Understanding The Intra And Inter-Cellular Interaction Complexities And Flexibilities Using Systems And Sequence Analysis Approach

Corrected Copy
Thesis submitted for the Degree of
Doctor of Philosophy (Science)
in
Biochemistry

By
Ishita Mukherjee

Department of Biochemistry
University of Calcutta
2020

Acknowledgements

I would like to extend my gratitude to my advisor Dr. Saikat Chakrabarti; I appreciate his contributions of time, thoughts, and funding to make my Ph.D. productive. He has been instrumental in involving me in multiple collaborative projects which have helped me to shape a broader perspective. I would also like to acknowledge Dr. Suvendranath Bhattacharryya and Dr. Sudipto Roy for their valuable inputs in these collaborative projects.

I would like to thank my parents, sister in-law and especially my brother Anirban, for their support and encouragement during this journey. I am also grateful to my friends Sumanta and Anand for their encouragement, ideas and inspiration.

I would like to thank the institute for infrastructural support and for making my journey eventful with informative talks from eminent personalities and diverse conferences. I would like to acknowledge my professors, Director (IICB) for their guidance. I am also grateful to my co-authors Avijit, Dipayan, Sudarshanadi, Susanta and my lab members particularly Abhijitda, Madhumitadi, Aneeshadi and project students Sunandan, Wridhisom for their support.

Ishita Mukherjee, M.Sc.
CSIR-Indian Institute of Chemical Biology

Dedicated to the alpha family

Table of Contents

Abstract

The present thesis work has been undertaken to gain an understanding of intra-cellular or inter-cellular interactions between bio-molecular entities utilizing either a systems analysis based perspective or different sequence analysis approaches. During this study different principles likely to be prevalent among intra-cellular and inter-cellular interactions have been studied with the help of computational approaches. Broadly, the complexities in intra-cellular interactions have been studied by determining the effect of perturbations such as over-expression or down-regulation of a key regulator on the intra-cellular interaction network architecture or its components. In particular, network analysis of regulatory network proteins in association with the intra-cellular protein-protein interaction network, led to a key observation that topologically important effector proteins in the regulatory network could be important signaling proteins. Identification of such important effector proteins essential for the regulatory network integrity of a key regulator may be performed by network analysis. It is likely that alterations in these important effector proteins may lead to disruptions in cellular physiology and as such in this manner probable disease associated entities can be determined. Alternately, the flexibility among protein-protein interactions has been studied by analyzing homologous sequence families of interacting proteins with the help of information theory based measures like mutual information and Bhattacharyya co-efficient. Since interacting proteins may co-evolve, co-variation may allow the preservation of a functional interaction between co-evolving proteins and inter-dependent residue pair alterations may occur as a result of evolutionary pressure. Analysis of molecular co-evolution in inter-cellular protein interaction complexes determined that co-evolutionary pairings may be present among interface and non-interface residue pairs and such positions are likely to be crucial for a functional interaction between these sets of proteins. Therefore, utilising information contained in biological sequences, co-evolutionary pairings involving structurally or functionally crucial residue positions in disease associated inter-cellular protein-protein interaction complexes were predicted. Thus, different computational approaches have been utilised to study a particular hypothesis in a disease scenario in order to delineate certain themes prevalent in intra-cellular or inter-cellular interactions among bio-molecular entities while predicting disease associated entities or studying interaction patterns among them.

Introduction

The primary objective of this work lay in understanding intra-cellular or inter-cellular interactions between bio-molecular entities either at the network level or within an interaction complex. Interactions between bio-molecules results in the culmination of cellular processes or functions. Thus, systematic analyzes of physical interaction networks such as protein-protein interaction network, transcriptional regulatory network, miRNA-mRNA regulatory network etc. may illuminate different properties of intracellular biological networks. For this purpose, certain hypothesis regarding biological intra-cellular interaction network topology and robustness have been taken into consideration. In particular, the applicability of scale free network properties to transcriptional regulatory network in association with protein-protein interaction network has been studied and analyzed to identify essential proteins for a cellular process. Additionally, the effect of down-regulation of an essential entity on the network architecture in terms of alterations in network components or their regulatory relationships has been studied to determine disease associated entities or altered network components. Further, protein-protein interactions may govern inter-cellular communication or signaling cascades and as such insights into evolutionarily conserved interaction paradigms may be derived based on sequence and structural analyses of interacting protein families. In this respect, with the help of extensive homology searches proteins likely to be involved in inter-cellular interactions may be determined. Additionally, utilizing sequence analysis methods residue pairing combinations important for functional conservation of inter-protein interactions may be identified and this information might allow one to modulate these interactions in disease processes.

The complexities in intra-cellular interactions have been explored by modulating the expression levels of a key regulator and studying the effect of this perturbation in the intra-cellular network with the help of systems based computational analyses approaches with the following objectives,

To identify essential proteins in intra-cellular interaction network architecture with the help of graph theory based network analysis concepts in turn to predict probable disease associated proteins.

To determine the effect of gradual alterations in levels of an essential regulator on the intra-cellular regulatory network components in order to envisage network component changes associated with altered cellular responses over time.

To study unusual regulatory relationships among intra-cellular miRNA-mRNA

interactions in disease conditions in order to identify disease associated network components.

To determine and analyze intra-cellular protein-RNA regulatory network in order to determine probable disease associated changes in network architecture.

To predict and verify probable regulatory relationships in miRNA-mRNA interaction networks

Alternately, flexibilities in inter-cellular interactions have been explored with the help of multiple sequence analysis approaches to achieve the following objectives,

To determine proteins likely to be involved in inter-cellular interactions based on the concept of lateral gene transfer with the help of sequence analysis approaches

To predict co-evolutionary residue pairings in interacting proteins involved in inter-cellular interactions in order to identify functionally relevant residue positions

Thus, in turn different disease associated entities have been determined or interaction patterns among them have been studied either at the systems level or at the structural/molecular interaction level in intra-cellular or inter-cellular bio-molecular interactions. These inferences regarding intra-cellular interactions and the patterns observed in inter-cellular interactions during this discourse have been discussed in the present thesis subsequently.

Literature Review

Intra-cellular interaction networks

The intra-cellular environment contains different bio-molecules that interact and network to carry out cellular processes and these biological processes may be represented as networks containing the different molecular components within a cell as nodes and their interactions as edges or links. A network in terms of interaction maps might hold information about how different intra-cellular molecules operate in a coordinated manner with other bio-molecules to enable different biological processes within the cell. Typically, protein-protein interaction networks (PPIN) may outline which proteins interact under a given physiological condition whereas a regulatory network might detail the regulator (transcription factor or miRNA) and its targets (genes or mRNA).

Systems biology approach in biomedical research

Systems biology involves utilizing experimental and computational approaches to provide a framework for studying biological problems with the help of systematic measurements (Chuang, Hofree, & Ideker, 2010). Systematic assessment of disease conditions utilizing high throughput technologies that measure genome wide transcriptomic changes allows one to study interactions between the cells constituent parts at a systems or global level. Such an approach attempts to capture or study cellular functions that are governed by a large and complex network of combinatorial interactions among genes or proteins (Chuang et al., 2010). Therefore, a combinatorial systems analysis based approach is primarily useful for identifying multiple genes that may be associated with complex disorders that arise as a result of a combination of factors (polygenic, environmental and lifestyle). Further, network analysis allows one to study and predict changes in network components or architecture that may arise as a result of these disease associated genes. A few examples of complex diseases include Alzheimer's disease, cancer, non-alcoholic fatty liver disease etc.

Alzheimer's disease: A progressive and debilitating neurodegenerative disorder, Alzheimer's disease (AD) clinically presents as memory deficits and cognitive decline. However, pathological presentation includes intracellular neurofibrillary tangles and extracellular amyloidal protein deposits leading to senile plaques (amyloid-beta (Aβ) plaques) within the brain (DeTure and Dickson, 2019). It is believed that neuronal function and structure degenerates progressively ultimately leading to neuronal death. A number of factors may be involved in the pathogenesis of AD but the molecular mechanism underlying the disease is not clear (Moradifard et al., 2018). One of the underlying causes may be aberrant spatial and temporal expression or dysfunction of brain-enriched miRNAs which under normal circumstances fine tune the expression of a wide range of target mRNAs essential for neuronal and glial development (Satoh 2012).

MicroRNAs play a key role in regulating the expression of APP and BACE1 which suggests that they have a role in Aβ production. In addition, multiple miRNA have been identified with roles in Aβ clearance, Aβ induced synaptic failure and Tau phosphorylation imbalance. Thus, a number of studies have suggested that miRNA such as miR-9, miR-146a, miR-107, miR-106b, let-7, miR-15a, miR-485-5p, miR-101, miR-128 etc. may be involved in the Aβ hypothesis and the Tau hypothesis of AD pathogenesis (Wang et al., 2019; Satoh 2012). In particular, one miRNA can concurrently down-regulate hundreds of target mRNAs and each mRNA may be co-regulated by multiple miRNA resulting in a biologically integrated network of functionally associated molecules. Some of these functionally associated molecules may be involved in cellular processes such as cell cycle progression that are aberrant as a result of deregulated miRNA target networks contributing to the pathogenesis of AD (Satoh 2012). Thus, characterization of de-regulated miRNA-mRNA interaction networks in AD might broaden our understanding of miRNA-mediated molecular mechanisms underlying AD.

Cancer: Genetic instabilities such as nucleotide substitutions, insertions or deletions, chromosomal copy number alterations, DNA rearrangements etc. may allow normal cells to acquire additional biological capabilities to establish a tumor or neoplastic disease such as cancer (Garraway and Lander, 2013). Certain hall mark capabilities of cancerous cells include sustained proliferative signalling, evasion of growth suppressors, resistance to cell death, replicative immortality, angiogenesis, active invasion and metastasis, reprogrammed energy metabolism and immune destruction evasion (Hanahan and Weinberg, 2011). Cancer encompasses over a hundred distinct diseases which originate from most of the cell types and organs of the human body, however, with diverse risk factors and epidemiology. Large-scale cancer genomics performed on a tissue site–specific basis to study the cancer genome landscape have identified a basic catalog of DNA-based alterations which are likely to be involved in the onset and progression of cancer. Further, several possible types of alterations may occur at a given gene or locus in the cancer genome and they have different functional consequences (Mardis 2018). Alternately, depending on their target genes, miRNA are also likely to function as oncogene or tumor suppressor. Moreover, miRNA profiling and deep sequencing studies have provided evidences that miRNA expression is dysregulated in cancer via various mechanisms. Amplification or deletion of miRNA genes, abnormal transcriptional control of miRNAs, dysregulated epigenetic changes and defects in miRNA biogenesis machinery have been shown to affect the hallmarks of cancer, such as sustained proliferative signalling, evasion of growth suppressors, resistance to cell death, invasion and metastasis, and angiogenesis (Peng and Croce, 2016).

Non-alcoholic fatty liver disease: A liver disorder, non-alcoholic fatty liver disease (NAFLD) encompasses a wide spectrum of pathological conditions ranging from steatosis (accumulation of fat in the liver without inflammation) to steatohepatitis with inflammation and percicellular fibrosis, liver fibrosis, cirrhosis and hepatocellular carcinoma (HCC) (Berná et al., 2020). Nonalcoholic fatty liver (NAFL) or hepatic steatosis and non-alcoholic steatohepatitis (NASH) are associated with multiple risk

factors which could be genetic, environmental, microbiome related or metabolism associated including dyslipidemia (Friedman et al., 2018). Toxic lipid species may accumulate when the liver's capacity to handle primary metabolic energy substrates such as carbohydrates and fatty acids is overwhelmed. Thus, upon excess fatty acids intake or their impaired disposal lipotoxic species may be generated which contribute to ER stress and hepatocellular stress, injury and death. Stress in turn may lead to fibrogenesis and genomic instability which serve as a predisposition to cirrhosis and hepatocellular carcinoma (Friedman et al., 2018). mRNA expression profiling can be utilised to uncover molecular mechanisms and explore diagnostic or predictive biomarkers in complex diseases such as NAFLD. It has been reported that metabolism-related genes and specific metabolic and repair pathways get activated whereas uptake transporter genes get down-regulated in NAFLD. Further, lipid metabolism and cholesterol biosynthesis genes have mainly been found to be altered during the pathological development of NAFLD (Zhu et al., 2018). Additionally, miRNA which regulate the expression of multiple target genes such as miR-122, miR-451, miR-33a/miR-33b*, miR-200b and miR-27 have altered expression profiles and are likely to be associated with the pathophysiology of NAFLD. Moreover, one of the hallmarks of the metabolic syndrome, dyslipidemia, has been shown to be strongly associated with NAFLD. Further, dietary fat and cholesterol may lead to the development of hepatic steatosis and NASH. In particular, animal model studies have shown that high-fat feeding induces development of fatty liver, associated hepatic inflammation and hepatic steatosis in a time-dependent manner (Jensen et al., 2018). Therefore, systematic meta analysis of hepatic mRNA and miRNA expression profiles in high fat fed animal models may enhance our understanding of molecular mechanisms associated with NAFLD disease progression.

Primary ciliary dyskinesia: Genetic abnormalities in one or more ciliary genes leading to developmental defects in motile cilia may give rise to a common ciliopathy known as primary ciliary dyskinesia (PCD). Motile cilia are microtubule based hair-like organelles composed of a structural core, basal body, transition zone, ciliary membrane, ciliary tip and axoneme having 9+2 microtubular architecture with dynein arms (Fliegauf, Benzing, & Omran, 2007). Further, cilia have diverse tissue specific roles in different physiological and developmental processes such as fluid clearance, cellular motility, sensory reception or signaling and as such mutation or defects in one or more ciliary genes or proteins can result in abnormalities in formation or function of these organelles (Fliegauf et al., 2007; Horani, Ferkol, Dutcher, & Brody, 2016). Defects in the structural organization of cilia or regulation of their assembly can result in disrupted development of body pattern or physiology of a number of organ systems (Bisgrove & Yost, 2006).
Clinically associated with chronic respiratory infections, bronchiectasis, infertility and in certain cases, laterality defects or hydrocephalus PCD happens to be one of the most prevalent of ciliopathies (Zariwala, Omran, & Ferkol, 2011). However, PCD which is inherited as an autosomal recessive trait exhibits variability in clinical phenotype and not all mutations in disease causing genes are exhibited as defects in ciliary ultra-structure. The genetic basis of PCD has conventionally been studied utilizing family based or genome-wide linkage studies, homozygosity mapping and candidate gene testing to identify causal PCD genes; however, locus heterogeneity has been a major challenge. In

addition, genome and exome sequencing studies have identified multiple disease associated genes in families of PCD patients during the last decade (Zariwala et al., 2011; Horani et al., 2016). Presently, the OMIM database documents about 35 disease causing genes with mutations associated with PCD (McKusick, 2007; Amberger, Bocchini, Schiettecatte, Scott, & Hamosh, 2014). Since, defects in cilia formation or function may result in disrupted development of body pattern or physiology of multiple organ systems, determination of a genetic screening test for PCD causing genes will help in disease diagnosis (Zariwala et al., 2011). For this purpose, disease causing variants may be identified by sequencing PCD patients. Alternately, identifying genes and proteins crucial for motile cilia biogenesis and mis-expressing them during ciliary development studies in model organisms allows the identification of additional genes or proteins possibly associated with abnormal ciliary development or PCD. In this respect, large scale studies have identified thousands of proteins in the ciliary proteome that coordinately interact to form these microtubule based hair-like organelles (Boldt et al., 2016). Moreover, macromolecular synthesis and assembly of all of these ciliary structures is a complex and coordinately regulated process that involves cascades of transcription factors like E2f4, E2f5, Rfx1, Rfx2, GemC1, McIdas, Myb, Rfx3, Rfx4, and FoxJ1 (Arbi et al., 2016; Choksi, Lauter, Swoboda, & Roy, 2014; Danielian, Hess, & Lees, 2016; Vladar & Mitchell, 2016). These transcriptional regulators organize the development and function of cilia wherein Rfx factors regulate both immotile and motile cilia genes; however FoxJ1 typically regulates motile cilia biogenesis and appears to be its master regulator (Choksi, Lauter, et al., 2014). Since, FoxJ1 regulates a set of genes [FoxJ1 induced genes (FIG)] that are sufficient for motile cilia development and function it has been coined as the motile cilia master regulator (Stubbs, Oishi, Izpisúa Belmonte, & Kintner, 2008; X. Yu, Ng, Habacher, & Roy, 2008; Choksi, Babu, Lau, Yu, & Roy, 2014). FoxJ1 over-expression, down-regulation or knock-out in experimental model systems might allow the identification of genes essential for motile cilia development (Choksi, Babu, et al., 2014). Finally, additional targeted gene knock-down or knock-out studies in experimental model systems considering the genes identified to be essential for motile cilia formation or function may help in delineation of PCD causal genes and further our understanding of pathophysiology in PCD.

Transcriptomic profiling to study human diseases

Transcriptomics involves studying the gene expression of the complete set of RNA transcripts which may include messenger RNA (mRNA) or other non-coding RNAs like micro-RNA (miRNA) in cells or a tissue under a specific physiological condition (Dong & Chen, 2013). Whilst mRNA serve as a transient intermediary molecule to proteins and the non-coding RNAs regulate gene expression, changes in their expression profiles may reflect changes in protein expression profiles (Lowe, Shirley, Bleackley, Dolan, & Shafee, 2017). With the help of next-generation high-throughput sequencing technologies and further computational analyses, transcriptome analysis may be utilized to improve our understanding of gene regulatory or protein interaction networks.

Bioinformatics analysis of expression profiling or sequencing analysis data: Gene expression in cells is highly dynamic and varies with cell type wherein the RNA complement

changes in response to signals or stresses in a qualitative and quantitative manner over time. The RNA complement (transcriptome) encodes proteins that in turn determine the phenotype of the cell and transcriptional analysis aids in the analysis of molecular changes that underlie cellular behavior and disease. After pre-processing and expression quantification of data obtained from microarray or RNA-seq analysis one can perform differential expression analysis, alternative splicing analysis, gene fusion discovery, study roles of small RNA, functional profiling etc. (Conesa et al., 2016). Differential expression analysis between the conditions being analyzed allows the identification of differentially expressed genes (DEG) which can then be analyzed with a number of techniques to identify expression patterns or essential candidate genes. Clustering gene expression data allows statistical grouping of samples to reduce data complexity and dimensionality for predicting functions or identifying shared regulatory mechanisms (Spies and Ciaudoa, 2015). Further, functional enrichment analysis (FEA) allows the identification of candidates sharing biological function or pathway by statistical over-representation analysis with the help of annotated databases such as Gene Ontology, Reactome or KEGG. Thus, analysis of genome-scale (omics) experiments may allow one to identify gene lists or genes associated with a particular process or disease. Further, pathway enrichment analysis may provide additional mechanistic insights by identifying biological pathways that are enriched more than that would be expected by chance among these genes (Reimand et al., 2019).

Biological protein-protein interaction network analysis

Cellular protein interaction networks may be generated from experimentally identified protein-protein interactions (PPI) with the help of high-throughput yeast two-hybrid screens (Y2H), mass spectroscopy, small scale interaction studies and proteomics data (Stelzl et al., 2005). Additionally, databases containing predictive and experimental PPI information like Biological General Repository for Interaction Datasets (BioGRID), Database of Interacting Proteins (DIP), Molecular Interactions Database (MINT), and Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) may be utilized for determining PPI networks (Kuzmanov and Emili, 2013). In general, interactome networks can be constructed by compiling already existing data available in the literature or by considering “orthogonal” information (sequence similarities, gene-order conservation, co-presence and co-absence of genes) apart from physical or biochemical interactions. Further, with the help of systematic, unbiased high-throughput experimental mapping strategies involving genomes or proteomes these networks can be further enriched (Vidal et al., 2011). Network analysis of interactome data has suggested that PPI network structure is related to whether a given protein is essential or not. In this respect, a scale-free structure of biological networks implies robustness with respect to random component failure. Biological interactome network models can be studied to determine global relationships between human disorders, associated genes and interactome networks and predict disease-associated genes. Network clustering algorithms can be utilized to determine topological modules or regions within a global network diagram that are likely to carry specific cellular functions. Moreover, aggregation of nodes of related function in the same network neighborhood can be utilized for identification of disease modules whose disruption may result in a particular disease phenotype in humans. Thus, utilizing interaction data it is possible to

derive information regarding biological systems from network structure (Hakes et al., 2008).

Graph theory based measures in network analysis: Studies on the topology of networks of different biological origin have demonstrated that in general they exhibit certain characteristics like scale-free architecture and a hierarchical arrangement of modules (Wuchty, Ravasz, & Barabási, 2006). The signature of scale-free networks is that the node degree follows a power-law distribution. Further, scale free networks contain few well-connected or high degree nodes and abundant small degree nodes. Such nodes that have much higher degree than the average degree in the network are referred to as hubs (Albert, 2005). Additionally, scale free networks also contain proteins that have many 'shortest paths' going through them. Such proteins that have a high betweenness centrality (H. Yu, Kim, Sprecher, Trifonov, & Gerstein, 2007) are referred to as bottleneck proteins. Further, centrality is another property that may be used to study scale free networks and it captures a networks compactness or capability of relaying information (Pavlopoulos et al., 2011). Therefore, network analysis utilizing graph theory based metrics that determine hubs, bottlenecks or central proteins may be capable of identifying essential proteins in a network (Jeong, Mason, Barabási, & Oltvai, 2001; H. Yu et al., 2007).

Regulatory networks in gene expression modulation

Gene expression in part determines organismal phenotype and is controlled by intricate, integrated regulatory circuits. Multiple molecular levels of regulatory circuits such as transcriptional regulation via transcription factor (TF) binding; noncoding RNA interaction with regulatory DNA; chromatin remodeler's interaction with regulatory DNA; post-transcriptional control of RNA (RNA transport, processing, modification, sequestration, and degradation); signal transduction networks involved in control of protein translation, stability, and activity; post-translational protein modifications (phosphorylation, acetylation); higher order chromosome organization are involved in gene expression modulation (Thompson et al., 2015). A transcriptional regulatory network specifies the manner in which *trans* regulators (e.g., proteins and noncoding RNAs) influence the expression of their target genes in different contexts. This information may be represented in a graph in which nodes represent regulators (e.g., TFs, signalling proteins, and chromatin remodelers) and their target genes while edges represent physical interactions between a regulator and a target (Thompson et al., 2015). Experimental strategies such as RNA profiling (RNA-seq), measurement of TF-DNA interactions (ChIPseq, ChIP-chip assays) help in the identification of components and interactions in a regulatory network, its molecular structure, function and causality.

Transcription factor regulatory networks: Gene regulatory networks may have a role in different cellular processes like cell differentiation, metabolism, cell cycle and signal transduction. A gene regulatory network (GRN) may be inferred from gene expression data such as mRNA abundance and transcription factor binding information (Emmert-Streib et al., 2014). Further, transcription factor binding sites may generally be predicted by scanning a position weight matrix (PWM) against DNA using a pattern matching algorithm (Bulyk 2004). Therefore, GRNs prepared in this manner may

represent causal biochemical interactions where the predicted links may correspond to physical binding events between molecules. Further, by studying the dynamics of these networks in deregulated cellular processes with the help of network models and determining their functionality analysis may help in understanding disease mechanisms.

RNA regulatory networks in gene regulation: Non-coding regulatory RNAs use diverse mechanisms including RNA-RNA, RNA-ligand, RNA-protein, RNA-DNA, and RNA-substrate interactions to carry out housekeeping functions or important regulatory roles in both eukaryotic and prokaryotic cells (Vandevenne et al., 2019). Non-protein coding DNA increases with organism complexity to reach about 98% in human beings and these non coding RNA participate in complex regulatory networks to achieve spatio-temporal control of gene expression in organismal homeostasis. Further, given the genome sizes and number of coding genes (proteins) in complex eukaryotes global regulation of all genome components is largely achieved by non-coding RNA which are capable of forming a much higher number of interactions than proteins (Vandevenne et al., 2019). Thus, different classes of regulatory RNAs (e.g., miRNAs, siRNAs, long-non-coding RNAs) are likely to be key players in the cell biology and dysregulated RNA or their regulatory networks may be associated with numerous diseases ranging from cancers to neurological disorders (Barta et al., 2017, Vandevenne et al., 2019).

MicroRNA (miRNA) mediated gene regulation: miRNAs (22 nucleotide long endogenous non-coding RNAs) influence fundamental biological processes such as cell death, cell proliferation, differentiation and apoptosis by interacting with their target mRNA preferably at 3′ UTR or at coding regions and 5′ UTR to fine tune target gene expression (Pu et al., 2019). In addition to the canonical mode of miRNA-mediated gene regulation wherein post-transcriptional repression occurs through miRNA seed region binding to 3′UTR of target mRNA other regulatory modes which lead to gene silencing or activation have been identified (Vasudevan et al., 2007). MiRNAs may affect nearly one third of human genes and a single mRNA may be influenced by various miRNAs while a single miRNA can potentially regulate very large sets of genes which results in complex miRNA-mRNA regulatory networks (Friedman et al., 2009). Utilizing miRNA target gene information from different databases such as TarBase or miRTarBase (Vlachos et al., 2014, Chou et al., 2015) or from experiments such as high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP, CLIP-seq, PAR-CLIP), miRNA-mRNA regulatory networks may be determined (Thomson et al., 2011). Correspondingly, co-expression network analysis of miRNA and mRNA may then be utilized for integrative network analysis of miRNA-mRNA regulatory networks which can then be followed with pathway analysis of differentially expressed target mRNA to study disease associated molecular mechanisms.

MiRNA synthesis regulation during biogenesis: In general, miRNAs are intragenic and mostly processed from introns while some are processed from exons of protein coding genes or are intergenic which are transcribed independently of a host gene and regulated by their own promoters. The biogenesis of miRNA can occur via canonical and non-canonical pathways where polymerase II/III transcripts are processed post- or co-

transcriptionally (O'Brien et al., 2018). In the canonical biogenesis pathway which is the dominant pathway by which miRNAs are processed, pri-miRNAs are transcribed from their genes and processed into pre-miRNAs by the microprocessor complex (RNA binding protein, DiGeorge Syndrome Critical Region 8 (DGCR8) and ribonuclease III enzyme, Drosha). The non-canonical miRNA biogenesis may occur via Drosha/DGCR8-independent and Dicer-independent pathways. During miRNA generation by Drosha/DGCR8-independent pathway mirtrons generated from the introns of mRNA during splicing or 7-methylguanosine (m^7G)-capped pre-miRNA are directly exported to the cytoplasm through exportin 1 without cleavage by Drosha and processed with the help of AGO2. Dicer-independent miRNAs are processed from endogenous short hairpin RNA (shRNA) transcripts by Drosha to generate pre-miRNAs which are later processed by AGO2 within the cytoplasm (O'Brien et al., 2018).

RNA binding proteins in gene expression regulation: RNA-binding proteins (RBPs) act in concert with transcription factors, epigenetic regulators, and signal transduction networks to form a global regulatory network. RBPs are involved in the post-transcriptional regulation of RNA editing, location, stability, and translation and in turn influence gene expression. Proteins such as polypyrimide track protein 1 (PTBP1), embryonic lethal abnormal vision like protein 1 (ELAVL1) or zinc finger protein 36 family (ZFP36, ZFP36L1 and ZFP36L2) are involved in modulating networks of RNA molecules (RNA regulons) to trigger a response. RBP may influence RNA processing, mRNA translation, RNA stability and decay, non-sense mediated decay, mRNA sub-cellular location and miRNA-biogenesis. RBP editors (e.g., RNA methyltransferases and deaminases) edit the sequence content of the transcriptome while RBP readers (e.g., RNA granule components, eukaryotic translation factors) may influence sub-cellular location and translation. Finally, RNA decay may be induced by RBP erasers (e.g., destabilizing factors and nucleases). Thus, global mechanisms of RBP-mediated control may be studied with the help of an integrative analysis of the protein:RNA interactome, transcriptome, and translatome in normal or diseased cell types (Díaz-Muñoz and Turner 2018).

Bioinformatics approaches for sequence based or structural analyses of proteins

A number of computational methods or principles can be utilized to identify protein function, identify and characterize domains, predict secondary or tertiary protein structure or identify and characterize interactions.

Sequence similarity assessment to determine protein orthologs: Database searches performing sequence-sequence comparisons (BLASTp), sequence-profile comparisions (HMMscan) or profile-profile comparisions (HHblits) can be utilised for identifying homologs (Altschul et al., 1990; Eddy 1998; Söding 2005). Sequence databases such as non redundant (NR) protein database, NCBI Reference Sequence (RefSeq) database or Universal Protein (UniProt) resource database can be used as reference databases to determine close-homologs that have statistically significant similarities with the query

sequence (The UniProt Consortium, 2014; Pruitt et al., 2007). A number of widely utilized sequence similarity searching programs such as Basic Local Alignment Tool (BLAST), Position Specific Iterated BLAST (PSI-BLAST), Domain enhanced lookup time accelerated BLAST (Delta-BLAST), HMMER package or HMM-HMM–based lightningfast iterative sequence search (HHBlits) can be utilized for this purpose (Boratyn et al., 2012, Altschul et al., 1997, Remmert et al., 2011, Altschul et al., 1990, Eddy 1998; Söding 2005). In this respect, profile-profile and sequence-profile based searches are more sensitive than searches utilizing sequence-sequence comparisons for identifying orthology and in particular remote orthologs. This is because profiles which are derived from a multiple sequence alignment (MSA) based on probability theory include more information than one sequence alone about the degree of conservation of domains in a related family of proteins (Pearson and Lipman, 1988).

Phylogenetic analysis to determine evolutionary relationship between sequences: Phylogenetic analysis utilising protein sequence data is useful for studying protein evolution and functions, establishing orthology and paralogy relationships among proteins, detecting horizontal gene transfers (HGT), predicting functional interactions among coevolving genes, genome annotation, gene function prediction, identification and construction of gene families and gene discovery (Gabaldón, 2007; Zhang et al., 2019). A phylogenetic tree is determined from a multiple sequence alignment considering homologous protein sequences of the protein under consideration. An optimization principle such as the maximum likelihood, maximum parsimony, minimum evolution principle or distance based methods etc. are generally utilised to estimate the topology (branching pattern of a tree) or branch lengths for a given tree topology (Nei, 1996).

Comparative modeling of protein structures: Comparative modelling is utilised to predict reliable and accurate protein structure models considering the fact that evolutionarily related sequences may have similar three-dimensional (3D) structures. In particular, a 3D model of a protein of interest (target) can be determined from related protein(s) of known structure [template(s)] with reliable or statistically significant sequence similarity. Several consecutive steps like determining template protein(s) related to the target, aligning target and template(s) sequences, identifying structurally conserved regions and structurally variable regions, including insertions and missing N and C termini, modelling side chains and refining and evaluating the resulting model are performed iteratively (Ginalski, 2006). HHpred may be utilized for remote protein homology detection and structure prediction by pair-wise comparison of profile hidden Markov models (HMMs). An HMM profile contains position specific information regarding the degree of conservation of each column in the multiple alignment where an HMM column describes the probabilities of each of the 20 amino acids at that position and transition or emission probabilities which describe how often amino acids are inserted and deleted at that position. During an HHpred analysis, the query sequence or a multiple alignment can be compared across sequence profile databases such as protein data bank (PDB-HMM), protein families database (Pfam), conserve domain database (CDD) etc. Herein, based on local or global alignment and secondary structure similarity score between query-template a number of possible

query-template alignments, multiple alignments of the query with a set of templates or 3D structural models from MODELLER software considering these alignments are provided (Söding et al., 2005). Finally, utilizing a suitable query-template alignment and template PDB structure 3D models for the query may be generated in MODELLER (Eswar et al., 2006) and with the help of different programs available for model evaluation the most likely 3D structure of the query sequence can then be predicted.

Molecular docking programs: Docking allows one to study the most likely interaction pose between two or more molecular structures such as proteins in protein-protein docking or between ligand and protein in protein-ligand docking (Morris and Lim-Wilby, 2008). Molecular docking is usually performed with the help of a search method and a scoring function wherein a search space defined by the molecular representation utilized in the docking method is explored computationally. Protein-protein docking can be performed with the help of a number of methods such as PatchDock, FireDock, HADDOCK, ClusPro etc. PatchDock (Duhovny et al., 2002; Schneidman-Duhovny et al., 2005) performs geometric shape complementarities matching by utilizing geometric hashing and pose-clustering techniques whereas FireDock (Andrusier et al., 2007) performs optimization of side-chain conformations and rigid-body orientation followed by high-throughput refinement to determine docked conformations. HADDOCK (Vries et al., 2007; Dominguez et al., 2003; Van Zundert et al., 2016) is a rigid-body docking process involving solvent based refinement of complexes in a data driven (ambiguous interaction restraints) manner where residues which are likely to interact during complex formation are comsidered as active and passive residues. Further, ClusPro (Comeau et al., 2004a; Comeau et al., 2004b; Kozakov et al., 2006) utilizes a Fast-Fourier Transform (FFT)-based approach to determine native binding conformation and during the docking process correlation between binding site free energy attractor and cluster size are also taken into consideration. Subsequently, the docked solutions are analysed with the help of a scoring function that could be empirical, knowledge based, or molecular mechanics-based to rank the different docked conformations obtained. For instance, FireDock utilizes a glob score that represents the binding energy score of the complexes calculated considering stacking and aliphatic interactions, desolvation energy (atomic contact energy, ACE), partial electrostatics, hydrogen and disulfide bonds, van der Waals interactions, rotamer's probabilities etc. (Duhovny, Nussinov, & Wolfson, 2002; Schneidman-Duhovny, Inbar, Nussinov, & Wolfson, 2005; Andrusier, Nussinov, & Wolfson, 2007). Additionally, HADDOCK utilizes a score which is a weighted sum of intermolecular van der Waals, electrostatic, desolvation, ambiguous interaction restraints (AIR) energies, and a buried surface area (BSA) term (Vries et al., 2007; Dominguez, Boelens, & Bonvin, 2003; Zundert et al., 2016). Further, ClusPro utilizes a balanced score that is representative of shape complementarity (repulsive and attractive interactions), electrostatic and desolvation contributions [in terms of ACP (atomic contact potential) and pair-wise potentials such as DARS (Decoys as the Reference State)] (Kozakov, Brenke, Comeau, & Vajda, 2006; Comeau, Gatchell, Vajda, & Camacho, 2004b, 2004a). Thus, steric and physicochemical complementarities at the protein-protein interface may be utilized to determine the likely interaction conformation between proteins important for a variety of physiological processes and further characterization of the interaction interface can be useful in modulating these

interactions in disease associated processes.

Virulent proteins in host-pathogen interactions

Prediction of host-pathogen interactions computationally may involve utilising structural information of proteins and previous known interactions with other proteins to search for homologous proteins in the pathogen or host that may interact with known proteins (interologues). Further, characterisation of the novel host-pathogen interactions determined in this manner may aid in the development of novel drugs, vaccines or other therapeutics (Southwood and Ranganathan, 2019).

Host-parasite interactions in Leishmaniasis: Leishmaniasis is caused by an intra-cellular protozoan parasites belonging to *Leishmania* spp. (family Trypanosomatidae, domain Kinetoplastida). Clinical presentation of the disease depends on the host's immune response and the species infecting the host and ranges from cutaneous (*L. mexicana*, *L. major*, *L. braziliensis, L. amazonensis*), muco-cutaneous (*L. panamensis* and *L. braziliensis*) to visceral leishmaniasis (*L. infantum* and *L. donovani*) (Kaye & Scott, 2011). The parasites have two morphologically distinct variants during their life-cycle, an amastigote form (host: mammalian cells) and a promastigote form (host: phlebotomine sand flies).

A complex array of processes are involved during disease establishment wherein the parasite attaches and invades host cells such as neutrophils, macrophages, monocytes or dendritic cells with the help of a number of different virulent proteins. *Leishmania* spp. may utilize several virulence factors (polypeptide/polysaccharide ligands) to interact with the host cell receptors and these interactions are likely to be alternatively or synergistically involved in receptor-mediated endocytosis of the pathogen. In particular, the leucine-rich repeat (LRR)-containing proteins such as parasite surface antigen 2 and proteophosphoglycan participate in parasite attachment and invasion of host cells. However, the proteins lipophosphoglycan and leishmanolysin in addition to attachment and invasion of host cells aid in intracellular survival of the parasites (Kedzierski et al., 2004; Joshi, Kelly, Kamhawi, Sacks, & McMaster, 2002). Further, A2 protein is important for pathogen survival in visceral organs (Zhang & Matlashewski, 2001) while amastin, amastin-like surface protein and cysteine proteases include other virulence factor proteins that attribute virulence to these parasites (Rochette et al., 2005; Silva-Almeida, Pereira, Ribeiro-Guimarães, & Alves, 2012). Thus, utilizing these proteins during host-pathogen interaction *Leishmania* spp. either as avirulent or metacyclic promastigotes can interact with different receptors (mannose receptor, fibronectin receptors, complement receptors, and dendritic cell-specific intercellular adhesion molecule-3-grabbing non-integrin) on primary macrophages or dendritic cells for invasion. Additionally, they may interact with Fc gamma receptors in primary macrophages and dendritic cells for phagocytosis in the amastigote stage (Ueno & Wilson, 2012). Following the interaction with host cell receptors the pathogen, *Leishmania* spp. becomes internalized and utilizes different mechanisms to effectively suppress and evade host immune responses in order to establish long term infections within the host cells. Different mechanisms of evasion that *Leishmania* spp. might utilize include strategies to

survive within phagosomes, interference in antigen presentation, modification of the complement system and phagocytosis, modification of T cell responses and modulation of cytokine and chemokine levels. Further, the protozoan parasite might alter host cell signaling pathways like JAK/STAT pathway activation via IFN-γ, toll-like receptor pathways, protein kinase C mediated signaling, MAPK signaling pathway etc. to alter host immune responses for establishing infection (Gupta, Oghumu, & Satoskar, 2013).

Co-evolution in proteins or protein-protein interactions

Co-evolution typically refers to compensatory or coordinated changes that occur in biomolecules to maintain or refine functional interactions (Juan et al., 2013). Such compensatory substitutions may restore the functionality of protein by sustaining the fitness or functionality of a protein under constraints imposed by physico-chemical interaction forces and structural or folding associated factors (Chakrabarti and Panchenko 2010). Moreover, proteins may interact fleetingly in transient complexes for different types of interactions such as signaling–effector; enzyme–inhibitor, enzyme–substrate, hormone–receptor etc. or some proteins could act as parts of multi-subunit enzymes in permanent obligate interactions (Mintseris and Weng, 2005). These protein interactions may contribute to sequence evolution between proteins that physically interact or have a functional association. Additionally, when two proteins interact with each other, a mutation in one protein may result in a compensatory mutation in the other protein to maintain the stability or function of the interaction during the course of evolution due to evolutionary selection pressure (Lovell and Robertson, 2010). Applications of co-evolution analysis include protein contact prediction for structure prediction or modeling, determining ligand or protein interaction specificity and identifying physical or functional interactions in or among biomolecules (Juan et al., 2013).

Intra-molecular co-evolution: Co-evolution may be studied with the help of multiple sequence alignment (MSA) for a protein family of homologues to detect pairs of positions (columns of the MSA) that have interdependent amino acid frequencies or similar patterns of amino acid substitutions. Intra-molecular co-evolution has been widely studied with multiple approaches like substitution correlations (McLachlan-based substitution correlation [McBASC]), mutual information of amino acid frequencies or a global statistical model of the MSA such as in direct coupling analysis (DCA) or protein sparse inverse covariance (PSICOV). Briefly, mutual information (MI) which measures the dependence of one position in an alignment on another can be utilized to study co-evolution within a protein family. However, a few limitations such as requirement of large number of sequences (>125) in alignments, phylogenetic relationships among organisms represented in the alignment and random and nonrandom MI at high entropy positions in comparison to lower entropy positions may restrict its applicability (Dunn et al., 2008). Additional methods such as co-evolution analysis using protein sequences (CAPS) detect correlations in amino acid distance matrices by correcting for the influence of phylogenetic background (Juan et al., 2013). Intra-protein co-evolution studies have observed that interacting residues in close proximity have a tendency to co-evolve with only one or two other positions whereas

residues involved in functionally important interactions or conformational changes are generally involved in forming many co-evolutionary connections with other residues which tend to be more conserved in evolution (Chakrabarti and Panchenko 2009).

Inter-molecular co-evolution: Molecular co-evolution analysis in inter-protein complexes may determine co-evolving residue pairs among interface residues and it is likely that coordinated changes at these residue positions are likely to be crucial for a functional interaction between a set of interacting proteins. A few methods are presently available to study inter-protein co-evolution like MirrorTree, CAPS and EV-complex. MirrorTree quantifies similarities between the MSA-derived trees of orthologous set of sequences of the two protein families by extracting inter-orthologue distance matrices and calculating the linear correlation (Ochoa and Pazos, 2010). In contrast, CAPS identifies amino acid co-variation with the help of BLOSUM corrected amino acid distances and phylogenetic sequence relationships (Fares and McNally 2006). However, EV-complex predicts inter-evolutionary couplings using a global probability model of sequence co-evolution utilizing pseudo-likelihood maximization (PLM) based calculation of coupling parameters on the alignment of concatenated sequences (Hopf et al., 2014). Analysis of inter-molecular coevolution with the help of these methodologies has shown that residues in close spatial proximity at the interaction interface generally exhibit a higher tendency to co-evolve than other spatially separated predicted co-evolving residue pairs (Anishchenko et al., 2017; Mintseris and Weng, 2005).

Chapter 1

Deriving inferences regarding intra-cellular interactions utilizing systems analysis approaches

Deriving inferences regarding intra-cellular interactions utilizing systems analysis approaches

The intra-cellular environment of a cell contains different bio-molecular entities (proteins and nucleic acids) that interact in a number of ways to carry out different cellular processes that collectively result in the manifestation of a particular cellular function. In general, a cellular interactome may be comprised of different physical interaction networks among bio-molecules like protein-protein interaction network, transcriptional regulatory network, miRNA-mRNA regulatory network etc. Systematic assessment of genome wide expression changes under disease conditions utilizing high throughput technologies might allow one to study a large and complex network of combinatorial interactions among proteins, miRNA-mRNA or protein-mRNA within the cells constituent parts. Typically, intracellular biological networks contain different molecular components within a cell as nodes and their interactions as links or edges and these network components or the relationship between these components may be altered in disease conditions. Therefore, studying this repertoire of cellular interactions in terms of interaction networks might facilitate one in understanding the changes in network architecture or its components that may be associated with a cellular function under a disease condition. In this respect, I have studied complexities in intra-cellular interactions by modulating the expression level of a key regulator and studying the effect of this perturbation on the intra-cellular network components or its architecture with the help of systems based computational analyses approaches. For this purpose, to determine inferences regarding intra-cellular interactions, I have utilized different case scenarios as mentioned below:

In order to study the inter-relationship between the gene regulatory and protein-protein interaction networks, a scenario wherein over-expressing a transcriptional activator is necessary and sufficient for a cellular process has been considered. Now, given that an essential regulator is over-expressed, delineate the intra-cellular regulatory network governing the cellular process and identify essential proteins within the regulatory network by inferring their roles within the intra-cellular protein interactome for the cellular process.

Flexibility in miRNA-mRNA interaction networks may preserve cellular homeostasis by re-adjusting network components in response to gradual changes in intra-cellular concentration of an essential regulator (miRNA). Thus, given that the intra-cellular concentration of an essential regulator (miRNA) changes gradually over time, determine whether there are changes in the regulatory network components (mRNA) in relation to changes in cellular physiology.

MiRNA-mRNA interaction networks may be highly complex, however, disease associated alterations in intra-cellular concentrations of regulators (miRNA) may lead to adjustments

in the network organization or regulatory relationship to preserve cellular function. Therefore, given that multiple miRNA expression levels could be altered, determine whether the regulatory network components (mRNA) or the relationship between regulator-target (miRNA-mRNA) pairs may be affected in association with changes in cellular phenotype.

RNA binding proteins (RBP) may additionally regulate gene expression, consequently down-regulating post-transcriptional mRNAs regulating proteins may alter the protein-mRNA interaction network. However, not all target mRNA may be affected in a similar manner because of the versatile nature of regulatory relationships among interactions. In this context, given a stimulus that affects the RBP, determine the regulatory network of the essential mRNA regulator and study the relationship between the regulator-target (protein-mRNA) pairs under these different cellular conditions.

1.1 Results

1.1.1 Important effector proteins within regulatory networks act as topologically important signaling proteins in order to regulate a cellular process

Synopsis: The present investigation is an approach to determine inferences for regulatory network components based on a topological analysis of an intra-cellular protein-protein interaction network representing a cellular process. This is because such an analysis might enable one to identify important effector proteins in a regulatory network essential for a cellular process. Furthermore, alterations in an important effector protein which is likely an essential regulatory network component might result in changes in cellular physiology resulting in a disease associated with the dysregulation of the cellular process (ciliogenesis).

Problem Statement: Given that a cellular process can be experimentally replicated in the laboratory by over-expressing a key transcriptional activator (master regulator), delineate whether the regulatory network proteins could be topologically important in the cellular protein interactome? Alternately, identify key network components or essential proteins for the cellular process which when dysfunctional or deregulated could be associated with a disease process.

Hypothesis: Studying the regulatory network known to be essential for a cellular process in association with the cellular protein interaction map might elucidate topologically important proteins essential for the cellular process within the regulatory network. Additionally, graph theory based properties attributable to scale free protein-protein interaction networks may be utilized to identify essential network components within a transcriptional network.

System of study: A cellular process may be represented in terms of a protein interaction map comprised of proteins known to be involved in the process as nodes and their known associations represented as interaction edges. Such an interaction network may exhibit scale free network architecture and graph theory can be utilized to study the architectural features of this interaction network. This study is based on the assumption that changes in messenger RNA expression profiles upon over-expression of the key regulator (transcription factor) would be directly reflected in the expression profile of the proteins that they encode for. The representative interactome of the cellular process may be constructed by taking these regulatory network proteins and their protein interactors into consideration. Studying the topological properties of this interactome may yield insights into which proteins are crucial for the integrity of the network. Herein, a cellular process namely motile ciliogenesis has been taken into consideration to study the regulatory network of the motile cilia master regulator (FoxJ1) in association with the cellular protein-protein interactome (Figure 1.1).

Graphical Summary:

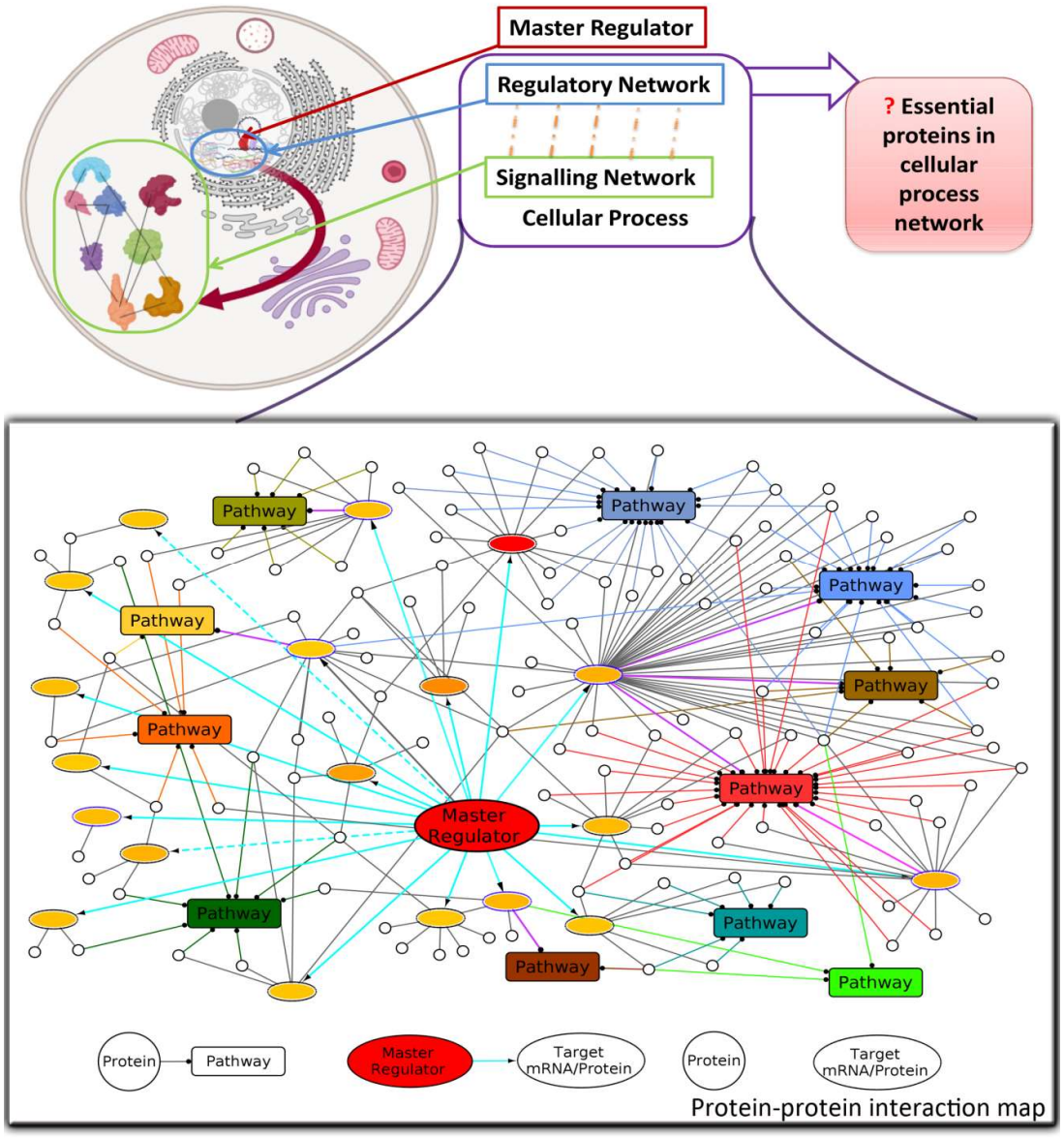


Figure 1.1: **Network analysis may be utilized to identify topologically important proteins in intra-cellular protein-protein interaction networks.** Topologically important proteins so identified are likely to be essential regulatory network proteins that act at the interface of regulatory and signaling networks to mediate a cellular process.

Analyzing intra-cellular protein-protein interaction networks to determine effector proteins that bridge the regulatory and signaling networks governing a cellular process

Macromolecular synthesis and assembly of ciliary structures during the ciliogenesis process is a complex co-coordinately regulated process by multiple proteins. Mutation(s) or defect(s) in one or more proteins involved in the structural organization of cilia, regulation of its assembly or functions of the ciliary organelle may result in a range of disorders. Particularly when motile cilia dysfunction it results in primary ciliary dyskinesia (PCD). Therefore, identifying effector proteins which form a crucial link between the regulatory and signaling components in the motile cilium with the help of this analysis might elucidate proteins essential for ciliary development or maintenance of ciliary function. Additionally, screening for defects in this repertoire of proteins might enable one to determine possible causal or etiological genes for PCD.

1.1.1.1 Delineation of the FOXJ1 regulatory network for motile cilia biogenesis in humans

Determining and studying the FOXJ1 regulatory network which mediates the development of motile cilia might help one to understand ciliogenesis and in turn ciliopathies associated with abnormal ciliary differentiation and function in humans. FoxJ1 protein acts as the master regulator of motile ciliogenesis, since over-expression of FoxJ1 in model systems such as zebra fish and *Xenopus* is necessary and sufficient for the development of motile cilia (Stubbs et al., 2008; X. Yu et al., 2008; Choksi, Babu, et al., 2014). Further, over-expression of FOXJ1 induces the expression of 572 genes which bring about the assembly of motile cilia and are likely to be involved in its functions. Thus, in order to determine the regulatory network that is essential for motile cilia development, genes that are likely to be transcriptionally regulated either directly or indirectly by the transcriptional activator, FOXJ1, have been predicted. FOXJ1 cis-regulatory sites were found within ±6kb of the upstream or downstream region of transcription start site (TSS) in 424 FoxJ1 induced genes (FIG) and such genes have been classified as FOXJ1 directly regulated genes (Supplementary Figure S1). However, the other 148 FIG that lack FOXJ1 cis-regulatory sites have been termed as FOXJ1 in-directly regulated genes. These indirectly regulated genes might have FOXJ1 binding sites in distant enhancer regions or could be indirectly regulated by FOXJ1 via other transcription factors activated by FOXJ1. Additionally, based on the predicted binding sites of the FOXJ1 protein it appears that FOXJ1 has a binding preference towards the consensus sequence NNN[GA]TAAACAAANNNN (Figure 1.2).

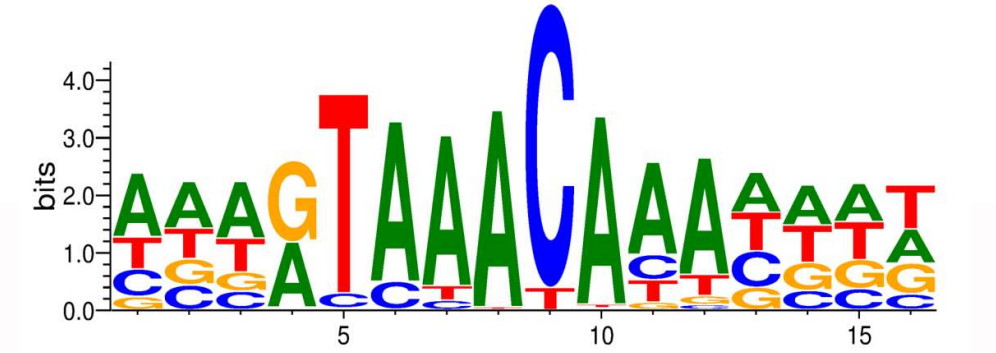


Figure 1.2: Predicted FOXJ1 binding specificity. Logo plot of predicted human FOXJ1 binding sites among FOXJ1 directly regulated genes.

1.1.1.2 Functional annotation and ciliary associations for FOXJ1 regulatory network genes

The most likely function that a set of proteins may have under a particular context may be ascertained with the help of functional enrichment analysis considering proteins associating with one another under a particular context. Known ciliary functional annotations for directly and indirectly regulated FOXJ1 genes were identified based on the data contained within the CCR and these functional associations were grouped into functional cohorts (Assigned Ciliary Role) manually. Based on the CCR dataset possible ciliary roles could be assigned to 102 (directly) and 35 (in-directly) regulated genes (Figure 1.3, 1.4). Briefly, the functional cohort classification elucidated that FOXJ1 primarily influences 'ciliary structural assembly or motility' by regulating proteins that 'act as structural constituents of cilia' (axoneme assembly, IFT complex, centrosome component, basal body associated), 'regulate the structural assembly of cilia' (cilia assembly, ciliogenesis) or 'have a role in ciliary function' (ciliary transport or motility) (Figure 1.3, 1.4). In particular, most of the directly (82) and in-directly (26) regulated FOXJ1 target genes had ciliary associations belonging to the 'ciliary structural assembly or motility' cohort (Figure 1.3, 1.4). The most likely function of a particular protein under a specific cellular context maybe inferred with the help of functional enrichment analysis. Herein, probable collective activities of proteins associating with one another under a particular context is utilized to get an idea about the most likely function that these proteins may be involved in collectively. GO mapping determined groups of genes having similar functions (co-associated genes) in multiple GO annotation categories, such co-associated genes in multiple annotation clusters were categorized into 'GO based Ciliary Association(s)' (Table S1). Further, with the help of GO enrichment analysis predicted ciliary associations could be assigned to 17 (directly) and 6 (indirectly) regulated FOXJ1 network genes (Table 1.1). Moreover, GO analysis was utilized to ascertain additional ciliary associations for FOXJ1 regulatory network genes. In this regard, GO mapping elucidated that 20 proteins were associated with DNA binding or the transcription factor ontology class and hence these 20 FIG encoded proteins (FIGp) are

likely to be transcription factors.

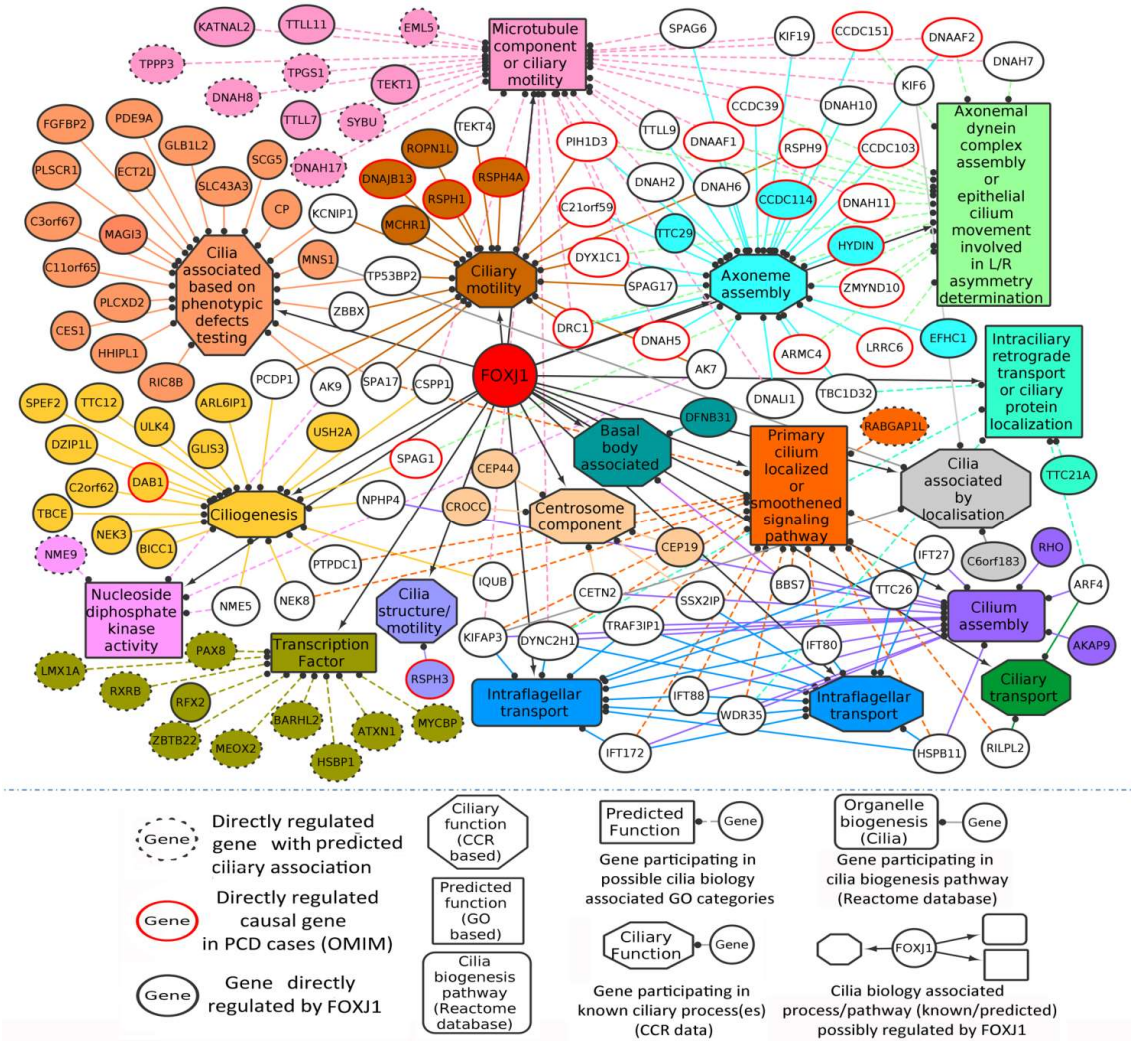


Figure 1.3: **FOXJ1 directly regulated genes within the FOXJ1 regulatory network and their possible involvement in ciliary processes.** FOXJ1 is involved in regulating ciliary structure assembly and motility by influencing the expression of directly regulated genes as depicted here. Functional associations of the FOXJ1 directly regulated genes either known (derived from CCR) or probable (predicted based on GO analysis) are shown here. The genes and edges are color coded to indicate the processes the genes regulate or are involved in (genes associated with more than one process are not colored). [Note: published in Mukherjee et al., 2019].

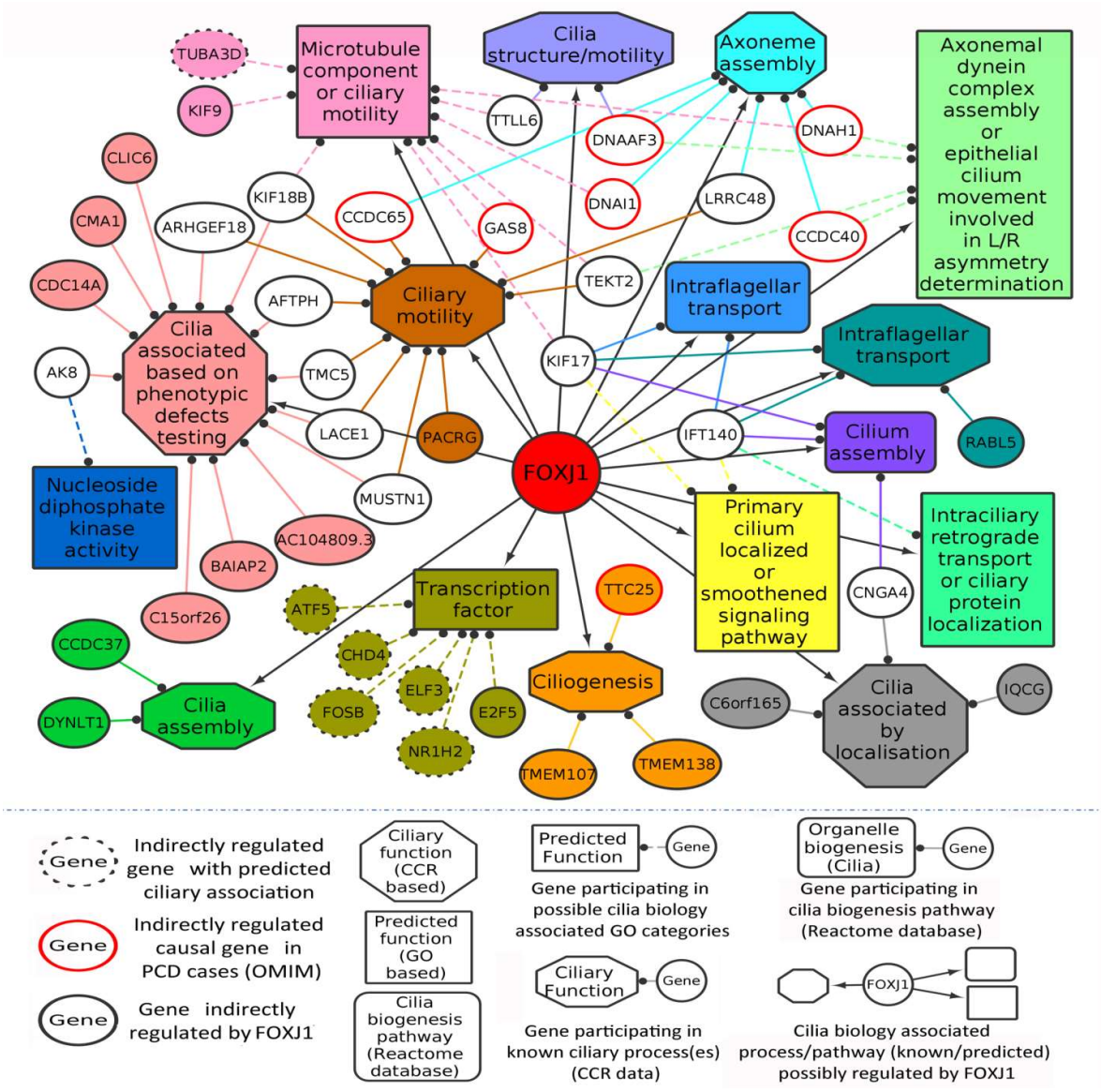


Figure 1.4: **FOXJ1 indirectly regulated genes within the FOXJ1 regulatory network and their possible involvement in ciliary processes.** The indirectly regulated FOXJ1 genes regulate ciliary assembly or motility. The ciliary roles of FOXJ1 indirectly regulated genes either known (based on CCR) or probable (predicted from GO analysis) are depicted here. The genes and edges are color coded to denote the processes they regulate or are involved in (genes associated with more than one process are not colored). [Note: published in Mukherjee et al., 2019].

Table 1.1: **Novel predicted functions of FOXJ1 regulated genes.** Assigned ciliary roles of FOXJ1 regulatory network genes based on gene ontology analysis are listed here.

FOXJ1 Target Gene	**[1]Regulatory Effect**	**[2]Assigned Ciliary Role**
RABGAP1L	Direct	Cilia associated by localization
DNAH8	Direct	Ciliary structure/ motility
TPPP3	Direct	Ciliary structure/motility
NME9	Direct	Ciliary structure/motility
DNAH17	Direct	Ciliary structure/motility
TPGS1	Direct	Ciliary structure/motility
EML5	Direct	Ciliary structure/motility
SYBU	Direct	Ciliary structure/motility
HSBP1	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
MYCBP	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
ATXN1	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
BARHL2	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
LMX1A	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
MEOX2	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
PAX8	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
RXRB	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
ZBTB22	Direct	Regulates genes involved in ciliary assembly/motility (Transcription factor)
TUBA3D	Indirect	Ciliary structure/motility
NR1H2	Indirect	Regulates genes involved in ciliary assembly/motility (Transcription factor)

FOSB	Indirect	Regulates genes involved in ciliary assembly/motility (Transcription factor)
ATF5	Indirect	Regulates genes involved in ciliary assembly/motility (Transcription factor)
CHD4	Indirect	Regulates genes involved in ciliary assembly/motility (Transcription factor)
ELF3	Indirect	Regulates genes involved in ciliary assembly/motility (Transcription factor)

[1]**Regulatory Effect:** Direct/Indirect depending on whether the gene is directly or indirectly regulated by FOXJ1.
[2]**Assigned Ciliary Role:** Predicted ciliary role based on gene ontology (GO) analysis.

Thus, construction and analysis of the FOXJ1 regulatory network for motile cilia biogenesis in humans suggested that majority of the regulatory network genes that have been identified or extensively characterized to date are involved in ciliary structural assembly or motility. Further, it was observed that while 84.67% of directly and 81.76% of indirectly regulated genes are expressed in multiple motile ciliated tissues and some are differentially expressed in PCD (Figure 1.5). However, only about 24% of the directly and indirectly regulated FOXJ1 network genes shared an overlap with the CCR dataset. In this context, it may be possible to obtain an idea about the probable functions of additional FIGp based on the principle of 'guilt by association', since it is likely that FIGp participate in motile cilia assembly or function in a coordinated manner in association with other proteins (signaling) of the ciliary milieu.

In the corresponding section, the regulatory network proteins have been studied in a broader context wherein the protein–protein interaction associations that the regulatory network proteins are likely to share with the other proteins in the intra-cellular compartment are taken into consideration. This exercise might allow one to identify the key connector proteins (regulatory network proteins) that relay the information onto the signaling component within the cell. It is ideal to study the signaling component in association with the FOXJ1 regulatory network because previous studies have identified signaling pathways like Notch, Fgf and Wnt to be involved in motile cilia biology (Neugebauer, Amack, Peterson, Bisgrove, & Yost, 2009; Lopes et al., 2010; Caron, Xu, & Lin, 2012). Moreover, alterations in proteins involved in the ciliary protein-protein interactome or disruptions in network interactions might alter the motile cilia interactome in turn leading to ciliopathies like PCD. Therefore, studying the probable signaling network(s) acting concurrently or in response to FOXJ1 activation might aid in the identification of essential proteins for cilia biology within the network.

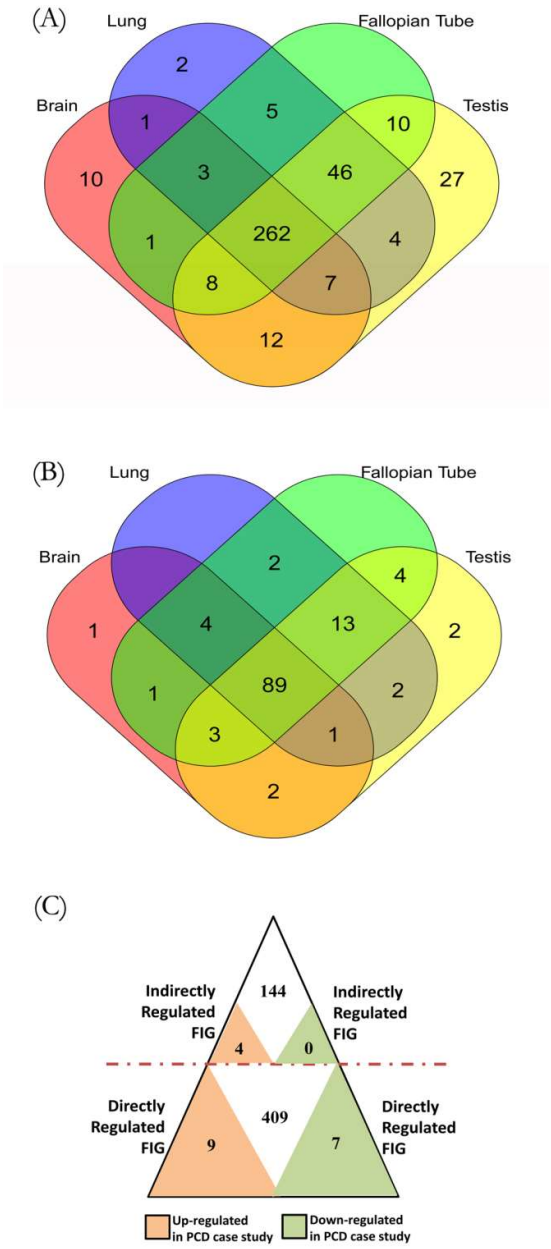


Figure 1.5: **Associations of FOXJ1 regulatory network genes with ciliary biology based on expression analysis.** (A) The Venn diagram depicts the expression pattern of FOXJ1 directly regulated genes among multiple motile ciliated tissues based on the Human Protein Atlas data (Uhlén et al., 2015; Thul et al., 2017; Uhlen et al., 2010) (B) The Venn diagram depicts the prevalence of FOXJ1 indirectly regulated genes in different motile ciliated tissues based on the Human Protein Atlas data (Uhlén et al., 2015; Thul et al., 2017; Uhlen et al., 2010) (C) The number of FOXJ1 regulatory network genes that could be associated with PCD based on differential expression analysis of bronchial biopsies from PCD patients is outlined.

1.1.1.3 Essential proteins in representative motile cilia protein-protein interactome

Topologically important proteins such as hub, bottleneck or central proteins could be essential for a cellular process (Barabási & Oltvai, 2004; H. Yu et al., 2007; Pavlopoulos et al., 2011; Jeong et al., 2001) and as such identifying important interacting proteins (IIP) by performing a computational analysis of the PPIN associated with FIGp might help one in identifying essential proteins involved in motile cilia assembly or function. Thus, in order to obtain inferences regarding the PPIN associated with FIGp, the relevant network was determined taking high confidence physical interaction data from different protein-protein interaction databases including cilia specific datasets into consideration. The re-constructed PPIN hereafter referred to as FIG-sub-network was comprised of 6493 primary interactors of FIGp (434) and their first level interactors participating in 40,608 interactions (Figure 1.6A). Networks that conform to the power law in general have p-value>0.1 (Clauset, Shalizi, & Newman, 2009) and therefore FIG-sub-network was considered as a scale free network since the p-value of the estimated fit of the network's degree distribution to the power law was found to be 0.71 (Figure 1.6B).

1.1.1.4 Network analysis of primary PPIN of FIGp (representative motile cilia interactome)

In order to determine essential proteins for cilia assembly or function, topologically important proteins for network integrity or function in a scale free PPIN were identified with the help of different graph theory based measures. The in-silico node deletion analysis identified 85 local network perturbing and 13 global network perturbing proteins which when removed introduced significant changes in network topology in terms of centrality (Figure 1.6C). The other topologically important proteins identified included 243 hub, 86 bottleneck and 166 central proteins. Considering the overlap among these topologically important proteins, 122 proteins identified in two or more metrics were considered as IIP (Figure 1.6C, Table S2).

Cilia biogenesis or function associated genes may be identified with the help of gene mis-expression, targeted gene knock-out or knock-down studies in experimental model systems (Stubbs et al., 2008; X. Yu et al., 2008; Choksi, Babu, et al., 2014; May-Simera et al., 2016; Terré et al., 2016). Additionally, large scale gene expression analysis studies or sequencing analysis of patient samples have also identified PCD associated genes (Zariwala et al., 2011; Horani et al., 2016). Such expression based evidences from multiple studies might provide an initial experimental support in regarding the identified IIP as essential proteins in motile cilia biology or function. In this regard, the 'cilia associated expression analysis' (Figure 1.6D) provided ciliary expression support to 121 IIP.

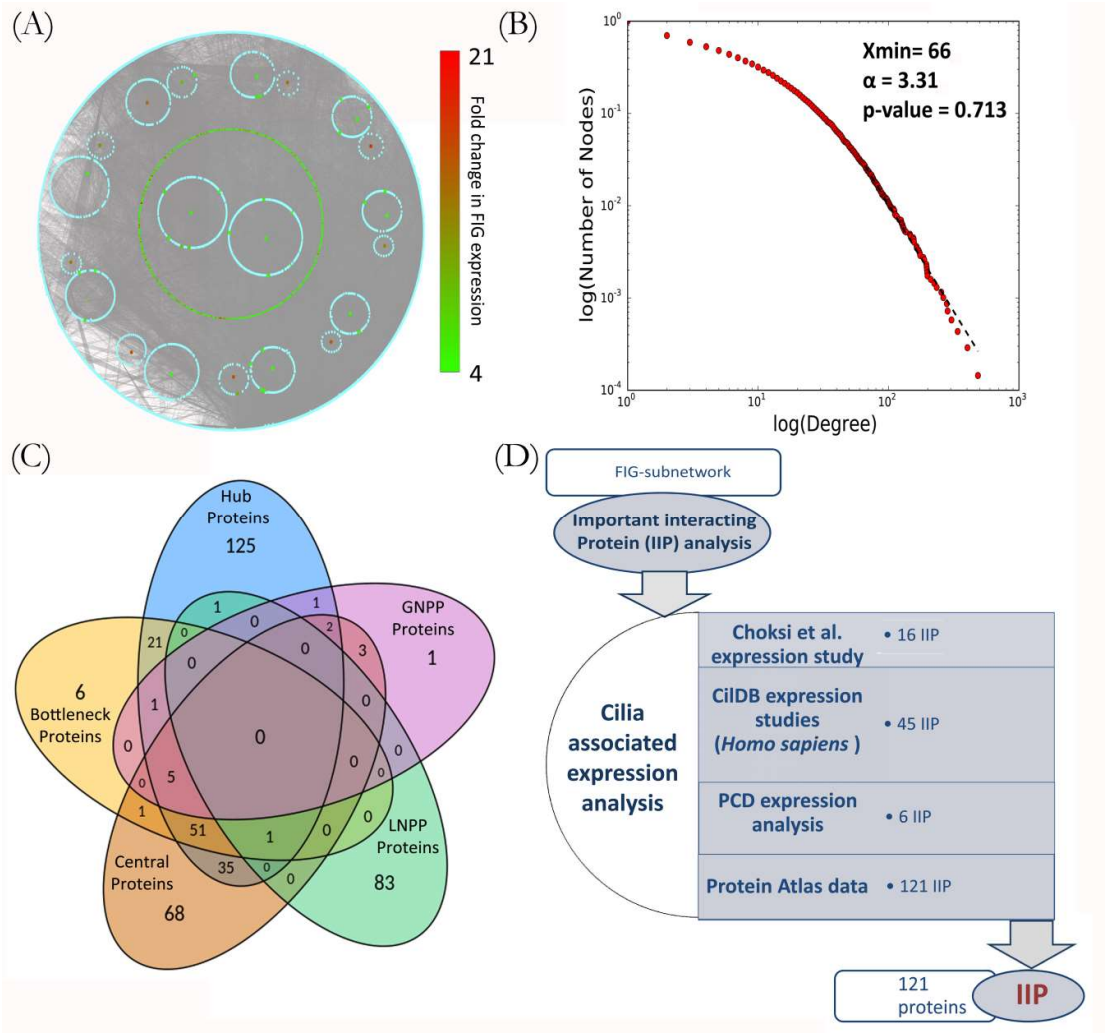


Figure 1.6: **Network analysis of FIG-sub-network.** (A) FIG-sub-network which was the largest connected component of primary interaction network of FIGp considered for the IIP analysis is shown here. (B) Degree distribution plot of the FIG-sub-network and the p-value for goodness of fit of the networks degree distribution to the power law as determined by the Kolmogorov-Smirnov test is represented here. (C) The number of proteins identified as hub, bottleneck, central, local network perturbing (LNPP) and global network perturbing proteins (GNPP) during the IIP analysis are indicated in the Venn diagram. (D) 'Cilia associated expression analyses for IIP is shown here. Abbreviations: FIG- FoxJ1 induced genes, IIP- Important interacting proteins. [Note: published in Mukherjee et al., 2019].

Moreover, considering the tissue specific distribution pattern of these 121 IIP in multiple motile ciliated tissues, it was found that 114 among them were expressed in all ciliated tissues considered here (Figure 1.7A, Table S2). Further, 6 IIP were found to be differentially expressed in the PCD case study (Figure 1.7B, Table S2) and based on the randomization analysis performed this observation is significant at 10% level of significance. In addition, the observation that 33 IIP had established roles in ciliogenesis and/or cilia function according to the CCR (significant at 1% level of significance according to the randomization analysis) further suggests that the identified IIP could indeed have essential roles in motile cilia or PCD pathogenesis.

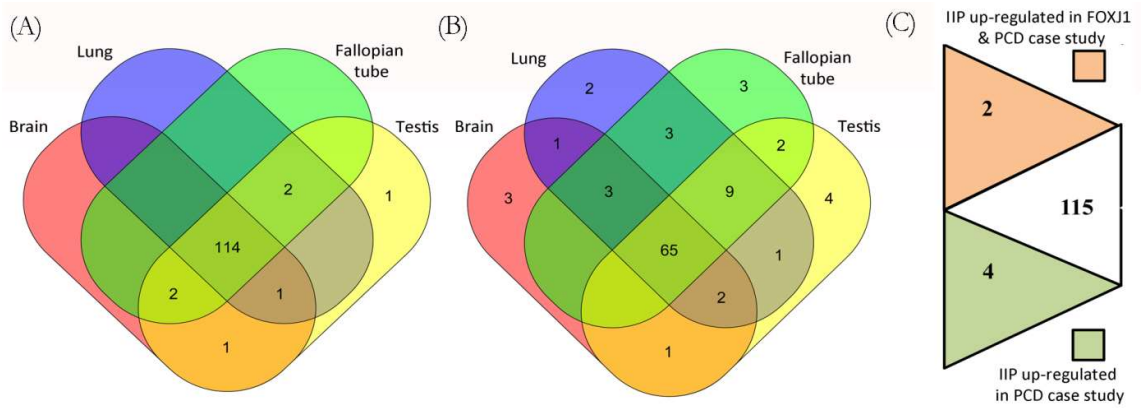


Figure 1.7: **'Cilia associated expression analysis' taking IIP into consideration.** (A) The Venn diagram depicts the mRNA expression pattern of IIP among multiple motile ciliated tissues based on expression data in the Human Protein Atlas (Uhlén et al., 2015; Thul et al., 2017; Uhlen et al., 2010). (B) Protein expression pattern of IIP among multiple motile ciliated tissues according to the Human Protein Atlas (Uhlén et al., 2015; Thul et al., 2017; Uhlen et al., 2010) is depicted in the Venn diagram. (C) The number of IIP that may be associated with PCD based on differential expression analysis in PCD tissue samples. Abbreviations: PCD: primary ciliary dyskinesia, IIP: important interacting proteins.

1.1.1.5 IIP and their probable roles in motile cilia interactome

Network analysis of the primary interaction network of FIGp (representative motile cilia interactome) identified 121 IIP among thousands of interacting proteins in the motile cilia interactome. The IIP (120) along with their primary interactors (246 FIGp, 1666 motile cilia expressed proteins) form an inter-connected module comprising of 2060 interactions among the protein nodes (Figure 1.8A). Such IIP having extensive connections with FIGp are likely to be involved in the coordinated assembly of functional motile cilia in association with FIGp. These interacting proteins are likely to participate in similar cellular pathways to regulate ciliogenesis and ciliary functions. Thus, utilizing the concept of 'guilt by association' (Oliver, 2000; Schwikowski, Uetz, & Fields, 2000), it is likely that these IIP are involved in signal transduction, developmental biology, cell cycle, generic transcription pathway, immune system etc. since these cellular

pathways were identified to be significantly enriched (Figure 1.8B, Table S3) in the Reactome pathway enrichment analysis (G. Yu & He, 2016). This data suggests that in addition to acting as structural constituents of the ciliary organelle, FIGp, in association with IIP, are likely to participate in multiple signaling pathways, cell cycle and developmental biology related activities in addition to generic transcriptional regulation of genes during motile cilia development.

1.1.1.6 Important effector proteins in FOXJ1 regulatory network

Extensive analysis of the FOXJ1 regulatory network associated PPIN (probable motile cilia interactome) identified IIP which are probably essential for the development of motile cilia and their involvement in physiological and developmental processes. Additionally, the IIP along with FIGp (246) form an interconnected module comprised mainly of signaling proteins. However, having identified probable essential proteins in the ciliary milieu which are likely to be signaling proteins, it would be interesting to observe whether FOXJ1 regulatory network proteins act as crucial modulators that relay the information onto the signaling component. In this regard, 16 regulatory network proteins have been identified as IIP and these proteins were found to interact with FIGp (26), IIP (68) and multiple ciliary expression associated proteins (1255) (Figure 1.8A). Such regulatory network proteins which share extensive interactions with topologically important proteins in the motile cilia interactome are likely to be key mediators acting downstream of FOXJ1 activation modulating ciliogenesis. Additionally, a set of 4 IIP were found to be associated with PCD based on differential expression of their mRNA in a PCD case study. Such IIP found to be directly involved in the FOXJ1 regulatory network and IIP that had associations with PCD along with direct connections to FIGp have been categorized as IIP-effector in the FOXJ1 regulatory network associated protein interactome (Table 1.2). It was interesting to note that some effector proteins (HSP90AA1, CDC42, ACTN2, SSX2IP, PLSCR1, PIAS4) have prior documented roles in cilia biology (Choi et al., 2013; Choksi, Babu, et al., 2014; Croft et al., 2013; Klinger et al., 2014; Ramachandran, Herfurth, Grosschedl, Schäfer, & Walz, 2015; Kohli et al., 2017; Fabregat et al., 2017), however, a set of 14 novel IIP-effector proteins that may act as crucial mediators in the FOXJ1 regulatory network have been identified herein (Table 1.2).

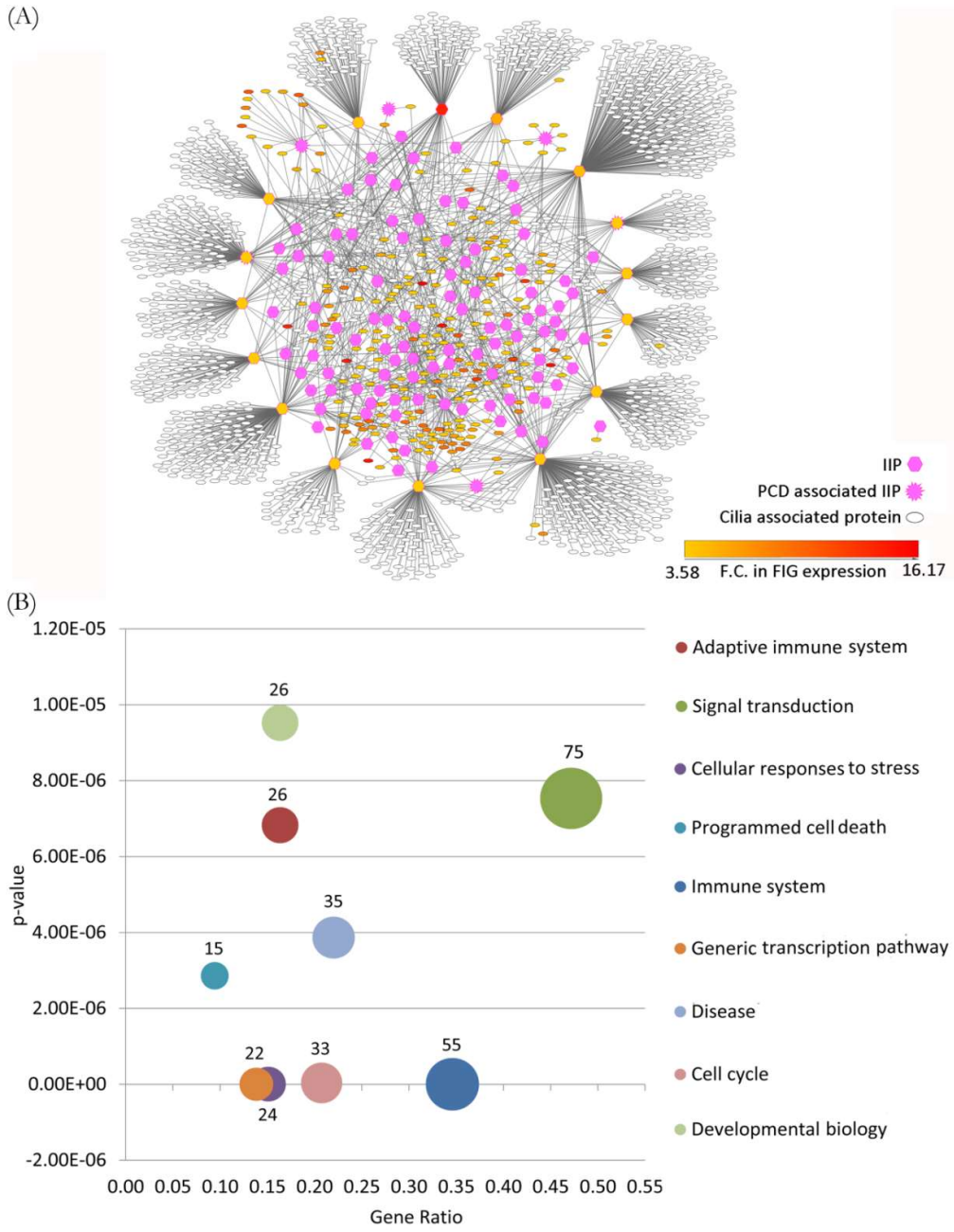


Figure 1.8: **Inter-relationship among IIP and FOXJ1 regulatory network proteins or FIGp.** (A) The inter-connected module within the motile cilia interactome that the IIP (120) and their primary interactors form with FIGp (246) is depicted here. (B) Cellular pathways that the IIP and FOXJ1 regulatory network proteins comprising the inter-connected module are likely to be involved in are shown (enriched pathways with p-value lower than 1e-05). Abbreviations: F.C. - Fold change, FIG- FoxJ1 induced genes, IIP- Important interacting proteins, PCD- Primary ciliary dyskinesia. [Note: published in Mukherjee et al., 2019].

Table 1.2: **IIP-effector proteins in FOXJ1 regulatory network associated protein interactome.** IIP-effector proteins that lie at the interface of the FOXJ1 regulatory network and its associated PPIN are listed here. The possible role of some of these IIP-effectors along with their primary interactors has been outlined.

IIP-Effector Category	Gene Name	[1]IIP category	[2]GO association (Ciliary)	Comment
IIP directly regulated by FOXJ1	ATXN1	HUB & BP & CP		
	EEF1A1	HUB & BP & CP		Associated with BBSome proteins, could be involved in cargo trafficking to cilia
	FKBP5	HUB & BP	Protein folding (GO) [GO:0006457~protein folding]	
	PIAS4	HUB & CP		PIAS4 regulates SUMOylation of Glis2/NPHP7, which is a transcriptional regulator mutated in type 7 nephronophthisis (Ramachandran et al, 2015)
	PLSCR1	HUB & BP		Cilia phenotype defects occur upon knockdown (Choksi, Lauter et al., 2014)
	SOCS3	HUB & CP		
	SSX2IP	HUB & CP	Cilium morphogenesis (GO) [GO:0042384~cilium assembly, GO:0060271~ cilium morphogenesis, GO:0035735~intraciliary transport involved in cilium morphogenesis	SSX2IP targets Cep290 to the ciliary transition zone. Cep290 takes a central role in gating proteins to the ciliary compartment (Klinger et al., 2014)
	STOM	HUB & BP & CP		
	SYNCRIP	HUB & BP		

	MEOX2	HUB & BP & CP		
	NLGN3	HUB & BP		
	FAM19A3	HUB & BP		
IIP indirectly regulated by FOXJ1	APOE	HUB & CP	Cytoskeleton organization (GO)[GO:0007010~ cytoskeleton organization]	
	DLG4	HUB & BP	Establishment or maintenance of epithelial cell apical/basal polarity (GO) [GO:0045197]	
FOXJ1 directly regulated IIP showing differential expression (mRNA) in PCD	ACTN2	CP & GNPP	Actin filament organization (GO)[GO:0051695~actin filament uncapping,GO:0005884~ actin filament], Cytoskeleton organization (GO) [GO:0005856~ cytoskeleton]	RhoA dependent actin remodelling for establishment of an apical web-like structure of actin for basal body docking and axoneme growth (Kohli et al., 2017)
	CASP8	HUB & BP & CP		
IIP showing differential expression (mRNA) in PCD	BTRC	HUB & BP & CP	Protein ubiquitination (GO)[GO:0000209~ protein polyubiquitination,GO:0006511~ubiquitin-dependent protein catabolic process,GO:0043161~ proteasome-mediated ubiquitin-dependent protein catabolic process],Cell cycle (GO)[GO:0051726~ regulation of cell cycle]	
	HSP90AA1	HUB & BP & CP	Protein folding (GO)[GO:0006457~protein folding]	Participates in cilium assembly according to Reactome database (Fabregat et al., 2017; Croft et

				al.,2013)
	TERF1	HUB & CP		
	CDC42	HUB & CP	Establishment or maintenance of cell polarity (GO)[GO:0007163~ establishment or maintenance of cell polarity], Small GTPase mediated signal transduction (GO) [GO:0007264~small GTPase mediated signal transduction],Cytoskeleton organization (GO)[GO:0030036~actin cytoskeleton organization],Actin filament organization (GO) [GO:0051017~actin filament bundle assembly]	Cdc42 docks vesicles carrying ciliary proteins and localizes the exocyst to primary cilia. CDC42 deficiency results in deranged ciliogenesis and polycystic kidney disease. (Choi et al, 2013)

[1]**IIP category:** Different graph theory metric categories (e.g., hubs: HUB, central proteins: CP, bottleneck proteins: BP and global network perturbing proteins: GNPP) that have identified the protein as important interacting protein (IIP).
[2]**GO association (Ciliary):** Predicted GO association of IIP-effector.
[Note: published in Mukherjee et al., 2019]

Moreover, GO mapping and pathway enrichment analysis in accordance with the concept of 'guilt by association' determined that IIP-effectors might be participating in cellular pathways/processes related to cilia biogenesis or function. In particular, GO mapping identified the involvement of IIP-effectors, DLG4 and CDC42 in maintenance of cell polarity based on their presence in GO categories such as 'establishment or maintenance of epithelial cell apical/basal polarity [GO: 0045197]' and 'establishment or maintenance of cell polarity [GO: 0007163]' respectively (Table 1.2, Figure 1.9). Further, based on the Reactome pathway enrichment analysis (G. Yu & He, 2016), IIP-effectors were found to be associated with 'cilia associated signaling pathways' and as such could be topologically important signaling proteins in the FOXJ1 regulatory network associated PPIN. Particularly, IIP-effector proteins, SYNCRIP and BTRC participate in pathways that regulate motile cilium development like cell cycle, Wnt, Fgfr and Notch signaling (Table S4, Figure 1.9). Further, IIP-effectors CASP8, SOCS3, BTRC, PIAS4 and their primary interactors are likely to be involved in pathways related to primary cilium development and function like Hedgehog, TGF-beta and Toll-like receptor signaling (Table S4, Figure 1.9). Therefore, in addition to coding for ciliary structural component proteins FOXJ1 regulatory network genes either code for topologically important

signaling proteins (CASP8, SOCS3, SYNCRIP) or form extensive cross-talk with topologically important signaling proteins (BTRC, HSP90AA1, CDC42). Moreover, some of these 'protein-pathway' associations have not previously been studied in the context of ciliogenesis and it would be interesting to determine whether these IIP-effectors are causal or etiological genes for PCD in further studies.

1.1.1.7 Alterations in important effector proteins under disease conditions might alter the PPIN

The IIP-effectors and their interacting protein partners in the inter-connected module could be involved in Fgfr, Hedgehog, Wnt, Notch, Tgf-beta and Toll-like receptor signaling pathways downstream of FOXJ1 activation (Figure 1.9). This is in accordance with previous reports which have been shown that Notch, Wnt and Fgf signaling pathways regulate processes like cilia length or number in motile cilia bearing cells and left-right patterning (Neugebauer et al., 2009; Lopes et al., 2010; Caron et al., 2012). It is likely that alterations in the 'topologically important signaling proteins' could be disease associated. In particular, IIP-effectors, BTRC and CASP8 have been found to be PCD associated based on differential expression in PCD patients. In addition these IIP-effector proteins along with their primary interactors in the ciliary interactome are possibly involved in mediating Toll-like receptor signaling. Moreover, other IIP and FIGp also have possible involvement in innate immune responses and Toll-like receptor signaling cascades. Additional experimental studies will help establish their causal link to PCD.

Conclusion

The transcriptional network of an essential regulator (FOXJ1) of a cellular process (ciliogeneis) has been predicted and the network of protein-protein interactions that are likely to be governed by this regulatory network has been analyzed. Directly (424) and indirectly regulated genes (148) constitute the predicted FOXJ1 regulatory network. Further, based on GO analysis certain transcription factors and other proteins having ciliary structure or motility related functions were identified among the directly (17) and indirectly (6) regulated genes. For instance, based on GO analysis we could predict that directly regulated genes (MYCBP, HSBP1) and indirectly regulated genes (ATF5, CHD4) likely act as transcription factors probably involved in ciliogenesis.

However, the functionally annotated (nearly 24%) regulatory network genes mainly encode for proteins acting as ciliary structural components, mutations in which may be associated with ciliary ultra structure defects occurring in PCD. Alternately, FIGp might participate in motile cilia assembly or function in a coordinated manner in association with other proteins of the ciliary milieu. Subsequently, an extensive computational analysis of the FOXJ1 regulatory network and the PPIN associated with it allowed the identification of IIP in the motile cilia interactome, among which some proteins (IIP-effectors) are likely to be topologically important signaling proteins in the analyzed interactome (6927 proteins, 40,608 interactions). Interestingly, while 33 IIP have

previously reported ciliary roles, some of these topologically important proteins have been predicted to be involved in multiple signaling pathways, generic transcription, cell cycle, developmental biology etc. This suggests that IIP are likely essential proteins with roles in cilia development or function. Moreover, the IIP-effector proteins that are transcriptionally regulated by FOXJ1 and/or differentially expressed in PCD are likely to have crucial roles in motile cilia. A total of 20 de-regulated topologically important effector proteins were identified in the FOXJ1 regulatory network associated PPIN. Some of these IIP-effectors (PLSCR1, SSX2IP, ACTN2, CDC42, HSP90AA1, PIAS4) have previous literature studies (Choi et al., 2013; Choksi, Babu, et al., 2014; Croft et al., 2013; Klinger et al., 2014; Ramachandran et al., 2015; Kohli et al., 2017; Fabregat et al., 2017) supporting their functional roles in motile cilia.

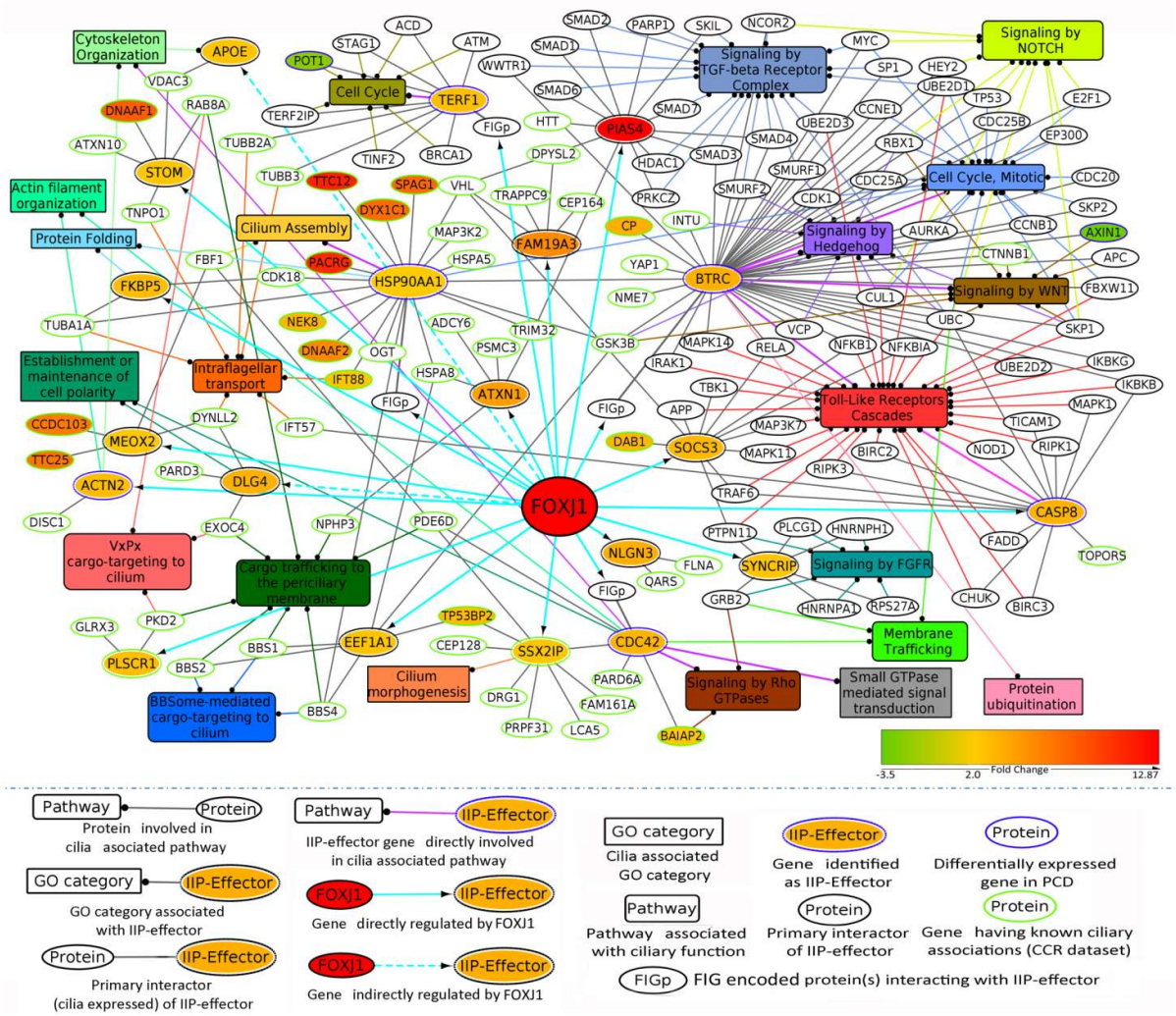


Figure 1.9: **IIP-effector proteins in FOXJ1 regulatory network associated PPIN and enriched 'protein-pathway' associations.** 'Protein-pathway' associations between IIP-effectors and their associated enriched pathways were determined utilizing pathway enrichment and GO analysis among the IIP-effector proteins and their network interactors. These possible ciliary role(s) of IIP-effector proteins are depicted here. Note: Edge color denotes the ciliary process(es) or pathway(s) the gene/protein is/are associated with. Fold changes in genes differentially expressed in PCD case study (Geremek et al., 2014) or Choksi et al. expression study (Choksi, Babu, et al., 2014) are mapped onto the protein nodes. [Note: published in Mukherjee et al., 2019]

Therefore, in this study we have utilized an in-silico knock-out strategy to determine the effects on the motile cilia interactome by considering whether the effective change in a centrality measure as a result of the knock-out varied significantly. Moreover, additional standard graph theory measures computed based on shortest path, degree and centrality of the network were utilized to determine topologically important effector proteins that lie at the interface of the FOXJ1 regulatory network and the linked protein interactome. Thus, proteins acting at the interface of the regulatory and signaling networks could be essential for a cellular process such as ciliogenesis. Furthermore, based on pathway enrichment of the IIP-effectors and their primary interactors, it is likely that 5 among these novel proteins that are involved in cilia associated signaling pathways (like Wnt, Hedgehog, Notch, Toll-like receptor etc.) are likely to act as 'topologically important signaling proteins'.

Inference: Topological analysis of the regulatory network of a master regulator in association with the intra-cellular protein interactome connections of the regulatory network proteins allows the identification of important effector proteins within the regulatory network. The important effector proteins that act as topologically important signaling proteins could be essential for the network integrity and in turn for the cellular process considered herein i.e. ciliogenesis. Thus, in this manner one may determine probable disease associated entities in the regulatory network of a key regulator which when perturbed may lead to disruptions in cellular physiology.

1.1.2 Variation in intra-cellular levels of an essential regulator results in time-dependent adjustments among network component levels significantly impacting cellular homeostasis

Synopsis: The present investigation is an approach to determine whether the regulatory network (miRNA-mRNA interaction network) of an essential regulator is altered upon gradual changes in the intra-cellular concentration of an essential regulator. Such an analysis might enable one to determine which regulatory network components change in a time dependent manner and whether they are likely to be associated with deregulated cellular pathways resulting in a disease phenotype. Herein, we have studied the alterations in miR-122 regulatory network based on the observation that high fat exposure leads to reduced miR-122 levels within hepatocytes.

Problem Statement: Given that the intra-cellular concentration of an essential regulator (miRNA) changes gradually over time, determine whether there are changes in the regulatory network components (mRNA). Additionally, ascertain whether changes in the network components over time could be associated with deregulation in cellular pathways and alterations in cellular phenotype.

Hypothesis: miRNAs may have crucial roles in maintaining robustness of biological networks and loss of certain essential miRNA may contribute to phenotypic changes. Further, gradual changes in the intra-cellular levels of an essential regulator may be reflected in the regulatory network components or associated cellular pathways that the targets (mRNA) participate in.

System of study: miRNA mediate gene transcriptional and post-transcriptional regulation of protein-coding mRNAs and the synergistic effects among multiple miRNA and their actions on the targets forms a regulatory network between miRNA and mRNA targets (Pu et al., 2019). Cellular networks are highly resilient towards perturbations (Albert, Jeong, & Barabási, 2000) and while miRNAs play crucial roles in maintaining robustness of biological networks, loss of certain miRNA may lead to phenotypic changes (Peláez & Carthew, 2012). Interestingly, it has been observed that the hepatocyte-specific miRNA, miR-122 which constitutes 70% of the total miRNA pool has metabolic, anti-inflammatory, and anti-tumorigenic functions and therefore is essential in the maintenance of liver homeostasis (Jopling, 2012; Hsu et al., 2012). In this context, our collaborator identified that high fatty acid or cholesterol exposure leads to exosome mediated exocytosis of miR-122 from hepatocyte cells. This exosome mediated release is likely to result in a gradual decrease in intra-cellular levels of the essential regulator (miR-122). However, the effect of this alteration in cellular levels of miR-122 on its target gene network and associated cellular pathways is not well understood. It is likely that there occurs a time dependent change in the regulatory network of this essential regulator that primarily governs liver homeostasis. The present investigation has been undertaken to understand the response of biological networks towards perturbations in essential regulators. Herein, time dependent changes in expression levels of the target genes of the essential regulator have been identified at the systems level and pathways

that might probably be altered as a result of these changes in hepatic cells have been predicted. Thus, resilience of cellular networks towards perturbations may be lost under certain conditions and changes in network components under such conditions are likely to be associated with altered cellular phenotype(s).

Graphical Summary:

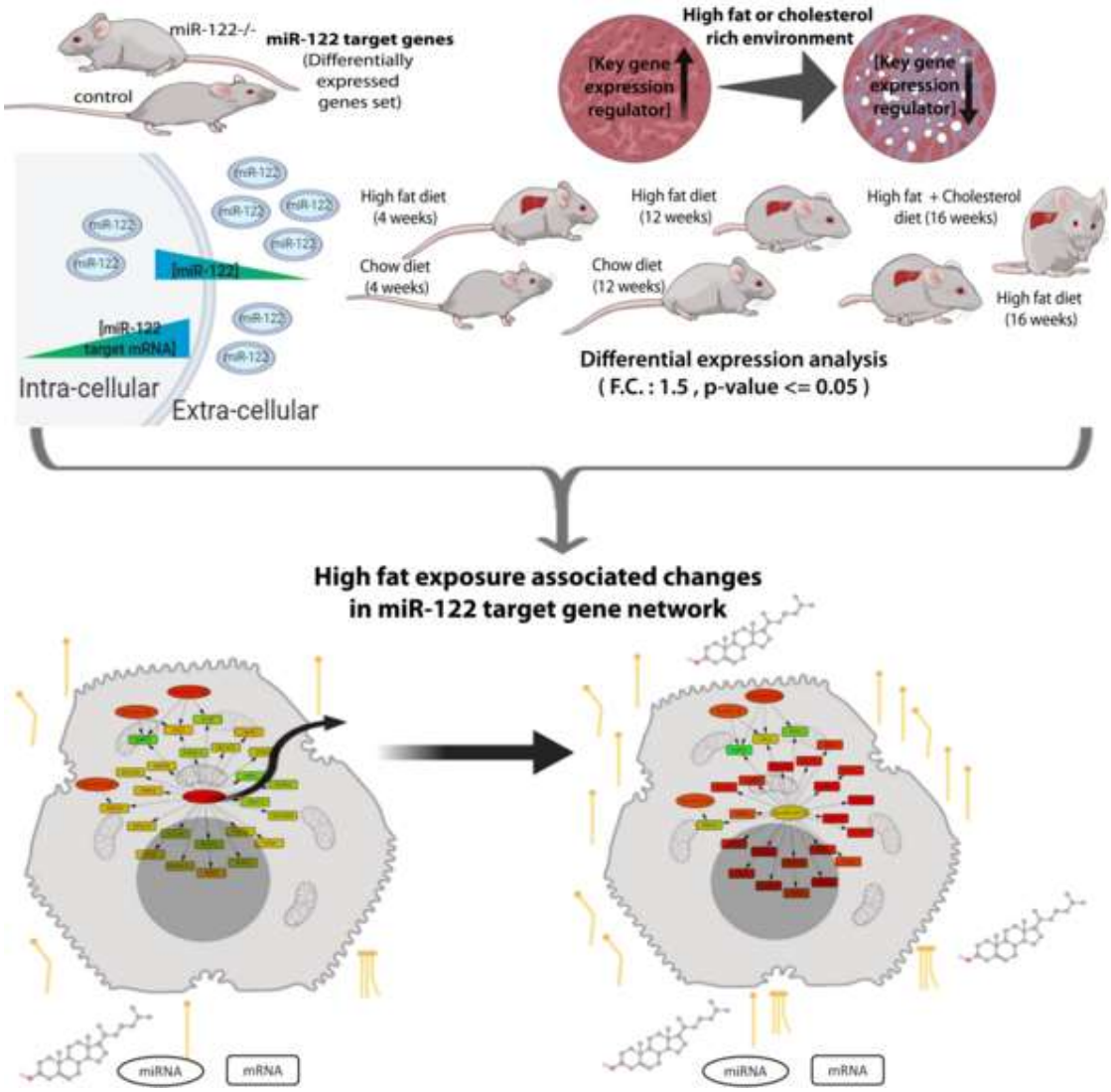


Figure 1.10: **Analyzing the effect of alteration in intra-cellular composition of an essential regulator (miRNA).** High fat exposure associated changes in hepatocyte specific essential regulator (miR-122) network have been studied with the help of expression analysis to determine regulatory network components that are likely to be altered under this condition.

Alteration in the intra-cellular composition of an essential regulator (miRNA) might gradually affect some regulatory network components, particularly signaling cross-talk mRNA (proteins)

1.1.2.1 High fat diet exposure associated changes in intra-cellular hepatic gene expression profile

Exposure of hepatocyte cells to excessive lipid content may lead to the exocytosis of miR-122 from hepatocytes and as a result changes in the expression patterns of miR-122 target genes is expected. This has been studied with the help of expression profiling of hepatocyte tissues isolated from high fat diet fed mice over a course of time (4-16 weeks). It would be interesting to note whether changes occur in miR122 target gene levels on short term or long term exposure of liver cells to high fat content and whether high cholesterol (CH) exposure also results in changes in miR122 target gene levels. Thus, in order to get overviews regarding the changes in the mRNA levels of genes in hepatic tissues under conditions of high fat exposure, differential expression analysis in high fat diet (HFD) fed mouse for 4 weeks (HFD4), 12 weeks (HFD12) or cholesterol diet (CH) fed mouse for 16 weeks (CH16) was performed. On short term or long term exposure of liver cells to high fat or cholesterol content, a higher number of genes were found to be commonly up-regulated in comparison to the commonly down-regulated genes (Table 1.3). Generally, high fat diet exposure may induce a re-distribution in the levels of miR-122, (a hepatocyte specific microRNA which predominates the microRNA content in the adult liver (Chang et al. 2004; Jopling et al. 2012)) in the intra-cellular milieu of hepatocytes resulting in large scale expression analysis changes in the hepatic cells. A comparison among high diet exposure associated up-regulated hepatocyte genes and predicted miR-122 target genes revealed that nearly 29.37% (178) of the total possible miR-122 target genes (606) show a significant change in their expression levels under high fat diet treatment conditions. Moreover, among the 256 genes that were found to be commonly up-regulated in two or more high fat diet or cholesterol treatment cases considered herein 18.36% (47) of them were found to be miR-122 targets (Figure 1.11A). However, very few (7) high fat diet exposure associated down-regulated genes were found to be miR-122 target genes (Figure 1.11B). Thus, we observed that high fat diet exposure associated stress results mainly in the up-regulation of miR-122 regulatory network target genes.

Table 1.3: Differentially expressed genes in hepatic tissues under cholesterol or fatty acid exposure associated stress

	HFD 4 weeks (GSE53381)	HFD 12 weeks (GSE93819)	CH 16 weeks (GSE58271)
Up-regulated genes	347	2550	184
Down-regulated genes	143	368	38
Up-regulated miR-122 target genes	37	160	35
Down-regulated miR-122 target genes	3	6	1

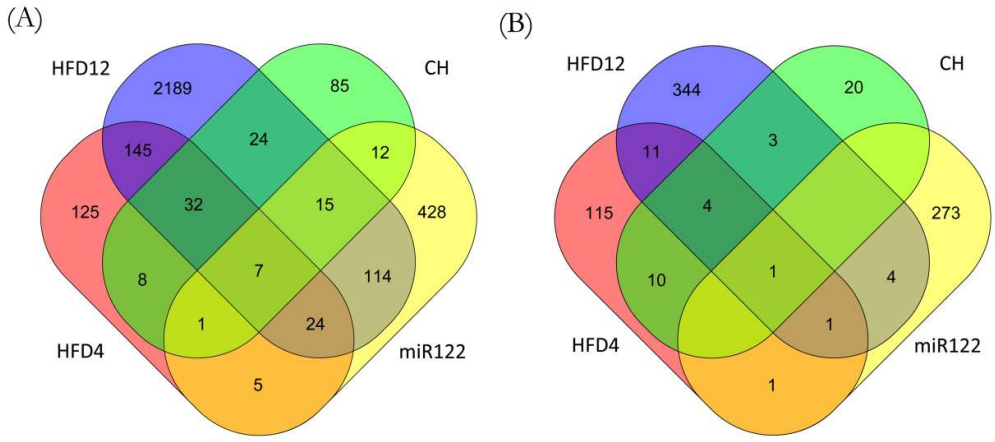


Figure 1.11: **Hepatocyte cells in high fat diet treated mouse models exhibit alterations in expression levels of miR-122 target genes.** Comparison of differentially expressed genes in mouse livers exposed to high fat or cholesterol concentrations for different time periods 4, 12 and 16 weeks results in alterations in miR-122 target genes. (A) The overlap in number of up-regulated genes that occur upon high fat/cholesterol treatment and upon miR-122 knock-out in mice models as represented in the Venn diagram depicts the number of miR-122 target genes that show up-regulation upon exposure of hepatocyte cells to elevated lipid content. (B) The overlap in number of down-regulated genes that occur upon high fat/cholesterol treatment and upon miR-122 knock-out in mice models as depicted in the Venn diagram illustrates the number of miR-122 target genes that get down-regulated under high fat diet treatment conditions.

1.1.2.2 miR-122 target gene network in hepatocytes is altered in a time dependent manner upon exposure to high lipid content

Upon excess fatty acids intake the liver's capacity to handle primary metabolic energy substrates such as fatty acids may be overwhelmed and under this circumstance ER stress or hepatocellular stress may serve as a predisposition to cirrhosis and

hepatocellular carcinoma (Friedman et al., 2018). Since, the hepatocyte-specific miRNA, miR-122 constitutes about 70% of the total miRNA pool in the liver, it is likely to be essential in the maintenance of liver homeostasis by regulating processes such as de-novo lipogenesis, fatty acid oxidation and triglyceride export (Jopling 2012, Hsu et al., 2012, Liu et al., 2015). Thus, herein based on the assumption that under conditions of excessive fatty acid exposure exocytosis of miR-122 may occur from hepatocytes, the probable miR-122 target gene network in hepatocytes has been studied under different conditions of high fat exposure. A number of miR-122 target genes have been found to be up-regulated within hepatic cells of mice models upon exposure to high fat concentrations for 4 weeks and 12 weeks or upon high cholesterol exposure for 16 weeks (Figure 1.12A) while very few miR-122 target genes get down-regulated under these conditions (Figure 1.12B). Further, some (7) target genes were found to be consistently up-regulated under conditions of lipotoxic stress while 47 miR-122 target genes were found to be differentially expressed in two or more conditions of lipotoxic stress studied herein (Figure 1.12C, D).

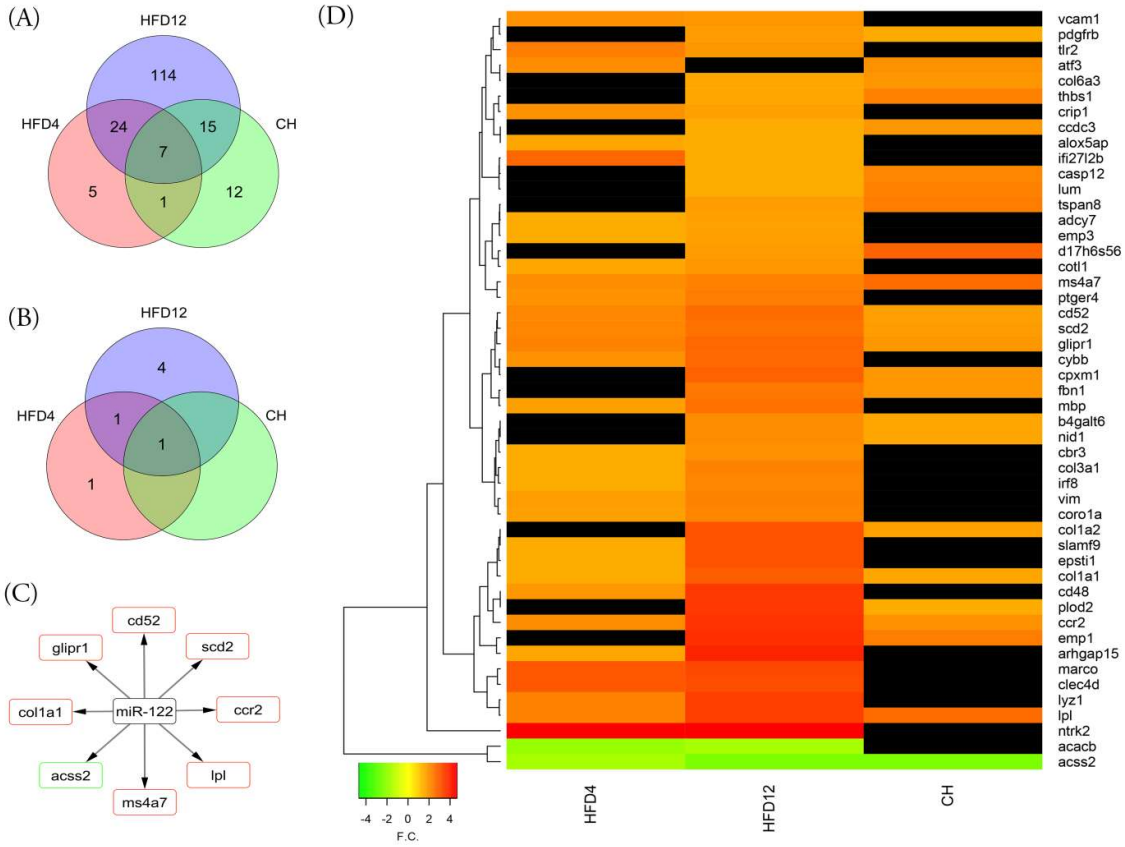


Figure 1.12: **Alterations in miR-122 target genes occurs in a time dependent manner upon exposure of hepatocyte cells to high lipid content.** (A) Venn diagram representing number of up-regulated miR-122 target genes in hepatocyte tissues of mice models on high fat diet or cholesterol treatment (B) Venn diagram representing number of down-regulated miR-122 target genes in hepatocyte tissues of mice models

on high fat diet or cholesterol treatment (C) Network of miR-122 target genes in hepatocytes that are consistently de-regulated upon short term or long term high fat or cholesterol exposure (D) Heatmap representing miR-122 target genes that exhibit changes in two or more HFD analysis case studies considering HFD 4 weeks, HFD 12 weeks and CH 16 weeks. Abbreviations: HFD12 – High fat diet treatment for 12 weeks, HFD4 – High fat diet treatment for 4 weeks, CH – Cholesterol treatment for 16 weeks

Further, it was observed that short term exposure leads to changes in nearly 6% of the probable miR-122 target genes while nearly 26% of the probable mR-122 target genes get de-regulated upon long term exposure (Figure 1.13). This minor change in miR-122 regulatory network is expected because biological networks are highly resilient towards perturbations and as such miRNAs can play crucial roles in maintaining the robustness of biological networks, however, a change in the level of an essential miRNA such as miR-122 and its corresponding target genes may be detrimental for cellular function. Thus, miR-122 target gene network exhibits time-dependent alterations in network architecture under conditions of hepatic lipotoxic stress (Figure 1.12D, 1.13).

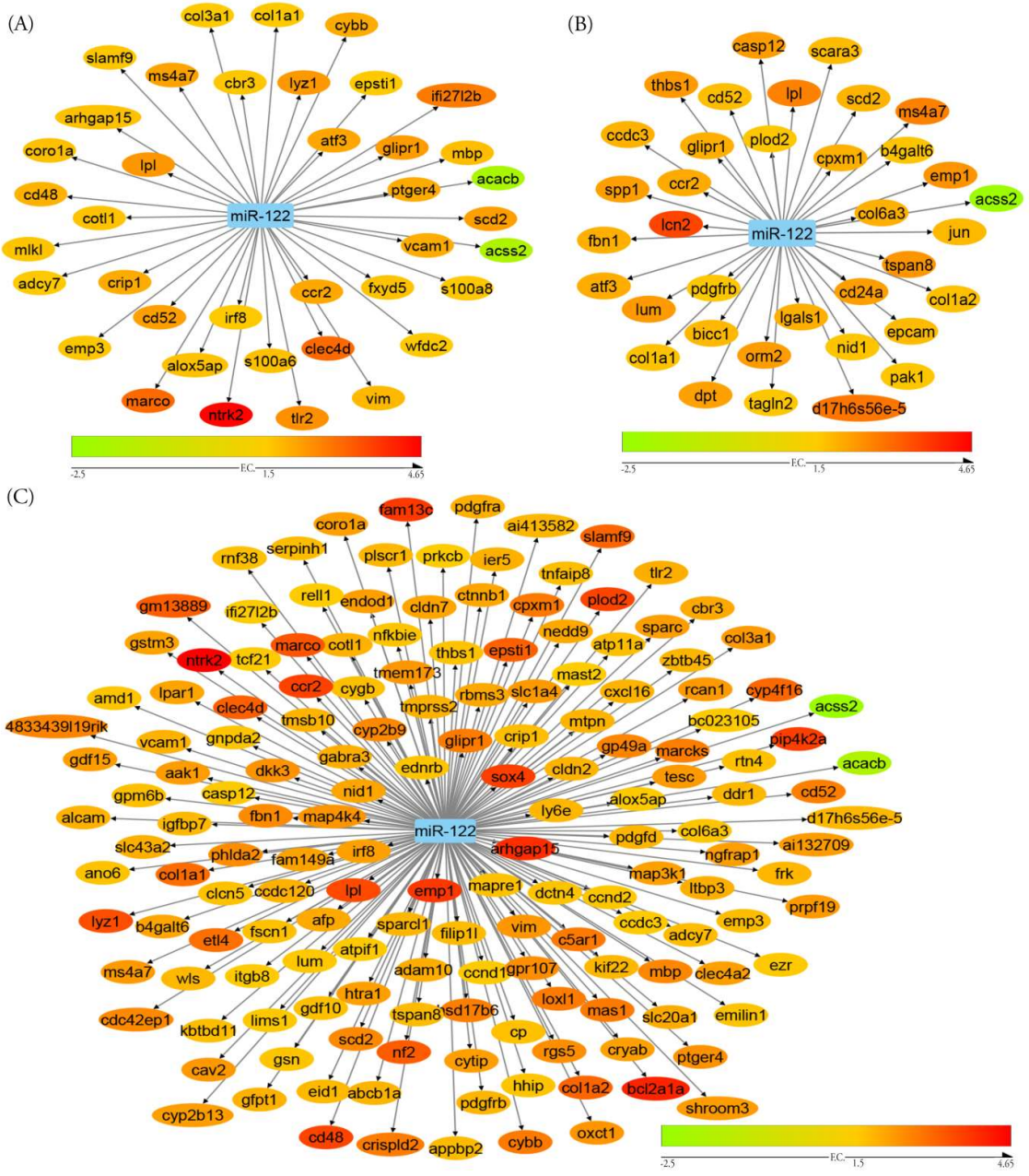


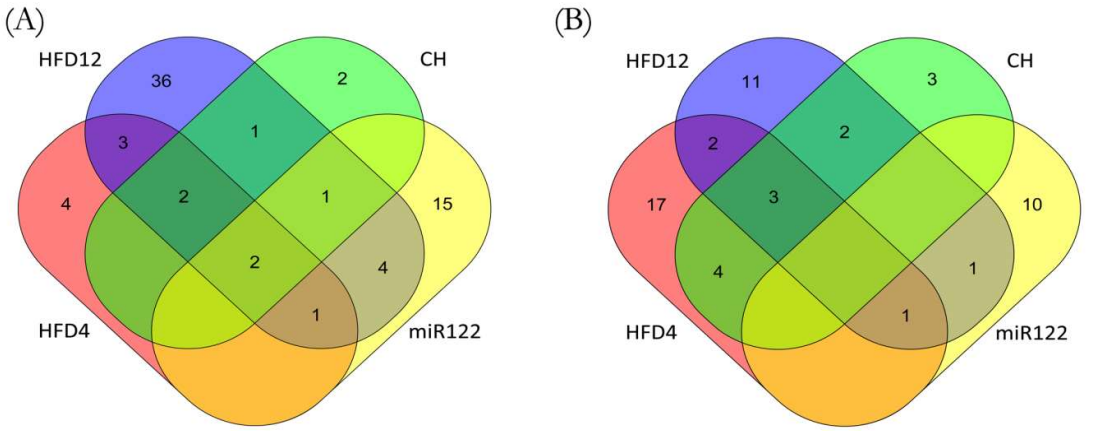
Figure 1.13: High fat diet exposure associated miR-122 target gene network in hepatocyte cells. High fat diet associated miR-122 network genes that exhibit differential expression in HFD case studies were taken into consideration and network genes that exhibit changes in expression profile have been outlined. (A) Altered miR-122 network genes in hepatocytes upon HFD exposure (4 weeks) (B) Altered miR-122 target gene network in hepatocytes under CH exposure (16 weeks) (C) Altered miR-122 target gene network in hepatocytes under HFD exposure (12 weeks).

1.1.2.3 Probable altered network pathways as a result of changes in miR-122 regulatory network within hepatocytes under lipotoxic stress

In order to envisage the probable physiological changes that could be arising as a result of this change in intra-hepatic miR-122 regulatory network architecture, the repertoire of differentially expressed miR-122 target genes and the pathways that they participate in have been taken into consideration. At the outset, it was identified that miR-122 target genes which participate in metabolism associated processes like fatty acid biosynthesis (ACSS2, ACACB) become down-regulated during initial stages of high fat treatment (4 weeks) while other target genes (44) such as CTNNB1, PDGFRA, PDGFRB, CCND1, PRKCB etc. which are mainly associated with cellular process or signaling pathway categories become up-regulated only upon prolonged exposure to high fat content (Figure 1.14).

Figure 1.14: **High fat exposure associated de-regulated miR-122 target genes may lead to alterations in multiple cellular pathways.** Pathway mapping of differentially expressed miR-122 target genes occurring under high fat diet treatment conditions in mice models has been performed. Differentially expressed miR-122 target genes under different high fat diet treatment conditions participate in different cellular pathway categories like carbohydrate metabolism, lipid metabolism, cellular processes and cellular signalling pathways as elucidated here. Abbreviations: HFD12 – High fat diet treatment for 12 weeks, HFD4 – High fat diet treatment for 4 weeks, CH – Cholesterol treatment for 16 weeks, miR122- miR122 knockout mice model

Further, during this investigation certain lipid metabolism associated miR-122 target genes were found to be differentially expressed under conditions of high fat or cholesterol exposure (Figure 1.15), this observation is consistent with previous studies wherein miR-122 knockdown in the liver resulted in changes in lipid metabolism (Wen et al., 2012).



(C)

	Lipid Metabolism							
	Fatty acid biosynthesis	Steroid hormone biosynthesis	Synthesis and degradation of ketone bodies	Glycerolipid metabolism	Sphingolipid metabolism	Arachidonic acid metabolism	Biosynthesis of unsaturated fatty acids	Fatty acid metabolism
Acacb								
Ugt2b38								
Oxct1								
Lpl								
Scd2								
Cyp2b13								
Cyp2b9								
Hsd17b6								
Cbr3								
B4galt6								

HFD12 & miR122 & HFD4 & CH | HFD4 & miR122 & HFD12 | CH & miR122 | Down-regulated genes
HFD12 & CH & miR122 | HFD12 & miR122 | HFD4 & miR122 | Up-regulated genes

Figure 1.15: **Lipid metabolism related genes may be involved in regulating hepatic cell stress arising as a result of high lipid exposure.** (A) The number of lipid metabolism associated miR-122 target mRNA that were found to be up-regulated upon exposure to high fat conditions for different time periods are shown here. (B) The number of lipid metabolism associated miR-122 target mRNA that becomes down-regulated under high fat diet treatment conditions in mice models is depicted here. (C) Differentially expressed miR-122 target genes under conditions of high fat diet treatment that are involved in lipid metabolism associated pathways have been depicted here. Abbreviations: HFD12 – High fat diet treatment for 12 weeks, HFD4 – High fat diet treatment for 4 weeks, CH – Cholesterol treatment for 16 weeks, miR122-miR122 target gene set identified in miR122 $^{(-/-)}$ knockout mice model.

In particular, certain miR-122 target genes like Scd2 (a key regulator of energy metabolism) that participates in fatty acid metabolism was found to be up-regulated. Additionally, miR-122 target genes involved in downstream processing of cholesterol (CYP2B13; CYP2B9 and HSD17B6), ketone body catabolism (OXCT1) and arachidonic acid metabolism (CYP2B13; CYP2B9, CBR3) were found to be up-regulated. Moreover, miR-122 target gene, ACACB (which regulates the rate-limiting step of fatty acid synthesis) was found to be down-regulated and as a consequence fatty acid biosynthesis could be limited in this scenario (Figure 1.15C). Thus, miR-122 target genes involved in regulating hepatic lipid homeostasis, for instance, genes involved in carbohydrate metabolism (ACSS2) or lipid metabolism (fatty acid biosynthesis, cholesterol synthesis, unsaturated fatty acid synthesis eg. ACSS2, ACACB) associated processes get down-regulated upon short term (HFD 4 weeks) or long term (HFD 12 weeks, CH 16 weeks) high fat exposure.

Based on the differential expression analysis of mice livers exposed to lipotoxic stress, it was observed that substantial changes occur in the expression profile of the miR-122 target genes. Additionally, pathway mapping of these differentially expressed miR-122 target genes elucidated that miR-122 target genes involved in metabolism are mainly down-regulated whereas miR-122 target genes involved in cellular process or cellular signalling pathways are up-regulated. In particular, multiple de-regulated miR-122 target genes are involved in cellular pathways associated with different biological processes like cell motility, proliferation, differentiation, regulation of gene expression and survival like focal adhesion, tight junction or actin cytoskeleton regulation (Figure 1.16).

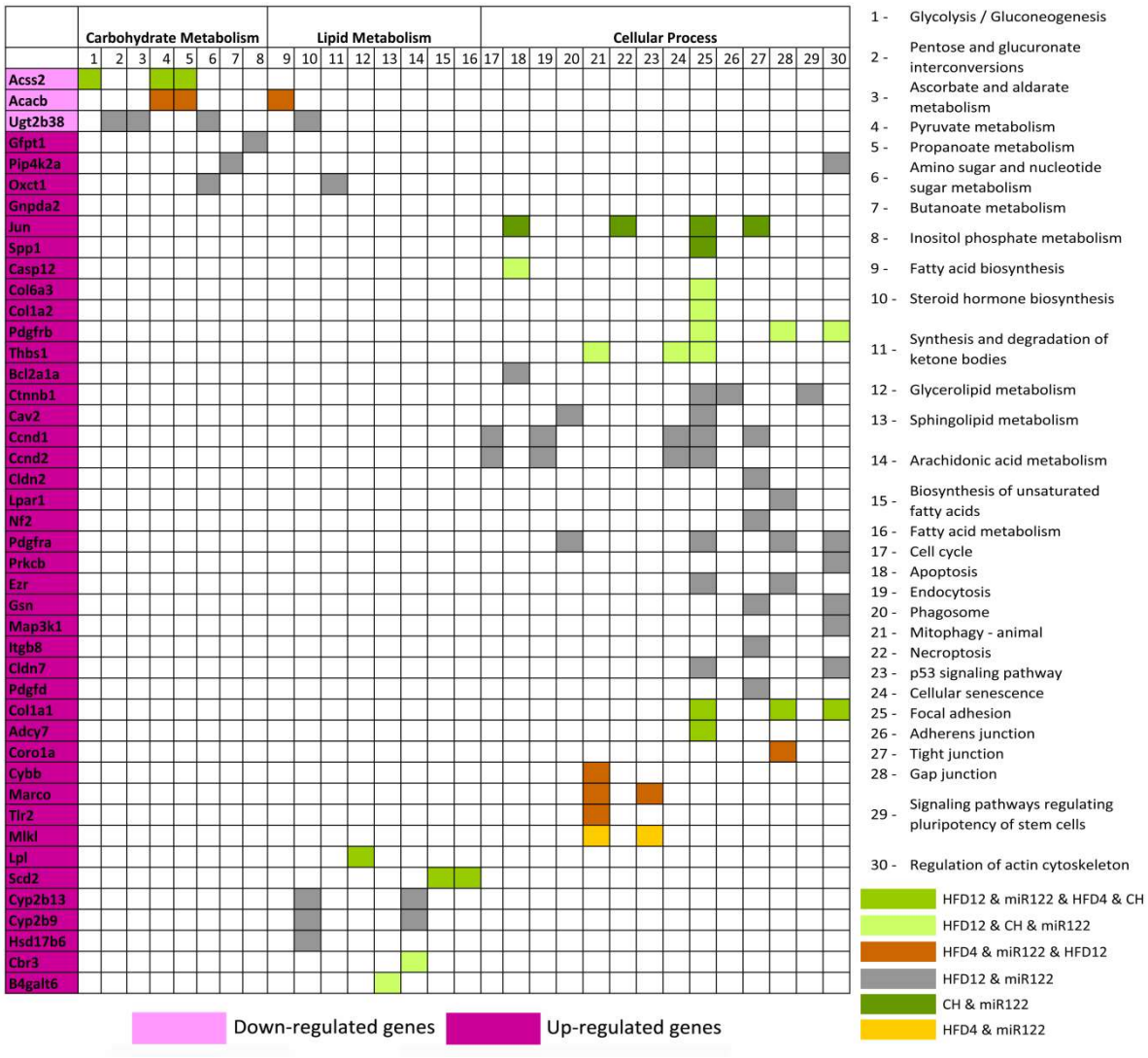


Figure 1.16: **Cellular metabolism and cellular process associated pathways may be involved in regulating high fat exposure related hepatic cell stress.** Differentially expressed miR-122 target genes that occur upon high fat exposure in hepatic cells participate in multiple pathways belonging to carbohydrate metabolism, lipid metabolism or cellular process as elucidated here. Abbreviations: HFD12 – High fat diet treatment for 12 weeks, HFD4 – High fat diet treatment for 4 weeks, CH – Cholesterol treatment for 16 weeks, miR122- miR122 knockout mice model

Additionally, many miR-122 target genes that show an altered expression profile participate in the PI3K/AKT signalling, ECM-receptor interaction pathway while some up-regulated miR-122 target genes were found to encode for important cross-talk proteins (PDGFRA, PDGFRB, ADCY7, CTNNB1, PRKCB and CCND1) which participate in multiple signalling or cellular process pathways (Figure 1.17). Thus, based on this analysis it appears that the initial export of miR-122 from hepatocyte cells in response to lipotoxic stress may have protective effects on the cells by aiding in re-adjustment of

cellular lipid levels.

However, at later time points a shift in miR-122 equilibrium due to prolonged intake of high fat diet if not corrected may have detrimental effects on the overall liver physiology as envisaged by the substantial changes observed in expression profile of miR-122 target genes encoding signalling cross-talk proteins or multiple proteins involved in a range of cellular process and signalling pathways.

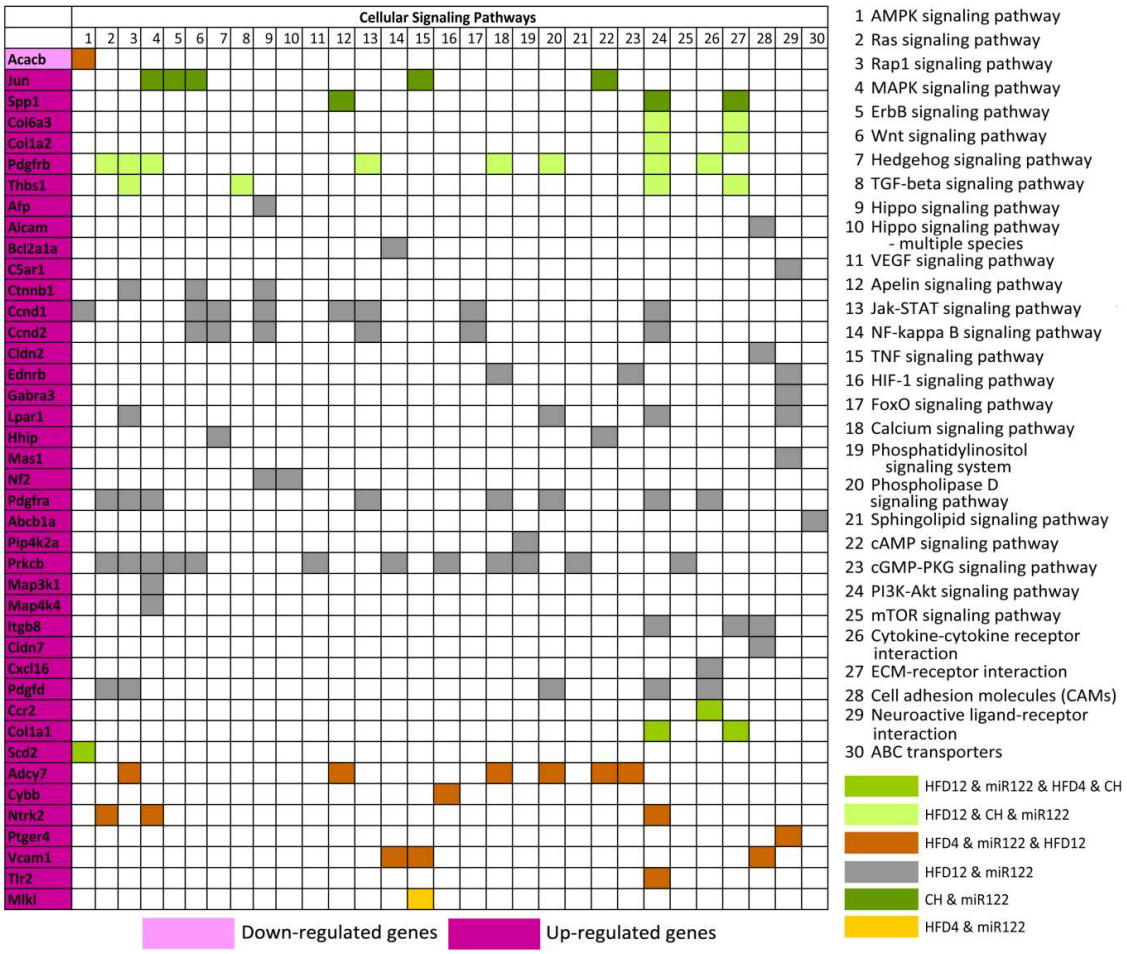


Figure 1.17: **Chronic high fat exposure associated stress in hepatic cells may lead to the deregulation of multiple cellular signalling pathways.** Pathway mapping of differentially expressed miR-122 target genes that occur upon prolonged high fat diet treatment in mice models determined that multiple proteins encoded by these genes participate in a range of cellular signalling pathways as exemplified here. Abbreviations: HFD12 – High fat diet treatment for 12 weeks, HFD4 – High fat diet treatment for 4 weeks, CH – Cholesterol treatment for 16 weeks, miR122- miR122 knockout mice model

Conclusion

In this study, it has been considered that exposure to high cellular lipid content may lead to the exocytosis of miR-122 from hepatocytes and as a result changes in the expression patterns of miR-122 target genes may occur. Herein, the miR-122 regulatory network has been studied with the help of differential expression analysis over a course of time in hepatocyte tissues isolated from high fat diet fed mice. In this context, it was observed that initial high fat exposure results in deregulation of miR-122 regulatory network genes (10) involved in regulating cellular metabolism, however, multiple components in miR-122 regulatory network are altered as a result of prolonged fatty acid exposure associated stress. In particular, miR-122 target genes (47) involved in multiple cellular processes or signaling pathways become up-regulated upon prolonged high fat exposure and alterations in key signaling pathway cross-talk proteins could be associated with liver pathobiology. Interestingly, in this respect similar to the finding in this study PI3K/AKT signalling pathway has been previously implicated to be associated with metabolic dysfunctions in obesity, metabolic syndrome and NAFLD (Matsuda et al., 2013). Moreover, among the up-regulated important cross-talk proteins some proteins such as PDGFRA, PDGFRB and CCND1 have also been previously associated with liver pathobiology (Kocabayoglu et al., 2015; Kikuchi et al., 2015; Deane et al., 2001). Therefore, based on these analyses it was identified that high fat exposure may lead to initial alterations in miR-122 target gene network that might promote re-adjustment of cellular lipid levels. However, over time prolonged high fat exposure may result in substantial alterations in the miR-122 regulatory network architecture that may have detrimental effects on the overall liver physiology since it could potentially lead to a de-regulation of multiple cellular process and signaling pathways.

Inference: High fat or cholesterol rich diet leads to an increase in hepatic lipid load and alterations in miR-122 network components in the intra-cellular (hepatic) environment wherein miR-122 is an essential regulator of liver homeostasis. However, alterations in the levels of the essential regulator result in changes in a minor fraction of the network components albeit gradually over a period of time but the de-regulated expression of certain mRNA that encode for signaling cross-talk proteins may lead to disruptions in cellular network robustness.

1.1.3 Regulatory relationship between network components may vary in a cellular phenotype specific manner

Synopsis: The present investigation is an approach to study regulatory relationships prevalent in miRNA-mRNA interaction networks under disease conditions. MiRNA regulatory networks are comprised of miRNAs that can concurrently down-regulate multiple target mRNAs and mRNAs that are tightly regulated by these multiple miRNA. Multiple miRNA and their target mRNA may be deregulated in disease conditions leading to changes in the regulatory network behavior and disease associated changes in cellular phenotype. Therefore, by analyzing miRNA-mRNA interaction networks one can determine whether unusual regulatory relationships for instance up-regulation of target mRNA upon up-regulation of their cognate regulatory miRNA is prominent under disease conditions.

Problem Statement: Given the intra-cellular expression levels of regulators (miRNA) and their targets (mRNA), determine which regulatory network components (mRNA) or the relationship between regulator-target (miRNA-mRNA) pairs are likely to be involved in disease associated changes in cellular phenotype.

Hypothesis: MicroRNAs primarily down-regulate target mRNA levels or expression post-transcriptionally or translationally, however, they may additionally stimulate gene expression by direct and indirect mechanisms. This alternate regulatory relationship wherein miRNA up-regulate target mRNA expression may be present in different cell types including tissues with de-regulated microRNA expression under disease conditions.

System of study: Non coding RNAs like miRNAs (20-22 nucleotides long) modulate gene expression by binding to the 3' UTR of their target mRNAs and bring about either mRNA degradation or translational suppression (Filipowicz et al., 2008). However, some reports have suggested that miRNA may function in mRNA stabilisation or upregulation as well (Vasudevan et al., 2007). Utilising miRNA and mRNA expression profiling data we have studied the possibility of existence of unusual regulatory relationships among miRNA and their corresponding target mRNA. For this purpose, we have considered different cell types such as degenerative (Alzheimer's disease) or highly proliferative cell types (breast cancer, colorectal cancer and oral squamous cell carcinoma). This is because miRNAs are involved in regulating biological processes like cell cycle, differentiation, proliferation, apoptosis etc. (Si et al., 2019). Aberrant spatial and temporal expression or dysfunction of brain-enriched miRNAs that regulate a wide range of target mRNAs essential for neuronal or glial development has been observed in Alzheimer's disease (Satoh 2012). Moreover, with the help of miRNA profiling, it has been observed that miRNA expression is dysregulated in cancer wherein amplification or deletion of miRNA genes, abnormal transcriptional control of miRNAs or defective miRNA biogenesis may contribute to the development of cancer hallmarks (Peng and Croce, 2016). Thus, by determining miRNA-mRNA interaction networks under these disease conditions the relationship between expression patterns of deregulated miRNA

and their target mRNA has been studied (Figure 1.18). Based on this analysis, it could be determined that unusual regulatory relationships between miRNA and their target mRNA may be prevalent in degenerative or highly proliferative cell types in disease conditions such as Alzheimer's and cancer.

Graphical Summary:

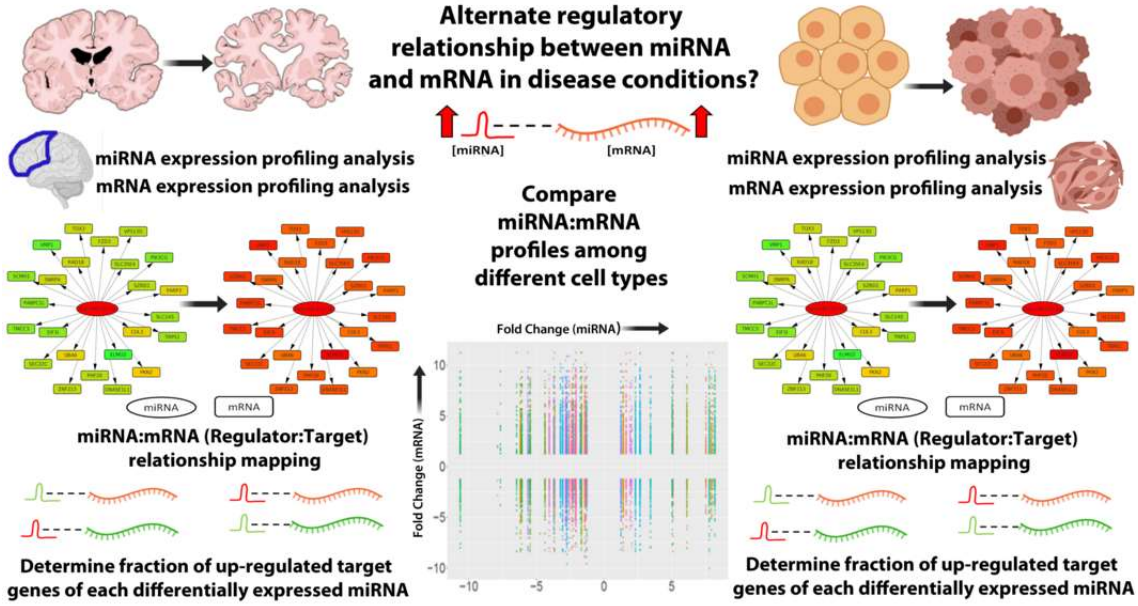


Figure 1.18: **Studying probable regulatory relationships in miRNA-mRNA interaction networks.** Disease associated changes in regulatory network components and regulatory relationships between them may be studied with the help of large scale expression profiling datasets.

Alternate regulatory relationships may be prevalent among miRNA-mRNA target pairs in regulatory networks governing different cellular physiologies

1.1.3.1 miRNA-mRNA interaction networks may be deregulated in frontal cortex of Alzheimer's disease patients

Generally, miRNA can fine tune the expression of multiple mRNA targets and aberrant miRNA expression profiles may contribute to dysregulated expression patterns of target mRNA and in turn de-regulated miRNA-mRNA interaction networks. Based on differential expression analysis, a set of 30 up-regulated miRNA and 23 down-regulated miRNA that are likely to be related to Alzheimer's disease aetiology or progression were identified in the pre-frontal cortex of Alzheimer's disease patients (Figure 1.19 and Table S5). Additionally, considering mRNA expression profiling data multiple mRNA

were found to be up-regulated and down-regulated within the superior frontal gyrus, frontal cortex and dorso-lateral prefrontal cortex of Alzheimer's disease (AD) patients respectively (Table 1.4). Subsequently, in order to study de-regulated miRNA-mRNA interaction networks in AD, the miRNA:mRNA interaction data and "regulator(s):target(s)" expression profiles particularly the differentially expressed miRNAs and their corresponding target mRNAs were utilised to predict the probable de-regulated miRNA-mRNA interaction network (Table 1.5).

Table 1.4: De-regulated mRNA in cortical regions of Alzheimer's disease patients. Numbers of de-regulated mRNA identified based on differential expression analysis considering brain tissue from multiple cortical regions of Alzheimer's disease patients has been enumerated here.

	Frontal cortex (GSE15222)	**Superior Frontal gyrus (GSE5281)**	**Dorsolateral pre-frontal cortex (GSE53697)**
Up-regulated mRNA	2166	2092	292
Down-regulated mRNA	699	1253	676

Table 1.5: Probable altered miRNA-mRNA interaction networks in Alzheimer's disease. De-regulated miRNA and their target mRNA in cortical regions of Alzheimer's disease patients were utilised to determine the dysregulated miRNA-mRNA interaction networks in AD. Numbers of differentially expressed miRNA and mRNA that comprised the altered miRNA-mRNA interaction networks in different cortical regions have been tabulated. Abbreviations: UP- upregulated, DN-downregulated.

GSE48552	**Frontal cortex (GSE15222)**		**Superior frontal gyrus (GSE5821)**		**Dorsolateral pre-frontal cortex (GSE53697)**	
	mRNA UP	**mRNA DN**	**mRNA UP**	**mRNA DN**	**mRNA UP**	**mRNA DN**
miRNA UP (30)	872	329	701	928	128	292
miRNA DN (23)	438	133	370	366	43	142

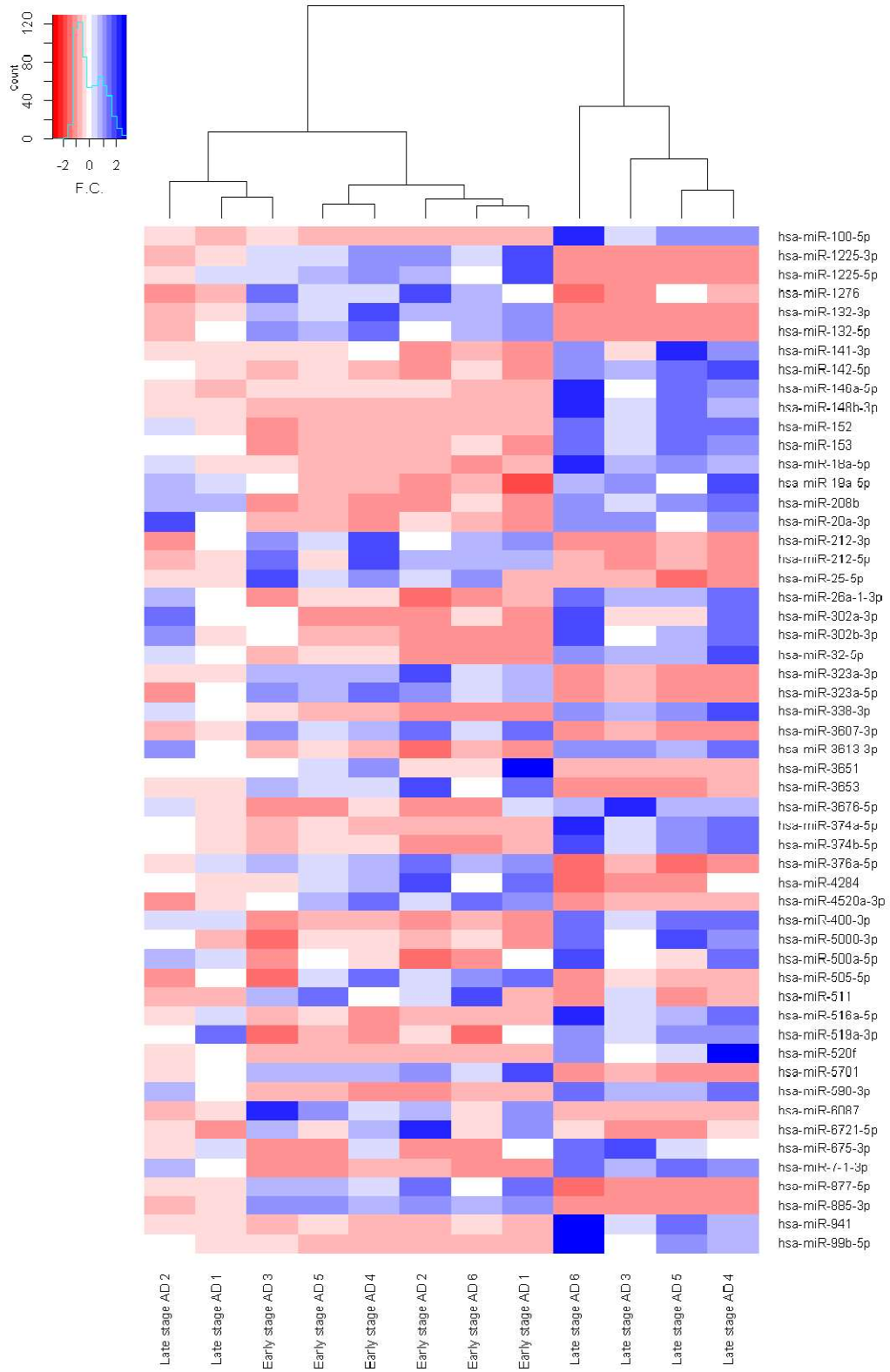


Figure 1.19: **Heat map representing the differentially expressed miRNAs in pre-frontal cortex of late stage Alzheimer's disease (AD) patients in comparison to early stage AD patients.**

Moreover, 253 genes (mRNA) were found to be commonly up-regulated while 350 genes (mRNA) were found to be commonly down-regulated among two of the different brain cortical regions considered herein for the analysis. Therefore, to obtain an insight regarding the set of miRNA and mRNA that get altered in disease conditions, the "regulator(s) to target(s)" (miRNA:mRNA) expression profiles were determined in different brain regions of Alzheimer's disease patients considering the de-regulated miRNA-mRNA interaction networks. Based on these altered miRNA-mRNA interaction networks, preliminarily it was observed that up-regulated miRNA may also regulate up-regulated mRNA (Figure 1.20). The corresponding fraction of up-regulated target mRNA of each of the differentially expressed miRNA was determined in different cortical regions (frontal cortex, superior frontal gyrus and dorsolateral pre frontal cortex) of the human brain (Figure 1.21A). Herein, it was found that in comparison to the down-regulated miRNA, nearly 79% of the up-regulated miRNA had a substantial fraction (0.6-0.8) of their target mRNA as up-regulated in the frontal cortex (Figure 1.21B, Table S6). Further, the observation that in comparison to the down-regulated miRNA, a higher percentage of the up-regulated miRNA have a substantial fraction of their target mRNA as up-regulated was consistently observed within the other cortical regions considered in this study as well. In particular, in comparison to the down-regulated miRNA, 88% and 48% of the up-regulated miRNA had between 0.4-0.6 or 0.2-0.4 fraction of their target mRNA as up-regulated within the superior frontal gyrus and dorsolateral pre-frontal cortex, respectively (Figure 1.21A, Table S6).

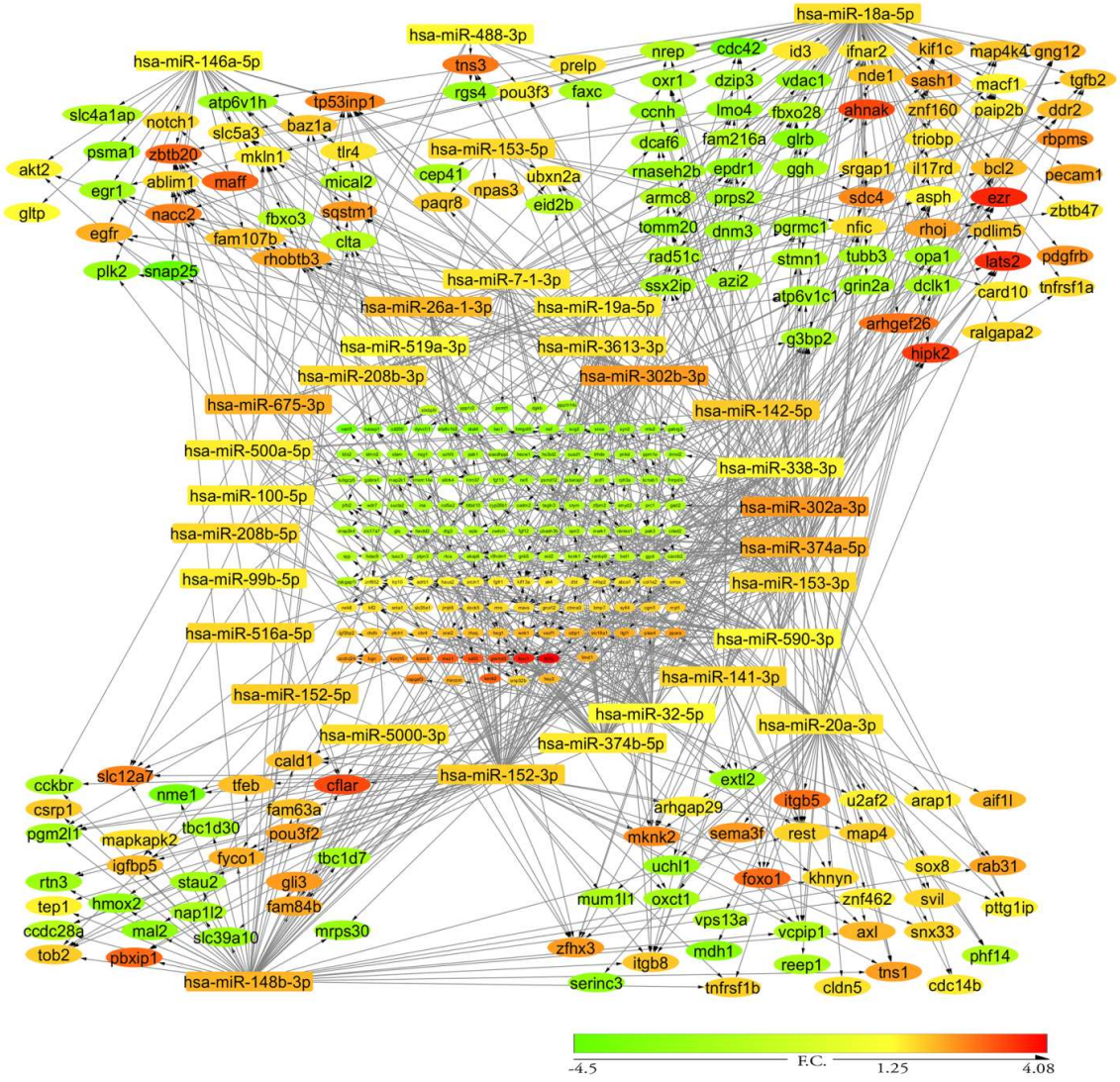


Figure 1.20: De-regulated miRNA-mRNA interaction network in Alzheimer's disease patients. Disease associated miRNA and their de-regulated target mRNA in the altered miRNA-mRNA interaction network have been depicted here. The set of target genes that were found to be commonly differentially expressed in multiple cortical regions of Alzheimer's disease patients have been utilised to determine a representative de-regulated miRNA-mRNA interaction network in AD. MiRNA-target mRNA interactions has been highlighted for few sets of miRNA to indicate that target mRNA of up-regulated miRNA may also be up-regulated. Further, mRNA expression profiling data from the frontal cortex has been utilised to delineate the up-regulated and down-regulated target mRNA.

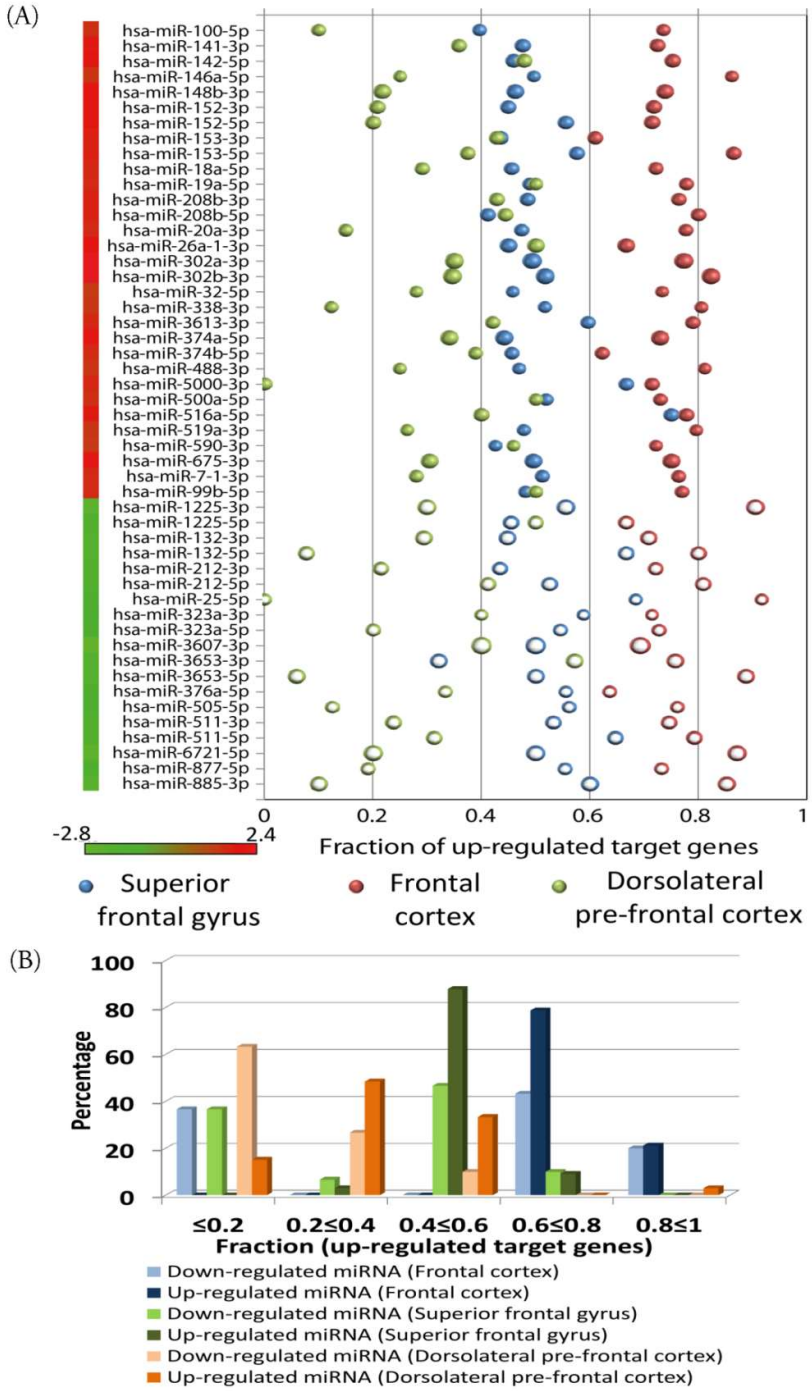


Figure 1.21: Possible miRNA:mRNA expression relationships in Alzheimer's disease. (A) The probable relationship between differentially expressed miRNA and mRNA in Alzheimer's disease considering the possible fraction of up-regulated target mRNA of each of the miRNA as studied in different cortical regions (frontal cortex, superior frontal gyrus and dorsolateral pre-frontal cortex) has been depicted here. (B) Percentages of up-regulated or down-regulated miRNA that possibly have a fraction of their target genes as up-regulated within the frontal cortex, superior frontal gyrus and dorsolateral pre-frontal cortex have been represented here.

However, it is possible that similar "regulator(s) to target(s)" relationship patterns may be prevalent in other diseases or cell types. In this respect, initially by determining whether these de-regulated mRNA could be related to any other disease condition other cell types that can be considered for analysis have been determined. Utilizing a disease enrichment analysis considering the set of commonly up-regulated mRNA, it was determined that these genes are de-regulated in cancer cells and thus are also known to be associated with 'diseases of cellular proliferation' or 'cancer' (Figure 1.22).

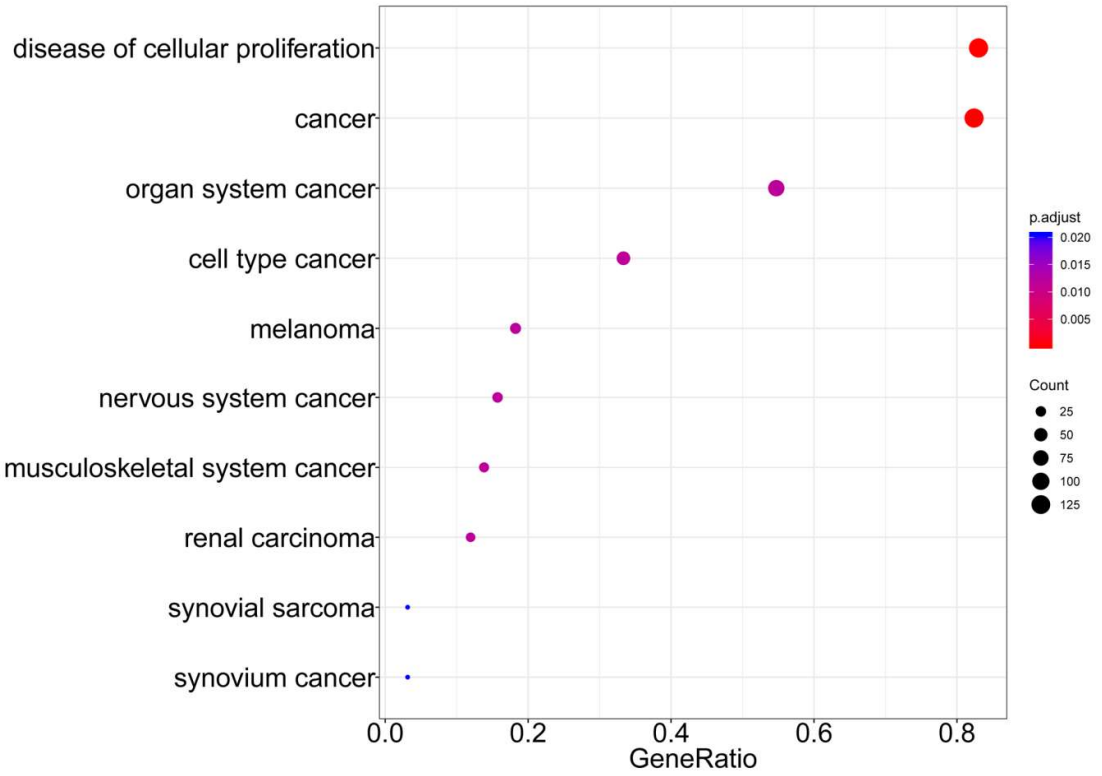


Figure 1.22: Disease associations of genes up-regulated among multiple brain cortical regions in Alzheimer's patients. P-value and gene ratio of diseases (top 10) in which the common up-regulated genes among different cortical regions in Alzheimer's disease patients are also de-regulated were determined based on disease enrichment analysis and are depicted here.

1.1.3.2 Aberrant miRNA-mRNA interaction networks could also be prevalent in other diseases such as cancer

Joint miRNA and mRNA expression profiling datasets that sampled multiple cancer cell types were assimilated and differential expression analysis considering these datasets suggested that multiple miRNA and mRNA are de-regulated in these cancer tissues (Table 1.6, Table 1.7).

Table 1.6: **Multiple miRNAs are differentially expressed in cancerous cell types.** Number of differentially expressed miRNA identified in multiple cancerous tissue types has been tabulated.

	Breast Cancer (GSE40056)	**Colorectal Cancer (GSE35982)**	**Oral squamous cell carcinoma (GSE70664)**
Up-regulated miRNA	46	366	116
Down-regulated miRNA	100	25	72

Table 1.7: **Multiple mRNAs are differentially expressed in cancerous cell types.** Number of differentially expressed mRNA identified in multiple cancerous tissue types has been tabulated.

	Breast Cancer (GSE40057)	**Colorectal Cancer (GSE35982)**	**Oral squamous cell carcinoma (GSE70665)**
Up-regulated mRNA	1362	786	1101
Down-regulated mRNA	1119	1059	3010

Thus, de-regulated miRNA and mRNA expression profiles in different cancer cell types suggested that aberrant miRNA-mRNA interaction networks are prevalent in these tissue types. Subsequently, the aberrant miRNA-mRNA interaction networks prevalent in cancerous cell types were determined based on "regulator(s) to target(s) relationships" (Table 1.8) and these networks were then analyzed with the help of miRNA and mRNA expression profiles in these cancer tissues.

Table 1.8: Probable altered miRNA-mRNA interaction networks in Cancer. De-regulated miRNA and their target mRNA in multiple tissue types from cancer patients were utilised to determine dysregulated miRNA-mRNA interaction networks in cancer cells. Numbers of differentially expressed miRNA and mRNA that comprised the altered miRNA-mRNA interaction networks in different cancerous tissues have been tabulated.

	Breast Cancer		**Colorectal Cancer**		**Oral squamous cell carcinoma**	
	Up-regulated miRNA (34)	**Down-regulated miRNA (73)**	**Up-regulated miRNA (410)**	**Down-regulated miRNA (34)**	**Up-regulated miRNA (83)**	**Down-regulated miRNA (56)**
Up-regulated mRNA	867	1068	630	413	371	319
Down-regulated mRNA	617	784	733	407	1141	1007

Further, analysis of regulator(s) to target(s)" (miRNA:mRNA) expression profiles in different cancer tissues identified that a substantial fraction of up-regulated miRNA had their target mRNA as up-regulated in oral squamous cell carcinoma, colorectal and breast cancer as well (Figure 1.23A, B, C). However, in this scenario, it was found that a similar percentage of down-regulated miRNA and up-regulated miRNA had their target mRNA fractions as up-regulated (Figure 1.23D). Briefly, the up-regulated miRNA (~65%) and down-regulated miRNA (~67%) had similar fraction (between 0.4-0.6 and 0.2-0.4) of their corresponding target mRNA as up-regulated in the cancer case studies (Figure 1.23).

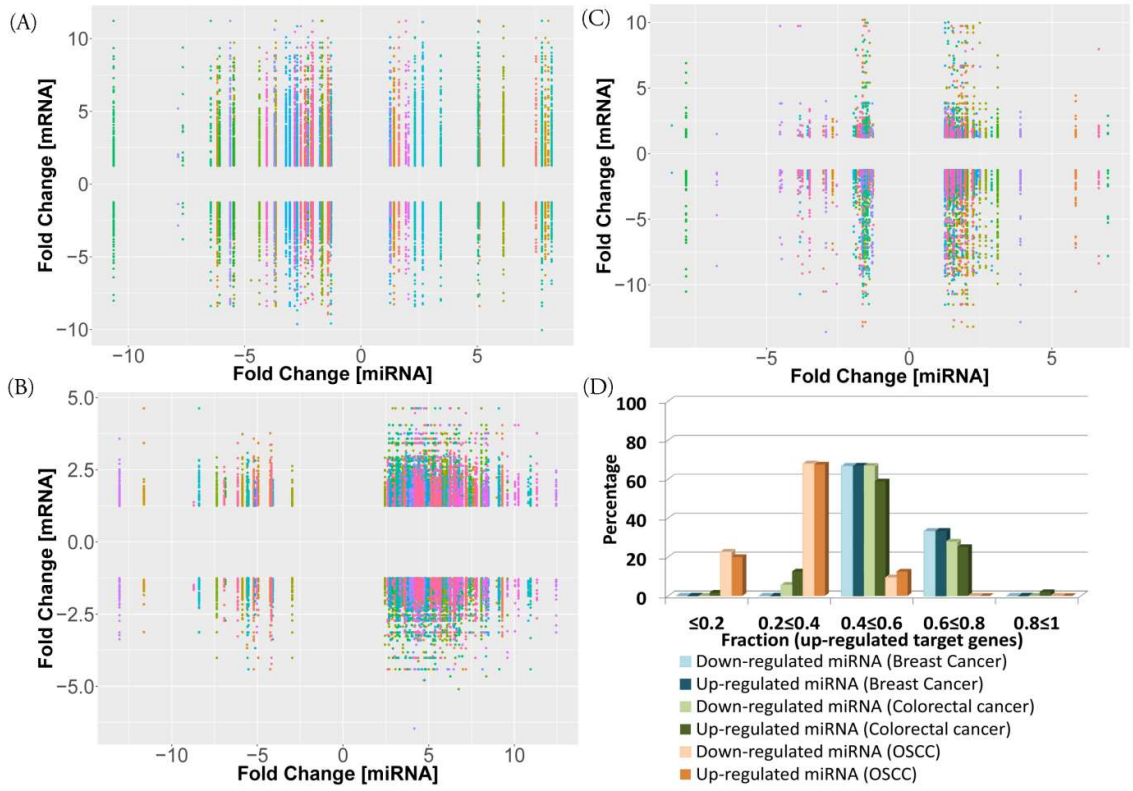


Figure 1.23: Probable miRNA:mRNA expression relationships in cancer patient samples. (A-C)Fold changes in differentially expressed miRNA and their corresponding differentially expressed target mRNA in oral squamous cell carcinoma (A), colorectal (B) and breast cancer (C) are depicted here. (D) Comparison among percentage of differentially expressed miRNA that have different subsets or fractions of their target mRNA as up-regulated in highly proliferative tissues (different cancer datasets).

1.1.3.3 Unusual miRNA-mRNA regulatory relationships could be prevalent in Alzheimer's disease

During the analysis of altered miRNA-mRNA interaction networks in AD, it was observed that a substantial percentage of up-regulated miRNA have a large fraction of

their corresponding target genes as up-regulated. This observation is contrary to the general regulatory relationship prevalent between miRNA and their target genes wherein the target genes of up-regulated miRNA generally get down-regulated. However, this inverse relationship between the target genes and up-regulated miRNA has also been noted in other cell types other than degenerative neurons. In cancer cells as well this trend of target gene upregulation in the backdrop of up-regulated cognate miRNA expression has been noted (Figure 1.24).

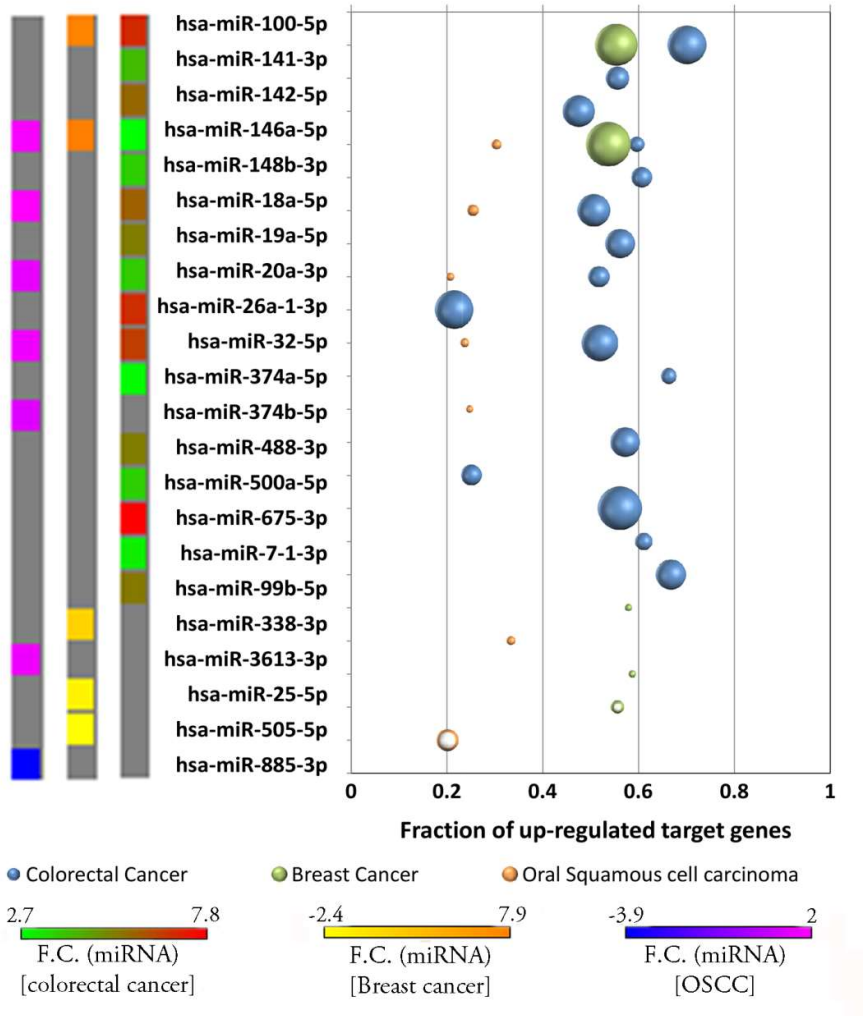


Figure 1.24: Comparison of unusual miRNA:mRNA interactions observed in cancer and Alzheimer's. The probable relationship between differentially expressed miRNA and their target mRNA was studied in different cancer tissues considering the possible fraction of up-regulated target mRNA of the miRNA that were found to be commonly differentially expressed in AD and cancer. The possible fraction of up-regulated target mRNA of each of these differentially expressed miRNA in breast cancer, colorectal cancer and oral squamous cell carcinoma has been depicted here.

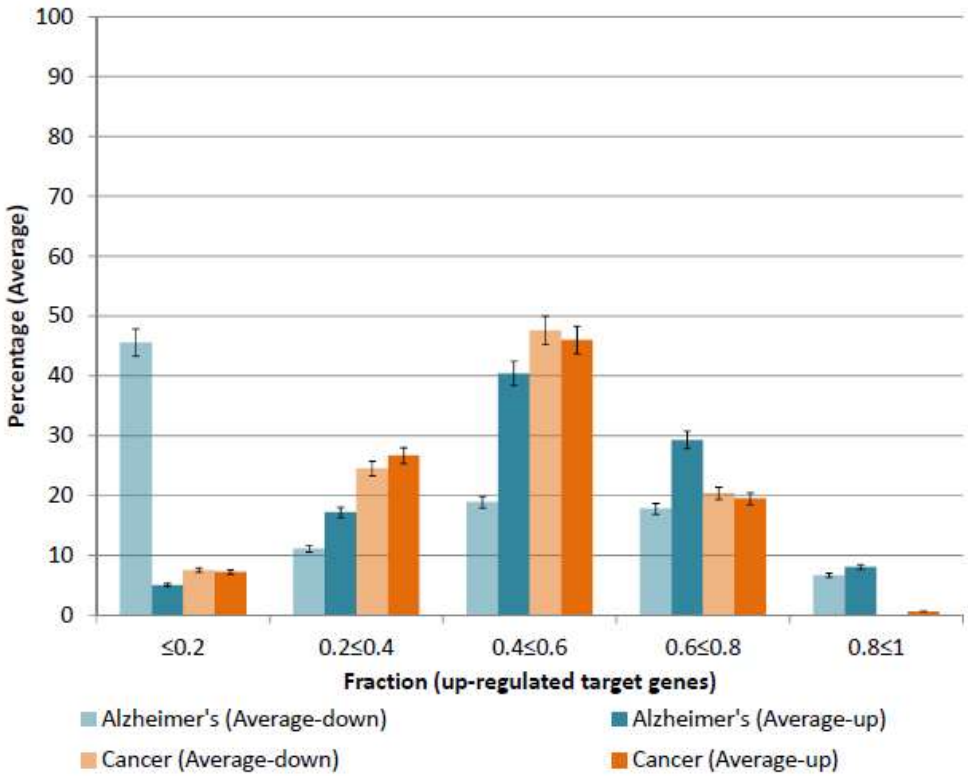


Figure 1.25: Comparison of miRNA-mRNA regulatory relationship in Alzheimer's and Cancer. The average of the percentage of differentially expressed miRNA with different subsets or fractions of their target genes as up-regulated among the three datasets in Alzheimer's and Cancer disease scenarios has been compared here.

However, the upregulation of target gene with increase in cognate miRNAs has been more prominently observed in AD rather than in cancer (Figure 1.25). Therefore, we observed trends that in comparison to the down-regulated miRNA, target mRNA of up-regulated miRNA are mainly up-regulated and this phenomenon is more prominent in Alzheimer's disease condition than in cancer (Figure 1.25).

Additionally, miR-146a-5p has consistently been found to exhibit this phenomenon across all the conditions analyzed herein (Figure 1.21, Figure 1.24). Further, expression of all the probable miR-146a targets was also studied. Interestingly, majority of the miR-146a targets were found to be upregulated in AD brain which further indicated a possible unusual regulatory function of miR-146a on its targets in AD brain (Figure 1.20 and Table S7). Moreover, in a study analyzing the role of miR-146a in neurodevelopmental disorders (Nguyen et al., 2018), miR-146a over-expression in differentiated neural stem cells lead to upregulation of around 40.62% mRNA which is again indicative of the fact that up-regulation of cognate miRNA (miR-146a-5p) can lead to upregulation of target genes. Further, similar observation with respect to miR-146a has been made in our collaborators laboratory considering $A\beta_{1-42}$ treated astrocyte cells.

Conclusion

In the present study, de-regulated miRNA and mRNA within the cortical regions of AD patients which are likely to be associated with AD development or progression were identified. Analysis of aberrant miRNA-mRNA interaction networks in AD yielded an interesting observation that in comparison to the down-regulated miRNA in Alzheimer's disease, a substantial percentage of up-regulated miRNA had a higher fraction of their corresponding target mRNAs as up-regulated. This trend was consistently observed in late stage Alzheimer's disease cases considered from multiple studies. Moreover, the up-regulated miRNA had their target mRNA fractions as up-regulated in patient samples from cancer patients (oral squamous cell carcinoma, colorectal and breast cancer) as well. However, the trend that in comparison to the down-regulated miRNA, target mRNA of up-regulated miRNA are mainly up-regulated is more prominent in Alzheimer's disease condition than in cancer. These observations suggest that in contrary to the general regulatory relationship between miRNA and their target mRNA wherein miRNA mainly down-regulate mRNA, other regulator:target (miRNA:mRNA) regulatory relationships might also exist in tissues under disease conditions. In particular, miR-146a-5p is one such miRNA that exhibited this phenomenon wherein increased production of miR-146a-5p resulted in the up-regulation of multiple target mRNAs.

Inference: Alteration in regulatory relationships between network components may occur in disease conditions wherein unusual miRNA-mRNA regulatory relationships are likely to become prevalent. Disease associated changes in the regulatory relationship between network components such as up-regulation in target mRNA upon up-regulation in miRNA are likely to occur in a cellular phenotype specific manner. Computational analysis of "regulator(s) to target(s)" (miRNA:mRNA) expression profiles in different cortical regions of Alzheimer's disease patients suggested that upregulation of target gene may occur with increase in cognate miRNAs levels. In particular, this altered regulatory relationship between miRNA and their target mRNA is more prominent in degenerative cells (Alzheimer's disease) in than highly proliferative cells (cancer).

1.1.4 Modifications in the network components of an essential regulator or alternate regulatory relationship between network components may result in a varied cellular response

Synopsis: The present section outlines a study wherein the regulatory network of an RNA binding protein has been determined under different cellular conditions to determine the network components that are likely to be regulating a cellular response. It has been observed that immune responses may be regulated by RNA binding proteins such as HuR by means of post-transcriptional regulation of mRNA. Herein, the HuR regulatory network in macrophages has been predicted to study whether the network or its components get modulated during macrophage immune response particularly in the case of an *L. donovani* infection.

Problem Statement: Given that during an infection, a protein which regulates mRNA levels in macrophages is down-regulated and multiple mRNA get differentially expressed, determine the regulatory network of this RNA binding protein. In this context, study the probable regulator-target network in infected and non-infected cells to identify genes involved in regulating disease establishment during an immune response.

Hypothesis: An RNA binding protein might modulate the stability or translation of its targets by associating with them based on its binding specificity. Target mRNA may be predicted by determining the presence of binding site(s) of the protein within the cellular mRNA and these regulator binding sites within the target mRNA may be similar in multiple cell types. However, the expression profile of these mRNA and the regulator-target regulatory relationship may vary during an infection scenario.

System of study: RNA binding proteins (RBP) by means of post-transcriptional regulation of messenger RNA (mRNA) transport/localization, stability or translation may regulate mRNA levels and in turn protein levels. It would be intriguing to study the regulatory network of RNA binding proteins and their mRNA targets in the context of modulation of macrophage immune response since RBP are essential in the control of gene expression. In particular, HuR protein, a member of the ELAV/Hu family of RBP is expressed broadly across tissues and may have a role in immune response by regulating mRNA metabolism. HuR-RNA associations may generally occur at pre-mRNA introns, proximity of 3' splice sites, 3'-untranslated regions (UTRs) and have been identified with the help of ribonucleoprotein (RNP) immunoprecipitation (RIP-chip), photoactivatable ribonucleoside enhanced cross-linking and immunoprecipitation (PAR-CLIP) etc. Herein, based on the observation from our collaborators laboratory that HuR expression goes down in *L. donovani* infected macrophages, the probable HuR targets in the human transcriptome were determined. Therefore, the HuR regulatory network in *L. donovani* infected and non-infected macrophages was studied to determine whether macrophage immune response may be regulated by the RBP HuR (Figure 1.26).

Graphical Summary:

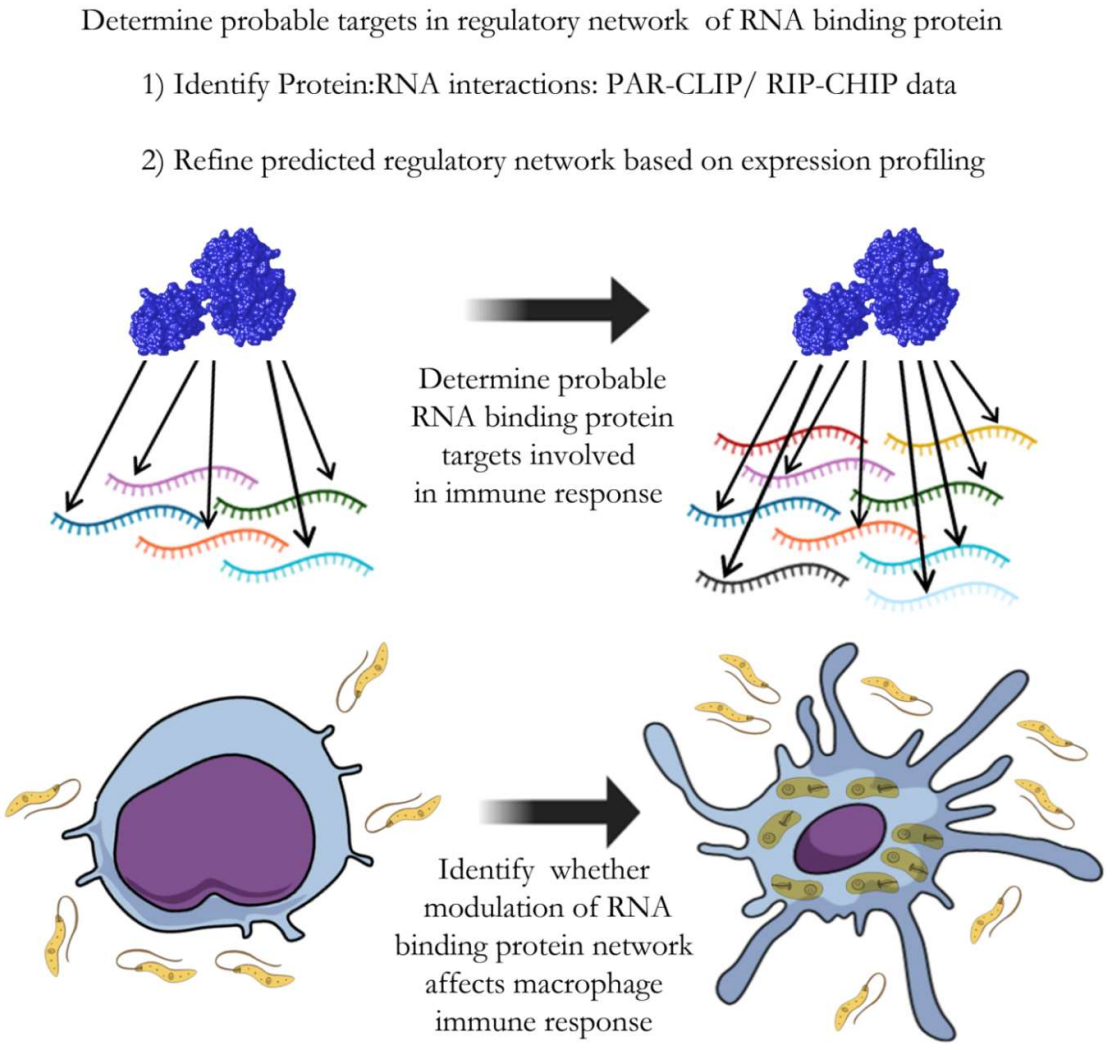


Figure 1.26: **Studying gene expression modulation by the regulatory network of an RNA binding protein under different cellular conditions**

Regulatory network of an RNA binding protein might exhibit condition specific differences in regulatory relationships between network components resulting in varied cellular responses

1.1.4.1 Predicted regulatory network of RNA binding protein HuR

HuR is primarily known to associate with specific upstream/downstream sequences of target mRNAs, for instance, U/AU-rich sequences either at 3'-untranslated regions (UTRs) or within pre-mRNA introns (Srikantan and Gorospe, 2011). Based on an

underlying assumption that HuR binding sites in the target mRNA would be common across multiple cell types with differences in the regulatory network arising as a result of transcript expression and abundance differences among the cell types, the regulatory network of the RNA binding protein has been predicted. Initially, HuR target genes were determined by considering data from previous PAR-CLIP analysis studies that have extensively characterized RNA-HuR interactions (Mukherjee et al., 2011, Lebedeva et al, 2011). Target mRNA (1103) that had binding site information and which scored a log odds ratio (LOD) >= 0, [where LOD is indicative of probability of an mRNA being associated with HuR] were considered. Further, only mRNAs that were found to be differentially expressed either at protein/mRNA level upon HuR knockdown in HEK293 cells were considered as HuR targets. Similarly, another set of target genes (875) identified based on PAR-CLIP analysis that showed differential expression either at mRNA/protein level upon HuR knockdown in HeLa cells was also considered as HuR targets. Thus, 1877 HuR target mRNAs have been predicted to be involved in the regulatory network of the RNA binding protein, HuR.

1.1.4.2 HuR regulatory network in macrophages is altered during *L. donovani* infection

Differential expression analysis considering *L. donovani* infected and non-infected macrophages identified a set of up-regulated (636) and down-regulated (276) genes. Previously, a list of probable HuR target mRNAs in macrophages has been determined and with a list of differentially expressed mRNA in *L. donovani* infected macrophages the set of HuR target mRNAs that are likely to be modulated in *L. donovani* infection scenario may be identified. The expression status of probable HuR target mRNAs in Ld-infected human macrophages suggested that certain HuR targets (56 mRNA) that were found to be up-regulated on HuR knock-down in other cell types were also found to be up-regulated in *L. donovani* infected human macrophages. Similarly, certain HuR targets (16 mRNA) that were found to be down-regulated on HuR knock-down in other cell types were also found to show reduced expression in *L. donovani* infected isolated human macrophages (Table S8). Thus, by determining whether any of these probable HuR targets get differentially expressed when human macrophages are infected with *L. donovani*, it was observed that 165 likely mRNA targets of HuR may have a role in macrophages in *L. donovani* infection scenario (Table S8, Figure 1.27). Moreover, considering RNAseq profile of mRNA from HuR (*Elavl1*) knockout murine bone marrow-derived macrophages (Lu et al., 2014), it was identified that among the previously predicted 165 targets 129 show differences in their expression profile under these conditions (Table S8). This observation, ascertains additional supportive evidence that HuR target mRNA network in macrophages is indeed altered in *L. donovani* infected macrophages. Further, modification of HuR target gene network by *L. donovani* might modulate the immune response of infected macrophages (Table 1.9).

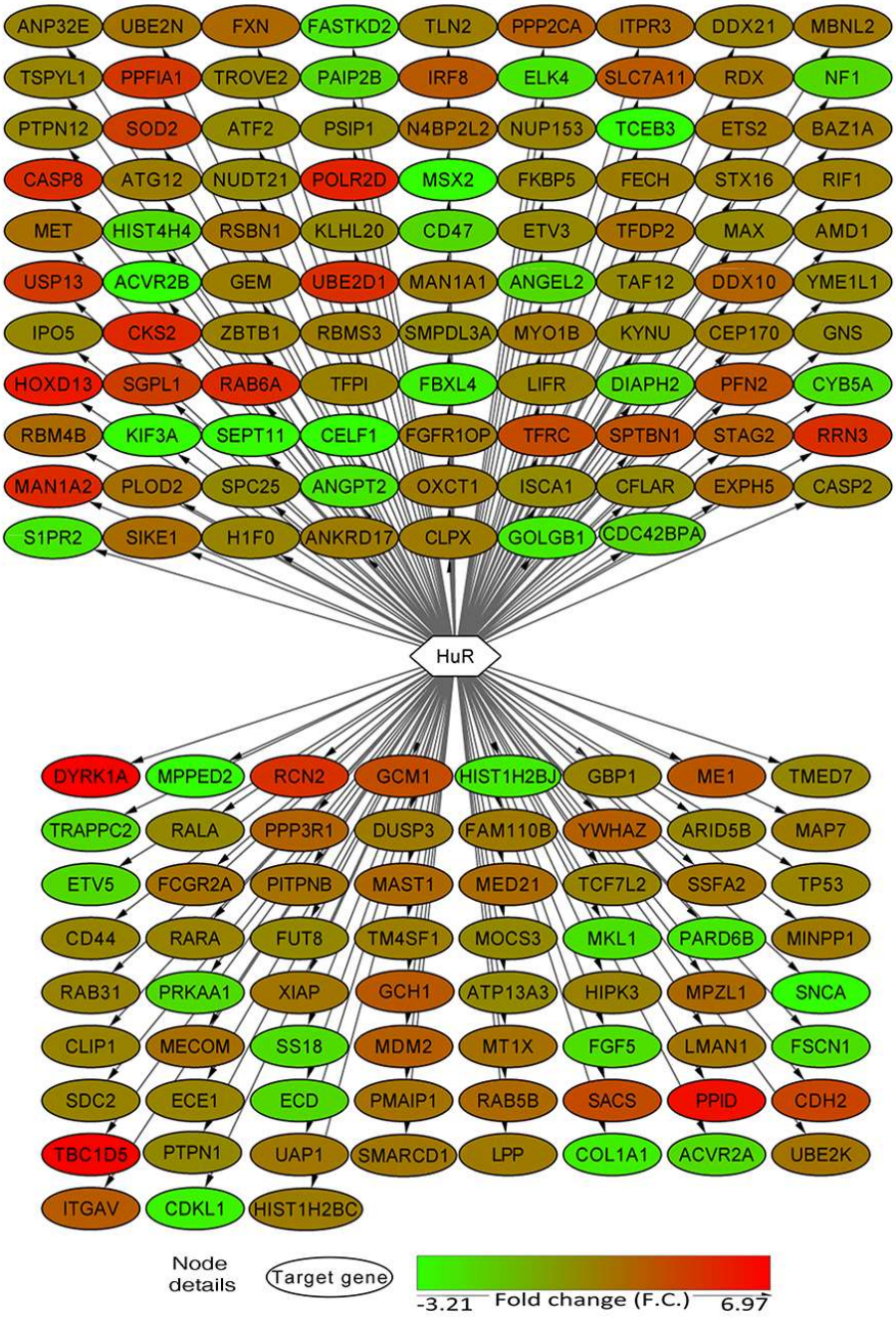


Figure 1.27 **Status of HuR target mRNA in *L. donovani* infected macrophages.** The altered HuR regulatory network in macrophages upon *L. donovani* infection has been depicted here. HuR regulated genes that may be targeted in Ld infected cells based on their differential expression upon *L. donovani* infection in macrophages is outlined here. Note: The node color is indicative of fold change (F.C.) in gene expression of respective HuR target genes in *L. donovani* infected macrophages. [Note: published in Goswami et al., 2020].

Table 1.9: **HuR target mRNA that are likely to get de-regulated during *L. donovani* infection (LDI) infection might alter macrophage immune responses.** A few HuR regulatory network mRNA in macrophages that have involvement in immune signalling pathways like cytokine signalling or anti-inflammatory responses have been mentioned.

	Gene Name	**log Fold Change (LDI macrophage)**	**HuR (*Elavl1*) knockout murine bone marrow-derived macrophages**	**Immune response related pathway Involvement of target mRNA**
HuR target genes that show up-regulation upon HuR knockdown in other cell types and are up-regulated in *L. donovani* infection in macrophages	CD44	1.73	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	CRKL	3.04	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	DUSP3	2.1	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	EIF4E	2.06	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	GBP1	1.76	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	PTPN1	1.53	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	RALA	1.59	Down-regulated	Cytokine Signaling in Immune system (R-HAS-

				1280215)
	TP53	1.76	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	YWHAZ	3.18	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
HuR target genes that show down-regulation upon HuR knockdown in other cell types but are up-regulated in *L. donovani* infection in macrophages	ATF2	1.52	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	CRKL	3.04	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	EIF4E	2.06	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	IRF8	3.44	Up-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	LIFR	1.81	Up-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	SOD2	4.49	Not found	Cytokine Signaling in Immune system (R-HAS-1280215)
	NUP153	1.8	Not found	Cytokine Signaling in Immune system (R-HAS-

				1280215)
	PPP2CA	3.18	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	PTPN12	1.67	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
	ITPR3	2.77	Up-regulated	Anti-inflammatory response favoring *Leishmania* parasite infection(R-HAS-9662851)
	PRKAR1A	1.92	Not found	
	UBE2N	2.39	Down-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)
HuR target genes that show up-regulation upon HuR knockdown in other cell types but are down-regulated in *L. donovani* infection in macrophages	FSCN1	-2.02	Up-regulated	Cytokine Signaling in Immune system (R-HAS-1280215)

1.1.4.3 Alterations in HuR regulatory network in infected macrophages may alter macrophage immune response

Regulatory network of the RNA binding protein, HuR that regulates cellular mRNA levels has been analyzed in terms of its network components and regulatory relationship between network components in order to characterize a possible change in cellular response. A set of HuR target mRNA (80) that were found to be down-regulated in other cell types (HEK293, HeLa) were found to be up-regulated in *L. donovani* infected macrophages. Additionally, another set of HuR target mRNA (13) that were found to be up-regulated in other cell types (HEK293, HeLa) were found to be down-regulated in *L. donovani* infected macrophages (Table S8). This observation suggests

that HuR regulatory network exhibits cell type specific differences in network components. Further, it has been observed that HuR expression is down-regulated during *L. donovani* infection in macrophages (Goswami et al., 2020) and based on this assumption the predicted HuR regulatory network in infected macrophages has been compared with the network in HuR knockout bone marrow derived macrophages. Herein, it was observed that the regulatory network components which are down-regulated in HuR knockout macrophages may be up-regulated in *L. donovani* infected macrophages (Table 1.9). Moreover, HuR regulatory network mRNA (47) that has been found to be consistently up-regulated or down-regulated in multiple cell types are de-regulated in *L. donovani* infected macrophages (Figure 1.28).

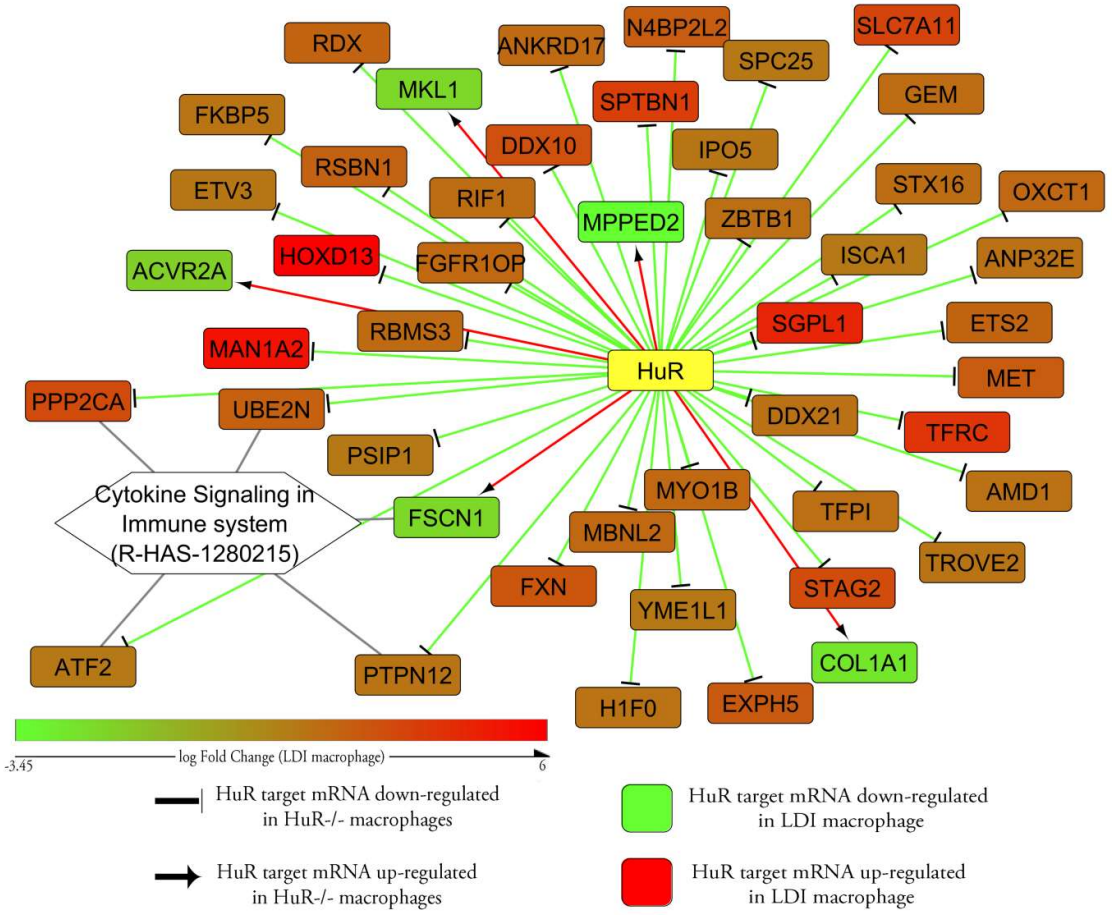


Figure 1.28 **Modifications in HuR regulatory network in *L. donovani* infected macrophages.** HuR regulatory network mRNA that are generally down-regulated in multiple cell types have been found to be up-regulated while HuR target mRNA that are generally up-regulated in multiple cell types have been found to be down-regulated. Note: Edge color denotes the general regulatory relationship observed between HuR and network components in HuR knock out macrophages and other cell types. Node color denotes the expression status of the HuR target in *L. donovani* infected

macrophages.

Thus, HuR regulatory network components may exhibit condition specific differences and this alteration in regulatory relationship between network components may result in modulation of cellular response (Figure 1.28). In particular, some de-regulated mRNAs encode proteins involved in cytokine signaling and in turn macrophage immune response may differ in infected cells promoting favorable conditions for parasite survival. For instance, during the infection process *Leishmania* targets HuR down-regulation in turn reducing its inhibitory effect on PPP2CA (protein phosphates 2A) expression and upregulates the HuR target PPP2CA to establish an anti-inflammatory response in infected macrophages (Goswami et al., 2020).

Conclusion

The regulatory network of the RNA binding protein, HuR, has been predicted in infected and non-infected macrophages with the help of HuR binding site information and expression profiling analysis. The basic assumption was that the binding sites within the target mRNAs would be present in all cell types, however, the network may differ based on transcript specific expression and abundance differences. Herein, it was observed that network components and regulatory relationship between the network components may differ. This is because on HuR knockdown or knockout at mRNA or protein levels multiple target mRNAs were found to be up-regulated and down-regulated. Further, the regulatory relationship with the same target mRNA also differed among the cell types considered. In particular, during *L. donovani* infection in macrophages, wherein HuR expression goes down a significant fraction of HuR regulatory network genes get differentially expressed. Thus, alterations in network components of HuR regulatory network or modulations in regulatory architecture of the network may result in modification of cellular responses such as immune response in the case of macrophages. Thus, the pathogen*, L. donovani* can alter HuR target mRNA in macrophages during an infection and this modulation can in turn promote pathogen survival in macrophages. In particular, HuR target PPP2CA is up-regulated upon HuR down-regulation and this axis promotes an anti-inflammatory response in macrophages.

Inference: The HuR regulatory network in macrophages was predicted and expression profile of multiple HuR target mRNA was found to be significantly altered during *L. donovani* infection in macrophages. It is likely that *L. donovani* modulates macrophage activation or immune response by de-regulating the regulatory network of the RNA binding protein HuR.

1.2 Methodology

1.2.1 Analyzing intra-cellular protein-protein interaction networks to determine effector proteins bridging the regulatory and signaling networks

1.2.1.1 Genes associated with ciliary biology based on expression analysis

Generally, genes involved in ciliary biology may be identified with the help of gene mis-expression, targeted gene knock-down or knock-out studies in experimental model systems (Stubbs et al., 2008; X. Yu et al., 2008; Choksi, Babu, et al., 2014; May-Simera et al., 2016; Terré et al., 2016). Similarly, with the help of functional genomics studies, it has been identified that FoxJ1 protein is crucial for basal body docking, ciliary axoneme assembly and ciliary motility (Stubbs et al., 2008; X. Yu et al., 2008). Moreover, FoxJ1 regulates a set of genes known as FoxJ1 induced genes (FIG), which together are sufficient for motile cilia development and function (Choksi, Babu, et al., 2014). In this manner, FoxJ1 has acquired the title of master regulator of motile ciliogenesis since its over-expression is sufficient to drive the motile ciliogenic program and generate functional ectopic motile cilia (Stubbs et al., 2008; X. Yu et al., 2008; Choksi, Babu, et al., 2014). Genes which are likely to have possible associations with ciliary biology based on their differential expression in studies probing ciliogenesis or ciliary function in experimental model systems have been reported in different ciliary databases or datasets. Data from Choksi et al. expression study (Choksi, Babu, et al., 2014), PCD expression analysis case study (Geremek et al., 2014), CilDB (Arnaiz et al., 2009; Arnaiz, Cohen, Tassin, & Koll, 2014) and Human Protein Atlas (Uhlen et al., 2010; Uhlén et al., 2015; Thul et al., 2017) providing evidence regarding RNA or protein expression abundance in ciliary cells have been taken into consideration for 'cilia associated expression analysis'. Subsequently, genes that had expression information in the 'cilia associated expression analyses were considered to have possible associations with ciliary biology.

Moreover, changes in multiple genes or proteins might result in defects in motile cilia formation or function resulting in a syndrome known as primary ciliary dyskinesia (PCD). Therefore, disease (PCD) associated genes determined based on their differential expression in ciliated tissues of PCD patients may also have possible associations with cilia biogenesis, function or motility. For this purpose, a previous study that had determined the expression profile of bronchial tissue of PCD patients (Geremek et al., 2014). Considering the data available in Gene Expression Omnibus (GEO) series dataset (GSE25186) (Edgar, Domrachev, & Lash, 2002; Barrett et al., 2012; Geremek et al., 2014) differentially expressed genes were determined with the help of limma (Ritchie et al., 2015) package in R (R Core Team, 2016). Differentially expressed genes that had fold change >=2 and p-value<=0.05 were considered as PCD associated and are likely to be involved in ciliary biology or function as well.

1.2.1.2 Constructing and understanding the FOXJ1 regulatory network for ciliogenesis in humans

Target genes of a regulator such as a transcription factor may be predicted by determining the presence of transcription factor binding site(s) in the vicinity of transcription start site (TSS) of genes. This may generally be accomplished by scanning genomic DNA with the help of a pattern matching algorithm for the presence of binding motifs represented in the position weight matrix (PWM) describing the general binding specificity of a transcription factor (Bulyk, 2003). In addition to the pattern matching algorithm and a PWM for the transcription factor, other pre-requisites for predicting target genes of a transcription factor include, a background matrix representative of general base frequencies around the TSS of genes and sequences (upstream and downstream around TSS) of genes to be scanned. In this regard, PWM for human FOXJ1 was unavailable, however, a PWM for mouse FoxJ1 [PB0016.1] was collected from footprintDB database (Sebastian & Contreras-Moreira, 2014) since it has been suggested that orthologous transcription factors may share similar binding specificities (Jolma et al., 2013). It was ascertained that FOXJ1 (human) and FoxJ1 (mouse) proteins are 92.6% identical while their DNA binding domains are 100% identical and as such it is likely that these orthologous proteins share similar binding specificities (Figure 1.29A). Further, genomic DNA sequences for a set of genes previously reported as inducible or over-expressed upon FoxJ1 over-expression in the zebra fish model system referred to as FoxJ1 induced genes (FIG) were collected for scanning with the help of Rsat (Turatsinze, Thomas-Chollier, Defrance, & Helden, 2008). Moreover, utilizing the Ensembl Compara database (Herrero et al., 2016) it was determined whether the FIG have high confidence orthologs in *Homo sapiens* and *Mus musculus*. Subsequently, with the help of Rsat ±6kb of the FIG were scanned for the presence of FOXJ1 binding motif(s) utilizing the FoxJ1 PWM, a background model (Markov order) representative of ±6kb of random *Homo sapiens* genes. Genes for which binding sites could be predicted in the vicinity of TSS with a p-value <= $1e^{-04}$ are likely to be transcriptionally regulated by FOXJ1 and have been predicted as FOXJ1 target genes (Figure 1.29B) (Turatsinze et al., 2008; Medina-Rivera et al., 2015). In this manner, probable FoxJ1 target genes that exhibit differential expression upon perturbations in FoxJ1 expression levels and which have probable cis-regulatory FOXJ1 binding sites have been termed as directly regulated genes while probable FoxJ1 target genes that only exhibit differential expression upon alterations in FoxJ1 expression levels have been coined as indirectly regulated FOXJ1 genes. Therefore, a FOXJ1 regulatory network for ciliogenesis in humans which is comprised of FOXJ1, its direct and indirect target genes has been determined.

1.2.1.3 Determining the functions of FOXJ1 regulatory network genes in the context of ciliogenesis

A number of ciliary reference databases and studies like the SysCilia gold standard database (Dam et al., 2013), Reactome pathway database [R-HSA-5617833] (Fabregat et al., 2017; Croft et al., 2013), cilliary proteome related studies (Boldt et al., 2016; Gupta et

al., 2015), FoxJ1 induced genes study (Choksi, Babu, et al., 2014) and OMIM database (McKusick, 2007; Amberger et al., 2014) were taken into consideration for this part of the analysis. Utilizing these resources a 'collated ciliary resource (CCR)' detailing information regarding genes experimentally probed and identified to be involved in ciliogenesis or ciliary function was prepared. This resource prepared from multiple literature resources then served as a reference to annotate the FoxJ1 directly and indirectly regulated network genes and outline their cilia specific functions.

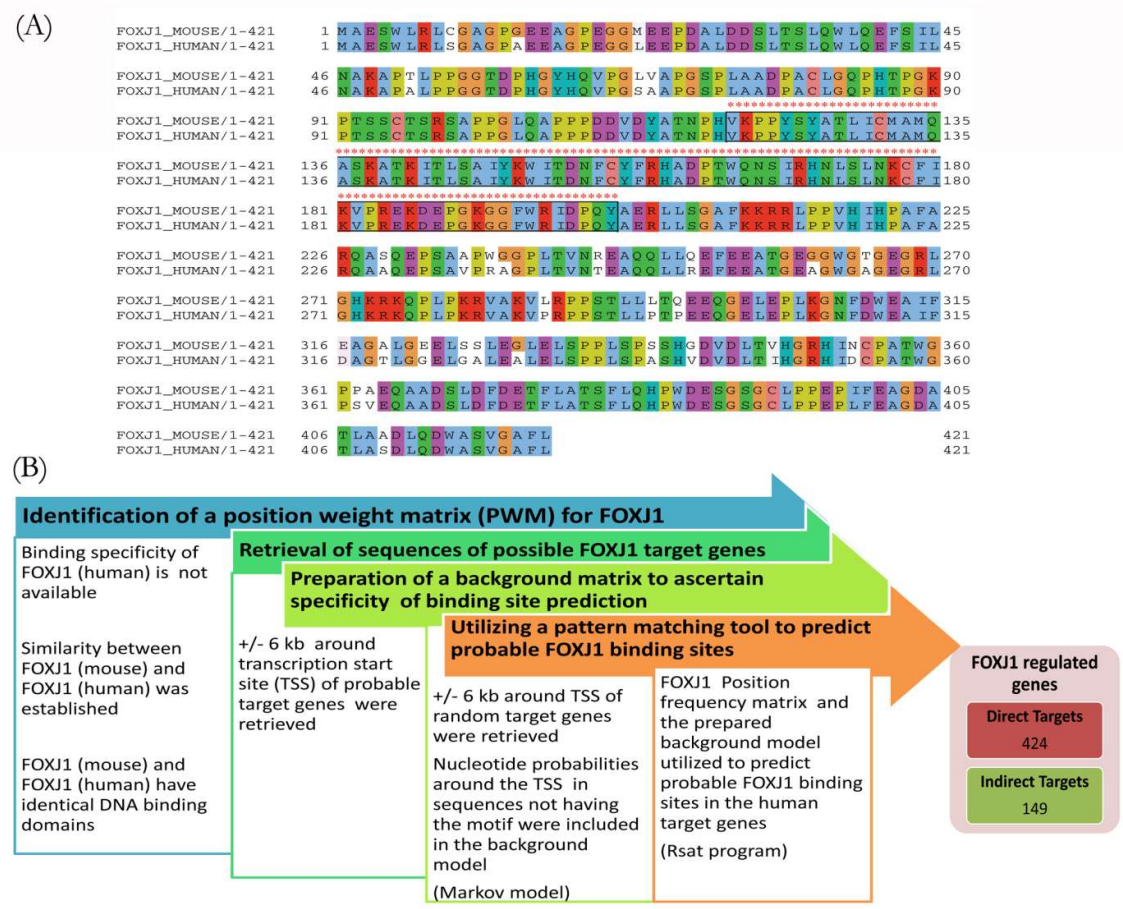


Figure 1.29: **Predicting the FOXJ1 regulatory network for ciliogenesis.** (A) The DNA binding domain of FOXJ1 from *Homo sapiens* and *Mus musculus* is evolutionarily conserved (denoted by *) as depicted here in the pair wise sequence alignment between them. (B) Methodology utilized to predict FOXJ1 directly regulated genes is outlined here.

In addition to data mining of experimental data available in literature, other computational methods such as GO analysis and GO enrichment analysis were utilized to assign probable function(s) to the FOXJ1 regulatory network genes. GO analysis using the DAVID web server (Huang, Sherman, & Lempicki, 2008, 2009) and GO enrichment analysis with the help of FGNet (Aibar, Fontanillo, Droste, & De Las Rivas, 2015) allowed identification of certain GO based enriched clusters among the FOXJ1 regulatory network

genes. Based on this analysis, genes belonging to clusters having p-value < = $1e^{-02}$, cluster enrichment score >= 2, fold enrichment >= 4 could be associated with predicted functions.

1.2.1.4 Determining and analyzing the FOXJ1 regulatory network associated protein-protein interaction network (PPIN)

Constructing the regulatory network associated protein-protein interaction network (PPIN)
In order to identify proteins connecting the regulatory network and the signaling network within the intra-cellular milieu during motile cilia biogenesis, the probable PPIN associated with the FOXJ1 regulatory network representative of proteins and connections likely to be important for cilia structure or function in relation to FOXJ1 activation was determined considering the FIG encoded protein (FIGp) as seed proteins. A PPIN comprised of proteins as nodes and interactions as edges was constructed around these seed proteins by obtaining high confidence experimentally reported interactions among the FIGp and other cellular proteins from different databases such as SysCilia, Bioplex, STRING and BioGrid (Dam et al., 2013; Huttlin et al., 2015; Szklarczyk et al., 2014; Stark et al., 2006; Chatr-aryamontri et al., 2016). Once a network of primary interactors of FIGp was obtained the largest connected component of this network hereafter referred to as FIG-sub-network was extracted. Scale free biological networks wherein the degree distribution follows a power law conform to certain graph theory based topological properties (Barabási & Oltvai, 2004) and in order to ascertain whether the FIG-sub-network also exhibits these characteristics, the degree distribution of this network was analyzed. The degree (k: number of proteins each protein is connected to) of each protein in the network was calculated and a power law [$P(k) \sim k^{\alpha}$ where α is the degree exponent] was fitted to the resulting degree distribution. Further, Kolmogorov–Smirnov test was utilized to determine the p-value for goodness of fit of the degree distributions to the power law (at 0.1 level of significance) (Clauset et al., 2009; R Core Team, 2016).

A number of reports based on computational analyses have suggested that hub, bottleneck or central proteins within a scale free biological PPIN could be crucial for network integrity or function and these proteins have indeed been identified to be essential for cellular function based on subsequent gene deletion studies (Barabási & Oltvai, 2004; H. Yu et al., 2007; Pavlopoulos et al., 2011; Jeong et al., 2001). Therefore, network topology analysis of the representative motile cilia interactome (FIG-sub-network) utilizing graph theory based methods and a computationally faster implementation (in parallel perl) of a previously proposed methodology (Bhattacharyya & Chakrabarti, 2015) suitable for analyzing human interactome data was performed. During this analysis, different graph theory metrics like degree, shortest path and centrality have been taken into consideration to determine important interacting proteins (IIP) (a combination of local network perturbing, global network perturbing, hub, bottleneck and central proteins) in the FIG-sub-network as outlined below.

Node perturbation analysis of the FIG-sub-network

Based on the observation that a PPIN is resilient towards the removal of random nodes while removal of hub proteins had a significant effect on the topology of a PPIN, a centrality based metric was derived that studies the effect of in-silico node removal on the network topology. Subsequently, a node perturbation analysis considering the global FIG-sub-network and local sub-graphs (comprised of proteins and their 2nd level interactors) of the FIG-sub-network was performed to determine global network perturbing proteins (GNPP) and local network perturbing proteins (LNPP) respectively (Bhattacharyya & Chakrabarti, 2015). LNPP were determined by comparing the local network centrality score (LNCS) (equation 1.1) of the nodes before and after node removal in the local sub-graphs whereas GNPP were determined by removing a single node at a time from the global network and studying the global network centrality score (GNCS) (equation 1.4) before and after the global network perturbation. It was assumed that higher the difference in the scores (LNCS/GNCS), higher is the perturbation ability, and thus, proteins important for maintaining the integrity of the FIG-sub-network locally or globally were determined in this manner.

$$LNCS = 1/N \sum_{1}^{N} CCS \quad (1.1)$$

Equation 1.1 where LNCS is the local network centrality score, N is the number of nodes in the local sub graph and CCS (equation 1.2) is the cumulative centrality score.

$$CCS = \sum_{1}^{n} CS \quad (1.2)$$

$$CS = \sum C(betweeness) + C(closeness) + C(clustering\ coefficient) \quad (1.3)$$

Equation 1.2 where n is the number of first degree interactors, CS (equation1.3) is the combined score and CCS is the cumulative centrality score

$$GNCS = 1/N \sum_{1}^{N} CCS \quad (1.4)$$

Equation 1.4, where GNCS is the global network centrality score, N is the number of nodes in the global network and CCS (equation 1.2) is the cumulative centrality score.

The LNCS scores were normalized into z-score and nodes having z-score>=1 were considered as LNPP. The GNCS scores were normalized into z-score and nodes having z-score>=0.5 were considered as GNPP.

Identifying hub, bottleneck and central proteins

Graph theory calculations that measure inherent properties of scale free networks can also be utilized to determine hub or bottleneck proteins which may also be essential for cellular processes (Barab´asi & Oltvai, 2004; H. Yu et al., 2007). Hubs by definition are protein nodes that have much higher degree (number of connections that each node has in the network) than the average degree in the network (Albert, 2005). In order to calculate hubs the node degrees were normalized into z-scores and the fraction of degree population that had z-score>=2 were considered to have significantly higher degree than

the rest of the population. Thus, protein nodes having degree 57 or higher were considered as hub proteins. Bottleneck proteins that can act as key connector proteins are those proteins which have a high betweenness centrality value as a result of multiple "shortest paths" passing through them (H. Yu et al., 2007). Proteins that had betweenness centrality indices higher than two standard deviations from the mean of the betweenness centrality distribution were termed as bottleneck proteins. Centrality may be utilized to assess the compactness of a network and determine central proteins that play a role in relaying information within the network (Pavlopoulos et al., 2011). Moreover, central proteins could also be essential for a cellular process since removal of central proteins by gene deletion leads to lethal phenotypic consequences (Jeong et al., 2001).Thus, by utilizing a range of centrality indices like load, Eigen, closeness centrality and clustering coefficient an average centrality parameter (combined score, CS [equation 1.5]) was determined for each node to identify central proteins within the FIG-sub-network.

$$CS = \sum C(load) + C(closeness) + C(eigen\ vector) + C(clustering\ coefficient) \quad (1.5)$$

$$CCS = \sum_{1}^{n} CS \quad (1.6)$$

Equation 1.6, where n is the number of first degree interactors, CS is the combined score and CCS (equation 1.6) is the cumulative centrality score. The CCS scores (equation 1.6) were normalized into z-score and nodes having z-score>=2 were considered as central proteins.

IIP in FIG-sub-network

Topologically important proteins identified in two or more categories as hub, central, bottleneck, local network perturbing or global network perturbing proteins which are likely to be essential proteins in FIG-sub-network have been termed as IIP. However, only proteins that had additional expression information support in ciliated cells either at the mRNA or protein level according to the previous analysis (genes associated with ciliary biology based on expression analysis) were retained as IIP. Additionally, IIP together with FIGp are likely to be involved in maintaining functional motile cilia or function. In this regard, one may utilize the concept of 'guilt by association' wherein it's suggested that proteins known to interact with one another are likely to participate in same or similar cellular functions (Oliver, 2000; Schwikowski et al., 2000). Thus, in order to determine the probable functional role(s) of these IIP in association with FIGp, a pathway over-representation analysis was performed considering interacting FIGp, IIP and their primary interactors to estimate the enriched pathways or likely roles of these proteins in this context (G. Yu & He, 2016).

1.2.1.5 Ascertaining the relevance of the predicted IIP within the ciliary interactome to ciliary biology

Literature based evidences regarding the involvements of the IIP in ciliary biology or gene expression based association of the IIP with PCD might provide an idea about whether IIP identified in this manner could indeed be essential for cilia biogenesis or function. Additionally, this analysis ascertains the significance of utilizing the devised

intra-cellular protein-protein interaction network analysis approach to predict essential proteins for a cellular process. In this respect, to determine the significance of the finding that some of the IIP were found to be differentially expressed in PCD patients a randomization analysis was performed. Herein, in each trial, 121 proteins were randomly selected from the complete set of proteins in FIG-sub-network and compared with the set of FIG-sub-network proteins that get differentially expressed in PCD. In order to determine whether the association between IIP and PCD associated genes is significant a z-test was performed considering the number of matches that were obtained in a 1000 trials. Additionally, a similar randomization analysis was performed considering matches between the CCR and IIP to determine the significance of the observation that some of the IIP have previously reported ciliary roles in literature.

1.2.1.6 Determining important effector proteins in the FOXJ1 transcriptional network and its associated PPIN

Topological analysis of the FOXJ1 regulatory network associated PPIN may be performed to predict essential proteins within the interactome and alterations in an essential regulatory network component could be disease associated. In this respect, IIP that could be associated with ciliogenesis or PCD based on their differential expression status upon ectopic FoxJ1 expression in zebra fish to model cilia biogenesis or in PCD have a higher likelihood of being essential proteins for ciliogenesis. Therefore, IIP found to be differential expressed upon FoxJ1 over-expression (Choksi, Babu, et al., 2014) have been classified as important interacting protein effector (IIP-effector) in the FOXJ1 regulatory network. Additionally, IIP that are associated with PCD based on differential expression of their corresponding genes in PCD patients (Edgar et al., 2002; Geremek et al., 2014) have also been considered as IIP-effector in the FOXJ1 regulatory network associated PPIN. Such genes associated with motile cilia biology or function based on expression analysis figure at the interface of the FOXJ1 regulatory network and the related PPIN. Moreover, such topologically important effector proteins participate in a number of cellular pathways to bring about a particular function. Subsequently, to determine the pathways in which the IIP-effector proteins are likely to be involved in, pathway enrichment analysis considering a p-value threshold of $1e^{-06}$ was performed in ReactomePA (G. Yu & He, 2016) by taking the IIP effectors and their primary interactors into consideration. Herein, proteins participating in enriched pathways in the 'cilia associated signaling pathways' category were taken into consideration. Pathways which have been experimentally identified to be involved in ciliary biology or function have been considered in the 'cilia associated signaling pathways' category. These pathways include cell cycle (Quarmby & Parker, 2005; Izawa, Goto, Kasahara, & Inagaki, 2015), TGF-beta (Clement et al., 2013), Hedgehog, PDGF, WNT (Goetz & Anderson, 2010), FGF (Neugebauer et al., 2009), RHO GTPase (Kim et al., 2015), TLR signaling (Baek et al., 2017) and vesicle mediated transport (Nachury et al., 2010). In addition, a complementary GO analysis was performed utilizing DAVID (Huang et al., 2009) for GO mapping wherein proteins involved in GO categories likely to be associated with cilia biology were predicted to have possible ciliary association. GO categories such as cilium morphogenesis, cell cycle (Quarmby & Parker, 2005; Izawa et al., 2015), protein

ubiquitination, centrosome cycle, protein folding (heat shock proteins), actin organization (cytoskeleton organization, actin filament organization) and establishment or maintenance of cell polarity were considered for this analysis (Stephens & Lemieux, 1999; Pan, You, Huang, & Brody, 2007; Nachury, Seeley, & Jin, 2010; Bettencourt-Dias, Hildebrandt, Pellman, Woods, & Godinho, 2011; Jones et al., 2012; May-Simera et al., 2016; Shearer & Saunders, 2016; Prodromou et al., 2012; Kasahara et al., 2014; Kohli et al., 2017).

1.2.2 Determining alterations in a regulatory network (miRNA-mRNA) given intra-cellular levels of an essential regulator (miRNA) is perturbed

1.2.2.1 Predicting the regulatory network of miR-122 in hepatic tissues

MiR-122 target mRNAs are likely to exhibit changes in their expression profile in case miR-122 is over-expressed or down-regulated. In order to determine a probable set of miR-122 target mRNA a previous study (Wen et al., 2012) wherein livers of miR122a$^{-/-}$ mice had been compared with wild-type mice has been taken into consideration. Differentially expressed mRNA which had fold change in expression higher than 1.5 or lower than -1.5 with significant p-values have been considered for further analysis. This set of mRNAs has been considered as probable miR-122 targets (Table 1.10) in hepatocytes in the absence of confirmatory binding site or RT-PCR data.

Table 1.10: List of miR122 target mRNA in hepatic cells (mice)

	miR-122 Target Genes	
	Up-regulated mRNA	**Down-regulated mRNA**
Differential expression analysis considering liver samples of miR122a$^{-/-}$ mice v/s wild-type mice (age-2 months)	606	280

1.2.2.2 Identifying genes involved in regulating lipotoxic species associated stress

Differential expression analysis was performed considering high fat diet fed mouse versus chow diet fed mouse for 4 weeks, 12 weeks and mice having high fat along with cholesterol diet as opposed to only high fat diet for 16 weeks. Samples from GSE53381 [high fat diet treatment for 4 weeks (HFD4)] and GSE93819 [high fat diet treatment for 12 weeks (HFD12)] (Kobori et al., 2017;) datasets were considered to identify genes that are involved in regulating short term or long term high fat diet exposure associated stress. Additionally, samples from GSE58271 (Lorbek et al., 2015) were considered to

identify genes involved in responding to high cholesterol (CH) exposure associated stress. For this purpose, differentially expressed genes were determined with the help of limma (Ritchie et al., 2015) R package in Gene Expression Omnibus (GEO) series datasets (Edgar et al., 2002; Barrett et al., 2012). Genes having fold change >= 1.5 and p-value <= 0.05 were considered as differentially expressed and are likely to be involved in regulating lipotoxic species associated stress.

1.2.2.3 Studying whether miR-122 target gene network exhibits alterations in response to high fat diet exposure related stress

In order to get overviews regarding the changes in the mRNA levels of miR-122 target genes in hepatic tissues under conditions of high fat exposure, the expression profile of the predicted miR-122 target genes were studied under different conditions of high fat exposure in mice models. Differentially expressed genes identified upon high fat diet treatment for 4 weeks (HFD4) and high fat diet treatment for 12 weeks (HFD12) were compared with the probable miR-122 targets. Similarly, de-regulated genes determined on high cholesterol (CH) exposure were also compared. Utilising this information miR-122 target gene networks that get de-regulated under multiple conditions of high fat exposure were determined. Venn diagrams were utilised to determine network components in miR-122 target gene network that showed altered levels under two or more of these conditions. Further, a heatmap was prepared to determine the miR-122 target genes that exhibited time dependent changes in expression profile.

1.2.2.4 Determining cellular pathways that may be altered as a result of changes in miR-122 target gene network

Pathway mapping of miR-122 target genes differentially expressed in two or more of the high fat diet treatment conditions was performed. The genes were mapped onto KEGG database (Kanehisa et al., 2000; Cokelaer et al., 2013) pathways with the help of a python script. Cellular pathways belonging to the categories such as ‘carbohydrate metabolism’, ‘lipid metabolism’, ‘cellular process’ and ‘signal transduction and signalling molecules and interaction were considered for the analysis’. A plot was prepared to determine whether the high fat diet exposure associated miR-122 network genes participated in multiple signalling or metabolic pathways and as such whether alterations in miR-122 network components may bring about time dependent physiological changes. Network components in miR-122 regulatory network that exhibited high fat diet exposure associated alterations were represented in rows with their corresponding pathway involvement(s) being shown in columns and each cell depicted the conditions under which they exhibit alterations.

1.2.3 Identifying whether alternate regulatory relationships between network components (miRNA-mRNA) may be prevalent in different cells

1.2.3.1 Determining and analyzing altered miRNA-mRNA interaction network in Alzheimer's disease

Differential expression analysis was performed to determine de-regulated miRNA within the brain cortex of Alzheimer's patients. Utilizing raw data provided in a GEO series dataset and with the help of R packages (R Core Team., 2017; Anders and Huber, 2010) differentially expressed (fold change: +/-1.25 and p-value <= 0.05) miRNA were determined within the pre-frontal cortex of Alzheimer's disease patients (GSE48552; Lau et al., 2013). Further, miRNA-target regulatory information from TarBase (Vlachos et al., 2014) and miRTarBase (Chou et al., 2015) databases was gathered to identify "regulator(s) to target(s)" (miRNA:mRNA) interaction maps for the differentially expressed miRNA. Subsequently, Further, raw data from GEO series datasets GSE5281, GSE53697, GSE15222 (Liang et al., 2007; Liang et al., 2008; Readhead et al., 2018; Scheckel et al., 2016; Webster et al., 2009) were analyzed in R (R Core Team., 2017; Ritchie et al., 2015) to determine the differentially expressed (fold change: +/-1.25 and p-value <= 0.05) mRNA in different regions of the frontal cortex in Alzheimer's disease patients. The differentially expressed miRNA and mRNA identified in this manner considering patient samples were utilized to ascertain "regulator(s) to target(s)" interactions that are likely to be altered in frontal cortex of Alzheimer's patients. In this manner, miRNA:mRNA interaction networks that are probably altered in different regions of the frontal cortex in Alzheimer's were identified and such altered interaction networks could be associated with the development or progression of Alzheimer's disease. These aberrant interaction networks have been studied by utilizing expression profiles of de-regulated miRNA and their target mRNA to determine the regulatory relationships possible between miRNA and their corresponding target mRNA. In this respect, the fractions of target mRNA of each differentially expressed miRNA that are significantly up-regulated or down-regulated has been determined for comparison. Subsequently, to ascertain whether such "regulator(s) to target(s)" patterns could be occurring in other disease conditions, we have performed a disease enrichment analysis considering only the genes that were commonly up-regulated among the Alzheimer's disease datasets considered (Yu et al., 2014).

1.2.3.2 Determining and analyzing altered miRNA-mRNA interaction network in Cancer

In order to study "regulator(s) to target(s)" patterns in cancer, initially miRNA and mRNA expression profiling data from different cancer tissue samples has been analyzed. With the help of large scale expression profiling analysis in R (R Core Team., 2017; Anders and Huber, 2010; Ritchie et al., 2015), differentially expressed (fold change: +/-1.25 and p-value <= 0.05) miRNA and mRNA were determined in cancer. For this purpose, relevant raw data provided in GEO datasets [GSE40056/GSE40057, GSE35982,

GSE70664/GSE70665] corresponding to miRNA and mRNA expression profiling in breast cancer, colorectal cancer and oral squamous cell carcinoma was considered (Luo et al., 2013; Fu et al., 2012; Shi et al., 2015; R Core Team.,2017; Ritchie et al., 2015). Subsequently, the aberrant miRNA:mRNA interaction networks in these cancer sub types have been determined by considering the expression profiles of these de-regulated miRNA, mRNA and regulator(miRNA)-target interaction information from TarBase (Vlachos et al., 2014) and miRTarBase (Chou et al., 2015) databases. Moreover, "regulator(s) to target(s)" interaction patterns have been analyzed by calculating the corresponding fractions of significantly up-regulated or down-regulated target mRNAs of each of the differentially expressed miRNA under each condition.

1.2.3.3 Comparing "regulator(s)-target(s)" interaction patterns prevalent in degenerative cell types and highly proliferative cell types

The miRNA that are commonly differentially expressed in Alzheimer's disease and cancer were identified. The fractions of significantly up-regulated or down-regulated target mRNAs of each of these differentially expressed miRNA were compared. Up-regulated miRNAs that had substantial fractions of their target mRNAs as up-regulated were determined. Based this observation, it was postulated that unusual regulatory relationships could be occurring between miRNA and their target mRNA in diseased cell types. Additionally, experimental evidence to support this concept was identified for miR-146a by analyzing expression profiles of mRNA upon mR-146a over-expression in differentiated neural stem cells (GSE100670; Nguyen et al., 2018).

1.2.4 Studying whether alternate regulatory relationships between network components (protein-mRNA) may alter cellular response

1.2.4.1 Determining the regulatory network of the RNA binding protein (HuR)

HuR target genes may be identified with the help of PAR-CLIP or RIP-CHIP analysis and since HuR-mRNA interaction have been extensively characterized in previous studies (Lebedeva et al, 2011; Mukherjee et al, 2011) either in HeLa cells or HEK293 cells, these data have been utilized. Target genes that had binding site information, scored a log odds ratio (LOD) >= 0, [where LOD is indicative of probability of an mRNA being associated with HuR], were differentially expressed either at protein/mRNA level upon HuR knockdown in HEK293 cells were compiled from Mukherjee et al., 2011. Similarly, another set of target genes identified based on PAR-CLIP analysis that showed differential expression either at mRNA/protein level upon HuR knockdown in HeLa cells was obtained from Lebedeva et al., 2011. Since, HuR binding sites in the target genes is likely to be present in the macrophage DNA/RNA as well, the assumption that HuR may have similar target genes in macrophages may be considered to predict the HuR regulatory network in macrophages.

1.2.4.2 HuR regulatory network in *L. donovani* infected macrophages

It has been observed that HuR expression level decreases in *L. donovani* infected macrophages (Goswami et al., 2020) and thus alterations in HuR regulatory network that may occur in this scenario have been characterized. Initially, differentially expressed HuR target genes in isolated *L. donovani* infected human macrophages (16 hours) [Fold Change: 1.5, p-value: 0.05] have been determined considering the GEO dataset GSE360 (Chaussabel et al, 2003; R Core Team. 2016). Additionally, mRNA expression profiles in HuR (Elavl1) knockout murine bone marrow-derived macrophages (Lu et al, 2014) were determined considering the RNAseq analysis GEO dataset (GSE63199) (Afgan et al., 2018). Utilizing these datasets the probable alterations in HuR regulatory network in infected macrophages has been determined. Additionally, the network components that exhibited changes were mapped onto 'Leishmania infection (R-HAS-9658195)' and 'cytokine signaling in immune system (R-HAS-1280215)' pathways (Fabregat et al., 2017) in order to determine the physiological relevance of the alterations in network components. Further, comparison of HuR regulatory network components among HeLa, HEK293, non-infected and infected macrophages has been performed.

1.3 Discussion

Detailed analysis of intra-cellular interactions is likely to provide insights into the complex and flexible nature of these cellular interactions. In particular, different molecular entities (proteins and nucleic acids) in the intra-cellular milieu may interact in a number of ways to bring about different cellular processes or functions. In this study, the repertoire of intra-cellular interactions has been studied primarily from a systems point of view to determine or predict how the interactome networks (protein-protein interaction network, miRNA-mRNA regulatory network, RNA-protein regulatory network) might change under a diseased scenario. Observations about this system at hand may include alterations in network components or architecture and modifications in regulatory relationship between network components.

Complexities in intra-cellular interactions have been studied by modulating the expression level(s) of key regulator(s) and determining the effect of this perturbation in the intra-cellular network architecture or its components with the help of systems based computational analyses approaches. In this regard, different hypothesis had been proposed and their applicability or validity has been determined in a disease perspective utilizing a set of case studies. One of the key observations obtained during this analysis entailed essential proteins for a cellular process may lie at the interface of the gene regulatory and protein-protein interaction network. This particular observation was determined by over-expressing a transcriptional activator (necessary and sufficient for the cellular process), delineating the intra-cellular network governing the process and performing topological analysis of the network. Similarly, by outlining the miRNA-mRNA interaction network pertaining to an essential miRNA regulator, it was identified that substantial changes in network architecture may only occur upon prolonged alterations in the level of the essential regulator. Further, it was also identified that alterations in network architecture or regulatory relationship in miRNA-mRNA interaction networks or protein-mRNA interaction networks may be associated with alterations in cellular phenotype or response. Thus, analysis of intra-cellular interactions networks in different contexts has led to the identification of essential network components (proteins/mRNA) and important network modules that when de-regulated are associated with changes in cellular function or phenotype. Finally, by identifying disease associated entities and analyzing interaction patterns among them I could identify certain phenomena that are likely to be prevalent in intra-cellular interactions among bio-molecular entities.

Chapter 2

Identifying patterns in inter-cellular interactions with the help of sequence analysis approaches

Identifying patterns in inter-cellular interactions with the help of sequence analysis approaches

Cellular communication or signaling cascades may involve inter-cellular protein-protein interactions. The interaction between proteins may be transient or obligate; however, it is maintained throughout the course of evolution, despite alterations in the protein sequences. Moreover, since evolutionarily similar sequences tend to have similar functions or protein interaction partners, protein sequence analysis approaches may be utilized to predict the function of a protein, interaction domains or likely protein interaction partner(s). In this respect, homology searching protocols including BLASTp and Hidden Markov Model (HMM) based searches may be utilized to determine orthologous proteins involved in inter-cellular interactions. This may then be followed with a molecular docking analysis including probable interaction partners of the protein to identify the likely interacting partner(s) by studying the binding poses with the probable protein interactors. Additionally, in order to co-evolve or conserve the functional interaction between interacting proteins compensating alterations among critical residues in protein-protein interaction complexes may occur. Inter-dependent residue pairs that tend to co-evolve in inter-protein interaction complexes may be studied with the help of information theory based measures like mutual information. In this context, to determine inferences regarding inter-cellular interactions utilizing sequence analysis approaches, I have utilized different case scenarios as mentioned below:

In order to study the complex nature of inter-cellular interactions, I have considered the assumption that lateral gene transfers might allow the conservation of complex inter-cellular interactions between proteins throughout the course of evolution. Now, given that sequence similarity measures may be utilized to determine orthologous proteins, determine whether phylogenetically distant organisms may have remote orthologous proteins that exhibit a tendency to participate in similar inter-cellular interactions.

In order to study the flexible nature of inter-cellular interactions, I have assumed that co-variation allows the preservation of a functional interaction between co-evolving proteins that interact. However, residue pair alterations beyond a certain threshold may be detrimental for functional conservation. Therefore, given that interacting proteins may co-evolve, utilizing information contained in biological sequences, determine co-evolving residue pairs that could be structurally or functionally relevant for an inter-cellular protein-protein interaction to occur.

2.1 Results

2.1.1 Phylogenetically distant organisms may share similar virulence factors participating in inter-cellular interactions

Synopsis: In this section, sequence analysis measures have been utilized to determine orthologous virulence factor proteins involved in inter-cellular interactions and study whether they may have arisen as a result of lateral gene transfer between phylogenetically distant organisms. A computational approach for predicting novel virulence factor proteins within a pathogen proteome and identifying its (their) likely host interaction partner(s) is proposed herein. This methodology comprising of rigorous homology searching protocols including Hidden Markov Model (HMM) based profile-profile or profile-sequence comparisons and BLASTp-based sequence-sequence comparisons (Remmert, Biegert, Hauser, & Söding, 2012; Altschul, Gish, Miller, Myers, & Lipman, 1990) has been utilized to identify novel virulence factor proteins similar to bacterial virulence proteins in eukaryotic parasites belonging to Kinetoplastida. Additionally, molecular docking studies of the predicted virulence factors with a set of likely host receptor proteins might help identify the likely host interaction partner of the novel virulence factor proteins.

Problem Statement: Given that sequence similarity analysis may be utilized to determine orthologous proteins, determine whether similar virulence factor proteins may be present in Bacteria and Eukarya? Additionally, determine whether these remote virulence factor orthologs exhibit a tendency to interact with host proteins in a similar manner as their bacterial counterparts?

Hypothesis: Lateral gene transfers of bacterial proteins (virulence factors) to parasitic eukaryotes might promote the existence of remote virulence factor orthologs that may share similar interaction mechanisms with host cell surface receptors during inter-cellular interactions.

System of study: In order to establish parasitic infections, multiple proteins might mediate attachment and invasion of host cells and contribute to host-pathogen interactions. In this context, the present analysis was undertaken to identify whether orthologous proteins such as virulence factors in distant organisms could be involved in host-cell pathogenic interactions (Figure 2.1). Sequence and structure analyses might allow one to study the possibility that remote orthologs may share a significant evolutionary relationship and share similar subversion mechanism in host cells. Further, it is possible that lateral gene transfer from Bacillales to Trypanosomatidae could have played a role in their evolution. Therefore, sequence-based studies and phylogenetic analyses have been utilized to identify potential virulence factors in *L. donovani* and other *Leishmania spp.* based on their similarity to known bacterial virulence factor proteins. Moreover, with the help of homology modeling and docking studies the possibility that the novel virulence factor proteins so identified can interact with the class of host cell receptors that the bacterial counterparts are known to interact with was explored.

Graphical Summary:

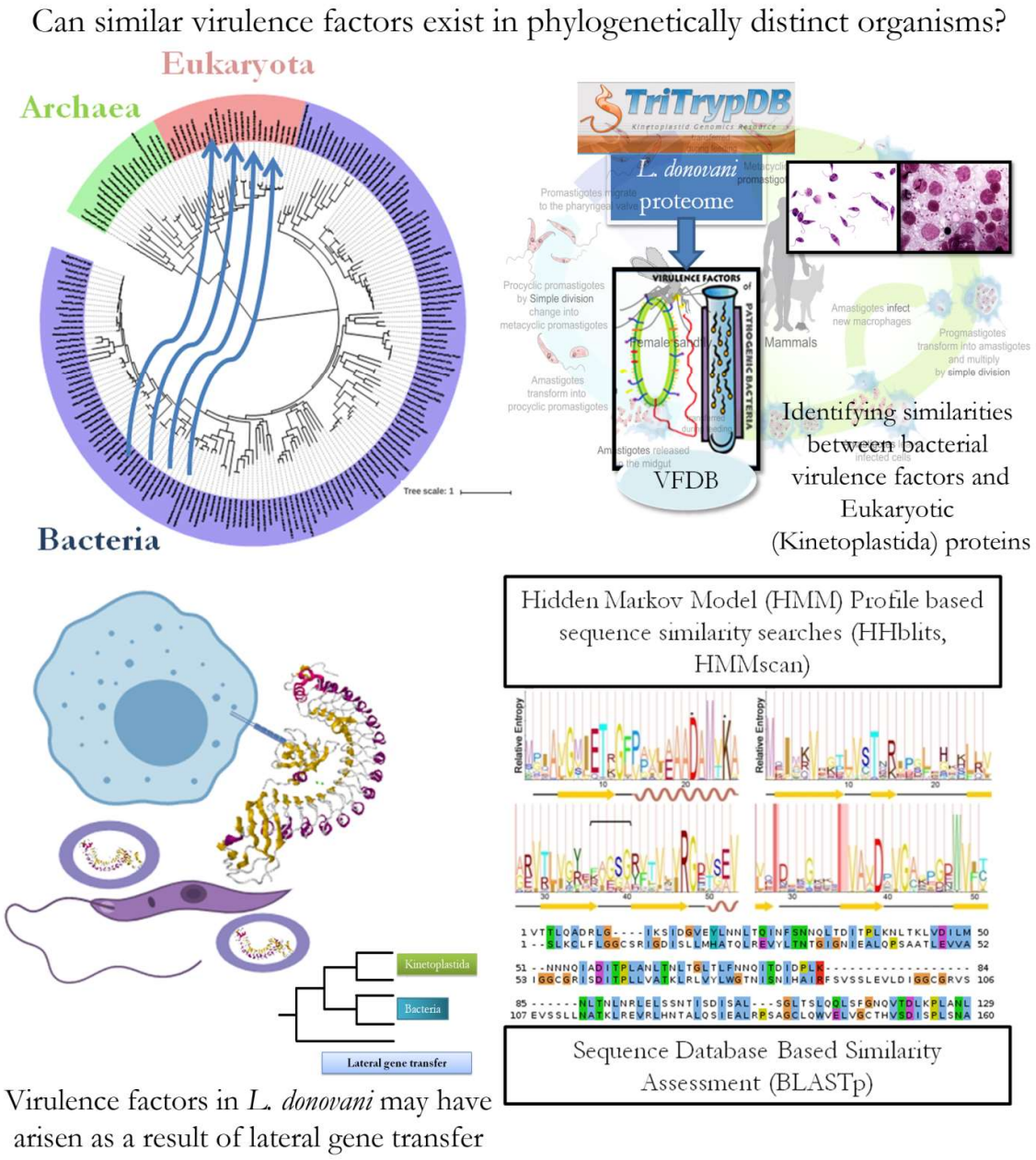


Figure 2.1: **Sequence analysis strategies may be utilized to identify virulence factors likely to be important in host-pathogen interactions.** Virulence factors may be identified based on sequence similarity approaches that determine orthology. Such orthologous virulence factors in phylogenetically distant organisms may be identified with the help of extensive homology identification strategies.

Predicting virulence factor proteins likely to be involved in host-pathogen inter-cellular interactions

Leishmania spp. (family Trypanosomatidae) are intracellular protozoan parasites that can interact with a number of different host receptors leading to invasion of the host cells and establishment of parasitic infections in mammalian hosts such as leishmaniasis. Identification of virulence factors in eukaryotic parasites such as *Leishmania* spp. has been performed utilizing 'forward and reverse search analysis' involving profile versus sequence (HMM (Eddy, 1998) based) and sequence versus sequence (BLASTp (Altschul et al., 1990)) comparisons. This search strategy identified remote but significant similarities between certain bacterial virulence factors and Leishmanial proteins. Further, the significance of evolutionary relationship between the bacterial and the predicted virulence factors was studied and a probable subversion mechanism that these predicted proteins may utilize was explored. Determining a new group of virulence factors in *Leishmania* spp. and a likely host invasion mechanism may help in the identification of new effective drug targets and tackling the problem of emerging drug resistance.

2.1.1.1 Bacterial virulence factor-like proteins may exist in eukaryotic parasites from *Leishmania* spp.

Lateral gene transfer from microbes to Trypanosomatidae may have played a role in their evolution (Alsmark et al., 2013) and therefore an initial search aiming to establish similarity between bacterial virulence factors and *L. donovani* proteins was undertaken. The search for identification of remote orthologs carrying representative virulence factor-like domains in the *L. donovani* proteome found 232 *L. donovani* proteins that had significant sequence similarity with bacterial virulence factor-like proteins in the virulence factor database (VFDB) (Chen, Xiong, Sun, Yang, & Jin, 2011) (Figure 2.2B). Herein, some of the *L. donovani* proteins (108) have been functionally characterized (Figure 2.2A); however, some *L. donovani* proteins without any known functions were found to be similar to VFDB proteins (Figure 2.2C, Table 2.1). Further, nearly 32% of the uncharacterized proteins were found to be orthologous to bacterial *L. monocytogenes* internalin-A (Inl-A) proteins possibly sharing an internalin domain.

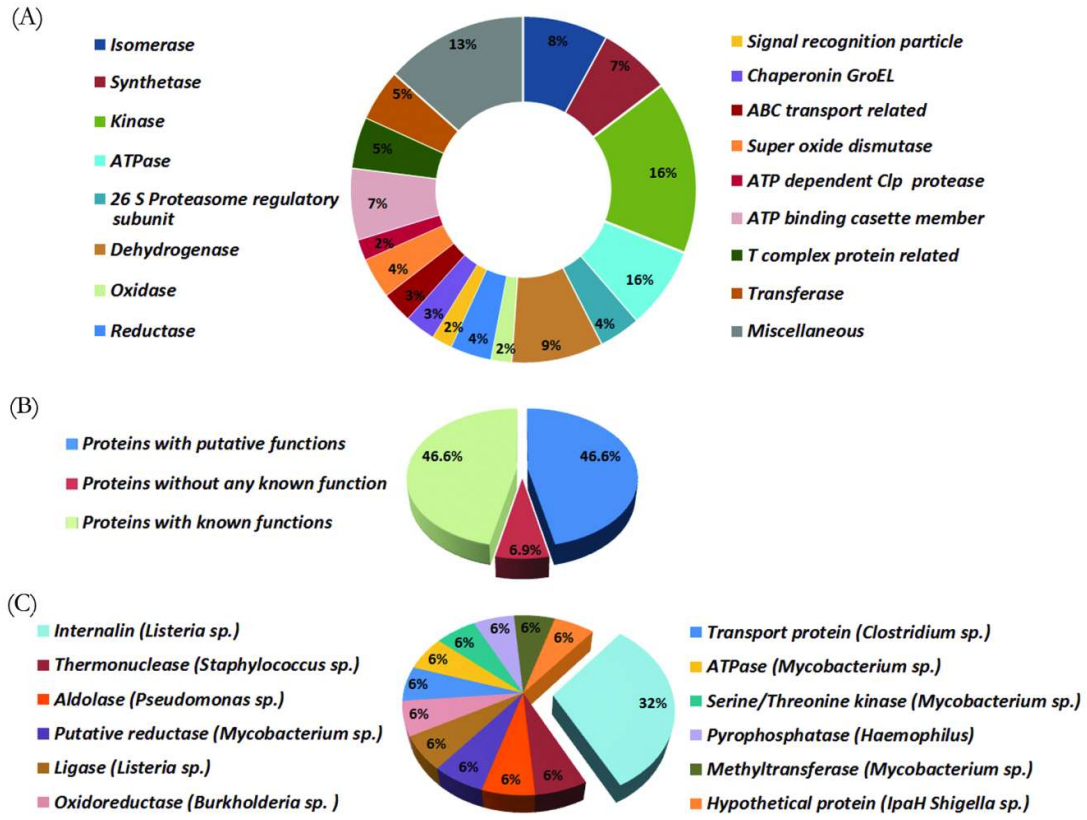


Figure 2.2: **Identifying the presence of bacterial virulence factor-like proteins in *L. donovani* proteome.** The virulence factor database (VFDB) containing bacterial virulence factor protein sequences was searched to identify whether any *L. donovani* proteins had significant similarity with bacterial virulence factor proteins. (A) Functional distribution among *L. donovani* proteins listed in the UniProt database that bears significant similarity to bacterial virulence factors. (B) Distribution of *L. donovani* proteins having significant matches with the VFDB proteins according to the search. (C) The *L. donovani* proteins without any known functions determined by BLASTp based search in VFDB exhibit similarity with bacterial proteins having the functional distribution shown here. [Note: published in Mukherjee et al., 2016]

Table 2.1: **List of *L. donovani* proteins that were found to be similar to virulence factors enlisted in VFDB database is reported here.** The VFDB proteins that the uncharacterized *L. donovani* proteins share similarity with according to the BLASTp search have been tabulated here. The functions of the VFDB proteins (as enlisted in VFDB database) that share similarity with *L. donovani* proteins have also been mentioned.

[1]Query ID (Gene Symbol)	Alternate Protein ID (UniProt)	[2]Subject details (VFDB-ID and Species)	VFDB Protein Function	[3]E-value	[1]Query Coverage (%)	[2]Subject Coverage (%)	Sequence Identity (%)
LdBPK_030010.1	E9B7L9	R029514 : *Listeria monocytogenes*	Internalin-A	1.98E-17	38.95	73.19	25.25
LdBPK_060020.1	E9B8L3	R029514 : *Listeria monocytogenes*	Internalin-A	6.98E-12	61.5	70.99	22.29
LdBPK_061120.1	E9B8X2	R002169 : *Staphylococcus saprophyticus*	Thermonuclease	1.30E-16	61.16	76.14	33.79
LdBPK_191670.1	E9BEF7	R029514 : *Listeria monocytogenes*	Internalin-A	3.42E-09	42.86	54.73	25.77
LdBPK_252090.1	E9BHW3	R012204 : *Pseudomonas syringae*	Achromobactin biosynthetic protein	3.79E-21	75.27	78.99	26.54
LdBPK_282190.1	E9BJU0	R030300 : *Listeria monocytogenes*	D-alanine activating enzyme	2.04E-22	46.66	95.81	27.92
LdBPK_311630.1	E9BMT7	R029514 : *Listeria monocytogenes*	Internalin-A	5.86E-13	47.35	71.21	24.42

LdBPK_340010.1	E9BQF5	R022463 : *Burkholderia cenocepacia*	Short chain dehydrogenase/ reductase	1.80E-11	70.59	80.84	25.34
LdBPK_342480.1	E9BR46	R008569 : *Clostridium perfringens*	CBS/transporter associated domain protein	2.99E-16	41.61	55.06	28.94
LdBPK_351270.1	E9BS03	R028655 : *Mycobacterium tuberculosis*	Putative ATPase	6.04E-12	36.29	44.74	31.16
LdBPK_361580.1	E9BTL7	R023384 : *Mycobacterium tuberculosis*	Serine/threonine protein kinase	1.16E-17	74.9	94.99	22.69
LdBPK_364860.1	E9BUJ4	R014950 : *Haemophilus somnus*	Non-canonical purine NTP pyrophosphatase	7.50E-13	93.91	98.48	29.68
LdBPK_365070.1	E9BUL5	R029515 : *Listeria ivanovii*	Putative internalin A	3.58E-09	26.07	40.23	25.35
LdBPK_365670.1	E9BUS4	R027032 : *Mycobacterium smegmatis*	Macrocin-O-methyltransferase	1.33E-12	63.47	80.42	27.67

[1]**Query :** refers to the *L. donovani* proteins.
[2]**Subject:** corresponds to VFDB proteins having significant similarity with the respective *L. donovani* proteins according to the BLASTp search.
[3]**E-value (expect-value):** The E-value is the average expected number of non-homologous proteins with a score higher than the one obtained for the database match. E-values closer to 0 are statistically significant.

2.1.1.2 Determining remote bacterial internalin-A orthologs in Kinetoplastida proteome

Identification of remote orthologs can be performed based on database searches comprising of sequence-sequence and profile-profile comparisons. A preliminary search for proteins similar to bacterial internalin-A in Kinetoplastida was performed utilizing the forward and reverse search analysis. In the forward search, some Kinetoplastida proteins (118 sequences) were found similar to bacterial virulence factor (Inl-A) sequence particularly the virulence region comprised of the signature LRR motif. However, since these sequences had sequence identities with Inl-A in the range of 22-36%, the probable Inl-A-like regions in the proteins selected based on the forward search were considered for a reverse search. In the HHblits search (reverse search analysis), ranks of internalin proteins obtained against the forward search selected sequences were utilized to determine probable Inl-A-like proteins in Kinetoplastida (Figure 2.3A). Probable Inl-A-like proteins were identified in *L. donovani*, *L. braziliensis*, *L. major*, *L. mexicana*, *L. panamensis*, *S. culicis*, *A. deanei* and *Phytomonas* sp. (Figure 2.3A). Thus based on this analysis, it is likely that an internalin-A like class of proteins is present in *Leishmania sp*. and other Kinetoplastida species which are remote orthologs of bacterial Inl-A-like proteins. Subsequently, during the course of this analysis the class of novel Inl-A-like virulence factors in Kinetoplastida has been studied particularly in the visceral leishmaniasis causative agent, *L. donovani*.

Bacterial virulence factor-like sequences in L. donovani

Database searches involving sequence-sequence, sequence-profile, and profile-profile comparisons may be utilized to identify orthologous proteins. In this respect, a forward and reverse search analysis was performed in order to ascertain whether bacterial Inl-A shared an orthologous relationship with certain *L. donovani* proteins. During the forward search analysis, sequence-sequence comparisons in BLASTp (Altschul et al., 1990) suggested two protozoan proteins (UniProt ID: E9B7L9, E9BMT7) to be Inl-A-like (Table 2.2).

Table 2.2: **Identified orthologs of a bacterial virulence factor in *L. donovani* based on a forward search analysis** (BLASTp search of the *L. donovani* proteome).

[1]Subject ID (Gene symbol)	Alternate protein IDs ([1]Subject) (Ensembl, UniProt)	[2]E-value	[3]Query (aligned region)	Query coverage (%)	Subject (aligned region)	Subject coverage (%)	Sequence identity (%)
LdBPK_030010.1	emb|CBZ31242 ; E9B7L9	$3.94e^{-19}$	10–342	96.8	439–831	38.95	25.25
LdBPK_311630.1	emb|CBZ36565.1; E9BMT7	$6.35e^{-14}$	16–339	94.19	301–684	47.35	24.68

[1]Subject: corresponds to *L. donovani* proteins which are probably Inl-A-like.

[2]E-value (expect-value): The average expected number of non-homologous proteins with a score higher

than the one obtained for the database match; E-values closer to 0 are statistically significant.
[3]**Query:** corresponds to Inl-A LRR region.
[Note: published in Mukherjee et al., 2016]

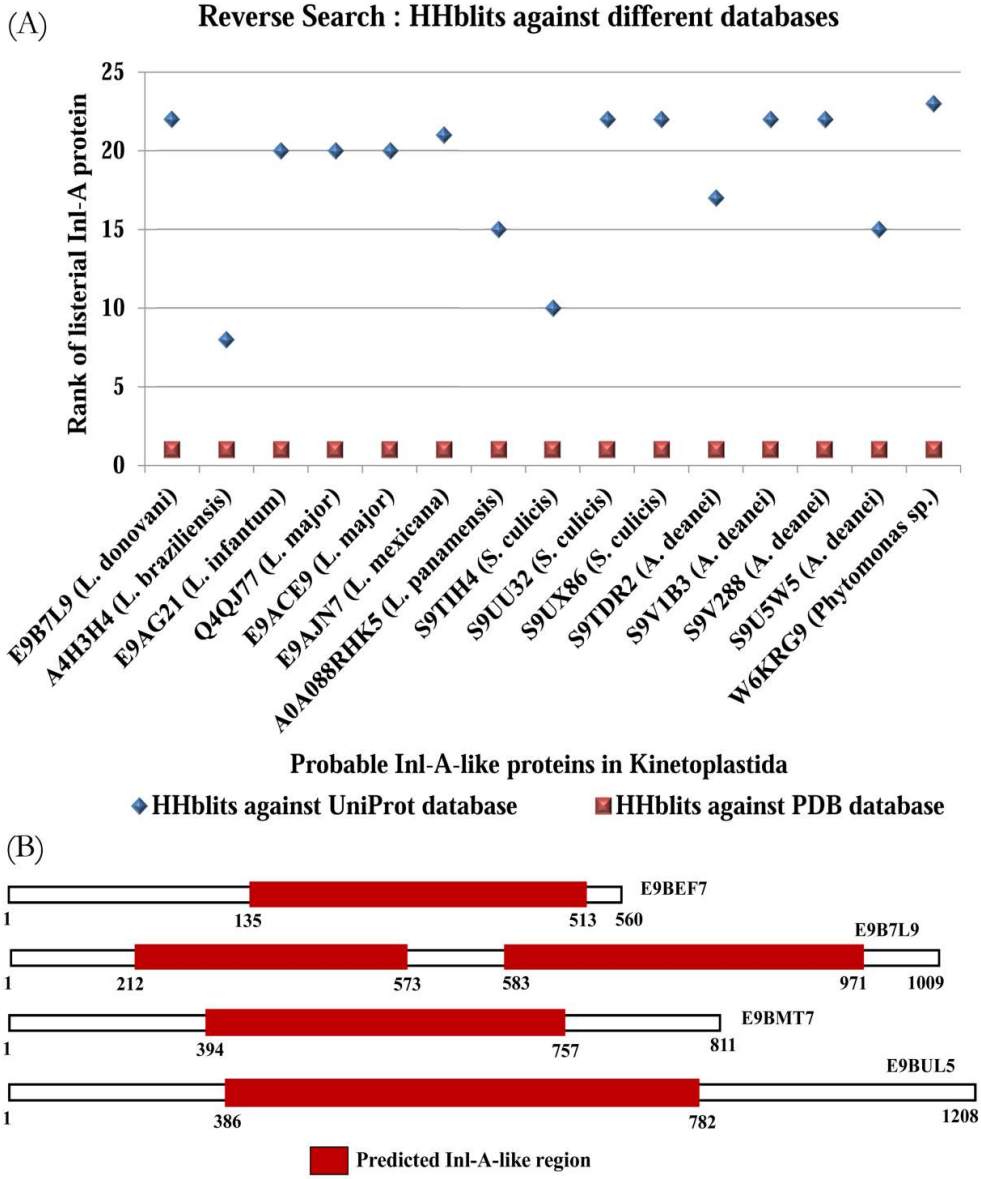


Figure 2.3: **Identification of Inl-A-like proteins in Kinetoplastida** (A) Forward search (BLASTp search in NR (Kinetoplastida) taking Internalin-A (*Listeria monocytogenes*) LRR region as query) and reverse search (HHblits search of forward search selected sequences against PDB and UniProt HMM databases) was performed. Ranks of probable Inl-A-like proteins in Kinetoplastida identified according to the reverse search analysis have been shown here. (B) Predicted *L. donovani* Inl-A-like proteins identified in a forward and reverse search analysis specifically considering *L. donovani* proteins.

However, it has been observed that sequence-profile and profile-profile based searches are more sensitive than sequence-sequence searches in detecting remote orthology. This is likely since profiles include more information than one sequence alone about a related family of proteins (Söding, 2004; Remmert et al., 2012). In this respect, an HMM profile was prepared to represent the virulence region in bacterial Inl-A. Subsequently, during the 'hmmscan' (Eddy, 1998, 2009) comparison of the *L. donovani* proteome with the modified Pfam-A (Finn et al., 2013) database (containing the Inl-A HMM profile) a number of domains were predicted including LRRs in many *L. donovani* proteins. However, in order to establish the existence of Inl-A-like proteins in *L. donovani*, proteins that were predicted to contain an Inl-A-like LRR region were considered for further analysis. Herein, five *L. donovani* proteins (UniProt ID: E9BUL5, E9BEF7, E9B7L9, E9B8L0, E9BMT7) were found to share significant similarity with the Inl-A profile (Table 2.3). Since the Inl- A LRR region is mainly known to interact with the host receptor E-cadherin (hEC1) during the invasion of host cells in *Listeria* pathogenesis (Lecuit, Ohayon, Braun, Mengaud, & Cossart, 1997), the presence of Inl-A-like LRR region(s) in *L. donovani* proteins suggests that these proteins are likely to be Inl-A orthologs. These *L. donovani* proteins were predicted to contain one Inl-A LRR-like region; however, E9B7L9 is likely to contain two possible Inl-A LRR-like regions. Additionally, an HMM-HMM-based search using HHblits algorithm (Remmert et al., 2012) and MAC realignment allowed refinement of the predicted LRR regions in the *L. donovani* Inl-A-like proteins (Table 2.4).

Table 2.3: **Identified bacterial virulence factor orthologs in *L. donovani* utilizing a forward search analysis (hmmscan).** Probable distant Inl-A orthologs identified in *L. donovani* proteome based on a sequence-profile comparison involving an hmmscan against a modified Pfam database including an Inl-A profile.

[1]Query ID (Gene symbol)	Alternate protein IDs [Query] (UniProt)	[2]Subject coverage (%)	Query length	[3]Domain score	[4]c-Evalue	[5]i-Evalue	[1]Query (Aligned region)	Sequence identity (%)	[6]Conservation (%)
LdBPK_365070.1	E9BUL5	81.49	1208	117	$3.30e^{-37}$	$6.10e^{-34}$	394–776	28.18	75.26
LdBPK_191670.1	E9BEF7	55.74	560	141.6	$1.10e^{-44}$	$2.00e^{-41}$	265–526	26.98	77.78
LdBPK_030010.1	E9B7L9	56.6	1009	117.4	$3.10e^{-37}$	$4.50e^{-34}$	202–467	23.05	75.78
LdBPK_030010.1	E9B7L9	60.64	1009	145	$1.30e^{-45}$	$1.90e^{-42}$	563–847	28.52	78.15
LdBPK_051200.1	E9B8L0	59.57	396	146	$5.70e^{-46}$	$9.40e^{-43}$	77–356	29.06	76.23

LdBPK_311630.1	E9BMT7	72.13	811	213.8	$1.60e^{-66}$	$2.70e^{-63}$	394–732	30.65	84.21

[1]**Query:** Corresponds to *L. donovani* proteins
[2]**Subject:** Corresponds to Internalin-A internalin domain profile.
[3]**Domain Score**: The calculated domain score for the best aligned profile (Inl-A)-sequence (*L. donovani*) matched states of the Inl-A-like internalin domain predicted in *L. donovani* proteins.
[4]**c-Evalue (conditional E-value):** It measures the statistical significance of each domain, given that the target sequence is a true homolog.
[5]**i-Evalue (independent E-value):** It represents the significance of the sequence in the whole database search, if this were the only domain one had identified.
[6]**Conservation**: It indicates estimation of identical or similar matching states over the alignment length.
[Note: published in Mukherjee et al., 2016].

The predicted Inl-A-like LRR regions of the 'forward search' selected *L. donovani* proteins were considered for the 'reverse search' to confirm reliable similarity between bacterial internalins and the predicted Inl-A-like *L. donovani* proteins. Herein, the predicted Inl-A-like LRR regions in the *L. donovani* proteins were compared against different sequence databases [NCBI non redundant database (NR), UniProt database] (Consortium, 2014; Pruitt, Tatusova, & Maglott, 2006; Altschul et al., 1990; Remmert et al., 2012) and structure database (PDB) (Rose et al., 2014). This search indicated that E9BMT7 (region: 394–757) and E9B7L9 (region: 212–573) are highly similar to different internalins since about 71% and 70% of their significant orthologous 'hits' retrieved via BLASTp belong to internalins (Figure 2.4A). Further, E9B7L9 (region: 583–971), E9BUL5 (region: 386–782) and E9BEF7 (region: 135–513) also have LRR rich regions which are similar to internalins. Moreover, the HMM-HMM-based search results were ranked according to E-value and Inl-A/Inl-A-like proteins appeared among the top ranked candidate proteins that are similar to *L. donovani* Inl-A-like regions (Figure 2.4B). However, based on the reverse search analysis it became apparent that E9B8L0 (region: 39–381) is less likely to possess an internalin domain. Therefore, based on the analyses performed herein involving different sequence-sequence, sequence-profile and profile-profile comparisons it is likely that E9BUL5, E9BEF7, E9B7L9 and E9BMT7 are virulent proteins possibly contain Inl-A-like LRR region within their sequence (Figure 2.3B). Interestingly, 13 leishmanial orthologs of *L. donovani* Inl-A-like proteins were also identified indicating that these Inl-A-like proteins are not restricted to *L. donovani* and an Inl-A-like class of proteins is present in *Leishmania* spp. (Figure2.5A, Table 2.5) and other Kinetoplastids.

Table 2.4: **Probable Inl-A-like LRR regions within the *L. donovani* Inl-A-like proteins.** HMM-HMM-based search utilizing HHblits and MAC realignment predicted LRR regions in each of the *L. donovani* Inl-A-like proteins as mentioned below.

[1]Query (UniProt ID, Gene ID TriTrypDB)	[2]Subject	[3]Probability (%)	[4]E-value	[5]p-value	[6]Score	Query HMM (Matched States Region)	Subject HMM (Matched States Region)
E9B7L9, LdBPK_030010.1	Internalin-A (1O6V chain A)	100	$6.00e^{-29}$	$1.70e^{-33}$	288.7	583-971	42-391
	Internalin-A (1O6V chain A)	100	$3.60e^{-28}$	$1.00e^{-32}$	282.1	212-573	37-384
E9BMT7, LdBPK_311630.1	Internalin-A (1O6V chain A)	100	$2.70e^{-32}$	$7.60e^{-37}$	309.2	394-757	42-389
E9BUL5, LdBPK_365070.1	Internalin-A (1O6V chain A)	100	$9.30e^{-31}$	$2.70e^{-35}$	317.8	386-782	44-390
E9BEF7, LdBPK_191670.1	Internalin-A (1O6V chain A)	100	$1.80e^{-32}$	$5.10e^{-37}$	284.2	135-513	42-391
E9B8L0, LdBPK_051200.1	Internalin-A (1O6V chain A)	100	$1.30e^{-28}$	$3.70e^{-33}$	235.3	39-381	47-387

[1]**Query:** refers to corresponding *L. donovani* proteins (UniProt ID).
[2]**Subject:** HMM profile having similarity with the corresponding *L. donovani* proteins.
[3]**Probability:** The secondary structure score (SS) is taken into account, together with the score to calculate the probability of a true positive homolog.
[4]**Expect-value (E-value):** The E-value gives the average number of false positives with a score better than the one for the template when scanning the database. E-values near 0 signify a very reliable hit.
[5]**p-value:** The p-value is the E-value divided by the number of sequences in the database. It is the probability that in a pair wise comparison a wrong hit will score at least this good.

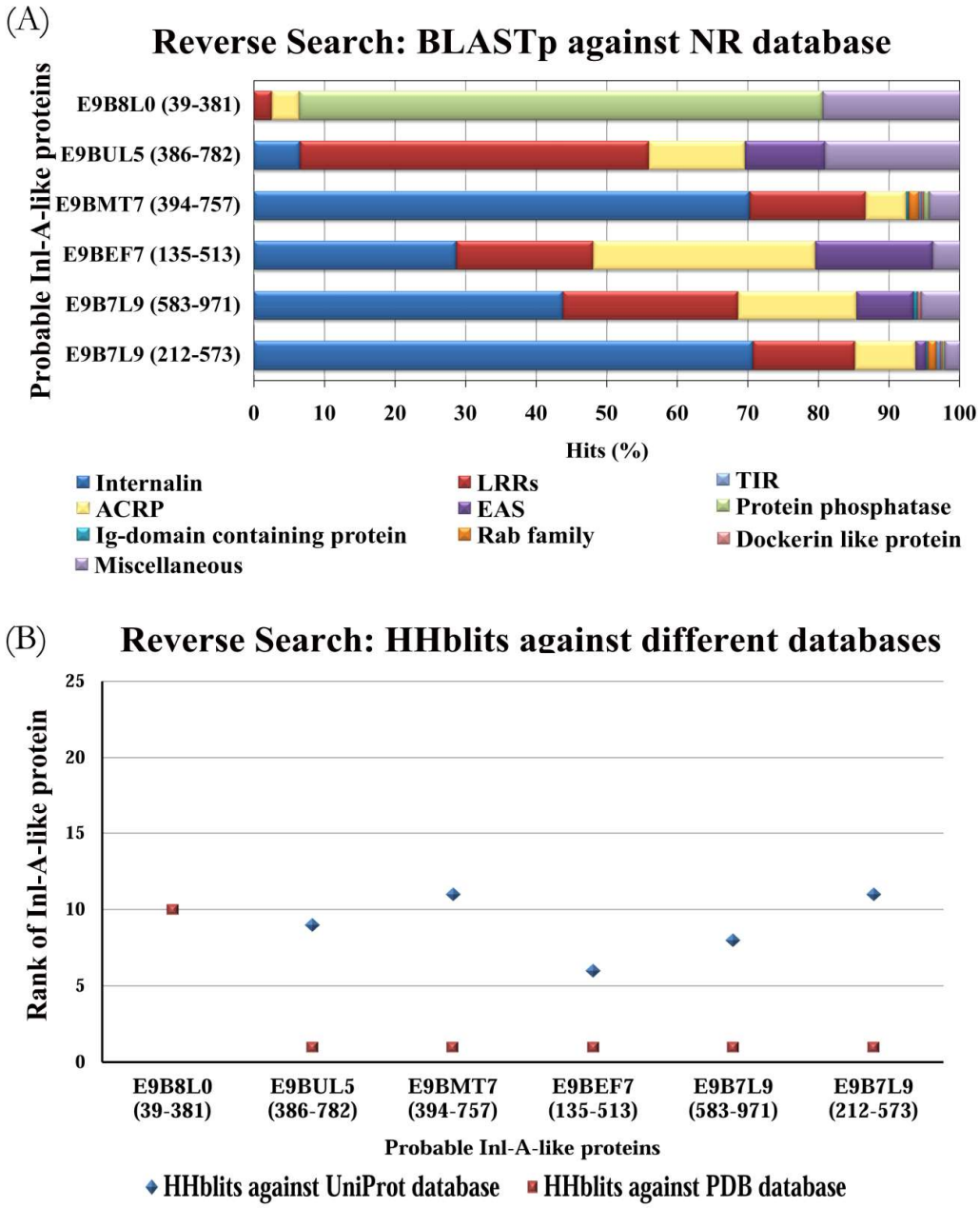


Figure 2.4: **Reverse search analysis identified Inl-A-like proteins in *L. donovani.*** (A) Major functional classes (percentage) of similar sequences (hits) obtained from a BLASTp search considering probable Inl-A-like region in *L. donovani* proteins as query. (B) The ranks of Inl-A/Inl-A-like proteins obtained in a HHblits (HMM-HMM-based lightning fast iterative sequence search) search by querying probable Inl-A-like region in *L. donovani* proteins against PDB and UniProt HMM databases. Ranks of the Inl-A and Inl-A-like proteins obtained within top 25 similar proteins are shown. [Note: published in Mukherjee et al., 2016].

Table 2.5: **Orthologs of *L. donovani* internalin-A like proteins in other Kinetoplastida**

	NCBI RefSeq ID/GenBank ID	Protein	Sequence Identity (%) (whole length)
E9B7L9 orthologs	XP_001561593.1	Hypothetical protein (*Leishmania braziliensis*)	68.6
	XP_00392173.1	Hypothetical protein (*Leishmania infantum*)	99.1
	XP_003721680.1	Hypothetical protein (*Leishmania major*)	87.3
	XP_003871673.1	Hypothetical protein (*Leishmania mexicana*)	84.7
	XP_010700943.1	Hypothetical protein (*Leishmania panamensis*)	68.3
E9BMT7 orthologs	XP_001467457.1	Hypothetical protein (*Leishmania infantum*)	99.6
	XP_001685146.1	Hypothetical protein (*Leishmania major*)	94.5
	XP_003877680.1	Hypothetical protein (*Leishmania mexicana*)	93
	Gb|EPY27897.1	Internalin-A-like protein (*Strigomonas culicis*)	32
E9BUL5 orthologs	XP_001469881.1	Hypothetical protein (*Leishmania infantum*)	99.3
	XP_001687069.1	Hypothetical protein (*Leishmania major*)	92.5
	XP_001569126.2	Hypothetical protein (*Leishmania braziliensis*)	75.3
	XP_003874842.1	Hypothetical protein (*Leishmania mexicana*)	89.9
	XP_010703427.1	Hypothetical protein (*Leishmania panamensis*)	73.9

2.1.1.3 Establishing orthology between the virulence factor proteins in Bacteria and Eukaryota

Pair-wise sequence identities between Inl-A and *L. donovani* Inl-A-like proteins ranges between 13-21% (Figure 2.5B), however a phylogenetic analysis considering orthologs of bacterial Inl-A and *L. donovani* Inl-A-like proteins suggests distant orthology among them with E9BMT7 being the most closely related (Figure 2.5C). Thus, in order to determine whether these remote orthologs share a significant evolutionary relationship the signature leucine rich repeat (LRR) motifs between the two sets of proteins were compared in-depth.

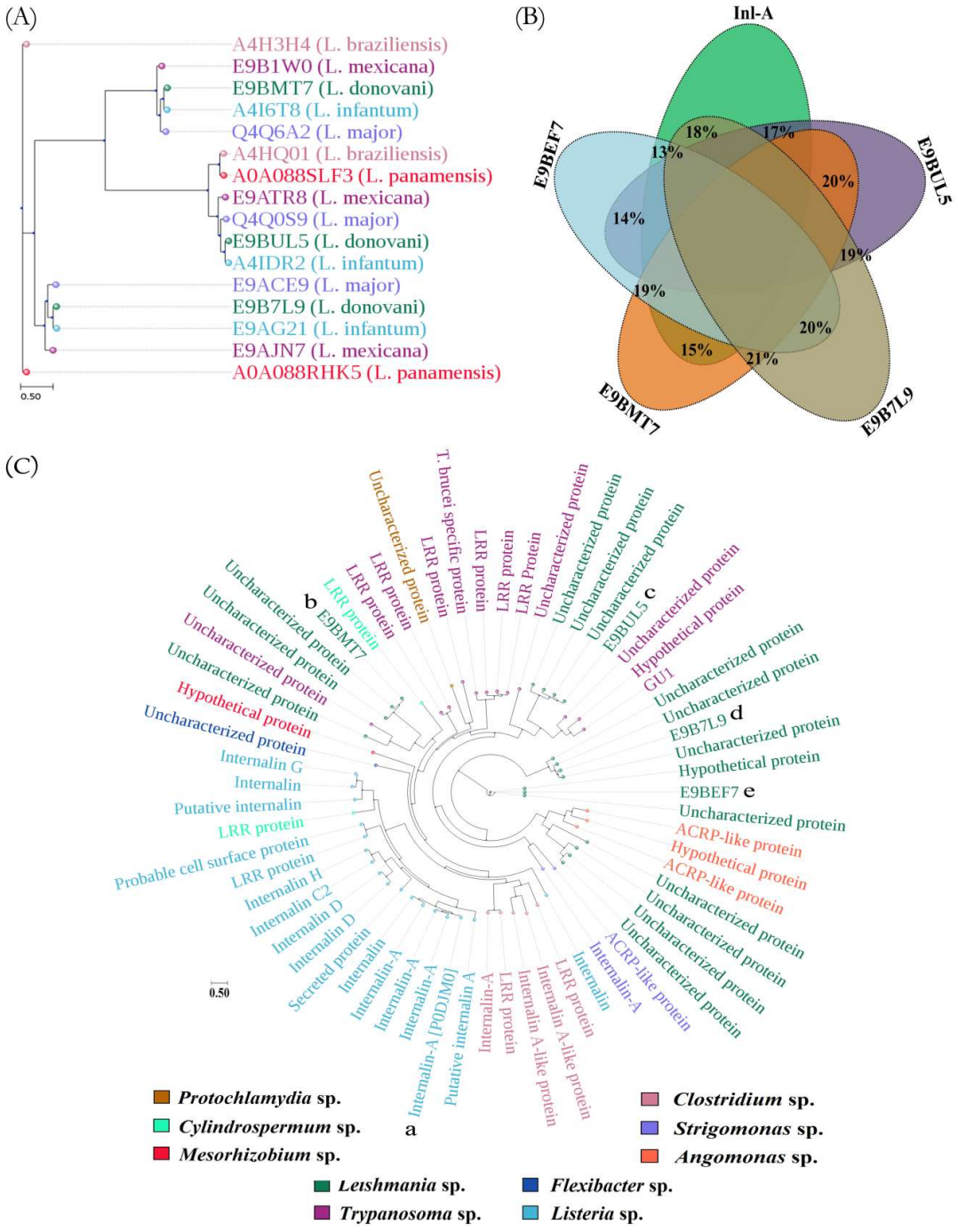


Figure 2.5: **Inl-A-like proteins in *Leishmania* spp.** (A) Phylogenetic tree of orthologs of *L. donovani* Inl-A-like proteins in *Leishmania* spp. prepared using the maximum likelihood algorithm in RAxML. (B) Pair-wise sequence identities between *L. monocytogenes* Inl-A and *L. donovani* Inl-A-like complete sequences. (C) Orthologs of Inl-A and *L. donovani* Inl-A-like proteins were compared to establish the extent of orthology they share. The generated

tree in circular layout shown here exemplifies the distant orthology between Inl-A (a) and E9BMT7 (b), E9BUL5 (c), E9B7L9 (d), E9BEF7 (e), respectively.

2.1.1.4 Establishment of LRR motif similarity among *L. donovani* Inl-A-like proteins and bacterial Inl-A

Average sequence identities between the equivalent LRR motifs (LRR1-15) in *L. donovani* Inl-A-like proteins and *L. monocytogenes* Inl-A were observed to vary in the range of 8–42% while sequence identities between the LRR motifs in distant bacterial Inl-A orthologs and Inl-A varied between 21-40% (Figure 2.6A). HMM-HMM comparison between LRR motifs in bacterial Inl-A distant orthologs and *L. donovani* Inl-A-like proteins indicated that the probability of an orthologous relationship between bacterial Inl-A LRRs and predicted LRRs of E9BMT7 (region: 394–757), E9B7L9 (region: 212–573) and E9BUL5 (region: 386–782) are reasonably high (Figure 2.6B). The profile-profile comparison resulted in probability >= 90% and an E-value<= $1e^{-05}$ for all the LRRs except one which had significant E-value but probability <= 90%. However, profiles of bacterial Inl-A LRR motifs and profiles of predicted LRRs in E9B7L9 (region: 583–971) and E9BEF7 (region: 135–513) on comparison either did not show significant E-values (E-value>=$1e^{-05}$) or probabilities (probability<=90%) for some of the LRRs (Figure 2.6B). Thus, *L. donovani* Inl-A-like proteins which had a reasonable degree of similarity with Inl-A in all of the LRRs particularly E9B7L9 (region: 212–573), E9BMT7 (region: 394–757) and E9BUL5 (region: 386–782) were considered for further study.

2.1.1.5 Lateral gene transfers of bacterial proteins to eukaryotes may have facilitated the existence of remote virulence factor orthologs

A phylogenetic tree considering orthologs of *L. donovani* Inl-A-like proteins identified that the E9BMT7 sequence (*L. donovani* protein) formed a cluster with some bacterial and archael internalin proteins along with some other bacterial and Trypanosomal proteins (Inl-A like and LRR protein) (Figure 2.7). This cluster of proteins is closely related to another cluster primarily comprised of E9B7L9, E9BUL5 and other Trypanosomal proteins including Internalin-A (*Strigomonas culicis* and some LRR proteins. In particular, the tree topology showed that one or more eukaryotic sequences (*Leishmania* spp.) clustered with prokaryotic sequences (*Listeria* sp., *Clostridium sp.*, *Bacillus* sp. etc.). Additionally, another phylogenetic analysis revealed that 16s rRNA genes from the order Bacillales and Lactobacillales (such as *Salinibacillus* sp., *Brochothrix* sp., *Listeria* sp., *Enterococcus* sp., *Vagococcus* sp. *Carnobacterium* sp. etc.) clustered more closely with 18s rRNA sequences from the order Kinetoplastida instead of 16s rRNA fom other bacterial origins. Briefly, 18s rRNA genes from *Strigomonas* sp., *Angomonas* sp., *Trypanosoma* sp., *Leishmania* sp. etc. which clustered together were found to be related to bacterial species particularly from the family Listeriaceae which included *Listeria monocytogenes*, *Listeria grayi*, *Listeria seeligeri*, *Listeria fleischmannii* etc. (Figure 2.8). Therefore, based on the clustering pattern of ribosomal RNA (rRNA) genes it is possible that lateral transfer of genes may have occurred between some prokaryotes and eukaryotes. Moreover, the phylogenetic analysis considering protein sequences also suggested that similar Internalin-A like sequences that are present in *Listeria* sp. and *Leishmania* sp. may have arisen as a result of lateral gene transfer.

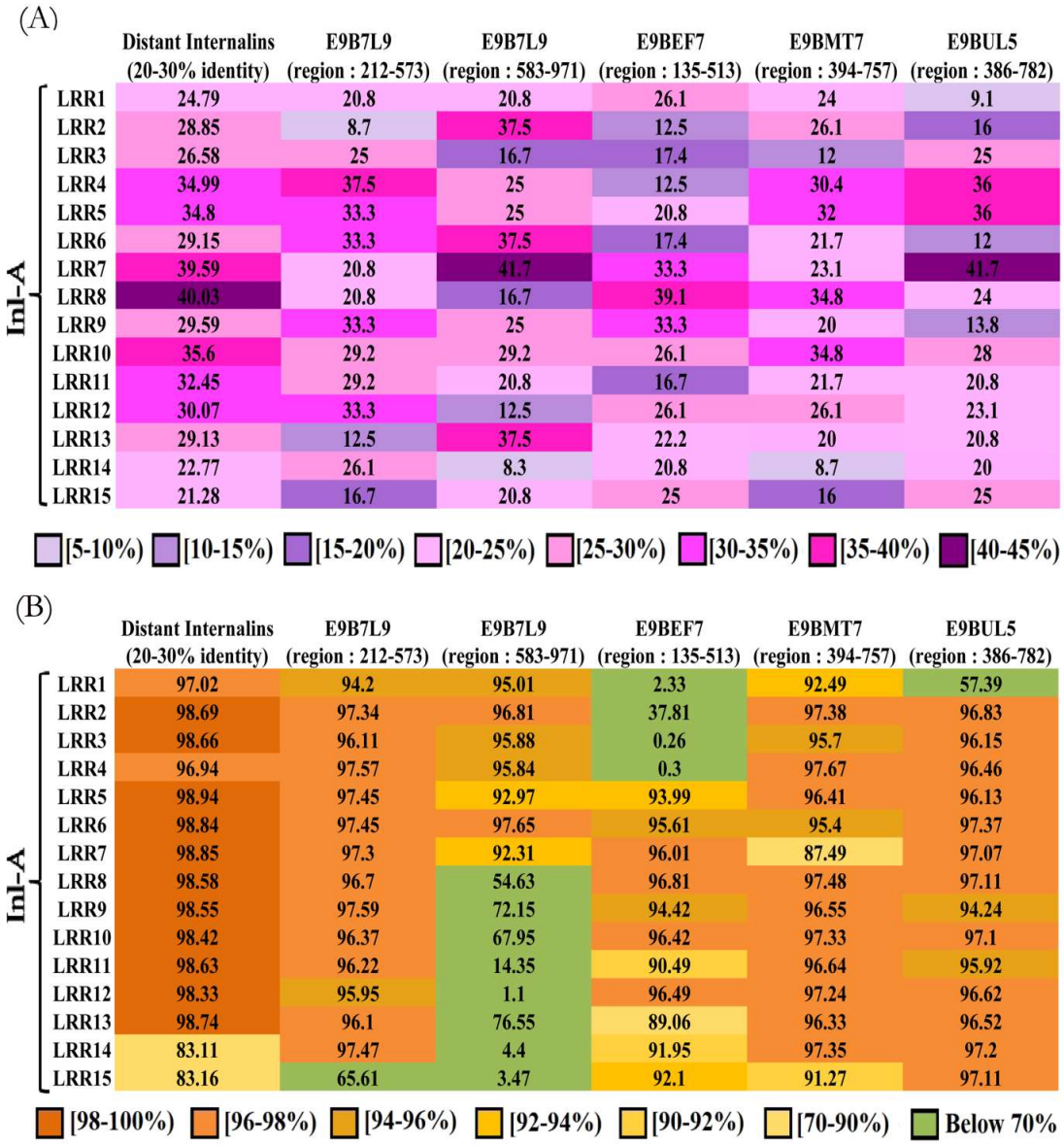


Figure 2.6: **LRR motif similarity between *L. donovani* Inl-A-like proteins and bacterial Inl-A.** (A) Pair-wise sequence identities between each Inl-A LRR and the corresponding predicted LRR in *L. donovani* Inl-A-like proteins has been compared with the average LRR sequence identities between Inl-A LRRs and distant internalin-like sequences. (B) Probability of similarity [identified based on (HMM) profile-(HMM) profile comparisons] between Inl-A LRRs and the corresponding predicted LRRs in *L. donovani* Inl-A-like proteins have been outlined. Average probability of similarity between LRR regions in distant internalin-like sequences and bacterial Inl-A LRRs has been considered for comparison. Probability>90% indicates an orthologous relationship that is either globally or locally similar in structure to the Inl-A LRRs. [Note: published in Mukherjee et al., 2016].

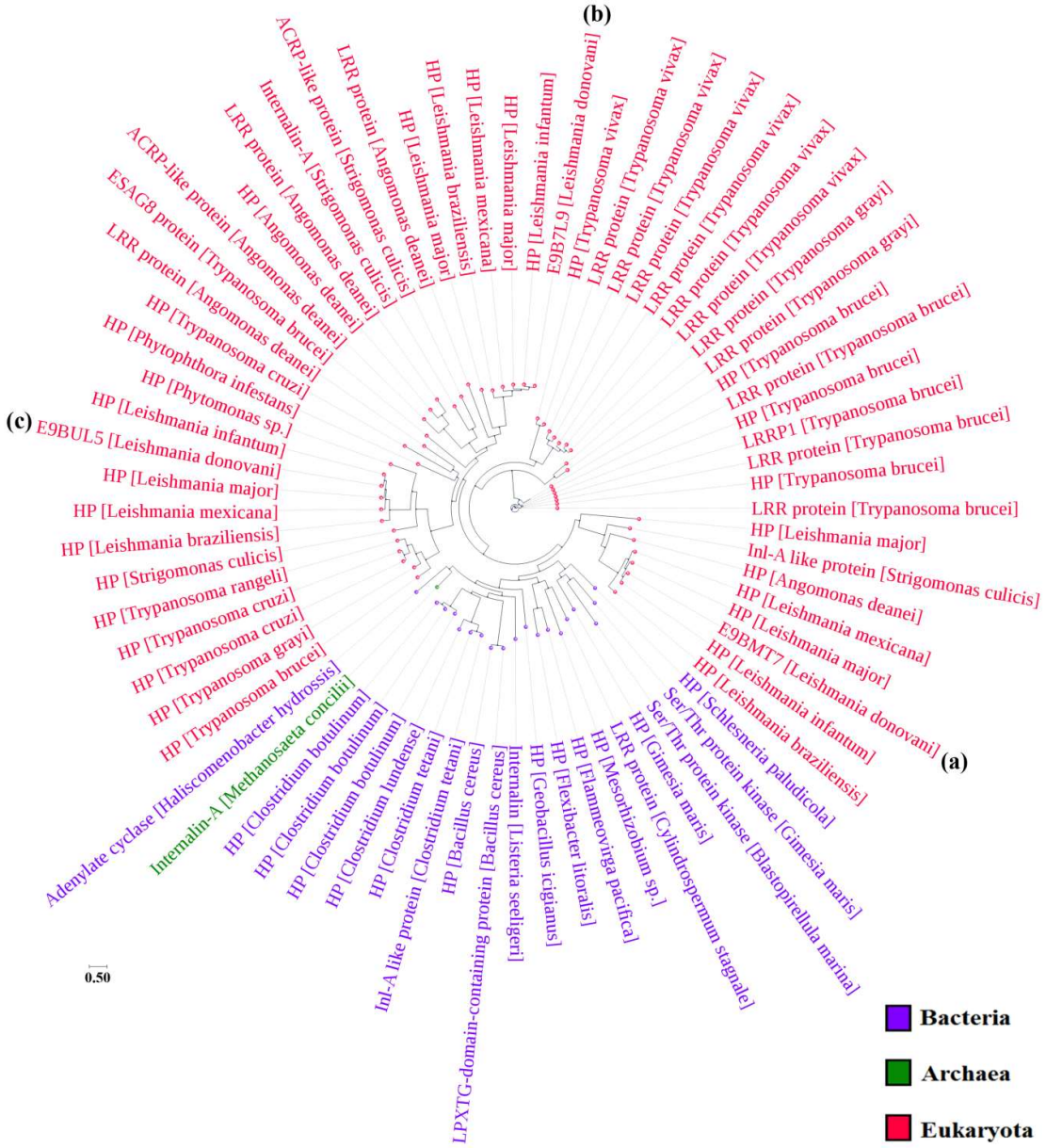


Figure 2.7: **Lateral gene transfer of bacterial Inl-A protein to Kinetoplastida** Phylogenetic tree considering orthologs of Inl-A and *L. donovani* Inl-A-like proteins was determined to analyze the evolutionary relationship between *Leishmanial* Inl-A-like proteins [(a),(b), (c)] and *Listerial* Inl-A.

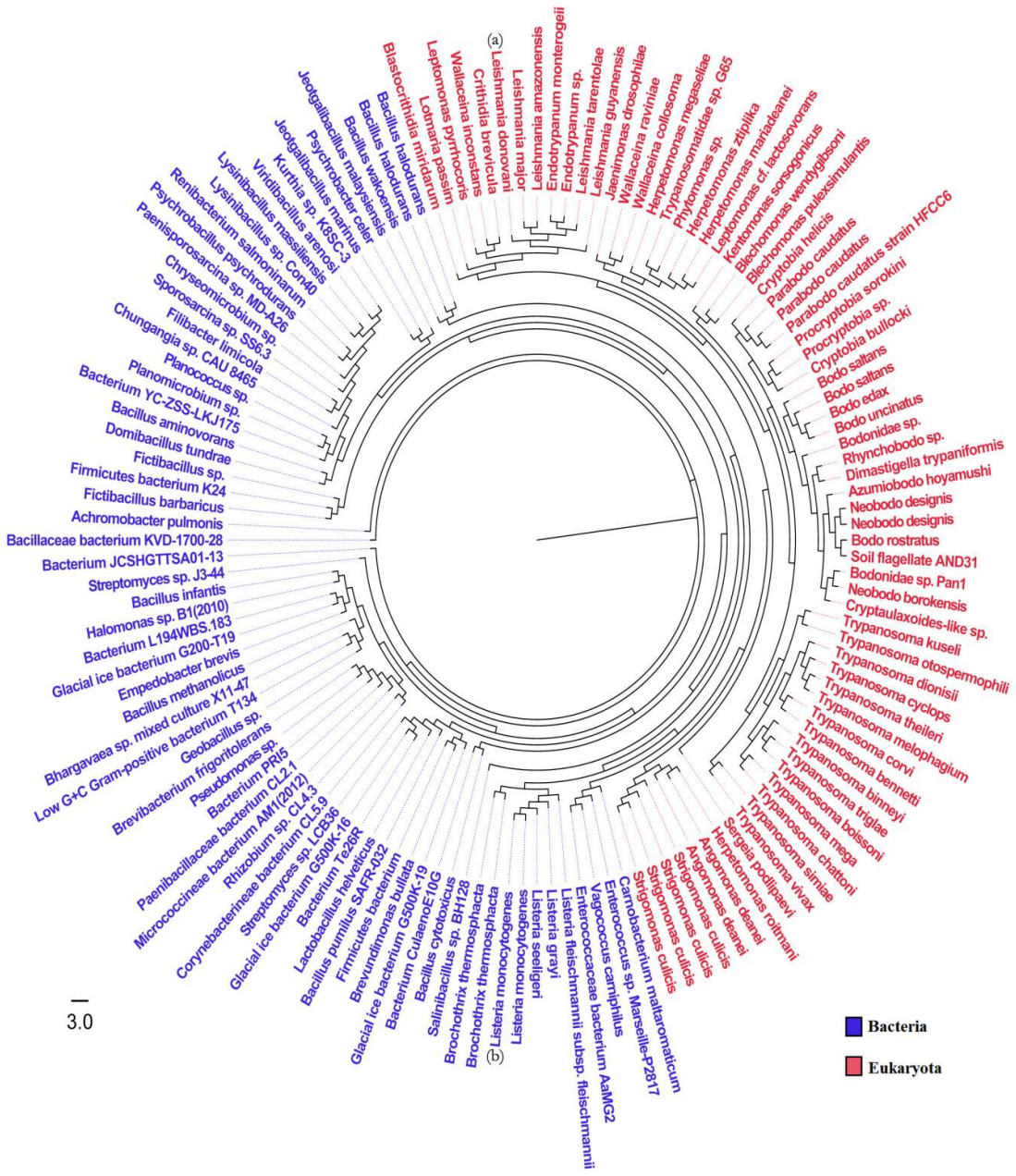


Figure 2.8: **Lateral gene transfer between prokaryotic and eukaryotic organisms.** Phylogenetic tree determined considering orthologs of 16S ribosomal RNA (*Listeria monocytogenes*) and 18S ribosomal RNA (*Leishmania donovani*) is depicted here.

2.1.1.6 Determining the structure of the predicted virulence factors (*L. donovani* Inl-A-like proteins) and identifying the probable interacting host protein

Homology modeling of bacterial Inl-A-like virulence factor proteins in L. donovani (L. donovani Inl-A like proteins)

Three-dimensional (3D) model structures for E9B7L9 (region: 212–573), E9BMT7 (region: 394–757) and E9BUL5 (region: 386–782) proteins were prepared based on the LRR region of *L. monocytogenes* Inl-A crystal structure (PDB ID: 1O6S, chain A). Since, the similarity between E9B7L9, E9BMT7, E9BUL5 and *L. monocytogenes* Inl-A (P0DJM0) falls within the twilight zone (20–30% sequence identity) of sequence similarity high quality alignments based on HMM-HMM comparison from HHpred was considered for template-based modeling (Söding, 2004; Söding, Biegert, & Lupas, 2005). In particular, the internalin domain (region 36–496) of *L. monocytogenes* Inl-A is comprised of a helical domain (residues: 36–78), LRR domain (residues: 79–414) and an immunoglobulin-like domain (residues: 415–495). The LRR domain comprises of 15 full and a half 22 residue repeats wherein each repeat contains a strand (xxLxL, L: leucine, valine, or isoleucine; x: any amino acid), a loop (xxNxLxx), a 3_{10}-helix (LxxLx), and a second loop (xLxxL). The arrangement of these repeats result in a right-handed solenoid with a stretch of conserved aliphatic hydrophobic residues and an asparagine directed toward the solenoid core (Schubert et al., 2002). 3D homology models for E9B7L9 (region: 212–573), E9BMT7 (region: 394–757), E9BUL5 (region: 386–782) were determined (Figure 2.9) and parameters describing these model structures are outlined in Table 2.6. The predicted structures of *L. donovani* Inl-A-like proteins mostly have 23 residue repeats in contrast to the 22 residue repeats present in *L. monocytogenes* Inl-A. Further, each of the 15 LRR repeats in the models consecutively contain a sheet, a loop, a helical region and another loop creating a solenoid. The solenoid generated by these repeats exhibits a similar pattern of aliphatic hydrophobic residues towards its core (Figure 2.9).

Table 2.6: **Validation of homology models of *L. donovani* Inl-A-like proteins**

***L. donovani* Inl-A like proteins**	**E9B7L9**	**E9BMT7**	**E9BUL5**
Predicted Inl-A like Region	212–573	394–757	386–782
Sequence similarity index			
Whole Sequence Identity with Inl-A (UniProt ID: PODJMO)	18 %	15 %	16.8 %
LRR Region Sequence Identity with Inl-A (UniProt ID: PODJMO)	21 %	20 %	22 %
LRR Region Sequence Similarity with Inl-A (UniProt ID: PODJMO)	41.7 %	41.2 %	33.4 %
Structure Similarity Index			
Root Mean Square Deviation (RMSD) [Å] with Inl-A (PDB ID:	1.2	1	1.5

1O6S, chain A)				
Stereo-chemical Evaluation				
Ramachandran Plot	Favored	320 (88.9 %)	319 (88.1 %)	334 (84.6 %)
	Allowed	37 (10.3 %)	39 (10.8 %)	53 (13.4 %)
	Outlier	3 (0.8 %)	4 (1.1 %)	8 (2.0 %)
Verify3d		Passed (93.65 % of the residues had an averaged 3D–1D score >= 0.2)	Passed (91.21 % of the residues had an averaged 3D–1D score >= 0.2)	Passed (88.92 % of the residues had an averaged 3D–1D score >= 0.2)

[Note: published in Mukherjee et al., 2016].

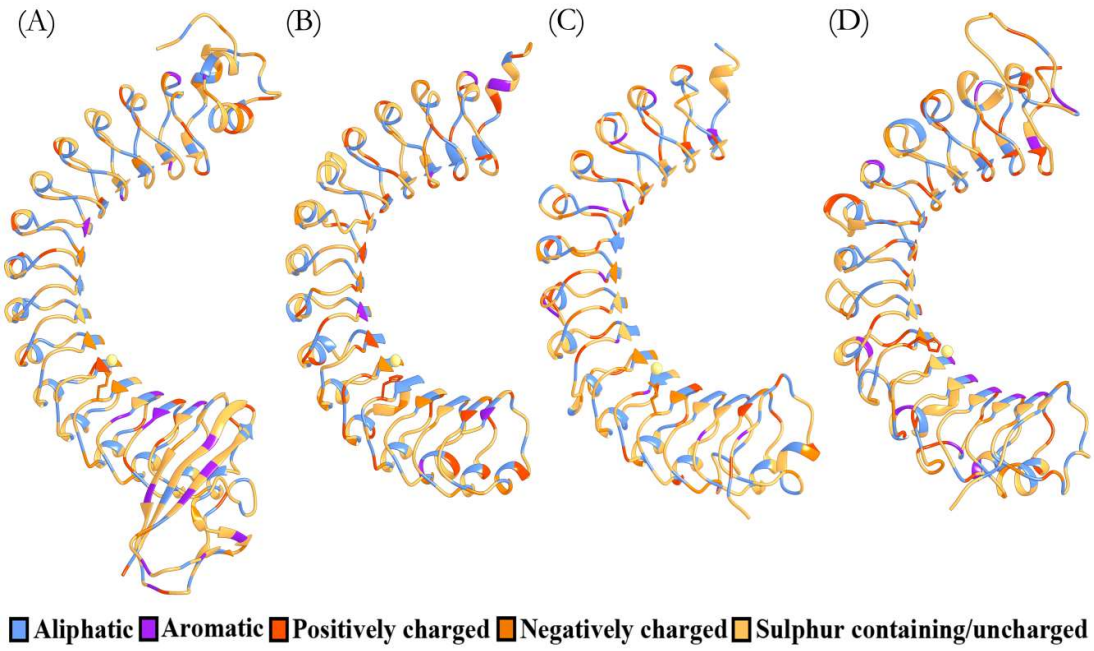


Figure 2.9: **Identified remote bacterial virulence factor protein orthologs in *L. donovani.*** Homology models of *L. donovani* Inl-A like proteins were determined and cartoon representations of E9B7L9 [region: 212–573] (B), E9BMT7 [region: 394–757] (C), E9BUL5 [region: 386–782] (D) are shown in comparison with *L. monocytogenes* Inl-A internalin domain (PDB ID: 1O6S, chain A) (A). Note: Amino acid residues are colored based on the nature of residues. [Note: published in Mukherjee et al., 2016].

Molecular docking analysis of L. donovani Inl-A-like proteins with cadherin superfamily members

Homologous proteins are likely to share protein interaction partners (Todd, Orengo, & Thornton, 2001) and in this regard we have explored the possibility that *L. donovani* Inl-A-like proteins might interact with E-cadherin (hEC1), the native *L. monocytogenes* Inl-A receptor or other members of cadherin superfamily. *L. monocytogenes* Inl-A is known to form a complex with E-cadherin wherein hEC1 fills the cavity created by the curved repeat domain of Inl-A forming hydrogen bonds with most of the repeats except LRR3 and LRR10 while LRR12-15 that have a patch of aromatic amino acids form extensive hydrophobic interactions with hEC1 (Schubert et al., 2002). A comparative molecular docking analysis was performed considering the possibility that *L. donovani* internalin-A like proteins (E9B7L9, E9BMT7 and E9BUL5) may interact with human cadherin superfamily members. In this respect, structural similarity assessment within representative members of this superfamily indicated that these cadherin proteins are very similar in structure having RMSD with E-cadherin in the range of 0.7-1.1Å (Table 2.7). The HADDOCK docking scores of representative cadherins [human P-cadherin (PDB ID: 4OY9), ectodomains of human Desmocollin 1 and 3 (PDB ID: 5IRY and 5EQX, respectively)] and E-cadherin (PDB ID: 1O6S) with the identified *L. donovani* internalin-A like proteins (E9B7L9, E9BMT7 and E9BUL5) are comparable. However, despite the receptors being structurally very similar the average l-RMSDs between *L. donovani* internalin-A like proteins docked to cadherin superfamily members varied significantly. The l-RMSD were calculated with respect to hEC1-Inl-A complex when hEC1 was docked with E9B7L9, E9BMT7 and E9BUL5 while docked complexes of *L. donovani* Inl-A-like proteins with type I cadherin and desmosomal cadherin were compared with respective hEC1 complex with E9B7L9, E9BMT7 and E9BUL5 to calculate the average l-RMSD. This analysis suggested that docked poses of *L. donovani* internalin-A like proteins with E-cadherin were similar to the native Inl-A and E-cadherin complex. Further, quality of protein-protein interaction predictions may be assessed based on the critical assessment of predicted interactions (CAPRI) evaluation criteria that utilizes fraction of native contacts (Fnat) and ligand root mean square deviation (l-RMSD) to evaluate performance of docking programs. Generally, the model and the reference protein are fitted on the backbone atoms of the larger protein to determine the l-RMSD on the backbone atoms of the smaller protein for identifying near native predictions. CAPRI criteria identifies $0.1 \leq Fnat < 0.3$ & l-RMSD ≤10Å, as acceptable predictions, $0.3 \leq Fnat < 0.5$ & l-RMSD≤5Å as medium quality predictions and Fnat≤0.5 & l-RMSD≤1Å as high quality predictions (Vries et al., 2007; M´endez, Leplae, De Maria, & Wodak, 2003). In particular, most of the top docking cluster poses had a reasonably similar l-RMSD (within 10Å) with the Inl-A-hEC1 complex while the average l-RMSDs of other docked cadherins such as P-cadherin and desmosomal cadherin indicated that these docked conformations are highly dissimilar to the Inl-A-hEC1 complex (Figure 2.10, Table S9). Therefore, it is possible that *L. donovani* internalin-A like proteins might interact with the cadherin superfamily members particularly E-cadherin.

Table 2.7: **Structural similarity assessment within the cadherin superfamily** Root mean square deviation between E-cadherin (1O6S chain B) and members of different cadherin family classes were computed to determine structural similarity among cadherin family members (human proteins or close orthologs were considered when structures for human protein were unavailable). Further, their expression pattern in different cells according to Protein Atlas database is also reported.

	PDB-ID	Description	RMSD with E-cadherin (1O6S chain B) [Å]	Expression (Human Protein Atlas)
Type I cadherin	2O72	Human E-cadherin	0.81	Epidermal cells, Langerhan cells, Melanocytes, Keratinocytes
	4OY9	Human P-cadherin	0.82	Epidermal cells, Langerhan cells, Melanocytes, Keratinocytes
Type II cadherin	3PPE	Chicken VE-cadherin	0.952	Epidermal cells, Langerhan cells, Melanocytes, Keratinocytes
	2A4C	Mouse Cadherin 11	0.821	Keratinocytes (Low)
Desmosomal	5EQX	Human Desmoglein 3 ectodomain	1.02	Keratinocytes
	5ERD	Human Desmoglein 2 ectodomain	1.05	Low /Absent in skin
	5IRY	Human Desmocollin1 ectodomain	0.786	Keratinocytes
Cadherin related	5DZW	Mouse Protocadherin alpha4	1.043	Absent in skin
	2YST	Human Protocadherin 7	1.12	Absent in skin

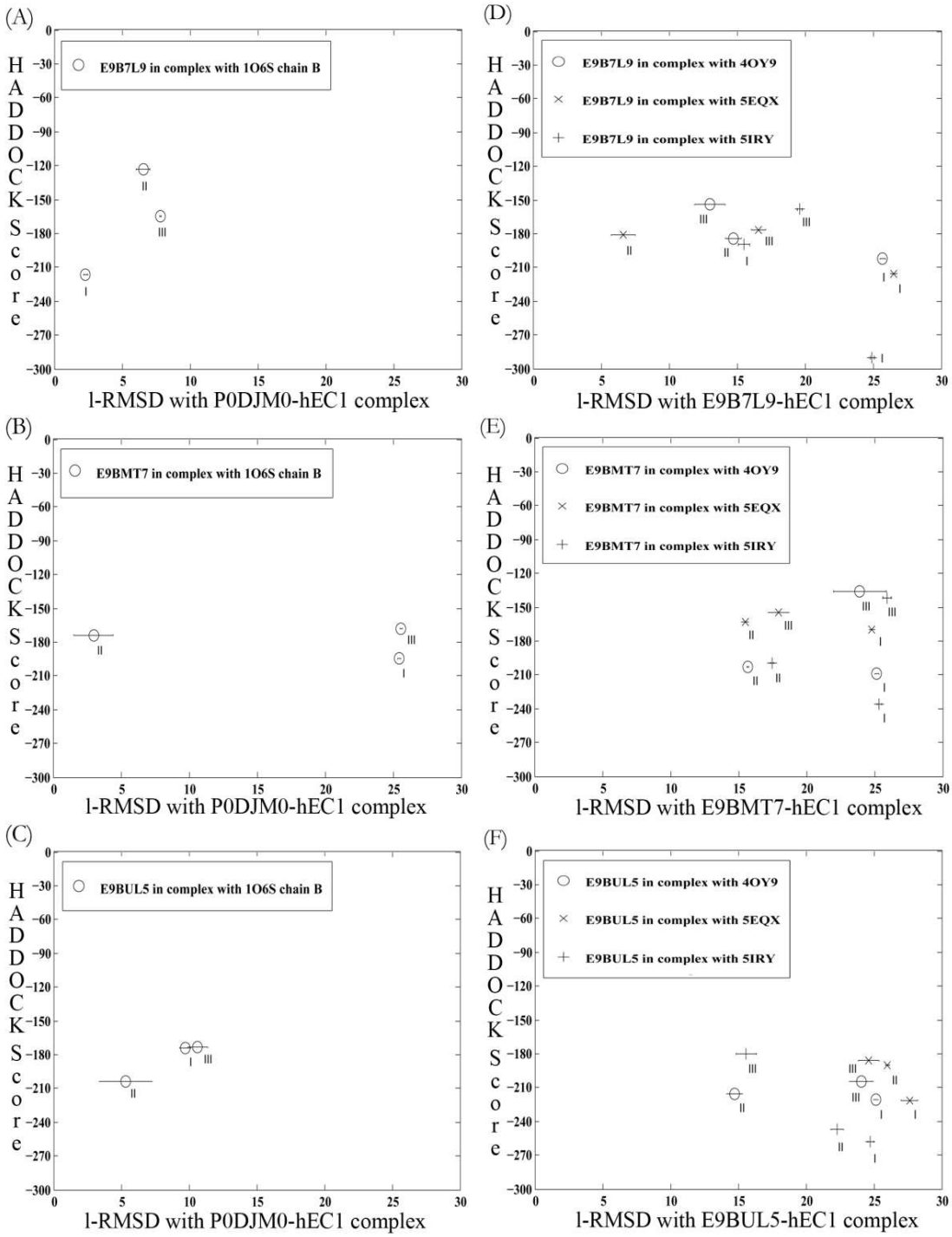


Figure 2.10: **Molecular docking analysis of *L. donovani* internalin-A-like proteins with representative members of cadherin superfamily.** Average HADDOCK scores of top three docking clusters (I, II, III) of *L. donovani* internalin-A like proteins with E-cadherin (hEC1) and their corresponding average l-RMSD with respect to hEC1 complex with *Listeria monocytogenes* Inl-A (P0DJM0) are shown. (A-C) Average HADDOCK scores and l-

RMSD (calculated with respect to hEC1-Inl-A complex) of hEC1 docked with E9B7L9, E9BMT7 and E9BUL5, respectively are enumerated. (D-F) Type I cadherin (PDB ID: 4OY9) and desmosomal cadherin (PDB ID: 5EQX, 5IRY) were docked with E9B7L9, E9BMT7 and E9BUL5, respectively. The corresponding HADDOCK scores and l-RMSD (calculated with respect to hEC1 complex with corresponding *L. donovani* internalin-A like protein) of the complexes are depicted.

Multiple approaches involving varying protocols and different scoring functions have been utilized for docking *L. donovani* Inl-A-like proteins with E-cadherin. Different scoring functions with various parameters and ranked clusters based on higher number of similar frames allow one to evaluate the docked conformations. Representative poses from clusters containing higher number of frames or better docking scores namely Cluster I, Cluster II, Cluster III were determined from the docking solutions of all the programs (Table S10). Similar poses of *L. donovani* Inl-A-like proteins docked with hEC1 among the top three possible interaction poses predicted from the multiple programs is suggestive of the most plausible interaction conformation between two proteins. During this analysis, most of the top docking cluster poses that were identified had a reasonably similar ligand RMSD with Inl-A-hEC1 complex indicating that the *L. donovani* internalin-A like proteins might interact with E-cadherin in a similar manner as *L. monocytogenes* Inl-A does. Docked conformations when compared to L. monocytogenes Inl-A structure in complex with hEC1 were found to be within 5Å for E9B7L9 (region: 212–573) and within 10Å for E9BMT7 (region: 394–757) and E9BUL5 (region: 386–782), respectively (Figure 2.11). Thus, if the *L. donovani* internalin-A like proteins are expressed during early stages of an infection it is likely that *L. donovani* internalin-A like proteins may interact with E-cadherin participating in host-parasite interactions during pathogen internalization.

Probable interaction pose of L. donovani Inl-A-like proteins with E-cadherin

Structural modeling and docking studies suggests the possibility that the interaction of *L. donovani* Inl-A-like proteins with host cell receptors such as E-cadherin may facilitate host cell infection. This is because multiple docking programs have suggested similar interaction poses between *L. donovani* Inl-A-like proteins and hEC1. Moreover, the predicted interaction conformation is similar to the native Inl-A-hEC1 interaction conformation. Further, the free energies of complex formation between E9B7L9 (region: 212–573)-hEC1, E9BMT7 (region: 394–757)-hEC1 and E9BUL5 (region: 386–782)-hEC1 are -7.4, -8.8 and -10.4 kcal/mol respectively. Additionally, the probable interaction forces that make these complexes stable include hydrogen bond interactions as well along with hydrophobic interactions in the predicted E9B7L9-hEC1, E9BMT7-hEC1 and E9BUL5-hEC1 docked complexes (Figure 2.12B-D, Table S11, S12). Therefore, all the molecular docking based observations indicated that the selected *L. donovani* Inl-A-like proteins and *L. monocytogenes* Inl-A might interact with hEC1 in a similar manner (Figure 2.12).

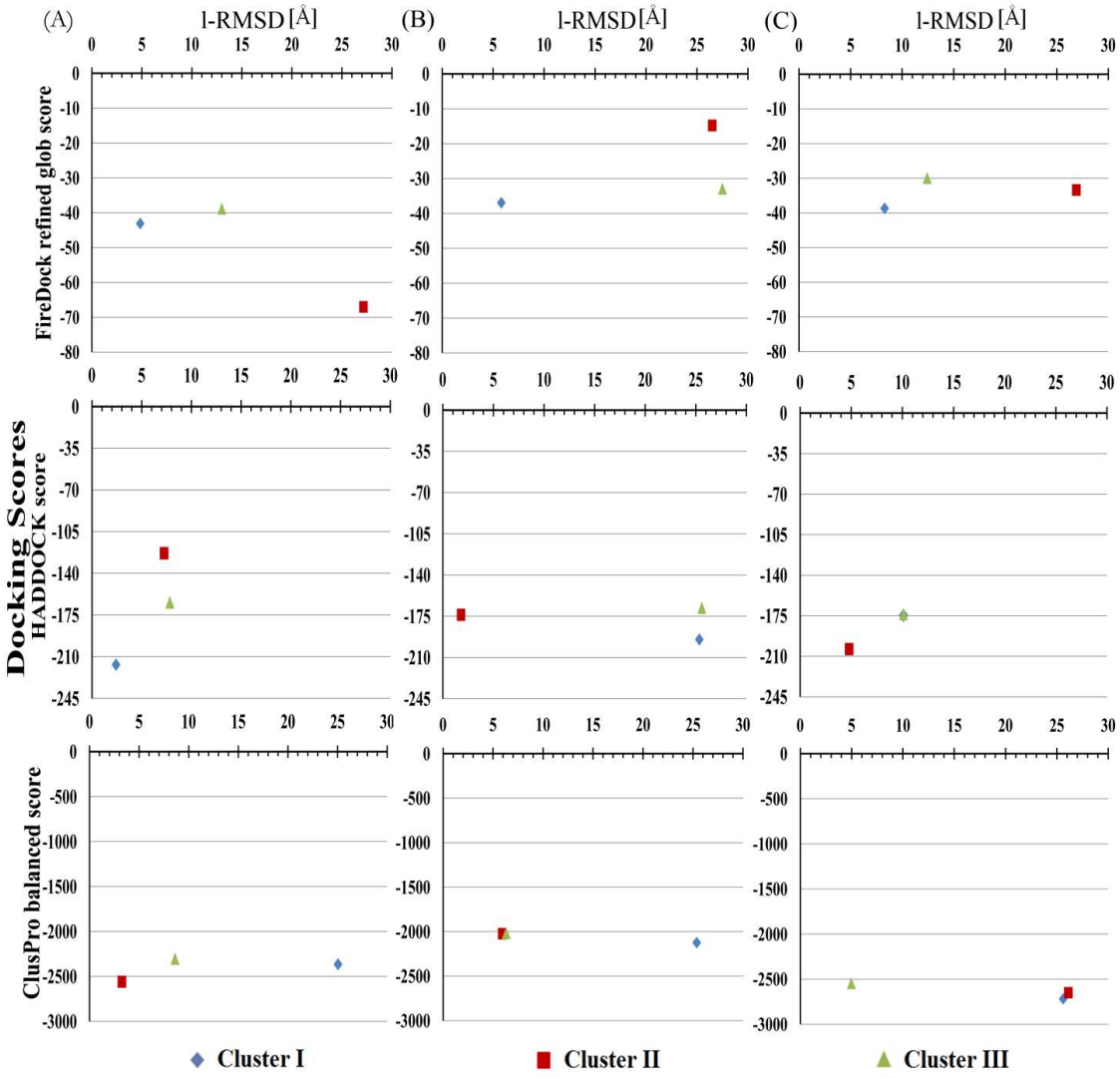


Figure 2.11: **Molecular docking of *L. donovani* Inl-A-like proteins with human E-cadherin (hEC1).** Directed docking solutions from each program were ranked and compared to identify consensus poses from multiple docking programs (PatchDock followed by FireDock refinement, HADDOCK and ClusPro). (A-C) Plot of docking scores for the top three significant clusters (Cluster I, Cluster II and Cluster III) with the mean of average root mean square deviations (RMSD) between the crystal complex (PDB ID: 1O6S) and docked poses of hEC1 along with *L. donovani* Inl-A-like proteins E9B7L9 [region: 212–573] (A), E9BMT7 [region: 394–757] (B) and E9BUL5 [region: 386–782] (C). [Note: published in Mukherjee et al., 2016].

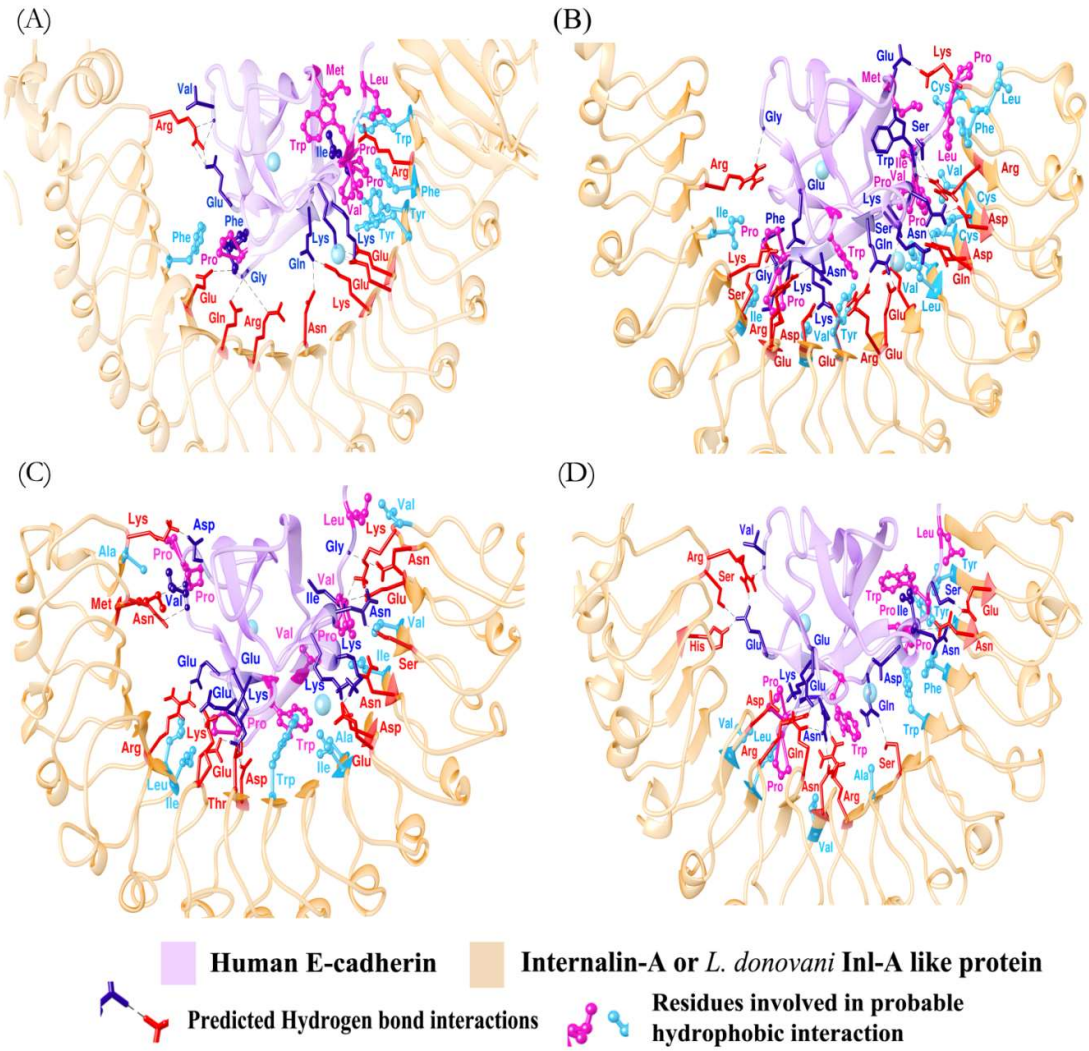


Figure 2.12: **Probable docking conformations of E-cadherin with *L. donovani* Inl-A-like proteins.** Ranked solutions from each program having ligand RMSD (l-RMSD) with Inl-A crystal structure within 10Å have been compared and best poses obtained from HADDOCK predictions are shown here as representatives. (A) *L. monocytogenes* Inl-A internalin domain crystal structure in complex with hEC1 [PDB ID: 1O6S]. (B-D) Docked conformation of *L. donovani* Inl-A-like protein E9B7L9 (region: 212–573) with hEC1 (B), E9BMT7 (region: 394–757) with hEC1 (C), and E9BUL5 (region: 386–782) with hEC1 (D). Key interacting residues likely to be involved in forming hydrogen bond interactions or hydrophobic interactions in the complexes (E9B7L9-hEC1, E9BMT7-hEC1 and E9BUL5-hEC1) are also shown. [Note: published in Mukherjee et al., 2016].

Conclusion

Leishmania spp. (family Trypanosomatidae) are intracellular protozoan parasites having two morphologically distinct variants during their life-cycle, a promastigote form (phlebotomine sand flies) and an amastigote form (mammalian cells) (Kaye & Scott, 2011). These parasites can utilize several polypeptide and polysaccharide ligands to interact with the host cells and some of these interactions could be synergistically or alternatively involved in receptor-mediated endocytosis. Further, once phagocytosed, *Leishmania* spp. may establish long term infections within phagosomal vesicles by subverting or altering host immune responses (Kaye & Scott, 2011). In general, *Leishmania donovani* utilizes lipophosphoglycan, leishmanolysin, parasite surface antigen 2, A2 protein, amastin, amastin-like surface protein, cysteine protease B, etc. during the establishment of visceral leishmaniasis (Kedzierski et al., 2004; Joshi et al., 2002; Zhang & Matlashewski, 2001; Rochette et al., 2005; Silva-Almeida et al., 2012). However, identifying additional virulence factors or invasion mechanisms might aid in the discovery of more effective drug targets to combat visceral leishmaniasis. In this respect, utilizing sequence similarity assessment strategies novel virulence factor(s) have been identified in *Leishmania donovani* based on the assumption that lateral gene transfer between prokaryotes and eukaryotes may have facilitated the existence of the same.

The forward search and reverse search analysis strategies employing extensive sequence-profile and profile-profile comparisons have been utilized to predict the existence of *L. donovani* Inl-A-like proteins, which are possibly remote orthologs of *L. monocytogenes* Inl-A. Elucidation of a class of Inl-A-like virulence proteins in *L. donovani* and/or other *Leishmania* spp. and further experimental characterization of these proteins in the context of host-parasite interactions may provide a better understanding of the mechanism of infection(s) mediated by *L. donovani* or *Leishmania* spp. in particular. A comparative LRR motif analysis of *L. donovani* Inl-A-like proteins with the Inl-A protein of *L. monocytogenes* revealed the existence of characteristic consensus LRR regions and a phylogenetic analysis suggested a reliable evolutionary relationship between them. Therefore, based on sequence analysis and phylogenetic analysis this study suggests the existence of an internalin-A-like class of proteins in *L. donovani* and some other species in Kinetoplastida. Internalin-A (Inl-A) protein of *Listeria monocytogenes* happens to be a surface LRR protein which mediates host cell invasion by interacting with E-cadherin on host cells (Schubert et al., 2002). However, such an invasion mechanism has not been previously reported in *Leishmania donovani*, the causative agent of visceral leishmaniasis. Further, three dimensional (3D) modeling of *L. donovani* Inl-A-like proteins and subsequent molecular docking studies with the host interaction partner (hEC1) of their bacterial ortholog (Inl-A) suggested that an interaction between the *L. donovani* Inl-A-like proteins and E-cadherin may occur.

Pathogens may use diverse strategies to subdue host defenses during the establishment of parasitic infections and a number of different proteins are likely to contribute to host-pathogen interactions. In this respect, it is plausible that Inl-A-like proteins in *Leishmania* sp. can interact with E-cadherin or other receptors structurally similar to E-cadherin (such as cadherin superfamily members). However, key factors

determining such an interaction may include the stage of expression of the virulent proteins and proximal presence of host receptor proteins. A recent observation suggests that leishmanial Inl-A-like proteins if expressed in the promastigote stage may promote cell invasion during the initial stages of infection by interacting with the host cell receptors. The fact that E-cadherin is widely expressed in the skin, specifically in epidermal cells, keratinocytes, langerhans cells and melanocytes (Uhlén et al., 2015; Uhlen et al., 2010) supports this possibility. Additionally, the experimental finding that leishmanial Inl-A-like proteins are expressed in promastigote stage (Atayde et al., 2015), provides preliminary support that these predicted virulent proteins are expressed, and may be involved in host-pathogen interaction. In particular, it has been reported that a close *L. major* homolog of E9BMT7 (95% identical to E9BMT7) is present in the exosomes co-egested in the inocula with the parasite during the sand fly's bite (Atayde et al., 2015). Moreover, interaction with host cell receptors leads to internalization of *Leishmania* spp. and subsequently by effectively suppressing and evading host immune responses they utilize different mechanisms to re-infect and/or establish long term infections within host cells. In this context as well if the *L. donovani* Inl-A-like proteins are expressed in the amastigote stage they might play a role in suppressing or evading host immune responses. Preliminary supportive evidence towards this possibility has been obtained in an RNAseq study comparing the expression profile of virulent and less virulent parasites. Herein, the three *L. donovani* Inl-A-like genes were found to be over-expressed in more virulent parasites (intracellular amastigotes) as compared to less virulent parasites isolated from infected murine resident peritoneal macrophages nearly 12 h post-infection (unpublished data). Thus, it is possible that *L. donovani* Inl-A-like proteins such as E9BMT7 could be associated with higher infectivity of *L. donovani*. However, direct experimental evidence to support these hypothesis, for instance, lower infectivity upon knock down of these protein(s) and other mechanistic studies such as receptor binding interaction analysis will aid in he further characterization of these virulent proteins, the plausible invasion mechanism and eventually help in the development of better intervention strategies.

Inference: Sequence analysis based approaches may be utilized to determine virulence factor proteins that are likely to be involved in inter-cellular interactions occurring between host cells and pathogens. Further, phylogenetically distant organisms may have acquired similar virulence factors by means of lateral gene transfer during the course of evolution. The remote protein orthologs may have low sequence similarities but contain similar signature motifs or virulence factor interaction domains that are likely to interact with host protein interaction partners in a similar manner as their bacterial orthologs. Thus, with the help of extensive sequence-profile and profile-profile comparisons a class of proteins termed as 'internalin-A-like proteins' which are remote orthologs of *L. monocytogenes* Inl-A has been identified in *L. donovani* and other *Leishmania* spp. in Kinetoplastida.

2.1.2 Inter-dependent residue changes in interface and non-interface co-evolving residues preserve functional interaction among inter-cellular protein interaction complexes

Synopsis: The present analysis has been performed to determine residues or co-evolutionary pairings likely to have important structural or functional roles in protein-protein interaction complexes. Interaction between protein interfaces is likely to be maintained by evolutionary pressure wherein selection may preserve a degree of conservation in the binding interface(s) or restrict amino acid replacements for functional preservation (Lovell and Robertson, 2010). Therefore, in order to identify co-evolutionary pairings in protein-protein interaction complexes a co-variation measure based on mutual information and Bhattacharryya coefficient has been developed. In this manner, molecular co-evolution in inter-cellular protein interaction complexes involved in cancer metastasis has been studied and co-evolutionary pairings have been determined. Further, these residue positions involved in co-evolutionary pairings occurred among residues that occur at the interface or non-interface regions and co-ordinated changes at these positions could be crucial for preservation of native interactions.

Problem Statement: Given that information theory may be utilized to study interacting proteins that co-evolve, determine residue pairs that could have crucial roles in protein-protein interactions.

Hypothesis: Proteins involved in interactions throughout the course of evolution tend to co-evolve and pairs of residues in these proteins may co-evolve to allow the preservation of a functional interaction between these proteins. However, certain residue pair alterations could be detrimental and determining co-evolving residue pairs that could be structurally or functionally relevant for conservation of inter-protein interactions might allow one to modulate these interactions in disease processes.

System of Study: Interacting proteins may co-evolve and exhibit compensatory or coordinated changes to maintain functional interactions. Analysis of co-evolution in protein-protein interaction complexes may be performed utilizing measures based on information theory. Evolutionary interaction between protein sites in different molecules that undergo compensatory changes to maintain the stability or functions of the interaction over the course of evolution results in co-evolutionary pairings between the inter-protein residue sites. A method named Co-Var has been developed to identify such inter-molecular residues that could have important structural or functional roles in interacting proteins as a result of co-evolutionary pairings among them (Figure 2.13). Protein-protein interaction complexes involved in inter-cellular communication in cancer have been studied to determine whether they exhibit co-evolution. Co-evolutionary pairings that are likely to be crucial for the native interactions in these complexes have been identified to deduce functionally important residue positions, modifications at which may result in aberrant complex functionality.

Graphical Summary:

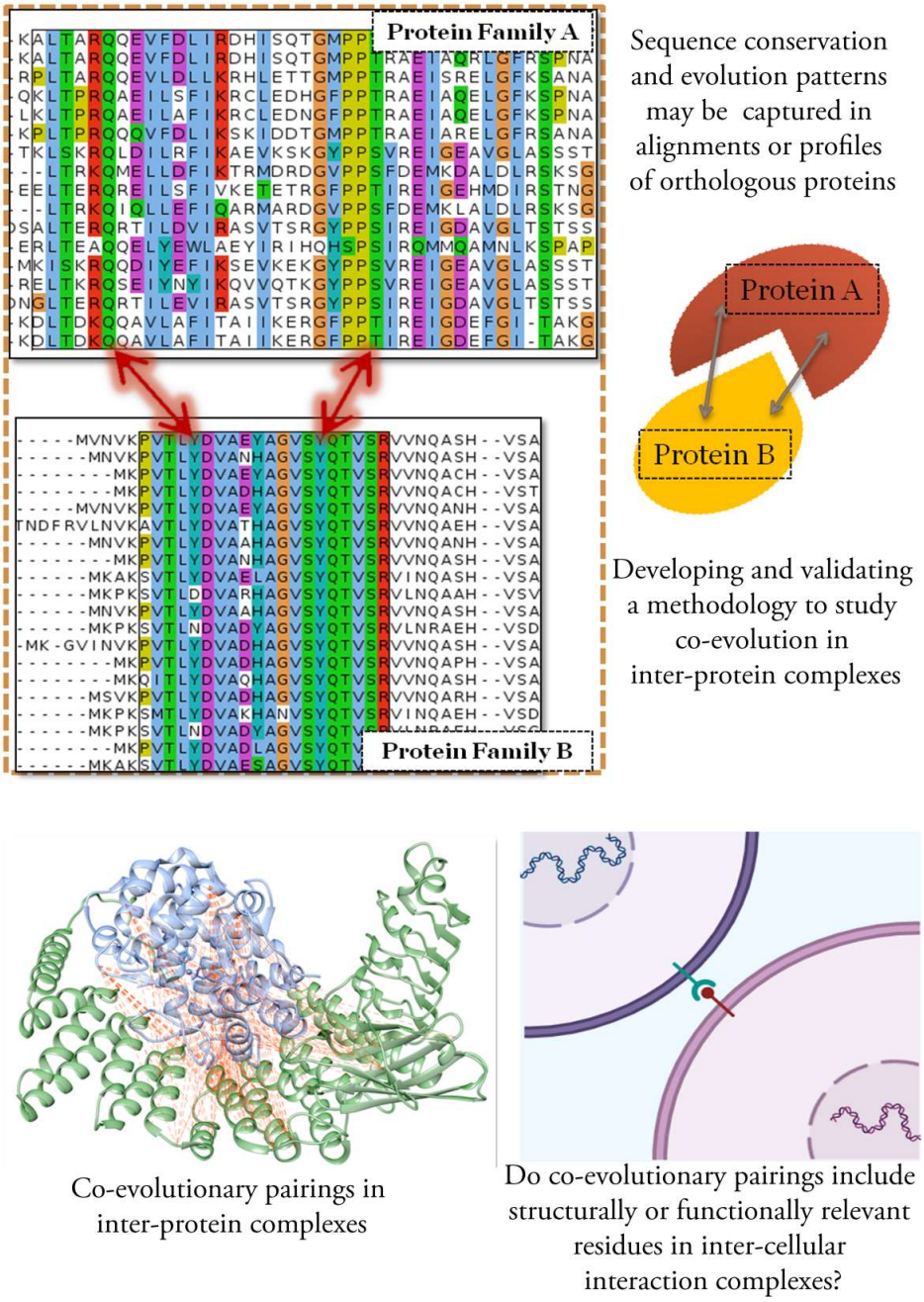


Figure 2.13: Identifying functionally relevant co-evolutionary pairings in inter-cellular protein-protein interaction complexes. An evolutionary approach based on information theory utilising sequence analysis can be utilised to identify co-evolutionary pairings. Such inter-dependent protein residues may be structurally or functionally crucial for conservation of interaction between a set of proteins.

Co-ordinated changes at interface and non-interface co-evolving residues preserve functionality in protein-protein interaction complexes

Evolutionary selection pressure in interacting proteins may promote co-evolutionary interaction wherein changes in one protein molecule may affect or require a reciprocal change in the other molecule involved in the complex. This has generally been observed in the case of inter-molecular residues that have crucial structural or functional roles in protein-protein interactions particularly the ones occurring at the interface. In this respect, an evolutionary approach based on information theory may be utilised to identify co-evolutionary pairings among inter-protein residue positions that exhibit inter-dependent changes as a result of evolutionary interaction. Such an approach termed as Co-Var has been developed and is outlined below. This approach has additionally been utilised to study co-evolutionary pairings in protein interaction complexes involved in cancer metastasis with a view to identify inter-protein residue positions crucial for a functional interaction in these inter-cellular complexes.

2.1.2.1 Identifying co-evolving residue pair interactions in protein-protein interaction complexes

Inter-protein co-evolution analysis was studied among interacting and non-interacting proteins to derive an index 'percentage of co-evolved pairs' which is likely to be higher for interacting complexes with a tendency to co-evolve in comparison to non-interacting proteins which are less likely to co-evolve. In order to determine the applicability of Co-Var methodology in studying inter-protein co-evolution its performance has been compared with inter-protein co-evolution analysis methods such as MirrorTree (Ochoa and Pazos, 2010), CAPS (Fares and McNally 2006; Fares and Travers 2006), and EV-complex (Hopf et al., 2014) methods respectively. The median of the MirrorTree correlation co-efficient distribution for the positive dataset was found to be higher than 0.8 suggesting that these complexes are more likely to co-evolve than the negative set of complexes with the median of distribution lower than 0.8 (p-value <= 0.0001) (Figure 2.14A). The Co-Var methodology predicted that interacting (positive) complexes have 'percentage of co-evolved pairs' index in the range of 6-10% while the non-interacting (negative) complexes have 'percentage of co-evolved pairs' index in the range of 0-2%. This indicated that interacting complexes are more likely to co-evolve than non-interacting proteins and as such this methodology could segregate the two set of complexes substantially (p-value <= 0.0001) (Figure 2.14B, Table 2.8). In general, interacting complexes had a higher percentage of co-evolving pairs than the non-interacting complexes in CAPS and EV-complex but the segregation in the distributions of the index calculated for the positive and negative set of complexes is not as significant as that obtained in Co-Var (Figure 2.14C-D, Table 2.8). Therefore, the Co-Var methodology may be utilised to determine whether protein-protein interaction complexes are likely to co-evolve and in addition to identify residue pair positions that exhibit inter-dependent changes.

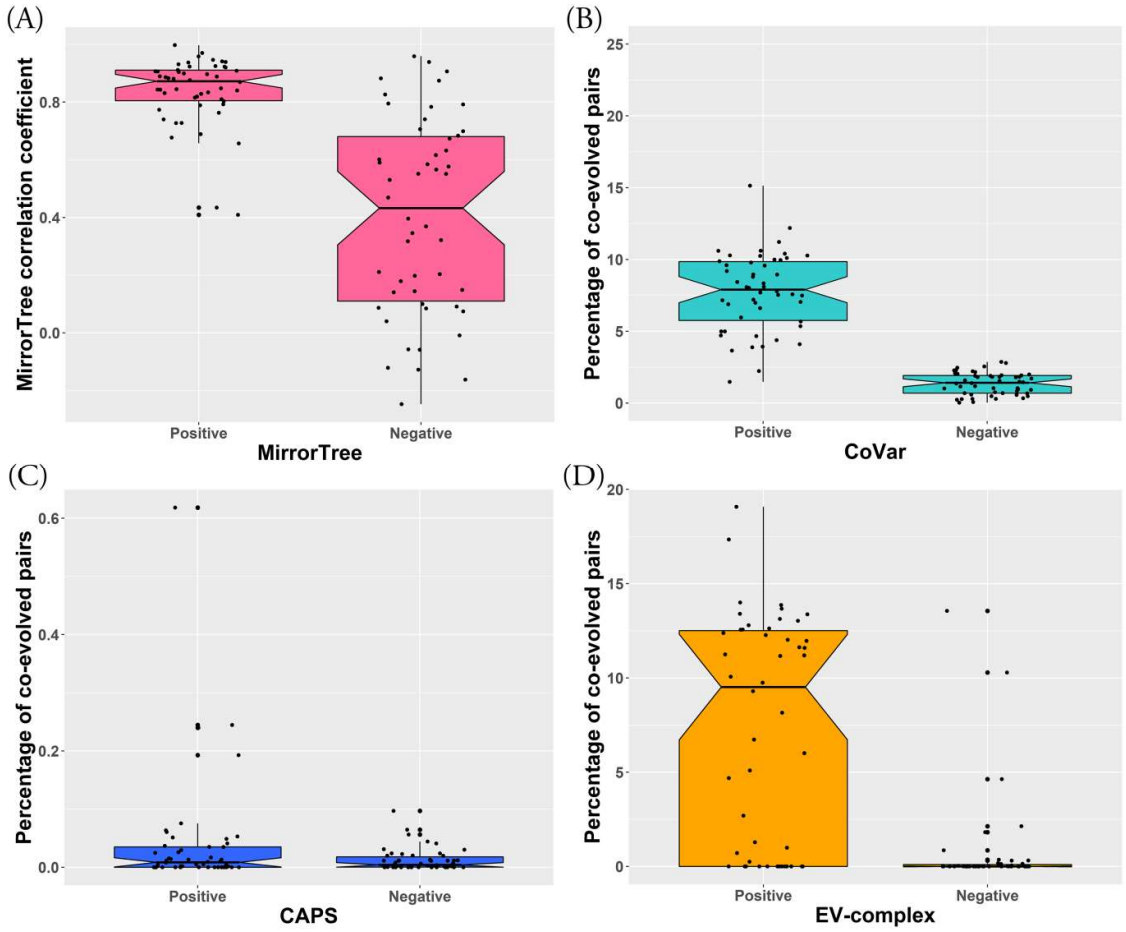


Figure 2.14: Studying inter-protein co-evolution with the Co-Var methodology. A set of interacting proteins (Positive) and a set of non-interacting proteins (Negative) were studied with the help of multiple programs available to study inter-protein co-evolution. (A) MirrorTree based analysis of positive and negative dataset to predict co-evolving and non co-evolving proteins (B) Inter-protein co-evolution analysis in the positive and negative dataset with the help of Co-Var (C) Co-evolution analysis of the positive and negative dataset in CAPS (D) Evolutionary coupling analysis (EVcomplex) to study co-evolution in the positive and negative dataset.

Table 2.8. Inter-protein co-evolution analysis in Co-Var and other inter-protein co-evolution analysis programs. Co-evolution in interacting and non-interacting proteins was studied with the help of Co-Var, MirrorTree, CAPS and EVcomplex. Statistics for student's t-test utilised to analyze the significance of the distribution for the calculated index of the positive and negative set is outlined.

	CoVar		**MirrorTree correlation coefficient**		**CAPS**		**EV-complex**	
	Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Mean	7.73	1.26	0.84	0.42	0.01	0.04	7.275	0.729
Standard Deviation	2.71	0.75	0.12	0.34	0.019	0.1	6.034	2.5
T-test statistics	t=16.2703 df=98		t = 8.2369 df = 98		t=2.0840 df=98		t = 6.8875 df = 96	
P-value	<= 0.0001		<= 0.0001		0.0398		<=0.0001	

2.1.2.2 Non-interface residue pairs which are not in close spatial proximity could also be co-evolving

The co-evolutionary pairings predicted based on the Co-Var methodology in the interacting protein complexes (Table S13) were further analyzed by determining inter-residue distances among them based on representative structures. Herein, it was observed that a large percentage of co-evolutionary pairings (nearly 70%) do not lie in close spatial proximity (>10 Å). This trend was consistently observed in the co-evolving pairs determined in these complexes utilising multiple inter-protein co-evolution analysis methodologies (Figure 2.15 A, C, D). Moreover, in general very few interface pairs (13.6% on average) tend to co-evolve in these protein-protein interaction complexes with varying rates among obligate and transient interaction complexes (Figure 2.15B, Table S13). Therefore, considering inter-residue distances among inter-protein co-evolved pairs it was found that some co-evolutionary pairings occur in close spatial proximity whereas a high proportion of pairs did not occur in close spatial proximity (Figure 2.15).

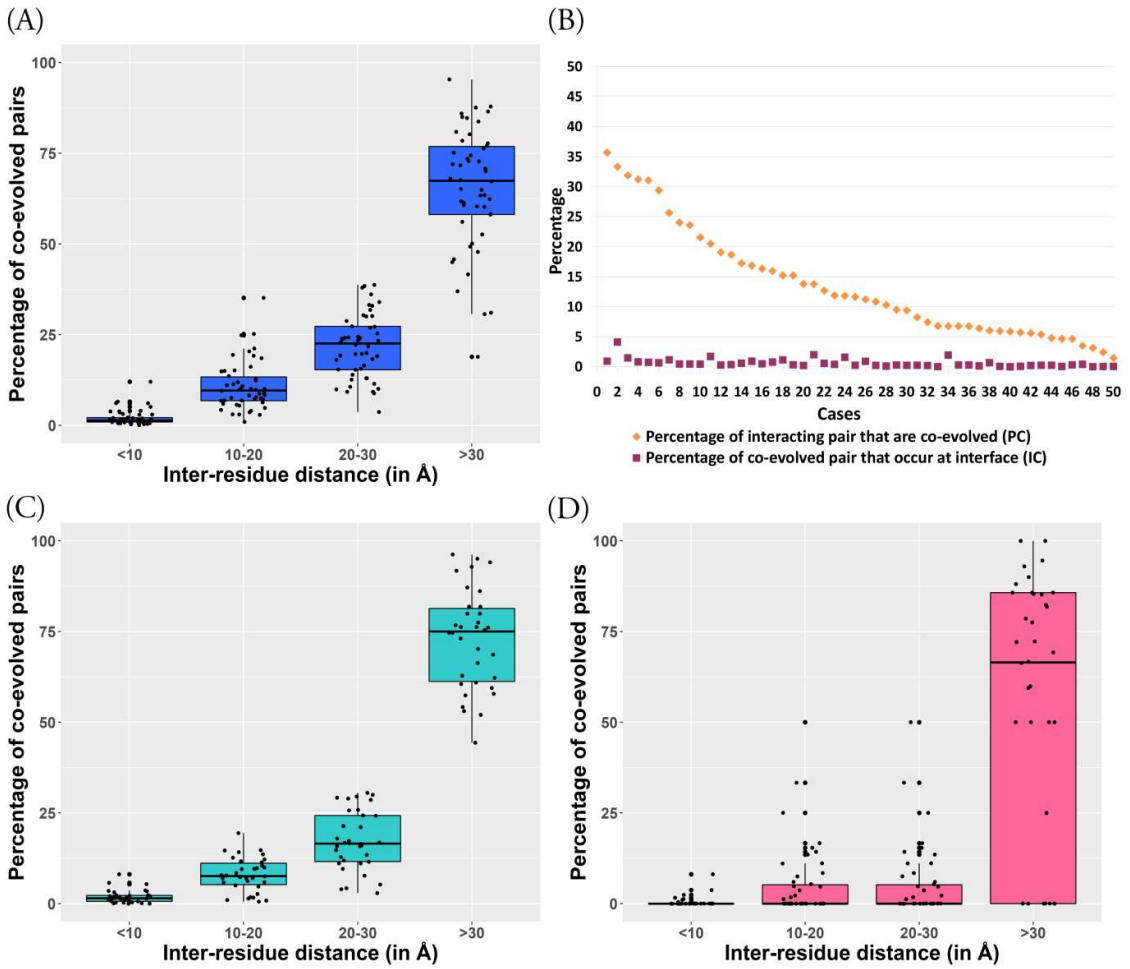


Figure 2.15 Inter-residue distances in co-evolutionary pairings among protein-protein interaction complexes. Co-evolutionary pairings were mapped onto a reference structure and to analyse whether the residues were in close spatial proximity inter-residue distances were determined. (A) Distribution of 'percentage of co-evolved pairs' predicted by Co-Var (B) Predicted co-evolved residue pairs that occur at the interface (IC) or interacting pairs that were found to co-evolve (PC) based on co-evolution analysis of positive set in Co-Var (C) Inter-residue distance distribution for EVcomplex predicted co-evolved pairs (D) Distance distribution analysis for CAPS based prediction of co-evolved pairs in interacting complexes.

2.1.2.3 Absence of co-ordinated changes at interface and non-interface co-evolving residues may lead to disruption of inter-cellular protein-protein interactions

Co-evolutionary pairings in protein complexes were found to occur in interface and non-interface regions and it would be interesting to study whether residue positions involved in these pairings have crucial structural or functional roles in these complexes. If this were the case then residue propensities at these residue positions and co-evolutionary pairings among them could be crucial for native interactions. Studies of

distribution pattern of disease associated variants have identified that disease causing mis-sense mutations frequently occur at the core region of protein-protein interaction interfaces or at active sites or ligand-binding sites (Gao et al., 2015; David and Sternberg, 2015). Based on this observation it can be stated that interaction interfaces are involved in determining complex functionality and absence of co-ordinated changes at interface residue positions are likely to contribute to altered interaction profiles and aberrant complex functionality. It is plausible that co-evolutionary pairings identified at the interface could indeed identify interface residues important for complex functionality in this manner. Further, residue positions in non-interface regions involved in inter-protein co-evolutionary pairings could also be functionally relevant and it would be interesting to note whether mis-sense mutations frequently occur at these positions wherein absence of co-ordinated changes at these positions could be functionally detrimental.

2.1.2.4 Co-evolutionary pairings predicted in hetero-dimeric protein complexes involved in inter-cellular interactions occur among functionally relevant residues

The likely structural or functional roles of co-evolving residue positions that do not lie in close spatial proximity in inter-protein interaction complexes has been explored in selected inter-cellular protein-protein interaction complexes that have roles in cancer metastasis. Co-evolutionary pairings were predicted in selected protein-protein interaction complex case studies and a large fraction of them did not occur at the interface (nearly 0.1-0.4% of the total co-evolved pairs lie at the interface) (Table 2.9). Further, only about 1.5-16% (mean= 5.56%) of the interface pairs were found to co-evolve in the complexes considered which is in correlation with the observation that transient interfaces exhibit a low degree of co-evolution (Mintseris et al., 2005).

Table 2.9: Co-evolutionary pairings in hetero-dimeric protein complexes.

	Reference PDB structure	Reference sequence (Protein family A)	Reference sequence (Protein family B)	Number of co-evolutionary pairings (z<=-1)	Percentage of interface pairs that are predicted to be co-evolved (PC)	Percentage of co-evolved pairs that occur at the interface (IC)
Case 1	2D9Q	CSF3 (P09919)	CSF3R (Q99062)	7065	16.883	0.184
Case 2	1MOX	EGFR (P00533)	TGFA (P01135)	3115	2.203	0.161
Case 3	1EVT	FGF1 (P05230)	FGFR1 (P11362)	3343	3.185	0.15

Case 4	1KTZ	TGFB3 (P10600)	TGFBR2 (P37173)	5634	5.66	0.053
Case 5	1NUN	FGF10 (O15520)	FGFR2 (P21802)	2402	1.429	0.125
Case 6	1DJS	FGF1 (P05230)	FGFR2 (P21802)	2140	4	0.374

Moreover, a large fraction of co-evolutionary connections between proteins involved in the inter-cellular interaction complexes did not occur in close spatial proximity and were found to occur among residues within interface regions and non-interface regions (Table 2.9, Table 2.10).

Co-evolutionary pairings in inter-protein interactions were predominantly identified within functional domain regions within each protein involved in the interaction complex (Figure 2.16). Further, in most of the inter-cellular protein interaction complex case studies analyzed, about 70-80% of the positions involved in co-evolutionary pairings were found to occur within functional domain regions of the interacting proteins (Table 2.11). This observation suggests that these residue positions may be biologically relevant for functional integrity of the complex.

Table 2.10: Co-evolutionary pairings in inter-cellular protein interaction complexes lie in interface and non-interface regions

	Reference PDB structure	**Reference sequence (Protein family A)**	**Reference sequence (Protein family B)**	**Reference structure mapped co-evolved positions**	**Inter-residue distances among co-evolved positions**			
					<=10 Å	**10-20Å**	**20-30Å**	**>30 Å**
Case 1	2D9Q	CSF3 (P09919)	CSF3R (Q99062)	3144	1.40	9.32	18.35	70.93
Case 2	1MOX	EGFR (P00533)	TGFA (P01135)	470	3.83	17.02	37.02	42.13
Case 3	1EVT	FGF1 (P05230)	FGFR1 (P11362)	1157	2.42	17.55	33.28	46.76
Case 4	1KTZ	TGFB3 (P10600)	TGFBR2 (P37173)	129	7.75	14.73	17.83	59.69
Case 5	1NUN	FGF10 (O15520)	FGFR2 (P21802)	1149	1.83	15.49	33.42	49.26
Case 6	1DJS	FGF1 (P05230)	FGFR2 (P21802)	831	3.73	17.57	33.45	45.25

2.1.2.5 High degree co-evolutionary pairings occur at interface and non-interface regions in inter-cellular protein interaction complexes

Analysis of molecular co-evolution in inter-protein complexes allowed the identification of co-evolutionary pairings in close spatial proximity (<7Å) and in non-interface regions (inter-residue distances>7Å). Herein, it has been observed that certain residue positions in one complex constituent exhibit a tendency to have a large number of co-evolutionary connections with positions in the interacting protein partner. For instance, in the CSF3-CSF3R interaction complex, co-evolutionary pairings were obtained between 68 out of total 207 residue positions in CSF3 and 187 out of 836 residue positions in CSF3R, respectively. Further, most co-evolutionary pairings were obtained between certain CSF3 (58) and CSF3R (68) residues only. Additionally, 7064 co-evolutionary pairings exist between residues in CSF3 and CSF3R complex in which 58 residue positions exhibit this tendency to have large number of co-evolutionary connections with 68 positions in the interacting protein partner (Figure 2.17A). Co-evolutionary connections occurring involving high degree co-evolved positions were found to occur at the interface as well as non-interface regions in this complex and these could be important for the interaction between these proteins (Figure 2.17B).

In particular, in the TGF-A and EGFR interaction complex, co-evolutionary pairings were obtained between 23 out of total 161 residue positions in TGF-A and 251 out of 1210 residue positions in EGFR, respectively. Further, most co-evolutionary pairings were obtained between certain TGF-A (9) and EGFR (91) residues. Moreover, 3114 co-evolutionary pairing combinations exist between residues in TGF-A and EGFR complex among these residue positions that exhibit this tendency to have large number of co-evolutionary connections in the interacting protein partner (Figure 2.18A). However, co-evolutionary connections among high degree residue positions occur at the interface as well as non-interface regions in this protein pair as well (Figure 2.18B).

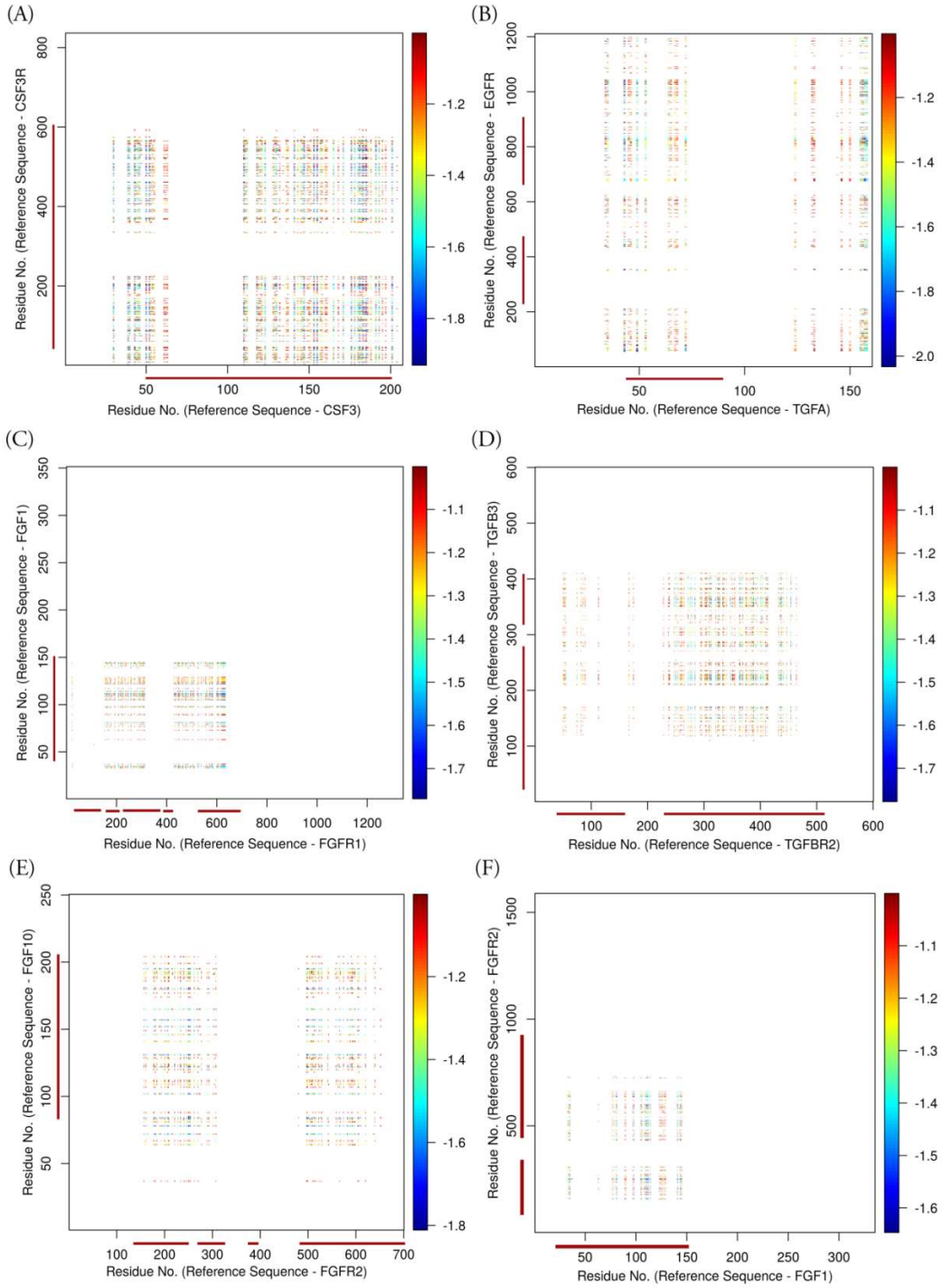


Figure 2.16: Co-evolutionary pairings in hetero-dimeric protein interaction complexes. (A)Predicted co-varying positions (Z-score<=-1) observed between CSF3 and CSF3R. (B) Predicted co-varying positions (Z-score<=-1) in TGFA and EGFR complex (C) Predicted co-varying positions (Z-score <= -1) occurring between FGF1 and FGFR1 (D) Predicted co-varying positions (Z-score <= -1) in TGFBR2 andTGFB3 complex (E) Predicted co-varying positions (Z-score<=-1) observed between FGF10 and FGFR2 (F) Predicted co-varying positions (Z-score<=-1) in FGF1 and FGFR2 complex.

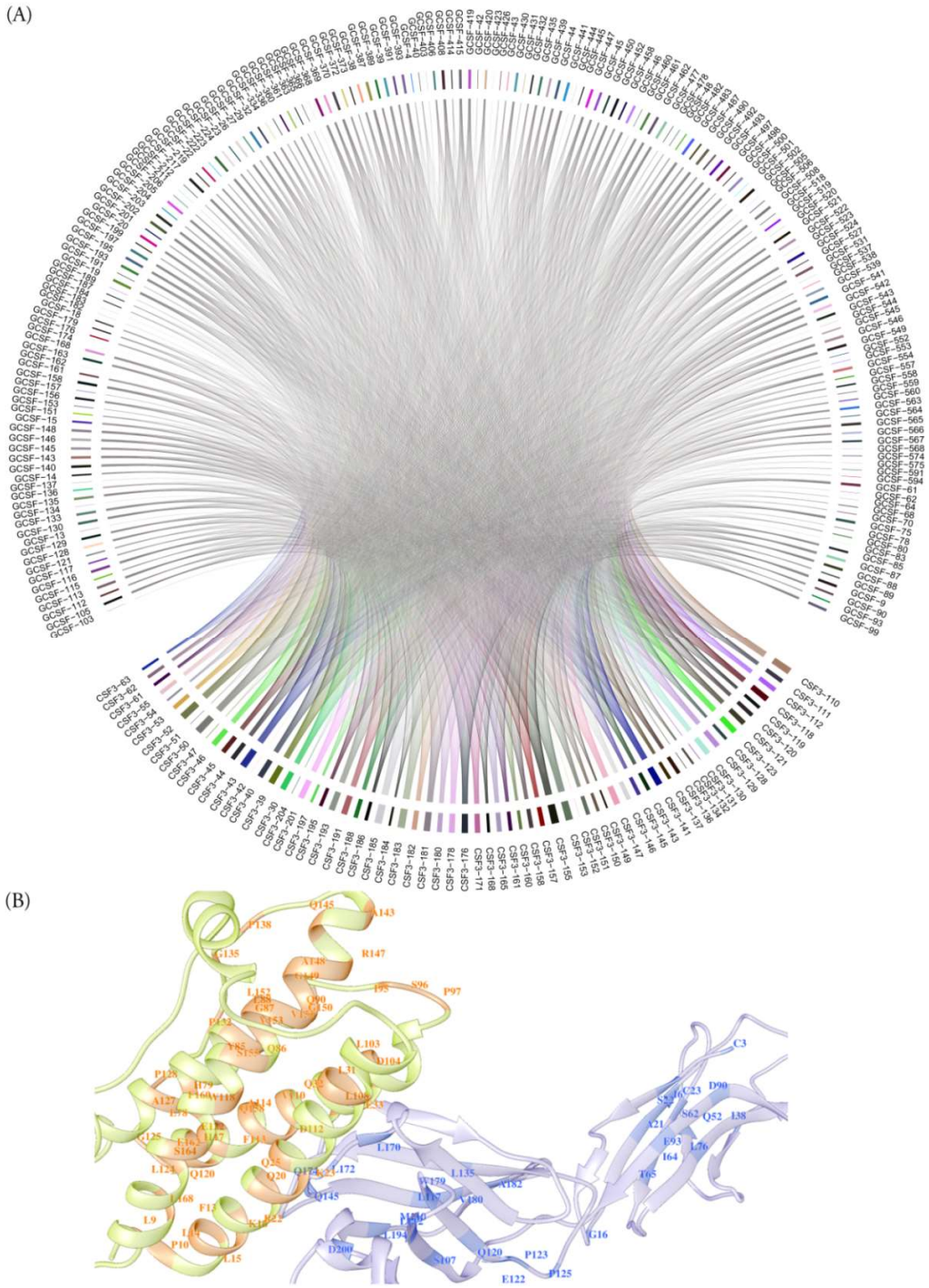


Figure 2.17: High degree co-evolved positions in CSF3-CSF3R inter-cellular protein interaction complex. (A) Co-evolving residues in CSF3 (GCSF) and CSF3R (GCSFR) that tend to have a large number of co-evolutionary connections (high degree co-evolved positions) or pairings among them are shown here. (B) High degree co-evolved positions that lie in spatially proximal and distal regions are depicted on the reference structure (PDB ID: 2D9Q). In the structural representation of co-evolved positions, CSF3 chain: light green, CSF3R chain: light blue.

Further, in the TGFB3 and TGFBR2 interaction complex, co-evolutionary pairings were obtained between 104 out of total 412 residue positions in TGFB3, and 95 out of 567 residue positions in TGFBR2 respectively. Further, most co-evolutionary pairings were obtained between certain TGFB3 (53) and TGFBR2 (45) residues. Moreover, 5491 co-evolutionary pairing combinations exist between residues in TGFB3 and TGFBR2 complex among these residue positions that exhibit this tendency to have large number of co-evolutionary connections in the interacting protein partner (Figure 2.19A). Additionally, co-evolutionary connections were noted among high degree residue positions at the interface and non-interface regions in this protein pair as well (Figure 2.19B).

Table 2.11: High degree co-evolved positions in inter-protein interaction complexes.

	Protein A	**Protein B**	**Number of co-evolved pairs (z<=-1)**	**Number of co-evolved pairs in functional domains**	**Co-evolved pairs in functional domains (%)**	**High Degree Co-Evolved Position pairs**	**High Degree co-evolved positions with mutations**	**High Degree co-evolved as mutated (%)**
Case 1	CSF3 (P09919)	CSF3R (Q99062)	7065	5563	78.74	7064	3204	45.36
Case 2	EGFR (P00533)	TGFA (P01135)	3115	1020	32.74	3114	1955	62.78
Case 3	FGF1 (P05230)	FGFR1 (P11362)	3343	2655	79.42	3342	1791	53.59
Case 4	TGFB3 (P10600)	TGFBR2 (P37173)	5634	4757	84.43	5491	3059	55.71
Case 5	FGF10 (O15520	FGFR2 (P21802)	2402	1890	78.68	2387	1261	52.83
Case 6	FGF1 (P05230)	FGFR2 (P21802)	2140	1641	76.68	2138	1270	59.40

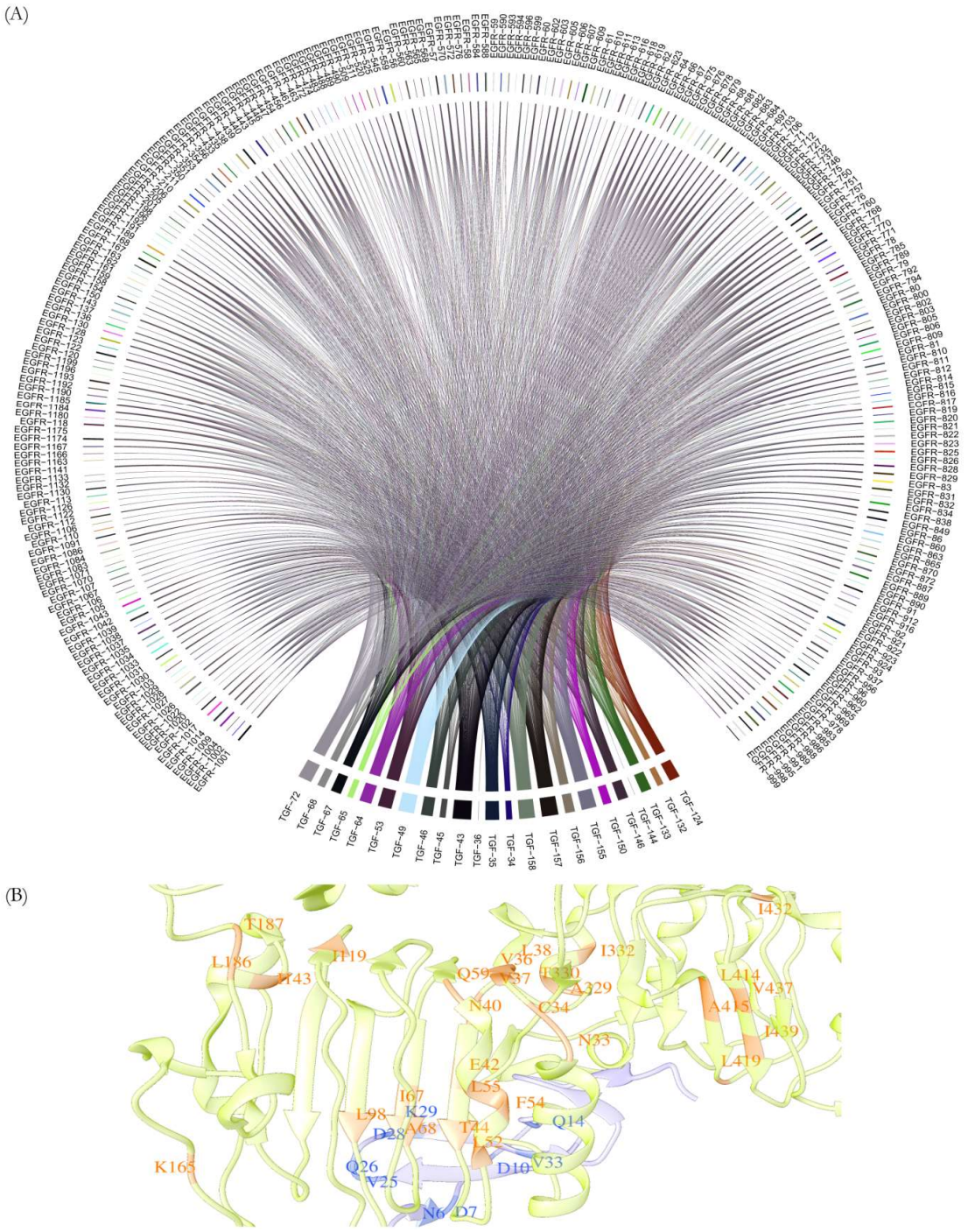


Figure 2.18: High degree co-evolved positions in TGF-A and EGFR protein interaction complex. (A) Co-evolving residues in TGF-A and EGFR that tend to have a large number of co-evolutionary connections (High degree co-evolved positions) or pairings among them are shown here. (B) High degree co-evolved positions have been represented on the reference structure (PDB ID: 1MOX). Note: In the structural representation, EGFR chain: light green and TGFA chain: light blue.

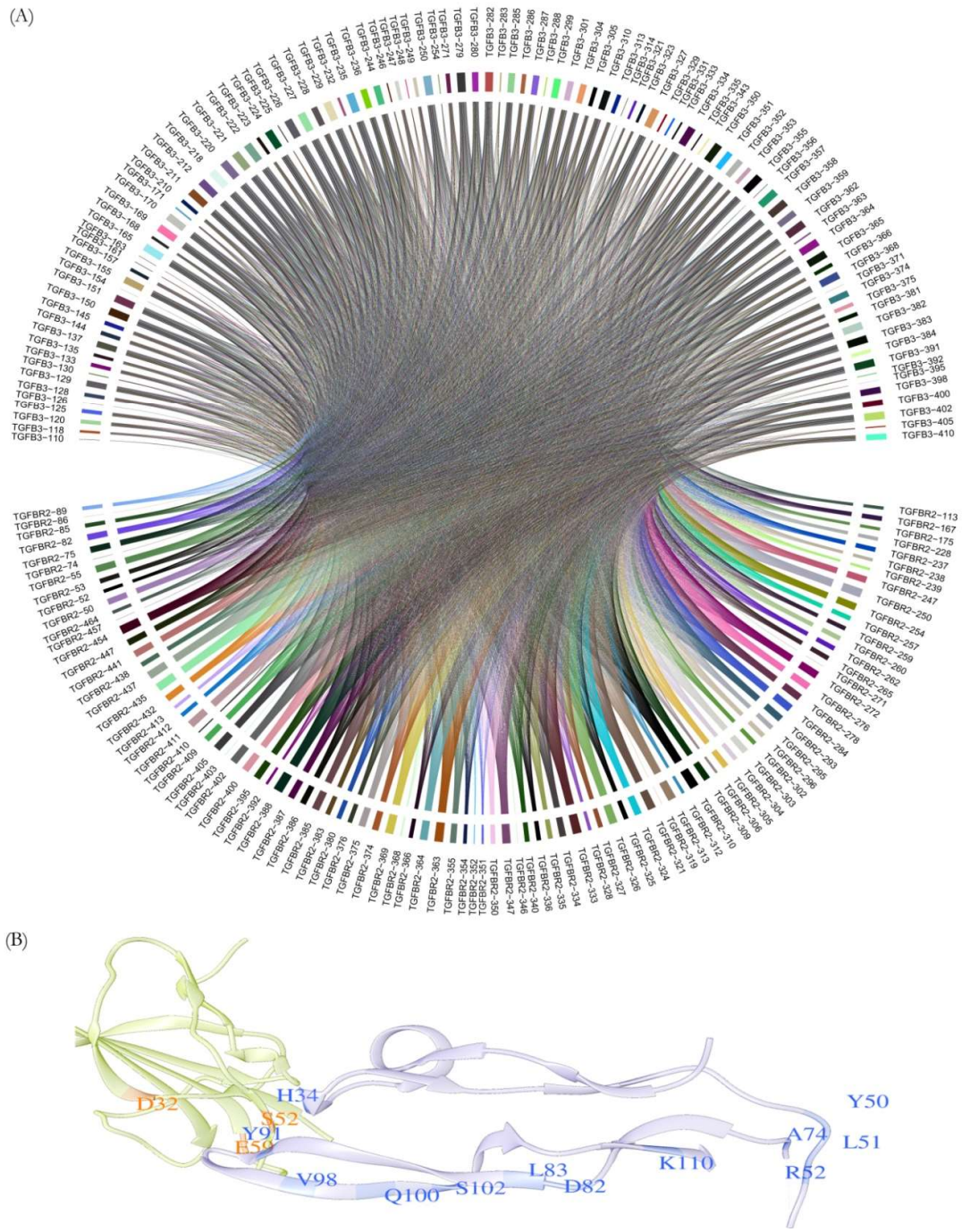


Figure 2.19: Predicted high degree co-evolved positions in inter-protein interaction complex between TGFB3 and TGFBR2. (A) Co-evolving residues in TGFB3 and TGFBR2 complex that share a large number of co-evolutionary pairings among them (High degree co-evolved positions) are depicted here. (B) High degree co-evolved positions that occur in spatially proximal and distal regions have been depicted on the reference structure (PDB ID: 1KTZ). Note: In the structural representation, TGFBR2 chain: light green while TGFB3 chain: light blue.

Similarly, the co-evolutionary pairings obtained in the other inter-cellular interaction complexes studied namely FGF1 and FGFR1, FGF10 and FGFR2, FGF1 and FGFR2 also exhibit this trend, wherein certain residue positions exhibit a tendency to have larger number of co-evolutionary connections with positions in the interacting protein partner (Figures S3-5). Moreover, these residue positions involved in these co-evolutionary pairings occur in interface (spatially proximal<7Å) and non-interface (spatially distal>7Å) regions (Figure 2.20). Interestingly, a large fraction of residues among the high degree co-evolved positions or residue positions with large number of co-evolutionary connections are frequently prone to substitution mutations prevalent in cancer (Table 2.11).

2.1.2.6 Residue positions important involved in intra-protein stability or function may additionally influence inter-protein interactions

Intra-molecular co-evolution analysis has identified that residues in close proximity important for structural stability and distant sites having a functional dependence could be highly co-evolving (Fares and Travers 2006, Chakrabarti and Panchenko 2009). Herein, intra-molecular co-evolution was studied in a set of proteins to identify the pattern of inter-residue distances among Co-Var predicted intra-protein co-evolved positions. An initial comparison of inter-residue distances among predicted co-evolving residues within proteins utilising Co-Var, CAPS, MI and PsiCov suggested that Co-Var may be utilised to study intra-protein co-evolution. This is because it predicts a relatively higher percentage of co-evolved pairs occurring in close proximity in comparison to the other existing methodologies (Figure 2.21). However, the trend that intra-protein co-evolved positions occur in close proximity and distant regions was observed in the intra-molecular co-evolution analysis of 252 CDD protein families utilising all the methodologies (Figure 2.21). Therefore, considering inter-residue distances among intra-protein co-evolved pairs we found that a higher proportion of pairs occur in close proximity whereas inter-protein co-evolved pairs had a significant fraction of co-evolved pairs that do not lie in close spatial proximity (Figure 2.15A, Figure 2.21A). Subsequently, intra-molecular co-evolution in each constituent protein of the inter-cellular complex case studies was performed. This analysis suggested that residue positions important for intra-protein stability or function may additionally influence inter-protein co-evolution interactions as well (Table 2.12). Further, positions important for protein stability or function within a protein were also found to be involved in forming extensive high degree co-evolutionary pairings in inter-protein interactions as well.

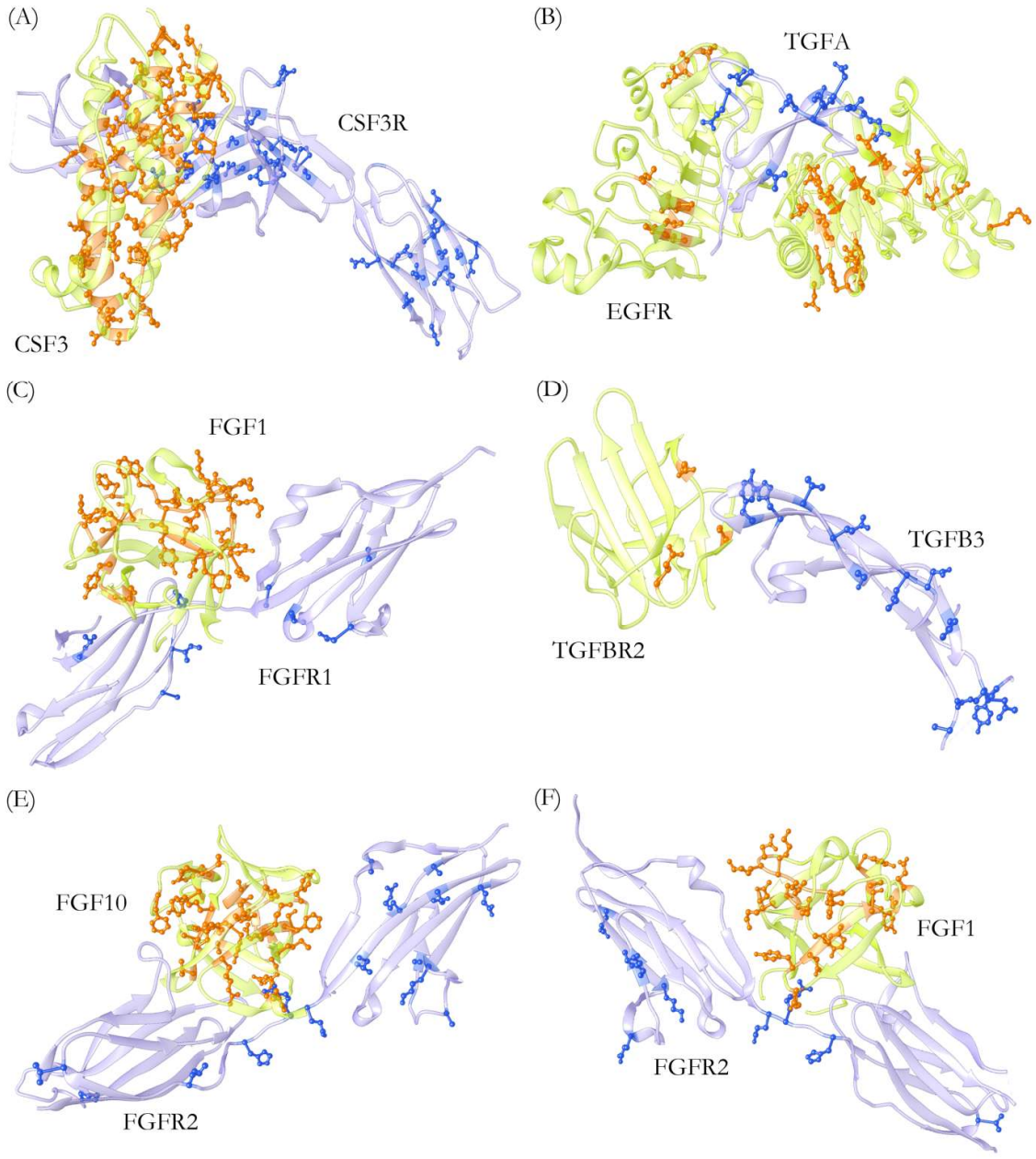


Figure 2.20: High degree co-evolved positions observed in interface and non-interface regions. High degree co-evolved positions predicted in inter-cellular protein interaction complexes have been depicted on respective reference structures. High degree co-evolved positions in (A) CSF3-CSF3R [PDB ID: 2D9Q] (B) TGFA and EGFR [PDB ID: 1MOX] (C) FGF1 and FGFR1 [PDB ID: 1EVT] (D) TGFBR2 and TGFB3 [PDB ID: 1KTZ] (E) FGF10 and FGFR2 [PDB ID: 1NUN] (F) FGF1 and FGFR2 complex [PDB ID: 1DJS], respectively are shown in ball and stick models.

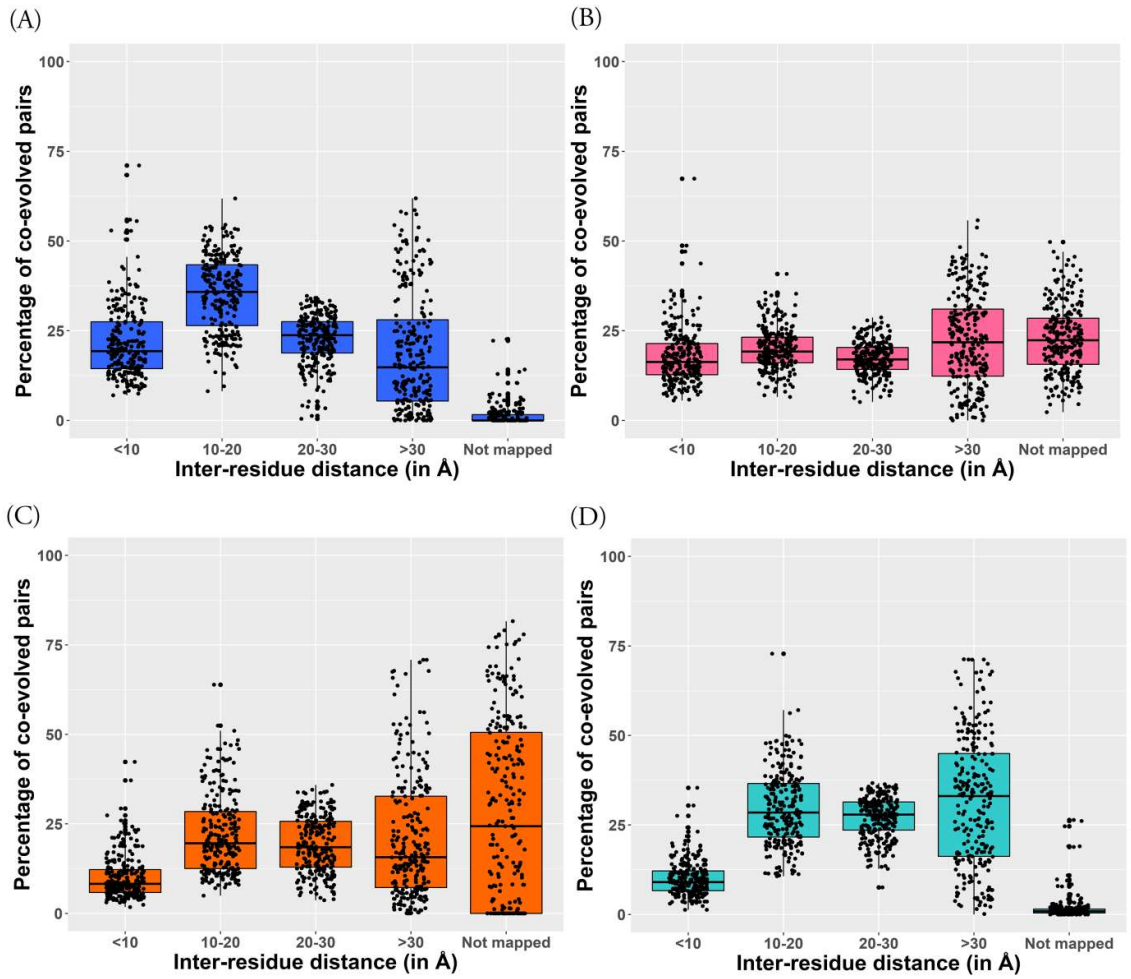


Figure 2.21: Studying intra-protein co-evolution utilising Co-Var. Distance distribution analysis of co-evolved positions predicted in intra-protein interaction complexes was considered to evaluate the performance of Co-Var against existing methods for studying intra-protein co-evolution. (A) Distance distribution analysis of intra-protein co-evolved pairs based on predictions from Co-Var (B) Inter-residue distance among predicted co-evolved pairs within proteins based on CAPS (Fares and Travers 2006) (C) Distance distribution analysis considering intra-protein co-evolving pairs determined utilizing mutual information (Korber et al, 1993) (D) Inter-residue distance among predicted co-evolved pairs within proteins utilising PsiCov (Jones et al., 2011).

Table 2.12: Intra-protein co-evolving positions may also be predicted as high degree inter-protein co-evolved positions (CP) in proteins constituting a complex.

	Reference sequence (Protein family A)	Reference sequence (Protein family B)	No. of intra-protein CP (A)	No. of intra-protein CP (B)	No. of inter-protein CP (A)	No. of Inter-protein CP (B)	Inter-protein and intra-protein CP (A)	Inter-protein and intra-protein CP (B)
Case 1	CSF3 (P09919)	CSF3R (Q99062)	79	272	68	187	58	67
Case 2	EGFR (P00533)	TGFA (P01135)	40	120	23	251	37	13
Case 3	FGF1 (P05230)	FGFR1 (P11362)	56	166	36	143	34	23
Case 4	TGFB3 (P10600)	TGFBR2 (P37173)	116	161	104	95	53	45
Case 5	FGF10 (O15520)	FGFR2 (P21802)	85	107	42	88	34	28
Case 6	FGF1 (P05230)	FGFR2 (P21802)	55	171	33	101	28	21

2.1.2.7 Disease associated mutations in inter-cellular protein-protein interaction complexes involved in cancer metastasis may be predicted utilising Co-Var

In order to ascertain whether the high degree co-evolutionary pairings have a biological significance, the functional relevance of the predicted co-evolutionary pairings has been studied by performing mutation analysis utilizing data from the COSMIC database (Tate et al., 2018). By determining whether residue positions identified in high degree co-evolutionary pairings important from the standpoint of inter-molecular co-evolution tend to exhibit frequent substitution mutations in disease conditions such as cancer. Interestingly, it was observed that the high degree co-evolved positions or the residue positions with large number of co-evolutionary connections is that a large fraction among them are frequently prone to substitution mutations prevalent in cancer (Table 2.11). Moreover, the residue pairing propensity at the high degree co-evolved positions varies substantially in mutated protein complexes wherein pairs containing amino acids such as glycine, proline, aspartate, glutamate, tryptophan, tyrosine, histidine and glutamine are more frequent (Figure 2.22). Such substitutions or alterations in pairing

propensity at positions that tend to co-evolve are likely to have deleterious functional characteristics in the absence of coordinated compensatory changes. A similar trend is observed in most of the intercellular protein interaction complexes considered herein where the residue pairing propensity among the co-evolved positions gets altered frequently as a result of mutations under diseased conditions (Table 2.11, Figure 2.22). Based on these observations it can be postulated that high degree co-evolved residue positions could be frequently mutated in cancers and as such changes at these residue positions may not always be compensatory changes which may result in a perturbed interaction between these proteins. This finding suggested that residue positions involved in high degree co-evolutionary pairings may frequently show substitution mutations in cancer conditions which may result in alterations in complex functionality and such positions could potentially be important for conservation of the functional interaction between the interacting proteins. Therefore, with the help of the Co-Var methodology we can predict residue positions in interacting proteins at interface and non-interface positions but are likely to be functionally relevant or important for an interaction to occur between the proteins involved in an inter-cellular interaction. Further, absence of co-ordinated changes at these at interface and non-interface co-evolving residue positions may lead to disruption of inter-cellular protein-protein interactions, thus such alterations could be disease associated.

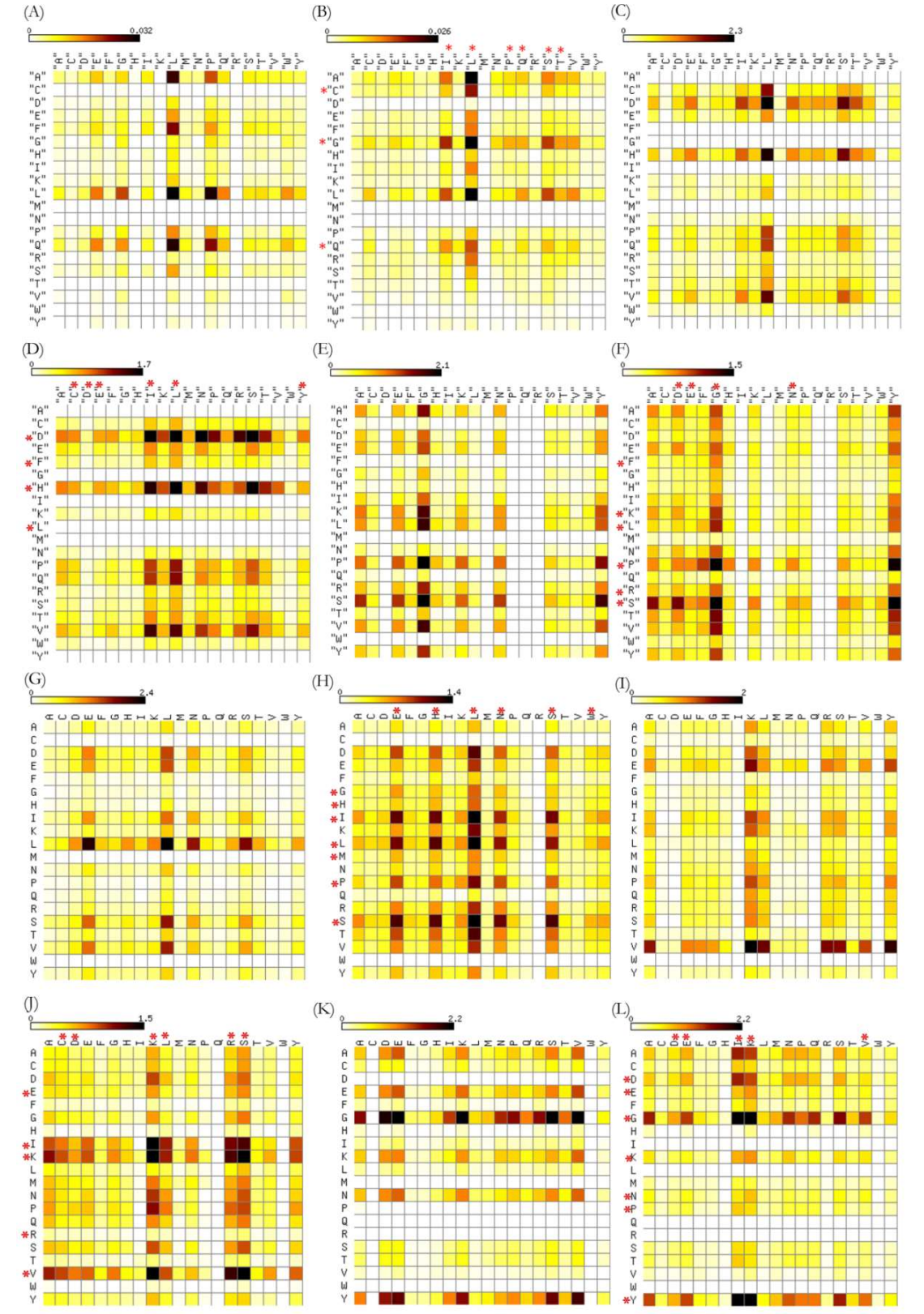


Figure 2.22: **Predicted high degree co-evolved positions may be functionally**

relevant in protein-protein interactions. Residue pairing propensity at the high degree co-evolved positions in reference protein sequences of the complex and the altered pairing propensity at these pairings based on the observed substitution mutations have been compared. (A) Pairing propensity in native CSF3-CSF3R complex (B) Altered pairing propensity in CSF3-CSF3R complex (C) Pairing propensity in native TGFA-EGFR complex (D) Altered pairing propensity in mutated TGF-EGFR complex (E) Pairing propensity in native FGF1-FGFR1 complex (F) Altered pairing propensity in mutated FGF1-FGFR1 complex (G) Pairing propensity in native TGFBR2-TGFB3 complex (H) Altered pairing propensity in mutated TGFBR2-TGFB3 complex (I) Pairing propensity in native FGF10-FGFR2 complex (J) Altered pairing propensity in mutated FGF10-FGFR2 complex (K) Pairing propensity in native FGF1-FGFR2 complex (L) Altered pairing propensity in mutated FGF1-FGFR2 complex.

2.1.2.8 Co-Var web server for studying inter-molecular co-evolution

A web server for analyzing inter-molecular co-evolution is available online at http://www.hpppi.iicb.res.in/ishi/covar/index.html (Figure 2.23). During the inter-protein co-evolution analysis, co-evolutionary pairings are determined based on the Co-Var methodology and reference sequence mapped co-evolving positions are reported. The Co-Var score and Z-score for threshold selected co-evolutionary pairings are depicted with the help of a surface plot representation. High degree co-evolved positions and/or co-evolved positions in close spatial proximity are displayed on the reference structure provided and the list of co-evolutionary pairings in close spatial proximity may be downloaded. Additionally, a distance distribution plot of the inter-residue distances among residue positions involved in co-evolutionary pairings is provided and can be utilized to get an idea about whether co-evolutionary pairings occur in close proximity or among spatially distant residue positions. The final results of the analysis are mailed to the e-mail address provided and can be easily available for downloaded for further analysis.

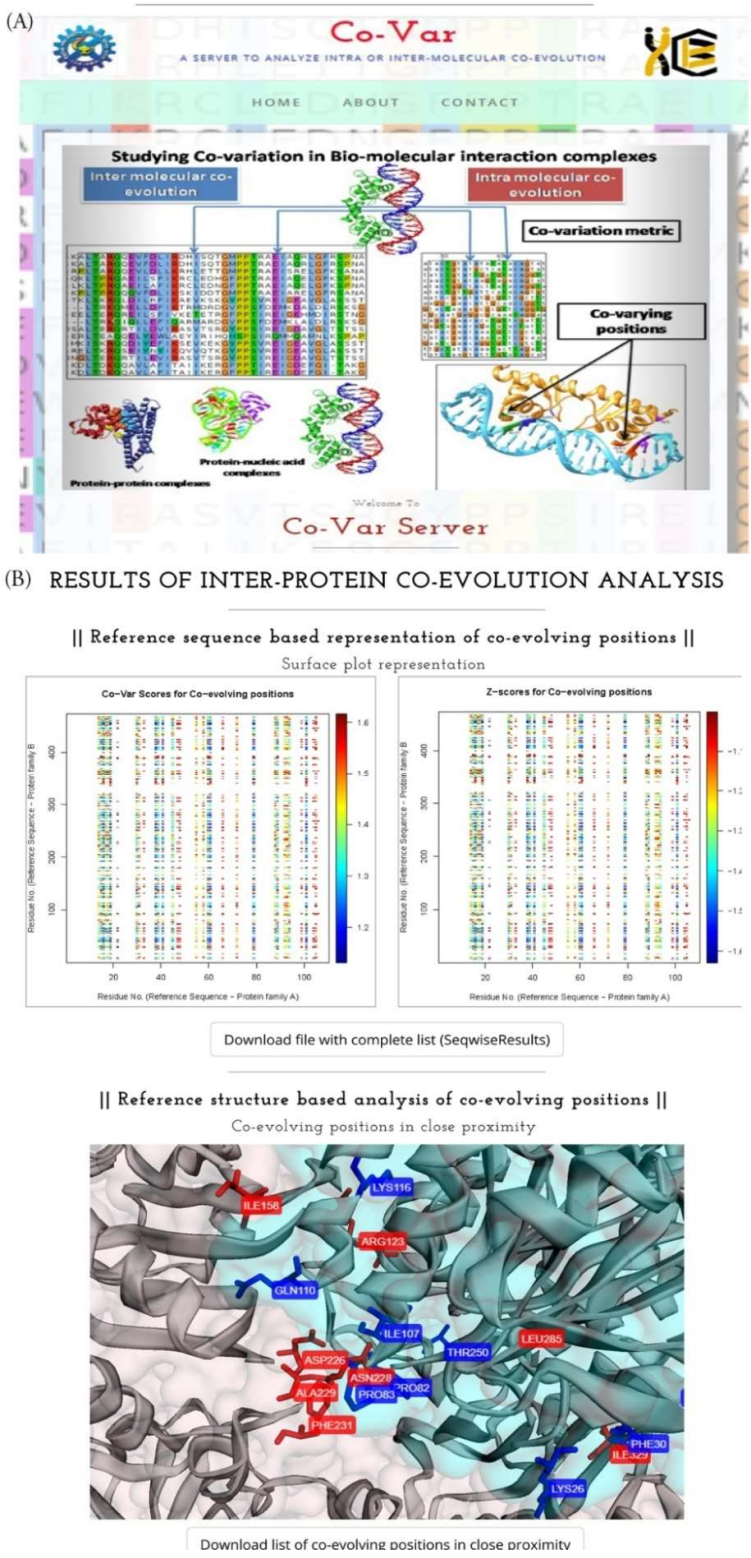


Figure 2.23: **Co-Var web server to study inter-protein co-evolution.** (A) User interface to Co-Var web-server (B) Sample inter-protein co-evolution analysis results provided by Co-Var web-server

Conclusion

During this analysis, a method named Co-Var (Correlated Variation) has been developed to determine inter-molecular residues likely to carry out important structural or functional roles in protein-protein interactions. Initially, the applicability of the Co-Var measure in studying inter-protein co-evolution has been verified and consequently it has been utilised it to study co-evolution in certain ligand-receptor proteins involved in inter-cellular interactions. The assumption herein was that co-evolutionary pairings may occur in interface and non-interface regions and such pairings could be crucial for native interactions. However, absence of co-ordinated changes at these positions might contribute to altered interaction profiles and aberrant complex functionality. Analysis of molecular co-evolution in inter-protein complexes may be useful for determining co-evolutionary pairings among interface residues that are likely to exhibit co-ordinated changes as a result of evolutionary pressure and such positions are likely to be crucial for a functional interaction between these sets of proteins. Co-evolutionary pairings in inter-protein interaction complexes were found to occur in close spatial proximity (<7Å) as well as in non-interface regions (inter-residue distances>7Å), an observation concordant with previous studies (Anishchenko et al., 2017; Marino Buslje et al., 2010). The likely structural or functional roles of these co-evolutionary pairings that do not lie in close spatial proximity has been studied in detail in selected inter-cellular protein-protein interaction complexes having roles in cancer metastasis. Herein, co-evolutionary pairings were identified within interface or non-interface regions; mainly occur among residues present in functional domains or residues that have a role in intra-protein co-evolution in individual constituents of the complex. Additionally, certain receptor positions exhibited a tendency to share many co-evolutionary pairing connections with certain ligand positions only. Such high degree co-evolving positions are likely to show frequent substitution mutations in cancer and as such absence of co-ordinated changes at these positions may contribute to disease associated altered interaction in these complexes. Thus, utilising these inter-cellular protein interaction complex as case studies, it was identified that co-evolutionary pairings that are likely to be crucial for functional interactions between these proteins may be predicted utilising the Co-Var methodology. Additionally, the observation that these residue positions exhibit mis-sense mutations in cancer provides an indication that modifications at these positions may result in altered protein interactions in disease conditions. Additionally, a web server named Co-Var has been developed which is freely accessible at http://www.hpppi.iicb.res.in/ishi/covar/index.html for inter-protein co-evolution analysis.

Inference

Co-evolution analysis utilising the Co-Var measure may allow one to determine inter-dependent residue pairs for an inter-protein interaction based on information contained in sequences. Such co-evolutionary pairings may occur among residues in close proximity or in distant regions in the inter-protein interaction complexes could be biologically or functionally relevant.

2.2 Methodology

2.2.1 Determining virulence factor proteins likely to be involved in inter-cellular interactions with the help of sequence analysis measures

Rigorous homology searching methods including Hidden Markov Model (HMM) and BLASTp based searches may be employed to detect remote but significant similarities between existing virulence factors and uncharacterized proteins. Herein, in order to identify novel virulence factors in Eukaryotic parasitic proteomes particularly among *Leishmania spp.*, initially, a representative proteome (*L. donovani* sequences) has been compared with sequences in the virulence factors database (VFDB) (Chen et al., 2011). This approximation may be considered since comparison among *L. donovani* genome and other *Leishmania* species has shown genetic differentiation exists at the species level between *L. donovani* and *L. infantum* (Downing et al., 2011). Further, genome comparisons among *L. major*, *L. infantum* and *L. braziliensis* demonstrated great conservation of synteny among the genomes with only small number of species specific genes that are differentially distributed (Peacock et al., 2007). A strategy comprising of profile-profile, profile-sequence and sequence-sequence comparisons termed as the 'forward search analysis' and 'reverse search analysis' has been utilized herein to identify a class of virulent proteins. Subsequently, the Kinetoplastida proteome was analyzed to identify bacterial virulence factor-like proteins therein. Additionally, structural modeling and docking studies were performed to determine the host protein(s) that are likely to interact with this class of virulence factor proteins in *Leishmania spp.*, by studying the *L. donovani* bacterial virulence factor-like proteins in detail.

2.2.1.1 Identifying whether bacterial virulence factor-like proteins are prevalent in eukaryotic parasites

In order to identify virulence factors in *L. donovani*, utilizing TriTrypDB (release 8) (Aslett et al., 2009) *L. donovani* protein sequences (8,083) were obtained and compared with the virulence factors database (VFDB) (Chen et al., 2011) with the help of BLASTp algorithm (Basic Local Alignment Search Tool) (Altschul et al., 1990). Possible bacterial virulence factor-like proteins within *L. donovani* proteome were determined considering a threshold of E-value <= $1e^{-08}$, sequence identity >= 20%, query protein length coverage >= 25% and subject protein length coverage >= 40%. During this analysis, it was determined whether certain *L. donovani* proteins had reasonable similarity with bacterial virulence factors. Subsequently, further studies (Figure 2.24) as outlined below were performed to determine the extent of this similarity.

Forward search analysis: During the forward search analysis, the known virulence factor identified previously was considered as the query and either a BLAST (Altschul et al., 1990) or hidden markov model (HMM) profile based search (Söding et al., 2005; Remmert et al., 2012) was performed against the *L. donovani* proteome.

Reverse search analysis: During the reverse search analysis, the *L. donovani* proteins that demonstrated similarity with bacterial virulence factors via forward search approaches were further compared with sequence and profile databases utilizing sequence-sequence

and sequence-profile comparison tools.

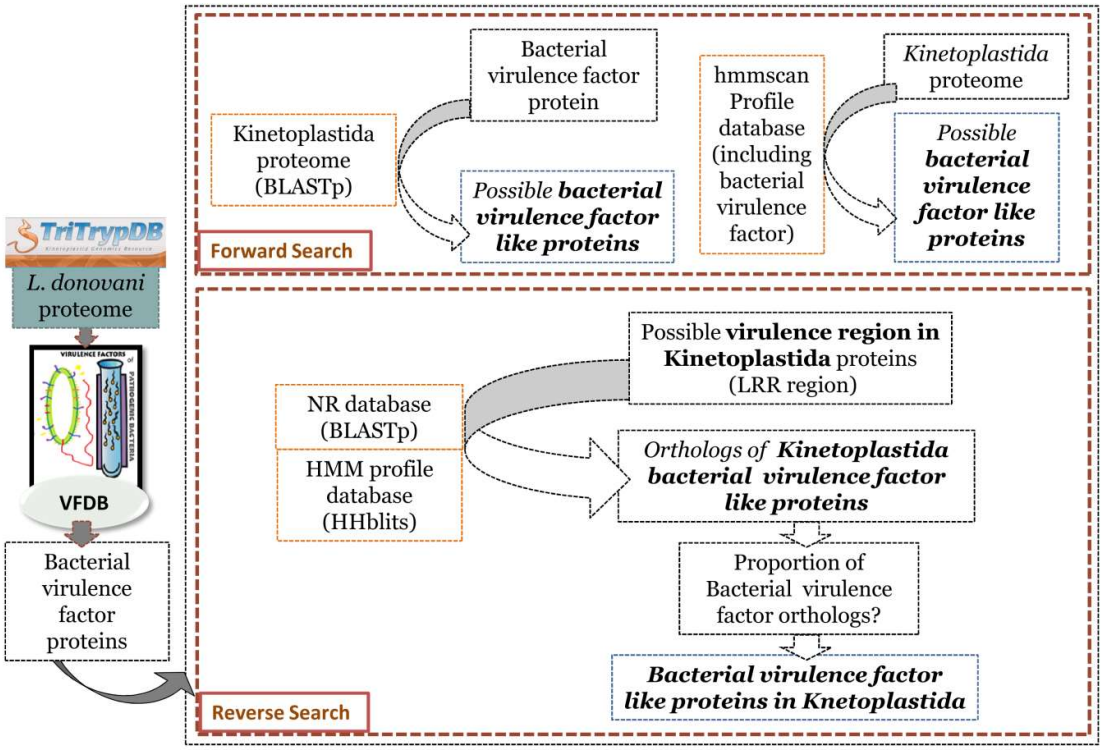


Figure 2.24: **Basic methodology for identification of bacterial virulence factor proteins in *L. donovani*.** A search strategy comprised of 'forward search' and 'reverse search' analysis was utilised to identify remote orthology between bacterial virulence factors and candidate *L. donovani* proteins. Abbreviations: Inl-A, Internalin-A; LRR, Leucine Rich Repeat; HMM, Hidden Markov Model; BLAST (Altschul et al., 1990), Basic Local Alignment Search Tool; HHblits, HMM-HMM-based lightning-fast iterative sequence search tool (Remmert et al., 2012).

2.2.1.2 Determining remote bacterial virulence factor orthologs in Kinetoplastida

Identification of remote bacterial virulence factor orthologs in Kinetoplastida was done with the help of a 'forward search' analysis and a 'reverse search' analysis. Herein, the 'forward search' comprised of a search of the Kinetoplastida proteome utilizing Internalin-A (*Listeria monocytogenes*) LRR region as query in BLAST (Altschul et al., 1990) and selected sequences (E-value<=$1e^{-10}$, query coverage>=60%) were considered as probable Inl-A-like proteins. Further, the 'forward search' analysis was complemented with a 'reverse search' analysis involving profile-profile comparisons to specifically obtain a set of remote bacterial Inl-A orthologs. An HHblits (Remmert et al., 2012) analysis was performed against the UniProt and PDB database to identify whether the subset of probable Inl-A-like Kinetoplastida proteins shared similarities with bacterial Inl-A-like sequences based on the ranks of listerial Inl-A proteins obtained in the 'reverse search' analysis (Figure 2.25).

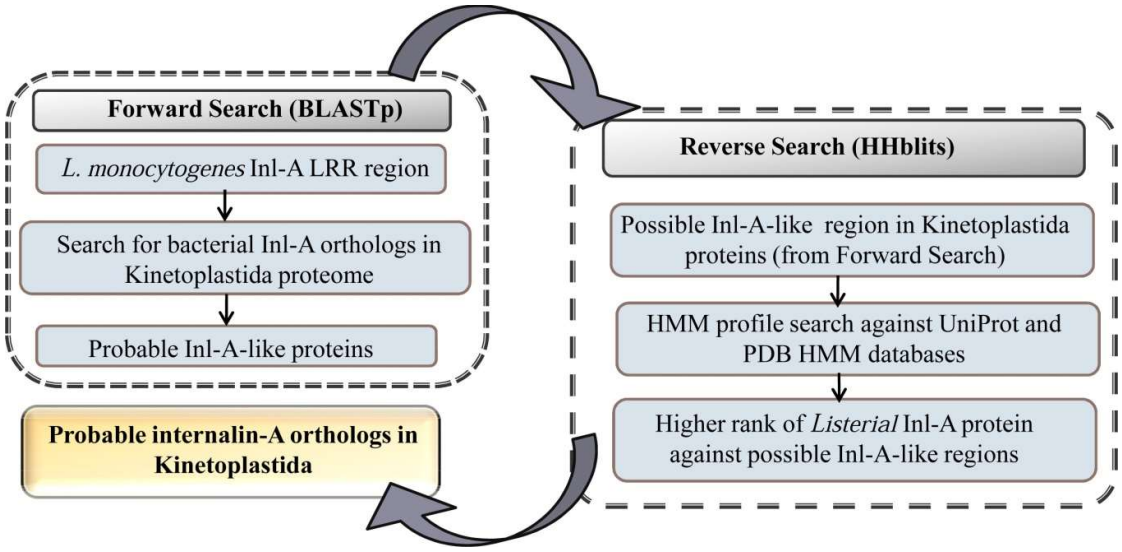


Figure 2.25: **Methodology for identification of remote bacterial protein orthologs in Kinetoplastida.** 'Forward search' and 'reverse search' analysis strategies employed to identify Inl-A-like proteins in Kinetoplastida are outlined here.

Identification of virulence factors in L. donovani

Briefly, the initial search against VFDB identified that certain *L. donovani* proteins had reasonable similarity with bacterial virulence factor, Internalin-A (Inl-A) suggesting the probability that some *L. donovani* proteins could be distant orthologs of Inl-A. The forward search analysis (Figure 2.26) consisting of a BLASTp search was performed against the *L. donovani* using the virulence factor domain (*L. monocytogenes* Inl-A [UniProt ID: P0DJM0] LRR region [77-405]) as query (Altschul et al., 1990). Protein segments that satisfied the threshold E-value <= $1e^{-10}$, query length coverage >=60% and sequence identity>=20% with Inl-A LRR region were assumed to be similar to the Inl-A domain. Close homologs of Inl-A were determined via BLASTp search (homolog selection criteria: E- value<=1e-06, query length coverage >=70%, sequence identity >=40%) against all bacterial genome databases to prepare a hidden Markov model (HMM) profile representative of the Inl-A virulence domain (Altschul et al., 1990; Tatusova, Ciufo, Fedorov, O'Neill, & Tolstoy, 2013). Additionally, representative non-redundant (<95%identical) sequences were selected for further analysis and the Inl-A-like region (10 residues) was extracted from these bacterial orthologs were determined for further analysis (Li, Jaroszewski, & Godzik, 2001, 2002). Utilizing an iterative refinement method (MAFFT- L-INS-I version 7) (Katoh, Misawa, Kuma, & Miyata, 2002; Katoh & Standley, 2013) which gives preference to local alignment Inl-A and its bacterial orthologs (81 sequences) were aligned such that the conserved LRR motif region is well aligned. The multiple sequence alignment (MSA) was then utilized to prepare an HMM profile via HMMER 3.0 (Eddy, 1998, 2009). The prepared Inl-A profile was added to the Pfam database (Finn et al., 2013) to prepare a modified Pfam-A database which was then used as the profile database against which each *L. donovani* protein was scanned. The 'hmmscan' program from HMMER 3.0 package was utilized to search for similarity between *L. donovani* sequences and profiles contained in the modified HMM profile database (Eddy, 1998, 2009). Sequences containing domain(s) with an E-value <=$1e^{-08}$,

target coverage >=50%, sequence identity >=20%, conservation >=70% with the Inl-A profile could be probable Inl-A-like sequences. Further, the proteins having considerably high scoring predicted Inl-A LRR domains (domain score >=110) were then considered for the reverse search analysis. In addition, considering the protein data bank (PDB) HMM database, an HHpred (Söding et al., 2005) analysis based on the HHblits (HMM-HMM-based lightning-fast iterative sequence search) algorithm (Remmert et al., 2012) followed by maximum accuracy (MAC) alignment (MAC realignment threshold =0.3) was performed to define the LRR regions in the predicted Inl-A-like proteins more reliably.

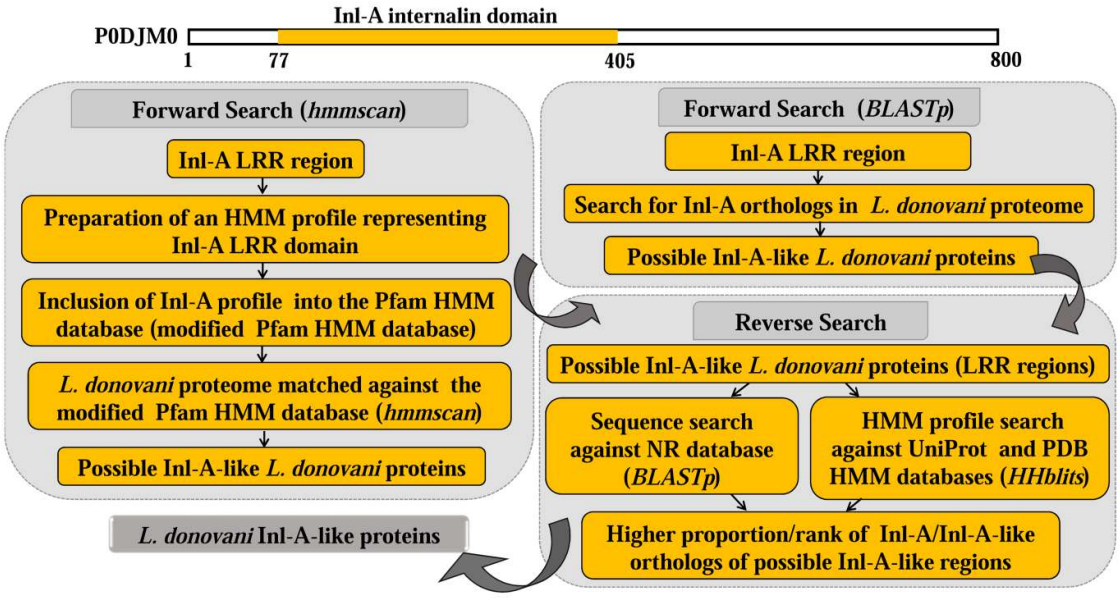


Figure 2.26: **Identifying bacterial virulence factor-like proteins (Inl-A) in *L. donovani*.** Basic methodology utilized to identify Inl-A-like virulence proteins in *L. donovani* is outlined here. Abbreviations: Inl-A, Internalin-A; LRR, Leucine Rich Repeat; HMM, Hidden Markov Model; BLAST (Altschul et al., 1990), Basic Local Alignment Search Tool; HHblits, HMM-HMM-based lightning-fast iterative sequence search tool (Remmert et al., 2012).

L. donovani protein sequences that exhibited similarity to bacterial Inl-A sequence in the 'forward search' analysis were then subjected to the 'reverse search' analysis (Figure 2.26). The predicted Inl-A-like proteins were further compared with NCBI non-redundant (NR) sequence database (Pruitt et al., 2006) using BLASTp (Altschul et al., 1990; Camacho et al., 2009) and against the UniProt and PDB databases using HHblits (Remmert et al., 2012). Homologous sequences in the BLASTp based sequence-sequence comparisons were determined with the threshold, E-value <=$1e^{-06}$, query length coverage>=70% and sequence identity>=20%. Similarly, sequences having probability>=97% in the (HMM) profile-(HMM) profile comparison using HH-blits search were considered homologous, since probability denotes a percentage score which indicates whether a database match is a true positive. A number of proteins were found with regions similar to the predicted Inl-A-like regions of the *L. donovani* proteins and among these proteins only well characterized non-Kinetoplastida proteins were

retained to get a better idea of a relationship between the *L. donovani* proteins and bacterial internalins. The homologous proteins were then ranked considering E-value and only the top 25 subject hits were utilized to determine whether the *L. donovani* proteins had any significant similarity to LRR proteins or Inl-A. The *L. donovani* proteins that had significant sequence similarity with bacterial Inl-A protein(s) according to the 'forward search' and 'reverse search' analysis were considered as *L. donovani* Inl-A-like proteins. Subsequently, in order to identify orthologs of *L. donovani* Inl-A-like proteins in *Leishmania spp.* each *L. donovani* Inl-A-like protein were subjected to database similarity search considering the Kinetoplastida proteome with the help of BLASTp (Altschul et al., 1990). The non-*L. donovani* proteins that were found similar to *L. donovani* Inl-A-like proteins (threshold criteria: E-value <=$1e^{-12}$, query coverage>=70%, identity >= 70%) were considered as orthologs.

2.2.1.3 Establishing orthology among virulence factor proteins in Bacteria and Eukaryota (Kinetoplastida)

Considering *L. donovani* Inl-A-like proteins as a representative of bacterial virulence factor proteins, a number of comparisons were made to establish the orthology among Inl-A class of virulence factor proteins in Bacteria and Kinetoplastida. Similarity between bacterial Inl-A and *L. donovani* Inl-A-like proteins was studied utilizing pair-wise alignments, HMM-HMM comparison of virulence domains and an evolutionary similarity assessment (phylogenetic tree). Sequence similarity among bacterial Inl-A and *L. donovani* Inl-A-like proteins was assessed by determining the global sequence identities in EMBOSS (v6.2.0) (Rice, Longden, & Bleasby, 2000). Further, similarity between the sequence regions important for receptor based internalization of a pathogen could be important for determining orthologous relationship between virulent proteins from phylogenetically distant organisms. Since the LRR region is known to promote invasiveness in the case of bacterial internalins (Lecuit et al., 1997), the similarity between bacterial Inl-A LRR regions and the predicted LRR regions in *L. donovani* Inl-A-like proteins were compared with the help of the HHalign (Söding, 2004) program. An HMM profile for each LRR region in *L. donovani* Inl-A-like proteins was prepared in a similar manner as before (criteria for selecting orthologs: E-value<=$1e^{-06}$, query length coverage>=70% and sequence identity>=30%). Additionally, a PSI-BLAST (version 2.2.28+) (Altschul et al., 1997) search was utilized to determine distant orthologs of Inl-A in Bacteria. These bacterial internalin sequences had sequence identities in the range of 20–30% with *L. monocytogenes* Inl-A and an HMM-HMM comparison between the LRR regions in distant internalin sequences and Inl-A was also performed. Since HMM-HMM comparisons can effectively determine remote homologues (Söding, 2004) the scores obtained in the above analyses were compared to determine the similarity within the LRR region of known internalins and predicted internalins. Moreover, a phylogenetic tree (Huerta-Cepas, Dopazo, & Gabaldón, 2010) considering close orthologs of *L. donovani* Inl-A-like and *L. monocytogenes* Inl-A was prepared utilising the maximum likelihood algorithm of Randomized Axelerated Maximum Likelihood (RAxML version 8.1.6) (Stamatakis, 2014) program. Close ortholog sequences were determined based on a BLAST search (Altschul et al., 1990) of the query in NCBI non-redundant (NR) database (Pruitt et al., 2006) considering a threshold criteria of E-value <=$1e^{-06}$, query length coverage>=70%, sequence identity >=45% and a redundancy check of <=95% sequence identity (Li & Godzik, 2006; Fu et al., 2012).

2.2.1.4 Studying whether remote virulence factor orthologs arose as a result of lateral gene transfer

In order to identify whether the predicted Inl-A-like virulence factors in eukaryotic parasites belonging to Kinetoplastida may have arisen as a result of lateral gene transfer from bacterial pathogens, a phylogenetic analysis was performed to study the evolutionary relationship between these protein sequences. Since, lateral gene transfer from prokaryotes to eukaryotes may have occurred in the case of enzymatic genes belonging to core pathways like amino acid and sugar metabolism (Alsmark et al., 2013; Hirt, Alsmark, & Embley, 2015); the assumption of lateral gene transfer of virulence factors is a likely one. In this respect, orthologs of *L. donovani* Inl-A-like proteins were determined by a BLASTp (Altschul et al., 1990) search in NR (Pruitt et al., 2006) database restricted to non-plant proteins. Orthologous proteins were determined based on the search criteria of E-value<= $1e^{-04}$, query coverage >= 60%, identity>= 20% and representative sequences sharing sequence identity <=95% were aligned in MAFFT (Katoh et al., 2002) with the help of linsi algorithm. Utilizing RAxML (Stamatakis, 2014) the best trees were obtained for the generated multiple sequence alignment consisting of orthologs of *L. donovani* Inl-A-like proteins. Additionally, another phylogenetic tree was prepared by considering orthologs of 16S ribosomal RNA (*Listeria monocytogenes*) and 18S ribosomal RNA (*Leishmania donovani*). Orthologs were determined with BLAST (Altschul et al., 1990) against NT (Pruitt et al., 2006) database and sequences having E-value <= $1e^{-04}$, query coverage >=60%, identity>=40% were considered for the analysis. MSA of representative sequences sharing identity<=95% were utilized to generate the best tree in RAxML (Stamatakis, 2014). The phylogenetic trees were then utilized to study the possibility of lateral gene transfer of bacterial Inl-A to eukaryotic parasites such as *L. donovani*.

2.2.1.5 Ascertaining structures of predicted virulence factors and determining probable host protein interacting partner

Structural modeling was performed in order to identify the likely structure of the predicted Inl-A-like proteins. Herein, *L. donovani* Inl-A-like proteins have been considered as representatives of the predicted Inl-A-like proteins in Kinetoplastida. Further, molecular docking analysis considering some likely host receptor proteins and the predicted virulence factors might provide an indication regarding the feasibility of an interaction complex between the receptor(s) and the identified *L. donovani* internalin-A like proteins. Based on the assumption that clusters having higher number of similar frames and better energetics are more likely to contain the best possible interaction pose between two proteins that are likely to interact; the top three clusters (Cluster I, Cluster II and Cluster III) from the docking analysis were selected for further evaluation.

Homology modeling of L. donovani Inl-A-like proteins

Selected *L. donovani* Inl-A-like proteins were modelled in MODELLER 9.15 (Webb & Sali, 2016; Mart´ı- Renom et al., 2000; S˘ali & Blundell, 1993; Fiser, Do, & S˘ali, 2000) utilizing *L. monocytogenes* Inl-A [PDB ID: 1O6S, chain A] as a template and structure-sequence alignment from the HHpred server (S¨oding et al., 2005). Subsequently, structure refinement and loop modeling was performed in MODELLER for certain selected models to predict the most plausible 3D coordinates with least stereo-chemical

violations in the models. Since calcium (Ca^{2+}) is important for the interaction of Inl-A with its host receptor (Schubert et al., 2002), ligand Ca^{2+} was coordinated with corresponding residues in the modelled proteins with a coordination sphere of 2.45˚A (similar to Inl-A structure) using MODELLER. Three dimensional (3D) models were then ranked on the basis of energy parameters such as DOPE score and the structures were validated with the help of VERIFY3D (Lüthy, Bowie, & Eisenberg, 1992; Bowie, Luthy, & Eisenberg, 1991) and RAMPAGE (Lovell et al., 2003). These homology models were visualized and analyzed in chimera 1.10.1 (Pettersen et al., 2004).

Molecular docking analysis of L. donovani Inl-A-like proteins with selected host receptor proteins

It has been observed that homologous proteins may share protein interaction partners (Todd et al., 2001) and thus the possibility that the *L. donovani* Inl-A-like proteins might interact with E-cadherin (hEC1), the known host receptor for *L. monocytogenes* Inl-A has been explored. The modeled *L. donovani* Inl-A-like proteins were superimposed onto the template Inl-A structure and residues that were found to be within 5˚A of hEC1 were taken as possible interacting residues of the predicted virulence factor. Similarly, hEC1 residues that are known to interact with Inl-A were taken as possible interacting residues of the receptor. Docking between *L. donovani* Inl-A-like proteins and hEC1 was guided with this a priori interaction information and molecular docking was performed via multiple programs namely, PatchDock (Duhovny et al., 2002; Schneidman-Duhovny et al., 2005), HADDOCK 2.0 (High Ambiguity Driven protein-protein Docking) (Vries et al., 2007; Dominguez et al., 2003; Zundert et al., 2016) and ClusPro 2.0 (Kozakov et al., 2006; Comeau et al., 2004b, 2004a). Additionally, re-docking of Inl-A with hEC1 was also performed with these docking programs to check whether the native Inl-A-hEC1 interaction conformation can be captured reliably in these docking programs (Figure S2). Utilizing geometric shape complementarities matching based on geometric hashing and pose clustering techniques docked poses between *L. donovani* Inl-A-like proteins and hEC1 were obtained in PatchDock (Duhovny et al., 2002; Schneidman-Duhovny et al., 2005) followed with further refinement of the complexes in FireDock (Andrusier et al., 2007). The best scoring solutions (top 100) were obtained from PatchDock based on the receptor-ligand geometric score and were clustered according to root mean square deviation (RMSD) in chimera (version 1.10.1) (Pettersen et al., 2004) to determine the largest docked clusters. Subsequently, *L. donovani* Inl-A-like proteins and hEC1 were docked in HADDOCK (Vries et al., 2007; Dominguez et al., 2003; Zundert et al., 2016) where the predicted interacting residues were taken as active residues and passive residues were defined automatically around the active residues. Here, a data driven (ambiguous interaction restraints) rigid-body docking process along with solvent based refinement of complexes is utilized to determine likely interaction conformations between the proteins. Additionally, a Fast-Fourier Transform (FFT)-based approach available in ClusPro (Kozakov et al., 2006; Comeau et al., 2004b, 2004a) was utilized to dock *L. donovani* Inl-A-like proteins and hEC1 assuming that the residues determined as possible interacting residues previously share attractive forces. Thus, correlation between binding site free energy attractor and cluster size was also considered to provide an approximation of the native binding conformation between these proteins which are likely to interact. Top three clusters (Cluster I, Cluster II and Cluster III) with more average negative docking scores were selected from the docking solutions for further evaluation. A representative frame from each cluster

obtained in the multiple programs was aligned with Inl-A crystal structure (PDB ID: 1O6S) and ligand RMSD (l-RMSD) was calculated in chimera (Pettersen et al., 2004). For better enumeration of similar interaction poses, the docking scores (FireDock refined glob score, HADDOCK score, ClusPro balanced score) of the ranked clusters from each program were plotted with l-RMSD. A consensus pose predicted from all three programs that utilize different methodologies and scoring functions was considered as the best docked conformation. This ascertains more confidence to such an interaction pose between *L. donovani* Inl-A-like proteins and hEC1 and the likelihood that these proteins interact. Finally, the free energy of binding of *L. donovani* Inl-A-like proteins and hEC1 were determined using PISA (available in ccp4mg (version 2.8.1) (Krissinel & Henrick, 2007) and the probable hydrogen-bonding interaction patterns of these complexes was also determined. Further, residues involved in probable hydrophobic interactions were predicted in Protein Inter-Chain Interaction (PICI) web server (Das, Sharma, Kumar, Krishna, & Mathur, 2013). Additionally, in order to explore the possibility that *L. donovani* Inl-A- like proteins can interact with other members of the cadherin superfamily initially the structural similarity within this superfamily was assessed. Structures for representative members of the cadherin superfamily like Type I, Type II, desmosomal, and cadherin related classes were selected based on a previous analysis that studied structural diversity within the mammalian cadherin superfamily (Sotomayor, Gaudet, & Corey, 2014). Molecular docking between representative cadherin family members and the identified *L. donovani* Inl-A-like proteins was performed in a similar manner as discussed above. Herein, docking was performed in HADDOCK (Vries et al., 2007; Dominguez et al., 2003; Zundert et al., 2016) and this might provide an idea regarding the feasibility of a complex between the proteins under consideration and a likely interaction conformation between them. The l-rmsd values were obtained utilizing a python script interface to chimera wherein the docked structure was compared with the best docked conformation of the respective *L. donovani* Inl-A-like protein bound to hEC1 (previous analysis). The plots of docking scores versus l-RMSD for a possible interaction between *L. donovani* Inl-A-like proteins and cadherin superfamily members (type I cadherin and desmosomal cadherin) were prepared and compared with the plots of docking scores versus l-RMSD for *L. donovani* Inl-A-like proteins and hEC1.

2.2.2 Utilizing co-evolution analysis to determine inter-dependent residue pairs that could be functionally relevant in inter-cellular protein-protein interactions

2.2.2.1 Identifying co-evolving residue pairs in protein-protein interaction complexes

In order to determine co-evolutionary residue pairings in inter-protein interactions an information theory based measure (Co-Var) has been developed. This score is computed considering the position-wise amino acid frequencies in the multiple sequence alignments of the proteins involved in the protein-protein interaction. Correlations among evolutionary patterns between protein pairs may be determined based on the Co-Var score as outlined below:

$$Co-Var\ score(A,B) = BHC(A,B) - MI(A,B)$$

wherein, Co-Var score (A, B) represents the co-variation score between position A and B, MI (A,B) and BHC (A,B) denote mutual information and Bhattacharyya coefficient between positions A and B. Additionally, 'A' and 'B' represent all possible amino acids at position A in the alignment for the first protein family position while B in the second protein family of the complex. The mutual information value here measures the mutual dependence between the distributions of amino acids at particular positions between protein families. In information theory, mutual information represents the entropy-based formulation for quantifying the interdependence between the values of two random categorical variables such as position-wise amino acid frequency distributions (Juan et al, 2013). It is calculated considering the sum over all the possible combinations of di-residue frequencies (Dunn et al., 2007) and mutual information (MI) between two aligned columns A and B is calculated as

$$MI(A,B) = \sum_{a \in A} \sum_{b \in B} p(a,b) \times log\ \frac{p(a,b)}{p(a) \times p(b)}$$

wherein, 'p(a)' is the frequency of occurrence of each residue in the column 'A', 'p(b)' is the frequency of occurrence of each residue in the column 'B' and 'p(a,b)' represents the di-residue frequency. Additionally, Bhattacharyya coefficient quantifies the overlap between set of amino acids between a pair of columns. It is a measure of similarity between two distributions and is used to calculate the amount of overlap between them, by splitting the samples into several partitions.

$$BHC(A,B) = \sum_{a,b=1}^{20} \sqrt{p(a) * p(b)}$$

wherein, BHC(A,B) denotes Bhattacharyya coefficient between positions A and B, 'p(a)' and 'p(b)' are the amino acid residue frequencies present in the respective positions.

2.2.2.2 Utilizing the Co-Var measure to study protein-protein interaction complexes

Close orthologs or similar sequences may be determined using DELTA-BLAST (Domain enhanced lookup time accelerated BLAST) (Boratyn et al., 2012) and minimum 50 sequences may be considered for generating multiple sequence alignment representative of each protein family. While detecting co-evolution, the first sequence in each alignment is considered as the reference sequence, alignment shuffling is performed and multiple instances of the program are run such that additional statistical significance can be assigned to the predicted co-evolving positions. In order to reduce the influence of phylogenetic relationships, alignment shuffling was performed by randomly selecting equal number of orthologous sequences (from the entire set of orthologs) of each family for the alignment. This ensures that amino acids across each column exhibited variation during different runs of the Co-Var program. After the calculation of the Co-Var scores, the alignment column pairs and scores are mapped onto a corresponding reference sequence and structure for each set and average Co-Var score and the corresponding z-scores across the runs were determined. Co-evolving positions that have significant z-scores and which are reported in multiple runs (5) of the program with alignment shuffling may be considered for further analysis. The Co-Var methodology to study inter-protein co-evolution has been depicted in Figure 2.27. Co-evolutionary pairings may be selected based on a z-score threshold where z-scores are calculated on the Co-Var score. Further, Co-Var scores corresponding to lower (negative) z-scores are indicative of higher likelihood of co-variation.

2.2.2.3 Validating the applicability of Co-Var measure in studying inter-protein co-evolution

A set of protein-protein interaction complexes were collected from previously published data (Sowmya et al., 2015; Mintseris and Weng, 2003; Rodriguez-Rivas et al., 2016) and complexes for which sufficient number of homologs and a reference structure could be determined were considered during this analysis. Interacting complexes (50 protein complexes) that are likely to co-evolve were selected as the "positive set" (Table S13). Additionally, 50 protein pairs which are known to be non-interacting based on experimental analysis were randomly selected from the Negatome database (Smialowski et al., 2009) as the "negative set" (Table S14). The compiled positive and negative set were analyzed in Co-Var, CAPS (Fares and McNally 2006; Fares and Travers 2006), EV-complex (Hopf et al., 2014) and MirrorTree (Ochoa and Pazos, 2010) methods. For this purpose, close orthologs or similar sequences of each reference protein in the selected complexes were determined using DELTA-BLAST (Domain enhanced lookup time accelerated BLAST) (Boratyn et al., 2012). Taxonomy filtered non-redundant sequences having E-value <= $1E^{-04}$, query coverage >= 70%, sequence identity >= 45% were utilized for preparing multiple sequence alignments (MSA) representative of each sequence family in MAFFT (Katoh et al., 2002). Alignments for homologous sequences of the representative interacting and non-interacting proteins were analyzed in MirrorTree to determine whether the proteins were co-evolving. Further, utilizing Co-Var, CAPS and EV-complex co-evolving residue positions were determined in the interacting and non-interacting proteins. In order to determine the

applicability of Co-Var methodology in studying inter-protein co-evolution the 'percentage of co-evolved pairs' in interacting (positive) and non-interacting proteins (negative) has been considered as an index. The 'percentage of co-evolved pairs' predicted for each positive and negative set pairs by CAPS and EV-complex or the MirrorTree correlation coefficient have also been utilised to study these complexes. Based on these analyses, it has been determined whether these indices can successfully differentiate between the positive and negative set. Subsequently, considering a reference structure for each of the positive set of complexes the inter-residue distances between the Co-Var based predicted co-evolving residue positions in the interacting proteins has been calculated to prepare a distance distribution plot for analysis. Additionally, two measures that capture the co-evolving pairs in the overall complex [Percentage of co-evolved pair that occur at interface (IC)] and those that occur at the interface [Percentage of interacting pair that are co-evolved (PC)] were also computed.

$$\mathrm{IC} = \left(\frac{\mathrm{Co-evolved\ residues\ at\ interface}}{\mathrm{Total\ co-evolved\ pairs}}\right) * 100$$

$$\mathrm{PC} = \left(\frac{\mathrm{Co-evolved\ residues\ at\ interface}}{\mathrm{Total\ interface\ pairs}}\right) * 100$$

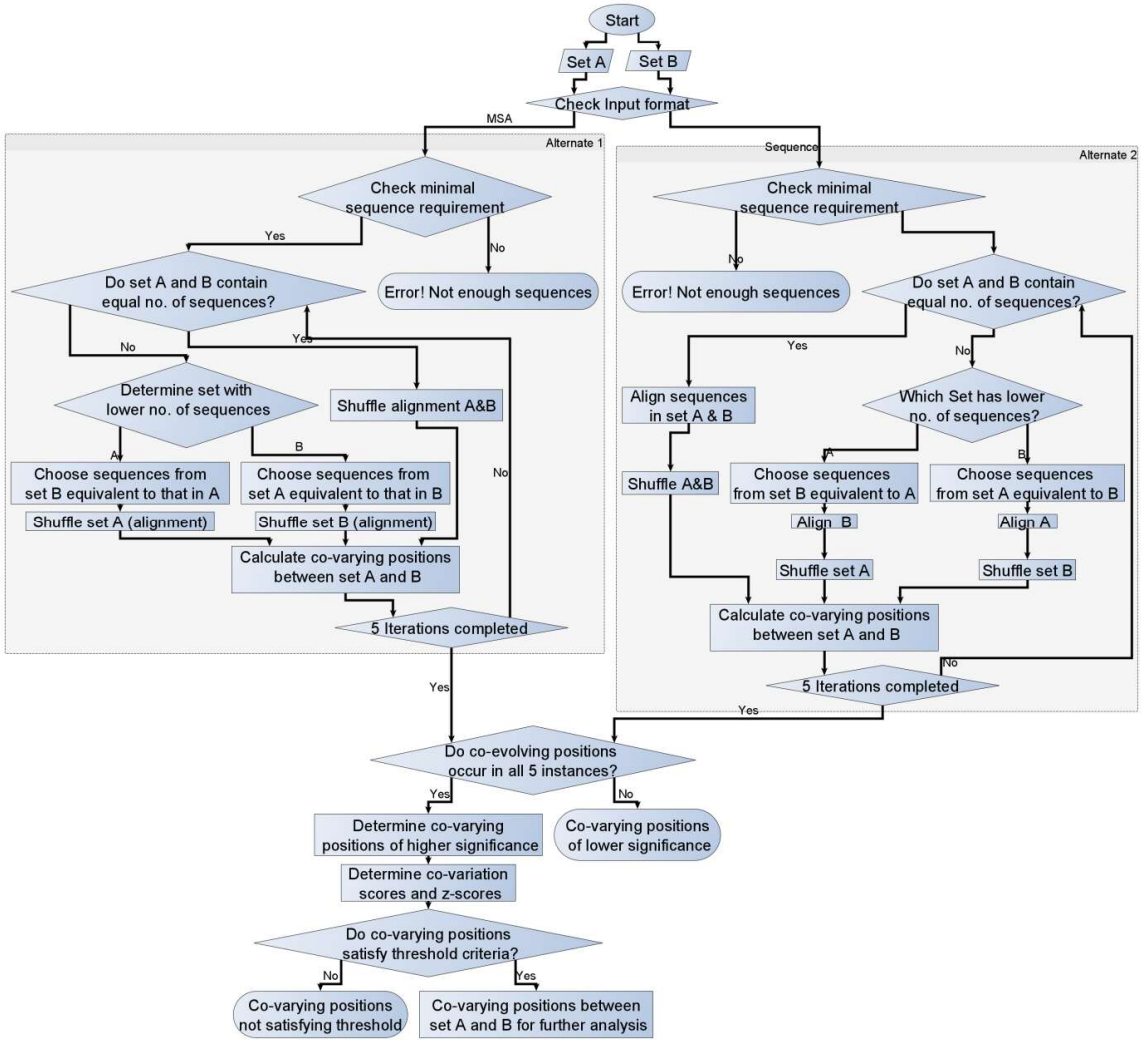


Figure 2.27: Co-Var methodology to study inter-protein co-evolution

2.2.2.4 Studying co-evolution in hetero-dimeric protein complexes involved in inter-cellular interactions

Inter-protein co-evolution has been studied in detail in certain inter-cellular protein interaction complexes such as CSF3-CSF3R, TGFA-EGFR, FGF1-FGFR1, TGFB3-TGFBR2, FGF10-FGFR2 and FGF1-FGFR2. The respective sequences and structures were obtained from the UniProt and PDB databases (Rose et al., 2014; Tamada et al., 2006; Breuza et al., 2016; Consortium, 2018). Orthologs of the representative sequence in each family were determined with the help of DELTA-BLAST (Boratyn et al., 2012) (E-value <= $1E^{-04}$, query coverage >= 70%, sequence identity >= 45%) and taxonomy filtered non-redundant sequences were retained. These sequences were utilized for preparing multiple sequence alignments in MAFFT (Katoh et al., 2002). Considering these alignments as input co-evolving residue positions among the representative sequences in each sequence family involved in the interaction complex were determined with the help of Co-Var methodology based on z-score <= -1 as the selection threshold. Subsequently, a distance distribution analysis was performed by determining the inter-residue distances among the predicted co-

evolved residue pair positions by mapping them onto the corresponding 3D structure of the complex involving the reference sequences. Further, the degree of each residue position involved in the co-evolutionary pairings were determined by analyzing a network representing residue positions as nodes and co-evolutionary pairing between positions as an edge. Based on whether a residue position had a degree higher than the median of the degree distribution, high degree co-evolved positions were selected for further analysis. Moreover, in order to determine whether the high degree co-evolved positions are biologically relevant or not, a number of additional analyses were performed. It was studied whether these positions are present within the functional domain regions of each protein, whether these positions are important from the stand point of intra-molecular co-evolution or whether these positions are frequently prone to mis-sense substitutions.

2.2.2.5 Studying the physiological and functional relevance of predicted co-evolutionary pairings in inter-cellular interaction complexes

In order to ascertain whether the high degree co-evolutionary pairings have a biological significance a number of analyses was performed. The functional relevance of the predicted co-evolutionary pairings was studied by performing domain mapping, mutation analysis and determining whether the residue positions are important from the standpoint of intra-molecular co-evolution as well. Details regarding the domains in each protein were obtained by querying the Pfam database (Finn et al., 2013) and it was determined whether the residue positions involved in co-evolutionary pairings occurred within the functional domain regions in the interacting proteins with the help of in-house perl programs. Additionally, mutation data from the COSMIC database (Tate et al., 2018) has been considered to map whether residue positions predicted as important with the help of inter-molecular co-evolution analysis exhibit frequent mis-sense substitution mutations in disease conditions such as cancer. It is plausible that as a result of mutation/s in critical residue positions the interaction between these proteins is perhaps compromised in conditions such as cancer. Amino acid pairing frequencies among the predicted co-evolved positions in the native reference sequences were compared to the ones obtained based on the assumption that the sequences exhibit the substitution mutations. Similarly, the Co-Var methodology was also utilised to study co-variation within each protein involved in the inter-cellular protein interaction complex case studies considered herein. This was done with a view to determine whether the high degree co-evolved residue positions identified in inter-molecular co-evolution could additionally have crucial roles within each protein in terms of their role in intra-molecular co-evolution. For this purpose, initially the applicability of Co-Var was tested in determining intra-protein co-evolution pairs. A set of 252 conserved domain database (CDD) protein family alignments (Marchler-Bauer et al., 2014) with at least 80 sequences in each alignment was utilised to study intra-molecular co-evolution. Intra-molecular co-evolution in these protein families was studied with the help of Co-Var, CAPS (Fares and Travers 2006), MI (Korber et al., 1993) and PSICOV (Jones et al., 2011), respectively. Herein, the inter-residue distances among intra-protein co-evolved pairs predicted utilising these programs was considered for comparison purpose. Having established the applicability of Co-Var in studying intra-protein co-evolution, it was determined whether the residue positions predicted to be involved in inter-protein co-evolution in inter-cellular protein interaction complexes are also predicted to be involved in intra-protein co-evolutionary pairings.

2.2.2.6 Co-Var web-server for studying inter-protein co-evolution

In order to provide a wider scope to the Co-Var methodology and a web server version (http://www.hpppi.iicb.res.in/ishi/covar/index.html.) of the method has been developed which is freely accessible to users. Herein, utilizing a set of homologous sequences or alignment(s) of proteins as inputs the Co-Var methodology may be utilized for studying inter-molecular (protein-protein) co-evolution. The front end of the server is HTML, PHP and java based while a perl based implementation of the Co-Var methodology works on the backend of the server to predict reference sequence (first sequence in the alignment) mapped co-evolved residue positions. Moreover, a reference structure may be uploaded and based on this structure structural mapping of pairings may be obtained along with inter-residue distances between the residues involved in co-evolutionary pairings. Co-evolutionary pairings in close structural proximity can be visualized in a viewer (Rego and Koes, 2015). Further, additional modules are available for functional interpretation of the inter-protein co-evolutionary pairings in terms of their frequency of occurrence among predicted co-evolved positions (high degree co-evolved positions). Finally, a list of reference sequence or structure mapped co-evolutionary pairings can be downloaded from the results link.

2.3 Discussion

Inter-cellular interactions may be governed by a range of proteins that could be involved in protein-protein interactions governing cellular signalling cascades and in turn the interaction between these proteins might dictate the cellular response. Such inter-cellular interactions governed by proteins may be studied with the help of different sequence analysis approaches. Evolutionary similarities in sequences may be utilized to identify the function of a protein or predict its protein interaction partner(s). Thus, sequence analysis techniques may be utilised to identify proteins possibly involved in inter-cellular interactions. Additionally, information contained in biological sequences may also help in identification of structurally or functionally relevant residues in inter-cellular protein interaction complexes.

In this respect, functional annotation of proteins possibly involved in inter-cellular interactions between host cells and pathogens has been performed. For this purpose, a methodology for predicting novel virulence factor proteins within a pathogen proteome and identifying its (their) probable host interaction partner(s) was outlined. This methodology involving rigorous homology searching protocols including Hidden Markov Model (HMM) and BLASTp-based searches (Remmert et al., 2012, Altschul et al., 1990) helped in the identification of probable virulence factors. This class of virulence factors, internalin-like class of proteins (which is essentially a bacterial (*Listerial*) virulence factor) was predicted in Kinetoplastida and in particular *Leishmania donovani.* Moreover, preliminary experimental evidence from an RNA-seq analysis that had identified differentially expressed genes associated with more virulent *L. donovani* suggested that the *L. donovani* Inl-A-like genes (LdBPK_030010.1, LdBPK_311630.1, and LdBPK_365070.1) are over-expressed. Therefore, utilizing different sequence analysis methods three potential candidates (UniProt ID: E9B7L9, E9BMT7 and E9BUL5) of Inl-A-like LRR containing proteins were predicted in *L. donovani* This class of virulence

factors, internalin-like class of proteins (which is essentially a bacterial (*Listerial*) virulence factor) was predicted in Kinetoplastida and in particular *Leishmania donovani.* Moreover, preliminary experimental evidence from an RNA-seq analysis that had identified differentially expressed genes associated with more virulent *L. donovani* suggested that the *L. donovani* Inl-A-like genes (LdBPK_030010.1, LdBPK_311630.1, and LdBPK_365070.1) are over-expressed. Therefore, utilizing different sequence analysis methods three potential candidates (UniProt ID: E9B7L9, E9BMT7 and E9BUL5) of Inl-A-like LRR containing proteins were predicted in *L. donovani.* This analysis was followed with a molecular docking analysis wherein the known interaction partners of the known virulence factors or their protein class were docked with the *L. donovani* Inl-A-like proteins. Further, the respective ligand root-mean-square deviation (l-RMSD) values and docking scores were analyzed to identify the most likely host interaction partner of the probable virulence factors. Thus, based on this analysis it was determined that phylogenetically distant organisms may share similar virulence factors or invasion mechanisms.

The interaction between proteins may be transient or obligate; however it is generally maintained throughout the course of evolution, despite alterations in the protein sequences (Mintseris et al., 2005). Thus, an evolutionary approach can be utilised to identify the inter-dependent protein residues or sequence positions which are crucial for the conserved interaction between these proteins throughout evolution. Such co-evolutionary pairings or co-varying positions may contain important structural or functional information regarding these interactions. Thus, a measure (Co-Var) has been developed to identify co-varying positions under evolutionary selection pressure that probably arise as a result of compensatory alterations in protein-protein complexes. Inter-molecular co-evolution studies in proteins led to the observation that a certain fraction of co-evolving residue pairs may occur in spatially separated positions. This observation was consistently identified via multiple inter-protein co-evolution analysis programs that we have utilized. This trend is similar to previous studies wherein spatially separated residue pairs have been predicted as co-evolving in addition to residue pairs in close spatial proximity at the interaction interface (Anishchenko et al., 2017; Mintseris and Weng, 2005). Inter-dependent co-evolutionary pairings have been determined in inter-cellular protein interaction complexes with the help of this co-variation measure that studies inter-protein co-evolution. Herein, co-evolutionary pairings that occur at non-interface regions occur among residues present in functional domains or residues that have roles in intra-protein co-evolution in individual proteins involved in the complex. Moreover, inter-protein co-evolution analysis in the inter-cellular protein interaction complexes involved in cancer metastasis identified that certain receptor positions share many co-evolutionary pairing connections with certain ligand positions only. Additionally, these high degree co-evolving positions were found to be frequently prone to mis-sense substitution mutations in cancer and as such absence of co-ordinated changes at these positions may contribute to disease associated altered interaction in these complexes. Thus, this sequence analysis based approach might allow one to envisage which residue pairing interactions could be functionally relevant or important for an interaction to occur between a pair of proteins.

Chapter 3

Studying intra-cellular meta-interaction networks

3.1 Results

3.1.1 Regulatory miRNA may influence mRNA expression or gene expression via intermediate miRNAs

Synopsis:

In the present analysis, a possible mechanism that could be important for miRNA biogenesis has been studied. It is possible that miRNA (miRNA-A) which has binding sites in tandem with the binding site of another miRNA (miRNA-B) may influence the biogenesis of miRNA-B. In this scenario, miRNA-A has larger number of binding sites than miRNA-B. In order to validate this hypothesis wherein miRNA can influence mRNA expression by regulating the biogenesis of intermediate miRNA, a computational analysis was performed. This hypothesis termed as "coordinate biogenesis" by our collaborator, has been the subject of this investigation. A set of miRNA that are likely to exhibit co-ordinate biogenesis has been determined and the possibility whether this biological phenomenon occurs in general has been explored. In particular, the probable miRNA co-ordinate biogenesis regulatory network for multiple miRNA in *Homo sapiens* and *Mus musculus* has been determined initially. Subsequently, this hypothesis was validated by identifying regulator-target relationships in the miR-146a-5p network (Figure 3.1). Further, some of the predicted regulatory relationships were studied in lipopolysaccharide exposed macrophages by our collaborator.

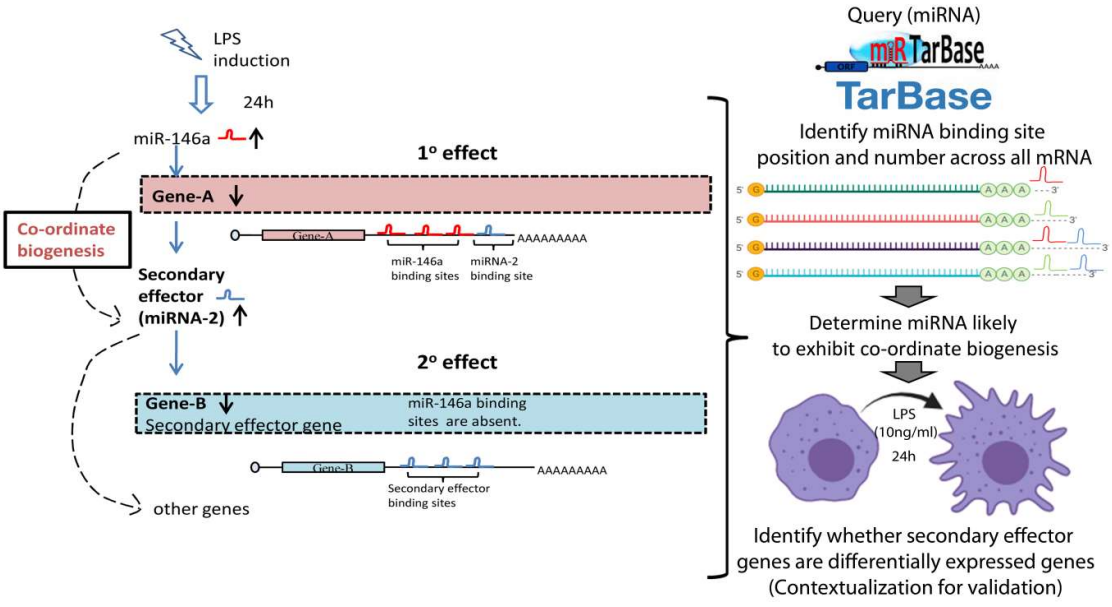


Figure 3.1: Determining CB regulator (miR-146a-5p)-target network in monocyte response under lipopolysaccharide (LPS) exposure (endotoxin response).

3.1.1.1 Identification of regulatory miRNAs and their targets (mRNAs/miRNA) to construct the meta-interaction network in miRNA biogenesis

Analysis of miRNA binding sites in mRNA may allow the prediction of possible miRNA pairs that exhibit co-ordinate biogenesis (CB). Experimentally determined miRNA binding site information such as genomic location, and number of sites collected from TarBase (Vlachos et al., 2014) and miRTarBase (Chou et al., 2015) database was utilized to predict the miRNA that are likely to act as the coordinate biogenesis regulator. Interactions (2,33,601) retrieved among human miRNA (2588) and genes (8808) from miRTarBase and interactions (2,82,181) retrieved among human miRNA (1026) and genes (13,811) from TarBase allowed the prediction of miRNA-A (1433) likely to act as CB regulators in humans with around 9,16,506 CB pairs. Further, interactions (24,650) retrieved among mouse miRNA (847) and genes (3940) from miRTarBase and interactions (1,34,997) retrieved among mouse miRNA (434) and genes (10,581) from TarBase predicted that miRNA-A (302) are likely to act as CB regulators in mice with around 68,783 CB pairs. This relationship ["miRNA-A:mRNA-A —> miRNA-B:mRNA-B"] derived between miRNA-A and miRNA-B considering the binding site information data has been referred to as coordinate biogenesis (CB) and the CB regulatory network for miR-146a-5p in *Mus musculus* and *Homo sapiens* has been analyzed subsequently.

3.1.1.2 Determining and analyzing regulator (miRNA)-target (mRNA) interaction network considering miR-146a-5p as CB regulator

Based on the assumption that miR-146a may exhibit coordinate biogenesis wherein it influences the generation of a cluster of miRNAs, the probable coordinate biogenesis network in humans and mice has been predicted. Herein, 33,214 CB regulator-target relationships were identified across 21 genes (Gene A) and 616 miRNA-B when hsa-miR-146a-5p (miRNA-A) acts as the regulatory miRNA. Considering the differential expression analysis dataset, a set of 62 CB regulator-target relationships were predicted in which both 'Gene A and B' are down-regulated without a direct regulatory relationship between 'Gene A and B' (Figure 3.2, Table S15). However, the coordinate biogenesis hypothesis has been studied in detail considering the mouse mmu-miR-146a-5p network in detail by identifying possible CB regulator-target relationships in lipopolysaccharide exposed macrophages. Based on these coordinate biogenesis conditions, 80,082 CB regulator-target (miRNA-1:mRNA-A —> miRNA-2:mRNA-B) relationships were obtained across 57 genes (Gene A) and 192 miRNA-B for mmu-miR-146a-5p (miRNA-A). However, considering the gene expression data about 116 CB regulator-target relationships were identified which are likely to have a role in macrophage response to LPS exposure. Thus, 91 CB regulator-target relationships are likely to be active in this scenario in which both Gene A and Gene B are down-regulated and in which there is no known direct regulatory relationship between Gene A and B (Figure 3.3). These final sets of 91 CB regulator-target relationships wherein both 'Gene A'& 'Gene B' are significantly down-regulated have been considered for experimental validation (Table S16). Thus, 21 possible candidate miRNAs (i.e., miRNA-A) exist that

are likely to be regulated by miR-146a-5p via coordinate biogenesis which can in turn regulate the expression of 42 miRNA-B (secondary effector genes) (Table S16).

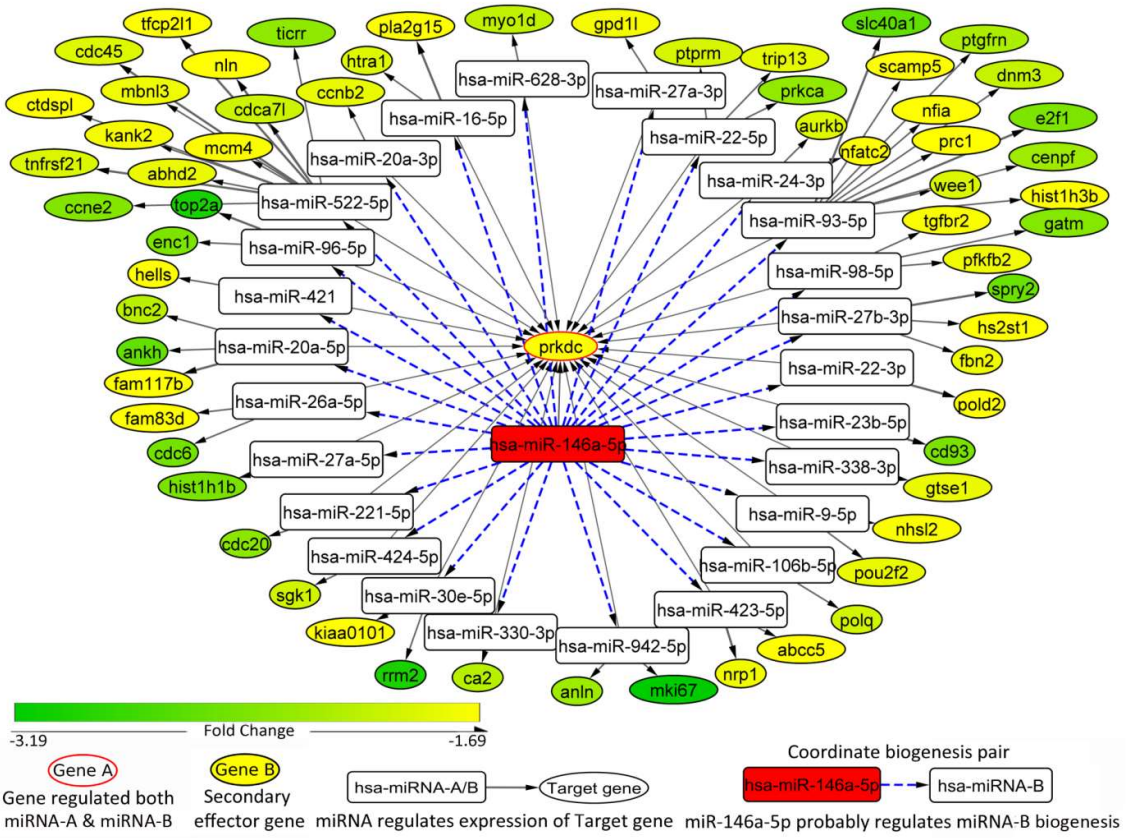


Figure 3.2: Probable CB regulator-target relationships in hsa-miR-146a-5p network. Possible set of miRNA-A:mRNA-A –> miRNA-B:mRNA-B co-ordinate biogenesis relationships identified considering hsa-miR-146a-5p as the regulatory miRNA (miRNA-A) have been exemplified here. 'GeneA' and 'Gene-B' that were found to be down-regulated based on the fold change status of differentially expressed mRNA in macrophages exposed to LPS (10ng/ml) for 24 hours (GSE85333) are shown.

CB regulator-target relationships may influence miRNA-mRNA interaction networks in cells

Based on conditions required for coordinate biogenesis and the binding site analysis, hsa-miR-146a-5p can possibly influence the expression of 187miRNA-B as a CB regulator. In this respect, based on miRNA sequencing analysis upon miR-146a-5p over-expression in macrophage cell line in-vitro (in our collaborators lab) validation of CB-regulator-target relationships has been performed. Herein, 32 miRNA were identified that get differentially expressed, wherein, 9 miRNA were found to be up-regulated and 23 miRNA were found to be down-regulated. Considering this miRNA expression profiling data it was validated that miR-146a over-expression is likely to influence the biogenesis of 2 miRNA (miR-345-5p and miR-16-5p) whose expressions were found to

be significantly up-regulated. Additionally, considering the previous analysis of macrophage response to LPS exposure, 8 CB regulator-target relationships (Table 3.1) could be validated wherein miR-16 (miRNA-B) is up-regulated and gene A and gene B are down-regulated. Thus, it is likely that some CB-regulator target relationships may be active in miRNA-mRNA interaction network in cells influencing macrophage response to LPS exposure.

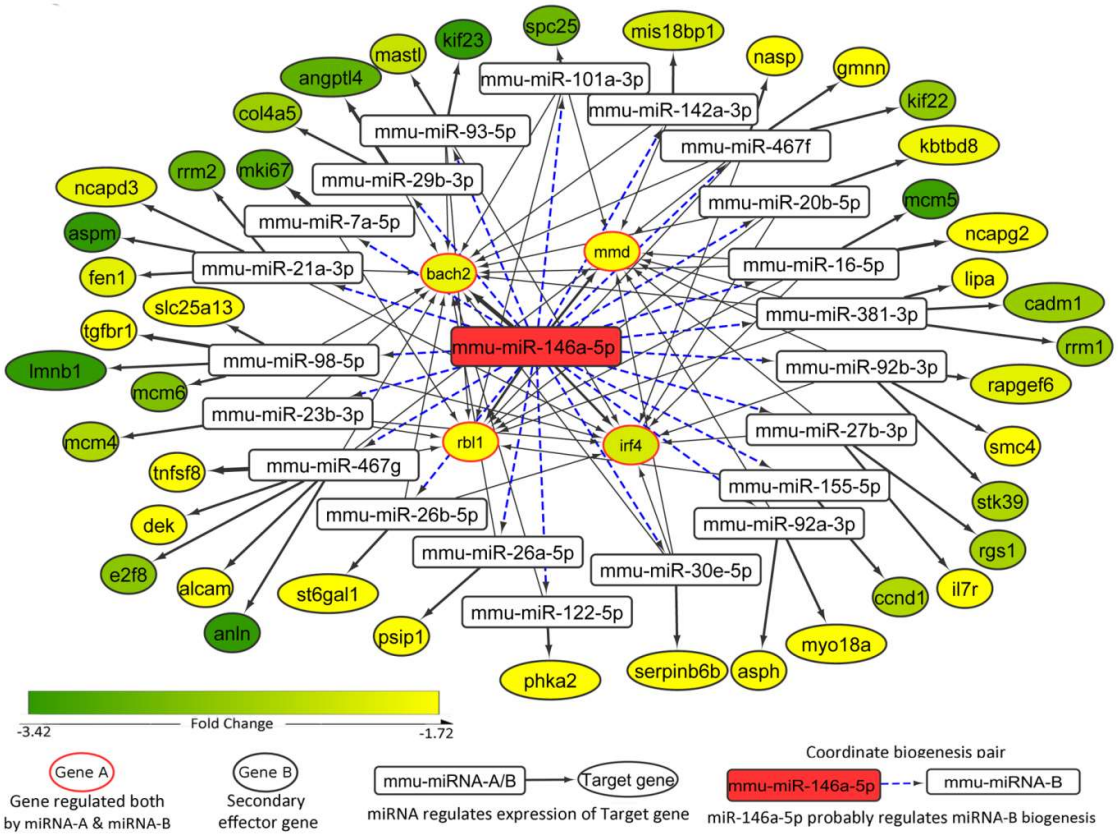


Figure 3.3: Possible CB regulator-target relationships considering mmu-miR-146a-5p as a regulator. The set of predicted co-ordinate biogenesis relationships (miRNA-1:mRNA-A —> miRNA-2:mRNA-B) with mmu-miR-146a-5p as the CB regulator miRNA are shown here. The fold change status of differentially expressed mRNA in murine macrophages exposed to LPS (10ng/ml) for 24 hours (GSE19490) have been included for down-regulated 'Gene A' and 'Gene-B'.

Table 3.1: **Validated CB regulator-target relationships considering miR-146a-5p as CB-regulator in macrophages responding to LPS exposure.**

Gene with tandem miRNA pairs (Gene A)	CB-pair Regulator (miRNA-A)	Binding sites of miRNA-A in Gene A	Secondary Effector miRNA (miRNA-B)	logFC (miRNA-B)	P-Value (miRNA-B)	Target gene of secondary effector miRNA (Gene B)	Binding sites of miRNA-B in Gene B	Counts of miRNA CB-pair (prevalence)
bach2	mmu-miR-146a-5p	4	mmu-miR-16-5p	0.51	0.0114	mcm5	2	39
Mmd	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	mcm5	2	39
rbl1	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	mcm5	2	39
irf4	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	mcm5	2	39
bach2	mmu-miR-146a-5p	4	mmu-miR-16-5p	0.51	0.0114	ncapg2	3	39
Mmd	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	ncapg2	3	39
rbl1	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	ncapg2	3	39
irf4	mmu-miR-146a-5p	2	mmu-miR-16-5p	0.51	0.0114	ncapg2	3	39

Note: FC- fold change

Conclusion

The computational approach has been utilized to predict probable regulatory relationships between miRNA wherein one miRNA is likely to influence the biogenesis of the other. Herein, in case a CB regulator (miRNA-A) is up-regulated it can in turn up-regulate the expression of another miRNA (miRNA-B) by co-ordinate biogenesis and as such its corresponding targets (mRNA-A) are likely to be down-regulated. Further, as a result of this biological phenomenon, a corresponding downstream effect in the known mRNA target (mRNA-B) of the secondary effector miRNA-B may also be observed. In order to study this mechanism, miRNA-mRNA regulatory interaction data such as genomic binding site location and number has been considered to predict a set of miRNA which are likely to act as CB regulator and influence the biogenesis of miRNA targets. A probable miRNA biogenesis regulatory or meta-interaction network for miR-146a-5p has been predicted in mouse and humans. Further, experimental analysis of this relationship was performed in a scenario wherein miR-146a-5p acts as a critical regulator of monocyte response under lipopolysaccharide (LPS) exposure (Nahid et al., 2009). In this model system, it is believed that among endotoxin-responsive microRNAs, miR-146a gradually increases up to 35-fold in response to LPS stimulation over a period of over 24 h whereas its target gene (TNF-α) expression increased and then gradually decreased exhibiting negative correlation with miR-146a expression (Nahid et al., 2009). In this respect, utilizing large scale expression analysis data of monocyte response to LPS exposure, a number of miRNA which are likely to be regulated by miR-146a-5p were identified. MiRNA such as miR-16-5p, miR-26a-5p, miR-27b-3p, miR-30e-5p, miR-93-5p, and miR-98-5p have been identified as probable miR-146a targets in miR-146a-5p coordinate biogenesis regulatory network in mice and humans. Moreover, 8 among the predicted coordinate biogenesis relationships (miRNA-1:mRNA-A —> miRNA-2:mRNA-B) for mmu-miR-146a-5p could be validated.

Inference:
Regulatory miRNA may influence mRNA gene expression via intermediate miRNAs by regulating their expression by means of co-ordinate biogenesis. This has particularly been observed in the case of CB-regulator target relationships identified considering miR-146a-5p in miRNA-mRNA interaction network in macrophage cells responding to LPS exposure.

3.2 Methodology

3.2.1 Determining the miRNA-mRNA meta-interaction regulatory network

3.2.1.1 Predicting possible miRNA pairs that exhibit co-ordinate biogenesis to form a meta-interactome regulatory network

Experimentally identified miRNA binding site information (genomic location, number of sites) collected from TarBase (Vlachos et al., 2014) and miRTarBase (Chou et al., 2015) database was utilised for identifying probable co-ordinate biogenesis pairs. Interaction data from TarBase (Vlachos et al., 2014) and miRTarBase (Chou et al., 2015) were combined to extract possible miRNA pairs that exhibit co-ordinate biogenesis. Additionally, other network components such as target genes of miRNA-A and miRNA-B (possible secondary effector genes) were identified according to the following conditions (Figure 3.4):

- Within any gene say 'Gene A', if miRNA-A had 2 or more binding sites in tandem with miRNA-B (with fewer number of binding sites than miRNA-A) on the same 3' UTR they were assumed to exhibit co-ordinate biogenesis.
- If target gene ('Gene B') of miRNA-B (secondary effector) had 2 or more binding sites of miRNA-B without any known binding sites of miRNA-A on 3' UTR, 'Gene B' may be considered as 'secondary effector gene'.

These two conditions generated multiple triads including 'miRNA-A', 'miRNA-B' (secondary effector) and 'Gene B' (secondary effector gene) (Figure 3.4). This coordinate biogenesis relationship ["miRNA-A:mRNA-A—> miRNA-B:mRNA-B"] derived between miRNA-A and miRNA-B considering the binding site information data may be utilized to determine the coordinate biogenesis network and in turn the miRNA–mRNA meta-interaction network for different CB regulators in *Mus musculus* and *Homo sapiens*. Additionally, a prevalence count corresponding to different miRNA coordinate biogenesis pairs that occur across multiple 'Gene A' may be determined to assign precedence to coordinate biogenesis pairs that are more relevant. Subsequently, a set of regulator-target coordinate biogenesis relationships have been predicted for validation and large scale expression analysis datasets have been considered for pruning the list of regulator-target relationships and validating the same.

3.2.1.2 Predicting and contextualizing coordinate biogenesis regulatory relationships for experimental validation

The coordinate biogenesis relationship ["miRNA-A:mRNA-A—>miRNA-B:mRNA-B"] has been derived considering miR-146a-5p as the CB regulator to determine the miRNA–mRNA meta-interaction network for miR-146a-5p in *Mus musculus* and *Homo sapiens*, respectively. Herein, the alterations in the expression level of mRNA-B (secondary effector gene) may indirectly indicate probable changes in miRNA-B biogenesis

(secondary effector) mediated by miRNA-146a-5p. Thus, expression information mapping may be performed in *Homo sapiens* and *Mus musculus* coordinate biogenesis regulator-target relationship data for miR-146a-5p to select some miRNA-B (secondary effector) and mRNA-B (secondary effector gene) for further study in this context.

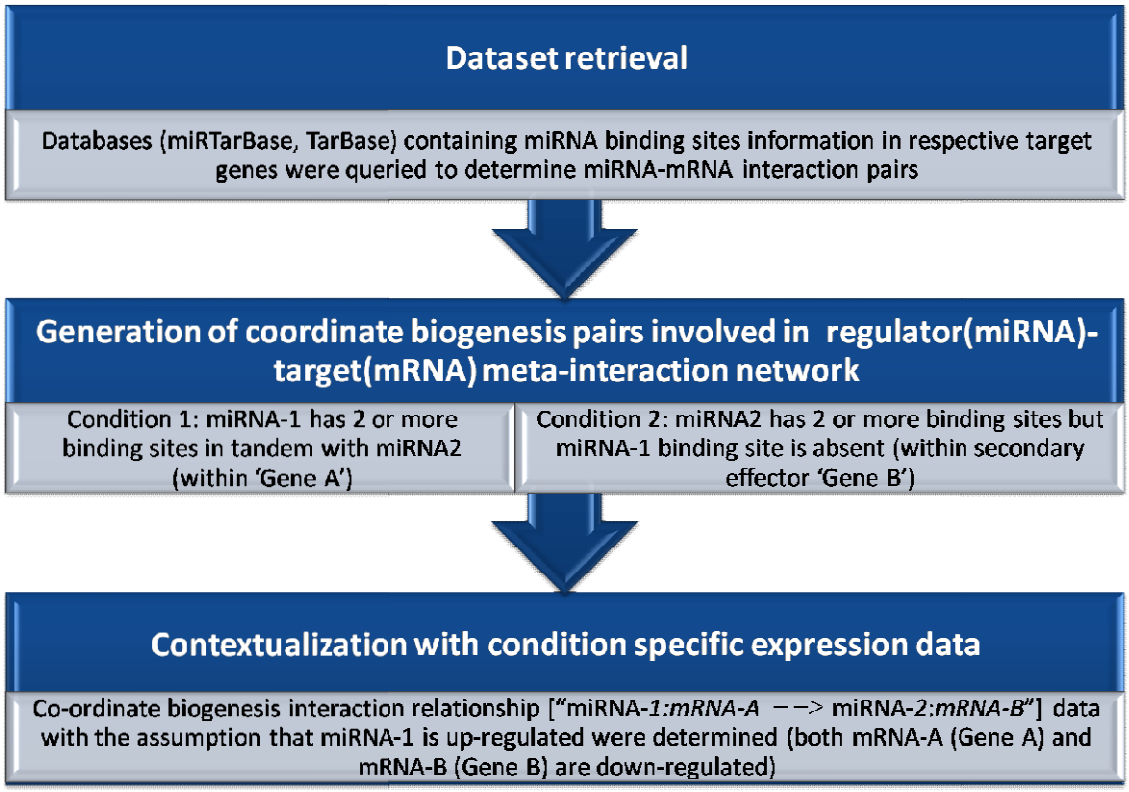


Figure 3.4: **Schematic representation of the methodology to predict model CB regulator-target relationships.** CB regulator-target relationships may be predicted based on miRNA-mRNA network analysis and miRNA binding site data analysis. Herein, mRNA- A (Gene A) bears two or more bindings sites of the CB-regulator and less number of binding sites of miRNA-B (secondary effector). Additionally, mRNA B (Gene B) harbours two or more miRNA-B binding sites but no miRNA-A binding sites.

CB regulator-target relationship has been studied in the context of endotoxin response in macrophages. Expression mapping has been performed by identifying differentially expressed genes in murine macrophages and isolated human monocytes upon exposure to LPS (10ng/ml) for 24 hours (GSE19490, Schroder et al., 2012; GSE85333, Regan et al., 2018). Differentially expressed mRNAs in each case were determined by considering a fold change threshold of 1.5 and p-value threshold of 0.05. CB regulator-target relationships in which secondary effector gene (Gene B) and primary target for miR-146a (Gene A) were both found to be down-regulated were selected considering "miR-146a mediated cooperative repression". Further, coordinate biogenesis miRNA pairs occurring in genes wherein 'Gene A' has a direct regulatory relationship with 'Gene B'

(for eg. 'Gene B' is a target gene of transcription factor 'Gene A') have been filtered out in the selected validation set.

3.2.1.3 Experimental validation of CB regulator-target relationships utilizing large scale expression profiling

Small RNA sequencing analysis was performed [at Nucleome Informatics by collaborator Dr. S. N. Bhattacharyya] to determine differentially expressed miRNA when miR-146a is over-expressed in RAW264.7 cells. This experiment was performed to identify miRNA-B whose expressions are influenced when the CB regulator (miR-146a) expression is up-regulated. Analysis of this sequencing data was performed to determine up-regulated miRNA that had fold change (F.C.) > 0 and p-value <= 0.05. Since CB regulator expression is up-regulated it should in turn up-regulate the expression of miRNA-B by co-ordinate biogenesis. Utilizing miRNA expression profiling data CB-regulator target relationships for miR-146a-5p were validated. CB regulator-target relationships observed in macrophages responding to LPS exposure wherein miR-146a is likely to be over-expressed has further been determined.

Discussion

Detailed analysis of intra-cellular and inter-cellular interactions may provide insights about the complex or flexible nature of these cellular interactions. In particular, different bio-molecules (proteins and nucleic acids) in the intra-cellular or inter-cellular milieu interact to bring about different cellular processes. In this thesis dissertation, these repertoires of cellular interactions have been explored in terms of either interaction networks or interacting complexes.

The repertoire of intra-cellular interactions has been studied primarily at the systems level to predict whether the interactome networks (protein-protein interaction network, miRNA-mRNA regulatory network, RNA-protein regulatory network) might change under a diseased scenario. Complexities in intra-cellular interactions have been studied by determining the effect of perturbations on the intra-cellular network architecture or its components. In this regard, different hypothesis were considered and their validity was studied in a disease perspective utilizing a number of case studies. During these analyses, a key observation obtained was that essential protein(s) for a cellular process may lie at the interface of the gene regulatory and protein-protein interaction network. This observation was determined by over-expressing a key transcriptional activator and delineating the intra-cellular network governing the particular cellular process by performing topological analysis of the network. Subsequently, the miRNA-mRNA interaction network pertaining to an essential regulator was studied to identify whether substantial changes in the regulatory network architecture may occur upon alterations in the level of the essential regulator. Herein, it was identified that mRNA encoding for crucial signalling cross-talk proteins in the regulatory network get altered only upon substantial changes in the levels of the essential regulator over a certain time period. Based on these analyses, it was ascertained that while intra-cellular networks have multiple levels of regulation and are highly resilient to perturbations however, over time gradual changes in the network components may result in cellular physiological changes. Further, it is also plausible that alterations in network architecture or regulatory relationship in miRNA-mRNA or protein-mRNA interaction networks may be associated with alterations in cellular phenotype or response. Moreover, the likelihood that miRNA can modulate mRNA expression levels by regulating the levels of intermediate miRNA was analyzed by constructing miRNA-mRNA meta-interaction networks. In this manner, by analysing intra-cellular interactions networks in multiple contexts essential network components (proteins/mRNA) and important network modules were predicted. Such essential network components or modules when de-regulated are likely to be associated with changes in cellular function or phenotype. Thus, by identifying disease associated entities and analyzing interaction patterns among them different themes likely to be prevalent in intra-cellular interactions among bio-molecular entities were obtained.

A range of proteins are likely to participate in inter-cellular interactions governing

cellular signalling cascades and in turn the interaction between these proteins leads to a cellular response. Evolutionary similarities in sequences may be considered to identify the function of a protein, predict protein interaction partner(s) or identify crucial residues for functional conservation. In this respect, functional annotation of proteins possibly involved in inter-cellular interactions between host cells and pathogens was performed with the help of a methodology to predict novel virulence factors in a pathogen proteome. Based on this analysis it could further be postulated that phylogenetically distant organisms may share similar virulence factors or invasion mechanisms. Moreover, an information theory based evolutionary approach was utilised to identify inter-dependent protein residues or sequence positions crucial for the conserved interaction between proteins involved in inter-cellular interactions. This, measure to study co-variation (Co-Var) has been validated in the context of protein-protein co-evolution. Inter-protein co-evolution studies predicted that a certain fraction of co-evolving residue pairs may occur in spatially separated positions. Subsequently, co-evolution analysis of inter-cellular protein interaction complexes involved in cancer metastasis identified that certain receptor positions share many co-evolutionary pairing connections with certain ligand positions only. These co-evolutionary pairings which occur at non-interface regions predominantly occur among residues present in functional domains or residues important in intra-protein co-evolution as well. Moreover, by studying altered residue pairing propensity among these high degree co-evolving positions considering mis-sense substitution mutations in cancer it may be suggested that absence of co-ordinated changes at these positions could contribute to disease associated alterations in these interaction complexes. Thus, with the help of this sequence analysis approach presented one can envisage residue pairing interactions important for an interaction to occur between a pair of proteins.

Therefore, in conclusion, different network based approaches can be utilised to determine important components in a network from a systems biology perspective and sequence analysis based approaches may be useful in determining important molecular connections between interacting network components.

Supplementary Figures and Tables

Figure S1: **Directly regulated FOXJ1 regulatory network genes.** Predicted FOXJ1 directly regulated genes with FOXJ1 binding motif(s) in +/-6kb of their transcription start site (TSS) are shown here. Only few genes among 424 directly regulated genes have been depicted here.

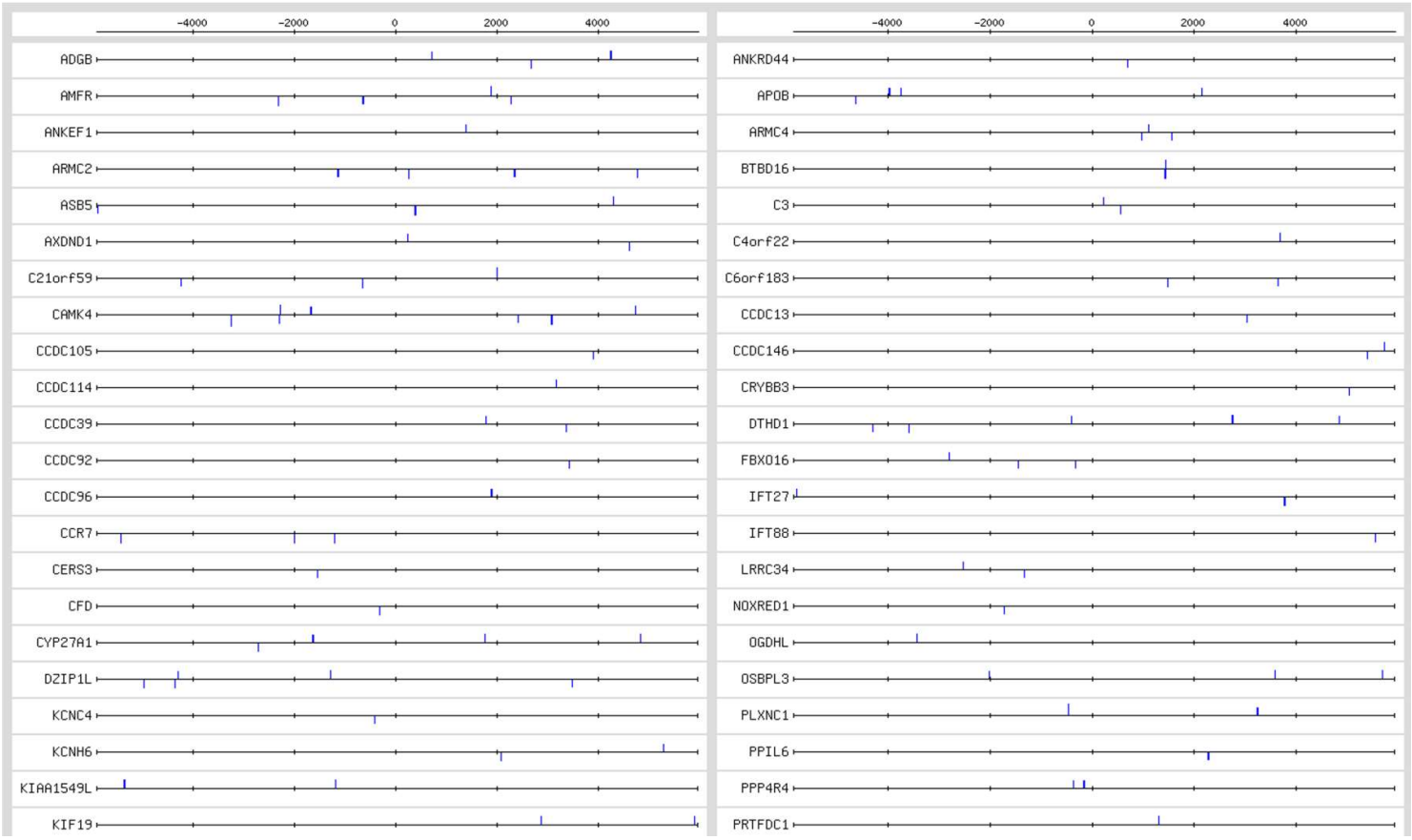


Table S1: **Predicted ciliary associations for FOXJ1 regulatory network genes based on Gene Ontology (GO) analysis.** Statistics for predicted ciliary functions of FOXJ1 regulatory network genes based on FGNet (Aibar et al., 2015) functional annotation clustering of FIG mapped to different GO categories are shown here.

Cluster[1]	GO based ciliary association	Category[2]	Terms[3]	Cluster Enrichment Score[4]	P-value[5]	Genes[6]	Fold Enrichment[7]
1	Microtubule component or ciliary motility	Molecular function	GO:0003777: microtubule motor activity	7.408	8.99E-10	DNAH11,DNAH10,KIF17,DNAH17,KIF18B,DNAH1,DNAH2,KIF9,DNAH7,DNAH8,DNAH5,DNAH6,DNALI1,KIF6,DYNC2H1,KIF19	8.096
		Cellular component	GO:0005874: microtubule		1.01E-12	DNAH11,DNAH10,TPPP3,TTLL9,DNAH17,TTLL6,DNAH1,TTLL7,DNAH2,KIF9,DNAH7,DNAH8,SPAG17,DNAH5,TTLL11,DNAH6,TPGS1,DNAI1,KIF6,TEKT1,SYBU,DYNC2H1,TEKT2,TUBA3D,TEKT4,KIF17,TBCE,KIF18B,GAS8,EML5,CSPP1,SPAG6,KIF19,KATNAL2	4.518
			GO:0005858: axonemal dynein complex		3.70E-08	DNALI1,DNAH17,DNAH1,DNAH2,DNAH7,DNAH8,DNAH6	28.927
			GO:0030286: dynein complex		1.03E-05	DNAH11,DNAH10,DNAH17,DYNC2H1,DNAH1,DNAH2,DNAH8	13.149
		Biological process	GO:0060285: cilium-dependent cell motility		1.32E-09	DNAAF2,DNAH17,DRC1,DNAH1,DNAH2,CCDC39,DNAH7,DNAH8	30.229
			GO:0007018:microtubule-based movement		4.92E-07	DNAH10,KIF17,DNAH17,KIF18B,DNAH2,KIF9,DNAH8,DNAH5,DNAH6,KIF6,KIFAP3,DYNC2H1,KIF19	6.671
2	Cilium localised or smoothened signaling pathway	Biological process	GO:0007224: smoothened signaling pathway	5.662	0.00131	IFT80,HSPB11,BBS7,TTC26,IFT172,IFT27,PTPDC1,IQUB	4.82
			GO:0042073:		0.00609	TRAF3IP1,HSPB11,TTC26,IFT27	10.391

			intraciliary transport				
		Cellular component	GO:0097542: ciliary tip		7.24E-08	TRAF3IP1,IFT80,HSPB11,TTC26,KIFAP3,IFT172,DYNC2H1,IFT27,WDR35,IFT88,IFT140	10.331
			GO:0072372: primary cilium		4.74E-10	IFT80,KIF17,TTC26,RABGAP1L,SPA17,TRAF3IP1,RILPL2,HSPB11,KIFAP3,IFT172,DYNC2H1,NEK8,WDR35,IFT27,IFT88,IFT140	8.477
			GO:0030992: intraciliary transport particle B		1.79E-07	TRAF3IP1,IFT80,HSPB11,KIF17,TTC26,IFT172,IFT27,IFT88	17.4
3	Axonemal dynein complex assembly or epithelial cilium movement involved in L/R assymmetry determination	Biological process	GO:0070286: axonemal dynein complex assembly	5.232	2.37E-14	CCDC103,CCDC40,DNAAF3,DNAAF2,DNAAF1,DRC1,CCDC151,SPAG1,PIH1D3,CCDC39	41.564
			GO:0036159: inner dynein arm assembly		1.59E-12	CCDC103,LRRC6,CCDC40,DYX1C1,DNAAF1,ZMYND10,TEKT2,DNAH1,CCDC39,DNAH7	31.973
			GO:0044458: motile cilium assembly		7.67E-07	LRRC6,CCDC40,DNAAF3,DNAAF1,ZMYND10,CCDC39,RSPH9	19.397
			GO:0003356: regulation of cilium beat frequency		1.60E-06	DNAH11,CCDC40,DNAAF1,ARMC4,CCDC39	41.564
			GO:0060287: epithelial cilium movement involved in determination of left/right asymmetry		2.12E-05	CCDC103,LRRC6,CCDC40,DNAAF1,CCDC39	25.978
			GO:0071907: determination of digestive tract left/right asymmetry		2.60E-04	CCDC103,CCDC40,DNAAF1,CCDC39	27.71
			GO:0035469: determination of pancreatic left/right asymmetry		0.00548	CCDC40,DNAAF1,CCDC39	24.939

			GO:0071910: determination of liver left/right asymmetry		0.00808	CCDC40,DNAAF1,CCDC39	20.782
4	Intraciliary retrograde transport or ciliary protein localization	Biological process	GO:0061512: protein localization to cilium	2.595	7.66E-04	ARF4,WDR35,TTC21A,TBC1D32,IFT140	11.546
			GO:0035721: intraciliary retrograde transport		0.00145	DYNC2H1,WDR35,TTC21A,IFT140	16.626
5	Nucleoside diphosphate kinase activity	Molecular function	GO:0004550: nucleoside diphosphate kinase activity	2.061	0.00129	NME5,AK7,NME9,AK9,AK8	10.121

[1]**Cluster**: Rank of the cluster or group of genes according to cluster enrichment score
[2]**Category**: Enriched gene ontology category
[3]**Terms**: Description of enriched gene ontology categories
[4]**Cluster Enrichment Score**: It is the geometric mean (in -log scale) of gene member's EASE score (Fischer exact p-values) in a corresponding annotation cluster
[5]**P-value**: It is the modified fisher exact p-value associated with the enriched GO annotation terms that belong to a gene group or cluster. Fisher exact probability weights significance in favor of GO themes supported by more genes
[6]**Genes**: Genes belonging to an enriched cluster or category
[7]**Fold enrichment**: Ratio based on the fisher exact probability of genes mapping to a particular GO category on the gene list as compared to the ratio of genes in that category within the background population.
[Note: published in Mukherjee et al., 2019]

Table S2: **Important interacting proteins (IIP) in the FIG-sub-network identified with the help of IIP analysis.** Multiple graph theory measures formulated on degree, shortest path and centrality were utilized to identify IIP among hub, bottleneck, central, local network perturbing and global network perturbing proteins.

IIP Categories	Gene Name (Official Gene Symbol)	Protein/Function	Choksi et al. Expression Study (Up-regulated)	PCD Case Study (Differentially Expressed)	Human Protein Atlas Expression Data				Ciliary Role Related Reference Paper
					Brain	Lung	Fallopian tube	Testis	
HUB & BP & CP & GNPP	NEDD4	E3 ubiquitin-protein ligase NEDD4							
	CREBBP	CREB-binding protein							
	REL	Proto-oncogene c-Rel							
	PTEN	Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase PTEN							Shnitsar et al., 2015
	CACNA1A	Voltage-dependent P/Q-type calcium channel subunit alpha-1A							
HUB & BP & CP & LNPP	UBE2I	SUMO-conjugating enzyme UBC9							Malicki & Johnson, 2017
HUB & CP & GNPP	ACTB	Actin, cytoplasmic 1							
	ABL1	Tyrosine-protein kinase ABL1							
HUB & BP & GNPP	CDC5L	Cell division cycle 5-like protein							
HUB & BP & CP	APP A4 AD1	Amyloid beta A4 protein							

	GRB2 ASH	Growth factor receptor-bound protein 2 (Adapter protein GRB2)							
	TP53 P53	Cellular tumor antigen p53							
	HSP90AA1 HSP90A HSPC1 HSPCA	Heat shock protein HSP 90-alpha							Wang et al., 2015
	BRCA1 RNF53	Breast cancer type 1 susceptibility protein							
	EGFR ERBB ERBB1 HER1	Epidermal growth factor receptor							Kasahara et al., 2018
	LNX1 LNX PDZRN2 UNQ574/PRO1136	E3 ubiquitin-protein ligase							
	ESR1 ESR NR3A1	Estrogen receptor (ER) (ER-alpha)							
	CSNK2A1 CK2A1	Casein kinase II subunit alpha							
	SUMO1 SMT3C SMT3H3 UBL1 OK/SW-cl.43	Small ubiquitin-related modifier 1 (SUMO-1)							
	NEDD4L KIAA0439 NEDL3	E3 ubiquitin-protein ligase NEDD4-like							
	EP300 P300	Histone acetyltransferase p300							
	UBC	Polyubiquitin-C							
	ATXN1 ATX1 SCA1	Ataxin-1							

	HSPB1 HSP27 HSP28	Heat shock protein beta-1							
	CTNNB1 CTNNB OK/SW-cl.35 PRO2286	Catenin beta-1 (Beta-catenin)							Gerdes et al., 2007; VanHook 2012; Caron et al., 2012
	SRPK2	SRSF protein kinase 2							
	MDM2	E3 ubiquitin-protein ligase Mdm2							
	HDAC1 RPD3L1	Histone deacetylase 1 (HD1)							Cao et al., 2009
	TRAF2 TRAP3	TNF receptor-associated factor 2							
	SUMO2 SMT3B SMT3H2	Small ubiquitin-related modifier 2 (SUMO-2)							
	UBE3A E6AP EPVE6AP HPVE6A	Ubiquitin-protein ligase E3A							
	AKT1 PKB RAC	RAC-alpha serine/threonine-protein kinase							Suizu et al., 2016
	PIN1	Peptidyl-prolyl cis-trans isomerase NIMA-interacting 1							
	CSNK2B CK2N G5A	Casein kinase II subunit beta (CK II beta)							
	SMAD3 MADH3	Mothers against decapentaplegic homolog 3 (MAD homolog 3)							Clement et al., 2013
	MYC BHLHE39	Myc proto-oncogene protein							Wu et al., 2013

	HGS HRS	Hepatocyte growth factor-regulated tyrosine kinase substrate							
	AR DHTR NR3C4	Androgen receptor (Dihydrotestosterone receptor)							
	SRC SRC1	Proto-oncogene tyrosine-protein kinase Src							Bershteyn et al., 2010
	IKBKG FIP3 NEMO	NF-kappa-B essential modulator (NEMO) (FIP-3)							Danescu et al., 2018
	FYN	Tyrosine-protein kinase Fyn							
	CASP8 MCH5	Caspase-8 (CASP-8)							
	RAC1 TC25 MIG5	Ras-related C3 botulinum toxin substrate 1							Epting et al., 2015
	MEOX2 GAX MOX2	Homeobox protein MOX-2 (Growth arrest-specific homeobox)							
	PRKN PARK2	E3 ubiquitin-protein ligase parkin (Parkin)							
	EEF1A1 EEF1A EF1A LENG7	Elongation factor 1-alpha 1 (EF-1-alpha-1)							
	LRIF1 C1orf103 RIF1	Ligand-dependent nuclear receptor-interacting factor 1							
	HSPA4 APG2	Heat shock 70 kDa protein 4							
	UBQLN1 DA41 PLIC1	Ubiquilin-1 (Protein linking IAP with cytoskeleton 1)							
	NFKB1	Nuclear factor NF-kappa-B p105 subunit (DNA-binding factor KBF1)							

	BTRC BTRCP FBW1A FBXW1A	F-box/WD repeat-containing protein 1A							
	UBB	Polyubiquitin-B [Cleaved into: Ubiquitin]							
	GOLGA2	Golgin subfamily A member 2							Hurtado et al., 2011
	STOM BND7 EPB72	Erythrocyte band 7 integral membrane protein (Stomatin)							
	RELA NFKB3	Transcription factor p65 (Nuclear factor NF-kappa-B p65 subunit)							
	TRAF6 RNF85	TNF receptor-associated factor 6							
	TRIM27 RFP RNF76	Zinc finger protein RFP							
	VCP	Transitional endoplasmic reticulum ATPase (TER ATPase)							Raman et al., 2015
	KAT5 HTATIP TIP60	Histone acetyltransferase KAT5							
HUB & LNPP	TRIM54 MURF RNF30 muRF3 MURF-3	tripartite motif containing 54							
HUB & GNPP	A2M CPAMD5, FWP007	Alpha-2-macroglobulin							
CP & GNPP	ACTN2	Alpha-actinin-2							Kohli et al., 2017
	LCAT	Phosphatidylcholine-sterol acyltransferase							Kim et al., 2010

	ACTN1	Alpha-actinin-1							Pan et al., 2007
HUB & CP	SMN1 SMN SMNT SMN2 SMNC	Survival motor neuron protein (Component of gems 1) (Gemin-1)							
	TK1 TK2	thymidine kinase 1							
	H2AFX H2AX H2A.X H2A/X	H2A histone family member X							
	CDKN1A P21 CIP1 SDI1 WAF1 CAP20 CDKN1 MDA-6 p21CIP1	cyclin dependent kinase inhibitor 1A							
	CDC37 CDC37A	Hsp90 co-chaperone Cdc37							
	KAT2B PCAF	Histone acetyltransferase KAT2B							
	RB1	Retinoblastoma-associated protein (p105-Rb)							
	E2F1 RBBP3	Transcription factor E2F1 (E2F-1) (PBR3)							
	PPP1CA PPP1A	Serine/threonine-protein phosphatase PP1-alpha catalytic subunit (PP-1A)							
	YWHAQ	14-3-3 protein theta (14-3-3 protein T-cell)							
	TRAF1 EBI6	TNF receptor-associated factor 1							
	CBL CBL2 RNF55	E3 ubiquitin-protein ligase CBL							Schmid et al., 2018
	XIAP API3 BIRC4 IAP3	E3 ubiquitin-protein ligase XIAP							Kim et al., 2010

	KPNA2 RCH1 SRP1	Importin subunit alpha-1 (Karyopherin subunit alpha-2)							
	VIM	Vimentin							
	DISC1 KIAA0457	Disrupted in schizophrenia 1 protein							Marley et al., 2010
	WDYHV1 C8orf32 NTAQ1	Protein N-terminal glutamine amidohydrolase							
	SOCS3 CIS3 SSI3	Suppressor of cytokine signaling 3 (SOCS-3)							
	HDAC2	Histone deacetylase 2 (HD2)							Kobayashi et al., 2017
	UBE2D1 SFT UBC5A UBCH5 UBCH5A	Ubiquitin-conjugating enzyme E2 D1							
	XRCC6 G22P1	X-ray repair cross-complementing protein 6							
	UBE2D3 UBC5C UBCH5C	Ubiquitin-conjugating enzyme E2 D3							
	PIAS4 PIASG	E3 SUMO-protein ligase PIAS4							Ramachandran et al, 2015
	CDK1 CDC2 CDC28A CDKN1 P34CDC2	Cyclin-dependent kinase 1 (CDK1)							Fabregat et al., 2018; Croft et al., 2014
	SSX2IP KIAA0923	Afadin- and alpha-actinin-binding protein (Afadin DIL domain-interacting protein)							Klinger et al., 2014
	JUN	Transcription factor AP-1 (Activator protein 1)							

	GSK3B	Glycogen synthase kinase-3 beta (GSK-3 beta)							Thoma et al., 2007
	YWHAZ	14-3-3 protein zeta/delta (Protein kinase C inhibitor protein 1)							
	VHL	Von Hippel-Lindau disease tumor suppressor							Lutz and Burk, 2006
	MAPK6 ERK3 PRKM6	Mitogen-activated protein kinase 6 (MAP kinase 6)							
	PIK3R1 GRB1	Phosphatidylinositol 3-kinase regulatory subunit alpha (PI3-kinase regulatory subunit alpha)							Han et al., 2014
	SMAD4 DPC4 MADH4	Mothers against decapentaplegic homolog 4 (MAD homolog 4)							Clement et al, 2013
	APOE	Apolipoprotein E							
	CDC42 TKS G25K CDC42Hs	cell division cycle 42							Choi et al, 2013
	TERF1 PIN2 TRBF1 TRF TRF1	Telomeric repeat-binding factor 1 (NIMA-interacting protein 2)							
BP & CP	BHLHE40 BHLHB2 DEC1 SHARP2 STRA13	Class E basic helix-loop-helix protein 40							
HUB & BP	MCM7 CDC47 MCM2	DNA replication licensing factor MCM7							
	APPBP2 KIAA0228 PAT1	Amyloid protein-binding protein 2 (Amyloid beta precursor protein-binding							

		protein 2)							
	FBXW7 FBW7 FBX30 SEL10	F-box/WD repeat-containing protein 7 (Archipelago homolog)							Maskey et al., 2015
	UBE2N BLU	Ubiquitin-conjugating enzyme E2 N (Bendless-like ubiquitin-conjugating enzyme)							
	HSPA8 HSC70 HSP73 HSPA10	Heat shock cognate 71 kDa protein							Bhowmick et al., 2009
	FKBP5 AIG6 FKBP51	Peptidyl-prolyl cis-trans isomerase FKBP5 (PPIase FKBP5)							
	PRPF40A FBP11 FLAF1 FNBP3 HIP10 HYPA HSPC225	Pre-mRNA-processing factor 40 homolog A (Fas ligand-associated factor 1) (Formin-binding protein 11)							
	BAG3 BIS	BAG family molecular chaperone regulator 3 (BAG-3) (Bcl-2-associated athanogene 3)							
	ACD PIP1 PTOP TINT1 TPP1	Adrenocortical dysplasia protein homolog (POT1 and TIN2-interacting protein)							
	FAM19A3 TAFA3	Protein FAM19A3 (Chemokine-like protein TAFA-3)							
	HSPA1A HSP72 HSPA1 HSX70	Heat shock 70 kDa protein 1A (Heat shock 70 kDa protein 1) (HSP70-1) (HSP70.1)							

	KDM1A AOF2 KDM1 KIAA0601 LSD1	Lysine-specific histone demethylase 1A							
	TRAF3 CAP1 CRAF1	TNF receptor-associated factor 3							
	NCOA3 AIB1 BHLHE42 RAC3 TRAM1	Nuclear receptor coactivator 3 (NCoA-3)							
	STUB1 CHIP PP1131	E3 ubiquitin-protein ligase CHIP							
	RBL1	Retinoblastoma-like protein 1							
	DLG4 PSD95	Disks large homolog 4 (Postsynaptic density protein 95)							
	CDC27 ANAPC3 D0S1430E D17S978E	Cell division cycle protein 27 homolog (Anaphase-promoting complex subunit 3)							Wheway et al., 2015
	NLGN3 KIAA1480 NL3	Neuroligin-3 (Gliotactin homolog)							
	SYNCRIP HNRPQ NSAP1	Heterogeneous nuclear ribonucleoprotein Q (hnRNP Q)							
	PLSCR1	Phospholipid scramblase 1 (PL scramblase 1)							Choksi et al., 2014

Abbreviations: BP - Bottleneck protein, CP - Central protein, LNPP - Local network perturbing protein, GNPP- Global network perturbing protein

IIP with previously reported ciliary roles [Note: published in Mukherjee et al., 2019]

Protein is differentially expressed in study or expressed in organs comprised of ciliated cells

Table S3: **FOXJ1 regulatory network proteins along with IIP participate in multiple cellular pathways.** IIP interact with a number of FOXJ1 regulatory network proteins (246 FIGp) and these clusters of proteins are likely to be involved in enriched cellular pathways listed here.

[1]Enriched Pathway ID	[2]Enriched Reactome pathways	[3]P-value	Gene Ratio	Genes
168256	Immune System	2.42E-11	0.3459119	ABL1/ACTB/ACTR3/AKT1/APP/BAIAP2/BTRC/C3/CAMK4/CASP8/CBL/CDC42/CDK1/CDKN1A/CREBBP/CTNNB1/CTSB/EGFR/EP300/FYN/GRB2/GSK3B/HSP90AA1/IKBKG/JUN/KIFAP3/KPNA2/MASP1/MDM2/NEDD4/NFKB1/PELI1/PIK3R1/PIN1/PTEN/RAC1/RASGRP2/REL/RELA/RNF125/SOCS3/SRC/SUMO1/TEC/TRAF2/TRAF3/TRAF6/UBB/UBC/UBE2D1/UBE2D3/UBE2N/XAF1/XRCC6/YWHAZ
1280218	Adaptive Immune System	6.83E-06	0.163522	AKT1/BTRC/C3/CBL/CDC42/CDKN1A/EGFR/FYN/GRB2/GSK3B/IKBKG/KIFAP3/MDM2/NFKB1/PIK3R1/PTEN/RAC1/RASGRP2/REL/RELA/SRC/TRAF6/UBB/UBC/UBE2N/YWHAZ
212436	Generic Transcription Pathway	3.80E-10	0.1383648	AKT1/AR/CREBBP/E2F5/ESR1/HDAC1/KAT2B/MYC/NEDD4L/NR1H2/PTEN/RBL1/RRM2B/SMAD3/SMAD4/TP53/UBB/UBC/UBE2D1/UBE2D3/YWHAQ/YWHAZ
162582	Signal Transduction	7.53E-06	0.4716981	A2M/ABL1/ACTB/ACTR3/AKT1/APOB/APOE/APP/AR/ARAP1/BAIAP2/BTRC/C3/CAMK4/CASP8/CBL/CDC37/CDC42/CDK1/CDKN1A/CREBBP/CSNK2A1/CSNK2B/CTNNB1/DLG4/E2F1/E2F5/EGFR/EP300/ESR1/FBXW7/FYN/GRB2/GSK3B/HDAC1/HDAC2/HGS/HSP90AA1/HSPB1/KAT2B/KAT5/KDM1A/MDM2/MYC/MYO9A/NEDD4/NEDD4L/NFKB1/PIK3R1/PPP1CA/PTEN/PTGER4/RAC1/RASGRP2/RBL1/RELA/RHO/SMAD3/SMAD4/SOCS3/SRC/STUB1/TEC/TJP2/TP53/TRAF2/TRAF6/UBB/UBC/UBE2D1/UBE2D3/VCP/XIAP/YWHAQ/YWHAZ
1643685	Disease	3.87E-06	0.2201258	AKT1/APP/BTRC/CBL/CDC37/CDKN1A/CREBBP/CTNNB1/EGFR/EP300/FBXW7/FYN/GRB2/GSK3B/GSN/HDAC1/HSP90AA1/IKBKG/KAT2B/MDM2/MYC/NEDD4L/NFKB1/PIK3R1/RAC1/RELA/SLC25A4/SMAD3/SMAD4/SYT2/TRAF3/UBB/UBC/VCP/XRCC6
2262752	Cellular responses to stress	4.91E-09	0.1509434	ACD/BAG3/CDC27/CDKN1A/CREBBP/E2F1/EP300/GSK3B/HSBP1/HSP90AA1/HSPA1A/HSPA8/JUN/MDM2/NFKB1/RELA/TERF1/TP53/UBB/UBC/UBE2D1/UBE2D3/VCP/VHL

1640170	Cell Cycle	3.45E-08	0.2075472	ACD/AKAP9/BRCA1/BTRC/CCNA1/CDC27/CDK1/CDKN1A/CETN2/CSNK2A1/CSNK2B/E2F1/E2F5/EP300/GOLGA2/H2AFX/HDAC1/HSP90AA1/LMNA/MCM7/MDM2/RB1/RBL1/SMC1A/SYNE1/TERF1/TP53/UBB/UBC/UBE2D1/UBE2I/YWHAQ/YWHAZ
5357801	Programmed Cell Death	2.86E-06	0.0943396	AKT1/CASP8/CTNNB1/E2F1/GSN/LMNA/TJP2/TP53/TRAF2/UBB/UBC/VIM/XIAP/YWHAQ/YWHAZ
1266738	Developmental Biology	9.52E-06	0.163522	ABL1/ACTB/ACTR3/AKT1/CDC42/CDK1/CREBBP/CRMP1/CSNK2A1/CSNK2B/CTNNB1/DLG4/EGFR/EP300/EZR/FYN/GRB2/GRIN2B/GSK3B/HSP90AA1/HSPA8/NCOA3/RAC1/SMAD3/SMAD4/SRC

[1]**Enriched Pathway ID:** Enriched reactome pathway ID
[2]**Enriched Reactome Pathway**: Enriched Reactome pathway description
[3]**P-value**: It denotes whether any terms annotated to a specified list of genes occurs at frequency greater than that would be expected by chance. The p-value is calculated using the hypergeometric distribution
[Note: published in Mukherjee et al., 2019]

Table S4: **Enriched pathways including 'cilia associated signaling pathways' that IIP-effectors are likely to be involved in.** Statistics for the pathway enrichment analysis identifying the involvement of IIP-effectors and their interacting proteins in cilia associated signaling pathways along their parent Reactome pathways (Fabregat et al., 2017; Croft et al., 2013) is exemplified here.

[1]IIP-effector	[2]Enriched/Parent Pathway	[3]Enriched Pathway (Child Process)	[4]P-value (Enriched Parent Pathway)	[5]P-value (Enriched Child Process)	[6]Genes
Btrc	Toll-Like Receptors Cascades (R-HSA-168898)		9.64E-12		MAPK14/FBXW11/IKBKB/IRAK1/NFKB1/NFKBIA/RELA/SKP1/UBC/UBE2D1/UBE2D2/UBE2D3/CUL1/IKBKG/BTRC
	Signaling by TGF-beta family members (R-HSA-9006936)	R-HSA-170834 (Signaling by TGF-beta Receptor Complex)	1.84E-09	4.57E-11	WWTR1/SMAD3/SMAD4/MYC/SMURF1/SMURF2/SP1/UBC/UBE2D1/UBE2D3/NCOR2

	Cell Cycle, Mitotic (R-HSA-69278)	R-HSA-453274 (Mitotic G2-G2/M phases); R-HSA-453279 (Mitotic G1-G1/S phases) ; R-HSA-453276 (Regulation of mitotic cell cycle)		2.68E-10, 8.5E-07, 5.0E-09	E2F1/EP300/FBXW11/HSP90AA1/SKP1/AURKA/TP53/UBC/CUL1/CCNB1/BTRC/CDK1/CDC25A/CDC25B/RBX1/MYC/SKP2/CCNE1/UBE2D1/CDC20
	Signaling by WNT (R-HSA-195721)	R-HSA-195253 (Degradation of beta-catenin by the destruction complex)		5.11E-08	CTNNB1/GSK3B/APC/SKP1/UBC/AXIN1/CUL1/BTRC/RBX1
	Signaling by NOTCH (R-HSA-157118)		6.04E-08		E2F1/EP300/HEY2/MYC/SKP1/TP53/UBC/CUL1/NCOR2/RBX1
	Signaling by Hedgehog (R-HSA-5358351)		9.62E-07		INTU/GSK3B/SMURF1/SMURF2/SKP1/UBC/VCP/CUL1/BTRC/RBX1
Cdc42	Signaling by Rho GTPases (R-HSA-194315)	R-HSA-195258 (RHO GTPase Effectors); R-HSA-194840 (Rho GTPase cycle)	2.12E-15	1.70E-07, 5.81E-10	BAIAP2/RALBP1/A2M/GRB2/ARHGDIA/MYO9A/NCF2/PAK1/PAK2/ITSN1/VAV1/WAS/WIPF1/IQGAP1/ARHGEF7/WASL/TRIP10/ARHGAP32/CDC42/CDH1
	Membrane Trafficking (R-HSA-199991)	R-HSA-8856828 (Clathrin-mediated endocytosis)		2.18E-07	GRB2/ARRB1/ARRB2/ITSN1/UBC/CBL/WASL/TRIP10
Casp8	Toll-Like Receptors Cascades (R-HSA-168898)		1.35E-09		NOD1/RIPK3/CHUK/TICAM1/BIRC2/BIRC3/IKBKB/MAPK1/TRAF6/UBC/CASP8/RIPK1/FADD
Pias4	Signaling by TGF-beta family members (R-HSA-9006936)	R-HSA-170834 (Signaling by TGF-beta Receptor Complex)	1.16E-09	2.44E-08	PARP1/HDAC1/SMAD1/SMAD2/SMAD3/SMAD4/SMAD6/SMAD7/PRKCZ/SKIL
Socs3	Toll-Like Receptors Cascades (R-HSA-168898)		1.67E-07		TBK1/APP/NFKB1/NFKBIA/MAPK11/PTPN11/RELA/MAP3K7/TRAF6

Syncrip	Signaling by FGFR (R-HSA-190236)	R-HSA-5654738 (Signaling by FGFR2)		5.98E-07	GRB2/HNRNPA1/HNRNPH1/PLCG1/ PTPN11/RPS27A
Terf1	Cell Cycle (R-HSA-1640170)	R-HSA-1500620 (Meiosis)		3.33E-07	STAG1/POT1/TINF2/ATM/TERF2IP/ ACD/BRCA1/TERF1

[1]**IIP-effector**: Gene name of IIP-effector protein
[2]**Enriched/Parent Pathway**: Enriched reactome pathway description or parent pathway description in which the enriched Reactome pathway participates
[3]**Enriched Pathway (Child Process)**: Reactome pathway ID's of enriched child processes and description.
[4]**P-value (Enriched Parent Pathway)**: This p-value denotes whether any parent Reactome pathways were found to be significantly enriched. The p-value which is calculated using a hypergeometric distribution denotes whether any terms annotated to a specified list of genes occurs at frequency greater than that would be expected by chance.
[5]**P-value (Enriched Child Process)**: This p-value denotes whether any child Reactome pathways were found to be significantly enriched. The p-value which is calculated using a hypergeometric distribution denotes whether any terms annotated to a specified list of genes occurs at frequency greater than that would be expected by chance.
[6]**Genes**: Gene list that participates in the enriched Reactome pathway or enriched child processes.
[Note: published in Mukherjee et al., 2019]

Table S5 : **Differentially expressed miRNAs in the pre-frontal cortex with their respective fold changes and p-value**

S.No	miRNA ID	log_2(FoldChange)	P-value
1	hsa-miR-100-5p	1.43	0.010015112
2	hsa-miR-141-3p	1.69	0.009813196
3	hsa-miR-142-5p	1.82	1.94116E-05
4	hsa-miR-146a-5p	1.29	0.022513004
5	hsa-miR-148b-3p	2.07	0.001036918
6	hsa-miR-152	1.78	4.1408E-06
7	hsa-miR-153	1.67	1.99302E-07
8	hsa-miR-18a-5p	1.58	5.09058E-05
9	hsa-miR-19a-5p	1.46	0.017524207
10	hsa-miR-208b	1.65	6.3016E-08
11	hsa-miR-20a-3p	1.52	0.000400745
12	hsa-miR-26a-1-3p	2.03	0.003336347
13	hsa-miR-302a-3p	2.41	0.030609753

14	hsa-miR-302b-3p	2.35	0.004783156
15	hsa-miR-32-5p	1.28	6.24695E-07
16	hsa-miR-338-3p	1.32	1.84384E-06
17	hsa-miR-3613-3p	1.56	3.82106E-05
18	hsa-miR-3676-5p	1.58	0.004546007
19	hsa-miR-374a-5p	2.15	0.000217979
20	hsa-miR-374b-5p	1.47	0.000168906
21	hsa-miR-488-3p	1.31	2.9572E-08
22	hsa-miR-5000-3p	1.58	0.010159335
23	hsa-miR-500a-5p	1.44	0.036769838
24	hsa-miR-516a-5p	1.69	3.02779E-05
25	hsa-miR-519a-3p	1.29	0.00193549
26	hsa-miR-590-3p	1.27	2.14893E-06
27	hsa-miR-675-3p	2.11	0.045356182
28	hsa-miR-7-1-3p	1.52	9.75075E-10
29	hsa-miR-941	2.2	0.003752442
30	hsa-miR-99b-5p	1.45	0.003692031
31	hsa-miR-1225-3p	-2.30	5.66163E-10
32	hsa-miR-1225-5p	-1.78	0.000139448
33	hsa-miR-1276	-1.66	0.001487482
34	hsa-miR-132-3p	-2.02	2.77195E-14
35	hsa-miR-132-5p	-1.89	3.19349E-10
36	hsa-miR-212-3p	-1.7	4.75848E-08
37	hsa-miR-212-5p	-1.87	2.07458E-06
38	hsa-miR-25-5p	-1.36	0.008877855
39	hsa-miR-323a-3p	-1.30	3.65802E-07
40	hsa-miR-323a-5p	-1.53	9.3328E-06
41	hsa-miR-3607-3p	-2.76	2.6923E-10
42	hsa-miR-3651	-1.95	0.040895517
43	hsa-miR-3653	-2.08	4.95899E-07
44	hsa-miR-376a-5p	-1.49	1.97435E-05

45	hsa-miR-4284	-1.43	0.021108295
46	hsa-miR-4520a-3p	-1.71	0.003275044
47	hsa-miR-505-5p	-1.43	0.019202847
48	hsa-miR-511	-1.846	0.015080741
49	hsa-miR-5701	-2.23	1.96902E-08
50	hsa-miR-6087	-2.56	0.001535109
51	hsa-miR-6721-5p	-2.39	0.009938169
52	hsa-miR-877-5p	-1.52	1.93574E-07
53	hsa-miR-885-3p	-2.146	9.5438E-15

Table S6: **Up-regulated miRNA and corresponding up-regulated target mRNA fractions in de-regulated miRNA:mRNA interaction network in Alzheimer's disease.**

	Frontal Cortex (GSE15222)		**Superior frontal gyrus (GSE5281)**		**Dorsolateral pre-frontal cortex (GSE53697)**	
Up-regulated miRNA in AD	**Total mRNA targets**	**Up-regulated mRNA targets (Fraction)**	**Total mRNA targets**	**Up-regulated mRNA targets (Fraction)**	**Total mRNA targets**	**Up-regulated mRNA targets (Fraction)**
hsa-miR-100-5p	34	0.74	63	0.40	10	0.10
hsa-miR-141-3p	138	0.72	250	0.48	53	0.36
hsa-miR-142-5p	93	0.75	148	0.46	23	0.48
hsa-miR-146a-5p	108	0.86	147	0.50	36	0.25
hsa-miR-148b-3p	206	0.74	290	0.46	60	0.22
hsa-miR-152-3p	241	0.72	318	0.45	77	0.21
hsa-miR-152-5p	7	0.71	9	0.56	5	0.20
hsa-miR-153-3p	105	0.61	179	0.44	42	0.43
hsa-miR-153-5p	37	0.86	33	0.58	8	0.38
hsa-miR-18a-5p	262	0.72	393	0.46	79	0.29

hsa-miR-19a-5p	81	0.78	135	0.49	18	0.50
hsa-miR-208b-3p	55	0.76	68	0.49	NA	NA
hsa-miR-208b-5p	15	0.80	17	0.41	NA	NA
hsa-miR-20a-3p	215	0.78	346	0.47	60	0.15
hsa-miR-26a-1-3p	27	0.67	40	0.45	12	0.50
hsa-miR-302a-3p	132	0.77	152	0.49	40	0.35
hsa-miR-302b-3p	96	0.82	114	0.52	26	0.35
hsa-miR-32-5p	176	0.73	258	0.46	50	0.28
hsa-miR-338-3p	164	0.80	211	0.52	65	0.12
hsa-miR-3613-3p	76	0.79	109	0.60	NA	NA
hsa-miR-374a-5p	133	0.73	233	0.44	44	0.34
hsa-miR-374b-5p	167	0.62	296	0.46	54	0.39
hsa-miR-488-3p	53	0.81	66	0.47	16	0.25
hsa-miR-5000-3p	7	0.71	6	0.67	NA	NA
hsa-miR-500a-5p	48	0.73	77	0.52	12	0.50
hsa-miR-516a-5p	9	0.78	12	0.75	NA	NA
hsa-miR-519a-3p	49	0.80	69	0.48	19	0.26
hsa-miR-590-3p	176	0.72	275	0.43	61	0.46
hsa-miR-675-3p	68	0.75	121	0.50	23	0.30
hsa-miR-7-1-3p	76	0.76	125	0.51	25	0.28
hsa-miR-99b-5p	26	0.77	27	0.48	NA	NA

Table S7: **Expression profile of miR-146a-5p target mRNAs in brain cortical regions of Alzheimer's disease patients.** Fold change and p-values of differentially expressed miR-146a-5p within cortical regions of Alzheimer's disease patients determined based on differential expression analysis of mRNA profiling data is represented here.

Frontal cortex			Superior frontal gyrus			Dorsolateral pre-frontal cortex		
Target gene name	**Fold change**	**p-value**	**Target gene name**	**Fold change**	**p-value**	**Target gene name**	**Fold change**	**p-value**
brca2	1.27	8.96E-03	cxcr4	4.45	5.39E-05	tlr4	-2.34	5.96E-03
cdkn1a	1.27	3.93E-09	egfr	1.89	5.03E-03	wasf2	-1.44	2.92E-02
egfr	1.48	1.36E-03	erbb4	2.76	4.37E-07	notch2	-1.58	2.12E-02
tlr4	1.27	3.03E-08	tlr4	1.33	2.16E-03	lfng	-1.75	2.13E-02
slpi	1.64	1.71E-03	stat1	-1.42	2.39E-02	tgfb1	-2.57	2.32E-02
card10	1.45	6.02E-11	card10	1.39	4.72E-02	klf4	-3.34	2.67E-03
icam1	1.40	2.87E-02	cops8	-1.26	4.64E-02	tcf7	-1.92	1.15E-02
ccl5	1.62	5.28E-06	elavl1	2.11	3.86E-03	col18a1	-2.04	3.82E-02
cnot6l	1.62	1.25E-02	casp7	1.69	3.86E-03	prkacb	1.66	4.21E-02
cpm	1.71	1.59E-13	notch2	2.26	1.48E-03	igf1	2.91	4.39E-03
tgfb1	1.72	5.13E-03	sox2	2.11	9.91E-04	slc3a1	2.20	2.66E-02
fancm	2.94	2.61E-02	lfng	1.74	2.26E-03	cpt1a	-1.32	3.16E-02
notch1	1.34	1.29E-19	mif	-1.81	4.21E-02	lrp5	-2.88	1.22E-02
nos2	1.49	4.48E-03	notch1	1.33	1.20E-03	rassf3	-1.67	3.20E-02
egr1	-2.05	9.95E-13	nfat5	1.55	3.48E-02	bcat2	-1.94	2.45E-02
tgif1	1.40	5.36E-04	med1	1.39	1.10E-02	necab1	2.11	1.77E-02
snap25	-3.55	7.86E-26	egr1	-1.95	7.87E-03	znrf3	-1.32	1.50E-02
st6gal2	1.46	1.39E-02	camk2d	-1.47	9.12E-03	maff	-2.54	3.95E-02
tmprss5	1.78	4.29E-11	snap25	-4.50	1.35E-04	gpr146	-1.88	4.60E-02
shcbp1	1.28	6.20E-11	klf4	2.49	1.28E-02	gpr116	-1.57	5.62E-03
rhobtb3	1.42	2.96E-16	mettl7a	1.31	3.75E-04	eid1	1.75	1.63E-02

mical2	-1.62	9.70E-24	kctd15	1.85	1.54E-02	pigb	1.41	3.53E-02
plekhg5	2.38	2.25E-02	rhobtb3	2.34	9.71E-03	tpd52	1.56	1.38E-02
lbr	1.50	1.31E-03	ppp1r11	-1.46	3.47E-03	sox4	-1.63	3.19E-02
cd93	1.43	8.11E-09	mical2	-1.52	4.29E-02	rhpn2	-2.20	2.83E-03
sgk3	1.53	5.53E-12	copa	-1.33	4.03E-02	btbd3	1.29	3.77E-02
hormad2	2.32	1.20E-03	pacs2	1.28	1.40E-03	pla1a	-4.08	3.65E-03
alg10b	1.90	4.30E-03	cybrd1	1.78	1.31E-02	fli1	-1.86	3.39E-02
tcf7	1.45	2.54E-05	brwd1	1.54	7.17E-05	synj1	1.58	4.35E-02
mid1ip1	1.37	3.61E-15	kat2b	1.68	7.78E-03	midn	-1.89	2.07E-02
kif11	-3.94	2.75E-02	Sep-07	-2.00	2.73E-03	znf768	-1.61	2.73E-02
cd274	1.88	4.86E-04	braf	-1.33	3.25E-02	metrnl	-2.40	3.26E-02
hnf4a	1.38	2.52E-02	cse1l	-2.28	4.27E-03	aen	-2.01	4.39E-03
ypel2	1.43	8.66E-03	crot	-1.30	1.05E-02	stk40	-1.80	1.09E-02
plk2	-2.95	5.39E-21	insig2	-1.29	4.06E-02	itpripl2	-1.99	2.38E-02
il6st	1.38	1.42E-02	hlf	-1.72	1.38E-02	pcdhgb6	-1.89	3.11E-02
c6	2.40	1.54E-03	smurf2	1.31	1.97E-04			
zbtb20	1.48	2.66E-20	ube2b	-1.62	1.93E-02			
cpt1a	1.46	8.40E-04	xbp1	2.01	3.44E-04			
lrp5	1.40	5.50E-14	plk2	-2.49	1.42E-02			
ablim1	1.31	1.02E-02	serinc5	1.29	4.75E-03			
mettl7b	1.57	4.35E-13	arf4	-1.58	3.46E-05			
creb1	1.32	4.52E-11	samm50	-1.65	1.01E-03			
cdc73	-1.38	6.84E-03	zbtb20	2.92	2.29E-04			
fam107b	1.38	2.65E-12	usp25	-1.40	4.23E-02			
hsd17b7	1.34	4.82E-10	slc3a1	-1.46	3.17E-02			
atrn	-1.36	9.68E-17	ablim1	1.56	9.72E-03			
rbl1	1.41	9.21E-05	atp5b	-2.15	3.13E-02			
clta	-2.97	6.19E-17	tm9sf2	-1.84	8.74E-03			

slc6a13	1.96	1.66E-13	mrpl30	-1.50	1.96E-03			
psma1	-1.39	1.08E-02	glul	2.14	1.77E-02			
tirap	1.53	1.17E-03	fam107b	1.46	3.43E-02			
ddx6	2.49	6.30E-05	camsap1	-1.41	2.77E-02			
pank1	1.52	3.67E-03	cers6	-2.30	8.53E-05			
atp6v1h	-1.56	1.79E-23	sp3	2.05	4.70E-04			
hsd17b13	1.63	1.08E-07	ndrg3	-1.57	6.69E-03			
ror1	1.30	6.14E-14	ralgapb	-1.29	1.84E-02			
ago2	1.70	1.91E-14	dhx36	-1.50	1.63E-02			
rcor1	2.01	4.13E-02	pnisr	1.74	3.90E-02			
necab1	-2.80	2.37E-23	clta	-1.44	3.60E-02			
maff	2.20	8.69E-04	akr1a1	-1.54	7.02E-03			
serpinb9	1.91	1.28E-04	psma1	-2.40	6.68E-04			
tnrc6a	1.33	8.46E-08	nek1	1.97	1.47E-03			
ddx17	1.79	7.05E-08	atp6v1h	-2.39	2.59E-02			
cfhr1	1.31	7.76E-03	dhcr24	-2.39	3.87E-04			
rbm47	1.46	9.45E-07	srrm2	2.02	4.06E-02			
ccr5	1.28	2.98E-07	necab1	-1.67	2.51E-02			
slc5a3	2.54	1.15E-09	znrf3	2.15	3.77E-04			
ndc1	1.29	7.13E-08	maff	1.47	1.72E-03			
cd84	1.26	1.52E-03	maml1	1.52	5.06E-03			
rora	1.40	9.37E-05	hnrnpc	-1.39	6.10E-03			
stxbp2	1.36	4.81E-09	dlgap4	1.50	7.99E-03			
itch	1.39	6.78E-09	rab18	-1.59	1.90E-03			
znf264	1.36	1.97E-12	slc5a3	1.56	1.79E-02			
ddhd1	-1.86	4.45E-03	eid1	-1.91	1.41E-02			
hm13	1.58	9.86E-05	rev3l	1.47	2.22E-02			
akt2	3.32	1.28E-03	gopc	-1.86	2.22E-03			

slc1a5	1.50	9.02E-14	kmt2c	1.29	2.54E-03			
fbxo3	-1.67	2.07E-10	rora	1.53	3.23E-02			
dnph1	2.00	1.25E-04	tpd52	-2.47	1.36E-03			
srsf11	1.48	1.04E-08	gsk3b	-1.32	2.44E-02			
mkln1	1.30	8.40E-19	papola	1.27	3.14E-02			
gtpbp3	1.43	1.45E-04	supt16h	-1.28	1.18E-03			
btbd3	1.32	6.22E-03	cecr2	2.34	5.76E-04			
ccnb1	-1.67	9.49E-14	chmp4b	-1.39	3.30E-02			
gltp	2.12	1.29E-18	mid1	1.65	1.16E-02			
pla1a	1.63	9.92E-04	akt2	1.35	1.02E-02			
tnfaip8	1.81	4.37E-10	card8	1.64	4.37E-04			
nacc2	1.44	3.42E-02	grpel1	-1.66	7.79E-03			
hnrnpu	1.31	1.47E-04	fbxo3	-1.44	1.51E-02			
slc26a2	2.69	3.00E-03	slc38a1	-1.39	3.76E-03			
elk4	1.44	3.37E-02	aak1	-2.10	9.46E-03			
synj1	2.16	7.42E-04	pds5a	-1.34	1.14E-02			
sqstm1	1.68	3.98E-16	mkln1	1.36	1.48E-02			
hipk1	1.80	5.20E-06	arpp19	-2.82	5.90E-05			
slc4a1ap	-1.60	9.21E-15	phf20l1	-1.28	2.74E-03			
tp53inp1	1.35	3.14E-18	clip1	1.56	1.94E-02			
znf367	1.73	3.19E-03	rhpn2	2.45	5.04E-05			
midn	1.55	5.59E-15	anxa7	-1.68	3.39E-03			
zdhhc16	1.38	5.82E-03	gltp	1.29	2.57E-02			
klhl15	1.46	2.18E-04	rfx5	-1.51	2.62E-02			
zbtb33	1.64	1.38E-03	abracl	-2.02	2.17E-03			
znf620	1.56	2.98E-06	nacc2	2.64	3.87E-03			
efna5	2.00	2.52E-02	eif4ebp2	1.44	2.45E-04			
baz1a	1.69	2.63E-06	aldoa	-1.26	3.53E-02			

ltb4r2	1.36	9.44E-04	nek7	1.48	1.78E-02			
txnip	1.78	1.79E-15	fli1	1.33	1.17E-03			
rassf5	1.72	2.93E-05	uhmk1	-2.06	3.05E-03			
			carhsp1	2.17	2.57E-03			
			azin1	-1.54	4.60E-02			
			rer1	-1.95	4.14E-06			
			synj1	-1.60	4.91E-02			
			sqstm1	2.22	8.84E-04			
			slc4a1ap	-1.47	8.42E-03			
			tp53inp1	2.36	9.87E-04			
			kbtbd6	-1.30	3.83E-02			
			usp54	1.53	4.09E-03			
			tmem41b	-1.45	3.51E-03			
			znf597	-1.44	3.56E-03			
			clic4	1.34	2.14E-02			
			jun	1.58	4.47E-02			
			ptar1	1.53	1.07E-02			
			peli1	1.30	4.59E-02			
			ythdf2	-1.45	2.15E-02			
			baz1a	1.73	4.95E-04			
			itpripl2	2.21	3.60E-03			
			sft2d2	2.07	1.26E-03			

Table S8: HuR target gene status in *L. donovani* infected (LDI) macrophages. HuR target mRNAs that were found to be differentially expressed in monocyte-derived macrophages exposed to the intracellular protozoan parasite *L. donovani* for 16 hours have been tabulated here.

	Gene Name	log Fold Change (LDI macrophage)	p-value (LDI macrophage)	No. of HuR sites (Lebedeva et al., 2011)	No. of HuR sites (Mukherjee et al., 2011)	HuR (Elavl1) knockout murine bone marrow-derived macrophages (Lu et al, 2014)
HuR target genes that show down-regulation upon HuR knockdown in other cell types and are down-regulated in L. donovani infection in macrophages	ANGEL2	-1.65	4.80E-04		9	Downregulated
	ANGPT2	-2.24	6.69E-04		1	Upregulated
	CD47	-1.58	1.63E-02		7	Upregulated
	CDC42BPA	-1.56	1.14E-03	14		Downregulated
	CELF1	-3.89	2.98E-04	15		Upregulated
	CYB5A	-2.02	7.43E-03	2		Downregulated
	DIAPH2	-1.77	9.52E-04	9		Downregulated
	ELK4	-2.2	6.88E-04		25	Upregulated
	FASTKD2	-2.21	6.85E-04	2		Downregulated
	FBXL4	-2.61	5.35E-04		10	Upregulated
	GOLGB1	-2.41	3.52E-02	2		Not found
	KIF3A	-3.11	4.14E-04		23	Upregulated
	MSX2	-4.43	2.45E-04		3	Upregulated
	NF1P9	-1.62	1.08E-03	13		Not found
	Sep-11	-3.11	1.63E-04	14		Not found
	TCEB3	-4.48	2.42E-04	7		Not found
HuR target genes that show up-regulation upon HuR knockdown in other cell types	ARID5B	1.69	4.74E-03		12	Upregulated
	ATP13A3	1.51	5.19E-05		30	Downregulated
	BCL10	2.32	1.21E-03	10		Not found

and are up-regulated in *L. donovani* infection in macrophages	CD44	1.73	9.78E-04	9	4	Downregulated
	CDH2	4.25	2.60E-04		1	Upregulated
	CLIP1	1.79	3.32E-04	6		Upregulated
	CRKL	3.04	8.47E-03		4	Not found
	DUSP3	2.1	3.51E-05	3		Downregulated
	DYRK1A	7.33	1.17E-04		61	Downregulated
	ECE1	1.75	1.55E-02	2		Upregulated
	EIF4E	2.06	9.79E-04		23	Not found
	FAM110B	2.12	2.18E-02		2	Upregulated
	FUT8	1.58	1.12E-03		16	Downregulated
	GBP1	1.76	1.18E-03	2		Not found
	GCH1	3.44	1.40E-03		6	Upregulated
	HIPK3	1.81	1.92E-02		9	Upregulated
	ITGAV	3.32	8.99E-04	20		Downregulated
	LMAN1	2.29	4.85E-03	10		Upregulated
	MED21	2.74	4.98E-04			Not found
	LPP	2.11	3.09E-02	42	46	Downregulated
	MAP7	2.07	9.05E-03	13		Upregulated
	MAST1	2.82	4.76E-04		2	Downregulated
	MDM2	3.23	4.51E-03		4	Downregulated
	ME1	3.52	1.48E-02	8		Downregulated
	MINPP1	2.13	4.29E-03	3		Upregulated
	MOCS3	1.53	1.17E-03	4		Downregulated
	MPZL1	2.74	1.03E-04	6		Downregulated
	PARD6B	2.7	1.20E-04		3	Not found
	PITPNB	2.42	1.15E-02	12	43	Downregulated
	PMAIP1	2.41	5.99E-06		15	Downregulated
	PPID	6.36	1.44E-04	3		Downregulated

	PPP3R1	3.06	5.16E-03	11		Downregulated
	PRKAR1A	1.92	4.58E-02		18	Not found
	PTPN1	1.53	1.17E-04	3		Downregulated
	RAB31	1.8	5.24E-03	10		Upregulated
	RAB5B	2.79	4.85E-04		5	Downregulated
	RALA	1.59	1.89E-04	8		Downregulated
	RARA	1.76	2.16E-03		2	Upregulated
	RBBP4	1.95	1.52E-02		13	Not found
	RCN2	5.07	1.65E-03	2		Downregulated
	SACS	3.99	4.66E-04	5		Downregulated
	SDC2	1.78	3.20E-02	10		Upregulated
	SMARCD1	2.25	5.58E-03	3		Upregulated
	SNAP23	2.14	3.89E-03		1	Not found
	MATR3	5.6	2.42E-04			Not found
	SSFA2	2.28	2.07E-02	25		Downregulated
	TBC1D5	8.24	2.37E-05	4		Downregulated
	TCF7L2	1.76	9.56E-04		20	Upregulated
	TM4SF1	1.99	1.46E-03	5		Upregulated
	TMED7	1.62	7.44E-03	23		Not found
	TMX1	1.97	1.83E-03	9		Not found
	TP53	1.76	3.09E-02		4	Downregulated
	UAP1	2.21	6.86E-04	3		Upregulated
	UBE2K	2.36	2.97E-02	24		Downregulated
	XIAP	2.27	6.58E-04		24	Downregulated
	YWHAZ	3.18	3.62E-03	10		Downregulated
HuR target genes that show up-regulation upon HuR knockdown in other cell types	ETV5	-1.65	3.21E-05		5	Downregulated
	SNCA	-4.43	2.45E-04		17	Downregulated
	MPPED2	-3.45	3.54E-04		1	Upregulated

but are down-regulated in *L. donovani* infection in macrophages	COL1A1	-2.57	5.47E-04	1		Upregulated
	PARD6B	-2.28	6.53E-04		3	Not found
	FGF5	-1.88	8.67E-04		4	Downregulated
	PRKAA1	-1.78	9.37E-04	25		Downregulated
	ECD	-1.65	1.05E-03	4		Downregulated
	SS18	-1.83	9.44E-04		58	Downregulated
	MKL1	-2	4.01E-03	4		Upregulated
	FSCN1	-2.02	8.38E-03	4		Upregulated
	TRAPPC2	-1.71	1.82E-02			Not found
	ACVR2A	-1.72	3.86E-02		10	Upregulated
HuR target genes that show down-regulation upon HuR knockdown in other cell types but are up-regulated in *L. donovani* infection in macrophages	AMD1	1.81	1.93E-03		17	Downregulated
	ANKRD17	2.08	1.83E-02	27		Downregulated
	ANP32E	1.8	9.01E-03	16		Downregulated
	ATF2	1.52	1.21E-04		41	Downregulated
	BAZ1A	2.07	3.07E-05	4		Not found
	BCL10	2.32	1.21E-03		15	Not found
	CASP2	1.74	5.39E-03		8	Upregulated
	CEP170	2.05	3.80E-05			Not found
	CKS2	5.32	2.70E-02	3		Upregulated
	CLPX	2.08	3.45E-03	2		Upregulated
	CRKL	3.04	8.47E-03	6		Not found
	DDX10	3.03	1.95E-06	5		Downregulated
	DDX21	1.81	4.92E-03	11		Downregulated
	EIF4E	2.06	9.79E-04	9		Not found
	ETS2	2.25	3.19E-03		2	Downregulated
	ETV3	1.6	1.10E-03		9	Downregulated
	EXPH5	2.67	7.36E-04		8	Downregulated
	FECH	2.03	7.32E-03	6		Not found

	FGFR1OP	2.15	1.00E-04	2		Downregulated
	FKBP5	1.66	1.04E-03	3	7	Downregulated
	FXN	2.78	4.86E-04		1	Downregulated
	GEM	1.97	2.35E-03		2	Downregulated
	GNS	1.54	2.52E-02	2		Upregulated
	H1F0	1.74	5.63E-05	4		Downregulated
	HOXD13	6	6.33E-06		1	Downregulated
	IPO5	1.62	2.23E-02	7		Downregulated
	IRF8	3.44	4.69E-03		3	Upregulated
	ISCA1	1.51	6.58E-03	2	7	Downregulated
	ITPR3	2.77	4.90E-04	1		Upregulated
	KLHL20	1.53	1.29E-05		15	Upregulated
	KYNU	1.59	4.86E-04	13		Upregulated
	LIFR	1.81	9.17E-04		3	Upregulated
	SOD2	4.49	1.48E-03	10		Not found
	RRN3	5.12	1.98E-04	5		Not found
	MAN1A1	2.03	6.67E-03		14	Upregulated
	MAN1A2	5.34	1.86E-04	9		Downregulated
	MAX	1.57	5.31E-03	6	5	Upregulated
	MBNL2	2.16	1.07E-02		1	Downregulated
	MET	2.58	3.26E-03	5		Downregulated
	MYO1B	2.16	7.64E-04	2		Downregulated
	N4BP2L2	2.57	5.48E-04		94	Downregulated
	NUDT21	1.54	1.79E-05	12		Upregulated
	NUP153	1.8	2.62E-04	5		Not found
	OXCT1	2.13	7.20E-04	10		Downregulated
	PFN2	3.69	4.27E-02	6	6	Upregulated
	PLOD2	2.5	1.58E-03	16	4	Upregulated

	POLR2D	5.63	1.66E-06		1	Upregulated
	PPFIA1	4.77	2.38E-04	2		Not found
	PPP2CA	3.18	3.99E-04	8		Downregulated
	PRKAR1A	1.92	4.58E-02	11		Not found
	PSIP1	1.5	4.25E-04	8		Downregulated
	PTPN12	1.67	2.56E-04	4		Downregulated
	RBBP4	1.95	1.52E-02	6		Not found
	RBMS3	2.04	7.71E-04		3	Downregulated
	RDX	2.28	4.21E-02	12		Downregulated
	RIF1	1.91	1.67E-03	8		Downregulated
	RSBN1	2.36	6.26E-03		9	Downregulated
	SGPL1	4.58	1.41E-02	4		Downregulated
	SIKE1	2.64	3.66E-03		7	Upregulated
	SLC7A11	3.43	3.59E-03	55		Downregulated
	SNAP23	2.14	3.89E-03	5		Not found
	SNHG4	5.6	2.42E-04	17		Not found
	SPC25	1.55	1.16E-03		7	Downregulated
	SPTBN1	3.7	3.20E-04	12		Downregulated
	STAG2	3.19	3.72E-03	9		Downregulated
	STX16	1.91	8.50E-04	5		Downregulated
	TCEB3	2.48	3.30E-03	7		Not found
	TFDP2	2.5	7.21E-06		19	Not found
	TFPI	1.9	1.06E-04		12	Downregulated
	TFRC	4.01	1.86E-02	10		Downregulated
	TLN2	1.96	1.03E-03	2		Upregulated
	TMX1	1.97	1.83E-03		5	Not found
	TROVE2	1.66	1.04E-03	27	29	Downregulated
	TSPYL1	1.73	3.99E-03	6	5	Upregulated

	UBE2D1	5.08	6.68E-03		21	Upregulated
	UBE2N	2.39	6.09E-04		20	Downregulated
	USP13	4.66	2.28E-04	2		Upregulated
	RAB6A	5.3	1.65E-02		7	Not found
	YME1L1	1.55	3.66E-02	7		Downregulated
	ZBTB1	1.88	8.64E-04		9	Downregulated

[Note: published in Goswami et al., 2020]

Table S9. Molecular docking analysis of *L. donovani* internalin-A like proteins with representative members of cadherin superfamily. Each *L. donovani* Inl-A-like protein was docked with representative members of cadherin superfamily like type I cadherin (1O6S.B, 4OY9) and desmosomal cadherin (5EQX, 5IRY) in HADDOCK. The corresponding poses determined in HADDOCK among the top three clusters are compared with a predicted representative pose between E-cadherin and *L. donovani* internalin-A like proteins and InlA-hEC1 complex in terms of docking scores, cluster size and l-RMSD.

	1O6S chain B				4OY9				5EQX				5IRY			
L. donovani Inl-A like proteins (Ldon InlA)	Cluster	Haddock Score	Clus-ter size	RMSD with InlA-hEC1 complex	Haddock Score	Clus-ter size	RMSD with LdonInlA - hEC1 complex	RMSD with InlA-hEC1 complex	Haddock Score	Clus-ter size	RMSD with LdonInlA - hEC1 complex	RMSD with InlA-hEC1 complex	Haddock Score	Clus-ter size	RMSD with LdonInlA - hEC1 complex	RMSD with InlA-hEC1 complex
E9B7L9	Cluster I	-216.7+/-12.1	20	2.32 +/- 0.149	-202.3 +/- 5.9	20	25.71 +/- 0.194	26.26 +/- 0.205	-215.6 +/- 6.1	49	26.52 +/- 0.255	26.96 +/- 0.257	-290.3 +/- 18.8	32	24.94 +/- 0.17	25.27 +/- 0.250

	Cluster II	-123.2+/-13.4	15	6.55+/-0.506	-184.2+/-1.9	36	14.73+/-0.604	14.96+/-0.466	-181.0+/-14.0	10	6.63+/-0.909	7.13+/-0.681	-189.7+/-16.4	9	15.51+/-0.395	15.97+/-0.442
	Cluster III	-164.5+/-18	5	7.82+/-0.05	-154.2+/-19.0	19	13.01+/-1.141	12.71+/-0.905	-176.4+/-18.4	29	16.57+/-0.538	16.35+/-0.500	-158.2+/-20.9	7	19.62+/-0.191	19.62+/-0.286
E9BMT7	Cluster I	-194.8+/-6.6	75	25.43+/-0.119	-209.3+/-30.6	14	25.16+/-0.139	26.09+/-0.291	-170.0+/-13.7	11	24.82+/-0.231	25.7+/-0.428	-236.1+/-16.7	16	25.35+/-0.133	24.89+/-0.238
	Cluster II	-173.8+/-14.5	55	2.98+/-1.48	-203.1+/-10.4	32	15.71+/-0.059	14.66+/-0.073	-163.4+/-13.8	25	15.49+/-0.259	15.38+/-0.342	-199.4+/-10.9	30	17.47+/-0.1	17.23+/-0.216
	Cluster III	-168+/-12	9	25.59+/-0.109	-136.3+/-6.0	77	23.92+/-1.926	20.09+/-2.436	-154.6+/-4.8	17	17.94+/-0.753	14.17+/-0.477	-141.6+/-8.2	39	25.91+/-0.289	22.34+/-0.096
E9BUL5	Cluster I	-174+/-10.6	11	9.71+/-0.411	-220.6+/-16.3	19	25.16+/-0.161	24.94+/-0.100	-221.3+/-10.9	73	27.63+/-0.635	25.88+/-0.543	-258.3+/-24.6	15	24.74+/-0.204	25.01+/-0.264
	Cluster II	-203.8+/-12.4	9	5.35+/-1.96	-215.6+/-18.4	21	14.73+/-0.57	12.53+/-0.429	-190.1+/-20.5	8	26.01+/-0.122	25.23+/-0.097	-247.4+/-10.1	33	22.29+/-0.493	21.75+/-0.248
	Cluster III	-173.6+/-15.6	7	10.63+/-0.748	-204.5+/-5.5	37	24.1+/-0.894	23.21+/-0.648	-186.1+/-13.1	19	24.62+/-0.763	23.58+/-0.822	-180.3+/-14.6	11	15.59+/-0.762	14.31+/-0.380

Table S10. **Statistics for docking analysis of *L. donovani* Inl-A-like proteins with human E-cadherin (hEC1).**

		Model	Cluster Size	Score	[1]RMSD [Å]	[2]ΔG [kcal/mol]	Interface Area [Å^2]	[3]N_{HB}	[4]N_{SB}
E9B7L9 (region : 212-573)	HADDOCK	Cluster I	20	-216.7	2.42	-7.4	1759.9	13	9
		Cluster II	15	-123.2	7.26	-12.7	1394.2	7	6
		Cluster III	5	-164.5	7.83	-10.9	1681.9	8	10
	ClusPro	Cluster I	231	-2365	25.04	-7.3	1622.9	12	19
		Cluster II	166	-2560.4	3.26	-7	1629.5	22	25
		Cluster III	79	-2305.2	8.61	-7.7	1670	13	20
	PatchDock & FireDock	Cluster I	31	-43.13	4.83	-7.8	1462.8	9	7
		Cluster II	25	-67.04	27.15	-8.2	1158.1	12	7
		Cluster III	11	-38.88	13	-2.9	1158.1	12	7
E9BMT7 (region : 394-757)	HADDOCK	Cluster I	75	-194.8	25.35	-11.6	1601.2	14	14
		Cluster II	55	-173.8	1.8	-8.8	1685.2	16	13
		Cluster III	9	-168	25.6	-11	1531	19	16
	ClusPro	Cluster I	201	-2123.2	25.37	-4.8	1483.9	13	20
		Cluster II	141	-2024.9	5.92	-5.1	1640.4	16	12
		Cluster III	100	-2019.8	6.32	-6.7	1397.7	14	9
	PatchDock & FireDock	Cluster I	30	-36.96	5.8	-4.5	1278.6	10	2
		Cluster II	16	-14.81	26.5	-4.6	1493.8	11	5
		Cluster III	16	-32.98	27.5	-9.4	900.4	4	4
E9BUL5 (region : 386-782)	HADDOCK	Cluster I	11	-174.7	10.1	-14.2	1693.8	10	5
		Cluster II	9	-203.8	4.8	-10.4	1657.6	9	11
		Cluster III	7	-173.6	10.1	-10.7	1679.5	10	8
	ClusPro	Cluster I	179	-2718.6	25.61	-7.5	1802.2	16	17

		Cluster II	131	-2651.2	26.12	-7.2	1747.3	9	8
		Cluster III	109	-2547.6	4.97	-10.3	1665.8	16	23
	PatchDock & FireDock	Cluster I	47	-38.7	8.3	-8.2	1516.6	9	0
		Cluster II	18	-33.42	26.9	-5.7	1533.3	16	12
		Cluster III	9	-29.98	12.43	-1.9	1458.8	14	3

[1]**RMSD**: It corresponds to ligand RMSD (l-RMSD) of each docked complex with Inl-A crystal structure (PDB ID: 1O6S).
[2]**ΔG**: Free energy of complex formation (kcal/mol) (PISA).
[3]**NHB**: Number of predicted hydrogen bonds (PISA).
[4]**NSB**: Number of predicted salt bridges (PISA).
[Note: published in Mukherjee et al., 2016]

Table S11: **Possible hydrogen bond forming residues in E9B7L9-hEC1, E9BMT7-hEC1 and E9BUL5-hEC1 docked complexes.** Probable hydrogen bond interactions in each representative *L. donovani* Inl-A-like protein in complex with hEC1 (as determined from docking analysis) were predicted in PISA. Residue pairs involved in the interaction, their corresponding atoms and distances between these atoms are shown.

E9B7L9-hEC1			**E9BMT7-hEC1**			**E9BUL5-hEC1**		
Chain A (E9B7L9 212-573)	Chain B (E-cadherin)	Distance	Chain A (E9BMT7 394-757)	Chain B (E-cadherin)	Distance	Chain A (E9BUL5 386-782)	Chain B (E-cadherin)	Distance
SER [HG] (130)	PHE [O] (17)	1.7	ASN [ND2] (356)	GLY [O] (0)	3	ASN [ND2] (341)	ILE [O] (4)	3
GLN [NE2] (293)	ASN [OD1] (27)	3.3	LYS [HZ3] (354)	ILE [O] (4)	1.9	ARG [HE] (150)	GLU [OE2] (54)	2.2
LYS [HZ3] (109)	GLU [OE2] (54)	1.6	SER [HG] (312)	ASN [OD1] (27)	1.8	SER [HG] (11)	GLU [OE1] (64)	2.4
LYS [HZ2] (356)	GLU [OE2] (93)	1.7	LYS [HZ2] (14)	ASP [OD2] (44)	1.6	HIS [HE2] (52)	GLU [OE1] (64)	1.7

GLU [OE1] (243)	SER [HG] (8)	2.2	THR [HG1] (176)	GLU [OE1] (54)	1.7	GLU [OE2] (367)	SER [HG] (1)	1.6
GLU [OE1] (151)	LYS [HZ1] (14)	1.6	LYS [HZ1] (151)	GLU [OE2] (54)	1.6	GLN [OE1] (173)	ASN [ND2] (20)	2.9
GLU [OE2] (151)	GLY [N] (15)	3.1	LYS [HZ2] (151)	GLU [OE2] (56)	2.2	ASN [OD1] (194)	ASN [ND2] (20)	3.6
ASP [OD2] (176)	LYS [HZ3] (19)	1.6	LYS [HZ3] (151)	GLU [OE2] (56)	2.4	GLU [OE2] (367)	ASN [ND2] (27)	3.7
GLU [OE2] (197)	LYS [HZ2] (19)	1.8	GLU [OE1] (330)	ILE [N] (4)	3.6	ASP [OD2] (125)	LYS [HZ3] (61)	1.7
GLU [OE1] (224)	GLN [NE2] (23)	2.7	ASP [OD1] (264)	LYS [HZ1] (25)	1.7	ARG [HH11] (12)	VAL [O] (48)	1.96
ASP [OD2] (291)	LYS [HZ1] (25)	1.6	ASN [OD1] (286)	LYS [HZ3] (25)	1.7	ARG [HH22] (216)	GLU [OE2] (13)	2.15
GLN [OE1] (293)	LYS [HZ3] (25)	1.7	GLU [OE2] (267)	LYS [HZ1] (30)	1.7	ARG [HH22] (150)	GLU [OE1] (54)	1.67
ASP [OD2] (316)	ASN [ND2] (27)	2.9	MET [SD] (35)	VAL [N] (48)	3.6	ARG [HH21] (12)	GLU [OE2] (64)	1.65
ARG [HH11] (336)	SER [O] (1)	1.7	GLU [OE2] (149)	LYS [HZ1] (61)	1.6	SER [OG] (248)	GLN [OE1] (23)	3.2
ARG [HH12] (336)	TRP [O] (2)	2	ASP [OD1] (172)	LYS [HZ2] (61)	2.3			
ARG [HH21] (153)	ASN [OD1] (20)	2.2	ASP [OD1] (172)	LYS [HZ3] (61)	2.2			
ARG [HH12] (225)	GLN [OE1] (23)	1.8	ASN [HD21] (37)	VAL [O] (48)	2.26			
ARG [HH11] (63)	GLY [O] (49)	1.9	ARG [HH21] (101)	GLU [OE1] (64)	1.61			

[Note: published in Mukherjee et al., 2016]

Table S12: **Possible hydrophobic interaction forming residues in E9B7L9-hEC1, E9BMT7-hEC1 and E9BUL5-hEC1 docked complexes.** Hydrophobic interactions in the representative complexes of *L. donovani* Inl-A-like protein with hEC1 were estimated using the PICI program and are tabulated here. (Numbers in parenthesis indicate the residue number from each chain).

E9B7L9-hEC1		**E9BMT7-hEC1**		**E9BUL5-hEC1**	
Chain A (E9B7L9 212-573)	Chain B (E-cadherin)	Chain A (E9BMT7 394-757)	Chain B (E-cadherin)	Chain A (E9BUL5 386-782)	Chain B (E-cadherin)
ILE (86)	PRO (18)	ALA (13)	PRO (46)	VAL (123)	PRO (18)
ILE (128)	PRO (16)	ALA (13)	VAL (48)	LEU (147)	PRO (16)
VAL (174)	PRO (16)	MET (35)	PRO (47)	LEU (147)	PRO (18)
TYR (199)	TRP (59)	MET (35)	VAL (48)	VAL (192)	PRO (16)
VAL (266)	PRO (5)	ILE (125)	PRO (18)	ALA (220)	TRP (59)
VAL (266)	PRO (6)	LEU (127)	PRO (18)	TRP (291)	PRO (6)
LEU (268)	VAL (3)	TRP (198)	TRP (59)	PHE (316)	PRO (5)
CYS (289)	PRO (5)	ALA (240)	TRP (59)	PHE (316)	PRO (6)
CYS (289)	PRO (6)	ILE (242)	TRP (59)	TYR (339)	PRO (6)
VAL (312)	PRO (5)	ILE (284)	VAL (3)	TYR (363)	LEU (-1)
VAL (312)	PRO (6)	ILE (284)	PRO (5)	TYR (363)	TRP (2)
CYS (314)	VAL (3)	ILE (284)	VAL (22)	TYR (363)	ILE (4)
CYS (357)	LEU (-1)	VAL (308)	VAL (3)		
CYS (357)	ILE (4)	VAL (308)	PRO (5)		
CYS (357)	MET (92)	VAL (353)	LEU (-1)		
PHE (359)	PRO (-2)				
PHE (359)	LEU (-1)				
LEU (360)	PRO (-2)				

[Note: published in Mukherjee et al., 2016]

Figure S2. **Docking analysis considering Inl-A and E-cadherin (hEC1)** Plot of docking scores against l-RMSD between the crystal complex (PDB ID: 1O6S) and docked poses obtained after re-docking Inl-A with hEC1 in PatchDock followed by FireDock refinement, HADDOCK and ClusPro respectively. (Legend-diamond: cluster 1; square: cluster 2; triangle: cluster 3). [Note: published in Mukherjee et al., 2016]

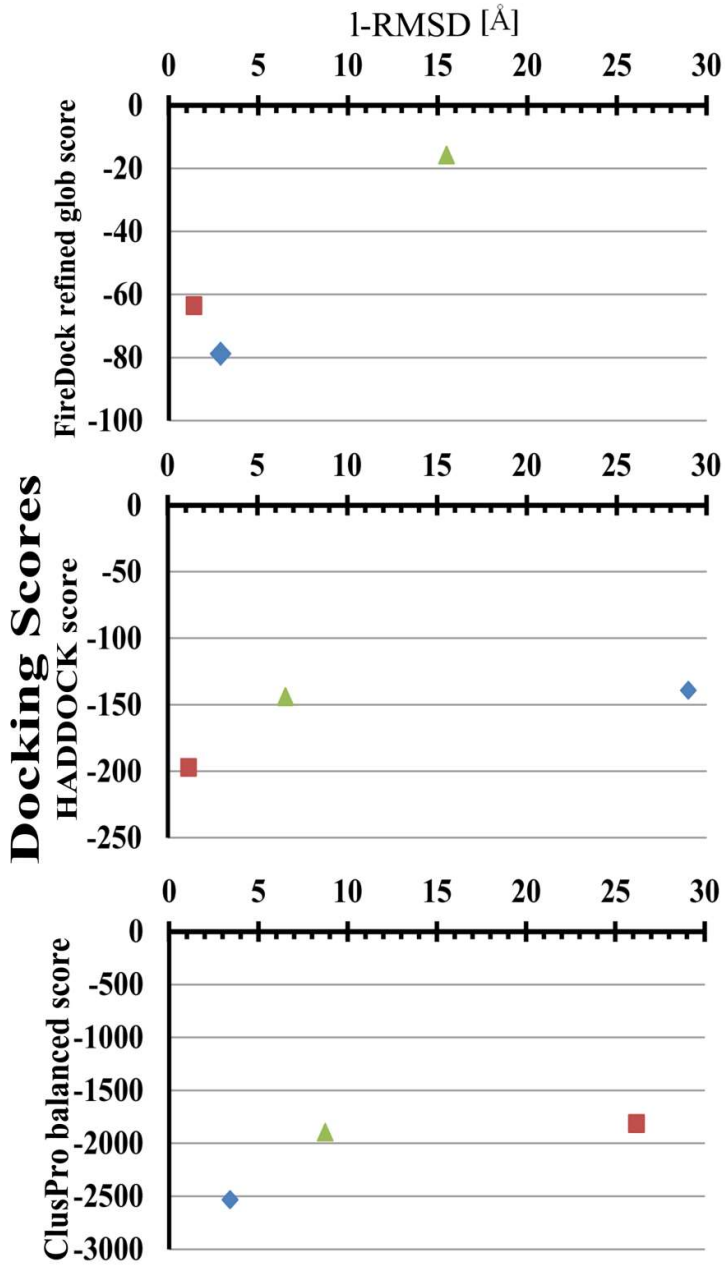


Table S13. **Dataset for determining the applicability of Co-Var in studying inter-protein co-evolution analysis**. Details about interacting protein complexes (Positive set) considered for validation of Co-Var methodology in studying protein-protein co-evolution and statistics for their analysis in Co-Var, Mirror Tree, CAPS and EV-complex is listed here.

Protein A (UniProt ID)	Generic description (Protein A)	Protein B (UniProt ID)	Generic description (Protein B)	Organism	Reference Structure (PDB ID)	Percentage of interacting pair that are co-evolved (PC)	Percentage of co-evolved pair that occur at interface (IC)
O54050	O54050_RHOCA Xanthine dehydrogenase xdhA	O54051	O54051_RHOCA Xanthine dehydrogenase xdhB	*Rhodobacter capsulatus*	1JRO	13.793	0.23
P06609	BTUC_ECOLI Vitamin B12 import system permease protein BtuC btuC	P06611	BTUD_ECOLI Vitamin B12 import ATP-binding protein BtuD btuD	*Escherichia coli (strain K12)*	1L7V	5.983	0.081
P09060	ODBA_PSEPU 2-oxoisovalerate dehydrogenase subunit alpha bkdA1	P09061	ODBB_PSEPU 2-oxoisovalerate dehydrogenase subunit beta bkdA2	*Pseudomonas putida*	1QS0	18.677	0.393
P09152	NARG_ECOLI Respiratory nitrate reductase 1 alpha chain narG	P11349	NARH_ECOLI Respiratory nitrate reductase 1 beta chain narH	*Escherichia coli (strain K12)*	1R27	10.313	0.133
P0A836	SUCC_ECOLI Succinate--CoA ligase [ADP-forming] subunit beta sucC	P0AGE9	SUCD_ECOLI Succinate--CoA ligase [ADP-forming] subunit alpha sucD	*Escherichia coli (strain K12)*	1SCU	15.234	0.344

Q8U093	TRPB1_PYRFU Tryptophan synthase beta chain 1 trpB1	Q8U094	TRPA_PYRFU Tryptophan synthase alpha chain trpA	*Pyrococcus furiosus (strain ATCC 43587 / DSM 3638 / JCM 8422 / Vc1)*	1WDW	31.064	0.769
Q9V0T9	GATD_PYRAB Glutamyl-tRNA(Gln) amidotransferase subunit D gatD	Q9V0U0	GATE_PYRAB Glutamyl-tRNA(Gln) amidotransferase subunit E gatE	*Pyrococcus abyssi (strain GE5 / Orsay)*	1ZQ1	5.738	0.094
P81303	NOP10_METJA Ribosome biogenesis protein Nop10 nop10	Q57612	TRUB_METJA Probable tRNA pseudouridine synthase B truB	*Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440)*	2APO	20.492	1.733
P0AC55	GLNK_ECOLI Nitrogen regulatory protein P-II 2 glnK	P69681	AMTB_ECOLI Ammonia channel amtB	*Escherichia coli (strain K12)*	2NS1	4.688	0.074
P08839	PT1_ECOLI Phosphoenolpyruvate-protein phosphotransferase ptsI	P0AA04	PTHP_ECOLI Phosphocarrier protein HPr ptsH	*Escherichia coli (strain K12)*	2XDF	25.641	1.18
P0A6T9	GCSH_ECOLI Glycine cleavage system H protein gcvH	P27248	GCST_ECOLI Aminomethyltransferase gcvT	*Escherichia coli (strain K12)*	3A8I	23.611	0.481
P30750	METN_ECOLI Methionine import ATP-binding protein MetN metN	P31547	METI_ECOLI D-methionine transport system permease protein MetI metI	*Escherichia coli (strain K12)*	3DHW	29.381	0.683

P0A6B9	ISCS_ECO57 Cysteine desulfurase IscS iscS	P0ACD6	ISCU_ECO57 Iron-sulfur cluster assembly scaffold protein IscU iscU	*Escherichia coli O157:H7*	3LVL	7.447	0.238
P07395	SYFB_ECOLI Phenylalanine--tRNA ligase beta subunit pheT	P08312	SYFA_ECOLI Phenylalanine--tRNA ligase alpha subunit pheS	*Escherichia coli (strain K12)*	3PCO	16.355	0.525
P76014	DHAL_ECOLI PEP-dependent dihydroxyacetone kinase, ADP-binding subunit DhaL dhaL	P76015	DHAK_ECOLI PEP-dependent dihydroxyacetone kinase, dihydroxyacetone-binding subunit DhaK dhaK	*Escherichia coli (strain K12)*	3PNL	35.678	0.951
O06662	VAPC_SHIFL tRNA(fMet)-specific endonuclease VapC vapC	O06663	VAPB_SHIFL Antitoxin VapB vapB	*Shigella flexneri*	3TND	6.787	1.969
P09143	SUCD_THETH Succinate--CoA ligase [GDP-forming] subunit alpha sucD	P25126	SUCC_THETH Succinate--CoA ligase [GDP-forming] subunit beta sucC	*Thermus thermophilus*	3UFX	10.87	0.26
P00929	TRPA_SALTY Tryptophan synthase alpha chain trpA	P0A2K1	TRPB_SALTY Tryptophan synthase beta chain trpB	*Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)*	1A50	9.406	0.254
P08559	ODPA_HUMAN Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial PDHA1	P11177	ODPB_HUMAN Pyruvate dehydrogenase E1 component subunit beta, mitochondrial PDHB	*Homo sapiens*	1NI4	11.233	0.938

Q9NSD9	SYFB_HUMAN Phenylalanine--tRNA ligase beta subunit FARSB	Q9Y285	SYFA_HUMAN Phenylalanine--tRNA ligase alpha subunit FARSA	*Homo sapiens*	3L4G	6.054	0.713
Q08602	PGTA_RAT Geranylgeranyl transferase type-2 subunit alpha Rabggta	Q08603	PGTB2_RAT Geranylgeranyl transferase type-2 subunit beta Rabggtb	*Rattus norvegicus*	1DCE	3.53	0.433
O19069	SUCA_PIG Succinate--CoA ligase [ADP/GDP-forming] subunit alpha, mitochondrial SUCLG1	P53590	SUCB2_PIG Succinate--CoA ligase [GDP-forming] subunit beta, mitochondrial (Fragment) SUCLG2	*Sus scrofa*	1EUD	5.396	0.274
O85040	RBL1_HALNC Ribulose bisphosphate carboxylase large chain cbbL	P45686	RBS_HALNC Ribulose bisphosphate carboxylase small chain cbbS	*Halothiobacillus neapolitanus (strain ATCC 23641 / c2)*	1SVD	15.979	0.796
Q7SIC0	HIS5_THET8 Imidazole glycerol phosphate synthase subunit HisH hisH	Q7SIB9	HIS6_THET8 Imidazole glycerol phosphate synthase subunit HisF hisF	*Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)*	1KA9	31.892	1.503
P13448	NHAA_RHOER Nitrile hydratase subunit alpha nthA	P13449	NHAB_RHOER Nitrile hydratase subunit beta nthB	*Rhodococcus erythropolis*	3A8O	13.78	2.024
Q9LA16	Q9LA16_STANO Sulfite:cytochrome c oxidoreductase subunit A sorA	Q9LA15	Q9LA15_STANO Sulfite:cytochrome c oxidoreductase subunit B sorB	*Starkeya novella*	2BLF	12.707	0.588

P69905	HBA_HUMAN Hemoglobin subunit alpha HBA1	P68871	HBB_HUMAN Hemoglobin subunit beta HBB	*Homo sapiens*	4N7N	11.842	1.603
Q939U1	SOXA_RHOSU L-cysteine S-thiosulfotransferase subunit SoxA soxA	Q939U4	Q939U4_RHOSU Cytochrome c soxX	*Rhodovulum sulfidophilum*	1H32	15.244	1.157
P54322	PYRDB_LACLM Dihydroorotate dehydrogenase B (NAD(+)), catalytic subunit pyrDB	P56968	PYRK_LACLM Dihydroorotate dehydrogenase B (NAD(+)), electron transfer subunit pyrK	*Lactococcus lactis subsp. cremoris (strain MG1363)*	1EP1	21.557	0.461
P13804	ETFA_HUMAN Electron transfer flavoprotein subunit alpha, mitochondrial ETFA	P38117	ETFB_HUMAN Electron transfer flavoprotein subunit beta ETFB	*Homo sapiens*	1EFV	16.883	0.941
P12694	ODBA_HUMAN 2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial BCKDHA	P21953	ODBB_HUMAN 2-oxoisovalerate dehydrogenase subunit beta, mitochondrial BCKDHB	*Homo sapiens*	1DTW	9.524	0.308
P07998	RNAS1_HUMAN Ribonuclease pancreatic RNASE1	P13489	RINI_HUMAN Ribonuclease inhibitor RNH1	*Homo sapiens*	1Z7X	4.808	0.279
P01241	SOMA_HUMAN Somatotropin GH1	P16471	PRLR_HUMAN Prolactin receptor PRLR	*Homo sapiens*	1BP3	6.41	0.169
P15153	RAC2_HUMAN Ras-related C3 botulinum toxin substrate 2 RAC2	P52566	GDIR2_HUMAN Rho GDP-dissociation inhibitor 2 ARHGDIB	*Homo sapiens*	1DS6	31.21	0.839

P39958	GDI1_YEAST Rab GDP-dissociation inhibitor GDI1	P01123	YPT1_YEAST GTP-binding protein YPT1 YPT1	*Saccharomyces cerevisiae (strain ATCC 204508 / S288c)*	1UKV	1.493	0.082
P02753	RET4_HUMAN Retinol-binding protein 4 RBP4	P02766	TTHY_HUMAN Transthyretin TTR	*Homo sapiens*	1RLB	2.439	0.068
P32851	STX1A_RAT Syntaxin-1A Stx1a	P61765	STXB1_RAT Syntaxin-binding protein 1 Stxbp1	*Rattus norvegicus*	3C98	4.659	0.341
P13489	RINI_HUMAN Ribonuclease inhibitor RNH1	P03950	ANGI_HUMAN Angiogenin ANG	*Homo sapiens*	1A4Y	6.771	0.322
P18510	IL1RA_HUMAN Interleukin-1 receptor antagonist protein IL1RN	P14778	IL1R1_HUMAN Interleukin-1 receptor type 1 IL1R1	*Homo sapiens*	1IRA	6.787	0.32
Q5SLR3	ODBB_THET8 2-oxoisovalerate dehydrogenase subunit beta TTHA0230	Q5SLR4	ODBA_THET8 2-oxoisovalerate dehydrogenase subunit alpha TTHA0229	*Thermus thermophilus (strain HB8 / ATCC 27634 / DSM 579)*	1UM9	11.899	0.427
P77499	SUFC_ECOLI Probable ATP-dependent transporter SufC sufC	P77689	SUFD_ECOLI FeS cluster assembly protein SufD sufD	*Escherichia coli (strain K12)*	2ZU0	19.084	0.315
O29689	ISCS2_ARCFU Cysteine desulfurase IscS 2 iscS2	P0DMG2	ISCU2_ARCFU Iron-sulfur cluster assembly scaffold protein IscU 2 iscU2	*Archaeoglobus fulgidus (strain ATCC 49558 / VC-16 / DSM 4304 / JCM 9628 / NBRC 100126)*	4EB5	8.271	0.27

P00968	CARB_ECOLI Carbamoyl-phosphate synthase large chain carB	P0A6F1	CARA_ECOLI Carbamoyl-phosphate synthase small chain carA	*Escherichia coli (strain K12)*	1A9X	3.161	0.034
P30748	MOAD_ECOLI Molybdopterin synthase sulfur carrier subunit moaD	P30749	MOAE_ECOLI Molybdopterin synthase catalytic subunit moaE	*Escherichia coli (strain K12)*	1FM0	33.333	4.136
P43895	EFTS_THET8 Elongation factor Ts tsf	P60338	EFTU1_THETH Elongation factor Tu-A tufA	*Thermus thermophilus*	1AIP	5.628	0.237
P0A6P1	EFTS_ECOLI Elongation factor Ts tsf	P0CE48	EFTU2_ECOLI Elongation factor Tu 2 tufB	*Escherichia coli (strain K12)*	1EFU	17.254	0.61
P27001	SYFA_THETH Phenylalanine--tRNA ligase alpha subunit pheS	P27002	SYFB_THETH Phenylalanine--tRNA ligase beta subunit pheT	*Thermus thermophilus*	1B70	24.044	0.483
P0A988	DPO3B_ECOLI Beta sliding clamp dnaN	P28630	HOLA_ECOLI DNA polymerase III subunit delta holA	*Escherichia coli (strain K12)*	1JQJ	6.838	0.052
Q87KD2	Q87KD2_VIBPA Potassium uptake protein TrkA VP3045	Q87TN7	TRKH_VIBPA Trk system potassium uptake protein TrkH trkH	*Vibrio parahaemolyticus serotype O3:K6 (strain RIMD 2210633)*	4J9U	5.882	0.018
P00897	TRPE_SERMA Anthranilate synthase component 1 trpE	P00900	TRPG_SERMA Anthranilate synthase component 2 trpG	*Serratia marcescens*	1I7Q	11.667	0.272

Table S14. **Dataset for determining the applicability of Co-Var methodology in studying inter-protein co-evolution analysis**. Details about non interacting proteins (Negative set) considered for validation of Co-Var methodology in studying protein-protein co-evolution and statistics for their analysis in Co-Var, Mirror Tree, CAPS and EV-complex is listed here.

Protein A (UniProt ID)	**Generic description (Protein A)**	**Protein B (UniProt ID)**	**Generic description (Protein B)**	**Organism**
O89100	GRAP2_MOUSE GRB2-related adaptor protein 2 Grap2	Q9Z1S8	GAB2_MOUSE GRB2-associated-binding protein 2 Gab2	*Mus musculus*
O60716	CTND1_HUMAN Catenin delta-1 CTNND1	P35222	CTNB1_HUMAN Catenin beta-1 CTNNB1	*Homo sapiens*
O60716	CTND1_HUMAN Catenin delta-1 CTNND1	P14923	PLAK_HUMAN Junction plakoglobin JUP	*Homo sapiens*
Q99626	CDX2_HUMAN Homeobox protein CDX-2 CDX2	P20226	TBP_HUMAN TATA-box-binding protein TBP	*Homo sapiens*
Q08877	DYN3_RAT Dynamin-3 Dnm3	Q62600	NOS3_RAT Nitric oxide synthase, endothelial Nos3	*Rattus norvegicus*
P39052	DYN2_RAT Dynamin-2 Dnm2	Q62600	NOS3_RAT Nitric oxide synthase, endothelial Nos3	*Rattus norvegicus*
O43395	PRPF3_HUMAN U4/U6 small nuclear ribonucleoprotein Prp3 PRPF3	O43447	PPIH_HUMAN Peptidyl-prolyl cis-trans isomerase H PPIH	*Homo sapiens*
Q16891	MIC60_HUMAN MICOS complex subunit MIC60 IMMT	Q9UKY1	ZHX1_HUMAN Zinc fingers and homeoboxes protein 1 ZHX1	*Homo sapiens*

Q9QXX8	NUFP1_MOUSE Nuclear fragile X mental retardation-interacting protein 1 Nufip1	Q9WVR4	FXR2_MOUSE Fragile X mental retardation syndrome-related protein 2 Fxr2	*Mus musculus*
Q9UM73	ALK_HUMAN ALK tyrosine kinase receptor ALK	P21246	PTN_HUMAN Pleiotrophin PTN	*Homo sapiens*
Q6EMB2	TTLL5_HUMAN Tubulin polyglutamylase TTLL5 TTLL5	P03372	ESR1_HUMAN Estrogen receptor ESR1	*Homo sapiens*
O15151-4	MDM4_HUMAN Isoform HDMX211 of Protein Mdm4 MDM4	P04637	P53_HUMAN Cellular tumor antigen p53 TP53	*Homo sapiens*
Q8CBD1	NRIP1_MOUSE Nuclear receptor-interacting protein 1 Nrip1	Q9JIF0	ANM1_MOUSE Protein arginine N-methyltransferase 1 Prmt1	*Mus musculus*
Q6EMB2	TTLL5_HUMAN Tubulin polyglutamylase TTLL5 TTLL5	Q92731	ESR2_HUMAN Estrogen receptor beta ESR2	*Homo sapiens*
Q96RE7	NACC1_HUMAN Nucleus accumbens-associated protein 1 NACC1	Q05516	ZBT16_HUMAN Zinc finger and BTB domain-containing protein 16 ZBTB16	*Homo sapiens*
Q12873	CHD3_HUMAN Chromodomain-helicase-DNA-binding protein 3 CHD3	Q9H213	MAGH1_HUMAN Melanoma-associated antigen H1 MAGEH1	*Homo sapiens*
P63167	DYL1_HUMAN Dynein light chain 1, cytoplasmic DYNLL1	Q7LDG7	GRP2_HUMAN RAS guanyl-releasing protein 2 RASGRP2	*Homo sapiens*

P15207	ANDR_RAT Androgen receptor Ar	Q96RL1	UIMC1_HUMAN BRCA1-A complex subunit RAP80 UIMC1	*Homo sapiens*
P06876	MYB_MOUSE Transcriptional activator Myb Myb	P02340	P53_MOUSE Cellular tumor antigen p53 Tp53	*Mus musculus*
P10242	MYB_HUMAN Transcriptional activator Myb MYB	P04637	P53_HUMAN Cellular tumor antigen p53 TP53	*Homo sapiens*
O15360	FANCA_HUMAN Fanconi anemia group A protein FANCA	Q00597	FANCC_HUMAN Fanconi anemia group C protein FANCC	*Homo sapiens*
Q15404	RSU1_HUMAN Ras suppressor protein 1 RSU1	Q91XD2	LIMS2_MOUSE LIM and senescent cell antigen-like-containing domain protein 2 Lims2	*Mus musculus*
P38532	HSF1_MOUSE Heat shock factor protein 1 Hsf1	Q99K43	PRC1_MOUSE Protein regulator of cytokinesis 1 Prc1	*Mus musculus*
Q16513	PKN2_HUMAN Serine/threonine-protein kinase N2 PKN2	Q99759	M3K3_HUMAN Mitogen-activated protein kinase kinase kinase 3 MAP3K3	*Homo sapiens*
P15531	NDKA_HUMAN Nucleoside diphosphate kinase A NME1	Q00987	MDM2_HUMAN E3 ubiquitin-protein ligase Mdm2 MDM2	*Homo sapiens*
P06493	CDK1_HUMAN Cyclin-dependent kinase 1 CDK1	P49207	RL34_HUMAN 60S ribosomal protein L34 RPL34	*Homo sapiens*

P68400	CSK21_HUMAN Casein kinase II subunit alpha CSNK2A1	Q9UBF6	RBX2_HUMAN RING-box protein 2 RNF7	*Homo sapiens*
O75192	PX11A_HUMAN Peroxisomal membrane protein 11A PEX11A	P28328	PEX2_HUMAN Peroxisome biogenesis factor 2 PEX2	*Homo sapiens*
O43543	XRCC2_HUMAN DNA repair protein XRCC2 XRCC2	Q9BXW9	FACD2_HUMAN Fanconi anemia group D2 protein FANCD2	*Homo sapiens*
P27699	CREM_MOUSE cAMP-responsive element modulator Crem	Q8R4I4	TF2AY_MOUSE TFIIA-alpha and beta-like factor Gtf2a1l	*Mus musculus*
P59190	RAB15_HUMAN Ras-related protein Rab-15 RAB15	Q15276	RABE1_HUMAN Rab GTPase-binding effector protein 1 RABEP1	*Homo sapiens*
P35606	COPB2_HUMAN Coatomer subunit beta' COPB2	Q08499-6	PDE4D_HUMAN Isoform 5 of cAMP-specific 3',5'-cyclic phosphodiesterase 4D PDE4D	*Homo sapiens*
P49597	P2C56_ARATH Protein phosphatase 2C 56 ABI1	Q38898	AKT2_ARATH Potassium channel AKT2/3 AKT2	*Arabidopsis thaliana*
P08069	IGF1R_HUMAN Insulin-like growth factor 1 receptor IGF1R	Q92990	GLMN_HUMAN Glomulin GLMN	*Homo sapiens*
P54748-1	PDE4A_RAT cAMP-specific 3',5'-cyclic phosphodiesterase 4A Pde4a	Q5FWY5	AIP_RAT AH receptor-interacting protein Aip	*Rattus norvegicus*

O88346	TNNT1_MOUSE Troponin T, slow skeletal muscle Tnnt1	Q61410	KGP2_MOUSE cGMP-dependent protein kinase 2 Prkg2	*Mus musculus*
P97471	SMAD4_MOUSE Mothers against decapentaplegic homolog 4 Smad4	Q62424	HXA13_MOUSE Homeobox protein Hox-A13 Hoxa13	*Mus musculus*
O60674	JAK2_HUMAN Tyrosine-protein kinase JAK2 JAK2	Q9JM90	STAP1_MOUSE Signal-transducing adaptor protein 1 Stap1	*Mus musculus*
A2AM29	AF9_MOUSE Protein AF-9 Mllt3	P30658	CBX2_MOUSE Chromobox protein homolog 2 Cbx2	*Mus musculus*
O55055	TRDMT_MOUSE tRNA (cytosine(38)-C(5))-methyltransferase Trdmt1	Q9UJW3	DNM3L_HUMAN DNA (cytosine-5)-methyltransferase 3-like DNMT3L	*Homo sapiens*
O15169	AXIN1_HUMAN Axin-1 AXIN1	P30153	2AAA_HUMAN Serine/threonine-protein phosphatase 2A 65 kDa regulatory subunit A alpha isoform PPP2R1A	*Homo sapiens*
O00206	TLR4_HUMAN Toll-like receptor 4 TLR4	Q8IUC6	TCAM1_HUMAN TIR domain-containing adapter molecule 1 TICAM1	*Homo sapiens*
P37840	SYUA_HUMAN Alpha-synuclein SNCA	P43004	EAA2_HUMAN Excitatory amino acid transporter 2 SLC1A2	*Homo sapiens*
O88777	PSN2_RAT Presenilin-2 Psen2	Q99569	PKP4_HUMAN Plakophilin-4 PKP4	*Homo sapiens*

P28651	CAH8_MOUSE Carbonic anhydrase-related protein Ca8	Q9ERP3	TRI54_MOUSE Tripartite motif-containing protein 54 Trim54	*Mus musculus*
P14619	KGP1_HUMAN cGMP-dependent protein kinase 1 PRKG1	Q5VSY0	GKAP1_HUMAN G kinase-anchoring protein 1 GKAP1	*Homo sapiens*
O75190	DNJB6_HUMAN DnaJ homolog subfamily B member 6 DNAJB6	P02533	K1C14_HUMAN Keratin, type I cytoskeletal 14 KRT14	*Homo sapiens*
Q9NV70	EXOC1_HUMAN Exocyst complex component 1	P61586	RHOA_HUMAN Transforming protein RhoA	*Homo sapiens*
P21246	PTN_HUMAN Pleiotrophin	P26641	EF1G_HUMAN Elongation factor 1-gamma	*Homo sapiens*
P17706	PTN2_HUMAN Tyrosine-protein phosphatase non-receptor type 2	P31946	1433B_HUMAN 14-3-3 protein beta/alpha	*Homo sapiens*

Figure S3: High degree co-evolutionary pairings in FGF1-FGFR1 complex (A) Residues predicted as co-evolved in FGF1-FGFR1 complex tend to have a large number of co-evolutionary connections or pairings among them (High degree co-evolved positions) as shown here. (B) High degree co-evolved positions that occur in spatially proximal and distal regions have been represented on the reference structure (PDB ID: 1EVT). In the structural representation of co-evolved positions, FGF1 chain: light green and FGFR1 chain: light blue.

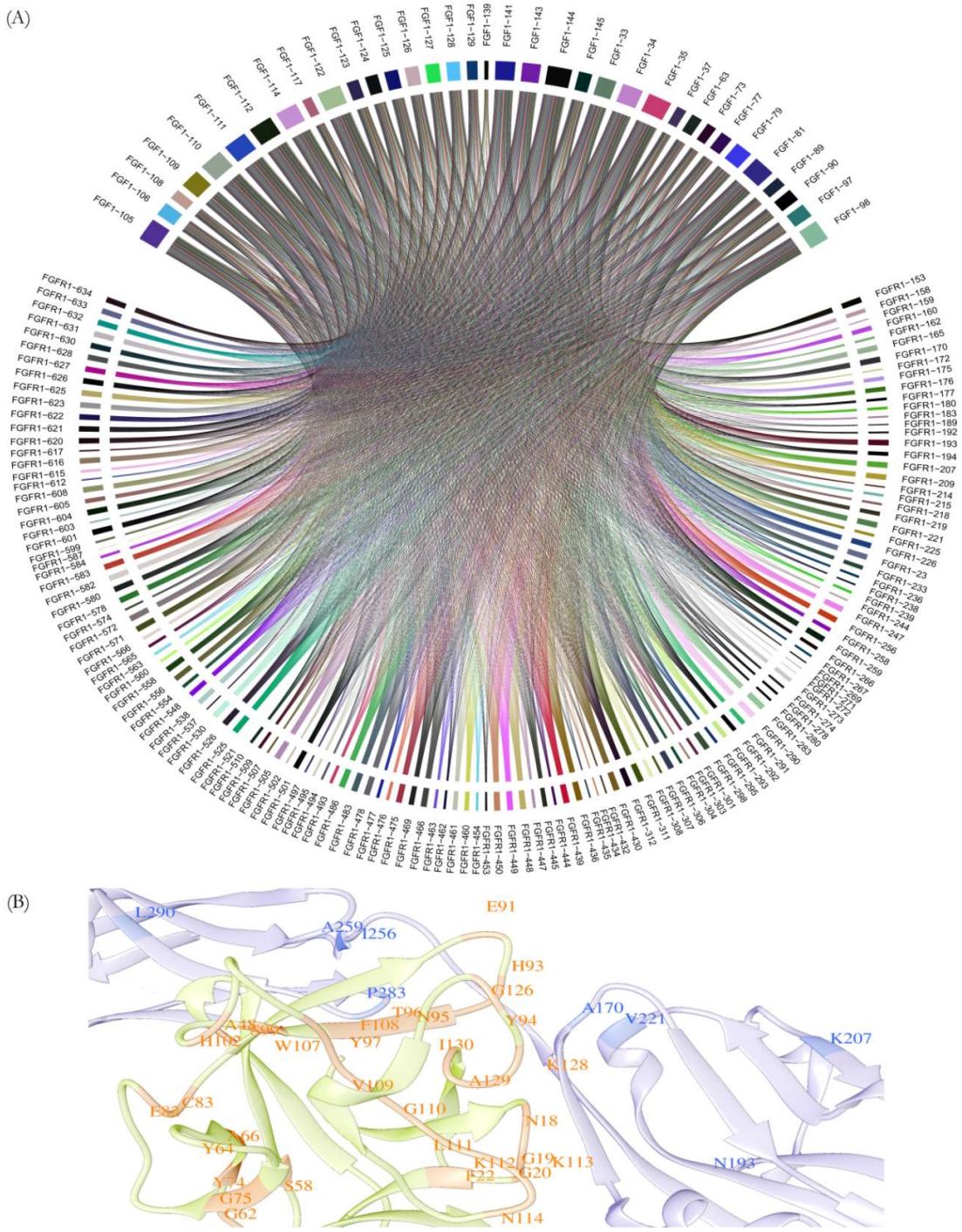


Figure S4: Inter-cellular protein interaction complex involving FGF10 and FGFR2 have a large number high degree co-evolved positions. (A) Co-evolving residues in FGF10 and FGFR2 form a large number of co-evolutionary connections (High degree co-evolved positions) or pairings among them and are shown here. (B) High degree co-evolved positions have been represented on the reference structure (PDB ID: 1NUN). In the structural representation of co-evolved positions, FGF10 chain: light green and FGFR2 chain: light blue.

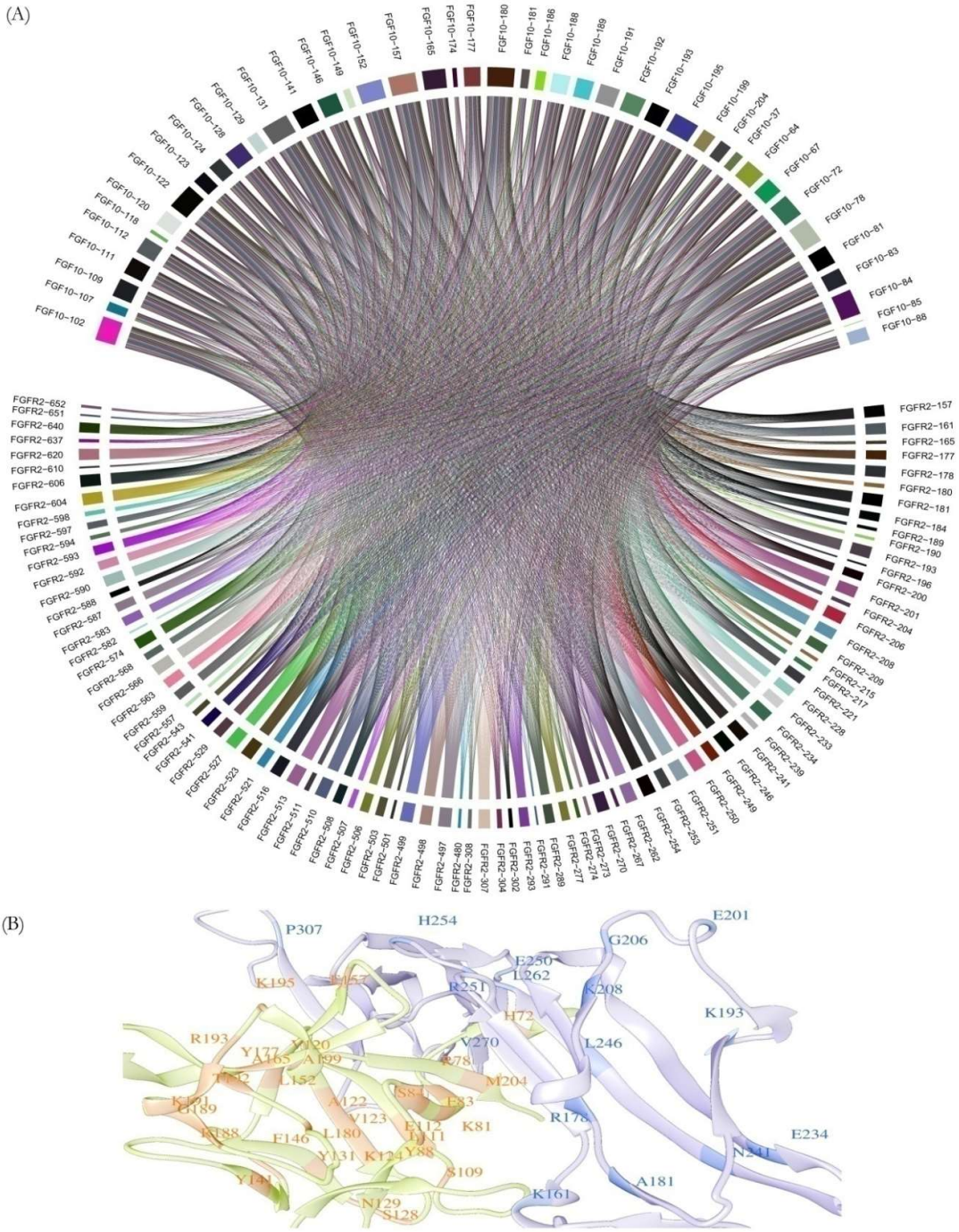


Figure S5: High degree co-evolved positions in FGF1-FGFR2 complex. (A) High degree co-evolved residue positions (Co-evolving residue positions that tend to have a large number of co-evolutionary connections or pairings among them) in FGF1 and FGFR2 are shown here. (B) High degree co-evolved positions mapped onto the reference structure (PDB ID: 1DJS) have been represented. In the structural representation of co-evolved positions, FGF1 chain: light green while FGFR2 chain: light blue.

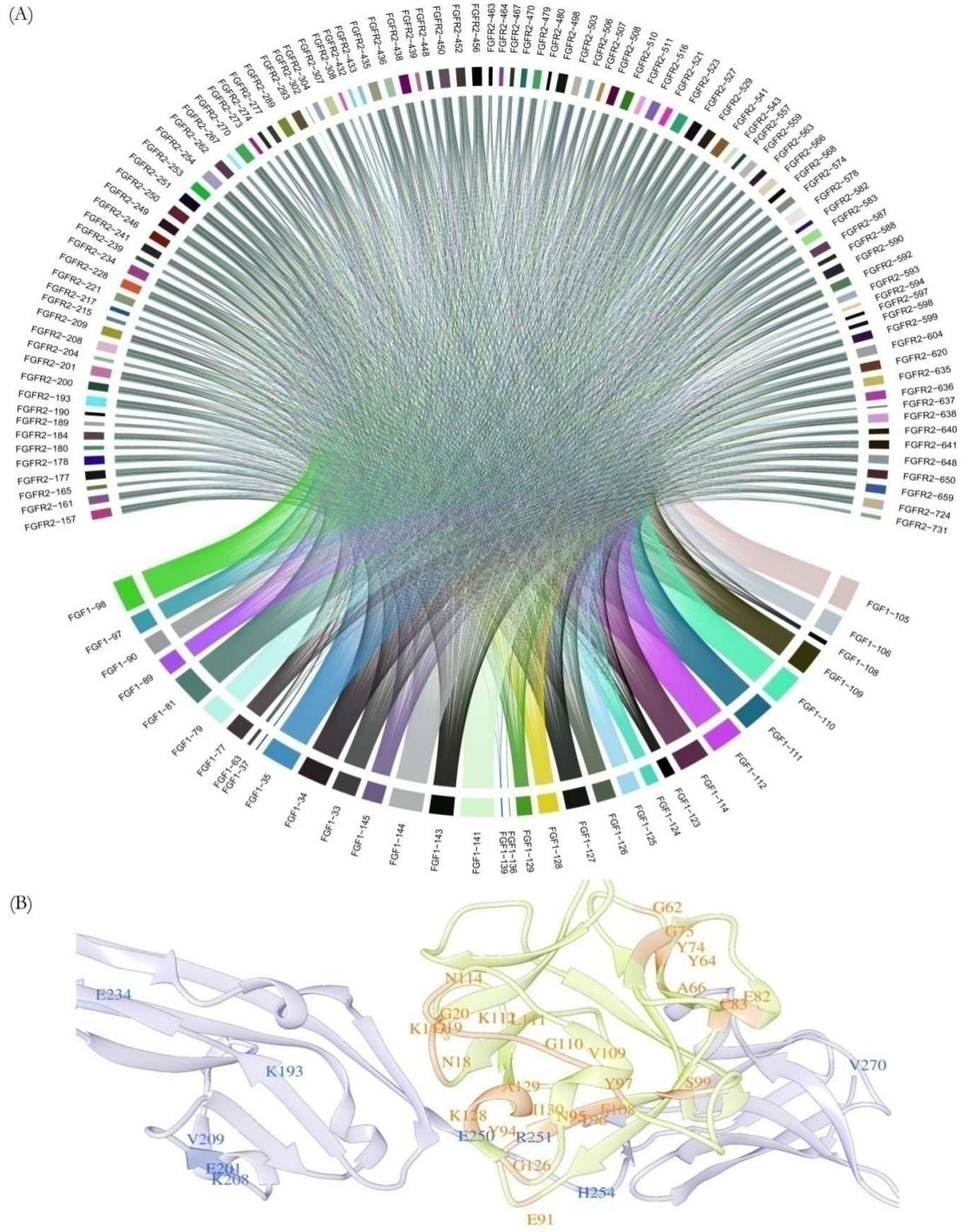


Table S15 **: List of probable CB regulator-target relationships identified considering hsa-miR-146a-5p as a regulator**

Gene with tandem miRNA pairs (Gene A)	CB-pair Regulator	Binding sites of CB-pair regulator in Gene A	Secondary Effector miRNA	Target gene of secondary effector miRNA Gene B (Secondary effector gene)	Binding sites of Secondary effector miRNA in Gene B (Secondary effector gene)	Prevalence of miRNA CB-pair (Count)	F.C. in Gene A	P-value (Gene A)	F.C. in Gene B	P-value (Gene B)
Prkdc	hsa-miR-146a-5p	2	hsa-miR-20a-5p	fam117b	3	12	-1.60	1.50E-08	-1.54	4.85E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-96-5p	top2a	3	6	-1.60	1.50E-08	-3.05	3.27E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-338-3p	gtse1	2	6	-1.60	1.50E-08	-1.80	4.57E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-96-5p	enc1	2	6	-1.60	1.50E-08	-2.45	4.51E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	ccne2	2	6	-1.60	1.50E-08	-2.41	8.73E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-20a-5p	ankh	2	12	-1.60	1.50E-08	-2.60	2.40E-11
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	scamp5	2	11	-1.60	1.50E-08	-1.56	1.13E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	slc40a1	3	11	-1.60	1.50E-08	-2.62	1.31E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-20a-3p	ccnb2	2	4	-1.60	1.50E-08	-1.82	1.67E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	e2f1	3	11	-1.60	1.50E-08	-2.45	2.95E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	cdc45	3	6	-1.60	1.50E-08	-1.88	9.46E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-22-3p	pold2	3	10	-1.60	1.50E-08	-1.62	2.99E-12
Prkdc	hsa-miR-146a-5p	2	hsa-miR-24-3p	nfatc2	2	8	-1.60	1.50E-08	-1.55	3.35E-03
Prkdc	hsa-miR-146a-5p	2	hsa-miR-30e-5p	kiaa0101	2	12	-1.60	1.50E-08	-1.63	8.71E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-27b-3p	hs2st1	2	12	-1.60	1.50E-08	-1.52	1.36E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-221-5p	cdc20	2	6	-1.60	1.50E-08	-2.37	3.92E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-16-5p	htra1	2	14	-1.60	1.50E-08	-1.88	4.08E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-423-5p	abcc5	2	3	-1.60	1.50E-08	-1.63	1.15E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-26a-5p	fam83d	2	10	-1.60	1.50E-08	-1.52	1.23E-03

Prkdc	hsa-miR-146a-5p	2	hsa-miR-423-5p	nrp1	3	3	-1.60	1.50E-08	-1.77	3.94E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-20a-5p	bnc2	2	12	-1.60	1.50E-08	-2.10	2.41E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-942-5p	anln	2	6	-1.60	1.50E-08	-2.26	1.43E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-98-5p	pfkfb2	2	8	-1.60	1.50E-08	-1.77	4.90E-10
Prkdc	hsa-miR-146a-5p	2	hsa-miR-22-5p	prkca	2	8	-1.60	1.50E-08	-2.32	6.41E-12
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	dnm3	2	11	-1.60	1.50E-08	-2.02	1.70E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-9-5p	pou2f2	2	7	-1.60	1.50E-08	-1.82	1.15E-11
Prkdc	hsa-miR-146a-5p	2	hsa-miR-9-5p	nhsl2	2	7	-1.60	1.50E-08	-1.71	2.55E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-22-5p	ptprm	2	8	-1.60	1.50E-08	-1.92	1.53E-08
Prkdc	hsa-miR-146a-5p	2	hsa-miR-27b-3p	spry2	3	12	-1.60	1.50E-08	-2.56	1.54E-12
Prkdc	hsa-miR-146a-5p	2	hsa-miR-24-3p	aurkb	2	8	-1.60	1.50E-08	-1.88	2.49E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	tnfrsf21	3	6	-1.60	1.50E-08	-1.90	4.68E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	mbnl3	2	6	-1.60	1.50E-08	-1.78	9.48E-10
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	ctdspl	2	6	-1.60	1.50E-08	-1.63	9.61E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	mcm4	2	6	-1.60	1.50E-08	-1.71	2.49E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	cenpf	2	11	-1.60	1.50E-08	-2.25	6.22E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-628-3p	myo1d	2	1	-1.60	1.50E-08	-2.06	1.44E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-942-5p	mki67	2	6	-1.60	1.50E-08	-3.38	5.80E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	cdca7l	2	6	-1.60	1.50E-08	-1.98	8.12E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-27a-3p	gpd1l	2	14	-1.60	1.50E-08	-1.74	6.27E-09
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	ptgfrn	2	11	-1.60	1.50E-08	-2.17	7.71E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-106b-5p	polq	2	12	-1.60	1.50E-08	-1.95	7.02E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-98-5p	tgfbr2	2	8	-1.60	1.50E-08	-1.72	5.56E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-22-5p	trip13	2	8	-1.60	1.50E-08	-1.84	4.90E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-26a-5p	cdc6	2	10	-1.60	1.50E-08	-2.50	1.90E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	wee1	2	11	-1.60	1.50E-08	-1.90	1.09E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	nfia	2	11	-1.60	1.50E-08	-1.50	9.73E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-27b-3p	fbn2	2	12	-1.60	1.50E-08	-1.77	1.41E-03
Prkdc	hsa-miR-146a-5p	2	hsa-miR-30e-5p	rrm2	2	12	-1.60	1.50E-08	-3.03	2.17E-07

Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	kank2	2	6	-1.60	1.50E-08	-1.73	1.82E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	abhd2	2	6	-1.60	1.50E-08	-1.80	1.06E-12
Prkdc	hsa-miR-146a-5p	2	hsa-miR-330-3p	ca2	2	6	-1.60	1.50E-08	-2.11	3.84E-06
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	prc1	2	11	-1.60	1.50E-08	-1.56	4.35E-03
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	tfcp2l1	3	6	-1.60	1.50E-08	-1.57	5.59E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-23b-5p	cd93	3	2	-1.60	1.50E-08	-2.47	1.50E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-93-5p	hist1h3b	2	11	-1.60	1.50E-08	-1.50	5.45E-03
Prkdc	hsa-miR-146a-5p	2	hsa-miR-98-5p	gatm	2	8	-1.60	1.50E-08	-2.40	4.27E-09
Prkdc	hsa-miR-146a-5p	2	hsa-miR-421	hells	2	6	-1.60	1.50E-08	-1.55	1.24E-03
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	ticrr	2	6	-1.60	1.50E-08	-2.34	1.54E-05
Prkdc	hsa-miR-146a-5p	2	hsa-miR-27a-5p	hist1h1b	2	2	-1.60	1.50E-08	-2.47	2.20E-04
Prkdc	hsa-miR-146a-5p	2	hsa-miR-16-5p	pla2g15	3	14	-1.60	1.50E-08	-1.59	2.47E-12
Prkdc	hsa-miR-146a-5p	2	hsa-miR-522-5p	nln	3	6	-1.60	1.50E-08	-1.59	8.27E-07
Prkdc	hsa-miR-146a-5p	2	hsa-miR-424-5p	sgk1	2	16	-1.60	1.50E-08	-1.96	3.10E-09

Table S16: **List of possible CB regulator-target relationships identified considering mmu-miR-146a-5p as a regulator**

Gene with tandem miRNA pairs (Gene A)	CB-pair Regulator	Binding sites of CB-pair regulator in Gene A	Secondary Effector miRNA	Target gene of secondary effector miRNA Gene B (Secondary effector gene)	Binding sites of Secondary effector miRNA in Gene B (Secondary effector gene)	Counts of miRNA CB-pair (prevalence)	FC (Gene A)	P-value (Gene A)	FC (Gene B)	P-value (Gene B)
mmd	mmu-miR-146a-5p	2	mmu-miR-101a-3p	spc25	2	20	-1.62	5.48E-05	-2.93	1.11E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-101a-3p	spc25	2	20	-1.85	5.54E-04	-2.93	1.11E-05

rbl1	mmu-miR-146a-5p	2	mmu-miR-101a-3p	spc25	2	20	-1.69	1.30E-04	-2.93	1.11E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-122-5p	phka2	2	13	-1.85	5.54E-04	-1.51	9.02E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-142a-3p	mis18bp1	2	25	-1.62	5.48E-05	-2.05	2.32E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-142a-3p	mis18bp1	2	25	-1.85	5.54E-04	-2.05	2.32E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-155-5p	ccnd1	2	24	-1.69	1.30E-04	-2.37	9.96E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-16-5p	mcm5	2	39	-1.85	5.54E-04	-3.33	5.14E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-16-5p	mcm5	2	39	-1.62	5.48E-05	-3.33	5.14E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-16-5p	mcm5	2	39	-1.69	1.30E-04	-3.33	5.14E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-16-5p	mcm5	2	39	-2.06	1.32E-05	-3.33	5.14E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-16-5p	ncapg2	3	39	-1.85	5.54E-04	-1.52	1.56E-03
mmd	mmu-miR-146a-5p	2	mmu-miR-16-5p	ncapg2	3	39	-1.62	5.48E-05	-1.52	1.56E-03
rbl1	mmu-miR-146a-5p	2	mmu-miR-16-5p	ncapg2	3	39	-1.69	1.30E-04	-1.52	1.56E-03
irf4	mmu-miR-146a-5p	2	mmu-miR-16-5p	ncapg2	3	39	-2.06	1.32E-05	-1.52	1.56E-03
irf4	mmu-miR-146a-5p	2	mmu-miR-20b-5p	kbtbd8	2	35	-2.06	1.32E-05	-1.82	7.82E-03
bach2	mmu-miR-146a-5p	4	mmu-miR-20b-5p	kbtbd8	2	35	-1.85	5.54E-04	-1.82	7.82E-03
rbl1	mmu-miR-146a-5p	2	mmu-miR-20b-5p	kbtbd8	2	35	-1.69	1.30E-04	-1.82	7.82E-03

irf4	mmu-miR-146a-5p	2	mmu-miR-21a-3p	rrm2	2	16	-2.06	1.32E-05	-2.99	2.48E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-21a-3p	rrm2	2	16	-1.85	5.54E-04	-2.99	2.48E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-21a-3p	aspm	2	16	-2.06	1.32E-05	-3.62	2.68E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-21a-3p	aspm	2	16	-1.85	5.54E-04	-3.62	2.68E-06
irf4	mmu-miR-146a-5p	2	mmu-miR-21a-3p	ncapd3	2	16	-2.06	1.32E-05	-1.89	2.29E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-21a-3p	ncapd3	2	16	-1.85	5.54E-04	-1.89	2.29E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-21a-3p	fen1	2	16	-2.06	1.32E-05	-1.94	4.95E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-21a-3p	fen1	2	16	-1.85	5.54E-04	-1.94	4.95E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-23b-3p	mcm4	2	21	-2.06	1.32E-05	-2.39	6.22E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-23b-3p	mcm4	2	21	-1.85	5.54E-04	-2.39	6.22E-06
rbl1	mmu-miR-146a-5p	2	mmu-miR-23b-3p	mcm4	2	21	-1.69	1.30E-04	-2.39	6.22E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-26a-5p	psip1	2	30	-1.85	5.54E-04	-1.64	4.57E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-26b-5p	st6gal1	2	30	-1.85	5.54E-04	-1.56	6.66E-04
irf4	mmu-miR-146a-5p	2	mmu-miR-26b-5p	st6gal1	2	30	-2.06	1.32E-05	-1.56	6.66E-04
mmd	mmu-miR-146a-5p	2	mmu-miR-27b-3p	il7r	2	33	-1.62	5.48E-05	-1.55	1.80E-03
irf4	mmu-miR-146a-5p	2	mmu-miR-27b-3p	rgs1	2	33	-2.06	1.32E-05	-2.46	3.13E-04

mmd	mmu-miR-146a-5p	2	mmu-miR-27b-3p	rgs1	2	33	-1.62	5.48E-05	-2.46	3.13E-04
bach2	mmu-miR-146a-5p	4	mmu-miR-29b-3p	angptl4	3	25	-1.85	5.54E-04	-3.04	7.37E-04
bach2	mmu-miR-146a-5p	4	mmu-miR-29b-3p	col4a5	2	25	-1.85	5.54E-04	-2.53	2.91E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-30e-5p	serpinb6b	2	25	-1.62	5.48E-05	-1.69	8.85E-04
irf4	mmu-miR-146a-5p	2	mmu-miR-30e-5p	serpinb6b	2	25	-2.06	1.32E-05	-1.69	8.85E-04
bach2	mmu-miR-146a-5p	4	mmu-miR-30e-5p	serpinb6b	2	25	-1.85	5.54E-04	-1.69	8.85E-04
rbl1	mmu-miR-146a-5p	2	mmu-miR-381-3p	rrm1	2	23	-1.69	1.30E-04	-2.57	3.15E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-381-3p	rrm1	2	23	-1.62	5.48E-05	-2.57	3.15E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-381-3p	cadm1	2	23	-1.69	1.30E-04	-2.56	2.46E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-381-3p	cadm1	2	23	-1.62	5.48E-05	-2.56	2.46E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-381-3p	lipa	2	23	-1.69	1.30E-04	-1.65	8.14E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-381-3p	lipa	2	23	-1.62	5.48E-05	-1.65	8.14E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-381-3p	lipa	2	23	-1.85	5.54E-04	-1.65	8.14E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467f	gmnn	2	30	-1.85	5.54E-04	-1.71	8.91E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-467f	gmnn	2	30	-2.06	1.32E-05	-1.71	8.91E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467f	gmnn	2	30	-1.69	1.30E-04	-1.71	8.91E-05

mmd	mmu-miR-146a-5p	2	mmu-miR-467f	gmnn	2	30	-1.62	5.48E-05	-1.71	8.91E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467f	nasp	2	30	-1.85	5.54E-04	-1.56	1.18E-04
irf4	mmu-miR-146a-5p	2	mmu-miR-467f	nasp	2	30	-2.06	1.32E-05	-1.56	1.18E-04
rbl1	mmu-miR-146a-5p	2	mmu-miR-467f	nasp	2	30	-1.69	1.30E-04	-1.56	1.18E-04
mmd	mmu-miR-146a-5p	2	mmu-miR-467f	nasp	2	30	-1.62	5.48E-05	-1.56	1.18E-04
bach2	mmu-miR-146a-5p	4	mmu-miR-467f	kif22	2	30	-1.85	5.54E-04	-2.63	1.96E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467f	kif22	2	30	-1.69	1.30E-04	-2.63	1.96E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467f	kif22	2	30	-1.62	5.48E-05	-2.63	1.96E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467g	anln	2	24	-1.62	5.48E-05	-3.47	2.15E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467g	anln	2	24	-1.69	1.30E-04	-3.47	2.15E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467g	anln	2	24	-1.85	5.54E-04	-3.47	2.15E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467g	e2f8	2	24	-1.62	5.48E-05	-2.68	1.95E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467g	e2f8	2	24	-1.69	1.30E-04	-2.68	1.95E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467g	e2f8	2	24	-1.85	5.54E-04	-2.68	1.95E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467g	alcam	2	24	-1.62	5.48E-05	-1.60	2.97E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467g	alcam	2	24	-1.69	1.30E-04	-1.60	2.97E-05

bach2	mmu-miR-146a-5p	4	mmu-miR-467g	alcam	2	24	-1.85	5.54E-04	-1.60	2.97E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467g	dek	2	24	-1.62	5.48E-05	-1.64	4.65E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467g	dek	2	24	-1.69	1.30E-04	-1.64	4.65E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467g	dek	2	24	-1.85	5.54E-04	-1.64	4.65E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-467g	tnfsf8	4	24	-1.62	5.48E-05	-1.67	4.15E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-467g	tnfsf8	4	24	-1.69	1.30E-04	-1.67	4.15E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-467g	tnfsf8	4	24	-1.85	5.54E-04	-1.67	4.15E-05
rbl1	mmu-miR-146a-5p	2	mmu-miR-7a-5p	mki67	4	27	-1.69	1.30E-04	-3.13	2.75E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-92a-3p	asph	2	24	-1.62	5.48E-05	-1.57	2.46E-03
mmd	mmu-miR-146a-5p	2	mmu-miR-92a-3p	myo18a	2	24	-1.62	5.48E-05	-1.63	2.92E-02
irf4	mmu-miR-146a-5p	2	mmu-miR-92b-3p	smc4	2	17	-2.06	1.32E-05	-1.67	4.06E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-92b-3p	smc4	2	17	-1.62	5.48E-05	-1.67	4.06E-05
mmd	mmu-miR-146a-5p	2	mmu-miR-92b-3p	stk39	2	17	-1.62	5.48E-05	-2.38	5.40E-03
mmd	mmu-miR-146a-5p	2	mmu-miR-92b-3p	rapgef6	2	17	-1.62	5.48E-05	-1.93	1.60E-03
irf4	mmu-miR-146a-5p	2	mmu-miR-93-5p	mastl	2	33	-2.06	1.32E-05	-2.18	1.24E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-93-5p	mastl	2	33	-1.85	5.54E-04	-2.18	1.24E-05

rbl1	mmu-miR-146a-5p	2	mmu-miR-93-5p	mastl	2	33	-1.69	1.30E-04	-2.18	1.24E-05
irf4	mmu-miR-146a-5p	2	mmu-miR-93-5p	kif23	2	33	-2.06	1.32E-05	-3.63	6.87E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-93-5p	kif23	2	33	-1.85	5.54E-04	-3.63	6.87E-06
rbl1	mmu-miR-146a-5p	2	mmu-miR-93-5p	kif23	2	33	-1.69	1.30E-04	-3.63	6.87E-06
bach2	mmu-miR-146a-5p	4	mmu-miR-98-5p	slc25a13	2	17	-1.85	5.54E-04	-1.65	1.30E-04
irf4	mmu-miR-146a-5p	2	mmu-miR-98-5p	lmnb1	2	17	-2.06	1.32E-05	-3.57	3.43E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-98-5p	lmnb1	2	17	-1.85	5.54E-04	-3.57	3.43E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-98-5p	tgfbr1	3	17	-1.85	5.54E-04	-1.69	9.29E-05
bach2	mmu-miR-146a-5p	4	mmu-miR-98-5p	mcm6	2	17	-1.85	5.54E-04	-2.77	1.86E-05

Bibliography

Afgan, E., Baker, D., Batut, B., vandenBeek, M., Bouvier, D., Čech, M., et al. (2018, 05). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *NucleicAcidsResearch*, *46*(W1), W537-W544. Available from https://doi.org/10.1093/nar/gky379

Aibar, S., Fontanillo, C., Droste, C., & De Las Rivas, J. (2015, 01). Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics*, *31*(10), 1686-1688. Avail- able from https://doi.org/10.1093/bioinformatics/btu864

Albert, R. (2005). Scale-free networks in cell biology. *JournalofCellScience*, *118*(21), 4947–4957. Available from https://jcs.biologists.org/content/118/21/4947

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378-382. Available from https://doi.org/10.1038/35019019

Alsmark, C., Foster, P. G., Sicheritz-Ponten, T., Nakjang, S., Martin Embley, T., & Hirt, R. P. (2013). Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *GenomeBiology*, *14*(2), R19. Available from https://doi.org/10.1186/gb-2013-14-2-r19

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *JournalofMolecularBiology*, *215*(3), 403 - 410. Available from http://www.sciencedirect.com/science/article/pii/S0022283605803602

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997, 09). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25* (17), 3389-3402. Available from https://doi.org/10.1093/nar/25.17.3389

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014, 11). OMIM.org: Online Mendelian Inheritance in Man (OMIM R), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, *43* (D1), D789-D798. Available from https://doi.org/10.1093/nar/gku1205

Anders, S., & Huber, W. (2010, Oct 27). Differential expression analysis for sequence count data. *GenomeBiology*, *11*(10), R106. Available from https://doi.org/10.1186/gb-2010-11-10-r106

Andrusier, N., Nussinov, R., & Wolfson, H. J. (2007). Firedock: Fast interaction refinement in molecular docking. *Proteins: Struture,Function,andBioinformatics*, *69*(1), 139-159. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21495

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., & Baker, D. (2017). Origins of coevolution between residues distant in protein 3d structures. *Proceedings of the National Academy of Sciences*, *114*(34), 9122–9127. Available from https://www.pnas.org/content/114/34/9122

Arbi, M., Pefani, D.E., Kyrousi, C., Lalioti, M.E., Kalogeropoulou, A., Papanastasiou, A. D., et al. (2016). Gemc1 controls multiciliogenesisin the airwayepithelium. *EMBO reports*, *17*(3), 400-413. Available from https://www.embopress.org/doi/abs/10.15252/embr.201540882

Arnaiz, O., Cohen, J., Tassin, A.-M., & Koll, F. (2014). Remodeling cildb, a popular database for cilia and links for ciliopathies. *Cilia*, *3* (1), 9. Available from https://doi.org/10.1186/2046-2530-3-9

Arnaiz, O., Malinowska, A., Klotz, C., Sperling, L., Dadlez, M., Koll, F., et al. (2009, 12). Cildb: a knowledgebase for centrosomes and cilia. *Database*, *2009*. Available from https://doi.org/10.1093/database/bap022 (bap022)

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., et al. (2009, 10). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research*, *38*(suppl$_1$), *D*457-*D*462. *Available from* https://doi.org/10.1093/nar/gkp851

Atayde, V. D., Aslan, H., Townsend, S., Hassani, K., Kamhawi, S., & Olivier, M. (2015, Nov 03). Exosome secretion by the parasitic protozoan *Leishmania* within the sand fly midgut. *Cell Reports*, *13* (5), 957-967. Available from https://doi.org/10.1016/j.celrep.2015.09.058

Baek, H., Shin, H. J., Kim, J.-J., Shin, N., Kim, S., Yi, M.-H., et al. (2017). Primary cilia modulate TLR4-mediated inflammatory responses in hippocampal neurons. *Journal of Neuroinflammation*, *14*(1), 189. Available from https://doi.org/10.1186/s12974-017-0958-7

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5* (2), 101-113. Available from https://doi.org/10.1038/nrg1272

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012, 11). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41* (D1), D991-D995. Available from https://doi.org/10.1093/nar/gks1193

Barta, A., & Jantsch, M. F. (2017). RNA in disease and development. *RNA Biology*, *14*(5), 457-459. Available from https://doi.org/10.1080/15476286.2017.1316929 (PMID: 28402218)

Berna, G., & Romero-Gomez, M.(2020). The role of nutrition in non-alcoholic fatty liver disease: Pathophysiology and management. *Liver International*, *40*(S1), 102-108. Available from https://onlinelibrary.wiley.com/doi/abs/10.1111/liv.14360

Bettencourt-Dias, M., Hildebrandt, F., Pellman, D., Woods, G., & God- inho, S. A. (2011). Centrosomes and cilia in human disease. *Trends in Genetics*, *27*(8), 307 - 315. Available from http://www.sciencedirect.com/science/article/pii/S0168952511000655

Bhattacharyya, M., & Chakrabarti, S. (2015). Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malaria Journal*, *14*(1), 70. Available from https://doi.org/10.1186/s12936-015-0562-1

Bisgrove, B. W., & Yost, H. J. (2006). The roles of cilia in developmental disorders and disease. *Development*, *133*(21), 4131–4143. Available from https://dev.biologists.org/content/133/21/4131

Boldt, K., Reeuwijk, J. van, Lu, Q., Koutroumpas, K., Nguyen, T.-M. T., Texier, Y., et al. (2016). An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nature Communications*, *7* (1), 11491. Available from https://doi.org/10.1038/ncomms11491

Boratyn, G. M., Schäffer, A. A., Agarwala, R., Altschul, S. F., Lipman, D. J., & Mad- den, T. L. (2012). Domain enhanced lookup time accelerated blast. *Biology Direct* , *7* (1), 12. Available from https://doi.org/10.1186/1745-6150-7-12

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, *253*(5016), 164–170. Available from https://science.sciencemag.org/content/253/5016/164

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M. L., Magrane, M., Tognolli, M., et al. (2016, 02). The UniProtKB guide to the human proteome. *Database*, *2016* . Available from https://doi.org/10.1093/database/bav120 (bav120)

Bulyk, M. L.(2003). Computational prediction of transcription-factor binding site locations. *Genome Biology*, *5*(1), 201. Available from https://doi.org/10.1186/gb-2003-5-1-201

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009, Dec 15). Blast+: architecture and applications. *BMC bioinformatics*, *10* , 421-421. Available from https://pubmed.ncbi.nlm.nih.gov/20003500 (20003500[pmid])

Caron, A., Xu, X., & Lin, X. (2012). Wnt/β-catenin signaling directly regulates Foxj1 expression and ciliogenesis in zebrafish kupffer's vesicle. *Development*, *139*(3), 514–524. Available from https://dev.biologists.org/content/139/3/514

Chakrabarti, S., & Panchenko, A. R. (2009, Apr). Coevolution in defining the functional specificity. *Proteins*, *75*(1), 231-240. Available from https://pubmed.ncbi.nlm.nih.gov/18831050 (18831050[pmid])

Chakrabarti, S., & Panchenko, A. R. (2010, 01). Structural and functional roles of coevolved sites in proteins. *PLOS ONE*, *5* (1), 1-10. Available from https://doi.org/10.1371/journal.pone.0008591

Chatr-aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., et al. (2016, 12). The BioGRID interaction database: 2017 update. *Nucleic Acids Research*, *45*(D1), D369-D379. Available from https://doi.org/10.1093/nar/gkw1102

Chaussabel, D., Semnani, R. T., McDowell, M. A., Sacks, D., Sher, A., & Nutman, T. B. (2003, 07). Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood* , *102* (2), 672-681. Available from https://doi.org/10.1182/blood-2002-10-3232

Chen, L., Xiong, Z., Sun, L., Yang, J., & Jin, Q. (2011, 11).VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Research*, *40* (D1), D641-D645. Available from https://doi.org/10.1093/nar/gkr989

Choi, S. Y., Chacon-Heszele, M. F., Huang, L., McKenna, S., Wilson, F. P., Zuo, X., et al. (2013). Cdc42 deficiency causes ciliary abnormalities and cystic kidneys. *Journal of the American Society of Nephrology*, *24* (9), 1435–1450. Available from https://jasn.asnjournals.org/content/24/9/1435

Choksi, S. P., Babu, D., Lau, D., Yu, X., & Roy, S. (2014). Systematic discovery of novel ciliary genes through functional genomics in the zebrafish. *Development*, *141*(17), 3410–3419. Available from https://dev.biologists.org/content/141/17/3410

Choksi, S. P., Lauter, G., Swoboda, P., & Roy, S. (2014). Switching on cilia: transcriptional networks regulating ciliogenesis. *Development*, *141* (7), 1427–1441. Available from https://dev.biologists.org/content/141/7/1427

Chou, C.H., Chang, N.W., Shrestha, S., Hsu, S.D., Lin, Y.L., Lee, W.H., et al. (2015, 11). miRTarBase 2016: updates to the experimentally validated miRNA- target interactions database. *Nucleic Acids Research*, *44*(D1), D239-D247. Available from https://doi.org/10.1093/nar/gkv1258

Chuang, H.-Y., Hofree, M., & Ideker, T. (2010). A decade of systems biology. *Annual Review of Cell and Developmental Biology*, *26*(1), 721-744. Available from https://doi.org/10.1146/annurev-cellbio-100109-104122 (PMID: 20604711)

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review* , *51* (4), 661-703. Available from https://doi.org/10.1137/070710111

Clement, C., Ajbro, K., Koefoed, K., Vestergaard, M., Veland, I., HenriquesdeJesus, M., et al. (2013). Tgf- signaling is associated with endocytosis at the pocket region of the primary cilium. *Cell Reports*, *3* (6), 1806 - 1814. Available from http://www.sciencedirect.com/science/article/pii/S2211124713002374

Cokelaer, T., Pultz, D., Harder, L. M., Serra-Musach, J., & Saez-Rodriguez, J. (2013, 09). BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics*, *29* (24), 3241-3242. Available from https://doi.org/10.1093/bioinformatics/btt547

Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. (2004a, 07). ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic Acids Research*, *32*(suppl$_2$),*W*96-*W*99. Availablefrom https://doi.org/10.1093/nar/gkh354

Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. (2004b, 01). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, *20*(1), 45-50. Available from https://doi.org/10.1093/bioinformatics/btg371

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016, Jan 26). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. Available from https://doi.org/10.1186/s13059-016-0881-8

Consortium, T. U. (2014, 10). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(D1), D204-D212. Available from https://doi.org/10.1093/nar/gku989

Consortium, T. U. (2018) UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research*, 47(D1),D506–D515, Available from https://doi.org/10.1093/nar/gky1049

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2013, 11). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42* (D1), D472- D477. Available from https://doi.org/10.1093/nar/gkt1102

Dam, T. J. van, Wheway, G., Slaats, G. G., Huynen, M. A., Giles, R. H., & Group, S. S. (2013). The SYSCILIA gold standard (SCGSV1) of known ciliary components and its applications within a systems biology consortium. *Cilia*, *2* (1), 7. Available from https://doi.org/10.1186/2046-2530-2-7

Danielian, P. S., Hess, R. A., & Lees, J. A. (2016). E2f4 and E2f5 are essential for the development of the male reproductive system. *Cell Cycle*, *15*(2), 250-260. Available from https://doi.org/10.1080/15384101.2015.1121350 (PMID: 26825228)

Das, A. A., Sharma, O. P., Kumar, M. S., Krishna, R., & Mathur, P. P. (2013). PepBind: A comprehensive database and computational tool for analysis of protein–peptide interactions. *Genomics, Proteomics & Bioinformatics*, *11*(4), 241 - 246. Available from http://www.sciencedirect.com/science/article/pii/S1672022913000739

David, A., & Sternberg, M. J. (2015).The contribution of missense mutations in core and rim residues of protein–protein interfaces to human disease. *Journal of Molecular Biology*, *427*(17), 2886 - 2898. Available from http://www.sciencedirect.com/science/article/pii/S0022283615003824

DeTure, M. A., & Dickson, D. W. (2019, Aug 02). The neuropathological diagnosis of Alzheimer's disease. *Molecular Neurodegeneration*, *14*(1), 32. Available from https://doi.org/10.1186/s13024-019-0333-5

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731-1737. Available from https://doi.org/10.1021/ja026939x (PMID: 12580598)

Dong, Z., & Chen, Y. (2013, Oct 01). Transcriptomics: Advances and approaches. *ScienceChina LifeSciences*,*56*(10),960–967. Available from https://doi.org/10.1007/s11427-013-4557-2

Downing, T., Imamura, H., Decuypere, S., Clark, T. G., Coombs, G. H., Cotton, J. A., et al. (2011, Dec). Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genomeresearch*,*21* (12),2143-2156. Available from https://pubmed.ncbi.nlm.nih.gov/22038251 (22038251[pmid])

Duhovny, D., Nussinov, R., & Wolfson, H. J. (2002). Efficient unbound docking of rigid molecules. In R. Guigó & D. Gusfield (Eds.), *Algorithms in bioinformatics* (pp.185–200). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dunn, S., Wahl, L., & Gloor, G. (2007, 12). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, *24* (3), 333-340. Available from https://doi.org/10.1093/bioinformatics/btm604

Díaz-Muñoz, M. D., & Turner, M. (2018). Uncovering the role of RNA-binding proteins in gene expression in the immune system. *Frontiers in Immunology*, 9 , 1094. Available from https://www.frontiersin.org/article/10.3389/fimmu.2018.01094

Eddy, S. R. (1998, 10). Profile hidden Markov models. *Bioinformatics*, *14* (9), 755-763. Available from https://doi.org/10.1093/bioinformatics/14.9.755

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. In *Genome informatics 2009* (Vol. 23, p. 205- 211). Published by imperial college press and distributed by world scientific publishing co. Available from https://www.worldscientific.com/doi/abs/10.1142/9781848165632_0019

Edgar, R., Domrachev, M., & Lash, A. E.(2002, 01). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30* (1), 207-210. Available from https://doi.org/10.1093/nar/30.1.207

Emmert-Streib, F., Dehmer, M., & Haibe-Kains, B. (2014). Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*, *2*, 38. Available from https://www.frontiersin.org/article/10.3389/fcell.2014.00038

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M.y., et al. (2006). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *15* (1), 5.6.1-5.6.30. Available from https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi05

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2017, 01). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46* (D1), D649-D655. Available from https://doi.org/10.1093/nar/gkx1132

Fares, M. A., & McNally, D. (2006, 09). CAPS: coevolution analysis using protein sequences. *Bioinformatics*, *22* (22), 2821-2822. Available from https://doi.org/10.1093/bioinformatics/btl493

Fares, M. A., & Travers, S. A. A.(2006). A novel method for detect- ing intramolecular coevolution: Adding a further dimension to selective constraints analyses. *Genetics*, *173*(1), 9–23. Available from https://www.genetics.org/content/173/1/9

Filipowicz, W., Bhattacharyya, S. N., & Sonenberg, N. (2008, Feb 01). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, *9* (2), 102-114. Available from https://doi.org/10.1038/nrg2290

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2013, 11).

Pfam: the protein families database. *Nucleic Acids Research*, *42* (D1), D222-D230. Available from https://doi.org/10.1093/nar/gkt1223

Fiser, A., Do, R. K. G., & Šali, A. (2000). Modeling of loops in protein structures. *Protein Science*, *9*(9), 1753-1773. Available from https://onlinelibrary.wiley.com/doi/abs/10.1110/ps.9.9.1753

Fliegauf, M., Benzing, T., & Omran, H. (2007). When cilia go bad: cilia defects and ciliopathies. *Nature Reviews Molecular Cell Biology*, *8* (11), 880-893. Available from https://doi.org/10.1038/nrm2278

Friedman, R. C., Farh, K. K.-H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19* (1), 92-105. Available from http://genome.cshlp.org/content/19/1/92.abstract

Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M., & Sanyal, A. J. (2018, Jul 01). Mechanisms of NAFLD development and therapeutic strategies. *Nature Medicine*, *24*(7), 908-922. Available from https://doi.org/10.1038/s41591-018-0104-9

Fu, J., Tang, W., Du, P., Wang, G., Chen, W., Li, J., et al. (2012). Identifying microRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis. *BMC Systems Biology*, *6*(1), 68. Available from https://doi.org/10.1186/1752-0509-6-68

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012, 10). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28* (23), 3150-3152. Available from https://doi.org/10.1093/bioinformatics/bts565

Gabaldón, T. (2005). Evolution of proteins and proteomes: A phylogenetics approach. *Evolutionary Bioinformatics*, *1*, 117693430500100004. Available from https://doi.org/10.1177/117693430500100004

Gao, M., Zhou, H., & Skolnick, J. (2015, Jul 07). Insights into disease-associated mutations in the human proteome through protein structural analysis. *Structure*, *23*(7), 1362-1369. Available from https://doi.org/10.1016/j.str.2015.03.028

Garraway, L. A., & Lander, E. S. (2013, Mar 28). Lessons from the cancer genome. *Cell*, *153* (1), 17-37. Available from https://doi.org/10.1016/j.cell.2013.03.002

Geremek, M., Zietkiewicz, E., Bruinenberg, M., Franke, L., Pogorzelski, A., Wijmenga, C., et al. (2014, 02). Ciliary genes are down-regulated in bronchial tissue of primary ciliary dyskinesia patients. *PLOS ONE* , *9* (2), 1-8. Available from https://doi.org/10.1371/journal.pone.0088216

Ginalski, K. (2006). Comparative modeling for protein structure prediction. *Current Opinion in Structural Biology*, *16*(2), 172 - 177. Available from http://www.sciencedirect.com/science/article/pii/S0959440X06000364 (Theory and simulation/Macromolecular assemblages)

Goetz, S. C., & Anderson, K. V. (2010). The primary cilium: a signalling centre during vertebrate development. *Nature Reviews Genetics*, *11* (5), 331-344. Available from https://doi.org/10.1038/nrg2774

Goswami, A., Mukherjee, K., Mazumder, A., Ganguly, S., Mukherjee, I., Chakrabarti, S., et al. (2020). MicroRNA exporter HuR clears the internalized pathogens by promoting pro-inflammatory response in infected macrophages. *EMBO Molecular Medicine*, *12*(3), e11011. Available from https://www.embopress.org/doi/abs/10.15252/emmm.201911011

Gupta, G., Coyaud Étienne, Gonçalves, J., Mojarad, B., Liu, Y., Wu, Q., et al. (2015). A dynamic protein interaction landscape of the human centrosome-cilium interface. *Cell*, *163* (6), 1484 - 1499. Available from http://www.sciencedirect.com/science/article/pii/S009286741501421X

Gupta, G., Oghumu, S., & Satoskar, A. R. (2013). Chapter five - mechanisms of immune evasion in leishmaniasis. In S. Sariaslani & G. M. Gadd (Eds.), (Vol. 82, p. 155 - 184). Academic Press. Available from http://www.sciencedirect.com/science/article/pii/B9780124076792000053

Hakes, L., Pinney, J. W., Robertson, D. L., & Lovell, S. C. (2008, Jan 01). Protein-protein interaction networks and biology—what's the connection? *Nature Biotechnology*, *26* (1), 69-72. Available from https://doi.org/10.1038/nbt0108-69

Hanahan, D., & Weinberg, R. (2011). Hallmarks of cancer: The next generation. *Cell* , *144* (5), 646-674. Available from http://www.sciencedirect.com/science/article/pii/S0092867411001279

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., et al. (2016, 02). Ensembl comparative genomics resources. *Database*, *2016*. Available from https://doi.org/10.1093/database/bav096 (bav096)

Hirt, R. P., Alsmark, C., & Embley, T. M. (2015, Feb). Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. *Current opinion in microbiology*, *23*, 155-162. Available from https://pubmed.ncbi.nlm.nih.gov/25483352 (25483352[pmid])

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., et al. (2014, sep). Sequence co-evolution gives 3d contacts and structures of protein complexes. *eLife*, *3* , e03430. Available from https://doi.org/10.7554/eLife.03430

Horani, A., Ferkol, T. W., Dutcher, S. K., & Brody, S. L. (2016, Mar). Genetics and biology of primary ciliary dyskinesia. *Paediatric respiratory reviews*, *18*, 18-24. Available from https://www.ncbi.nlm.nih.gov/pubmed/26476603 (26476603[pmid])

Hsu, S. hao, Wang, B., Kota, J., Yu, J., Costinean, S., Kutay, H., et al. (2012, 8). Essential metabolic, anti-inflammatory, and anti-tumorigenic functions of miR-122 in liver. *The Journal of Clinical Investigation*, *122* (8), 2871-2883. Available from https://www.jci.org/articles/view/63539

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008, 11). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, *37* (1), 1-13. Available from https://doi.org/10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4* (1), 44-57. Available from https://doi.org/10.1038/nprot.2008.211

Huerta-Cepas, J., Dopazo, J., & Gabaldón, T. (2010). ETE: a python environment for tree exploration. *BMC Bioinformatics*, *11* (1), 24. Available from https://doi.org/10.1186/1471-2105-11-24

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., et al. (2015). The BioPlex network: A systematic exploration of the human interactome. *Cell*, *162* (2), 425 - 440. Available from http://www.sciencedirect.com/science/article/pii/S0092867415007680

Izawa, I., Goto, H., Kasahara, K., & Inagaki, M. (2015). Current topics of functional links between primary cilia and cell cycle. *Cilia*, *4* (1), 12. Available from https://doi.org/10.1186/s13630-015-0021-1

Jensen, V. S., Hvid, H., Damgaard, J., Nygaard, H., Ingvorsen, C., Wulff, E. M., et al. (2018, Jan 24). Dietary fat stimulates development of nafld more potently than dietary fructose in sprague–dawley rats. *Diabetology & Metabolic Syndrome*, *10* (1), 4. Available from https://doi.org/10.1186/s13098-018-0307-8

Jeong, H., Mason, S. P., Barabási, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, *411* (6833), 41-42. Available from https://doi.org/10.1038/35075138

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K., Rastas, P., et al. (2013). DNA-binding specificities of human transcription factors. *Cell*, *152*(1), 327 - 339. Available from http://www.sciencedirect.com/science/article/pii/S0092867412014961

Jones, D. T., Buchan, D. W. A., Cozzetto, D., & Pontil, M. (2011, 11). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, *28* (2), 184-190. Available from https://doi.org/10.1093/bioinformatics/btr638

Jones, T. J., Adapala, R. K., Geldenhuys, W. J., Bursley, C., AbouAlaiwi, W. A., Nauli, S. M., et al. (2012). Primary cilia regulates the directional migration and barrier integrity of endothelial cells through the modulation of hsp27 dependent actin cytoskeletal organization. *Journal of Cellular Physiology*, *227*(1), 70-76. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/jcp.22704

Jopling, C. (2012, Feb). Liver-specific microRNA-122: Biogenesis and function. *RNA biology*, *9* (2), 137-142. Available from https://pubmed.ncbi.nlm.nih.gov/22258222 (22258222[pmid])

Joshi, P. B., Kelly, B. L., Kamhawi, S., Sacks, D. L., & McMaster, W. (2002). Targeted gene deletion in Leishmania major identifies leishmanolysin (gp63) as a virulence factor. Molecular and Biochemical Parasitology , 120 (1), 33 - 40. Available from http://www.sciencedirect.com/science/article/pii/S0166685101004327

Juan, D. de, Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, *14* (4), 249-261. Available from https://doi.org/10.1038/nrg3414

Kanehisa, M., & Goto, S. (2000, Jan). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research*, *28*(1), 27-30. Available from https://www.ncbi.nlm.nih.gov/pubmed/10592173 (10592173[pmid])

Kasahara, K., Kawakami, Y., Kiyono, T., Yonemura, S., Kawamura, Y., Era, S., et al. (2014). Ubiquitin-proteasome system controls ciliogenesis at the initial step of axoneme extension. *Nature Communications*, *5* (1), 5081. Available from https://doi.org/10.1038/ncomms6081

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*,*30* (14),3059-3066. Available from https://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013, 01). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772-780. Available from https://doi.org/10.1093/molbev/mst010

Kaye, P., & Scott, P. (2011). Leishmaniasis: complexity at the host-pathogen interface. *Nature Reviews Microbiology*, *9*(8), 604-615. Available from https://doi.org/10.1038/nrmicro2608

Kedzierski, L., Montgomery, J., Bullen, D., Curtis, J., Gardiner, E., Jimenez-Ruiz, A., et al. (2004). A leucine-rich repeat motif of leishmania parasite surface antigen 2 binds to macrophages through the complement receptor 3. *The Journal of Immunology*,*172*(8),4902–4906. Available from https://www.jimmunol.org/content/172/8/4902

Kikuchi, A., & Monga, S. P. (2015). Pdgfr*α* in liver pathophysiology: emerging roles in development, regeneration, fibrosis, and cancer. *Gene expression*, *16* (3), 109-127. Available from https://pubmed.ncbi.nlm.nih.gov/25700367 (25700367[pmid])

Kim, J., Jo, H., Hong, H., Kim, M. H., Kim, J. M., Lee, J.-K., et al. (2015). Actin remodelling factors control ciliogenesis by regulating yap/taz activity and vesicle trafficking. *Nature Communications*, *6*(1), 6781. Available from https://doi.org/10.1038/ncomms7781

Klinger, M., Wang, W., Kuhns, S., Bärenz, F., Dräger-Meurer, S., Pereira, G., et al. (2014, Feb). The novel centriolar satellite protein ssx2ip targets cep290 to the ciliarytransition zone. *Molecular biologyofthecell*,*25*(4),495-507. Available from https://pubmed.ncbi.nlm.nih.gov/24356449 (24356449[pmid])

Kobori, M., Takahashi, Y., Sakurai, M., Ni, Y., Chen, G., Nagashimada, M., et al. (2017). Hepatic transcriptome profiles of mice with diet-induced nonalcoholic steatohepatitis treated with astaxanthin and vitamin E. *International Journal of Molecular Sciences*, *18*(3). Available from https://www.mdpi.com/1422-0067/18/3/593

Kocabayoglu, P., Lade, A., Lee, Y. A., Dragomir, A.-C., Sun, X., Fiel, M. I., et al. (2015, Jul 01). β-PDGF receptor expressed by hepatic stellate cells regulates fibrosis in murine liver injury, but not carcinogenesis. *Journal of Hepatology*, *63*(1), 141-147. Available from https://doi.org/10.1016/j.jhep.2015.01.036

Kohli, P., Höhne, M., Jüngst, C., Bertsch, S., Ebert, L. K., Schauss, A. C., et al. (2017). The ciliary membrane-associated proteome reveals actin-binding proteins as key components of cilia. *EMBO reports*, *18* (9), 1521-1535. Available from https://www.embopress.org/doi/abs/10.15252/embr.201643846

Korber, B. T., Farber, R. M., Wolpert, D. H., & Lapedes, A. S. (1993). Co-variation of mutations in the v3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15), 7176–7180. Available from https://www.pnas.org/content/90/15/7176

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65 (2), 392-406. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21117

Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, *372* (3), 774 - 797. Available from http://www.sciencedirect.com/science/article/pii/S0022283607006420

Kuzmanov, U., & Emili, A. (2013, Apr 30). Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine*, *5* (4), 37. Avail- able from https://doi.org/10.1186/gm441

Lau, P., Bossers, K., Janky, R., Salta, E., Frigerio, C. S., Barbash, S., et al. (2013). Alteration of the microRNA network during the progression of Alzheimer's disease. *EMBO Molecular Medicine*, *5*(10), 1613-1634. Available from https://www.embopress.org/doi/abs/10.1002/emmm.201201974

Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M., et al. (2011). Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *MolecularCell*,*43*(3),340-352. Available from http://www.sciencedirect.com/science/article/pii/S1097276511004229

Lecuit, M., Ohayon, H., Braun, L., Mengaud, J., & Cossart, P. (1997, Dec). Internalin of *Listeria monocytogenes* with an intact leucine-rich repeat region is sufficient to promote internalization. *Infectionandimmunity*,*65*(12),5309-5319. Available from https://pubmed.ncbi.nlm.nih.gov/9393831 (9393831[pmid])

Li, W., & Godzik, A. (2006, 05). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22* (13), 1658-1659. Available from https://doi.org/10.1093/bioinformatics/btl158

Li, W., Jaroszewski, L., & Godzik, A. (2001, 03). Clustering of highly homologous sequences to reduce the size of large protein databases . *Bioinformatics*, *17* (3), 282-283. Available from https://doi.org/10.1093/bioinformatics/17.3.282

Li, W., Jaroszewski, L., & Godzik, A. (2002, 01). Tolerating some redundancy significantly speeds up clustering of large protein databases . *Bioinformatics*, *18* (1), 77-82. Available from https://doi.org/10.1093/bioinformatics/18.1.77

Liang, W. S., Dunckley, T., Beach, T. G., Grover, A., Mastroeni, D., Walker, D. G., et al. (2007). Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiological Genomics*, *28* (3),311-322. Available from https://doi.org/10.1152/physiolgenomics.00208.2006 (PMID: 17077275)

Liang, W. S., Reiman, E. M., Valla, J., Dunckley, T., Beach, T. G., Grover, A., et al. (2008). Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences*, *105*(11), 4441–4446. Available from https://www.pnas.org/content/105/11/4441

Lopes, S. S., Lourenço, R., Pacheco, L., Moreno, N., Kreiling, J., & Saúde, L. (2010). Notch signalling regulates left-right asymmetry through ciliary length control. *Development*, *137* (21), 3625–3632. Available from https://dev.biologists.org/content/137/21/3625

Lorbek, G., Perse, M., Jeruc, J., Juvan, P., Gutierrez-Mariscal, F. M., Lewinska, M., et al. (2015). Lessons from hepatocyte-specific cyp51 knockout mice: Impaired cholesterol synthesis leads to oval cell-driven liver injury. *Scientific Reports*, *5* (1), 8777. Available from https://doi.org/10.1038/srep08777

Lovell, S. C., Davis, I. W., Arendall III, W. B., Bakker, P. I. W. de, Word, J. M., Prisant, M. G., et al. (2003). Structure validation by Cα geometry: ϕ,ψ and Cβ deviation. *Proteins:Structure, Function,andBioinformatics*,50(3),437-450. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10286

Lovell, S. C., & Robertson, D. L. (2010, 06). An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Molecular Biology and Evolution*, *27* (11), 2567-2575. Available from https://doi.org/10.1093/molbev/msq144

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017, 05). Transcriptomics technologies. *PLOS Computational Biology*, *13* (5), 1-23. Available from https://doi.org/10.1371/journal.pcbi.1005457

Lu, Y.C., Chang, S.H., Hafner, M., Li, X., Tuschl, T., Elemento, O., et al. (2014, Dec 24). ELAVL1 modulates transcriptome-wide miRNA binding in murine macrophages. *Cell Reports*, *9*(6), 2330-2343. Available from https://doi.org/10.1016/j.celrep.2014.11.030

Luo, D., Wilson, J. M., Harvel, N., Liu, J., Pei, L., Huang, S., et al. (2013). A systematic evaluation of miRNA:mRNA interactions involved in the migration and invasion of breast cancer cells. *Journal of Translational Medicine*, *11*(1), 57. Available from https://doi.org/10.1186/1479-5876-11-57

Lüthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *356* (6364), 83-85. Available from https://doi.org/10.1038/356083a0

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., et al. (2014, 11). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222-D226. Available from https://doi.org/10.1093/nar/gku1221

Mardis, E. R. (2018). Insights from large-scale cancer genome sequencing. *Annual Review of Cancer Biology*, *2*(1), 429-444. Available from https://doi.org/10.1146/annurev-cancerbio-050216-122035

Marino Buslje, C., Teppa, E., Di Doménico, T., Delfino, J. M., & Nielsen, M. (2010, 11). Networks of high mutual information define the structural proximity of catalytic sites: Implications for catalytic residue identification. *PLOS Computational Biology*, 6(11), 1-8. Available from https://doi.org/10.1371/journal.pcbi.1000978

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., & Šali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29 (1), 291-325. Available from https://doi.org/10.1146/annurev.biophys.29.1.291 (PMID: 10940251)

Matsuda, S., Kobayashi, M., & Kitagishi, Y. (2013, Jan 30). Roles for PI3K/AKT/PTEN pathway in cell signaling of nonalcoholic fatty liver disease. *ISRN Endocrinology*, *2013*, 472432. Available from https://doi.org/10.1155/2013/472432

May-Simera, H., Gumerson, J., Gao, C., Campos, M., Cologna, S., Beyer, T., et al. (2016). Loss of Macf1 abolishes ciliogenesis and disrupts apicobasal polarity establishment in the retina. *Cell Reports*, *17* (5), 1399 - 1413. Available from http://www.sciencedirect.com/science/article/pii/S2211124716313651

McKusick, V. A. (2007, Apr). Mendelian inheritance in man and its online version, OMIM. *American journal of human genetics*, *80* (4), 588-604. Available from https://www.ncbi.nlm.nih.gov/pubmed/17357067 (17357067[pmid])

Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., et al. (2015, 04). RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, *43* (W1), W50-W56. Available from https://doi.org/10.1093/nar/gkv362

Mintseris, J., & Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 53 (3), 629-639. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10432

Mintseris, J., & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, *102*(31), 10930–10935. Available from https://www.pnas.org/content/102/31/10930

Moradifard, S., Hoseinbeyki, M., Ganji, S. M., & Minuchehr, Z. (2018, Mar 19). Analysis of microRNA and gene expression profiles in Alzheimer's disease: A meta-analysis approach. *Scientific Reports*, 8(1), 4767. Available from https://doi.org/10.1038/s41598-018-20959-0

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. In *Methods in molecular biology* (Vol. 443, pp. 365–382). Humana Press. Available from https://doi.org/10.1007/978-1-59745-177-2_19

Mukherjee, I., Chakraborty, A. & Chakrabarti, S. (2016). Identification of internalin-A-like virulent proteins in *Leishmania donovani* . *Parasites Vectors* 9 (1):557, Available from https://doi.org/10.1186/s13071-016-1842-5

Mukherjee, I., Roy, S. & Chakrabarti, S. (2019). Identification of Important Effector Proteins in the FOXJ1 Transcriptional Network Associated With Ciliogenesis and Ciliary Function. *Frontiers in Genetics* 10:23. doi: 10.3389/fgene.2019.00023

Mukherjee, N., Corcoran, D., Nusbaum, J., Reid, D., Georgiev, S., Hafner, M., et al. (2011). Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular Cell*, *43*(3), 327-339. Available from http://www.sciencedirect.com/science/article/pii/S1097276511004217

Méndez, R., Leplae, R., De Maria, L., & Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: Current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, *52*(1), 51-67. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.10393

Nachury, M. V., Seeley, E. S., & Jin, H. (2010). Trafficking to the ciliary membrane: How to get across the periciliary diffusion barrier? *Annual Review of Cell and Developmental Biology*, *26*(1), 59-87. Available from https://doi.org/10.1146/annurev.cellbio.042308.113337 (PMID: 19575670)

Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics*, *30*(1), 371-403. Available from https://doi.org/10.1146/annurev.genet.30.1.371 (PMID: 8982459)

Neugebauer, J. M., Amack, J. D., Peterson, A. G., Bisgrove, B. W., & Yost, H. J. (2009, Apr 02). FGF signalling during embryo development regulates cilia length in diverse epithelia. *Nature*, *458* (7238), 651-654. Available from https://www.ncbi.nlm.nih.gov/pubmed/19242413 (19242413[pmid])

Nguyen, L. S., Fregeac, J., Bole-Feysot, C., Cagnard, N., Iyer, A., Anink, J., et al. (2018, Jun 19). Role of miR-146a in neural stem cell differentiation and neural lineage determination: relevance for neurodevelopmental disorders. *Molecular Autism*, 9(1), 38. Available from https://doi.org/10.1186/s13229-018-0219-3

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology* , *9* , 402. Available from https://www.frontiersin.org/article/10.3389/fendo.2018.00402

Ochoa, D., & Pazos, F. (2010, 03). Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, *26* (10), 1370-1371. Available from https://doi.org/10.1093/bioinformatics/btq137

Oliver, S. (2000). Guilt-by-association goes global. *Nature*, *403* (6770), 601-602. Available from https://doi.org/10.1038/35001165

Pan, J., You, Y., Huang, T., & Brody, S. L. (2007). Rhoa-mediated apical actin enrichment is required for ciliogenesis and promoted by Foxj1. *Journal of Cell Science*, *120*(11), 1868–1876. Available from https://jcs.biologists.org/content/120/11/1868

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., et al. (2011). Using graph theory to analyze biological networks. *BioData Mining*, *4*(1), 10. Available from https://doi.org/10.1186/1756-0381-4-10

Peacock, C. S., Seeger, K., Harris, D., Murphy, L., Ruiz, J. C., Quail, M. A., et al. (2007, Jul). Comparative genomic analysis of three leishmania species that cause diverse human disease. *Nature genetics*, 39(7), 839-847. Available from https://pubmed.ncbi.nlm.nih.gov/17572675 (17572675[pmid])

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, *85*(8), 2444–2448. Available from https://www.pnas.org/content/85/8/2444

Peláez, N., & Carthew, R. W. (2012). Chapter nine - biological robustness and the role of microRNAs: A network perspective. In E. Hornstein (Ed.), *MicroRNAs in development* (Vol. 99, p. 237 - 255). Academic Press. Available from http://www.sciencedirect.com/science/article/pii/B9780123870384000094

Peng, Y., & Croce, C. M. (2016, Jan 28). The role of microRNAs in human cancer. *Signal Transduction and Targeted Therapy*, *1*(1), 15004. Available from https://doi.org/10.1038/sigtrans.2015.4

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004, Oct 01). UCSF chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605-1612. Available from https://doi.org/10.1002/jcc.20084

Prodromou, N. V., Thompson, C. L., Osborn, D. P. S., Cogger, K. F., Ashworth, R., Knight, M. M., et al. (2012). Heat shock induces rapid resorption of primary cilia. *Journal of Cell Science*, *125*(18), 4297–4305. Available from https://jcs.biologists.org/content/125/18/4297

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2006, 11). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl$_1$), *D*61-*D*65. *Available from* https://doi.org/10.1093/nar/gkl842

Pu, M., Chen, J., Tao, Z., Miao, L., Qi, X., Wang, Y., et al. (2019). Regulatory network of miRNA on its target: coordination between transcriptional and post-transcriptional regulation of gene expression. *Cellular and Molecular Life Sciences*, *76*(3), 441-451. Available from https://doi.org/10.1007/s00018-018-2940-7

Quarmby, L. M., & Parker, J. D. (2005, 05). Cilia and the cell cycle? *Journal of Cell Biology*, *169*(5), 707-710. Available from https://doi.org/10.1083/jcb.200503053

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org/

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org/

Ramachandran, H., Herfurth, K., Grosschedl, R., Schäfer, T., & Walz, G. (2015, Jun 17). SUMOylation blocks the ubiquitin-mediated degradation of the nephronophthisis gene product Glis2/NPHP7. *PloSone*, *10*(6), e0130275- e0130275. Available from https://pubmed.ncbi.nlm.nih.gov/26083374 (26083374[pmid])

Readhead, B., Haure-Mirande, J.V., Funk, C. C., Richards, M. A., Shannon, P., Haroutunian, V., et al. (2018, Jul 11). Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron*, *99*(1), 64-82.e7. Available from https://www.ncbi.nlm.nih.gov/pubmed/29937276

Regan, T., Gill, A. C., Clohisey, S. M., Barnett, M. W., Pariante, C. M., Harrison, N. A., et al. (2018). Effects of anti-inflammatory drugs on the expression of tryptophan-metabolism genes by human macrophages. *Journal of leukocyte biology*, *103*(4), 681–692. Available from https://doi.org/10.1002/JLB.3A0617-261R

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., et al. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *14*(2), 482-517. Available from https://doi.org/10.1038/s41596-018-0103-9

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9* (2), 173-175. Available from https://doi.org/10.1038/nmeth.1818

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276 - 277. Available from http://www.sciencedirect.com/science/article/pii/S0168952500020242

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015, 01). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47-e47. Available from https://doi.org/10.1093/nar/gkv007

Rochette, A., McNicoll, F., Girard, J., Breton, M., Leblanc Éric, Bergeron, M. G., et al. (2005). Characterization and developmental gene regulation of a large gene family encoding amastin surface proteins in *Leishmania* spp. *Molecular and Biochemical Parasitology*, *140*(2), 205-220. Available from http://www.sciencedirect.com/science/article/pii/S0166685105000241

Rodriguez-Rivas, J., Marsili, S., Juan, D., & Valencia, A. (2016). Conservation of coevolving protein interfaces bridges prokaryote–eukaryote homologies in the twilight zone. *Proceedings of the National Academy of Sciences*, *113*(52), 15018–15023. Available from https://www.pnas.org/content/113/52/15018

Rose, P. W., Prlić, A., Bi, C., Bluhm, W. F., Christie, C. H., Dutta, S., et al. (2014, 11). The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Research*, *43*(D1), D345-D356. Available from https://doi.org/10.1093/nar/gku1214

Satoh, J. ichi. (2012). Molecular network of microRNA targets in Alzheimer's disease brains. *Experimental Neurology*, *235*(2), 436 - 446. Available from http://www.sciencedirect.com/science/article/pii/S0014488611003104 (MicroRNAs—Human Neurobiology and Neuropathology)

Scheckel, C., Drapeau, E., Frias, M. A., Park, C. Y., Fak, J., Zucker-Scharff, I., et al. (2016, feb). Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. *eLife*, *5* , e10421. Available from https://doi.org/10.7554/eLife.10421

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., & Wolfson, H. J. (2005, 07). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research*, *33*(suppl$_2$), *W*363-*W*367. *Available from* https://doi.org/10.1093/nar/gki481

Schroder, K., Irvine, K. M., Taylor, M. S., Bokil, N. J., Le Cao, K. A., Masterman, K. A., et al. (2012). Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), E944–E953. *Available from* https://doi.org/10.1073/pnas.1110156109

Schubert, W.D., Urbanke, C., Ziehm, T., Beier, V., Machner, M. P., Domann, E., et al.(2002). Structure of internalin, a major invasion protein of *Listeria monocytogenes*, in complex with its human receptor E-cadherin. *Cell*, *111*(6), 825 - 836. Available from http://www.sciencedirect.com/science/article/pii/S0092867402011364

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, *18*(12), 1257-1261. Available from https://doi.org/10.1038/82360

Sebastian, A., & Contreras-Moreira, B. (2014, 1). footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, *30*(2), 258-265. Available from https://doi.org/10.1093/bioinformatics/btt663

Shearer, R. F., & Saunders, D. N. (2016). Regulation of primary cilia formation by the ubiquitin proteasome system. *Biochemical Society Transactions*, *44*(5), 1265–1271. Available from https://doi.org/10.1042/BST20160174

Shi, W., Yang, J., Li, S., Shan, X., Liu, X., Hua, H., et al. (2015). Potential involvement of miR-375 in the premalignant progression of oral squamous cell carcinoma mediated via transcription factor klf5. *Oncotarget*, *6*(37), 40172-40185. Available from http://www.oncotarget.com/index.php?journal=oncotarget&page=article&op=view&path

Si, W., Shen, J., Zheng, H., & Fan, W. (2019, Feb 11). The role and mechanisms of action of microRNAs in cancer drug resistance. *Clinical Epigenetics*, *11*(1), 25. Available from https://doi.org/10.1186/s13148-018-0587-8

Silva-Almeida, M., Pereira, B. A. S., Ribeiro-Guimarães, M. L., & Alves, C. R. (2012). Proteinases as virulence factors in *Leishmania* spp. infection in mammals. *Parasites & Vectors*, *5*(1), 160. Available from https://doi.org/10.1186/1756-3305-5-160

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., et al. (2009, 11). The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Research,* (suppl$_1$)*, D*540-*D*544.*Available from* https://doi.org/10.1093/nar/gkp1026

Söding, J. (2004, 11). Protein homology detection by HMM–HMM comparison. *Bioinformatics*, *21*(7), 951-960. Available from https://doi.org/10.1093/bioinformatics/bti125

Söding, J., Biegert, A., & Lupas, A. N. (2005, 07). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, (suppl$_2$)*, W*244-*W* 248. *Availablefrom* https://doi.org/10.1093/nar/gki408

Sotomayor, M., Gaudet, R., & Corey, D. P. (2014). Sorting out a promiscuous superfamily: towards cadherin connectomics. *Trends in Cell Biology*, *24*(9), 524 - 536. Available from http://www.sciencedirect.com/science/article/pii/S0962892414000609

Southwood, D., & Ranganathan, S. (2019). Host-pathogen interactions. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of bioinformatics and computational biology* (p. 103 - 112). Oxford: Academic Press. Available from http://www.sciencedirect.com/science/article/pii/B9780128096338200885

Sowmya, G., Breen, E. J., & Ranganathan, S. (2015). Linking structural features of protein complexes and biological function. *Protein Science*, *24*(9), 1486-1494. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.2736

Spies, D., & Ciaudo, C. (2015). Dynamics in transcriptomics: Advancements in RNA-seq time course and downstream analysis. *Computational and Structural Biotechnology Journal*, *13*, 469 - 477. Available from http://www.sciencedirect.com/science/article/pii/S2001037015000392

Srikantan, S., & Gorospe, M. (2011, Aug 05). Uneclipsing HuR nuclear function. *Molecular Cell*, *43*(3), 319-321. Available from https://doi.org/10.1016/j.molcel.2011.07.016

Stamatakis, A. (2014, 01). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313. Available from https://doi.org/10.1093/bioinformatics/btu033

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006, 01). BioGRID: a general repository for interaction datasets. *Nucleic AcidsResearch*, (suppl$_1$), *D*535-*D*539. *Available from* https://doi.org/10.1093/nar/gkj109

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., et al. (2005, Sep 23). A human protein-protein interaction network: A resource for annotating the proteome. Cell , 122 (6), 957-968. Available from https://doi.org/10.1016/j.cell.2005.08.029

Stephens, R., & Lemieux, N. (1999). Molecular chaperones in cilia and flagella: Implications for protein turnover. *Cell Motility*, *44* (4), 274-283. Available from https://doi.org/10.1002/(SICI)1097-0169(199912)44:4<274::AID-CM5>3.0.CO;2-O

Stubbs, J. L., Oishi, I., Izpisúa Belmonte, J. C., & Kintner, C. (2008). The forkhead protein FoxJ1 specifies node-like cilia in xenopus and zebrafish embryos. *NatureGenetics*, *40*(12), 1454-1460. Available from https://doi.org/10.1038/ng.267

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2014, 10). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic AcidsResearch*, *43*(D1), D447-D452. Available from https://doi.org/10.1093/nar/gku1003

Tamada, T., Honjo, E., Maeda, Y., Okamoto, T., Ishibashi, M., Tokunaga, M., et al. (2006). Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (GCSF) receptor signaling complex. *Proceedings of the National Academy of Sciences*, *103* (9), 3135–3140. Available from https://www.pnas.org/content/103/9/3135

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2018, 10). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic AcidsResearch*, *47*(D1), D941-D947. Available from https://doi.org/10.1093/nar/gky1015

Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K., & Tolstoy, I. (2013, 12). RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Research*, *42* (D1), D553-D559. Available from https://doi.org/10.1093/nar/gkt1274

Terré, B., Piergiovanni, G., Segura-Bayona, S., Gil-Gómez, G., Youssef, S. A., Attolini, C. S.O., et al. (2016). Gemc1 is a critical regulator of multiciliated cell differentiation. *The EMBO Journal*, *35*(9), 942-960. Available from https://www.embopress.org/doi/abs/10.15252/embj.201592821

Thompson, D., Regev, A., & Roy, S. (2015). Comparative analysis of gene regulatory networks: From network reconstruction to evolution. *Annual Review of Cell and Developmental Biology*, *31*(1), 399-428. Available from https://doi.org/10.1146/annurev-cellbio-100913-012908 (PMID: 26355593)

Thomson, D. W., Bracken, C. P., & Goodall, G. J. (2011, 06). Experimental strategies for microRNA target identification. *NucleicAcidsResearch*, *39*(16), 6845-6853. Available from https://doi.org/10.1093/nar/gkr330

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A subcellular map of the human proteome. *Science*, *356*(6340). Available from https://science.sciencemag.org/content/356/6340/eaal3321

Todd, A. E., Orengo, C. A., & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology*, *307*(4), 1113 - 1143. Available from http://www.sciencedirect.com/science/article/pii/S0022283601945139

Turatsinze, J.V., Thomas-Chollier, M., Defrance, M., & Helden, J. van. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, *3* (10), 1578-1588. Available from https://doi.org/10.1038/nprot.2008.97

Ueno, N., & Wilson, M. E. (2012, Aug 01). Receptor-mediated phagocytosis of *Leishmania*: implications for intracellular survival. *Trends in Parasitology*,*28*(8),335-344. Available from https://doi.org/10.1016/j.pt.2012.05.002

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., et al. (2015). Tissue-based map of the human proteome. *Science*,*347*(6220). Available from https://science.sciencemag.org/content/347/6220/1260419

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, *28* (12), 1248-1250. Available from https://doi.org/10.1038/nbt1210-1248

Vandevenne, M., Delmarcelle, M., & Galleni, M. (2019). RNA regulatory networks as a control of stochasticity in biological systems. *Frontiers in Genetics*, *10*, 403. Available from https://www.frontiersin.org/article/10.3389/fgene.2019.00403

Vasudevan, S., & Steitz, J. A. (2007, Mar 23). Au-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell*, *128*(6), 1105-1118. Available from https://doi.org/10.1016/j.cell.2007.01.038

Vidal, M., Cusick, M. E., & Barab´asi, A.L. (2011, Mar 18). Interactome networks and human disease. *Cell* , 144 (6), 986-998. Available from https://doi.org/10.1016/j.cell.2011.02.016

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Ver- goulis, T., Kanellos, I., et al. (2014, 11). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Research*, *43*(D1), D153-D159. Available from https://doi.org/10.1093/nar/gku1215

Vladar, E. K., & Mitchell, B. J. (2016). It's a family act: the geminin triplets take center stage in motile ciliogenesis. *The EMBO Journal*, *35*(9), 904-906. Available from https://www.embopress.org/doi/abs/10.15252/embj.201694206

Vries, S. J. de, Dijk, A. D. J. van, Krzeminski, M., Dijk, M. van, Thureau, A., Hsu, V., et al. (2007). HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Structure,Function,and Bioinformatics*, *69*(4),726-733. Available from https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.21723

Wang, M., Qin, L., & Tang, B. (2019). MicroRNAs in Alzheimer's disease. *Frontiers in Genetics*, *10*, 153. Available from https://www.frontiersin.org/article/10.3389/fgene.2019.00153

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, *54* (1), 5.6.1-5.6.37. Available from https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.3

Webster, J. A., Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P., et al. (2009, Apr 10). Genetic control of human brain transcript expression in Alzheimer disease. *The American Journal of Human Genetics*, *84*(4), 445-458. Available from https://doi.org/10.1016/j.ajhg.2009.03.011

Wen, J., & Friedman, J. R. (2012, 8). miR-122 regulates hepatic lipid metabolism and tumor suppression. *The Journal of Clinical Investigation*, *122*(8), 2773-2776. Available from https://www.jci.org/articles/view/63966

Wuchty, S., Ravasz, E., & Barabási, A.L. (2006). The architecture of biological networks. In T. S. Deisboeck & J. Y. Kresh (Eds.), *Complex systems science in biomedicine* (pp. 165–181). Boston, MA: Springer US. Available from https://doi.org/10.1007/978-0-387-33532-2$_5$

Yu, G., & He, Q.Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.*, *12*, 477-479. Available from http://dx.doi.org/10.1039/C5MB00663E

Yu, G., Wang, L.G., Yan, G.R., & He, Q.Y. (2014, 10). DOSE: an R/Bioconductor package for disease ontology semantic and enrich- ment analysis. *Bioinformatics*, *31*(4), 608-609. Available from https://doi.org/10.1093/bioinformatics/btu684

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007, 04). The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics. *PLOS Computational Biology*, *3*(4), 1-8. Available from https://doi.org/10.1371/journal.pcbi.0030059

Yu, X., Ng, C. P., Habacher, H., & Roy, S. (2008). FoxJ1 transcription factors are master regulators of the motile ciliogenic program. *Nature Genetics*, *40* (12), 1445- 1453. Available from https://doi.org/10.1038/ng.263

Zariwala, M. A., Omran, H., & Ferkol, T. W. (2011, Sep). The emerging genetics of primary ciliary dyskinesia. *Proceedings of the American Thoracic Society*, *8*(5), 430-433. Available from https://www.ncbi.nlm.nih.gov/pubmed/21926394 (21926394[pmid])

Zhang, W.W., & Matlashewski, G. (2001). Characterization of the A2–A2rel gene cluster in *Leishmania donovani*: involvement of A2 in visceralization during infection. *Molecular Microbiology*, *39*(4), 935-948. Available from https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2958.2001.02286.x

Zhang, Y., Wen, J., & Yau, S. S.T. (2019). Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. *Genomics*, *111* (6), 1298 - 1305. Available from http://www.sciencedirect.com/science/article/pii/S0888754318304336

Zhu, M., Wang, Q., Zhou, W., Liu, T., Yang, L., Zheng, P., et al. (2018, May 16). Integrated analysis of hepatic mRNA and miRNA profiles identified molecular networks and potential biomarkers of nafld. *Scientific Reports*, *8*(1), 7628. Available from https://doi.org/10.1038/s41598-018-25743-8

Zundert, G. van, Rodrigues, J., Trellet, M., Schmitz, C., Kastritis, P., Karaca, E., et al. (2016). The HADDOCK2.2 web server: User- friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology*, *428*(4), 720 - 725. Available from http://www.sciencedirect.com/science/article/pii/S0022283615005379 (Computation Resources for Molecular Biology)

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779 - 815. Available from http://www.sciencedirect.com/science/article/pii/S0022283683716268

Publications & Reprints

Publications

Mukherjee, I., Chakraborty, A., & Chakrabarti, S. (2016). **Identification of internalin-A-like virulent proteins in *Leishmania donovani*.** *Parasites & vectors*, *9*(1), 557. doi:10.1186/s13071-016-1842-5

Mukherjee, I., Roy, S. and Chakrabarti, S. (2019) **Identification of Important Effector Proteins in the FOXJ1 Transcriptional Network Associated With Ciliogenesis and Ciliary Function.** *Frontiers in Genetics* 10:23. doi: 10.3389/fgene.2019.00023

Goswami, A., Mukherjee, K., Mazumder, A., Ganguly, S., **Mukherjee, I.,** Chakrabarti, S., Roy, S., Sundar, S., Chattopadhyay, K., & Bhattacharyya, S. N. (2020). **MicroRNA exporter HuR clears the internalized pathogens by promoting pro-inflammatory response in infected macrophages.** *EMBO molecular medicine*, *12*(3), e11011. https://doi.org/10.15252/emmm.201911011

Manuscripts submitted

Mukherjee, I., and Chakrabarti, S. **Co-evolutionary landscape at the interface and non-interface regions of protein-protein interaction complexes.**

De, D., **Mukherjee, I.,** Guha, S., Paidi, R., Chakrabarti, S., Biswas, S., Bhattacharyya, S. N. **Rheb-mTOR activation rescues amyloid beta-induced cognitive impairment and memory function in Alzheimer's disease brain.**

Bandyopadhyay, D., Basu, S., **Mukherjee, I.**, Chakraborty, R., Mukherjee, K., Adak, M., Chattopadhyay, K., Chakrabarti, S., Chakrabarti, P., and Bhattacharyya, S.N. **Lipid droplets promote phase separation of Ago2 to stop miRNA-mediated metaflammation.**

Manuscripts under preparation

Chatterjee, S., **Mukherjee, I.,** Bhattacharyya, S., Chakrabarti, S., Bhattacharyya, S.N. Cooperative biogenesis of miRNPs augments inflammatory response in mammalian macrophages.

Mukherjee *et al. Parasites & Vectors* (2016) 9:557
DOI 10.1186/s13071-016-1842-5

Parasites & Vectors

RESEARCH Open Access

Identification of internalin-A-like virulent proteins in *Leishmania donovani*

Ishita Mukherjee, Abhijit Chakraborty and Saikat Chakrabarti*

Abstract

Background: An active immune surveillance and a range of barriers to infection allow the host to effectively eliminate microbial pathogens. However, pathogens may use diverse strategies to subdue such host defences. For instance, one such mechanism is the use of leucine-rich repeat (LRR) proteins by pathogens (microbial) to cause infection. In this study, we aimed at identifying novel virulence factor(s) in *Leishmania donovani*, based on the possibility of lateral gene transfers of bacterial virulence factor(s) to *L. donovani*.

Methods: Rigorous homology searching protocols including Hidden Markov Model (HMM) and BLASTp based searches were employed to detect remote but significant similarities between *L. donovani* proteins and bacterial virulence factors.

Results: We found that some *L. donovani* proteins are similar to internalin-A (Inl-A) protein of *Listeria monocytogenes*, a surface LRR protein that helps mediate host cell invasion by interacting with E-cadherin on the cell membrane. However, to date, no such invasion mechanism has been reported in *Leishmania donovani*, the causative agent of visceral leishmaniasis. Moreover, a comparative LRR motif analysis of *L. donovani* Inl-A-like proteins against the Inl-A protein of *L. monocytogenes* revealed existence of characteristic consensus LRR regions, suggesting a reliable evolutionary relationship between them. Further, through rigorous three dimensional (3D) modeling of *L. donovani* Inl-A-like proteins and subsequent molecular docking studies we suggest the probability of human E-cadherin binding with the *L. donovani* Inl-A-like proteins.

Conclusions: We have identified three potential candidates (UniProt ID: E9B7L9, E9BMT7 and E9BUL5) of Inl-A-like LRR containing proteins in *L. donovani* with the help of systematic whole genome sequence analysis. Thus, herein we propose the existence of a novel class of Inl-A-like virulence factor proteins in *L. donovani* and other *Leishmania* species based on sequence similarity, phylogenetic analysis and molecular modelling studies in *L. donovani*.

Keywords: Cell invasion, LRR proteins, Inl-A-like proteins, Homology-based search, HMM-profile based search, *Leishmania donovani*

Background

Leishmania spp. (family Trypanosomatidae) are intracellular protozoan parasites that when transmitted to a mammalian host can cause a range of infectious diseases, collectively referred to as leishmaniasis. These parasites have two morphologically distinct variants during their life-cycle, a promastigote form in phlebotomine sand flies and an amastigote form in mammalian cells [1]. A complex array of processes is involved in the attachment and invasion of host cells, initially mediated by the promastigotes and subsequently by the amastigotes. In general, many proteins of *Leishmania* spp. have been identified as possible virulence factors. For instance, lipophosphoglycan and leishmanolysin are important in attachment, invasion and intracellular survival of the parasites [2, 3]. In addition, the leucine-rich repeat (LRR)-containing proteins proteophosphoglycan and parasite surface antigen 2 also participate in parasite attachment and invasion of host cells [2]. Moreover, the A2 protein is important for the survival of the pathogen in visceral organs [4]. Amastin, amastin-like surface protein and cysteine proteases are other groups of proteins that attribute virulence to this group of parasites

* Correspondence: saikat@iicb.res.in
Structural Biology and Bioinformatics Division, Council for Scientific and Industrial Research (CSIR) - Indian Institute of Chemical Biology (IICB), Kolkata, West Bengal, India

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 March 2019
doi: 10.3389/fgene.2019.00023

Identification of Important Effector Proteins in the FOXJ1 Transcriptional Network Associated With Ciliogenesis and Ciliary Function

Ishita Mukherjee[1], Sudipto Roy[2,3,4] and Saikat Chakrabarti[1]**

[1] Translational Research Unit of Excellence, Structural Biology and Bioinformatics Division, Council for Scientific and Industrial Research – Indian Institute of Chemical Biology, Kolkata, India, [2] Institute of Molecular and Cell Biology, Singapore, Singapore, [3] Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, [4] Department of Biological Sciences, National University of Singapore, Singapore, Singapore

OPEN ACCESS

Edited by:
Jose Maria Carvajal-Gonzalez, Universidad de Extremadura, Spain

Reviewed by:
Angel Carlos Roman, Cajal Institute (CSIC), Spain
Georges Nemer, American University of Beirut, Lebanon

***Correspondence:**
Sudipto Roy
sudipto@imcb.a-star.edu.sg
Saikat Chakrabarti
saikat@iicb.res.in

Specialty section:
This article was submitted to Genetic Disorders, a section of the journal Frontiers in Genetics

Received: *27 July 2018*
Accepted: *15 January 2019*
Published: *01 March 2019*

Citation:
Mukherjee I, Roy S and Chakrabarti S (2019) Identification of Important Effector Proteins in the FOXJ1 Transcriptional Network Associated With Ciliogenesis and Ciliary Function. Front. Genet. 10:23. doi: 10.3389/fgene.2019.00023

Developmental defects in motile cilia, arising from genetic abnormalities in one or more ciliary genes, can lead to a common ciliopathy known as primary ciliary dyskinesia (PCD). Functional studies in model organisms undertaken to understand PCD or cilia biogenesis have identified 100s of genes regulated by Foxj1, the master regulator of motile ciliogenesis. However, limited systems based studies have been performed to elucidate proteins or network/s crucial to the motile ciliary interactome, although this approach holds promise for identification of multiple cilia-associated genes, which, in turn, could be utilized for screening and early diagnosis of the disease. Here, based on the assumption that FOXJ1-mediated regulatory and signaling networks are representative of the motile cilia interactome, we have constructed and analyzed the gene regulatory and protein–protein interaction network (PPIN) mediated by FOXJ1. The predicted FOXJ1 regulatory network comprises of 424 directly and 148 indirectly regulated genes. Additionally, based on gene ontology analysis, we have associated 17 directly and 6 indirectly regulated genes with possible ciliary roles. Topological and perturbation analyses of the PPIN (6927 proteins, 40,608 interactions) identified 121 proteins expressed in ciliated cells, which interact with multiple proteins encoded by FoxJ1 induced genes (FIG) as important interacting proteins (IIP). However, it is plausible that IIP transcriptionally regulated by FOXJ1 and/or differentially expressed in PCD are likely to have crucial roles in motile cilia. We have found 20 de-regulated topologically important effector proteins in the FOXJ1 regulatory network, among which some (PLSCR1, SSX2IP, ACTN2, CDC42, HSP90AA1, PIAS4) have previously reported ciliary roles. Furthermore, based on pathway enrichment of these proteins and their primary interactors, we have rationalized their possible roles in the ciliary interactome. For instance, 5 among these novel proteins that are involved in cilia associated signaling pathways (like Notch, Wnt, Hedgehog, Toll-like receptor etc.) could be 'topologically

Article

SOURCE DATA | TRANSPARENT PROCESS | OPEN ACCESS

EMBO Molecular Medicine

MicroRNA exporter HuR clears the internalized pathogens by promoting pro-inflammatory response in infected macrophages

Avijit Goswami[1,†], Kamalika Mukherjee[1,†], Anup Mazumder[1,†,‡], Satarupa Ganguly[1,†], Ishita Mukherjee[2], Saikat Chakrabarti[2], Syamal Roy[3], Shyam Sundar[4], Krishnananda Chattopadhyay[2] & Suvendra N Bhattacharyya[1,*]

Abstract

HuR is a miRNA derepressor protein that can act as miRNA sponge for specific miRNAs to negate their action on target mRNAs. Here we have identified how HuR, by inducing extracellular vesicles-mediated export of miRNAs, ensures robust derepression of miRNA-repressed cytokines essential for strong pro-inflammatory response in activated mammalian macrophages. *Leishmania donovani*, the causative agent of visceral leishmaniasis, on the contrary alters immune response of the host macrophage by a variety of complex mechanisms to promote anti-inflammatory response essential for the survival of the parasite. We have found that during *Leishmania* infection, the pathogen targets HuR to promote onset of anti-inflammatory response in mammalian macrophages. In infected macrophages, *Leishmania* also upregulate protein phosphatase 2A that acts on Ago2 protein to keep it in dephosphorylated and miRNA-associated form. This causes robust repression of the miRNA-targeted pro-inflammatory cytokines to establish an anti-inflammatory response in infected macrophages. HuR has an inhibitory effect on protein phosphatase 2A expression, and mathematical modelling of macrophage activation process supports antagonistic miRNA-modulatory roles of HuR and protein phosphatase 2A which mutually balances immune response in macrophage by targeting miRNA function. Supporting this model, ectopic expression of the protein HuR and simultaneous inhibition of protein phosphatase 2A induce strong pro-inflammatory response in the host macrophage to prevent the virulent antimonial drug-sensitive or drug-resistant form of *L. donovani* infection. Thus, HuR can act as a balancing factor of immune responses to curtail the macrophage infection process by the protozoan parasite.

Keywords Ago2 dephosphorylation; drug-resistant *Leishmania*; host–parasite interaction; miRNA export; protein phosphatase 2A

Subject Categories Immunology; Microbiology, Virology & Host Pathogen Interaction

DOI 10.15252/emmm.201911011 | Received 16 June 2019 | Revised 24 December 2019 | Accepted 8 January 2020 | Published online 7 February 2020

EMBO Mol Med (2020) 12: e11011

Introduction

Macrophages act as the first line of defence against the invading microbes in mammalian hosts which engulf the invading pathogens and kill them (Mogensen, 2009). However, the macrophages may fall prey to certain pathogens that inactivate the arsenals of the host macrophage through variety of complex mechanisms (Aderem & Underhill, 1999). The protozoan parasite *Leishmania donovani* (*Ld*), the causative agent of visceral leishmaniasis in human (Ready, 2014), can live and replicate within the host macrophages and thus can escape from the host immune system (Liu & Uzonna, 2012). Macrophage that otherwise gets activated by pathogen-derived molecules remains immunologically dormant when it is invaded by the pathogen *Leishmania* (Olivier *et al*, 2005). The parasite gets into the macrophage through host phagocytic activity and lives within the parasitophorous vacuoles, a special class of endosomal structures that get matured but do not get fused to lysosomes in infected cells through an unique mechanism induced by the internalized pathogen to prevent its own killing by host lysosomal machineries (Courret *et al*, 2002).

The inactivation of the macrophage defence machineries against invading pathogens is achieved through a selective action of pathogen-derived molecules on the host immune system (Contreras *et al*, 2010). *Leishmania* not only impairs the acquired immunity of the host by preventing processing of the pathogen-derived antigens and its presentation by infected macrophage or dendritic cells on

1 RNA Biology Research Laboratory, Molecular Genetics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India
2 Structural Biology and Bio-informatics Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India
3 National Institute of Pharmaceutical Educations and Research, Kolkata, India
4 Department of Medicine, Banaras Hindu University, Varanasi, India
*Corresponding author. Tel: +91 33 24995783; E-mails: suvendra@iicb.res.in; sb@csiriicb.in
†These authors contributed equally to this work
‡Present address: University of California Los Angeles, Los Angeles, CA, USA

List of Corrections

1. See redundant line on page 146. "Sequence similarity among bacterial Inl-A and *L. donovani* Inl-A-like proteins was assessed by determining the global sequence identities in EMBOSS (v6.2.0) (Rice, Longden, & Bleasby, 2000). Global sequence identities between bacterial Inl-A and *L. donovani* Inl-A-like proteins were determined in EMBOSS (v6.2.0) (Rice et al., 2000)."

Correction: According to the reviewer's suggestion, redundant line on page 146 has been corrected, as mentioned below.
"Sequence similarity among bacterial Inl-A and *L. donovani* Inl-A-like proteins was assessed by determining the global sequence identities in EMBOSS (v6.2.0) (Rice, Longden, & Bleasby, 2000)."

2. On page 99, the line "In this respect, probability theory was utilized to define an HMM profile representing specific features of the virulence region derived from a multiple sequence alignment (MSA) of the LRR region in bacterial Inl-A homologs.", it is better to say HMM was built rather than involving anything else.

Correction: On page 99, the line has been simplified (as mentioned below).
"In this respect, an HMM profile was prepared to represent the virulence region in bacterial Inl-A."

3. In Fig 2.5 ("Orthologs of *L.donovani* Inl-A-like proteins in *Leishmania* spp."), it is better to say phylogenetic tree of orthologs and refer method (program) used to construct such a tree rather than mentioning orthologs.

Correction: In Fig 2.5, figure legend has been corrected as per reviewer's suggestion (as mentioned below).
"Phylogenetic tree of orthologs of *L. donovani* Inl-A-like proteins in *Leishmania* spp. prepared using the maximum likelihood algorithm in RAxML."

4. The problem statement on page 119 (Chapter 3) is very convoluted [**"Problem Statement:** Given that information theory may be utilized to study interacting proteins that may co-evolve, determine co-evolving residue pairs that could be structurally or functionally relevant for an inter-cellular protein-protein interaction to occur. Alternately, identify whether alterations at these co-evolving residue pair positions could be functionally detrimental for the inter-cellular protein-protein interaction to occur?"]. This can be simplified to construct a better statement.

Correction: The problem statement on page 119 (Chapter 3) has been simplified as per reviewer's suggestion (as mentioned below).

"Problem Statement: Given that information theory may be utilized to study interacting proteins that co-evolve, determine residue pairs that could have crucial roles in protein-protein interactions."

5. On page 118, it is speculated that internalins are secreted ("Additionally, the experimental finding that leishmanial Inl-A-like proteins are likely to be secreted or exported in vesicles by promastigotes contained in infected sand fly inocula (Atayde et al., 2015) further hints at the likelihood of this interaction."). How does secreted protein help in virulence, when interaction between internalin and e-cadherin is important for adherence of pathogen.

Correction: On page 118, the statement about Inl-A being secreted (as cited above) has been mentioned to provide support that these virulent proteins are expressed. This was mentioned to provide additional support that an interaction between Inl-A and E-cadherin is possible. However, additional experiments that *L. donovani* Inl-A-like genes were found to be over-expressed in more virulent parasites (intracellular amastigotes) as compared to less virulent parasites, also supported our finding (Data from collaborators laboratory). In response to the reviewer's comment, the following correction has been incorporated.

"Additionally, the experimental finding that leishmanial Inl-A-like proteins are expressed in promastigote stage (Atayde et al., 2015), provides preliminary support that these predicted virulent proteins are expressed, and may be involved in host-pathogen interaction."

6. See the statement, "It is possible that certain miRNA (miRNA-A) that have multiple binding sites in tandem with another miRNA (miRNA-B) binding site in the region influences the biogenesis of miRNA-B that has fewer binding sites in tandem with miRNA-A." (Page 158 of Chapter 3). This can be improved to make it clear.

Correction: The statement on page 158 has been simplified (as cited below).
"It is possible that miRNA (miRNA-A) which has binding sites in tandem with the binding site of another miRNA (miRNA-B) may influence the biogenesis of miRNA-B. In this scenario, miRNA-A has larger number of binding sites than miRNA-B."

7. In chapter 3, "coordinate biogenesis" is referred to as phenomenon as opposed to hypothesis or as established biological phenomenon.
   [Previous version: "In order to validate this phenomenon wherein miRNA can influence mRNA expression by regulating the biogenesis of intermediate miRNA, this computational analysis has been performed. This phenomenon termed as "coordinate biogenesis" by our collaborator, has been the subject of this investigation. A set of miRNA that are likely to exhibit co-ordinate biogenesis has

been determined and the possibility that this phenomenon occurs in general has been explored. In particular, the probable miRNA co-ordinate biogenesis regulatory network for multiple miRNA in *Homo sapiens* and *Mus musculus* has been determined initially. Subsequently, this phenomenon has been studied in detail considering the miR-146a-5p network in detail by identifying regulator-target relationships therein (Figure 3.1) which have been studied in lipopolysaccharide exposed macrophages by our collaborator."
"However, the coordinate biogenesis phenomenon has been studied in detail considering the mouse mmu-miR-146a-5p network in detail by identifying possible CB regulator-target relationships in lipopolysaccharide exposed macrophages."
"Further, as a result of this phenomenon, a corresponding downstream effect in the known mRNA target (mRNA-B) of the secondary effector miRNA-B may also be observed."]

Correction: In chapter 3, according to the reviewer's suggestion, I have referred to "coordinate biogenesis" as a proposed hypothesis or biological phenomenon instead of a phenomenon, as applicable. The corrected version is mentioned below.
"In order to validate this hypothesis wherein miRNA can influence mRNA expression by regulating the biogenesis of intermediate miRNA, a computational analysis was performed. This hypothesis termed as "coordinate biogenesis" by our collaborator, has been the subject of this investigation. A set of miRNA that are likely to exhibit co-ordinate biogenesis has been determined and the possibility whether this biological phenomenon occurs in general has been explored. In particular, the probable miRNA co-ordinate biogenesis regulatory network for multiple miRNA in *Homo sapiens* and *Mus musculus* has been determined initially. Subsequently, this hypothesis was validated by identifying regulator-target relationships in the miR-146a-5p network (Figure 3.1). Further, some of the predicted regulatory relationships were studied in lipopolysaccharide exposed macrophages by our collaborator."

"However, the coordinate biogenesis hypothesis has been studied in detail considering the mouse mmu-miR-146a-5p network in detail by identifying possible CB regulator-target relationships in lipopolysaccharide exposed macrophages."

"Further, as a result of this biological phenomenon, a corresponding downstream effect in the known mRNA target (mRNA-B) of the secondary effector miRNA-B may also be observed."

8. A small description on alignment shuffling will help understand the step clearly in the methodology section of chapter 2.

Correction: As suggested by the reviewer, a small description of the term "Alignment shuffling" has been provided in chapter 2 methodology (page 151, as mentioned below).
In order to reduce the influence of phylogenetic relationships, alignment shuffling was performed by randomly selecting equal number of orthologous sequences (from the entire set of orthologs) of each family for the alignment. This ensures that amino acids

across each column exhibited variation during different runs of the Co-Var program [page 151].

9. It is more appropriate to use “Bioinformatics” rather than “Bio-informatics” in the Thesis. [Previous: “Bio-informatics analysis of expression profiling or sequencing analysis data: (Contents)”

“Bio-informatic approaches for sequence based or structural analyses of proteins (Contents)”
“Bio-informatics analysis of expression profiling or sequencing analysis data” (Page 7)
“Bio-informatics approaches for sequence based or structural analyses of proteins” (Page 11)]

Correction: According to the reviewer’s comment, the term “Bio-informatics” has been corrected to “Bioinformatics” throughout the Thesis (as mentioned below).
“Bioinformatics analysis of expression profiling or sequencing analysis data:”
“Bioinformatic approaches for sequence based or structural analyses of proteins”
“Bioinformatics analysis of expression profiling or sequencing analysis data”
“Bioinformatics approaches for sequence based or structural analyses of proteins”