification of draught response genes **[20].**

Cluster	Proteins present	Network Density	Molecular Function	Biological Process	Cellular Component
Cluster 1	KDM2A, SKP1, FBXW5, FBXW11, FBXW2, FBXW7, FBXW8, FBXL7, BTRC, WEE1	4.889	Catalytic Activity	Ubiquitin Mediated Proteolysis	Cytoplasm
Cluster 2	HDAC3, CORO2A, TBL1XR1, TBL1X, NCOR1,GPS2	2.0	Binding	Regulation of Transcription	*Macro Molecular Complex*
Cluster 3	TLE2, RFWD2, DDB1, DTL, EED	1.5	Binding	Regulation of cell cycle/transcription	Nucleus
Cluster 4	GRWD1, CXXC1, WDR5B, PAXIP1, CUL4A, WDR5, NCOA6, SETD1A, RBBP5, HCFC1, MLL3, RBBP7,MEN1, WDR82	1.25	Binding	Chromatin Organization	Nucleus

Table 2: This table depicts the list of nodes from MCODE dissected clusters. Only cluster 1 found to have catalytic activity whereas all the other clusters are having binding activity.

WDR5 being a member of MLL complex, exhibits different effects on different tissues involved in diabetic complications. MLL complex contains MLL and WDR5 as core complex together with ASH2L and RBBP5 as accessory proteins **[24]** and has been demonstrated to increase adipogenesis, **[22]** and WDR5 has been found to recruit KAT2B and increase the expression of gluconeogenetic genes in liver and thus promoting impairment of systemic glucose homeostasis **[21].** WDR5 also negatively modulates the insulin granule biogenesis and insulin secretion in pancreatic beta cells thus affecting the insulin release **[25].** However, the implication of other nodes in insulin resistance is not clear. Among the 10 ranked nodes only WDR5, GRWD1, RBBP4, RBBP5 were found to be conserved throughout Eukaryota as per NCBI homologene database **(Table 1).** Previous reports have shown the physical interaction between WDR5 and RBBP4, WDR5 and GRWD1 **[26]** but there exists no data about its significance. Reports that there exist a high probability that interacting proteins are co-expressed **[27].** Hence in the present study, these four genes were taken for reverse transcription semi quantitative PCR (rtPCR) analysis in which GAPDH was used as housekeeping gene. Palmitate induced insulin resistance model using HepG2 cells in the present study resembles more closely to the physiological changes in vivo **[28].** The expression study showed WDR5, RBBP5, GRWD1 genes being expressed at different time points, expression of WDR5 at 3hr time point, RBBP5 at 3hr, 6hrs, 12 hrs time points and GRWD1 at 24 hrs time point **(Figure 7)** Zero hour time point is as good as control or untreated.

Therefore, the expression patterns of these genes in response to the perturbation leading to insulin resistance substantiate the gene expression data of the present study. Also upregulation of RBBP5 and GRWD1, indicates that these proteins may also contribute to insulin resistance along with WDR5. WD40-repeat protein can possess multiple functions depending on its direct and indirect interactions partners **[20].** Thus we hypothesize the possibility of WDR5 and GRWD1 interaction in insulin resistance. However, we were unable to measure the expression levels of RBBP7 and RBBP4 was expressed at same level at all the time intervals, the reason is not clear hence need detailed study. The genes that showed upregulation in the semi quantitative rt-PCR analysis **(Figure 7)** may be taken further to evaluate them as possible diagnostic markers for diabetes as the normal cells are transformed to insulin resistant cells.

Conclusion:

We report the important nodes of WD protein family using protein interaction network that are associated with known pathological conditions. The pipeline described for the interactome is helpful in hub identification. The data gleaned was further validated *in vitro*. Thus, the implied role of WDR5 and GRWD1 interaction involved in insulin resistance is hypothesized.

References:

[1] Li D & Roberts R, *Cell Mol Life Sci*. 2001 **58**: 2085 [PMID: 11814058]

[2] van Nocker S & Ludwig P, *BMC Genomics*. 2003 **4**: 50 [PMID: 14672542]

[3] Wu XH *et al. Proteins*. 2010 **78**: 1186 [PMID: 19927323]

[4] von Bohl A *et al. Malar J*. 2015 **14:** 435 [PMID: 26537493]

[5] Yu L *et al. Protein Sci*. 2000 **9**: 2470 [PMID: 11206068]

[6] Collins SR *et al. Mol Cell Proteomics*. 2007 **6**: 439 [PMID: 17200106]

[7] Ori A *et al. J Biol Chem*. 2011 **286**: 19892 [PMID: 21454685]

[8] Stirnimann CU *et al. Trends Biochem Sci*. 2010 **35**: 565 [PMID: 20451393]

[9] Xu C & Min J, *Protein Cell*. 2011 **2**: 202 [PMID: 21468892]

[10] te Velthuis AJ *et al. PLoS One*. 2011 **6**: e16047 [PMID: 21283644]

[11] Jeong H *et al. Nature*. 2001 **411**: 41 [PMID: 11333967]

[12] Brandes UJ *Math. Soc.* 2001 **25**: 163 [DOI:10.1080/0022250X.2001.9990249]

[13] Barthelemy M, *Physical Journal B*, 2004 **38**: 163 [DOI: 10.1140/epjb/e2004-00111-4]

[14] Bader GD & Hogue CWV, *BMC Bioinformatics*. 2003 **4**: 2 [PMID: 12525261]

[15] Rivero O *et al. PLoS One*. 2010 **5**: e12254 [PMID: 20805890]

[16] Randhawa V *et al. OMICS*. 2013 **17**: 302 [PMID: 23692363]

[17] Raman K *et al. Mol Biosyst*. 2009 **5**: 1740 [PMID: 19593474]

[18] Hoffmann R & Valencia A, *Nat Genet*. 2004 **36**: 664 [PMID: 15226743]

[19] Scardoni G & Laudanna C, *InTech publishers*. 2012 ISBN: 978-953-51-0115-4.

[20] Mishra AK *et al. PLOS One*. 2014 **9**: e86852 [PMID: 24466268]

[21] Ravnskjaer K *et al. J Clin Invest*. 2013 **123**: 4318 [PMID: 24051374]

[22] Lee J *et al. Proc Natl Acad Sci U S A*. 2008 **105**: 19229 [PMID: 19047629]

[23] Marselli L *et al. PLoS One*. 2010 **5**: e11499 [PMID: 20644627]

[24] Song JJ & Kingston RE. *J Biol Chem*. 2008 **283**: 35258 [PMID: 18840606]

[25] Li H *et al. EMBO Rep*. 2014 **15**: 714 [PMID: 24711543]

[26] Higa LA *et al. Nat Cell Biol*. 2006 **8**: 1277 [PMID: 17041588]

[27] Ramani AK *et al. Mol Syst Biol*. 2008 **4**: 180 [PMID: 18414481]

[28] Lee JY *et al. Metabolism*. 2010 **59**: 927 [PMID: 20006364]

Edited by P Kangueane

Citation: **Kumar *et al.*** Bioinformation 12(2): 54-61 (2016)